SCUOLA DI DOTTORATO IN INGEGNERIA DELL'INFORMAZIONE
INDIRIZZO IN SCIENZA E TECNOLOGIA DELL'INFORMAZIONE, CICLO: XXIII

Design, Implementation and Evaluation of a Methodology for Utilizing Sources of Evidence in Relevance Feedback

Direttore della Scuola
CHIAR.MO PROF. MATTEO BERTOCCO

Supervisore
CHIAR.MO PROF. MASSIMO MELUCCI

Dottorando
EMANUELE DI BUCCIO

31 GENNAIO 2011

ABSTRACT

The objective of an Information Retrieval system is to support the user when he searches for information by predicting the documents relevant to his information need. Prediction is performed on the basis of evidence available during the search process. User interactions are examples of sources from which this evidence can be gathered.

This thesis addresses the problem of uniformly modeling heterogeneous forms of user interaction that are selected as sources for feedback. The problem of uniform source modeling is addressed by way of a complete methodology. The methodology aims at designing, implementing and evaluating a system that validates an experimental hypothesis. The hypothesis being validated regards the possible factors that can explain the user perception of relevance through the evidence gathered from the user interaction. The objective is to obtain and exploit a usable representation of the factors in the role of a new dimension of the information need representation.

The methodology aims at being general and not tailored to a specific source. The methodology defines the set of steps needed for obtaining a vector subspace-based representation of the information need dimensions to further exploit this representation for relevance prediction purposes. The set of steps identified are source selection, evidence collection, dimension modeling, document modeling and prediction.

This thesis shows how the methodology can be used for modeling two sources of evidence: term relationship in documents judged as relevant and the relationship between interaction features gathered from the behavior of the user when interacting with a set of documents. As for the term relationship dimension, this thesis shows that the current implementation of term relationship is feasible with a very large text collection delivered within the 2009 and 2010 Relevance Feedback tracks of the Text Retrieval Conference initiative. The methodology has supported the evaluation of term relationship for document re-ranking. As for interaction feature relationships, this thesis investigates the adoption of the user behavior dimension for document re-ranking both without query expansion and with query expansion.

SOMMARIO

L'obiettivo di un sistema di reperimento dell'informazione è quello di supportare l'utente in cerca di informazioni predicendo quali documenti siano rilevanti per la sua esigenza informativa. La predizione di rilevanza è effettuata sulla base dell'evidenza disponibile durante il processo di reperimento. Le interazioni che coivolgono l'utente sono esempi di sorgenti di evidenza.

Questa tesi affronta il problema della modellazione uniforme di forme eterogenee di interazione utilizzate come sorgenti di retroazione. Il problema della modellazione uniforme delle sorgenti è affrontato mediante l'introduzione di una metodologia, finalizzata alla progettazione, la realizzazione e la valutazione di un sistema per validare ipotesi sperimentali. Le ipotesi riguardano i possibili fattori che possano spiegare la percezione di rilevanza dell'utente sulla base dell'evidenza ottenuta da interazioni che coinvolgano l'utente stesso. L'obiettivo è quello di ottenere una rappresentazione dei fattori che possa essere utilizzata come una nuova dimensione della rappresentazione dell'esigenza informativa.

La metodologia si propone di essere generale e non specifica per una particolare sorgente. Essa definisce una serie di passi necessari per ottenere una rappresentazione in termini di sottospazi delle dimensioni della rappresentazione dell'esigenza informativa per poi utilizzare tale rappresentazione al fine della predizione.

La tesi applica la metodologia per modellare due sorgenti di evidenza: le relazioni tra i termini nei documenti giudicati rilevanti e la relazione tra attributi utilizzati per caratterizzare il comportamento dell'utente durante l'interazione con i documenti. In merito alla relazione tra i termini questa tesi mostra come la attuale implementazione per questa sorgente possa essere utilizzata per effettuare il reperimento su collezioni molto ampie, in particolare quelle adottate nelle campagne di valutazione dell'iniziativa Text Retrieval Conference, nello specifico nelle track di Relevance Feedback tenutesi nel 2009 e nel 2010. La metodologia ha consentito di supportare la valutazione del riordinamento dei documenti basato sulle relazioni tra i termini. In merito

alle relazioni tra attributi per caratterizzare il comportamento dell'utente questa tesi investiga l'utilizzo di una dimensione basata su tale sorgente per effettuare un riordinamento dei documenti sia unicamente basato sul comportamento, sia mediante espansione dell'interrogazione.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER

ONE

INTRODUCTION

1.1 Background and Motivation

The objective of an Information Retrieval (IR) system is to support the user when he searches for information. The user typically interacts with an IR system because he perceives a lack of information on a topic, necessary to address a problem or accomplish a task. The lack of information is the gap between what the user knows and what the user wants to know [Belkin et al., 1982]. This gap can be filled through the accumulation of knowledge, supported by the system providing the user with informative resources that he can perceive as relevant to his information need. Therefore the output of the process should be all and only the relevant informative resources. The output is usually a ranked list of results: indeed, optimal retrieval can be obtained when informative resources are ranked according to their probability of being relevant to the user or according to a score that preserves this ranking [Robertson, 1977]. The basic assumption underlying this principle, named Probability Ranking Principle (PRP), is that (i) the relevance of a document to a request is independent of the other document in the collection and (ii) the usefulness of this document may depend on the number of relevant documents the requester has already seen.

But on the basis of what kind of information can the system make predictions? In other words, what is the input of the prediction process? The prediction is based on the representation of the user information need and the informative resources. Both these representations are problematic. Informative resources can be heterogeneous, e.g. because of difference in terms of media, presence or absence of structure in the data, they can be aggregates or groups of informative resources.

The representation of the information need is crucial since it concerns the user perception of relevance with regard to his current need. In other words, the repre-

sentation of the information need should put in a usable form what the user wants to know (usable for the system in order to perform prediction). But which information available during the search process can be exploited to obtain the information need representation?

Possible evidence can be gathered from interactions that involve the user, informative resources that help define the context where search is performed, e.g. location [Göker and Myrhaug, 2008], time, or task information [Dragunov et al., 2005, White and Kelly, 2006]. Other informative resources could involve personal data, e.g. email or local collections [Teevan et al., 2005].

Information gathered from user interactions is the most adopted. The formulation of the information need as a query is among the possible evidence can be gathered from the interaction between the user and the system. The user can submit an informative resource, a sample of it, or a description of the problem he is addressing or the information he wants to know to accomplish his search task. In the remainder of this section we will discuss how diverse forms of interactions can help the system perform prediction and which issues should be addressed when dealing with the various kinds of information provided by these interactions.

Text-based Information Need Representation

Textual descriptions, hereafter named textual queries or more simply queries, are a well known instance of information need formulation. Most IR tools provide access to informative resources through graphical interfaces that consist of a text box where the user can specify a description of his need. The evidence that can be gathered from these representations, if considered in isolation, is basically the set of descriptors, namely terms, adopted by the user. Some of the most successful approaches, e.g. [Robertson and Zaragoza, 2009], are based on the statistical information of those descriptors to weight query terms in the informative resources, namely textual documents. But in some circumstances these representations may be not sufficient.

One cause can be the difference between the descriptors adopted by the user and those adopted by the author of the informative resources to explain the same concepts, or the descriptors to which these informative resources are associated by the system. This issue has been addressed using, for instance, both document and corpus-wide statistics, external sources, or combination of these to capture possible dependencies among terms, to address the synonymy and polysemy, or in general to refine the textual description of the information need provided by the user.

Query refinement is not necessarily only required because of the difference in the adopted descriptors. Being query formulation a cognitively demanding process it can result in "compromised" [Taylor, 1968] description. Besides being an approximation

because of the inherent loss of information due to the formalization and the expression of the need, in some circumstances the descriptive capability of the query is also affected by the anomalous state of knowledge [Belkin et al., 1982] of the user at the initial stage of search, when he needs to describe what he does not know and he is actually looking for. Finally, the query does not necessarily provide an exhaustive description of the user need. Prediction based on the query allows the user to be provided with informative resources that are topically relevant to his need. But the user perception of relevance is not necessarily explained only by topicality. Other variables can affect the way a user perceives a document. In principle the system should consider all these variables, model them and explicitly exploit them in the prediction process.

What kind of sources can provide additional evidence to model other variables in order to obtain a more exhaustive representation of the user information need? Query formulation, even if the most adopted evidence gathered from user interaction, is only an example of the evidence that can be gathered during the search process. The search process is indeed rarely limited to a single search episode since the user rarely retrieves the information he seeks in response to the first prediction [van Rijsbergen, 1986]. Therefore the user can be engaged in additional interactions with the systems, e.g. for query reformulation. But the system could also assist the user before a new query is issued, specifically on the basis of the evidence gathered in the post-search activities, e.g. examination of the results or rating explicitly provided by the user. The next sections will discuss how these kinds of interaction can be adopted to support prediction.

Exploiting User Interactions

In [Rocchio, 1966] the author discussed the problem of achieving the optimal query formulation, namely finding the query able to distinguish relevant from non relevant documents. Rocchio, on the basis of this idea, defines a formula to obtain this separation. The problem is that this formula requires information of all the relevant and non relevant documents in the collection, which is actually the target of the prediction process. Rocchio states that this kind of circularity suggests a strong analogy to feedback control theory. A possible approach is to ask the user to assess the retrieved documents or a subset of them. This approach is based therefore on the assumption that, even if not able to describe his need, the user is able to recognize a document that could be useful to satisfy his need, namely a relevant document. The relevant document set and the non relevant document set can be adopted as an error signal in the feedback process. The next input provided to the system is constituted both by the error signal and the original query. The objective is to obtain a new query, i.e. a new input,

such that the output of the prediction process is more closer to what the user desires. The separation between the relevant and non-relevant set can be achieved in multiple steps, where the convergence depends both on the goodness of the query issued by the user and the effectiveness of the refinement process. Basically, the first query, i.e. the approximate information need description, here aims at locating a region of the index space that should contain relevant documents. The user is directly involved after each prediction and is asked to indicate relevant documents in the result list. This type of approach is known in IR as Relevance Feedback (RF). The problem is that the user does not specify why he perceived the documents as relevant, i.e. which are the variables the system should consider in the prediction process. In the Rocchio algorithm the assumption is that important variables can be captured by modifying the textual description of the information need in order to be closer (in terms of representation, in this case vectors) to the relevant documents. But other variables should be taken into consideration and adopted for prediction.

Relevance information can be adopted in many ways. A possible approach is to learn directly ranking functions from the judged documents, basically exploiting them as training set. This approach has been shown to be effective for combining multiple document features and consequently explicitly considering multiple variables. The problem is that these approaches cannot support the investigation of the hypothesis on the variables that can affect the user perception of relevance. In principle, the approach to adopt should be, given a hypothesis, to obtain a usable model for it and include this model directly in the prediction process. This is, for instance, the basic rationale behind the Language Modeling (LM) framework, specifically Probabilistic Distance Retrieval Models [Zhai, 2008]. In this framework the query and the document are supposed to be generated from hidden models and prediction could be done on the basis of the comparison between the query and the document model. Feedback data can be explicitly included in this framework [Zhai and Lafferty, 2001]. Different variables can be introduced by considering a query model as a mixture, where each constituting model describes the impact of a hypothesis. This is, for instance, the case when modeling dependencies between query terms [Bai et al., 2005] or combining different query contexts [Bai and Nie, 2008]. Basically, the LM framework provides a principled approach for combining multiple evidence that can explain multiple variables from diverse hypotheses.

But query (re)formulation and explicit feedback are only a subset of the possible interactions that involve the user during the search process. Even when the user is in a familiar information space, i.e. his personal document collection, there is evidence that he tends to issue a first, usually short, query and then browse the information space [Teevan et al., 2004]. The keyword search, namely prediction based on a textual

query, is therefore more a tactic in a more complex search strategy than an actual strategy. The point here is that the system should be able to understand the user perception of relevance also from other forms of interaction and more specifically from the evidence that can be gathered from them. Browsing activities and post-search navigation activities are an instance of this kind of interaction. These activities can be considered as specific instances of a more general form of interaction: the behavior of the user when interacting with the documents.

During the last decade interest in user behavior has increased in the IR community. This interest is mainly due to the possibility of gathering large amount of evidence without direct involvement of the user. Post-search navigation features, for instance, can be gathered by unobtrusively monitoring the behavior of the user when interacting with the results returned by the system. If these features can provide information on the user perception of a document with regard to his current need, they could be adopted as cost-less evidence as input prediction. This information can be adopted to improve the information need representation, e.g. supporting query expansion [White and Kelly, 2006] or to estimate query–term probabilities [Bilenko and White, 2008]. The latter approach, for instance, exploits display-time to improve the textual representation of the information need, since they are used to estimate the probability of a term in a document using statistics of documents in the search trails and display-time as a weight. These kinds of approaches can be included in a probabilistic framework.

Even if the above approaches are based on diverse types of interactions, they all refer to strategies that exploit textual representation both of the user need and the informative resources. But textual descriptions are only a subset of the possible ways to characterize both the information need and the informative resources. Indeed, features that can be gathered from user interactions can be directly adopted as descriptors for informative resources instead of terms.

Beyond Text-based Representation

The representations the system can process to perform prediction do not necessarily need to be based on a textual description of the information need. For instance, in [Vassilvitskii and Brill, 2006] ratings provided by the users on a subset of the top ranked web results are adopted to re-rank web pages by exploiting web graph distance between documents. But this approach involves document-specific properties. Representations do not need to be restricted to these properties; they can be more general than that. This is, for instance, the approach adopted in [Agichtein et al., 2006b] where each document is described as a vector of post-search navigation features. Document observations are used as evidence from which, provided a set of explicit judgments

on document pairs, a ranking function can be learned. But these kinds of approaches rely on learning function, not learning hypothesis.

The adoption of different representations based on diverse descriptors obtained from diverse sources can be beneficial for addressing the problem of capturing the different variables can affect the user perception of relevance. This is suggested both by theoretical and experimental results in the IR literature as discussed in the next section.

Combining Evidence

The principle of polyrepresentation [Ingwersen, 1996] is based on a cognitive approach to IR and its basic rationale is that "overlaps between a variety of contexts associated with the interactive IR process can be exploited to reduce the uncertainty and thereby improve IR performance" [White et al., 2009]. Interaction context, concerning the evidence of interaction behavior during the search session, is one of these contexts. White et al. consider diverse sources, including interactions described by search trials, to support web site recommendation, not for retrieval of search results, and showed that the overlap outperformed the single sources. The work reported in [Agichtein et al., 2006a] learning the ranking functions from both document-specific features and behavior derived features show that the latter features are able to substitute hundreds of features used by commercial search engines to support prediction.

Earliest results when considering only document-specific evidence showed that combinations both of approaches or diverse document representations can provide more effective results [Croft, 1999]. Croft interpreted the problem of combining evidence and approaches as a problem of combining classifiers. In order to obtain best retrieval performance by combination, each classifier (retrieval algorithm or system) should be as accurate as possible and classifiers that are combined should be uncorrelated. When using different representation and retrieval algorithms, classifiers are more likely to be independent.

In general it seems to be crucial to investigate relationship between the evidence provided by the diverse sources, for instance diverse forms of user interaction, before combining them.

1.2 Research Statement

The main issues that emerge from above can be summarized by the two following questions: on the basis of what kind of information can the system make the prediction and how should the system utilize and combine these various kinds of information. These are actually the two research questions Robertson in [Robertson, 1977] identified

to be central in the IR problem and every approach should ultimately address. In the specific context of the interactions involving the user, the problem is to investigate hypotheses on the possible information which can be extracted from the gathered evidence and then used for prediction. What is required is methods and frameworks both able to support the investigation of the hypotheses on the possible variables which can affect the user perception of relevance, and to exploit models derived from these hypotheses for prediction. When considering combination, the diversity among the sources should be addressed. Indeed, it is one of the major sources for complexity both because the need to handle representation based on heterogeneous features and the need to compare their contribution in order to capture possible relationships should be explicitly taken into account before combination.

The problem of diversity in this thesis is addressed by obtaining a uniform representation among the diverse sources for evidence, in order to obtain a diverse but uniform representation of the information need based on diverse hypotheses. The diversity among the hypotheses is due to the fact that the variables that can provide information on the user perception of relevance when a source is considered, e.g. content-based properties of explicitly judged documents, can be different from that of another source, e.g. when considering forms of interactions like user behavior. Since the final objective of the retrieval process is to perform prediction, those representations need to be usable by the IR system. The uniformity among the representation could allow them to be exploited through a unique ranking function.

Since, as discussed in Section 1.1, the IR process is usually not limited to a single search episode, the forms of interaction that involve the user after a first search, e.g. feedback explicitly or implicitly provided by the user, are possible sources of evidence that can actually be exploited to support the user before a new query formulation. Let us consider the scenario of a user who having issued a first query, obtains a list of results and interacts with some of them, e.g. by providing judgments or examining them. The research question is:

How can the evidence provided by diverse forms of interactions involving the user be uniformly modeled and exploited for feedback through a unique ranking function?

In this thesis we will show how a methodology based on a common abstraction of the diverse sources can help achieve this objective and moreover can assist the design and the development on an IR system able to exploit and evaluate the effectiveness of the diverse source contributions. The methodology achieves the objective of the uniform representation by identifying a set of steps to exploit the geometric framework originally proposed in [Melucci, 2008] for a generic source. The basic rationale

of this framework is to exploit the mathematical construct of the vector space basis to model the contribution of a source and a projector-based ranking function, originally proposed in [van Rijsbergen, 2004], to rank informative resources according to their distance from the subspaces spanned by the basis and that models the information need representation corresponding to the source. The objective of the methodology, in terms of modeling and exploiting sources, is indeed to start from the common abstraction, obtain a usable representation of the sources as vector subspaces (those spanned by the basis) and exploit them for prediction through the projector-based ranking function. The next section will specifically describe the diverse contributions of this thesis.

1.3 Contribution

This thesis addresses the problem of uniformly modeling diverse forms of user interaction adopted as sources for feedback.

In order to achieve this objective we will:

- define a common abstraction for the diverse sources;
- define a methodology to obtain a usable representation of the sources through a geometric framework;
- investigate two methodology applications for two diverse sources, e.g. the behavior of the user when interacting with the results and the relationship between terms, modeled by local co-occurrence in the documents judged as relevant; these applications will be evaluated using experimental test collections.

In the remainder of this section we will briefly discuss each of the above points.

Abstraction

In this thesis we will define a common abstraction both for sources and informative resources. We will show that this abstraction is not only functional to the methodology but actual systems can be designed and implemented to support retrieval of informative resources at diverse resource levels (e.g. in a distributed architecture where the entire collection is constituted by collections, set of collections and possible set of collection sets) and for diverse media (e.g. text and music). Two systems have been developed on the basis of this abstraction: SPINA to address the problem of documents of diverse media distributed in a Peer-To-Peer (P2P) network, and FALCON, an open source search engine for cover song identification.

Methodology

We will identify a set of steps to model the contribution of a generic source on the basis of the gathered evidence and use this model as input for a new prediction. The steps are: (i) source selection, (ii) evidence collection, (iii) dimension modeling, (iv) document modeling, and (v) prediction. The *source selection* step consists of the selection of the source from which the feature values are distilled. The *evidence collection* step consists of the actual collection and selection of the features from the selected sources. The evidence gathered in the previous step is then adopted as input for the *dimension modeling* step. Here the term *dimension* is adopted to denote a usable representation of the way the source can contribute to explain the user perception of a document with regard to his need. The objective of the *document modeling* step is to obtain a representation of the document in terms of the source features. Once a representation has been obtained both for dimensions and documents, the last step consists of the prediction based on those representations. These steps are actually those constituting the proposed methodology.

This thesis explains how the methodology steps are considered with regard to the selected geometric framework. The dimension modeling step consists in the computation of a vector space basis by exploiting evidence gathered from the corresponding source. Document modeling corresponds to obtaining a vector representation for the documents to re-rank and prediction can be performed measuring the distance between the vector and the subspace spanned by the computed basis.

Methodology Applications

In this thesis the designed methodology is applied to two specific sources: term relationship in the feedback documents and behavior of the user when interacting with the results.

Term Relationship in Documents Judged as Relevant. Term relationships are modeled using local co-occurrence of terms appearing in the feedback documents. Among the steps constituting the methodology, the work reported in this thesis is mainly focused on source selection, evidence collection, and document modeling. The problem of source selection is investigated by varying the prediction strategy adopted in the first prediction, thus varying the document in the feedback set. The problem of evidence collection is investigated by paying particular attention to the term selection process. Indeed, not all the terms appearing in the feedback set are good descriptors of the user information need. Selecting all the terms in the feedback documents increases the chance of considering good terms, but simultaneously increases that of considering useless and potentially harmful terms. In contrast, selecting a subset of good terms

can be beneficial both in terms of effectiveness and efficiency. Several term selection strategies are adopted to address this issue. Finally, the problem of document modeling is investigated using diverse weighting schemes.

The application of the methodology to term relationships has been experimentally evaluated by the participation in IR evaluation campaigns, specifically the Relevance Feedback Track of the Text REtrieval Conference (TREC) both in 2009 and 2010. The participation in the TREC campaigns required the development of an experimental system to manage large test collections: the corpus adopted in the RF Track was constituted by fifty million web pages. The system developed and used for the experiments reported in this thesis extends the SPINA software architecture that provides a tool to support experimental evaluation.

User Interaction Behavior. The second source considered is the behavior of the user when interacting with documents obtained as results of a first query formulation. The behavior of the user is described in terms of post-search navigation features, e.g. the time a user spent on a document or scrolling activities performed when examining it. In particular, in this thesis the evidence gathered after visiting several documents is adopted for dimension modeling. A user behavior dimension is obtained by extracting possible behavioral patterns from the collected data. The modeled dimension is adopted to re-rank documents uniquely using the user behavior dimension, or to support query expansion by extracting terms from the top documents re-ranked by user behavior.

With regard to this source, the work reported in this thesis is specifically focused on the source selection step of the methodology. In particular individual users and groups of interrelated users searching for the same query are considered as possible sources for post-search navigation features. This issue is investigated because modeling a dimension tailored for each user can help address the problem of the variety of user intents and needs that is one of the motivations for the work reported in this thesis. In principle the evidence from the user who is interacting with the system should be considered when modeling user behavior. However, personal evidence is often unavailable, insufficient, or unnecessary. For this reason in this work the evidence gathered from the group whose users search for information useful for meeting similar requests is investigated as an alternative to personal evidence.

Because of the lack of available test collections where both interaction data and document specific features are present, a user study was carried out; the resulting test collection is that adopted in the experiments concerning the source user behavior.

1.4 Thesis Overview

This thesis is organized as follows.

- **Chapter 2** discusses previous works that exploit feedback strategies in IR. It is mainly focused on the two sources investigated in this work when considering specific applications of the methodology. The first source is possible relationships between terms in document judged as relevant. After a brief introduction to the general techniques that aim at modeling term relationships also in a non feedback scenario, the discussion is focused on techniques that model term relationships through vector space-based representations and the difference with this thesis where a geometric framework is adopted for modeling this source too. The second source concerns another form of interaction, i.e. the behavior of the user during the search process. The discussion is focused on two issues. The first issue is the predicting capability of features gathered from user behavior, i.e. whether behavioral features considered in isolation and combined can provide information on the user interests. The second set of works discussed concern approaches to include and therefore exploit behavioral information in the prediction process. The final part of the chapter discusses previous works that investigated methodologies to support evaluation of feedback strategies in the prediction process.

- **Chapter 3** introduces the methodology which is the central part of the thesis. First a set of definitions are presented. These definitions constitute the abstraction on which the methodology is based. In particular, two notions are discussed: the notion of factor and the notion of dimension. These notions concern the way the evidence gathered from a source can be adopted to support prediction. This chapter discusses how a geometric framework can be adopted to model source contributions to support prediction. The methodology is introduced at the end of the chapter and explains how, starting from a hypothesis of possible variables that affect the user perception of relevance, it is possible to uniformly model source contributions and informative resources by the geometric framework and exploit them for feedback.

- **Chapter 4** discusses two possible applications of the methodology to two diverse sources: term relationships in documents judged as relevant and behavior of the user when examining the results. A specific implementation is discussed for each methodology step. Some issues which should be addressed when considering the two applications are pointed out. These issues are the subject of the experimental investigation reported in the subsequent chapter.

- **Chapter 5** introduces the specific research questions addressed for the considered methodology applications. Then the evaluation methodology for each application is presented together with the adopted test collections and the experimental system developed on the basis of the abstraction for supporting the investigation of the posed research questions. This chapter also presents the test collection including user interaction behavior, which is one of the contributions of this thesis.
- **Chapter 6** describes the results obtained for the research questions concerning the two methodology applications.
- **Chapter 7** reports concluding remarks and discusses avenues for future work.

CHAPTER

TWO

SOURCES OF EVIDENCE TO SUPPORT FEEDBACK IN INFORMATION RETRIEVAL

The objective of feedback strategies is to improve the information need and the informative resource representation on the basis of the evidence gathered from user interactions during the search process. The first formalization of feedback is usually credited to Rocchio [Rocchio, 1966, Rocchio, 1971]. The idea was to ask the user to provide explicit judgments on the retrieved documents set and refine the query accordingly. The evidence gathered from the interaction can be adopted to modify the original input to the system: the original input is here the query and the evidence gathered from the user interaction can be interpreted as an error signal in the feedback process. The next input provided to the system is constituted both by the error signal and the original query. Rocchio investigated this approach with regard to a vector representation of document and queries. But the idea was general and other works investigated this strategy in different models.

Feedback can be adopted for different objectives. It can be adopted for query modification, as in Rocchio, i.e. to expand the query, to modify query term weights, or both. Feedback data can be adopted to estimate probability distribution or as evidence to directly learn ranking functions. But the main challenge when considering feedback data is to understand which are the factors that affect the user perception of relevance, since the user does not explain why the document is relevant.

Feedback does not only consist of explicit judgments provided by the user. Other forms of interactions can be adopted. An example is the evidence gathered from user browsing activities from which interaction features such as click-through data or display-time can be observed. The main challenge in this case is what kind of information these features can convey on the user intents. Strategies exploiting this

kind of evidence for feedback are known as Implicit Relevance Feedback (IRF) techniques [Kelly and Teevan, 2003].

In this chapter we will review some previous works in explicit and implicit feedback, since the two applications of the methodology introduced in this thesis concern these two feedback strategies. At the end of this chapter some works will be discussed that concern the adoption of methodologies or conceptual models to support the design or the evaluation of approaches that exploit feedback data.

2.1 Document-specific Properties

In [Rocchio, 1966] feedback strategies were introduced to achieve an optimal query formulation to distinguish relevant from non relevant documents. The ideal query is the one for which the system is able to rank all the relevant documents at higher ranks than those perceived as non relevant for achieving the user information goal. But the ideal query cannot exist. Indeed, the prediction is a function of the query and document representation, where documents are passed through an indexing procedure whose purpose is to reduce information rather than preserve it. The user perception of relevance is based on a function of the query and his perception of the documents. On the contrary, the optimal query is defined on representation. In that work documents as well as queries are represented as vectors, where each element refers to an indexing unit, e.g. a term. Rocchio defines a cost function to operationalize this concept:

$$C = \frac{1}{|D_R|} \sum_{d_i \in D_R} \rho(q, d_i) - \frac{1}{m - D_R} \sum_{d_i \notin D_R} \rho(q, d_i) \tag{2.1}$$

where m is the number of documents in the collection, D_R is the set of relevant documents, q denotes the query, and ρ denotes a query–document correlation function, e.g. cosine correlation.

The problem is that to obtain the optimal query, the system should know all relevant and non-relevant documents, which is in fact the objective of the prediction. Rocchio states that this circular dependence is similar to feedback in control theory. A possible approach is indeed to ask the user to assess the retrieved documents or a subset of them. The relevant document set and the non relevant document set can be adopted as an error signal in the feedback process. The next input provided to the system is constituted by both the error signal and the original query. The objective is to obtain a better query that is optimal in distinguishing between relevant and non relevant documents in the sample. The separation between the relevant and non-relevant set can be achieved in multiple steps, where the convergence depends both on the goodness of the query issued by the user and the effectiveness of the refinement process. The first query here aims at locating a region of the index space that should contain relevant

documents. Note that Rocchio explicitly discussed the problem of relevant documents in diverse regions — the approach is based on the assumption that relevant documents tend to cluster. If relevant documents are in different regions, then an approach based on a document-document correlation matrix can be adopted to identify diverse clusters and then two queries can be issued instead of one — the complexity is the same that a further refinement is a subsequent iteration. When multiple iterations are adopted, Rocchio also investigated the difference between refining the initial query or refining the last modified query. The latter approach on average performed better than the former, but when analyzing individual queries, some of them suffer drastically when exploiting refinements instead of the original query as starting point. The way in which the query could be refined is: $q' = \alpha_1 q + \alpha_2 C^*$ where C^* is the optimal score obtained by estimating C using feedback documents and the two parameters α_1 and α_2 denote the degree to which the initial weight should be modified (e.g. they could be a function of the amount of feedback, starting from the idea that the estimation of C could be more reliable when the size of the sample increases).

The Rocchio feedback algorithm was further investigated in subsequent works that analyzed possible variations. In [Buckley et al., 1994] the authors exploit a modified version of the Rocchio formula where the documents considered non relevant are all the unseen documents and those explicitly judged as non relevant by the user. Buckley et al. state that, even if this assumption is actually false, it might be better than the hypothesis of the original Rocchio formula where the documents judged as non relevant are representative of non-relevant documents in general. In that work they performed an in-depth investigation of the effect of the number of relevant documents used for feedback and the number of terms for query expansion. The experimental collection adopted was the TREC routing environments where the corpus was divided into a training set and a test set corpus. The baseline was SMART that exploited the *ltc* weighting scheme for queries and the *lnc* weighting scheme for the documents. Since relevant documents in the training set were obtained by runs of diverse retrieval systems, the authors investigated the effect of using only relevant documents in the top retrieved by SMART, and all the relevant documents or a randomized subset ($\frac{1}{4}$ of the whole relevant set). Several runs were performed varying the number of relevant documents in the feedback set and the number of expansion terms. A log linear relationship was found between the average recall-precision and these two variables. Using relevant feedback set including also documents from other systems were shown to be beneficial, but the improvement was not substantial, i.e. 3 - 4%. In this thesis we will explicitly investigate the effect of diverse feedback set in the considered methodology application. One of the reasons for this choice is that possible strategies for diversifying the feedback set could be beneficial for considering diverse aspect of relevance,

e.g. the diverse regions suggested by Rocchio.

The approach adopted in this thesis to obtain a new information need representation through feedback differs from Rocchio strategy. In [Efron, 2008] the author discusses the notion of optimality in the Rocchio formula. The approach proposed by Rocchio is based on an interpretation of IR as a two-class classification problem, where the objective is to "separate" relevant from non relevant documents: Rocchio formula is optimal to reduce the risk of misclassification. The approach adopted in this thesis, as discussed in the following section, shares the basic rationale underlying Latent Semantic Indexing (LSI) [Deerwester et al., 1990] and can be considered in the context of a linear regression interpretation of the IR problem [Story, 1996]. The objective in this case is improving the regression model by explicitly considering information on the covariance structure of the data.

Many works concerning feedback strategies and exploiting document-specific properties have been proposed in the IR literature since [Rocchio, 1971]. Alternative feedback strategies have been introduced, e.g. Pseudo-Relevance Feedback (PRF), where the top ranked documents are assumed to be relevant and feedback is performed on the basis of their properties, thus making feedback possible also when no explicit judgments are available. A survey on feedback strategies is reported in [Ruthven and Lalmas, 2003], where a wide range of approaches, ranging from explicit and pseudo feedback to interactive query modification, are discussed. In this chapter we will focus on previous approaches that exploit a specific property of documents in the feedback set: relationship among terms. Indeed, one of the methodology applications discussed in this thesis exploit this source of evidence. After a brief review of previous works modeling term relationship, next section will focus on those strategies that exploit vector space-based representation and model relationship among terms in feedback documents.

Term Relationship

The first application of the methodology investigated in this thesis concerns the relationship between terms. Most approaches in IR are based on the hypothesis that the occurrence of a term in a document is independent of the occurrence of the other terms, e.g. unigram language models. The Binary Independence Retrieval (BIR) [Robertson and Sparck Jones, 1976] model is also based on the assumption of statistical independence between terms, even though the arguments reported in [Cooper, 1991] show that it is actually based on a weaker assumption, namely the *linked dependence* assumption; as stated in [Gao et al., 2004], this can be one of the reasons why it performed better than further approaches proposed to explicitly include dependency information. Indeed, since the independence assumption does not

necessarily hold, several approaches have been proposed to capture term dependencies. One of the earliest approach was introduced in [van Rijsbergen, 1979] where the author investigates dependence trees to model term dependencies. The final objective is to compute a joint probability distribution by explicitly including dependencies. Since considering all dependencies is not feasible, the author proposed computing a decomposition of the joint probability where each term is conditioned by only one of the other terms. Since many decompositions are possible and are all equivalent, the objective is to find the one that best approximates the dependence between two terms. The measure or dependence is the Expected Mutual Information Measure (EMIM)[1]. Each possible decomposition is associated with a dependence tree whose nodes are terms t_i's and the edge between two terms t_i and t_j is weighted by $I(t_i, t_j)$. The objective is then to find the best dependence tree, which actually corresponds to finding the Maximum Spanning Tree (MST). But this approach is still computational expensive and did not provide promising results. Spanning trees were further used to capture term dependencies at the sentence level in [Nallapati and Allan, 2002] to address task of topic detection and tracking, more specifically capture possible links between stories and event. The proposed approach is to consider sentences as semantic units and hypothesize independence between sentences rather than independence between terms. A language model is computed for each sentence in a document; dependence trees are adopted to compute the dependency among terms occurring in the sentence, where dependency are measured by the Jaccard Coefficient. Even if the computational complexity for computing the MST is reduced through a greedy algorithm, the approach is still more computationally expensive than the language modeling approach, e.g. the unigram language model adopted in that work as baseline. The proposed approach alone is not able to improve performance, while a slight improvement is obtained when combined with the unigram language model, thus indicating that considering the sentence level provides additional useful evidence.

A generalization of language models able to include term dependence was proposed in [Song and Croft, 1999]. The idea is to consider a query as a sequence of terms, thus being able to include local context too, e.g. including information on bigrams in addition to unigrams. A bigram here is considered a sequence of terms, therefore the order of occurrence is explicitly taken into consideration. The basic rationale when considering bigrams is to substitute the hypothesis that terms are statistically independent with the hypothesis that each is statistically dependent on the one that precedes it. A model is associated with each bigram as well as with a unigram; therefore the two models can be combined, e.g. through interpolation. The model obtained by

[1] For two terms t_i and t_j, the degree to which they deviate from independence is measure by $I(t_i, t_j) = \sum_{t_i,t_j} P(t_{t_i,t_j}) \log[P(t_{t_i,t_j})/(P(t_i)P(t_j)])$

the combination of unigram and bigram models outperforms the Ponte-Croft LM on the Wall Street Journal test collection; but no significant improvement in the TREC4 dataset. A variant of this approach was proposed in [Srikanth and Srihari, 2002] where biterm are adopted, i.e. two terms co-occurring near each other are considered as an unordered pair. The bigram model has been shown to be a special case of a more general dependence model proposed in [Gao et al., 2004]. In this work a query is generated from a document in two steps, through a hidden dependency structure named linkage. The linkage assumes that term dependencies form an acyclic, planar graph: two related terms are linked in this graph. Given a document, the linkage is generated from the document with a certain probability distribution; the query is then generated in a second step from the linkage, thus including also possible dependencies with other terms in the document according to the linkage structure. An unsupervised approach is adopted to learn linkage for each document. The main motivation behind this approach was to provide a way to also include dependency for terms not occurring near to each other, as in the bigram or biterm case, since relationships can be present between more distant terms; moreover, the linkage structure allows the most significant dependencies to be captured. A successful model able to include term dependencies was the Markov Random Field (MRF) model proposed in [Metzler and Croft, 2005]. It is a general discriminative model that can be adopted to combine scores or diverse document representations. A MRF is a graph whose nodes are random variables and the edges define interdependence between random variables; a random variable in the graph is independent of its non-neighbors given observed values from its neighbors. Different edge configurations imply different dependencies between variables. The author explored three possible configurations, where nodes are query terms and the document. In the independence configuration the query nodes are not dependent on each other, but only on the document. In the sequential dependence configuration each query term is dependent only on adjacent terms. In the full dependence model, all query terms can be dependent on each other. Once the graph has been defined, a set of potential functions over the cliques of the graph should be defined and finally documents are ranked by their posterior distribution of being generated from the query — actually an estimate that does not alter the rank since it is rank equivalent. The potential functions are a term function (estimated like in the unigram model), an ordered and an unordered potential function that check respectively the degree to which an ordered sequences of query terms occur, and the degree to which two or more terms appear in close proximity in a text window of size N. The full dependence mode was the most effective. With regard to the potential functions, the results showed that ordered functions were more effective on smaller collections, while in web collections e.g. WT10g and GOV2, there were no significant differences. A further extension of the

MRF model was successfully adopted to include feedback information [Lease, 2008].

These works show that explicitly model term relationship can be beneficial to improve the predicting capability of the IR system. Term relationships are therefore a potential and effective source to model and exploit for prediction. For this reason, in this thesis we will investigate the effectiveness of modeling term relationship in the introduced methodology. In particular, term relationships will be extracted from the content of documents explicitly judged as relevant by the user. The methodology exploits a geometric framework to model source contributions, particular by the mathematical construct of the vector subspace. For this reason, in the remainder of this section we will discuss previous works that exploit a vector space-based representation to model relationship among terms, particularly focusing on those that exploit feedback strategies.

Vector-space based Approaches for Modeling Term Relationship

The idea of explicitly including term relationship in the prediction process was one of the motivations for a generalization of the traditional Vector Space Model (VSM), i.e. the IR model based on the model for computation of the IR process introduced in [Salton, 1968]. In the VSM both query and document are represented as vectors. Query and document are characterized by their constituting terms, namely descriptors, $\mathcal{T} = \{t_1, \ldots, t_{|\mathcal{T}|}\}$ and each term t_i is represented as a vector $\mathbf{t}_i \in \mathbb{R}^{|\mathcal{T}|}$ where $\mathbf{t}_i = \mathbf{e}_i$ denotes the occurrence of term t_i and $\mathbf{e}_i$ is the ith vector of the canonical basis, i.e. the vector whose ith entry is 1 and all the other entries are 0. A query q is represented as a vector $\mathbf{q} = \sum_{i=1}^{|\mathcal{T}|} \alpha_i \mathbf{t}_i$, α_i's are the coefficients, e.g. frequency or weights of query terms; analogously document is represented as a vector $\mathbf{d} = \sum_{i=1}^{|\mathcal{T}|} \phi_i \mathbf{t}_i$. Prediction is performed by measuring the distance between vectors by means of their inner product $s(q, d) = \mathbf{d}^T\mathbf{q}$.

The Generalized Vector Space Model (GVSM) generalizes the scoring function as $s(q, d) = \mathbf{d}^T\mathbf{C}\mathbf{q}$ with $\mathbf{C} \in \mathbb{R}^{|\mathcal{T}|\times|\mathcal{T}|}$, where $\mathbf{C} = \mathbf{I}$ in the VSM. $\mathbf{C}$ contains information on the relationship among terms. If the relationship among terms is symmetric, then $\mathbf{C}$ is symmetric; if $\mathbf{C}$ is positive definite then it can be expressed as $\mathbf{C} = \mathbf{A}^T\mathbf{A}$ and $\mathbf{A}$ is unique[2]. Therefore $s(q, d) = \mathbf{d}^T\mathbf{C}\mathbf{q} = \mathbf{d}^T(\mathbf{A}^T\mathbf{A})\mathbf{q} = (\mathbf{A}\mathbf{d})^T(\mathbf{A}\mathbf{q})$. Basically, a descriptor t_i, namely a term, is no longer represented as a vector $\mathbf{t}_i$ of the canonical basis $\mathbf{I}_{|\mathcal{T}|\times|\mathcal{T}|}$, but as a linear combination of the columns of the matrix $\mathbf{A}$. The

[2] $\mathbf{C}$ can be decomposed as $\mathbf{C} = \mathbf{A}^T\mathbf{A}$ using Cholesky decomposition; this decomposition can also be applied to positive semi-definite matrix, but in that case the decomposition is not unique. Alternatively, if $\mathbf{C}$ is positive definite a LDU factorization $\mathbf{C} = \mathbf{T}\mathbf{D}\mathbf{T}^T$ exists where $\mathbf{D} \in \mathbb{R}^{n\times n}$ is a diagonal matrix $diag(\lambda_1, \ldots, \lambda_n)$. λ_i's are eigenvalues of $\mathbf{C}$ and are all positive since it is positive definite. Therefore, $\mathbf{C} = (\mathbf{D}^{1/2}\mathbf{T}^T)^T(\mathbf{D}^{1/2}T^T) = \mathbf{A}^T\mathbf{A}$ where $\mathbf{D}^{1/2}$ is a diagonal matrix whose ith diagonal element is $\sqrt{\lambda_i}$.

columns of $\mathbf{A}$ basically capture possible dependencies among terms. In this thesis the columns of $\mathbf{A}$ will be interpreted as possible factors that explain the observed data, in this case the correlation between terms expressed by $\mathbf{C}$.

But the generalization introduced by the GVSM is based on the assumption that both the query and the document have been generated by the same factors, specifically represented by the basis vectors constituting $\mathbf{A}$. In [Melucci, 2005] the author proposes further generalizing this approach. Query vector and document vector are not necessarily generated by the same factors: that can be modeled using different vector space basis $\mathbf{A} = [\mathbf{a}_1 \cdots \mathbf{a}_{|\mathcal{T}|}]$ and $\mathbf{B} = [\mathbf{b}_1 \cdots \mathbf{b}_{|\mathcal{T}|}]$ for the query and the document. Using the above scoring function, the degree to which a document d satisfies a query q is measured by $s(q,d) = (\mathbf{Ad})^T(\mathbf{Bq}) = \mathbf{d}^T(\mathbf{A}^T\mathbf{B})\mathbf{q}$, where $\mathbf{A}^T\mathbf{B}$ takes into account the fact that the observations have been generated by diverse factors.

A vector space-based representation of information need and informative resources that aims at capturing relationships among terms is LSI [Deerwester et al., 1990]. Let $\mathcal{D}$ be the set of documents in the collection. The term–document matrix $\mathbf{X} \in \mathbb{R}^{|\mathcal{T}|\times|\mathcal{D}|}$ is decomposed by Singular Value Decomposition (SVD) as $\mathbf{X} = \mathbf{T\Sigma D}^T$, where $\mathbf{T} \in \mathbb{R}^{|\mathcal{T}|\times r}$, $\mathbf{D} \in \mathbb{R}^{|\mathcal{D}|\times r}$, and $\mathbf{\Sigma} \in \mathbb{R}^{r\times r}$ where r is the rank of the matrix $\mathbf{X}$. The matrix $\tilde{\mathbf{X}} = \tilde{\mathbf{T}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{D}}^T$ where $\tilde{\mathbf{\Sigma}} \in \mathbb{R}^{s\times s}$, with $s < r$ and $\tilde{\mathbf{T}}$, $\tilde{\mathbf{\Sigma}}$, and $\tilde{\mathbf{D}}$ matrices constituted by only the first s components respectively of $\mathbf{T}$, $\mathbf{\Sigma}$, and $\mathbf{D}$, can be considered the best s-approximation of $\mathbf{X}$ in the least-square sense. Terms and documents can be represented as vectors in a s-dimensional space. Basically, terms are considered in a reduced space "in a way that reflects the correlations in their user across documents" [Hull, 1994]. A possible way to obtain a correlation matrix that express relationship among terms is to compute $\mathbf{C} = \mathbf{T}^T\mathbf{T}$. But LSI allows relationship to be considered in the reduced space: relationship among terms can be captured for instance considering the matrix $\tilde{\mathbf{C}} = \tilde{\mathbf{T}}^T\tilde{\mathbf{T}}$. In [Kontostathis and Pottenger, 2006] the authors investigate the values of the term–term correlation reduced matrix and their relationship with high order term co-occurrence. For instance, a second order co-occurrence between two terms A and C exists when a term A co-occurs with B and B co-occurrs with C. A strong correlation was observed between second-order term co-occurrence and values obtained by SVD.

LSI was adopted to support feedback in past works. In [Dumais, 1991] a single relevant document, the weighted average of three relevant documents or the centroid of the entire relevant set are investigated as possible information need representations to enhance the first stage prediction in the reduced space. The first two cases refer to the scenario where a user submits a query, obtains a list of results, and then examine some of them. In the first case prediction is done after the user indicates a relevant document, in the second case when three relevant documents are identified. The last approach cannot be adopted in practice since the relevance set is the target of the

prediction. All the approaches provided an improvement in terms of Mean Average Precision (MAP). In [Dumais, 1996] LSI is adopted to support information filtering. The author describe two filters: (i) a *word filter* where a single vector is obtained as the weighted sum of the term vectors obtained from the topic content, and (ii) a *reldocs filter* obtained as the centroid of the documents known to be relevant. Documents are ranked according to the distance from the filter vectors. Combinations of the two filters were investigated, e.g. weighted sum with diverse weights per filter; moreover, also fusion techniques were applied. Experiments were carried out in the TREC-3 filtering task. Weighted sum perform better than data fusion. When considering less than ten relevant documents, the word filter outperformed the reldocs filter.

A different application of LSI was proposed in [Hull, 1994]. This technique, named Local Latent Semantic Indexing (LLSI), consists of applying a new LSI on the matrix whose rows are the document known to be relevant and whose columns are the Here,factors obtained from the original LSI representation. This technique was evaluated in the context of the routing task performing cross-validation on the Cranfield collection. LLSI was exploited in combination with Discriminant Analysis to separate relevant from non-relevant documents. The combination of these two techniques, named Text-based Discriminant Analysis (TDA), outperformed both VSM and LSI. In that work SVD was applied on a term–document matrix whose entries were term weights, not term frequencies. Investigation of LSI using a term weight matrix for different weighting scheme is reported in [Dumais, 1991]. Unlikely [Hull, 1994] where the local region was constituted by relevant documents, in [Schütze et al., 1995] LLSI was applied to the 2000 nearest documents to the query, where the distance was computed as the inner product between the document vectors and the query vector representation obtained by Rocchio algorithm. The obtained set can be potentially constituted both by relevant and not-relevant documents (documents without judgments were considered not relevant). The descriptors obtained by LLSI on the local region was then adopted as input for learning algorithms. The best performance was obtained using a linear neural network with LLSI representations and 200 expansion terms obtained by χ^2-based measure of dependence.

The work reported in this thesis shares with LSI the adoption of SVD to compute a vector space basis to model term relationship, and with LLSI the adoption of a local approach, namely it considers only the entries corresponding to relevant documents. But the decomposition is not applied to a term–document matrix or the reduced matrix of the relevant set descriptors as in [Hull, 1994], but on a local term correlation matrix obtained by an approach similar to that adopted to compute Hyperspace Analogue to Language (HAL) spaces [Lund and Burgess, 1996]. The objective of HAL spaces is to capture term co-occurrence relationships explicitly including the proximity among

terms. A term-by-term co-occurrence matrix is obtained from the text corpus using a sliding window; all the terms co-occurring within the window are considered as related to each other. The strength of the relationship is inversely proportional to the distance between the terms and in the original proposal was directional: information on co-occurrence with proceeding terms was stored in the row corresponding to the term, while information on the co-occurrence with following terms was stored in the columns. When direction is not considered, e.g. in [Bai et al., 2005], the word vector can be obtained as the sum of the row and the column vector. Each term is therefore associated with a vector; terms related to it can be selected considering entries with weights above a certain threshold. In [Bai et al., 2005] HAL space was adopted for query model expansion in the LM framework. From each HAL vector only the weights over a threshold, set to the mean weight, were retained. Then each HAL vector was normalized so that its entries sum to one. Therefore, given two terms t_1 and t_2, the normalized entry of the t_1 HAL vector corresponding to t_2 was adopted as estimate of $\Pr(t_2|t_1)$, that provides a measure of the degree to which t_2 is related to t_1. The authors in that work also investigated the adoption of Information Flow (IF) where the relationship is not necessarily between two terms, but can be among a set of terms and a new term (e.g. IR can infer a relationship between "space program" and "satellite"). One of the approaches tested by Bai et al. consisted in creating HAL spaces using only a set of feedback document, i.e. the top 50. The “local” approach generally outperformed the “global” approach where the HAL space was obtained from the entire corpus.

In this thesis we will exploit an approach similar to [Hull, 1994, Bai et al., 2005], namely obtain a vector-based representation of relationship on the basis of the feedback set only. Our approach differs from HAL space since the weights are not propagated by a decaying factor to the near terms. We will focus only on terms in the expanded query to build the space, and the term–term matrix will be the starting point to obtain the final vector-based representation, specifically reducing the dimensionality of the space by SVD. Moreover, HAL spaces are not adopted as input for a learning algorithm as in [Hull, 1994] or to estimate probability of dependence between terms to expand the query model as in [Bai et al., 2005]. The obtained vector subspace representation can be directly included in the ranking function and exploited to perform prediction.

The specific approach adopted in this thesis is based on the modeling procedure introduced in [Melucci, 2008]. In that work the author introduced a geometric framework, that is actually that adopted in this thesis. The basic idea underlying that framework is to exploit the analogy between the possibility of representing an informative resource with regard to diverse sources, and the fact that a vector can be generated by different vector space basis. Each vector space basis spans a vector sub-

space, where the subspace models the contribution of a source. In this thesis a source is associated to a specific hypothesis on the possible factors, or variables, that can explain the user perception of relevance with regard to the evidence gathered from the source. An example of hypothesis is that term relationships extracted from documents judged as relevant are a possible set of factors that explain the user perception of relevance. If the hypothesis is true the factors provide information on the user information need; therefore, if modeled they can be adopted as a new dimension of the information need representation. Using the framework proposed in [Melucci, 2008], if a vector subspace representation of the term relationship is obtained, then this representation can be exploited to support prediction through a projector-based function. Melucci showed how such function can be interpreted as a trace-based function and that the measure is a probability measure. The idea of using trace in IR, and in particular the density operators, was originally introduced in [van Rijsbergen, 2004], and one of its important consequence – subsequently exploited in [Melucci, 2008] – was to "establish a link between geometry and probability in vector spaces" [van Rijsbergen, 2004]. In this thesis we will focus on a linear regression interpretation of the adopted function, as discussed in Section 3.3.2.

In [Melucci, 2008] term relationship are modeled using SVD on a term correlation matrix exploiting local co-occurrence data. We will revisit the proposed approach in the context of the introduced methodology and exploit the diverse methodology steps to evaluate this approach in an explicit relevance feedback scenario instead of Pseudo-Relevance Feedback. We will test the effectiveness by varying the diverse methodology steps, i.e. the selection of the source for term relationship, namely the feedback set, the term selection strategy and the document representation. Moreover, the test will be performed using test collections constituted by much larger document corpora and through the participation in standard evaluation campaigns, i.e. TREC.

2.2 User Behavior

During the last decade another source of evidence has gained increasing interest: the behavior of the user when interacting with the system. The interest in this source is mainly due to the possibility of gathering large amounts of evidence without direct involvement of the user. Post-search navigation features, for instance, can be gathered by unobtrusively monitoring the behavior of the user when interacting with the results returned by the system. If these features can provide information on the user perception a document with regard to his current need, they could be adopted as cost-less evidence for input prediction. Approaches exploiting this evidence are known as IRF techniques, as opposed to explicit feedback where users are explicitly asked to provide

information on their need, e.g. through an assessment.

When considering behavioral features as implicit indicators of user interest, two issues to address are: (i) given an implicit feature or a set of features, what is their predicting power? (ii) How can the information extracted from these features be explicitly included in the prediction process? In the remainder of this section past works concerning the investigation of user behavior in IR will be briefly reviewed with a focus on these questions.

The work reported in [Kelly and Teevan, 2003] provides a survey of features gathered by observing user behavior and adopted as implicit indicators or as evidence for implicit feedback. They extended the framework proposed in [Oard and Kim, 2001] for characterizing observable user behavior. Oard et al. used a categorization based on two dimensions: observed behavior — examine, retain, reference, and annotation — and scope of the item involved in the user interaction – segment, object, class. Kelly et al. added "create" to the observed behavior and classified several works using this categorization. The features adopted in this thesis concern the "examine" observed behavior and the "object" scope. "Design", "Implementation" and "Evaluation" are other categories proposed in that work for classification. The first of the two questions shares the same intent underlying the "Design" category whose underlying question was "what are good implicit measures to use?".

Some of the works reported in [Kelly and Teevan, 2003] will be briefly reviewed in the following with regard to the above posed questions, and integrated with more recent works. The specific focus will be on the second question, namely approaches exploiting user behavior.

2.2.1 User Behavior Feature Predicting Power

Investigation of Single Behavioral Feature

The study of the predicting power of interaction features has gained particular interest since, if they are reliable indicators of relevance, they can provide a costless — in terms of user effort — evidence for substituting explicit judgments, e.g. for learning ranking functions. Three of the earliest works which investigated the predicting power of implicit indicators are [Morita and Shinoda, 1994], [Claypool et al., 2001] and [Kelly and Belkin, 2001]. The work reported in [Morita and Shinoda, 1994] is focused on information filtering and investigates the capability of the time the user spent when reading a NetNews article as possible indicator of the user interest. They modified the GNUS reader to capture reading time, content of articles and other interaction, e.g. saving or follow-up actions. Eight users were asked to read and rate articles with a four graded scale; a total of 8000 article were rated. Reading session and rating session

were considered separately. A correlation was observed between the time spent and the rating — users tended to spend a long time on interesting documents and not to spend a long time on uninteresting documents. Using a time threshold of 20 seconds, they were able to retrieve 30% of the interesting articles with a 70% precision. Moreover, they investigated variables other than user interest that can potentially affect reading time, i.e. document length, readability – characterized by different variables, e.g. the number of characters per line or the ratio of blank lines in the article – and number of unread articles, i.e. size of the back log; no significant correlation was found with those variables. Besides reading time, none of the other variables was adopted to support prediction — see Section 2.2.2. Findings on reading times has been confirmed in a further work reported in [Konstan et al., 1997] that discusses the adoption of the GroupLens collaborative filtering tool for Usenet news.

In [Kelly and Belkin, 2001] the authors investigate three different hypothesis, specifically that (h1) the user spends more time, (h2) scrolls more, and (h3) interacts more with documents of interest than not interesting documents. Interactions corresponded to click on button for navigating among document passages or showing keywords. The adopted data were gathered during participation in the TREC-8 Interactive Searching Study, particularly were extracted from the trace files. Since users were asked to save documents perceived as useful to the topic, saving actions were interpreted as positive judgments. The methodology adopted was to compute the average values for display-time, scrolling, and interaction both in relevant and non-relevant documents and compare them. No significant difference was found among the average for all the three features. The result differs from those in the two previous works, where reading time was found to be correlated with user preference. The type of task affected the outcome of the study: indeed, the users were asked to construct queries, evaluate, save and label documents within a fixed time. In contrast, in the user study performed in this work, we did not impose any time constraint to avoid this issue.

The work reported in [Claypool et al., 2001] investigates the relationship between explicit ratings and the values of some interaction features monitored by the *curious browser*. This browser was implemented by the authors and captures actions when the browser window was on focus. The specific actions captured and whose predicting power was investigated were: the time the user spent on the browser window, time spent moving the mouse and mouse clicks, scrolling activities performed by mouse clicks on the scroll bar or key-strokes – i.e. by Page Down, Page Up, Up Arrow and Down Arrow – and the time spent performing scrolling actions. A user study was carried out that involved seventy-five students. They were instructed to open up the browser and perform browsing activities for 20–30 minutes, without specifying the objective of the study. The browser, when the user left the page, displayed an evaluation

window with a five graded relevance scale. Explicit ratings were provided for 1823 over 2267 web pages visited. Exploiting explicit ratings and gathered features, they investigated by the degree of independence between the medians among each of the five explicit rating groups for each feature. The Kruskal-Wallis test was adopted. A positive relationship was found between explicit rating group and time spent on the page and time spent scrolling. Unlikely the previous two features, when considering time spent moving the mouse the only significant difference was found between values corresponding to the lower rating and those corresponding to the four higher ratings. No difference between rating groups was found when considering mouse clicks. This thesis exploits the same client-side approach to gathering user behavior features. Indeed, as discussed in [Claypool et al., 2001] the adoption of a client-side tool allows complete coverage of the possible feature to monitor, differently from a server-side tool — e.g. when capturing the time spent on a web page or retention features, e.g. bookmarking, saving, or printing actions. Moreover, even thought Claypool et al. investigated several user behavior features, those features were considered in isolation and with regard to a generic browsing activity, not in the context of a specific search task — e.g. the queries possibly issued by the users are not available. In this thesis we will investigate the effectiveness of multiple features, but considering specific needs expressed by specific query statements.

These works investigated the predicting capabilities of the directly gathered features. In contrast, in [Rafter and Smyth, 2001] the authors proposed to adopt derived features in order to remove possible noise in the interaction. The investigation was carried out in the context of job recommendation. In particular, revisit data, i.e. number of times the user re-visited the same job posting, and reading time were adopted as features. Revisit data were filtered by collapsing several revisits in quick succession in a single click; this approach was adopted to handle multiple clicks due to the late response of the system because of the network latency. The reading time was treated in order to remove outliers. In particular, a normalized time value was obtained by computing the average between the median of the median reading time for users and the median of the median reading time for jobs. This normalized time was adopted as threshold: if the observed time was greater than double the normalized value, the normalized value was adopted instead of the observed value. The adoption of this threshold was not motivated. Scores adopted to rank jobs in user profiles were derived by computing the number of standard deviations above or below the user's mean reading time. Derived scores were more effective than raw features, both in terms of recall and precision defined by considering the number of predicted jobs for which the user actually applied. That suggests exploiting the application of transformation or normalization based on task or user information to improve the predicting capability

of features.

The work reported in [White and Kelly, 2006] for instance explicitly investigated possible variables that affected the predicting capability of display-time thresholds. The specific variables under consideration were the task that the user was performing when examining the document and information on the user. In particular, the research questions investigated were if the display-time thresholds tailored for the search task, or personalized for each user, or considering both these variables outperformed a threshold computed by ignoring this information — i.e. a display-time threshold computed over all the users and all the tasks. Display-time thresholds were adopted to distinguish between relevant and non-relevant documents; documents identified as relevant were then adopted as input for query expansion using the top six expansion terms by *wpq* weight [Robertson, 1990]. The test collection adopted was obtained by the longitudinal user study carried out in [Kelly, 2004]. The study involved seven users, whose activities were monitored for fourteen weeks by client-side loggers; moreover, a proxy was adopted to gather the visited web pages. A graphical user interface was adopted to gather information on the task the user was performing when visiting a page, a judgment on the page in a seven-graded relevance scale, and the confidence in the judgment. Several implicit features were monitored besides display-time, e.g. scrolling activity or retention actions (saving and bookmarking). The study on display-time was limited to web pages. The set of web pages were divided into a training and a test set, respectively 412 and 2329 pages. The thresholds were learned on the training set. User defined tasks were classified in nine classes by the authors; the three most frequent terms in the user provided description was adopted as queries for the evaluation; the number of queries was 46. An examination of the distribution of the judgments suggested collapsing relevance judgments from the seven-graded scale to a binary scale customized to subjects, thus obtaining the flattest distribution per subject. The methodology adopted for the evaluation consisted the following steps. For each task/subject pair, the documents was considered in visited order by the user. If the document display-time was equal or greater than the threshold, it was used as source for query expansion; if it was not the first relevant document, query expansion was performed on the entire set of recognized relevant documents. Document ranking was performed using the TF·IDF weighting scheme. The iteration was performed using diverse size of feedback set, i.e. $n \in \{1, 2, 5, 10, 15, 20\}$; the evaluation measures adopted was MAP and P10. The obtained results showed that the most effective strategy in terms of both the adopted measures was that based on task-tailored thresholds when ten or more feedback iterations was adopted. Per user personalization or considering both the variables affected consistency in performance. This finding was one of the motivations for the investigation reported in this thesis both on the adoption

of per-user observed feature values and per-group derived feature values. Groups in this thesis are created on a per-query basis instead of on a per-task basis. Differently from [White and Kelly, 2006], in the experiments reported in this thesis queries and the underlying topics were assigned to the user before simulating search and performing the evaluation, and not created from task description using most frequent terms.

The findings obtained in [Rafter and Smyth, 2001, White and Kelly, 2006] suggest that, when investigating the predicting capability of implicit features in isolation, possible variables that allow a better characterization of the user interaction, e.g. task, user, or topic (e.g. the job in [Rafter and Smyth, 2001]), can help to determine whether or not the feature value is denoting interest by the user. When considering reading or display-time, the above approaches are all based on the hypothesis that users tend to spend longer on interesting documents than on uninteresting documents. Even if this thesis shares the same hypothesis, it aims at considering correlation with other features, e.g. document-specific features, to explicitly include additional variables which can affect the time length, e.g. document length.

Besides reading and display-time, one of the most investigated implicit features is click-through data. The work reported in [Joachims et al., 2007] provided several insights both on the way users interact with the results in a web search scenario and the reliability of click as absolute indicators of user preference. Two user studies were performed by monitoring through an eye-tracker the behavior of the users when examining Google[3] result pages obtained in response to queries to address five informational and five navigational assigned questions: recruited users were asked to search to accomplish the assigned questions using Google and exploiting self-generated queries. In the first study 34 users were recruited and asked to rank results in preference order, instead of providing judgments, since the former approach is cognitively easier. The second study involved 16 users who were asked to rank both abstract and whole pages, namely to consider also the content, were asked to be ranked. The research questions concerned the possible effect of trust in the search engine capability and the overall result list quality in the user clicks. With regard to the first question, users tended to click on the top results, that is clicks were biased by the trust in the search engine capability, also when the abstract (title-snippet-URL) on the top results were judged less relevant. In order to investigate the impact of the overall result list quality, the authors altered the presentation of the results considering two configurations: swapping the first two results and showing results in reverse order. The obtained results showed that in the altered settings the user tend to click on less relevant results. These findings suggest clicks cannot be interpreted as an absolute indicator of user interests. Therefore they investigated possible interpretation of click-through data to infer relative preference

[3] http://www.google.com

among the results, both when considering a data observed for a single query or for the entire query chains to accomplish the objective request to the assigned question. An example of effective strategy was *last click* > *skip above* where the last result clicked on a result page was considered more relevant than those presented at highest ranking but not clicked. They investigated the effectiveness of inferring relative preference both with regard to explicit judgments on abstracts and on the whole page. In both cases reasonable agreement was observed for preference predicted by click-through strategy and explicit relative preference. However, the study is exploratory, not applied to result re-ranking. Another finding concerns the mean number of abstracts (title+snippet+url) viewed above and below a clicked link depending on its rank. The users tend to adopt a depth-first strategy, that is clicking on an abstract when considered promising instead of performing a full scan of the entire result list. Despite that, the users on average tend to view the result immediately below the clicked results 50% of the time. The same result was observed in [Cutrell and Guan, 2007] — e.g. when users clicked on results at rank five they had viewed almost all the results presented above and 1.4 results below. The work reported in [Agichtein et al., 2006b] investigated click-through strategies not including query chains and their extension in real web search settings: results showed that they are less robust than approaches that include click information (e.g. click above or below a page) as features to directly learn ranking functions.

In this work we will consider a scenario where the user examines the top ten results, therefore no relative result preference are extracted. Moreover, we will assume that the first visited documents can be adopted as sources for feature values from which to obtain behavioral models. An analysis will be carried out on the effect of the capability of the user to select relevant documents among those used to obtain the model and the effect of the effectiveness of the models to support re-ranking and query expansion. Because of the findings reported in [Agichtein et al., 2006b] these models could be further improved to explicitly include click information.

The common objective of these works was to predict user preference of a document on the basis of a single behavioral feature. The only exception is the work by Claypool et al. where a combined scrolling time was obtained as the sum of the time spent scrolling by mouse and that scrolling by keystrokes. This could be considered as a form of combination, but it involves homogeneous features. More complex relationships could exist between features, and the combination by sum might not be suitable where multiple and heterogeneous features are considered simultaneously, as done in this thesis. The next section will specifically focus on strategies that combine multiple features observed from user behavior.

Investigation of Multiple Behavioral Features

The work reported in [Fox et al., 2005] is focused on the predicting capability of the combination of behavioral features. The main objective of that work was learning models to predict user satisfaction both at the result level and the session level. Models were not used to directly support prediction since the objective of the work was more exploratory. The data adopted were obtained by a user study that involved 146 employers of the Microsoft Corporation for a period of six weeks. Both explicit judgments and implicit features were collected, both at the result and session level using an instrumented browser. Explicit judgments for each result were gathered using a dialog window that appeared after a user left a page or when the browser was inactive for more than ten minutes. Judgments were in a three-graded relevance scale, and gave also the possibility to not evaluate the page. When typing a new query, a dialog window asked if the session was terminated – that allowed for an exact session segmentation – and, when a positive response was issued, the user was asked for a judgment (three-graded relevance scale) of the entire search session. They investigated nineteen behavioral features at result level, mainly concerning examination (e.g. time spent on the page[4], time spent on scrolling, scrolling action, exit type, ...) and retention (bookmarking and printing). Exploited features were basically query independent. Eleven features were instead considered at the session level, seven of which were average values of result level features (e.g. average maximum scroll, average printed, ...). Bayesian-network learning methods were adopted to learn user models and the analysis was supported by representation of conditional probability distribution of the network nodes through decision graphs (nodes are variables and arcs correspond to dependence among variables). The analysis based on models through these tools and techniques provided some insights on effective combinations of features. At the result level, the way the user exited a result, click-through data, and the difference between the time the user left the result page and returned to it, were the most effective combination for predicting user satisfaction. Also retention actions (printing and bookmarking) were shown to be effective, but they were observed with very low frequency. At the session level the most important features were average duration on results, number of result sets and end action. At session level the authors also explored gene sequences, i.e. a way to represent behavioral patterns as a string. Results are not reported on gene sequences but the author stated that some of these were predictive of user satisfaction in preliminary experiments.

[4]Two different features concerning time was adopted: duration and difference. Duration referred to the actual time the page was on focus. Difference referred to the difference between the time the user left the result list and returned. Difference is actually the feature adopted in this thesis and named as display-time in the following.

This thesis shares the same intent of [Fox et al., 2005] in investigating multiple features in order to obtain a more effective model of user behavior. The main difference is that, as mentioned above, the work by Fox et al. is more exploratory and aims at predicting explicit judgments; they did not propose an approach for exploiting the obtained model. In contrast, the models obtained in this work can be directly integrated in the adopted ranking function. Fox et al. stressed that an interesting point, previously discussed in [Nichols, 1997], is that not necessarily implicit and explicit feedback should be considered as alternatives, but approaches to combine them should be investigated, e.g. to understand or increase the reliability of explicit judgments. The objective of the methodology introduced in this thesis is to obtain a uniform representation of explicit and implicit evidence: that could help to exploit both this feedback evidence using the same ranking function.

Bayesian dependency networks were further adopted to learn model to predict query ambiguity [Teevan et al., 2008]. Query ambiguity in that work refers to the variability of what different individuals perceive as relevant with regard to the same query. Two measures of ambiguity were investigated: click entropy [Dou et al., 2007] and implicit potential for personalization. Potential for personalization [Teevan et al., 2010] was defined as the gap between the optimal rating for an individual and the optimal rating for a group. The measure of optimality adopted was Normalized Discounted Cumulative Gain (NDCG) [Järvelin and Kekäläinen, 2002], where the gain for a group was computed as the sum of the gains of the individuals constituting the group — see Section 5.2.3.3 for an example with data in the dataset collected in this thesis. Potential for personalization could be explicit or implicit: in the former case gains are explicit judgments provided from the user; in the latter case they can be clicks, interpreted as proxy for explicit user interests, or content-based implicit features, e.g those adopted in [Teevan et al., 2005]. The dataset adopted in [Teevan et al., 2008] was a query log gathered from the Live Search engine
constituted by 44,002 distinct queries and interactions performed by the users when formulating the query or visiting results. They first investigated the effectiveness of these indicators by comparison with explicit judgments obtained from a user study involving 128 people who were asked to assess 12 queries among those present in the logs. In particular, they used Fleiss Kappa and explicit potential for personalization curve to measure variability between user curves to investigate if the two implicit measures were effective predictors of variability. Since the two implicit measures were good predictors, they used query features (considering the query as issued a single time), result features and history features (when the query was issued multiple times) as input to build the model. They obtained the best accuracy when considering all the features simultaneously.

The work reported in [Qi Guo, 2010] recently proposed using behavioral features in addition to query features to predict query performance. The performance to predict was the Discounted Cumulative Gain (DCG)@3. Data were collected by server-side logging and a browser toolbar from the Bing search engine. The adopted features concerned results, i.e. features obtained by parsing the result page, interaction, i.e. features derived from log, and navigation post-result page. 2,834 queries were used. Explicit judgments were provided from human assessors in the same period the log data were gathered. Assessors and users in the logs differed from each other, unlikely the work reported in this thesis when considering personalized behavioral models. Multiple Additive Regression Trees were adopted to train a regression model to predict the query performance, namely DCG@3. The obtained results suggest that the adoption of interaction features can improve predicting capability: the measure of accuracy was the correlation between actual and predicted DCG, and the full model (using all the features among the interactions) produce a correlation of 0.70.

The works reported in [Fox et al., 2005, Teevan et al., 2008, Qi Guo, 2010] are not directly related to the work reported in this thesis since they aim at predicting respectively relevant judgments, query ambiguity and query performance, while the purpose of this thesis is to investigate a methodology to support re-ranking by the obtained models. Prediction of query ambiguity can be adopted to support the methodology proposed in this thesis, e.g. to predict whether to build personalized models or not.

Unlikely those works [Teevan et al., 2010] performed both an analysis of content-based and behavior-based implicit indicators and applied them to support re-ranking. The specific approaches adopted to support prediction will be discussed in the next section, while some of their findings will be discussed below. One of the main contributions of that work is the potential for personalization curve. The potential for personalization is obtained as the average of the best NDCG's for the users constituting the group, and the best NDCG can be obtained personalizing results for each user. The potential for personalization curve plots the computed average NDCG's for different group size. This curve was adopted to investigate the potential of personalization of implicit content-based and behavior features. When considering the content-base curve, larger variability was observed than for the curve curve based on clicks and on explicit judgments; the interpretation is that content-based implicit evidence can provide more information on variation in user intents, and can be adopted, for instance, to improve the score of relevant documents at low rank positions — clicks cannot help in this situation since they are focused on high rank positions. Results on the effectiveness to support prediction are described in Section 2.2.2. In this thesis we will adopt the group gains as defined in [Teevan et al., 2010] to compute the relationship between the agreement of the group and the individual in terms of ideal ranking in

order to investigate the effect of this agreement on the effectiveness of using group data instead of individual data to build behavioral models.

The work reported in [Agichtein et al., 2006b] investigated the effect of user behavioral features to support web search, explicitly addressing the problem of noise in real data. The behavioral model was based on the assumption that each observed feature value for a result and a given query is actually constituted by two components. The first component is the *relevance* component, namely the component that provides actual information on the user preference. The second component is a *background* component: e.g. in the event of clicks, that component corresponds to a user clicking on results indiscriminately. The approach for obtaining the relevant component is therefore to estimate the background component and subtract it from the observed feature value. The estimation of the background feature value is performed by computing the average feature value for all the queries submitted by all the users for a specific rank position. The feature value is obtained as the average over all the users and search sessions for a given query-URL pair; the average feature value was computed to reduce the effect of variations in behavior. In this work we will adopt average feature values across all the users who assessed a query to investigate if group derived feature values can substitute personal feature values when the latter are not available. A number of query-text features, browsing features, and click-through features were prepared as a vector and provided together with explicit judgments as input to learn a behavioral model by RankNet [Burges et al., 2005]. The dataset adopted was constituted by 3,500 queries sampled from query logs of a commercial search engine, for which explicit judgments were available for the top ten results. To obtain pairwise preferences, for each query the cross-product between all search results were computed and the preference was determined according to the relevance label — all pairs with equal label, namely ties, were discarded. With regard to the insights into behavioral features, two contributions are provided by this work. The first is that by using group data robust behavioral models can be obtained. The second is that browsing features are effective evidence when adopted to obtain behavioral model.

The work reported in this thesis also investigates the effectiveness of exploiting multiple features, but instead of learning a ranking function, it investigates if a specific relationship between interaction features, i.e. correlation, is a useful factor for performing prediction. In this sense, the work reported in this thesis aims at gaining additional insights into the possible variable can affect predicting capability of multiple features when considered simultaneously.

2.2.2 Algorithms Exploiting User Behavior

The second question concerns approaches that exploit user behavior features as evidence to support prediction. In the following we will consider two categories of approaches: those that exploit user behavior to enhance the textual representation of the information need, and those that directly exploit the user behavior to represent the information need. The approach adopted in this thesis can actually support both those kinds of representation as shown by the investigation discussed respectively in Section 5.1.2.1–5.1.2.2 and Section 5.1.2.3–5.1.2.4.

2.2.2.1 User Behavior to Improve the Textual Representation of the Information Need

In [Morita and Shinoda, 1994] the authors propose exploiting documents identified as useful in a set of sessions by reading time threshold as feedback for the proposed sub-string indexing method to filter interesting articles. The sub-string indexing method splits an article into substrings and then checks the occurrence of those substrings in a collection of documents known to be interesting; the final score is based on the substring matching. Using a threshold score, the system decides whether the document should be filtered or not. Implicit feedback is here adopted as a preliminary step to assist prediction, which is actually performed on the basis of content based features. This kind of approach is defined in [Oard and Kim, 2001] as "inference–prediction" strategy, since the objective is to infer explicit ratings on the basis of the observed behavior.

The work reported in [White et al., 2005] investigated how user interactions with diverse representations of a result can be adopted to iteratively improve the information need description, specifically extracting terms from representations which the user interacts with. The source here is the representation path derived from user interaction with the results. Result interaction is supported by a search interface and result presentation; each document in the result list is characterized by a variety of query relevant representations created at retrieval time. After a search is issued by the user and prediction is performed, the title of the top ten documents and the top ranked (according to the query) sentences extracted from the top thirty documents are displayed. Other representations are sentences in the document summary and each summary sentence in the document context, i.e. displayed between the preceding and the following sentence. The procedure for obtaining the diverse representations is described in [White et al., 2003]. Interacting with these representations the user can be supported in the exploration of the results. The diverse possible interactions describe possible paths. The objective of the paper is to investigate how, through the diverse

paths, different expansion methods can iteratively improve the textual description of the information need by extracting terms from the interaction paths. Different models are investigated. The binary voting model assumes that terms appearing in many representations can be useful: different weights are provided to the different representations in the path. This approach is quite heuristic and more principled approaches are investigated: variations of the *wpq* method [Robertson, 1990] that estimated the weight for terms using complete interaction path information (distribution of terms in the seen path considering all the constituting representation simultaneously), full document information (distribution of terms in the seen documents), and ostensive profiles (each representation in the path is considered separately and higher weights are provided to most recently viewed representation). Finally, Jeffrey's Conditioning model revises the probability of relevance of a term through the different representation in a path, providing exponentially decreasing weights to information accessed later; the basic rationale underlying this choice is to interpret further interactions in an exploratory perspective, that is considering the user being more confident on earlier access representation and less on the others. Simulations are adopted to investigate the effectiveness of the diverse query expansion techniques, considering best scenario (the user visited all relevant paths), worst scenario (the user visited all the non-relevant paths), and intermediate scenarios (both relevant and non-relevant paths are considered). The most effective and most robust techniques were the binary voting model and Jeffrey's Conditioning model. The latter strategy was less effective at the first iterations since it uses prior knowledge of terms, but after diverse iterations it outperformed the other in all three scenarios. The models reported in [White et al., 2005] are focused on the diverse degree of exploration of results, supported by the adopted interface, and how these interactions can provide information to extract expansion terms. Differently, the work reported in this thesis is focused on interaction with the full document to assist prediction both by re-ranking based on query expansion and by direct re-ranking using user behavior to represent informative resources and information need. One objective that the approach adopted in this thesis shares with [White et al., 2005], is to support the user in real-time, e.g. learning a model for user behavior directly from the first interactions of the user after a first search.

The work reported in [White and Kelly, 2006], besides providing insights into the robustness of display-time thresholds computed with regard to task, user information, or both, also proposed an approach to exploit display-time: documents with a display-time above the threshold were adopted as source for query expansion, where term selection was performed using the *wpq* weight. The final objective is therefore to identify useful results to improve the textual description of the information need through query expansion.

In [Teevan et al., 2010] an approach for behavior-based personalization was proposed that boosts previously viewed results at the top of the result list and results belonging to domains the user tends to visit. Ranking was based on URL matching, where the results whose URL matches the last three components were boosted more than those whose URL matches the last two components (e.g. http://www.**dei.unipd.it** VS http://www.dei.**unipd.it**). Therefore also in this work user behavior is adopted to support a better description of the information need, since previous interactions are used to perform prediction on the basis of term matching. This strategy outperformed the baseline, namely BM25, but not web-ranking. An optimal linear combination of content-based implicit score [Teevan et al., 2005], behavior based score, and web-ranking score (inverse of the log of the rank of the result) outperformed web-ranking. Experiments were carried out on a set of 699 queries with complete judgments.

The work reported in [Ruthven et al., 2003] investigated how user behavior can be adopted to support term selection and term ranking, then adopted for query expansion. With regard to the selection, they investigated a variation of the F_4 measure [Robertson and Sparck Jones, 1976]. Each term weight was constituted by two components: a *partial* component and an *ostensive* component. The *partial* component explicitly considered the degree of relevance indicated by the user — users could specify their perception of relevance through a slider, where the degree of relevance ranged from one to ten. If the user assessed the document as relevant with a degree k, it contributed as a $\frac{1}{k}$ relevant document in the *wpq* weight. The second component was inspired by the ostensive retrieval model [Campbell and van Rijsbergen, 1996]. Documents judged as relevant later were weighted more than earlier judged documents. The final weight was obtained as the product of the two components. Experiments were carried out using the topic of the TREC-6 interactive track. Six undergraduate students were involved in the study. Users could search using an interface that allowed query specification, results access (result titles were displayed) and assessments through the mentioned relevance slider. Their findings showed that the adopted weight was more successful in suggesting terms perceived as useful by the user, and actually used more heavily for expanding the query, even if the overall improvement in terms of effectiveness was not significant. With regard to the term ranking, they investigated the effectiveness of automatic selection of expansion methods based on properties of documents judged as relevant, i.e. precision of the search, position of the document within the ranking and similarity among document judged as relevant. While the previous approach performed term selection on a per document basis, the latter approach considers the entire feedback set. The automatic selection approach was compared with a standard approach where the query was expanded with the top six terms extracted from the feedback set. Also in this case, the proposed approach did not provide signif-

icant improvement. Besides the specific strategies adopted, the basic idea underlying this work is to gather more evidence from the user interactions, in this case relevance assessments, and exploit this evidence for feedback. The partial scores, the ostensive weighting, and the diverse properties of the documents in the relevance feedback sets basically are additional sources to characterize the user interaction and provide more evidence on the user perception of relevance. This is crucial since users do not explain why they perceive a document as relevant. The work reported in this thesis shares this idea and continues this line of investigation also exploiting interaction with the documents.

The work reported in [Bilenko and White, 2008] exploited post-search navigation activities to estimate term weights. In particular, the post-search activities are not limited to the navigation of the result, but on the entire search trails associated to a specific result — e.g. subsequent pages accessed starting from the result. The dataset obtained was a set of queries, where each query was associated to an ordered sequence of pages, specifically all those accessed in the trails when searching for query q. The criteria to rank those pages can be various, e.g. highest visitation counts or total dwell-time spent. Since many queries are unique, they did not consider the query as an atomic event, but as a sequence of terms. For a new query, the weight of a constituting term in a document is estimated on the basis of statistical information derived from query trails, similarly to the standard content-based approach where term weights are estimated both on the basis of the occurrence in the document and the collection. Actually, the term weight is not necessarily estimated only using occurrence of documents in the search trails: a variant is investigated where the term weight is based on the total dwell-time or its logarithm. User behavior features are therefore adopted to estimate term weights. They investigated (i) an heuristic approach, similar to TF·IDF but using statistics on query trails, (ii) a probabilistic approach inspired to language model (query trails information is adopted to estimate the probability a term is generated from a query and document probabilities for every term), and (iii) a random walk extension of the probabilistic approach (the probability of reaching a document took also into account the possibility of reaching the document from shared query terms with another document). They investigated the effectiveness both in terms of direct ranking, based on the scored provided by approaches (i-iii), and using these scores as additional features for learning to rank. The approaches outperformed a previous approach based on atomic interpretation of queries [Agichtein et al., 2006a] — see Section 2.2.2.2. Estimation of term weight based on the log of the dwell-time, exploited by the random walk approach was the most effective. The approach adopted in this work basically exploited dwell-time to heuristically estimate a term weight, starting on the hypothesis that total time spent on a page is indicative of user interest.

Two main contributions of this work are that prediction could benefit considering the query as a set of its constituting terms, especially for previously unseen queries, and that also interaction in the further exploration of the user can be beneficial.

In this thesis display-time is considered as a descriptor, not to characterize terms. Moreover, we will focus only on the documents in the result list, not considering the entire trails; indeed we will consider a generic scenario, where not necessarily hyperlink structure can be adopted to support prediction. In a web search scenario the approach investigated in this thesis could be extended considering entries in a search trial as additional evidence to model user behavior or target documents to re-rank.

2.2.2.2 User Behavior Descriptors for Information Need Representation

The above approaches exploit user behavior to obtain a most effective characterization of the information need when textual descriptors, namely terms, are adopted as input to perform prediction. Other works as well as the work reported in this thesis, investigate how to directly exploit user behavior features as descriptor. That is, how to represent information need directly on the basis of user behavior features.

The approach proposed in [Rafter and Smyth, 2001] in the context of job recommendation could be considered an example of this approach. Each document is represented by a normalized reading time that is adopted to compute scores based on the deviation from the time threshold, to rank jobs in the user profile. Here the hypothesis is that normalized reading time can be used as document descriptor.

The I-SPY system [Smyth and Balfe, 2006] is a meta-search engine that, besides aggregating scores by the different back-end search engine, support re-ranking based of pattern of previous issued searches in interest-specific communities. For each specific search community, an hit-matrix is maintained; that matrix stores for each issued query the page visited by the users in the community. When a new query is issued, a number of candidate similar queries are identified on the basis of the number of terms shared with the new query. For a given query, the score assigned to the page is given by the number of accesses (hits) to that page over the total number accesses. The final score of a page is given by the weighted sum of the page scores for each of the candidate query, where the weight is the normalized similarity between issued and candidate query. A study was carried out involving 92 real users, specifically undergraduate students. Users were divided in two groups: one for training, i.e. to populate the hit-matrix, and one for test. The ground-truth, namely explicit judgments, was independently prepared manually. Both an increment in terms of recall and precision was observed. The source adopted here is not only the page access: the effectiveness of the proposed approach relies on the conjunction between related queries, previous visited pages and the fact the hit-matrix is computed for each interest-specific community.

Most recent approaches investigated representation based on multiple behavioral features directly to support document ranking. In [Agichtein et al., 2006b] the authors exploited both observed features, aggregated across all the users interacted with a specific query–URL pair, and derived features — the rationale underlying a derived feature is described in Section 2.2.1. Each query–URL pair was described as a vector of all the features and the behavioral model is obtained by a supervised approach through machine learning technique, namely via RankNet [Burges et al., 2005]. RankNet is a modification to the standard neural network back-prop algorithm — e.g. see [Bishop, 2006]. The objective of the back-prop algorithm is "to minimize the value of a cost function by adjusting each weight of the network according to the gradient of the cost function with respect to that weight" [Richardson et al., 2006]. Differently to the standard algorithm, instead of trying to minimize the error between network output and desired out, RankNet attempts to minimize the difference among outputs of result pairs: if result i should be ranked higher than result j in the training set, a larger cost is associated to larger difference among the output of i and that of j. RankNet was used to learn ranking function only based on document specific features [Richardson et al., 2006], to learn behavioral models [Agichtein et al., 2006b], and to investigate effectiveness of functions learned by using all the features simultaneously [Agichtein et al., 2006a]. In [Agichtein et al., 2006b] obtained behavioral model outperformed the commercial search engine strategy, previous click-through strategy proposed in [Joachims et al., 2005] as well as further refinements of these strategies proposed by the authors. Among all the post-search navigation features, browsing features were shown to be the most effective; actually browsing features outperformed the combination of all the three sets of features. The approach performs prediction only for documents with available behavioral data, and this can cause a low recall. Therefore, they examined also the variation in recall as a function of the number of days of activity data gathered: the results for a fixed precision of 0.7 showed an improvement – from 0.05 to 0.15 in approximately ten days – when the amount of data increased.

The reason for using a learning based approach is that ad-hoc approaches for combining the diverse indicators' contribution may fail when the domain of application of the IR system changes, e.g. in the event of intra-net search. Another factor which prevent use ad-hoc approaches is that the indicators predicting power, as suggested by the results obtained in [White and Kelly, 2006], may vary when designing approaches personalized for each user. This is actually the research question underlying the work reported in [Melucci and White, 2007b]. In that work each document was represented as a vector of interaction features monitored when a particular user was performing a specific task; the features of the first n documents visited when

accomplishing a task were adopted to build a vector subspace-based model of the user behavior when visiting those documents, specifically using Principal Component Analysis (PCA) [Pearson, 1901]. The specific methodology they adopted consisted in ranking the documents seen by the user according to the distance between the obtained subspace and vector representation of the documents in terms of interaction features; then extracting the ten most frequent keywords in the top n and using them for query expansion. The proposed approach was compared with query expansion based on the first visited documents per task and expansion supported by user-behavior based re-ranking using the centroid of behavior feature vectors, instead that eigenvectors. The dataset adopted was that gathered in [Kelly, 2004]. The best results were achieved when the model was build personalized for each user and tailored on the specific search task.

The work reported in this thesis shares the same idea underlying these works that exploiting multiple behavioral features simultaneously can be beneficial. Differently from the work reported in [Agichtein et al., 2006b] we will investigate non personalized models both support direct user behavior-based re.ranking and to support query expansion. Moreover, following the idea proposed in [Melucci and White, 2007b], this thesis investigates models personalized for each user and based on the first visited documents. The work reported in [Agichtein et al., 2006b] aimed at learning directly the ranking function capturing relationship among features from the training data. Ranking functions based on diverse set of behavioral features are compared in terms of variation in performance. In this thesis we will model user behavior starting from specific hypothesis on the possible factors, e.g. feature correlation, that can affect prediction when user interaction behavior is adopted as evidence. The modeled factors can be included in a unique ranking function which is the same for all the factors of all the possible sources: a unique ranking function can be adopted because we will uniformly model all the source factors, from completely different sources as vector subspaces. Therefore, model derived from content-based features is modeled through the same mathematical construct on user behavior. The relationship among the features is determined by the hypothesis: that provides us a principled approach to investigate diverse hypothesis in the same framework.

The work reported in this thesis exploit the same strategy for modeling user behavior proposed in [Melucci and White, 2007b], but

- we will investigate a scenario where documents were ranked with regard to the topic on a first stage and explicitly compare the effectiveness of user behavior-based re-ranking and query expansion supported by user behavior with the first stage prediction;
- we will investigate the effect of the number of relevant documents among those

used to model user behavior both on the direct re-ranking by user behavior and query expansion supported by user behavior;

- we will investigate the effect on the retrieval effectiveness of using features gathered from the behavior of interrelated, where the effectiveness will be measured with regard to the individual users;
- we will test the effectiveness of user behavior-based re-ranking to support query expansion on a much larger test collection.

2.2.3 Behavioral Feature Granularity

In this thesis we will investigate the effect of exploiting post-search interaction feature values gathered by monitoring the behavior of a group of users in order to substitute feature values for the individual user, when not available. The exploration will be carried out by exploiting the effectiveness in terms of gains explicitly provided by the individual, namely in a personalized scenario — see the methodology application described in Section 4.2. The granularity is determined by the source from which the behavioral features are distilled. For instance, we can consider the value of a behavioral feature at result granularity for an individual, e.g. the observed value; alternatively the feature value can be obtained from all the values in a search session (the source is session behavior), or all the users that issued or assessed a specific document–query pair (in this case the source is group behavior). Previous works investigated the adoption of different granularities for feature values observed from the user behavior. In this section we will briefly discuss these approaches and the difference with this work.

In [Kelly and Teevan, 2003] the authors considered individual and group as two distinct dimensions for classification: individual's and group granularity levels refer to explicit judgments that the implicit feedback strategy should predict; for instance, if a reading time threshold obtained as average over a group of users is adopted to support an individual or predict individual preference, the approach is categorized as "individual".

Works in collaborative or community-aware search can be considered as related to the approach adopted in this thesis, when group behavior is adopted to support the individual. For instance, the data maintained in the hit-matrix of the I-SPY system [Smyth and Balfe, 2006] is at group granularity: page accesses is indeed not maintained at the individual level, but at group level. Another approach to community-aware search is proposed in [Almeida and Almeida, 2004] where Bayesian Belief Network is proposed to combine content-based score and community-based score. Communities were automatically identified from sessions interest graphs using the HITS algorithm [Kleinberg, 1999]. A session is the subset formed by accesses performed by

the user during a single interaction with a informative resource. Informative resources, e.g. online radios or pages, accessed in a session are considered as related. The content-based weight assigned to the informative resource is given by the cosine similarity of normalized TF·IDF vectors of informative resource and query. The community-based score is given by aggregating informative resource weight across all the community member session and removing the weight of the non members — community members and non-member sessions are identified on the basis a the authority score obtained by HITS. Here informative resource scores are aggregated per community. These works basically considered only access information or derived scores at community granularity; in this thesis we will focus on post-search navigation features.

The work reported in [Teevan et al., 2009] investigated the relationship between different criteria for group creation, and the variation in user profile based of personal index, explicit judgments and selection of queries of interest (i.e. potential topic of interest). The criteria for group creation was longevity, i.e. trait-based and task-based group, and the group membership strategy, e.g. explicit and implicit. One of the motivation for this work was to investigate the extent to which group derived evidence can be adopted to support the individual user. The approach they adopted to perform prediction was based on the aggregation of personalized scores over all the members of a group. The personalized algorithm exploited the content-based component based on personal index [Teevan et al., 2005], and a behavioral component that takes into account the similarity between the URL and URLs previously visited by the user. This strategy, named *groupization* algorithm, significantly outperformed the personalized version and also web ranking, when adopted as additional evidence and using all participant as group members. The groupization algorithm provided a larger improvement when considering within-group score aggregation. One finding of that work is that contribution to support prediction can be obtained when group member has different indexes: that further supports previous findings on the effectiveness of combining diverse evidence. Even if the work reported in [Teevan et al., 2009] and this thesis shares the same motivation when considering groups, our approach involve directly feature values aggregation, not scores.

The approach used in [White and Kelly, 2006] for display-time threshold computation is an example of investigation of user behavior feature at diverse granularities. They investigate both threshold personalized for the individual and/or tailored for the specific search task. White et al. considered diverse granularity when exploiting a single behavioral feature. Other works, as well as this thesis, investigate diverse granularities when using multiple behavioral features. In [Fox et al., 2005] behavioral models to predict explicit judgments both using feature at result and session levels. When considering features at session levels, seven of the eleven features considered

were average values computed over the results in a session — in particular for duration in seconds, maximum scroll, printing, bookmarking, number of result pages, position in the result page and in the entire ranking. Differently, in this work the aggregation will be performed on a per query basis. This approach was actually adopted in [Agichtein et al., 2006b] to obtain increase the robustness of behavioral features by reducing the impact of the variability when considering individuals' behavior. No comparison is performed explicitly with re-ranking based on model exploiting feature values gathered from individuals. We will perform explicitly this investigation in Section 5.1.2.1 and Section 5.1.2.2.

2.3 Other Works on Feedback Strategies

The works discussed in previous section mainly concern with approaches to exploit evidence gathered from feedback interactions or the investigation of the predicting capability of this evidence. This section will discuss some works that concern methodologies to support the design of approaches that exploit feedback data or their evaluation.

When considering methodologies to support evaluations of feedback strategies, the common objective is to investigate which are the variables can affect their effectiveness. For instance, in [Salton and Buckley, 1990] the authors investigate the effectiveness of diverse feedback strategies explicitly considering the impact of expansion terms. A more systematic study is reported in [Harman, 1992]. While Salton et al. expand the query using most frequent terms, Harman investigate diverse term selection strategies and their impact of retrieval effectiveness. Results suggest that term should not only be re-weighted, but also expansion is beneficial. The work reported in [Wong et al., 2008] proposed an Idealized Relevance Feedback framework whose objective was to investigate the validity of some assumptions made when considering a RF environment. The final aim was to build a framework to obtain also an upper bound of the possible effectiveness can be achieved by RF. The basis for their investigation was the Rocchio formula. The new query based on the feedback set was constituted by a number of terms k. In order to find the best query with k terms, they used a greedy algorithm. Given all the terms in the relevance feedback set the algorithm select the term that maximized the MAP among all the terms. At the hth iteration it selects the term that maximize the MAP when using in q_{h-1} among all the terms not already selected, where q_{h-1} was the best query of $h-1$ terms. As done in [Harman, 1992] they investigated a number of term selection strategies. The weight assigned to a term was the product of two components: an inter-document component, e.g. Inverse Document Frequency (IDF), and an intra-document component, e.g. TF. Intra-document components were classified according to normalization in the feedback set, in the en-

tire collection, or none. Best results were achieved for full collection normalization; $W4$ combined with $NMaxNTF$ was among the best performing. The problem of strategies that exploit full collection normalization is that they require corpus-wide statistics, whose computation could be slow on very large test collections as the one considered in this thesis. Another finding concerns the best query size: even if considering the best query size per topic provided the best performance, no significant difference was observed respect to the case when fixed-size query were used. Since this finding supports the user of fixed size query, also in this thesis we will consider this approach.

The approach underlying these works is to identify a set of steps common to the diverse feedback strategies, e.g. vector space-based or probabilistic, and investigate the impact of these steps on their predicting capability of such strategies. Query modification is framed in a term selection strategy and a term re-weighting strategy. Some approaches, e.g. Rocchio, perform these steps simultaneously: vectors used to represent relevant documents are merged with the initial query vector. Other approaches, e.g. [Robertson and Sparck Jones, 1976], are focused on term re-weighting and the expansion strategy is considered as a distinct step. In this thesis term selection and term re-weighting will be considered as part of two methodology steps for the methodology application that aims at modeling term relationship in documents judged as relevance. This thesis aims at generalizing these steps for generic sources, possibly described by diverse descriptors, e.g. post-search interaction features.

The main objective of the methodology introduced in this thesis is to support the design of an IR approach where diverse source of evidence are uniformly modeled and exploited for feedback. The work reported in [Bodoff, 2004] also presents a methodological approach to address the problem of exploiting feedback data, but it is specifically focused on relevance models. The approach adopted in that work shares with language models [Ponte and Croft, 1998] the modeling of parameterized document distributions and with the models proposed in [Zhai and Lafferty, 2001] the modeling of parameterized query distributions; relevance data is supposed to depend stochastically both from the query and the document parameters. Query and document representations are obtained simultaneously, not as two distinct steps, and together with the function to perform prediction. Bodoff identifies two main steps (named "stages"): a parameter estimation step and a prediction step. The objective of the first step is to estimate parameters, provided that a distribution for the query and one for document have been defined. The selection of the distributions and their parameters estimation can be considered as a modeling step. The Bodoff models assumes that distributions have been selected, therefore the modeling step is not explicitly considered. Since relevance data is supposed to depend stochastically from the parameters, a distribution

is defined also for the relevance function; the estimation of this function is part of the parameter estimation step. The prediction step consists in the actual prediction of relevance on the basis of the estimated relevance function. Bodoff discusses also the problem of feature selection that consists both of the selection of the actual feature to use as predictors and the selection of the dimensionality, i.e. how many features should be used. Bodoff therefore identifies three steps: the feature selection step, the parameter estimation step and the prediction step. The basic rationale here was to identify aspects common to all the relevance models and the common steps to perform in order to obtain query and document representation and exploit them. This thesis shares the same objective but in a completely different framework, namely a geometric one, and explicitly considering the issues to address when diverse sources of evidence are considered. With regard to the modeling procedure, this thesis considers information need and document representation as two distinct steps, therefore a one-sided approach is adopted.

CHAPTER

THREE

A METHODOLOGY TO MODEL SOURCES FOR FEEDBACK

3.1 Definition

In this Section we will provide some basic definitions to describe the specific abstraction of the IR problem considered in this work. The reason for the introduction of this abstraction is twofold. The first is that allows us to describe the specific interpretation of sources of evidence considered in this work and how their contribution can be adopted to support feedback. Besides being the basis for a terminology, this abstraction is the basis for the design of the methodology and of IR system able to support diverse types of informative resources and sources of evidence.

Definition 1 (Unit)**.** A *unit* is a thing that has its own existence and is involved in the search process; it is complete by itself but can also be part of something larger.

Definition 2 (Relationship)**.** A *relationship* describes a connection among units involved in the search process.

Definition 3 (Resource)**.** An *informative resource*, or more simply *resource*, is a unit, an aggregation or a group of units or a relationship between units.

Definition 4 (Source of evidence)**.** Resources are characterized by a number of properties. A *source of evidence* is a property of a resource.

Definition 5 (Descriptor)**.** Each source is described by a set of *descriptor*s.

Definition 6 (Feature)**.** Each descriptor can be characterized by a set of *feature*s.

Definition 7 (Value)**.** The result of the observation of a property, namely a source, corresponds to the measurement of the *value* associated with each descriptor adopted

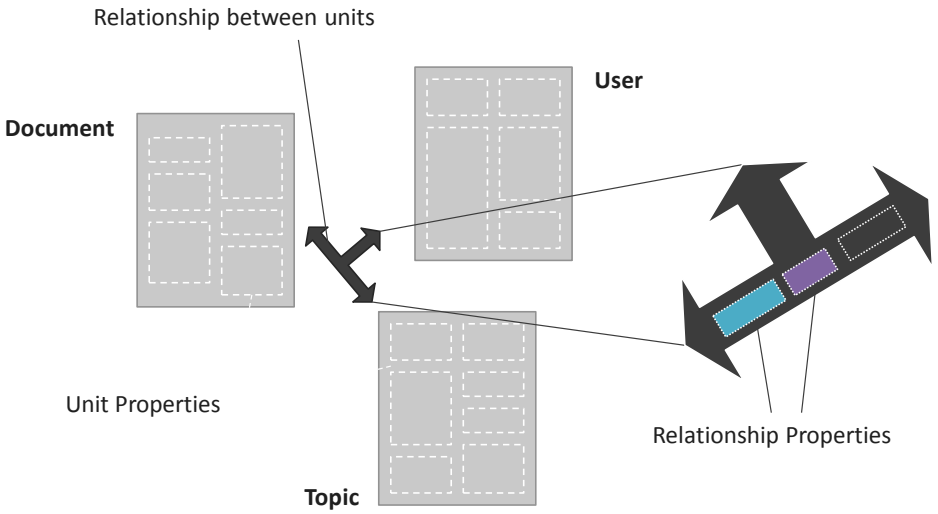


Figure 3.1: Examples of Informative Resources involved in the search process, i.e. units and relationship.

to characterize the source. If the source is characterized by a single descriptor, the result of the measurement/observation is a single value. If the source is characterized by a set of descriptors, the result of the measurement/observation is a set of values, each of them corresponding to a descriptor.

Definition 8 (Factor)**.** The possible values or value sets that can be obtained from the observation of a source can be explained by a number of (possibly) hidden variables or *factor*s.

Definition 9 (Dimension)**.** A *dimension* is a set of factors that can be used to represent the information need of the user with regard to the source from which the factors are obtained.

An Abstraction for the IR Problem

Let us consider the above definitions and show how they provide an abstraction of the search process. Example of units are the user who is searching for information, the topic the user is searching for, the task the user is performing when searching for the topic, the location where the user is, or the document. A pictorial representation of diverse units involved in the search process, i.e. document, user, topic, is reported in Figure 3.1. Units are represented as big rectangular forms with a thin dark gray border.

Each unit is characterized by a number of properties that, according to the above

definitions, are possible sources of evidence. For instance, a document can be characterized by a number of properties, e.g. metadata or, when terms are adopted as document descriptors, term occurrence or term relationship. Other examples of sources are the GPS position when considering the unit "location", or the type of task when considering the unit "task".

When defining a unit it was specified that, even if the unit has an its own existence and it is complete by itself, it can be part of something larger. A first reason for this choice is to explicitly consider in the abstraction the fact that a document can be interpreted as an aggregation of units, e.g. constituting fields – title, abstract and corpus – or passages. A second reason is that, when considering units that share common properties, they are all part of a unit set. For instance, all the documents can be considered as part of a document set D, since they all have a title, a content, an author[1]. A third reason is that this definition of informative resources allows modeling group of units, e.g. a group of interrelated users sharing common interests, searching for the same topic or performing the same task. Another example is the case when document collections are distributed among different providers, as in Distributed IR (DIR) or P2P IR. Document, collection (e.g. peer), set of collections (e.g. peer groups led by super-peers in a hybrid network) and sets of collection set (the entire network) can be considered as diverse levels of informative resources. In this case prediction should be done not only at the level of the documents, but also at higher resource levels. Indeed, the collections are distributed among diverse servers and searching all the servers could be too expensive both in terms of communication resources and computation [Callan et al., 1995]. A possible approach is to select the most promising servers and search only them; this problem is known in DIR as *resource selection* [Callan, 2002] and it is particularly important in the event of unstructured P2P networks when each peer has a limited knowledge of the other peers in the network [Lu and Callan, 2006, Nottelmann and Fuhr, 2006, Melucci and Poggiani, 2007]. This multi-resource level abstraction has been adopted in the design of SPINA [Di Buccio et al., 2008], is the basis of the IR system adopted for the experiments reported in this work and described in Section 5.2.4.

Units are not isolated from each other, but possible relationships can exist between them. Let us consider, for instance, a user examining the results returned by the IR system as response to a first query formulation. The way the user examines the results and interacts with some of them depends on the user himself, since each user can have his own style of interaction with the results, and on the task the user is performing since the task can affect the search strategy — for instance fact re-

[1] In a conceptual modeling perspective unit sets can be obtained by the application of a classification mechanism on units.

trieval or question answering generally imply look-up search activities, which differ from those adopted when the search activity is more exploratory, e.g. when driven by curiosity [White and Roth, 2009]. The main point here is that the user behavior is determined by the relationship between a number of units involved in the search process, e.g. document, topic, user, and task. In Figure 3.1 relationships are represented by a multi-directional arrow. Like units, relationships can be characterized by properties; an instance of relationship properties is the mentioned user behavior.

The concept of relationship is crucial in IR. The prediction of relevance of a document given an information need description can actually be interpreted as the prediction of a relationship, e.g. in the case of the Probabilistic Relevance Framework [Robertson and Zaragoza, 2009] where the basic assumption underlying is actually that relevance is a relationship that may or may not hold between a document and an information need.

Units and relationships are both considered as resources since their properties – depicted in Figure 3.1 as smaller rectangular boxes with a dotted border – can be exploited as a source of evidence to support prediction. Document specific properties are adopted by most IR systems as a source of information, e.g. document constituting fields or meta-data. An example of relationship property exploited to support prediction is the user behavior, adopted as a source of evidence in several IRF strategies [Kelly and Teevan, 2003].

Each resource property, namely source, can be described by a number of descriptors. When considering descriptors, two cases can be distinguished. The first case is when an observed feature is directly adopted as descriptor. For instance, if the unit "task" is considered and a measurement is performed to obtain information on the type of task, possible values can be "navigational", "informational" or "transactional" [Broder, 2002]; here the property is "task-type". An example where the result is a tuple can be the case when the user location is described by its GPS position. The GPS position is one of the possible properties of the unit location, the GPS coordinates are the features, and their values are the result of the measurement. As specified in the definition, a property does not necessarily characterize units, but also relationships between units. An example is the behavior of the user when examining one of the results obtained after a first information need formulation. The result of the measurement of this property can be a value or a set of values, according to the number of behavioral features adopted to describe property. For instance, if dwell time is adopted as a unique behavioral feature to describe user behavior when examining a document, the result of the measurement will be a value, e.g. 30 seconds. If other features are also considered, e.g. binary features indicating if the document has been printed or saved, a tuple $(30, 1, 0)$ can denote that the user spent 30 seconds on the document, he printed it but he did not

save it.

The second case is when a low level informative resource is adopted as descriptor for other resources. This is, for instance, the case of *term*s appearing in the documents. Each term has an its own semantic, e.g. it can express one or more concepts, but this semantic cannot be directly measured. A possible approach is to adopt a number of quantitative and observable features to characterize a term. These features can be observed when considering the document in isolation, as part of the collection or with regard to a topic, e.g. described by a textual query. When considering the document in isolation examples of features are the frequency of occurrence of the term or the positions where the term occurs in the document. When considering the document as part of a collection, examples of features are the IDF [Spark Jones, 1972] or the frequency of occurrence in the entire collection. When the document is considered with regard to the topic co-occurrence with query terms or pair of query terms are examples of features. Each of these features can be adopted as a distinct descriptor, similar to the case of task type, GPS coordinates or behavioral features. Alternatively, each term can be considered as a descriptor and characterized by a single value derived from the features observed for that term. The difference between the two cases is that in the latter case, i.e. the case of a term, a descriptor can be considered both as an informative resource – since it can be characterized by a number of features – and as a descriptor (and compound feature) — since it can be adopted to characterize other informative resources.

The work reported in this thesis is mainly focused on the last two definitions, namely factor and dimension. The reason for the introduction of the notion of factor is the need to understand why the user perceives a document as relevant on the basis of the evidence obtained from a source. A factor is one of the possible variables which explain the observations obtained from a source. Modeling factors corresponds to understanding the reason behind the user perception of relevance. The reason for the introduction of the notion of dimension is due to the need of modeling the user perception of relevance when multiple sources are exploited; each of the considered sources can be characterized by a subset of the possible factors obtained from the evidence gathered from the source. The following section will focus on these two notions, specifically their relationship with the notion of *source* and how they are adopted by the IR system to support prediction.

3.2 Source, Factor and Dimension

3.2.1 Recycling Scenario

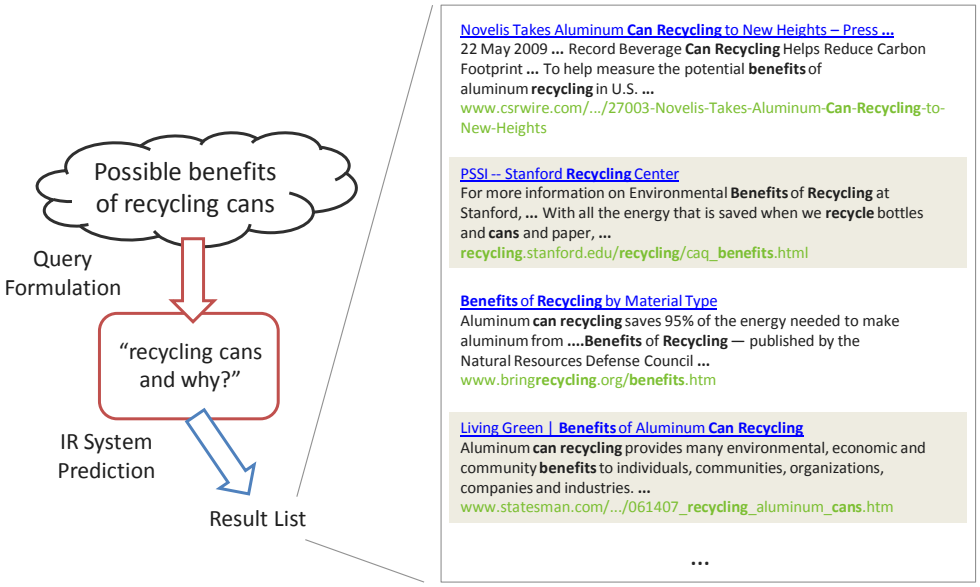
This section describes a search scenario. The scenario is introduced in order to support the description of the way sources are adopted to predict relevance. Moreover the scenario helps describe the methodology introduced in this work to obtain a usable representation of the diverse source contributions.

Let us consider a student that is writing a report of on possible benefits of recycling cans for a university class. In order to gather information to accomplish his task he submits a textual description of his information need to an IR system. A possible query could be "recycle cans and why?"[2]. As response to his request the searcher obtains a list of results — e.g. the top four could be those depicted in Figure 3.2a. The term "result" refers to the way a document is presented in the result list, e.g. title or abstract (title+snippet+URL) in Web search scenario. In the latter case the result can be considered a specific property of a document given an information need description. Indeed, the title or the URL are document-specific properties; but the snippet is query biased and constitutes a description of the document obtained on the basis of the information need description provided to the IR system by the user. In other words, the snippet can be interpreted as a property of the relationship between the topic and the document.

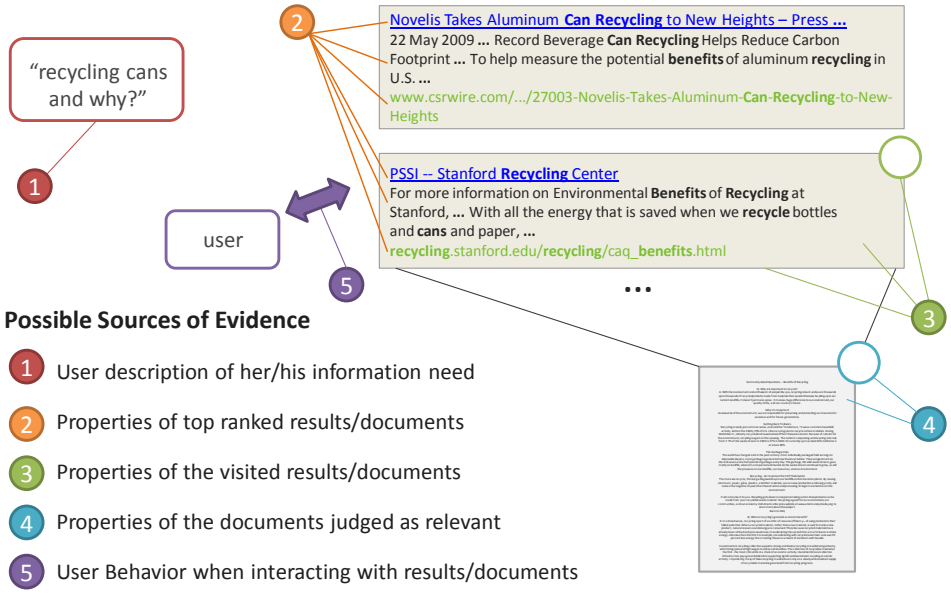
A subset of the results is usually displayed in the result page. Let us suppose that the user examines a small number of the top ranked results and visits the corresponding documents, e.g. the second and the fourth results, namely those with a dark background in Figure 3.2a. At this point, besides the initial information need formulation, e.g. its textual description, the following properties can be observed and used as sources for information to be exploited at the next stage:

- properties of the top retrieved and of the visited documents;
- properties of the relationship between the top retrieved and/or the visited documents and the topic, e.g., the snippet in the result;
- properties of the relationship among the document, the topic and the user:
  - the behavior of the user when interacting with the documents;
  - explicit judgments, if any, provided by the user on the documents.

[2]The example is based on an actual TREC topic adopted in the experiments, namely topic 546 of the TREC 2001 Web Track Test Collection.



(a)



(b)

Figure 3.2: Recycling Scenario. Results obtained after the submission of the query "recycling cans and why?" – Figure 3.2a. Possible sources after the first query formulation and the examination of the results – Figure 3.2b.

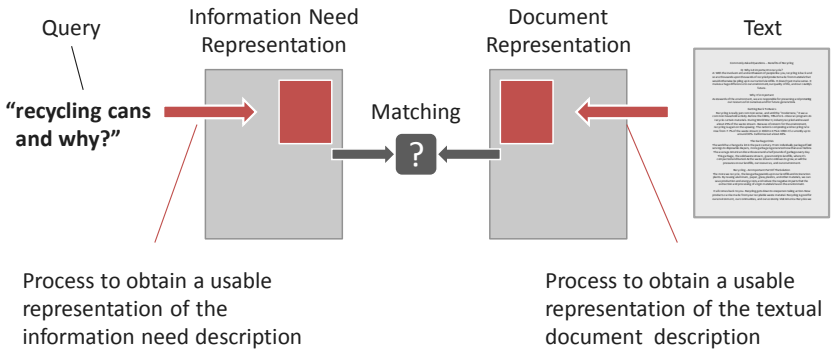


Figure 3.3: Classical interpretation of the IR process.

The objective of the methodology is to obtain a usable and uniform representation of these properties and exploit them as sources for feedback. The specific methodology applications investigated in this thesis will focus on the properties of documents judged as relevant among the top retrieved and the behavior of the user when interacting with the documents. The reason for investigating the former source is that, even if the user indicates a set of relevant documents, he does not explain why he perceives those documents as relevant; the possible factors that affect the user perception of relevance should therefore be investigated. If those factors will be identified and then modeled, they can be directly exploited in the prediction process. The latter source is even more challenging because, on the basis of the gathered evidence, it is necessary to understand whether or not a user perceives a document as useful. The main advantage of this source is that, different from explicit feedback, evidence can be gathered without an increment in terms of user effort. The two sources should not necessary be adopted independently: the diversity between them could be beneficial to model diverse aspects of the user perception of relevance, and therefore to obtain a better characterization of his need.

3.2.2 Dimension of the Information Need Representation

Let us consider the above scenario through a classical two-side interpretation, e.g. [Bates, 1989] of the IR problem — a pictorial description is reported in Figure 3.3. The two sides correspond to the user information need and the document. A representation for each of the two sides is required. Representations can be obtained from the evidence available during the search process. For instance, in the event of the scenario the information need is described through a query statement formulated by the user and submitted to the IR system, e.g. "recycle cans and why?". Terms constituting the

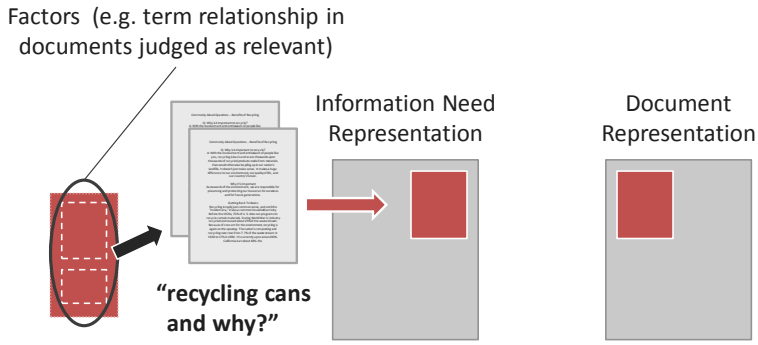
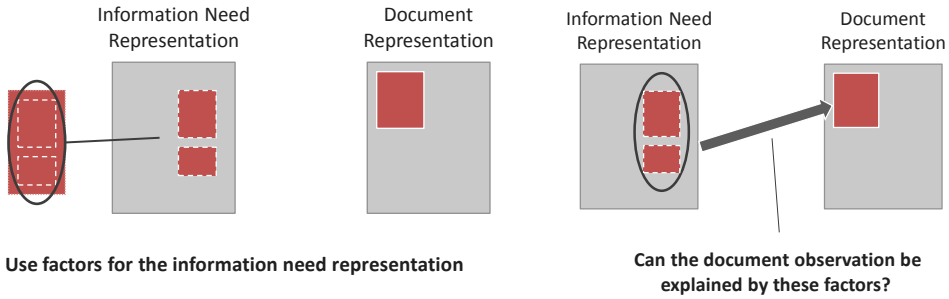


Figure 3.4: Term relationship in feedback documents.



(a) (b)

Figure 3.5: Prediction exploiting factors.

query are adopted to obtain a representation both of the information need side and the document side. The representations are based on an hypothesis of the possible factors that can affect the user perception of relevance. For instance, an hypothesis can be that the occurrence of query terms in a document can provide the system with information to the user perception of relevance.

While term occurrence is immediately observable, other factors could be hidden. Let us suppose that the user provides judgments on a subset of the top ranked documents. The objective of the system is to obtain a better characterization of the information need through the properties of the judged documents, specifically on the basis of possible factors that explain why the user perceived that documents as relevant — see Figure 3.4. For instance, on the basis of the assessed documents, the term "benefit" could be added to the descriptors, e.g. through a query expansion technique. Moreover, even if a document is about benefit of can recycling, it can be more focused on economical benefit while the user could be more interested in environmental benefit. Even if a query reformulation can help meet the user needs,

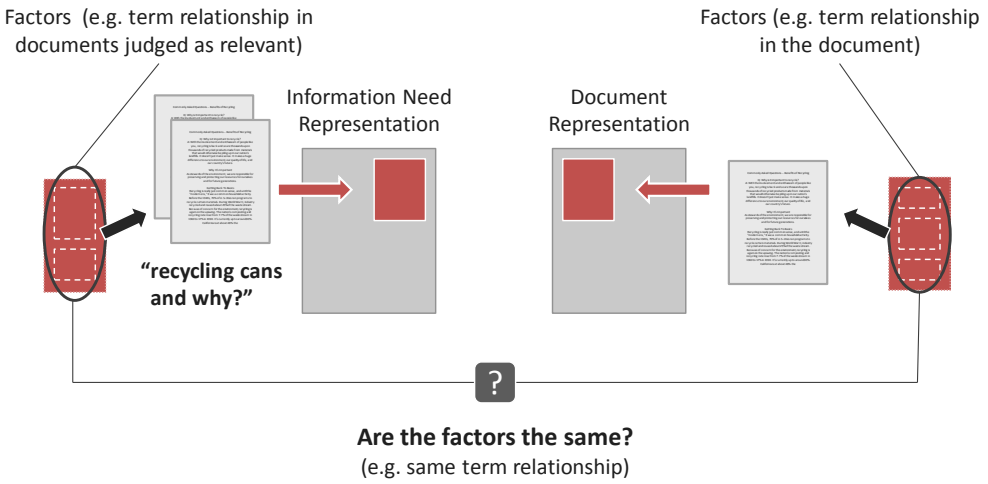


Figure 3.6: Prediction based on factors both of the information need and the document side.

exploiting evidence gathered from the first search, e.g. through explicit feedback, the system could automatically understand if the user is more interested in environmental than in economical benefits. Possible factors to capture this aspect are relationships among terms, e.g. described in terms of local co-occurrence data — e.g. if "benefit" tends to co-occur more near "environmental" than "economical" in documents judged as relevant, probably the user is more interested in environmental benefits. Once a model of the factors has been obtained, the idea is to use directly the models as a new dimension of the information need representation, instead of using observations — see Figure 3.5a. Given a document, the basic rationale underlying the prediction process is to measure the degree to which the document satisfies the new dimension of the information need representation. This is performed by measuring the degree to which the modeled factors explain the observation in the document — see Figure 3.5b.

In the above example we model factors only on the information need side, but the factors can be modeled for both the sides. For instance, if we extract term relationship both from the document feedback set and from each document in the collection, prediction could be done directly comparing factors, instead that using factors and observations — a pictorial description is reported in Figure 3.6. The basic rationale of this approach is similar to that underlying Probabilistic Distance Retrieval Models [Zhai, 2008] or the approach discussed in [Bodoff, 2004].

Although the underlying idea has been described using the example of the term relationships in feedback documents as factors, in this work the idea is generalized to sources of evidence other than content-based properties. In particular, the sources are the properties of informative resources involved in the search process. A set of factors

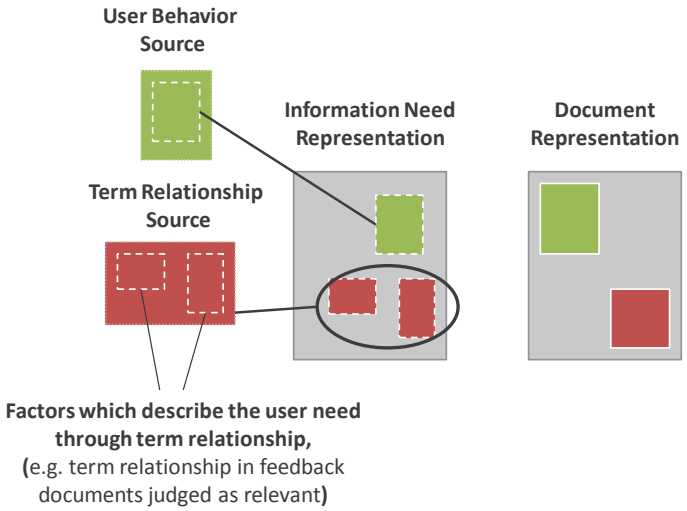


Figure 3.7: Factors from diverse sources to characterize the information need.

is obtained for each source and they, or a subset of them, constitute the dimension of the information need representation corresponding to the source. The selected factors are those that explain the relationship between the evidence obtained from the source and the user perception of relevance of a document with regard to his information need. For instance, in Figure 3.7 two sources are considered: term relationship in a document feedback set, and user behavior when visiting a subset of the obtained results. For each source a set of factors is obtained. The factors obtained from the term relationship source explain the relationship among the terms in the documents judged as relevant: these factors constitute the term relationship dimension. Analogously, factors obtained from the user behavior source constitute the user behavior dimension of the information need representation.

The motivation underlying this approach is to obtain a characterization of the information need not only based on the topic expressed by the query, but also on the basis of other units and relationship involved. This approach should allow the information need to be better characterized has suggested, for instance, by the polyrepresentation principle [Ingwersen, 1996].

The main problem is how to map this abstraction in a usable representation that can be adopted to support prediction using diverse sources of evidence. This objective is achieved by the methodology introduced in this work that models factors as basis vectors and document as vector according to the vector subspace framework proposed in [Melucci, 2008]. Recently, a geometric framework inspired by Quantum Theory (QT) [Frommholz et al., 2010], that is the dual of that adopted in this thesis,

was shown to be able to provide a usable representation of the ideas underlying the polyrepresentation principle, but focusing only on document-specific properties and ratings as form of interaction: the methodology introduced in this work aims at being more general and considering also other forms of interactions, e.g. the user behavior. Section 3.3 briefly describes the basic rationale underlying that framework; Section 3.4 describes a methodology to exploit this framework to model the diverse source contributions. Before focusing on the adopted geometric framework and the methodology to exploit it to model sources of evidence, some additional remarks on the notion of *factor* will be provided in Section 3.2.3.

3.2.3 Factor

Let us focus now on the notion of factor. A factor has been defined as one of the possible variables which explain the observations obtained from a source. The purpose of this section is to discuss the possible relationships among the diverse factors. Two different cases can be distinguished: relationship among factors concerning with the same source and relationship among factors concerning with diverse sources.

Factors concerning with the same source

Let us focus on a single source. The set of factors concerning with the source should be able to explain all the possible observations obtained from the source: factors should be *collectively exclusive*, that is, given an observation, at least one of the factors should be able to explain the observation. Among all the factors, those of interest are the factors that explain the user perception of relevance when considering a specific source of evidence. These factors are those that constitute the dimension of the information need representation corresponding with the considered source. The selection of the factors determines two subsets of the entire set of factors: the first subset is that constituting the dimension and the second that constituted by the remaining factors.

Factors should be collectively exclusive but not necessarily mutually exclusive. The relationship between factors depends on their meaning. For instance, if factors represent elementary concepts, then mutual exclusivity among factors should be modeled. Let us consider, for instance, the example of the student who is looking for information on benefits of cans recycling. Let us consider that, on the basis of the available evidence, the system is able to obtain a new characterization of the information need using possible relationships among the terms “recycle”, “cans”, “benefit”. Three possible cases are:

b_1 : there is a relationship between the three terms;

b_2 : there is a relationship between "recycle" and "cans", but no relationship with "benefit";

b_3 : there is no relationship between the term "recycle" and "cans".

The first case can be considered as the type of term relationship that characterizes documents of interest, where the main topic is possible benefits of recycling cans. The second case can refer to those documents that are focused on cans recycling, but not on the benefits, e.g. they talk about procedure for cans recycling. Finally, the last case can refer to documents not focused on cans recycling, e.g. that talk about recycling other materials than aluminum cans or benefits of recycling in general (when a relationship is present between "recycle" and "benefit"). If a user is looking for information on cans recycling and on possible benefits provided by it, an observation generated from the first two factors could interest him. If the user is not interested in methods for cans recycling but specifically on the benefits it can provide, it could be more interested in observations generated from the first factor only. The three possibilities above concern with the case of three possible topics covered in the documents, that are interpreted as elementary cases, where each one excludes the others. In this case the three factors are not only collectively exclusive, but also mutually exclusive. In general, factors should be not mutually exclusive.

Factors concerning with diverse sources

When considering diverse sources, the descriptor sets adopted to characterize them are not necessarily disjoint. In other words, sources can have descriptors in common. Let us consider, for instance, a dimension concerning with content-based descriptors and the behavior of the user when interacting with the results obtained from a first search. In the event of the recycling scenario, presence of query terms, e.g. "cans" and "recycling", in the title can affect the behavior of the user, e.g. click-through data [Clarke et al., 2007]. Therefore, the usable representation of factors should be able to represent also the cases of sources sharing common descriptors.

When considering factors concerning with diverse sources, it is required that factors modeling the two dimensions should be able to explain all the possible observations obtained when considered the two sources simultaneously. The relationship among the factors here depends on the relationship among the sources. For instance, if no relationship exist among the sources, they can be considered as distinct; then factors from the distinct sources can be extracted separately.

3.3 Utilizing Sources through Geometry

Section 3.1 and Section 3.2 describe an abstraction of the IR problem, how sources of evidence are interpreted and the basic rationale underlying the adoption of their contribution as new dimensions of the information need representation. The definition of the abstraction can be interpreted as addressing the problem at the conceptual level. The abstraction is not directly applicable to perform prediction: further steps are required. This section describes how using the mathematical construct of the vector space basis the problem can be addressed at the logical level. The abstraction defines a set of requirement should be met by the logical level. It should allow:

- addressing the complexity due the diversity among sources by uniformly modeling factors, the fact that they can be described by heterogeneous descriptors and that diverse sources can share descriptors;
- performing prediction by explicitly considering factors;
- measuring the degree to which an observation has been generated from the source.

Then starting from the logical level, specific instantiations for specific sources can be adopted to support the design and the development of an IR system. This is the approach actually adopted in this thesis and it constitutes the basic rationale underlying the methodology.

In this thesis it is suggested that the framework based on vector subspaces proposed in [Melucci, 2008] allows the logical level to be implemented for the problem described by the abstraction introduced in Section 3.1 and Section 3.2. The basic rationale of the framework is to exploit the fact that a vector can be generated by different vector space basis to model that an observation, e.g. concerning with a document, can be generated by different sources. With regard to the definitions introduced in Section 3.1, basis vectors can be adopted to model the diverse factors that explain the evidence gathered from a source; the dimension of the information need representation is the subspace spanned by those vectors.

In the remainder of this section we will show that the mathematical construct of the vector space basis allows the above requirements to be met. Section 3.3.1 will describe how the mathematical construct of the vector space basis can be adopted to represent a dimension and how a dimension can be adopted for document ranking. Section 3.3.2 will focus on the ranking function, specifically on the motivations behind the adoption of such function.

3.3.1 Modeling Dimension as Vector Space Basis

Preliminary Notation

Before focusing on the way the vector space basis can be adopted to model factor and dimension as defined in Section 3.2, some preliminary notation will be introduced.

Let us denote with $\mathcal{F} = \{f_1, \ldots, f_{|\mathcal{F}|}\}$ the set of all the descriptors can be adopted to characterize all the possible sources involved in the search process. A generic observation o_i can be represented by a vector $\mathbf{o}_i = \sum_{j=1}^{|\mathcal{F}|} f_{ij}\mathbf{e}_j = [f_{i1}, \ldots, f_{i|\mathcal{F}|}]^T \in \mathbb{R}^{|\mathcal{F}|}$ where f_{ij} is the value observed for the descriptor f_i when considering the observation o_i and $e_{ii} = 1$ and $e_{ij} = 0$ when $i \neq j$ — in other words each descriptor is represented by a vector of the canonical basis $[\mathbf{e}_1 \cdots \mathbf{e}_{|\mathcal{F}|}] = \mathbf{I}_{|\mathcal{F}|\times|\mathcal{F}|}$. When focusing on a specific source $\mathcal{S}$ let us denote with $\mathcal{F}_\mathcal{S} = \{f_1, \ldots, f_{|\mathcal{F}_\mathcal{S}|}\} \subseteq \mathcal{F}$ th set of descriptors adopted to characterize $\mathcal{S}$. A generic observation o_i can be represented by a vector $\mathbf{o}_i = \sum_{j=1}^{|\mathcal{F}_\mathcal{S}|} f_{ij}\mathbf{e}_j = [f_{i1}, \ldots, f_{i|\mathcal{F}_\mathcal{S}|}]^T \in \mathbb{R}^{|\mathcal{F}_\mathcal{S}|}$ where f_{ij} is the value observed for the descriptor f_i when considering the observation o_i.

Modeling factors concerning with the same source

When considering factors concerning with the same source, it is required that a generic observation obtained from the source can be explained by at least one of the factors of the source, that is factors should be collectively exclusive. That reminds the notion of *spanning set* of a vector space. A set of vectors $\mathcal{G} = \{\mathbf{g}_1, \ldots, \mathbf{g}_{|\mathcal{G}|}\}$ is a spanning set for the vector space V if each $\mathbf{v} \in V$ can be expressed as a linear combination of vectors in $\mathcal{G}$, namely $\mathbf{v} = \sum_{i=1} \gamma_i \mathbf{g}_i$. $V = span(\mathcal{G})$ denotes that $\mathcal{G}$ is a spanning set for V. A vector space basis $\mathcal{B} = \{\mathbf{b}_1, \ldots, \mathbf{b}_{|V|}\}$ of V is a minimal spanning set for V; in the context of this thesis, a basis is a minimal set of factors that allows to explain all the observations obtained from a source. For instance, when considering relationship among terms "recycling", "cans" and "benefit", possible basis vectors to represent the three relationships mentioned in the example are:

$$\mathbf{b}_1 = \begin{bmatrix} 0.685 \\ 0.685 \\ 0.247 \end{bmatrix} \qquad \mathbf{b}_2 = \begin{bmatrix} 0.174 \\ 0.174 \\ -0.969 \end{bmatrix} \qquad \mathbf{b}_3 = \begin{bmatrix} 0.707 \\ -0.707 \\ 0.000 \end{bmatrix} \tag{3.1}$$

The value of the jth element $\mathbf{b}_{i_j}$ of the vector $\mathbf{b}_i$ still refers to the term j but differently from $\mathbf{f}_i$'s here it represents the role of the term in the relationship. The concordance in sign can be adopted to model the existence of a direct relationship among related terms. For instance, $\mathbf{b}_1$ can be adopted to model the case where a direct relationship exists among all three terms, e.g. when the topic concerns with "benefits of can recycling". The vector $\mathbf{b}_2$ can represent the case where a direct relationship exists among the

terms "recycling" and "cans", but not with "benefit", e.g. the case where the topic under consideration is procedure for recycling cans or aspects other than benefits provided by this procedure. Finally, $\mathbf{b}_3$ can represent the case where the topic covered is recycling objects other than cans since for the first two elements of the vectors, namely those corresponding to the terms "recycling" and "cans" the relationship is negative. In the above example, each of the three relationships of the considered example excludes the other two, namely they are mutually exclusive. As proposed in [van Rijsbergen, 2004] mutual exclusivity can be modeled by orthogonality among vectors — $\mathbf{b}_i$'s, for example, are mutually orthogonal.

A generic document observation is a vector $\mathbf{d} \in \mathbb{R}^3$ and, being the $\mathbf{b}_i$'s a basis, all the possible observations can be generated from them. But the user is not interested in all the observations, namely the documents. For instance, the user can be interested in gaining knowledge on recycling cans and its possible benefits, that is the factors represented respectively by $\mathbf{b}_1$ and $\mathbf{b}_2$. The dimension of the information need representation corresponding to the need of the recycling scenario can be therefore modeled through the vector space basis constituted by $\mathbf{b}_1$ and $\mathbf{b}_2$. In particular, the dimension corresponds to the subspace $S_{12} = span(\{\mathbf{b}_1, \mathbf{b}_2\})$ — a pictorial description is reported in Figure 3.8c.

The prediction of the degree to which a document satisfies the modeled dimension of the information need representation can be performed by measuring the degree to which the document vector has been generated by the vector space basis, namely the factor vectors, that spans the dimension subspace. In particular, this is measured by the distance between the document vector and its projection onto the dimension subspace — the motivations for the adoption of this measure will be discussed in Section 3.3.2. For instance, the degree to which document $\mathbf{d}$ satisfies the dimension represented by S_{12} can be obtained measuring the distance between $\mathbf{d}$ and its projection $\mathbf{d}'$ onto S_{12} — a pictorial representation of the prediction process is reported in Figure 3.8d.

Using vector space basis to model source factors allow less stringent factor representation to be adopted. Indeed, as mentioned in Section 3.2.3, factors do not necessarily be mutually exclusive: whether or not adopting mutually exclusivity depends on the meaning of the factors. Two other possibilities are:

i. factors in the dimension are mutually exclusive with those not in the dimension;

ii. factors are not mutually exclusive.

Each factors in the framework is modeled as a basis vector. In the first of the two cases, the requirement is that factor vectors in the dimension should be orthogonal with those not in the dimension. For instance, in the event of the "concept factors",

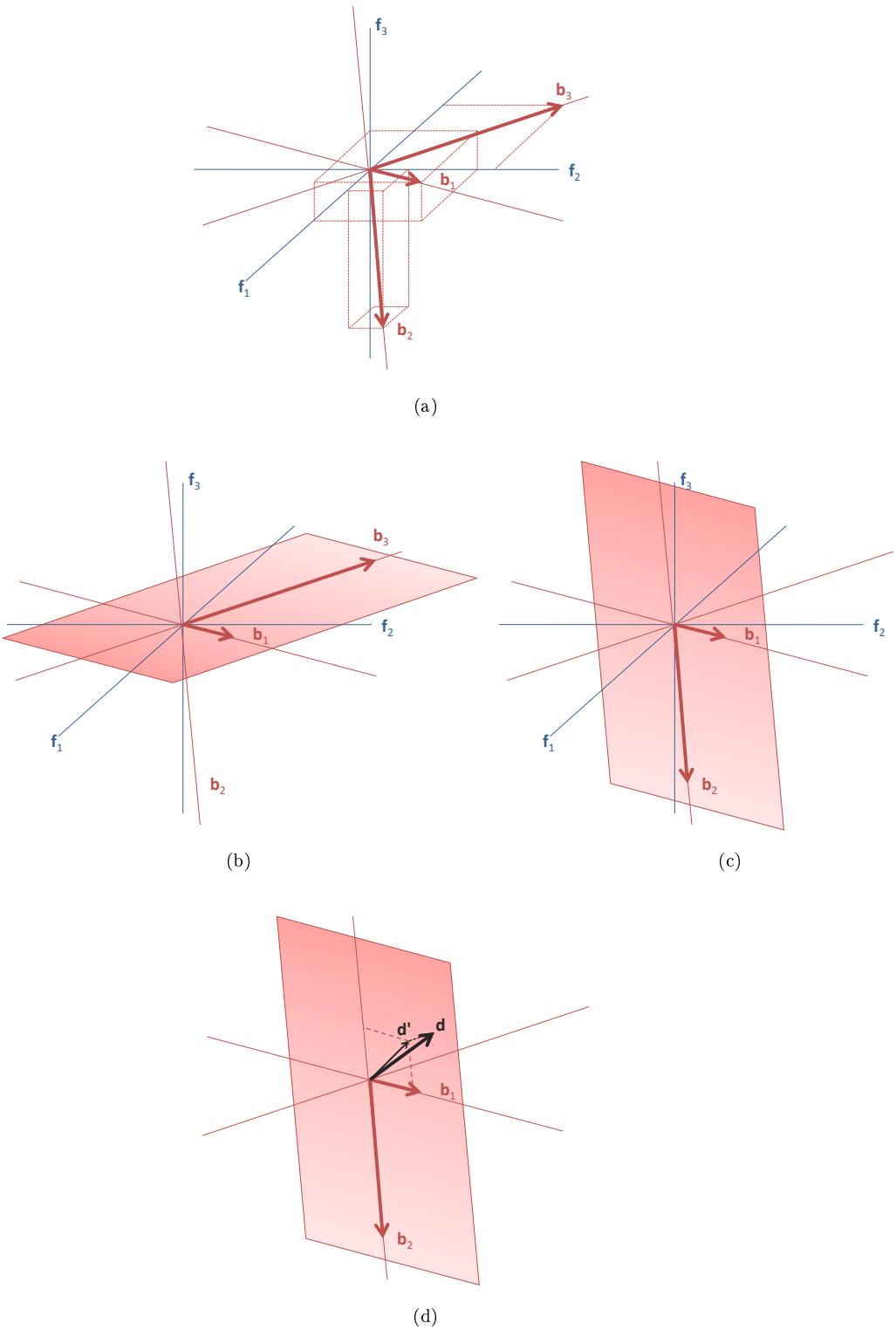


Figure 3.8: Pictorial description of factor and dimension modeling through vector space basis (Figures 3.8a–3.8c) and prediction by projectors (Figures 3.8d).

the dimension could be represented by diverse vectors $\{\mathbf{b}'_1, \mathbf{b}'_2\}$ not orthogonal each other but orthogonal with the factor not in the dimension represented by $\mathbf{b}_3$, e.g.:

$$\mathbf{b}'_1 = \begin{bmatrix} 0.577 \\ 0.577 \\ 0.577 \end{bmatrix} \qquad \mathbf{b}'_2 = \begin{bmatrix} 0.707 \\ 0.707 \\ 0.000 \end{bmatrix} \qquad \mathbf{b}_3 = \begin{bmatrix} 0.707 \\ -0.707 \\ 0.000 \end{bmatrix} \tag{3.2}$$

Here values in the factor vectors can be interpreted as correlation, e.g. among term occurrences, adopted to model term relationship. $\mathbf{b}'_1$ can refer to the case where the three terms, "recycle", "cans" and "benefit" are related to each other since positively correlated. $\mathbf{b}'_2$ can refer to the case where "recycle" and "cans" are related, but no relationship exists with "benefit" — the first two terms are correlated each other but uncorrelated with "benefit". This new dimension representation can be obtained, for instance, through a subsequent feedback step related to a new interaction between user and system, and adopted to further improve the information need representation.

The most general case is when factors are not mutually exclusive. In this case, vectors representing factors should be independent but not orthogonal each other. Independence is required since they should constitute a vector space basis, which is the mathematical construct adopted to model a dimension.

Modeling factors concerning with diverse sources

Let us focus now on the case where factors belonging to multiple sources are considered. Let us consider two sources $\mathcal{S}_1$ and $\mathcal{S}_2$, e.g. behavior of the user when interacting with the results or the corresponding documents and a content-based property. Let us consider the following descriptors (three descriptors have been considered thus allowing a representation in a three–dimensional space):

- f_1 : *display-time*, i.e. the time spend by the user on the document;
- f_2 : *query keywords in title*, i.e. the number of query keywords present in the title of the document;
- f_3 : *query keywords in document content*, i.e. the sum of the term weights in the document, e.g. TFIDF or BM25.

Let us suppose that $\mathcal{S}_1$ is described by $\mathcal{F}_{\mathcal{S}_1} = \{f_1, f_2\}$, while $\mathcal{F}_{\mathcal{S}_2} = \{f_2, f_3\}$, i.e. $\mathcal{F}_{\mathcal{S}_1} \cap \mathcal{F}_{\mathcal{S}_2} \neq \emptyset$. Each descriptor f_i can be modeled as a vector of the canonical basis, e.g. $[\mathbf{f}_1\ \mathbf{f}_2\ \mathbf{f}_3] = \mathbf{I}_9$. When considering a source at a time, the factors will lie on the subspace spanned by the descriptor vectors, e.g. $span(\{\mathbf{f}_1, \mathbf{f}_2\})$ in the event of $\mathcal{S}_1$ and $span(\{\mathbf{f}_2, \mathbf{f}_3\})$ in the event of $\mathcal{S}_2$. Therefore, modeling factors and predicting relevance by considering a single source will correspond to focus on a subspace of the entire space

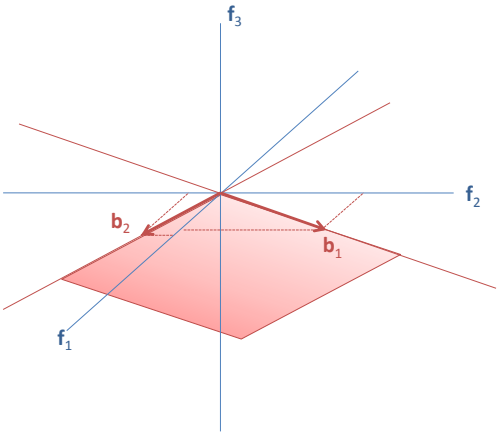
spanned by $\{\mathbf{f}_1, \ldots, \mathbf{f}_{|\mathcal{F}|}\}$ where $\mathcal{F} = \{f_1, \ldots, f_{|\mathcal{F}|}\}$ is the set of all the descriptors can be adopted to characterize all the sources involved in the search process. In other words, in the event of a single source $\mathcal{S}_i$ we will restrict to a $|\mathcal{F}_i|$-dimensional space instead of considering the entire $\mathcal{F}$-dimensional space. When considering multiple sources $\mathcal{S}_1, \ldots, \mathcal{S}_k$, we focus on a $|\mathcal{F}_{\mathcal{S}_1} \cup \cdots \cup \mathcal{F}_{\mathcal{S}_k}|$-dimensional space; for instance, in the event of the sources $\mathcal{S}_1$ and $\mathcal{S}_2$ a three-dimensional space is considered since the two sources have a descriptor in common.

When diverse source are considered a possible approach is to model factors for the different sources by considering the sources as distinct. This is actually the approach on which this thesis will be focused. In this case a set of factors will be identified for each source and a subset of them will be selected to model the dimensions corresponding to the source. For instance, $\{\mathbf{b}_1, \mathbf{b}_2\}$ could be possible factors obtained from the evidence gathered from source $\mathcal{S}_1$ and $\{\mathbf{c}_1, \mathbf{c}_2\}$ from evidence of the source $\mathcal{S}_2$ — a pictorial description for the two sources $\mathcal{S}_1$ and $\mathcal{S}_2$ is reported respectively is Figure 3.9a and Figure 3.9b. A subset of the factors is then selected to model the two dimensions, e.g. $\mathbf{b}_1$ for source $\mathcal{S}_1$ and $\mathbf{c}_2$ for source $\mathcal{S}_2$. Another possible approach is not considering the sources as independent, but modeling factors by exploiting all the descriptors in $\mathcal{F}_{\mathcal{S}_1} \cup \cdots \cup \mathcal{F}_{\mathcal{S}_k}$. When sources share descriptors the latter approach could be beneficial to capture possible relationships among the sources.

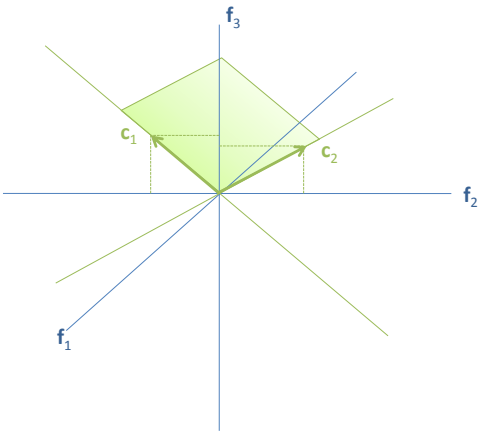
3.3.2 Prediction by Projectors

Let us assume that a basis $\mathcal{B} = \{\mathbf{b}_1, \ldots, \mathbf{b}_h\}$ has been computed to model a dimension or, equivalently the dimension is represented by the subspace $span(\mathcal{B})$. Each document among those to (re)rank can be described on the basis of the descriptors used to characterize the source corresponding to the dimension — the assumption is that the descriptors for that document are available; Section 4.2.1 will discuss this issue when the user behavior is adopted as source and personal behavioral descriptors are not available. If $\mathcal{F}_\mathcal{B} = \{f_1, \ldots, f_k\} \subseteq \mathcal{F}$, is the set of descriptor for the source $\mathcal{B}$, a generic document d can be described as a linear combination of the descriptor vectors $\mathbf{f}_i$'s as discussed in the previous section, i.e. $\mathbf{d} = \sum_{i=1}^{|\mathcal{F}_\mathcal{B}|} \gamma_i \mathbf{f}_i$; in the entire descriptor space $\mathbf{d} = \sum_{i=1}^{|\mathcal{F}|} \gamma_i \mathbf{f}_i$, where $\gamma_i = 0$ when $f_i \notin \mathcal{F}_\mathcal{B}$. The objective of the prediction is to measure the degree to which this observation has been generated by the basis $\mathcal{B}$.

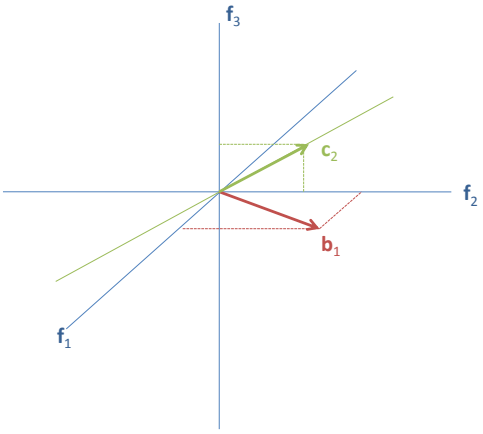
The adopted interpretation of the notion of dimension is similar to that underlying Factor Analysis where a set of unobservable variables are supposed to generate the observed data. The basis vectors can be interpreted as models of these variables. For a document not among those used to model the variables, we are interested in the degree to which such document has been generated by these variables. If $\mathbf{d}$ has been generated by the basis $\mathcal{B}$, then it lies in the subspace $span(\mathcal{B})$, so can be expressed as



(a)



(b)



(c)

Figure 3.9: Pictorial description concerning factors of multiple sources

a linear combination of the basis vectors constituting $\mathcal{B}$, namely $\mathbf{d} = \sum_{i=1}^{h} \beta_i \mathbf{b}_i$. In order to predict if $\mathbf{d}$ has been generated by $\mathcal{B}$ or in which degree it has been generated by $\mathcal{B}$, the approach we will adopt is to obtain the best predicted observation $\mathbf{d}_{\mathcal{B}}$ for d that can be produced by the basis $\mathcal{B}$. Then we measure the distance $m_{\mathbf{B}}(\mathbf{d})$ between $\mathbf{d}$ and $\mathbf{d}_{\mathcal{B}}$. The expression "best predicted observation" here is interpreted in a *least square* sense: the best predicted observation is the one that minimizes the difference between the actual observation $\mathbf{d}$ and the possible predicted observations among those generated by $\mathcal{B}$, namely $\mathbf{d}_{\mathcal{B}} = \sum_{i=1}^{h} \hat{\beta}_i \mathbf{b}_i$ with

$$\hat{\boldsymbol{\beta}} = arg\min_{\boldsymbol{\beta}} \; ||\boldsymbol{d} - \sum_{i=1}^{h} \beta_i \mathbf{b}_i||_2 = arg\min_{\boldsymbol{\beta}} \; ||\mathbf{d} - \mathbf{B}\boldsymbol{\beta}||_2 \tag{3.3}$$

where $||\cdot||_2$ is the L^2 norm, $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_h]^T$, and $\mathbf{B} \in \mathbb{R}^{n \times h}$ is the matrix whose columns are the basis vectors. The problem is actually reduced to find a solution to the *General Least Squares Problem* [Meyer, 2000]:

Definition 10 (General Least Squares Problem)**.** For a matrix $\mathbf{B} \in \mathbb{R}^{n \times h}$ and $\mathbf{d} \in \mathbb{R}^n$, let $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}(\boldsymbol{\beta}) = \mathbf{d} - \mathbf{B}\boldsymbol{\beta}$. The general least square problem is to find the vector $\boldsymbol{\beta}$ that minimizes the quantity

$$\sum_{i=1}^{h} \epsilon_i^2 = \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} = (\mathbf{d} - \mathbf{B}\boldsymbol{\beta})^T (\mathbf{d} - \mathbf{B}\boldsymbol{\beta}) \tag{3.4}$$

Any vector that provides a minimum value for the above expression is called *least squares solution.*

If $rank(\mathbf{B}) = h$ it can be shown — see [Meyer, 2000], Chapter 4, page 226 — that the unique solution is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{d} \tag{3.5}$$

In our case the condition $rank(\mathbf{B}) = h$ is satisfied since the columns of $\mathcal{B}$ are the basis vectors modeling the source, and being basis vectors they need to be independent. The estimated $\hat{\boldsymbol{\beta}}$ leads to the following value of the best predicted observation:

$$\mathbf{d}_{\mathcal{B}} = \mathbf{B}\hat{\beta} = \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{d} \tag{3.6}$$

which is the projector onto the subspace spanned by the basis $\mathcal{B}$. Thus the degree to which the document $\mathbf{d}$ has been generated by the basis $\mathbf{B}$ can be computed as

$$m_{\mathbf{B}}(\mathbf{d}) = \mathbf{d}^T\mathbf{d}_{\mathcal{B}} = \mathbf{d}^T\mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{d}. \tag{3.7}$$

In the special case where $\mathbf{B}$ is an orthonormal vector space basis, the expression reported in Equation 3.7 becomes

$$m_{\mathbf{B}}(\mathbf{d}) = \mathbf{d}^T\mathbf{B}\mathbf{B}^T\mathbf{d} = \mathbf{d}^T\mathrm{P}_{\mathbf{B}}\mathbf{d}. \tag{3.8}$$

Linear Regression Interpretation

The interpretation of the IR process as a multiple linear regression problem has been originally suggested in [Story, 1996], and further discussed also in [Efron, 2008] when comparing the notions of optimality adopted in the Rocchio feedback algorithm [Rocchio, 1971] and LSI [Deerwester et al., 1990]. In this thesis the linear regression interpretation is adopted only to motivate the choice of the ranking function adopted. One motivation for the adoption of this approach is that the obtained estimator $\hat{\boldsymbol{\beta}}$ under certain assumption is the Best Linear Unbiased Estimator (BLUE), where the meaning of "best" is that the estimator is the one with minimum variance among all the linear unbiased estimator. In a linear regression perspective the problem can be formulate as follows. A document d is interpreted as a variable D linearly dependent from a set of independent variables $\{B_i\}_{i=1\ldots h}$ plus an additional term ϵ which models the contribution of all the other variables which affect the dependent variable D; more formally:

$$D = \sum_{i=1}^{h} B_i \boldsymbol{\beta}_i + \epsilon \tag{3.9}$$

where β_i are named *regression coefficients*. Under the assumptions that $E[\boldsymbol{\epsilon}] = 0$ and $cov[\boldsymbol{\epsilon}] = \sigma^2 I_h$, namely the errors averaged out to zero[3] and have fixed finite variance, the Gauss–Markov Theorem states that the minimum variance linear unbiased estimator for β_i is given by the ith component $\hat{\beta}_i$ in $\hat{\boldsymbol{\beta}} = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{d}$, namely that the BLUE is the least square solution of $\mathbf{B}\hat{\beta} = \mathbf{d}$.

The remarks reported in this section suggest that diverse sources can be exploited through a unique ranking function, provided that a vector subspace representation of the corresponding dimensions has been obtained. But the complexity of dealing with diverse sources is not removed. The main cause of complexity is now the procedure to obtain a vector subspace representation for the dimension and the document (a document is actually represented as a vector that is a one dimensional subspace). The methodology introduced in the next section aims at addressing this issue.

3.4 Methodology Description

Previous sections described an abstraction for uniformly interpreting the contribution of diverse sources and a geometric framework that preserves this uniform interpretation

[3] We can reasonably assume that the first assumption is valid since it means that the other possible variables affecting the document observation has no impact: in the event of a single source they can be the other $k - h$ basis vectors not considered because not carrying additional information — for instance in LSI as suggested in [Efron, 2008] the eigenvectors corresponding to small eigenvalues are not considered since they are interpreted as random errors, that is they do not provide additional information respect to the selected eigenvectors.

by uniformly modeling the source contributions as vector space basis. Indeed, dimensions, factors and documents are modeled as subspaces — factors and documents are vectors that are actually one dimensional subspaces. The dimension corresponding to a generic source can be adopted to support document (re)ranking by means of the same projector-based ranking function.

The main issue now is how to obtain a dimension and a document representation in terms of the considered framework when a generic source is considered. This section introduces a methodology to achieve this objective. The methodology needs to be general enough to be applicable to a generic source. Moreover, since the final aim is the design and the development of an IR system able to exploit diverse sources, the methodology needs to assist the design of the system. The identified steps are: source selection, evidence collection, dimension modeling, document modeling, and prediction. Each of the following sections will be focused on a specific step, making explicit the issue should be addressed when applying it for a specific application. An instantiation of the methodology for two diverse source is then described in Chapter 4.

3.4.1 Source Selection

The first methodology step consists in the selection of the source. The term "source" refers to a property of an informative resource that can be adopted to support prediction. The selection of the source corresponds to the identification of the hypothesis on possible factors that affected the user perception of relevance. Let us consider, for instance, the case of relevance feedback strategies. Let us suppose that a set of documents judged as relevant by the user are available. Although they are known to be relevant the user provides no motivation that explains why he perceived those documents as relevant. Documents judged as relevant are not the actual source: the source corresponds to a possible class of factors that affected the user perception of relevance. For instance, as actually done in this work, we can hypothesize the relationship among terms in feedback documents is a possible class of factors. Then the validity of this hypothesis can be tested by means of the experimental approach.

The selected hypothesis determines the set of informative resources involved and from which features can be distilled. For instance, Figure 3.10 depicts possible informative resources from which features can be distilled when diverse feedback strategies are adopted. In the case of PRF the resources adopted are the top ranked documents, e.g. the top five in the figure, namely, those corresponding to the results with a highlighted background in Figure 3.10a. In the event of an IRF strategy, e.g. the specific methodology application investigated in Section 4.2, the informative resources are the relationship between the user and a subset of the visited documents, e.g. those corresponding to the results depicted in Figure 3.10b with a highlighted background. In

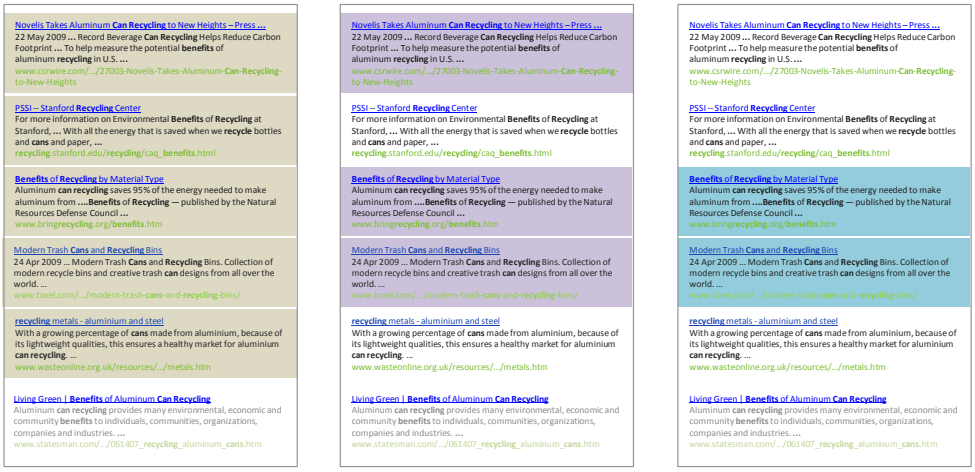


Figure 3.10: Instances of feedback sets adopted in different feedback strategies.

the event of explicit feedback, e.g. the methodology application investigated in Section 4.1, the resources could be the documents judged as relevant among those visited, e.g. those corresponding to the results depicted in Figure 3.10c with a highlighted background. In general, the sources for features considered in this work are a subset of the informative resources involved in the search process between the first and the second stage prediction.

The source selection step is affected by the particular domain of application which the IR system is designed for. In a Desktop Search scenario, properties of the local document collection, browsing and activity history can be crucial. In a Mobile Search scenario properties of the location of the user when searching for information could provide valuable information to support prediction. Let us consider a user who is looking for restaurant in London. If he is planning a travel or a dinner that is not happening very soon, an arbitrary restaurant can be interesting for the user. If is in London at dinner time, a restaurant near his current location can be useful — here both the property time and the property location, e.g. described by its GPS position, are involved. Another example is an Enterprise Search scenario where content of diverse document collections, e.g. e-mails, of the diverse employers are available and can be used as source, e.g. as done in [Teevan et al., 2009] as input for groupization algorithms. The selection of the source is affected also by the availability of the features. This is, for instance, the case of the interaction features adopted in this work to characterize the source user behavior. This issue will be specifically discussed in Section 4.2 when considering an application of the methodology for the user behavior

dimension.

By summarizing, the source selection step implies

- the specification of the hypothesis on the class of factors which can affect the user perception of relevance;
- the selection of the (instances of) informative resources then adopted as source from which distill feature values.

3.4.2 Evidence Collection

Once the source has been selected, the next step consists in the collection of the evidence then adopted to obtain a model of the dimension and of the document with regard to the considered source. In particular, this step requires two sub-steps:

- the actual collection of the descriptors and of the features used to characterize the descriptors or used as descriptors;
- the selection of the descriptors then adopted to extract the factors and to represent the documents.

The first of the above sub-steps should be considered in the design of the system since it determines the specific modules to be implemented. When considering document-specific descriptors, their features can be extracted at indexing time and stored in appropriate data structure for efficient access. But the main motivation behind the investigation of the methodology is to exploit diverse sources that are involved in the search process, and particularly properties that can be adopted to characterize the diverse ways in which a user interacts with the system. Indeed, the hypothesis is that user–system interaction, e.g. exploited through feedback techniques, is a valuable source for information to better characterize the user information need. Exploiting the diverse form of interaction requires appropriate modules to monitor features. An example, that is actually investigated in this work, is the interaction features that can be monitored by observing the behavior of the user when interacting with the results returned by the system or the system in general. When the user interaction behavior is adopted as source a monitoring tool should be integrated in the system or used to support the system. The monitoring functionality can be provided by a system logger, a browser extension or developed directly inside the application adopted to access the IR system. The latter approach is that adopted in this work to build the test collection described in Section 5.2.3.2.

The evidence collection step is not limited to the collection of the values of the possible descriptors (and/or features) can be observed when a specific source is considered. Indeed, not all the descriptors necessarily provide a useful contribution to

model the dimension. For this reason, after the actual collection of the descriptors, a subsequent *descriptor selection* process can be required — this could be considered an instance of the feature selection step adopted in statistics and machine learning. The features collected for each descriptor are adopted to provide quantitative information on it and to support the selection process. An instance of descriptor selection application is when, given a set of feedback documents, a subset of the terms should be selected, e.g. to expand the query. Indeed, not necessarily all the terms appearing in the feedback documents are good descriptor of the user interests or intents; including all of them could negatively affect the modeling procedure. The term selection procedure can be based on a single feature value, e.g. term frequency (TF) or IDF, on a score derived from different feature value, e.g. Local Context Analysis (LCA), or a number of features.

The selection strategy is related to the selected source not only because of the features adopted to characterize the source descriptors (or adopted as descriptors). The selection strategy could be determined also by the hypothesis underlying the source, namely the class of factors we are modeling. For instance, if the factors are possible behavior patterns modeled by the correlation among interaction features, a preliminary step is necessary to remove uncorrelated features, e.g. those that assume the same value in all the feedback set; these features are discarded since not meaningful to identify variation between document observations.

Both the abstraction and the methodology aim at being general. In the remainder of this section we will present a specific case study for descriptor selection when applied to another media, namely music. The module for descriptor selection has been developed for a cover identification engine whose design has been based on the abstraction introduced in Section 3.1.

FALCON: Descriptor Selection for Cover Identification

The problem of descriptor selection is not only limited to the textual case. An example is the *query pruning* strategy adopted in FALCON[4], an open source search engine for content-based cover song identification [Di Buccio et al., 2010b, Di Buccio et al., 2010a]. The objective of a cover song identification engine is the automatic identification of different performances of the same song. In FALCON this objective is achieved using one text retrieval approach by means of a methodology that allows a bag of feature representation of a song to be obtained. The methodology consists, firstly, in the representation of a song as sequence of excerpts; each excerpt is then represented as a sequence of chroma vectors, each of them then mapped in a hash, namely an integer value — the adopted hashing strategy is described in [Miotto and Orio, 2008]

[4] http://ims.dei.unipd.it/falcon/

and a complete description of the methodology implemented in FALCON is described in [Di Buccio et al., 2010b]. The final result of the representation steps is that each song is represented as a sequence of hashes, where each hash can be interpreted as an index term. The entire sequence is then divided in possibly overlapping segments: if an hash is treated as an index term, each segment can be interpreted as a passage. Basically, the methodology steps to obtain an hash-based representation allows the problem to be interpreted according the abstraction described by the definitions introduced in 3.1.

The "query" in FALCON is the entire song that, as described above, is a set of segments. Typically, the number of hash per segment is constituted by approximately one thousand and a song is constituted by multiple segments. Here, a strategy for descriptor selection is required in order to speed up the retrieval process, above all when no parallelization is adopted. The descriptor selection strategy adopted in FALCON characterizes each hash by a set of of features $\Phi = \{\phi_1, \ldots, \phi_k\}$, where each of them is normalized thus being a value in $[0, 1]$ — an instance of feature adopted is the term frequency normalized over the length of the excerpt, namely the number of hashes in the segment. Each feature is then characterized by an interval $I_\phi = [min_\phi, max_\phi]$ and a weight w_ϕ that are trained by a randomized hill climbing strategy — the objective function currently adopted privileges speed while maintaining sufficient accuracy results, where accuracy is measured in terms of Mean Reciprocal Rank (MRR). When parsing a segment of the song used as query, an hash is considered for the query processing step if the pruning strategy retains it. The criterion adopted by the pruning strategy to decide whether to retain an hash h or prune it is the following:

1. set the initial score of the hash h to zero, namely $s_h = 0$
2. for each feature $\phi \in \Phi$
   a. let $v_{h\phi}$ be the value of the feature ϕ for the hash h
   b. if $v_{h\phi} \geq min_\phi$ and $v_{h\phi} \leq max_\phi$ then
      $s_h \leftarrow s_h + w_\phi$
3. the hash is pruned h if $s_h < \rho$, where ρ is a predefined threshold.

3.4.3 Dimension Modeling

The complexity of dealing with the diversity among the sources and exploit them through a unique ranking function can be addressed by the adopted framework, provided that a representation of factors in terms of basis vector has been obtained. Therefore, a first constraint that the model poses is to represent factors as basis vectors. As discussed in Section 3.3.1, we can exploit the different ways in which the

vectors can be related to each other in order to model different class of factors. The class of factors is determined by the hypothesis on the user perception of relevance when a specific source is considered. For instance, orthogonal vectors can be adopted to model elementary factors, that exclude each other; differently, non-orthogonality can be be when a relationship among the diverse factors needs to be modeled.

Considering the hypothesis on the possible class of factors that affect the user perception of relevance is a crucial point. The motivation behind the methodology is indeed, not only exploiting diverse sources, but also investigating diverse hypotheses. We are not therefore interested in a generic basis that can be obtained from the collected data, i.e. descriptor feature values. This is actually a possible approach to obtain a dimension model. The main limitation is that a generic basis does not allow to immediately understand the nature of the user perception of relevance starting from the available evidence.

The general idea underlying the approach adopted in this thesis to model a dimension can be described as follows. A matrix $\mathbf{F} \in \mathbb{R}^{n \times k}$ can be prepared with all the observation that constitute the evidence gathered from one or more interaction episodes that involve the user; n is the number of observation and k is the number of descriptors. The descriptors are those selected during the evidence collection step. The target of the modeling step is a matrix $\mathbf{B} \in \mathbb{R}^{k \times s}$ whose columns are factors; $s < k$ is the number of factors selected to model the dimension among all the possible factors obtained from the source. The modeling step consists therefore in defining and applying a mapping $L : \mathbb{R}^{n \times k} \rightarrow \mathbb{R}^{k \times s}$, where the factors constituting the columns of B belong to the class of factors determined by the hypothesis.

The specific methodology applications considered in Chapter 4 will investigate relationship among descriptors as possible factors to model a dimension. The approach adopted is obtain a correlation matrix between descriptors and then obtain a basis from that matrix, applying matrix decomposition techniques — the aim of these techniques is also to reduce the noise in the gathered data. The result of this first step is a set of basis vectors that are able to explain all the possible observations can be obtained, when a specific source is considered and the selected set of descriptors is adopted. A subset of the factors is then selected among those extracted. The result of the modeling procedure is a set of s basis vectors; the subspace spanned by these vectors is the model of the dimension.

3.4.4 Document Modeling and Prediction

In the adopted framework, when sources are considered as distinct, a dimension model and a document representation with regard to the source descriptors is required for all of them. In this case the document is represented as a vector, where each entry

correspond to a descriptor and the value assumed by the entry could be the value of a feature of the descriptor or a value derived from a set of its features.

The document modeling step can consist, for instance, only in the representation of the document as a vector, where each entry corresponds to a descriptor. In this case the idea behind the prediction step is to find a representation of the document observation on the basis of the factors in the dimension, and then measure the degree to which this observation corresponds to the one actually observed.

The expression "document modeling" has been adopted since factors could be modeled not only on the information need side, both also in the document side, as discussed in Section 3.2.2. That implies a different prediction procedure, where the factors of both the sides are explicitly considered. A possible approach to perform this kind of prediction in the adopted geometric framework is through the function proposed in [Melucci, 2005] and briefly reviewed in Section 2.1: $s(q,d) = (\mathbf{Ad})^T(\mathbf{Bq})$. Factors obtained from the dimension modeling step can be adopted to model the basis for the information need side, i.e $\mathbf{B}$. The objective of the document modeling step is to obtain the basis $\mathbf{A}$ where each basis vector in the adopted framework represent one of the factors which explain the document observation with regard to the source.

One of the issues to address when exploiting the second approach is analogous to that affecting Model 1 in [Robertson et al., 1982], i.e. the model proposed in [Cooper and Maron, 1978]. In Model 1 documents are ranked by the probability that a document will be judged as relevant by a user who submitted a query q. That probability can be interpreted in a frequency sense: for instance, if a query constituted by a single term t is considered, the probability can be estimated as the ratio between the number of users that issued query q and judged the document relevance over all the user the issue query q. Following the interpretation proposed in [Bodoff and Robertson, 2004], that approach aimed at minimizing the error in document indexing, or more in general in the document representation. The problem is that this approach requires multiple judgments from diverse users on the same document when the same descriptor is adopted to characterize the information need, e.g. a term in the query. With regard to the work reported in this thesis, using a single observation as evidence to obtain factors on the document side can affect the reliability of document representation. The lack of availability of multiple judgments for the same document with regard the same query makes this estimation difficult, even impossible. A possible approach is to rely on other evidence. For instance, it could be easier to gather interaction features from diverse users with regard to the same document and the same query than gathering explicit judgments. This is actually the approach we will adopt in the methodology application for user behavior described in Section 4.2. Two representation of a document will be represented: the first representation will be

based on the feature values observed from the behavior of the individual user, while the second will exploit feature values obtained from a group of users that issued the same query. The comparison between the two representations will allow us to acquire some insights on their reliability.

CHAPTER

FOUR

METHODOLOGY APPLICATIONS

The methodology proposed in Section 3.4 aimed at being not tailored to a predefined dimension. In this chapter two applications are discussed with regard to two specific sources. Section 4.1 discusses an application of the methodology when relationships among terms in the feedback documents are adopted as source. Section 4.2 will focus on the behavior of the user described in terms of post-search interaction features observed from the first documents visited by the user when searching for information relevant to his formulated query.

4.1 Methodology for Term Relationship Dimension

The methodology application discussed in this section concerns with relationship among terms in the documents judged as relevant. The basic rationale is not to consider terms in the (possibly expanded) query as unrelated to each other, but capture and model possible relationships from statistical information on terms in the feedback documents. The dimension obtained from the modeling step aims at being a new representation of the information need that explicitly takes into account these relationships. The main research question is if the modeled dimension is an effective information need representation, i.e. it can help the system better understand the user intents. Starting from the original proposal reported in [Melucci, 2008], we will reconsider that approach in the context of the methodology introduced in this thesis and propose some possible variations of the diverse methodology steps. The objective is to investigate which of these steps can affect retrieval effectiveness and if possible variations can provide an improvement. The methodology is therefore used to support both the design and the evaluation of the methodology application by unveiling possible point of failures among the methodology step implementations.

The scenario where the methodology application will be evaluated considers a user who submits a query, obtains a ranked list of results, and provides judgments on the documents corresponding to the obtained results. Those documents are then used to model possible term relationships. The top ranked documents are not necessarily good sources for terms and relationship. Better results could be obtained by providing the user documents according to specific criteria, e.g. diversifying them. This issue will be discussed in Section 4.1.1.

Given a set of documents judged as relevant, the descriptors adopted to model the term relationship dimension are terms appearing in the feedback documents. But not necessarily all the terms are good descriptors of the user intents. For this reason, in Section 4.1.2 some term selection strategies will be investigated as a part of the evidence collection step.

Gathered terms will be then adopted to obtain a model of the dimension on the basis of weights obtained from local co-occurrence of these terms in the feedback set. The adopted approach shares intuitions underlying HAL spaces [Lund and Burgess, 1996] and LLSI [Hull, 1994] to obtain a correlation matrix. The description of the dimension modeling step is reported in Section 4.1.3. Finally, the specific representation adopted to obtain document vectors will be discussed in Section 4.1.4.

4.1.1 Feedback Source for Terms

As discussed in Section 3.4.1 the selection of the source is constituted by two sub-steps. The first is the definition of the specific hypothesis on the factors that can explain the user perception of relevance and that will be subject of the investigation. For the particular implementation considered in this section, the selected class of factors are possible relationships among terms in documents explicitly judged as relevant after a first stage of search. Relationships are modeled in terms of correlation on the basis of local co-occurrence data.

The second sub-step consists in the specific source for terms selected. A straightforward approach is to consider the top ranked documents obtained by the first search. In this case we can assume the user provides explicit judgments on the top n retrieved. However, not necessarily the top retrieved are good sources for expansion terms, relationship among terms, or both. It is often the case that top ranked documents are similar to each other. Varying the feedback set could allows us to gain some insights on the most effective criteria for the considered methodology application. For this reason, in the experiments we will investigate the effect of diverse criteria for feedback set selection. Moreover, another possible variable that could affect the effectiveness of the methodology application are specific properties of feedback documents when considered in isolation, and not as part of a feedback set. Possible criteria for se-

lecting document sets and the individual documents for feedback will be described in Section 5.2.3.1.

4.1.2 Term Selection

The descriptors adopted when considering this particular application of the methodology are terms. Given a set of feedback documents $\mathcal{R}_F$, a possible solution is to adopt only terms appearing in the user provided description and exploit the evidence extracted from the feedback set to capture possible relationships among them. But usually textual queries are short, as in the test collections adopted in this thesis. Queries can benefit from expansion based on other terms occurring in the document of the feedback set [Harman, 1992]. A possible approach is to consider all the terms in the feedback documents as good terms, but this could add too much noise. Therefore a term selection strategy, as instance of the descriptor selection problem, should be adopted. Previous works suggest that supervised strategies can be effective to support term selection. For instance, in [Cao et al., 2008] a supervised approach based on multiple term features was proposed that assume independence among terms; in [Cartright et al., 2009] dependence was explicitly takes into consideration to weight expansion terms. Part of the features adopted by these strategies requires corpus-wide statistics. The extraction of these features at query time can be quite slow when very large document collections, e.g. ClueWeb09, are adopted. A possible approach is to use a sample of the collection as done in [Cartright et al., 2009]. In this thesis we decided to focus on unsupervised approaches that do not require collection-wide statistics for feature extraction or feature normalization — see [Wong et al., 2008] for possible measures that exploit collection-wide normalization.

Given all the terms e's appearing the the feedback documents, they are ranked (in decreasing) order according to score assigned by one of the following functions:

- rTF·IDF$_e$, that is the product of the total frequency rTF of e in the feedback set $\mathcal{R}_F$, i.e. rTF$_e = \sum_{d \in \mathcal{R}_F} tf(e,d)$, and the IDF [Spark Jones, 1972] of the term e, i.e. $idf(e)$, defined as

  $$idf(e) = \log \frac{N - n_e + 0.5}{n_e + 0.5}$$

  where N is the total number of document in the collection, n_e is the number of documents in the collection where the term e appears. This measure was that actually adopted in [Melucci, 2008] and therefore it allows us to perform a comparison with the original method.

- LCA [Xu and Croft, 2000], originally applied to PRF its basic rationale is to expand the query with terms that tend to co-occur with query terms in the

feedback set. Expansion terms e's are ranked according to

$$f(e,q) = \prod_{t_i \in q} (\delta + co\_degree(e,t_i))^{idf(t_i)}, \tag{4.1}$$

where δ is a parameter adopted to avoid that the weight of the expansion term becomes zero when one of the query terms w_i contributes with a zero value and

$$co\_degree(e,t_i) = \log_{10}(co(e,t_i)+1)\ idf(e)/log_{10}(|\mathcal{R}_F|), \tag{4.2}$$

where $co(e,t_i) = \sum_{d \in \mathcal{R}_F} tf(e,d)\ tf(t_i,d)$ and $idf(e) = \min(1.0, \log_{10}(N/n_e)/5.0)$, where N is the number of documents in the collection, n_e is the number of document where e occurs and $|\mathcal{R}_F|$ is the number of feedback documents considered; in this methodology application it is the number of relevant documents. LCA can be applied both using passages and whole documents. The default value $\delta = 0.1$ was adopted in the experiments. Since results reported in [Xu and Croft, 2000] no difference was observed in terms of effectiveness, in this thesis we will use whole document, even if it could be more computational expensive. The experiments also showed that LCA was less effective for expansion terms when considering explicit feedback that using expansion based on the frequency of terms in the feedback set. Terms weights were based on the rank in the list provided by LCA score, i.e. $w_e = (1.0 - 0.9i)/k$, where k is the number of expansion terms. In this thesis we will adopt LCA in order to investigate if it will provide better terms for dimension modeling that rTF·IDF$_e$, since it implicitly takes into account co-occurrence with query terms.

In [Wong et al., 2008] when using the Rocchio formula the authors showed that better results can be obtained because of the initial query bias. Even if our feedback strategy is different, and therefore the result not necessarily generalizes, in this thesis terms in the original query will be maintained in the modified query since they constitute the descriptors explicitly chosen by the user and therefore that can provide us useful information on its interests.

The result of the evidence collection step is a set of terms $\mathcal{T}$ that constitutes the expanded query.

4.1.3 Modeling Term Relationship in Feedback Documents

The implementation of the dimension modeling step for term relationship consists in obtaining a vector subspace representation of local co-occurrence of terms selected from the feedback documents. Relationship among terms are considered symmetric, namely the relationship between term t_i and t_j is the same when considering terms t_j and t_i. Here the expression "local co-occurrence" indicates the co-occurrence of terms within

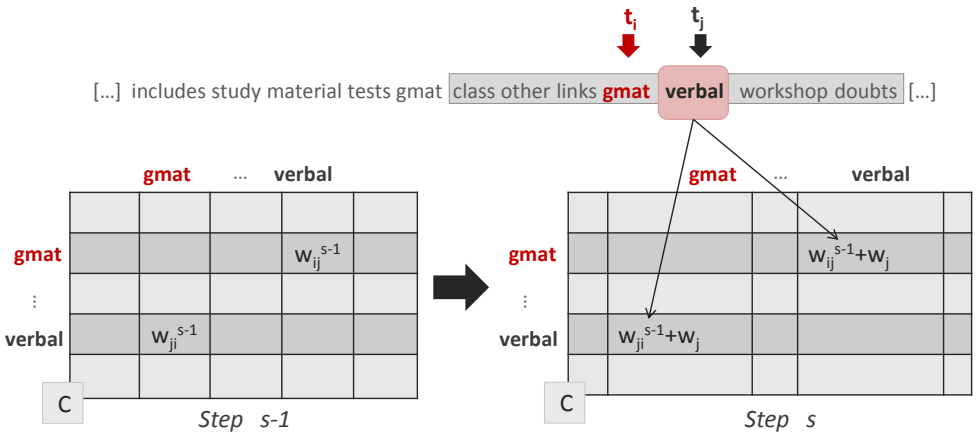


Figure 4.1: Matrix preparation for modeling term relationship.

windows of texts as described in the following. Two sub-steps can be distinguished when modeling the term relationship dimension, specifically the preparation of local co-occurrence data in a matrix and the computation of a vector space basis from the obtained matrix. More specifically:

- **Matrix Preparation**. Let $\mathcal{T}$ be the set of terms selected from the source and let $C \in \mathbb{R}^{|\mathcal{T}| \times |\mathcal{T}|}$ be a matrix whose elements are initially set to zero, namely $c_{ij} = 0$ for $1 \leq i, j \leq |\mathcal{T}|$. For each term $t_i \in \mathcal{T}$ a window of text centered around each occurrence of t_i is considered; if a term $t_j \neq t_i \in \mathcal{T}$ appears in the window of text, statistical information about t_j, e.g. its total frequency in the collection, or a weight w_j derived from such information, e.g. the TF·IDF, is added both to c_{ij} and c_{ji}. A pictorial description is reported in Figure 4.1 when a window of text of size 7 is considered. The text window is centered on the word t_i ="gmat". Since the term t_j ="verbal" belongs to the expanded query, the weight of "verbal" is added both to the entry c_{ij} and the entry c_{ji} of the matrix C; these entries refer to the correlation between t_i and t_j. The value w_{ij}^{s-1} and w_{ji}^{s-1} refer respectively to the weight in c_{ij} and c_{ji} at the step $s-1$.

- **Matrix Decomposition and Basis Vectors Selection.** A possible solution to obtain a vector space basis from the matrix C is to apply SVD. Once SVD has been applied, the matrix is decomposed as $C = U\Sigma V^T$, where $\Sigma \in \mathbb{R}^{n \times n}$ and $U, V \in \mathbb{R}^{|\mathcal{T}| \times n}$, with $U = V$ since C is symmetric; the columns of U constitute an orthonormal vector space basis. A subset of the basis vectors is adopted to model the dimension. Therefore, if $U = [\mathbf{b}_1, \ldots, \mathbf{b}_{\mathcal{T}}]$ and a subset $\{\mathbf{b}_r, \ldots, \mathbf{b}_{r+s}\}$ is selected, the subspace $L(\mathcal{R}_F) = span(\{\mathbf{b}_r, \ldots, \mathbf{b}_{r+s}\})$ is adopted as model of the dimension.

A criterion to select the eigenvectors is required. The approach adopted in this work is to consider the first s eigenvectors extracted by SVD starting from the eigenvector with the highest eigenvalue. The reason for this choice relies on the fact that the first s eigenvectors obtained by SVD provides the best $rank(s)$ approximation of the matrix C in terms of Frobenius norm; the resulting matrix $C' = \sum_{i=1}^{s} \lambda_i \mathbf{b}_i \mathbf{b}_i^T$ can be interpreted as an approximation where some noisy data have been filtered out. Components corresponding to lower eigenvalues can be removed without losing much information — see [Meyer, 2000], Chapter 5, page 418 for the interpretation of SVD as a Fourier expansion and the application to remove noisy data. In order to investigate the effect of the matrix approximation, we will consider different values of s, specifically satisfying the following constraints:

- if λ_s is the eigenvalue associated to the sth eigenvector, the maximum value of s is that $\frac{\sum_{i=1}^{s} \lambda_i}{\sum_{j=1}^{|\mathcal{T}|} \lambda_j} \leq 0.90$, i.e. the retained eigenvectors are those that explained no more than the 90% of the variance;

- $s < |\mathcal{T}|$, that is the maximum number of eigenvectors considered is less that all the possible eigenvectors.

The assumption underlying those constraints is that eigenvectors corresponding to the lowest eigenvalues are not useful to explain the feedback data, but only noise that needs to be filtered out.

The ranking function adopted in this thesis, i.e. Equation 3.8, is based on the projector onto the subspace spanned by the basis vectors selected to model the dimension. Instead of using directly the eigenvectors obtained by the decomposition technique, a possible approach is to build the projector as a weighted sum of the projectors corresponding to the individual vectors. For instance, if $\{\mathbf{b}_1, \ldots, \mathbf{b}_s\}$ have been selected, then the projector adopted could be $P_\omega = \omega_1 \mathbf{b}_1 \mathbf{b}_1^T + \cdots + \omega_s \mathbf{b}_s \mathbf{b}_s^T$, where the ω_i's are the weights assigned to the different projectors onto the one-dimension subspaces spanned by $\mathbf{b}_i$'s. In the experiments reported in this thesis we will adopt as weight for an eigenvector $\mathbf{b}_i$ the normalized value of the eigenvalue corresponding to that eigenvector, i.e. $\lambda_i / \sum_{j=1}^{|\mathcal{T}|} \lambda_j$. Therefore, the eigenvalue determines the relative contribution of the eigenvector (actually of the associated projector) in P_ω.

4.1.4 Document Modeling

The document modeling steps consists in obtaining a vector representation for each document, on the basis of the descriptors of the adopted source. We will investigate the following representations:

- binary-centered, where each document d is represented by a vector $\mathbf{d} \in \mathbb{R}^{|\mathcal{T}|}$ defined as $\mathbf{d} = \mathbf{d}' - \overline{\mathbf{d}}$ where the ith element of $\mathbf{d}'$ is defined as follows:

$$d'_i = \begin{cases} 1 & \text{if } t_i \text{ occurs in } d \\ 0 & \text{otherwise} \end{cases}$$

  and $\overline{\mathbf{d}} = \frac{1}{|\mathcal{T}|} \sum_{i=1}^{|\mathcal{T}|} d'_i$.

- TF·IDF, where each document d is represented by a vector $\mathbf{d} \in \mathbb{R}^{|\mathcal{T}|}$ of TF·IDF weights.

- BM25 [Robertson and Zaragoza, 2009], where a document d is represented by a vector $\mathbf{d} \in \mathbb{R}^{|\mathcal{T}|}$ of BM25 weights. A brief description is reported in the following. Let us denote with $\mathcal{T}_d$ the set of terms in a document d. The weight w_i assigned to the term $t_i \in \mathcal{T}_d$ is

$$w_i = \frac{tf'(t_i, d)}{k_1 + tf'(t_i, d)} \; idf(t_i)$$

  where k_1 is a parameter. The quantity $tf'(t_i, d)$ is defined as $tf'(t_i, d) = tf(t_i, d)/B$, where $tf(t_i, d)$ is the term frequency of t_i, and

$$B = (1 - b) + b\frac{dl}{avdl}$$

  where $dl = \sum_{t_i \in \mathcal{T}_d} tf(t_i, d)$ is the document length, and $avdl$ is the average document length in the collection.

- satTF, where a document d is represented by a vector $\mathbf{d} \in \mathbb{R}^{|\mathcal{T}|}$ of the term frequency component $\frac{tf'(t_i,d)}{k_1+tf'(t_i,d)}$ of the BM25 weight.

The first representation is that proposed in [Melucci, 2008]. The basic rationale is to capture the correlation among terms that occur in the document; only information about the presence of terms is adopted. The other three representations do not take into account correlation, but explicitly include other information, e.g. the discriminative power of the term in the document and in the collection or frequency saturation and normalization. Different document representations are considered in order to investigate if the correlation information is sufficient to obtain effective re-ranking, or exploit other information, e.g. normalized term frequency, will result in a more effective dimension-based re-ranking.

4.2 Methodology for User Behavior Dimension

The past decade has witnessed an increasing interest of the IR community in user behavior as a potential source for feedback. This interest is mainly due to the possibility of characterizing this source by inexpensive features in terms of user effort. For instance, when user behavior is described in terms of post-search interaction features, they can be gathered without the direct involvement of the user, e.g. by monitoring his behavior when interacting with a result list or with the corresponding individual documents.

The methodology application discussed in this section aims at exploiting the contribution of the relationship among post-search interaction features as source for feedback, specifically to support user-behavior based document re-ranking. The gap between what users perceive as relevant for achieving their information goal and what IR systems predict to be relevant suggests investigate approaches which are personalized for each user [Teevan et al., 2010]. Since the features are observed during user interaction and are data about user behavior, it is natural to investigate them in the context of the personalization of information access. When speaking about personalization, it is customary to consider the evidence from the user who is interacting with the system. However, personal evidence is often unavailable, insufficient or unnecessary. Therefore, a broader definition of personalization can be useful or necessary. Such a definition includes the situation in which the evidence is gathered from the group whose users search for information useful for meeting similar tasks or requests, thus making group-based evidence available for personalization purpose. As discussed in Section 3.4.4, considering group-data could allow us to investigate a document representation based on multiple observations. Indeed, group data for a document consists in multiple entries, where each entry corresponds to an observation of the behavior of a distinct user when interacting with that document. Even if based on a larger number of observations, the representation could not be effective to support the individual. Section 4.2.1 will specifically focus on this issue and discuss the impact of the selection of the source for post-search interaction features in the proposed methodology; additional remarks on the features, specifically on the collection and the selection of them are reported in Section 4.2.2. Finally, Section 4.2.3 and Section 4.2.4 respectively describe how to obtain a vector subspace representation for the source, namely a dimension, and for the document using post-search interaction features. In the remainder of this work we will refer to post-search interaction features simply as *interaction features*.

4.2.1 Feedback Source for Interaction Features

The need of personalizing search results suggests investigating individual user behavior as a source for interaction features. One problem is that these data can be unavailable. Let us come back to the considered scenario in order to point out some issues about the adoption of personal data. In this scenario the query formulation is considered as the first interaction of the user with the IR system when searching for information on the benefits of can recycling, e.g. in order to write a report for a university class. In the methodology proposed in Section 3.4 the documents to re-rank are represented in terms of the features gathered from the specific source considered, in this case interaction features. Because of the need to develop approaches personalized for each user, possible solutions are: (i) using interaction feature values observed from the same user when searching for the same query in the past, basically re-finding; (ii) using interaction feature values observed from the same user when searching for interrelated queries, e.g. those formulated when trying to accomplish the same task. In the event of the first search to accomplish a task or, in general, to satisfy an information need, if no prior information is available none of the above approaches can be adopted: features gathered from the first visited documents are the only additional information the system can actually use to support prediction. Therefore, in the considered scenario no other documents can be represented than those visited by the user.

A possible solution is to consider a broader interpretation of personalization which exploits interaction feature values distilled from a group of users interrelated to the user the system is supporting. Users can be interrelated because working on the same task, searching for the same query, or sharing the same interests. Therefore, individual users and user groups become possible sources from which interaction feature values can be distilled. In the former case, the features are those gathered during the user post-search interaction activity, e.g. when interacting with the results or the landing documents. In the latter case, some features can be distilled from the behavior of the group, e.g. the average dwell time spent on a page, while others can be group specific. The user/group behavior can be interpreted as a property of a relationship that involves document and user/group when a specific information need is considered. The specific unit considered determines the *granularity* at which the features can be distilled, i.e. individual user or group granularity.

In the adopted methodology, both interaction features for the dimension and the document representation are required. Since there are two the possible sources from which interaction feature values can be distilled, i.e. user behavior or group behavior, this leads to four possible combinations which are reported in Table 4.1.

The meaning of the labeling scheme reported in the third column of Table 4.1 is the following: the first letter denotes the source for interaction features adopted

Source for dimension modeling	Source for document modeling	Label
Personal behavior	Personal behavior	`P/P`
Personal behavior	Group behavior	`P/G`
Group behavior	Personal behavior	`G/P`
Group behavior	Group behavior	`G/G`

Table 4.1: Possible sources of features to model the user behavior dimension and to represent documents.

to model the dimension, while the second letter denotes the source for interaction features adopted to represent documents. The `P/P` combination refers to the case where the interaction feature values gathered from the individual user behavior — i.e. its *personal behavior* — when searching for a specific query are adopted both for modeling the dimension and for representing documents. The `P/G` combination refers to the case where personal behavior is adopted for modeling the dimension, while the data gathered when observing the behavior from a group of users searching for the considered query are adopted for document representation. The remaining two combinations, namely `G/P` and `G/G`, have analogous meaning.

The adoption of the diverse source combinations implies diverse assumptions on the availability of interaction feature values. As suggested by the above remarks not all the combinations are always applicable. The `P/P` combination is based on the assumption that documents to re-rank have been already visited: observations concerning personal behavior on the documents to re-rank should be available. In the event of the recycling scenario, where the first interaction of the user with the system and the documents in considered, these data is not available. As discussed at the beginning of this section, personal data for document representation can be available in the event of re-finding where the user has already searched for the query and interacted with the results, or when the user is searching for interrelated queries, e.g. those formulated when trying to accomplish the same task. An alternative solution is to use only a subset of the visited documents for dimension modeling, and re-rank all the visited documents to investigate if they provide better support for query expansion. This strategy is investigated for group-based dimension models in Section 5.1.2.3 but can be adopted to the personal behavior case as in [Melucci and White, 2007b]. A criterion for the selection of the subset of observations used for dimension modeling is required, e.g. considering the first visited documents.

The `P/G` combination "decreases" the granularity in the sense that document representation is no more based on personal behavior but on group behavior. Here, we are implicitly assuming that other users have searched using the same query or performed

the same task and visited the documents to be re-ranked. Task information can be elicited with adequate interface support, e.g. as proposed in [Dragunov et al., 2005].

When considering the `G/G` combination and a single group, the "lowest" level of granularity is actually adopted: in this case group distilled feature values are adopted for both dimension and document representation. Even if not explicitly investigated in this thesis, hierarchies of groups can be considered, thus allowing for more than two granularities.

The reason for investigating the `G/P` combination is that it will provide insights into the effectiveness of the dimension representation based on personal data, specifically when comparing `P/P` and `G/P` in order to investigate if group data is a more robust source of user behavior because it is less affected by the individual variations in the style of interaction.

A final remark concerns the two `P/G` and `G/P` combinations, namely those involving two diverse sources for the two representations. The assumption underlying the adoption of these two combinations is that, even though dimension and document representation are obtained from diverse sources, i.e. the individual and a group not including the individual, the contribution of the combination is still useful to support prediction since the individual and the group are interrelated.

4.2.2 Interaction Features Collection and Selection

Selection of the Features

As mentioned in Section 3.4.1 the *source selection* step should not be only interpreted as the selection of the source from which we can distill features, but it implies a specific hypothesis on the type of information that can be extracted from these features. A first basic hypothesis is that the features we are gathering can provide us useful information on the way the user perceives a document with regard to his current information need. The main issue is how to interpret the feature values, especially when interaction features are adopted. In accordance with what was proposed in [Melucci and White, 2007b], in this work it is hypothesized that useful information for supporting prediction can be extracted by exploiting the relationships among the interaction features or, more in general, the relationship among the features that can describe user behavior.

As in the recycling scenario, let us consider a user examining the results obtained after a first interaction with the IR system. Let us suppose that display-time thresholds are used as implicit indicators of relevance. When considered in isolation and with regard to the individual, past work [White and Kelly, 2006] showed that it is difficult

to consider display-time threshold as an absolute indicator of relevance; one of the reasons is the variation in the style of interaction among individuals. Moreover, the amount of display-time is affected not only by the user perception of relevance of a document, but it can be affected also by other features, e.g. by the document length: the meaning of a long time spent on a short document can provide different information than a long time spent on a long document; or a short time on a document that is bookmarked, saved or printed by the user differs from the same amount of time spent on a document on which no retention actions are performed. The amount of scrolling is another example: if the amount of scrolling is interpreted as an indicator of relevance and it is described in terms of number of actions performed by the mouse scroll or page up/down actions, a large amount of scrolling on a long page can be different from the same amount on a short page. Moreover, the style of interaction is not unique: each user has their own style and for this reason a personalized approach seems to be suitable [Melucci and White, 2007a, Melucci and White, 2007b]. The main idea of these works is that the useful information for supporting feedback is in the relationship among features, not only in the values when considered in isolation.

The above remarks provide a motivation for including diverse features; even if some of them are not strictly "behavioral features", e.g. number of query terms in the title or document length, they can contribute to explaining the specific observed behavior.

Personal and Group Feature Values

The adoption of diverse sources for features, namely individuals and group of users, requires some remarks on the way the values are obtained for the feature at the different granularities. When the features are considered with regard to the individual, the feature value for a document is that observed from the user in question when examining the document returned by the system in response to query. When features are considered at group granularity, a possible choice is to compute the feature value as the average value computed over all the users constituting the group. With regard to the scenario considered in this thesis, when computing the average value, the feature values for the user the IR system is supporting cannot always be included in the computation. Indeed, the feature values are not available at personal level for the unseen documents. For this reason in the experiments reported in Chapter 5 feature values of the user supported by the IR system will not be included, thus allowing us to investigate if features distilled from the group not including the user under consideration can substitute feature values for the documents unseen by the user.

Collecting and Retaining Interaction Features

One of the objectives of the methodology proposed in this work is to support the design and the development of an IR system able to exploit diverse sources modeled in a uniform way to support feedback. As a consequence some remarks are required on the approach adopted to collect and handle personal interaction features. Indeed, the main advantage of using interaction feature is that they are cost-less in terms of user effort, namely gathering feature values does not require direct involvement of the user. For instance, they can be obtained by a monitoring tool installed in the client operative system, e.g. as done in [Kelly, 2004], by implementing the IR application with functionalities to capture those features as done in this work — see Section 5.2.3.2, or through an extension of the browser in the event of a web search engine or in general an IR system accessible as a web application. Instances of the latter type of monitoring tool are the Lemur Query Log Toolbar and i-TEL-u. Lemur Query Log Toolbar[1] was developed for the Lemur Query Log Project[2] to support a study to gather the query logs and create a database of web search activities to be provided to the information retrieval research community. i-TEL-u [Agosti et al., 2010] is a query suggestion tool implemented as a browser extension that aims at supporting access to The European Library (TEL)[3]. This tool monitors queries submitted to the TEL portal through the extension and possible explicit judgments provided by the user. These data are stored locally and the user can set the extension for a manual or automatic upload. The gathered logs can be adopted as single source for query suggestion or in combination with other sources, i.e. exploiting term relationship taken from a ontology built on top of Wikipedia[4] and using frequency and co-occurrence data taken from most popular queries of two commercial search engines.

Despite the advantage in terms of user effort, one issue to address is how to manage personal usage data. Indeed, users can be reluctant to provide an external IR system, e.g. a web search engine, with his usage data, even if they are aware of the possible benefits in terms of personalization and support provided by the system. The adoption of source combinations, besides implying diverse assumptions on the availability of the features at the diverse granularities, also affects the way feature values need to be managed to perform prediction. Let us consider, for instance, the P/P case. If the search functionalities are provided by a local application or in an enterprise network, the user could be less reluctant to allow personal data to be retained to help the system predict relevance. If search functionalities are made available through an external provider, the user might have some privacy concerns. Client-side re-ranking could be

[1] http://www.lemurproject.org/querylogtoolbar/
[2] http://lemurstudy.cs.umass.edu/
[3] http://www.theeuropeanlibrary.org
[4] http://www.wikipedia.org/

a solution, but clearly implies an additional computational effort for the system, a requirement that could be critical in the case, for instance, of mobile devices.

The `P/G` combination has the advantage that it has no need to retain features at personal granularity: feature values can be adopted by the system to model the dimension, perform re-ranking, and then used to update the value of the group granularity features used to represent documents to be re-ranked. The user can be less reluctant to use this combination when the search functionalities are made available by an external system since it maintains only feature values at group granularity; however this approach also requires a certain level of trust in the IR system, that is that the system actually performs feature values aggregation. Similarly, re-ranking based on the `G/G` combination can be performed on the system side and the trust in the IR system is required if the aggregation is performed on the system side.

Some remarks on this issue are discussed in [Di Buccio and Melucci, 2009a]. For a critical review on personalized search system and their capability in protecting searchers privacy preservation the reader can refer to [Shen et al., 2007].

In the experiments reported in this thesis, the test collection adopted was obtained through a user study. The tool adopted is a Web application that collects feature values partly on the client side and partly on the server side. The need for a client-side tool depends on the kind of features exploited to model user behavior. For instance, if retention features are considered, a client-side tool is required since printing, saving, or bookmarking can be performed using, for instance, the browser instead of the web application. Without a client-side tool retention feature values will be lost.

Recycling Scenario

Let us denote with "user9" the user searching for information on benefits of can recycling. Let us suppose that the search functionalities are accessible via a web application equipped with a monitoring tool. After the submission of the query "recycling cans and why?" user9 obtains a list of results and he starts examining them. Possible entries in the log produced by the monitoring tool for the first four visited documents are the following:

```
user9 WTX087-B40-98 12000 5 0 1338 1
user9 WTX008-B38-175 38000 15 8 1821 1
user9 WTX091-B19-334 15000 38 0 1024 1
user9 WTX046-B37-24 11000 10 0 21400 2
```

where the first column reports the identifier of the user, the second column the identifier of the document, and the remaining columns values of the monitored features, i.e.

- the time (in milliseconds) user9 spent during the first visit of the document whose identifier is in the second column, e.g. 12000 ms;
- the number of scrolling down actions performed by mouse scroll or page down keystrokes, e.g. 5;
- the number of scrolling up actions performed by mouse scroll or page up keystrokes, e.g. 0;
- the length of the document measured in number of tokens, e.g. 1338;
- the number of query keywords in the document title, e.g. 1.

Since our objective is to obtain a vector subspace representation for user behavior, observed data can be prepared in a document-by-feature matrix as the matrix $\mathbf{F}_{\mathtt{P}} \in \mathbb{R}^{4\times 5}$ reported in the following:

$$\mathbf{F}_{\mathtt{P}} = \begin{bmatrix} 12000 & 5 & 0 & 1338 & 1 \\ 38000 & 15 & 8 & 1821 & 1 \\ 15000 & 38 & 0 & 1024 & 1 \\ 11000 & 10 & 0 & 21400 & 2 \end{bmatrix} \tag{4.3}$$

More in general, the feature values observed for n documents can be prepared in a matrix $\mathbf{F}_{\mathtt{P}} \in \mathbb{R}^{n\times k}$ where k is the number of feature selected to describe user behavior.

The issue that motivates the investigation of different combinations of source for interaction features is that personal data could be unavailable or the variation in the personal style of interaction can affect the effectiveness of the modeled dimension. For this reason, a possible solution is to consider feature values at group granularity. In this thesis the average feature value in the group is adopted as value at group granularity for the feature; in particular, the user under consideration, in this case user9, is not included in the average feature value computation. For instance, if the log file where the feature values are stored contains the following entries for the same query:

```
user12 WTX087-B40-98 19000 49 0
user12 WTX008-B38-175 71000 27 0
user12 WTX091-B19-334 13000 61 0
user12 WTX046-B37-24 39000 84 13
 ...
user3 WTX087-B40-98 226000 65 44
user3 WTX046-B37-24 539000 477 33
user3 WTX008-B38-175 275000 43 26
user3 WTX091-B19-334 16000 32 0
 ...
```

the value of the display-time at group granularity is computed as the average display-time observed for user3 and user12, e.g. 122500 milliseconds for document `WTX087-B40-98`. Using this approach the feature matrix in the event of group data is:

$$\mathbf{F}_{\mathtt{G}} = \begin{bmatrix} 122500 & 57 & 22 & 1338 & 1 \\ 173000 & 35 & 13 & 1821 & 1 \\ 14500 & 45.6 & 0 & 1024 & 1 \\ 289000 & 280.5 & 23 & 21400 & 2 \end{bmatrix} \tag{4.4}$$

Clearly document-specific features, being equal for all the users, are not affected by the granularity.

Although the rows of the matrix $\mathbf{F}_{\mathtt{P}}$ (as well as the rows of the matrix $\mathbf{F}_{\mathtt{G}}$) span a subspace, this straightforward vector subspace representation can be noisy and actually does not provide evident information on the relationships between features. For this reason, each of these matrix representations needs to be mapped in another one that clearly shows useful behavioral patterns in the observed data. The specific mapping adopted in the experiments is discussed in the next section.

4.2.3 Modeling User Interaction Behavior

As mentioned in Section 3.4.3 the mapping needs to be a matrix transformation technique which extracts information about our dimension from the collected data. For instance, if our hypothesis is that a dimension of the user information need can be represented by the correlation among the post-search interaction features, a technique like PCA [Pearson, 1901] can be adopted. This is actually the approach proposed in [Melucci and White, 2007b] and adopted in this thesis. There are different motivations for this choice. A first reason is that the considered features are measured in different units and have different variability: PCA standardizes the data so that features have zero mean and unit covariance. The second reason is that PCA allows diverse patterns to be extracted from the data, in particular, the principal components are mutually orthogonal axes along which the observed data cluster together: a subset of these patterns may be useful for representing the user behavior dimension and approximate the observed feature matrix.

In particular the procedure adopted is:

1. among the features considered and used to compute matrix $\mathbf{F}_{\mathtt{X}}$, where `X` could be either `P` or `G`, select only those for which the standard deviation computed on the sample in the column of $\mathbf{F}_{\mathtt{X}}$ is not null — if the standard deviation is null the feature value is the same for all the entries, thus not providing us any information on the variability of the feature;

2. compute the correlation matrix $C_{\mathtt{X}} \in \mathbb{R}^{k \times k}$ for the new matrix obtained at the first step;

3. apply SVD to the correlation matrix, thus obtaining $C_{\mathtt{X}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, where $\mathbf{\Sigma} \in \mathbb{R}^{n \times n}$ and $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{k \times n}$, with $\mathbf{U} = \mathbf{V}$ since $C_{\mathtt{X}}$ is symmetric; the columns of $\mathbf{U}$ constitute an orthonormal vector space basis.

With regard to the methodology, the columns of the matrix $\mathbf{U}$ obtained by applying PCA to the matrix $\mathbf{F}_{\mathtt{X}}$ constitute a set of possible user behavior factors that corresponds to behavioral patterns. A subset of the eigenvectors among the columns of $\mathbf{U}$ are adopted as factors to model the user behavior dimension. In the experiments reported in this work the eigenvectors associated with non-null eigenvalues are tested and the eigenvector which maximizes the retrieval effectiveness is manually chosen from among all the eigenvectors provided by PCA. The basic rationale is to investigate if at least one of the possible patterns provide an effective model for the user behavior dimension. In this sense, the modeling procedure is more exploratory and cannot be applied automatically. The automatic selection of the eigenvector is a challenging problem and will be matter of future investigation.

For instance, in the event of the recycling scenario possible dimensions for modeling personal user behavior are:

$$\mathbf{b}_{1_{\mathtt{P}}} = \begin{bmatrix} -0.49 \\ -0.18 \\ -0.45 \\ 0.51 \\ 0.52 \end{bmatrix} \quad \mathbf{b}_{2_{\mathtt{P}}} = \begin{bmatrix} 0.44 \\ -0.53 \\ 0.53 \\ 0.36 \\ 0.34 \end{bmatrix} \quad \mathbf{b}_{3_{\mathtt{P}}} = \begin{bmatrix} -0.24 \\ -0.83 \\ -0.18 \\ -0.33 \\ -0.34 \end{bmatrix} \tag{4.5}$$

The meaning of the first factor $\mathbf{b}_{1_{\mathtt{P}}}$ is that interaction features tend to cluster together and document specific features are not positively correlated with them. So in this case a great amount of scrolling actions or a long time spent on a page cannot be explained by the fact the the document is long. As well as for the personal case, a possible group dimension can also be computed using the same procedure. In particular, patterns associated with non-null eigenvalues are:

$$\mathbf{b}_{1_{\mathtt{G}}} = \begin{bmatrix} -0.45 \\ -0.47 \\ -0.36 \\ -0.47 \\ -0.47 \end{bmatrix} \quad \mathbf{b}_{2_{\mathtt{G}}} = \begin{bmatrix} 0.31 \\ -0.29 \\ 0.80 \\ -0.29 \\ -0.31 \end{bmatrix} \quad \mathbf{b}_{3_{\mathtt{G}}} = \begin{bmatrix} 0.82 \\ -0.31 \\ -0.47 \\ -0.02 \\ -0.10 \end{bmatrix} \tag{4.6}$$

For instance, pattern $\mathbf{b}_{3_{\mathtt{G}}}$ indicates that the time the users in the group spent on average on the document is negatively correlated with the length of the document. In

$\mathbf{b}_{2_{\texttt{G}}}$ display-time and scrolling up actions are correlated with each other but negatively correlated with the document length.

An alternative approach for dimension modeling is to apply PCA directly to data gathered from the diverse users instead of extract behavioral patterns from the average data. The underlying idea is to investigate if considering average feature values part of the possible useful variability in the data is removed. For this reason, besides the four combinations discussed in Section 4.2.1 an additional combination is considered labeled as `Gd/G`. In this combination, as for the other `-/G` cases, the evidence adopted to represent the documents with regard to a query is obtained by computing the average feature values over all the users other than the user in question who assessed that topic. The `Gd` label indicates that the model of the dimension is obtained by applying PCA to a document-by-feature matrix where the documents of the diverse users were considered as distinct evidence. For instance, if the system was supporting `user9`, the feature matrix adopted as evidence was $\mathbf{F} \in \mathbb{R}^{(n \cdot 2) \times k}$, where k is the number of features, n is the number of documents visited, and 2 is the number of users other than `user9` who searched for the same query q, namely `user3` and `user12`. In that case:

$$\mathbf{F}_{\texttt{Gd}} = \begin{bmatrix} 19000 & 49 & 0 & 1338 & 1 \\ 71000 & 27 & 0 & 1821 & 1 \\ 13000 & 61 & 0 & 1024 & 1 \\ 39000 & 84 & 13 & 21400 & 2 \\ 226000 & 65 & 44 & 1338 & 1 \\ 539000 & 477 & 33 & 21400 & 2 \\ 275000 & 43 & 26 & 1821 & 1 \\ 16000 & 32 & 0 & 1024 & 1 \end{bmatrix} \Rightarrow C_{\texttt{Gd}} = \begin{bmatrix} 1.00 & 0.83 & 0.78 & 0.47 & 0.46 \\ 0.83 & 1.00 & 0.46 & 0.71 & 0.72 \\ 0.78 & 0.46 & 1.00 & 0.30 & 0.30 \\ 0.47 & 0.71 & 0.30 & 1.00 & 1.00 \\ 0.46 & 0.72 & 0.30 & 1.00 & 1.00 \end{bmatrix}$$

$$\Rightarrow \mathbf{b}_{1_{\texttt{Gd}}} = \begin{bmatrix} -0.46 \\ -0.49 \\ -0.35 \\ -0.46 \\ -0.46 \end{bmatrix} \quad \mathbf{b}_{2_{\texttt{Gd}}} = \begin{bmatrix} 0.45 \\ 0.01 \\ 0.61 \\ -0.46 \\ -0.46 \end{bmatrix} \quad \mathbf{b}_{3_{\texttt{Gd}}} = \begin{bmatrix} 0.31 \\ 0.63 \\ -0.62 \\ -0.26 \\ -0.25 \end{bmatrix} \quad \mathbf{b}_{4_{\texttt{Gd}}} = \begin{bmatrix} -0.70 \\ 0.60 \\ 0.35 \\ -0.17 \\ -0.04 \end{bmatrix}$$

4.2.4 Document Modeling

The methodology proposed in this thesis requires both a representation for the information need, namely the dimension, and a representation for the document in terms of the features selected to characterize the source. Therefore each document is represented as a vector $\mathbf{d} \in \mathbb{R}^k$ where k is the number of features selected in the first step of the dimension modeling procedure; features with null standard deviation are not considered, where the standard deviation for a feature is computed using the values in

the corresponding column of $\mathbf{F}_{\texttt{X}}$. In the event of the `-/P` combinations, each document is represented by a vector whose components are the feature values observed for that user when examining the document with regard to the considered query. For instance, $\mathbf{d}_{\texttt{P}}$ = [12000 5 0 1338 1] for `user9` when examining document `WTX087-B40-98` with regard to the query "recycle cans and why?".

In the event of a `-/G` combination, as with the dimension modeling step, the feature values are computed as the average over all the feature values observed for the users in the group, not including the user in question. For instance, when considering document `WTX087-B40-98` with regard to the query "recycle cans and why?", the display-time was computed as the average between that observed for `user3` and that observed for `user12` when examining document `WTX087-B40-98`, i.e. $\mathbf{d}_{\texttt{G}}$ = [122500 57 22 1338 1].

CHAPTER

FIVE

EXPERIMENTS

Chapter 4 discussed possible applications of the methodology to two specific sources: term relationship in documents judged as relevant and user behavior described in terms of relationship among interaction features. Some issues were discussed when describing these methodology applications. Section 5.1 will frame these issues in a number of research questions that will be the subject of the experiments described in the remainder of the chapter. Section 5.2 will describe the experimental methodology adopted to investigate these questions. Each experiment involve a two stage prediction. The first stage corresponds to the prediction based on the initial query formulation. The second stage is the re-ranking based on a specific dimension. The user behavior dimension will be adopted both for direct re-ranking and to support query expansion. In the latter case the second stage prediction involves not only dimension based re-ranking, but also a subsequent application of a feedback algorithm on the re-ranked documents. Moreover, Section 5.2 describes the test collections adopted and the experimental system developed in this thesis to support the evaluation of the methodology applications.

5.1 Research Questions

5.1.1 Document Re-ranking through Term Relationship Dimension

5.1.1.1 Effect of Term Relationship in Relevant Documents on Re-ranking

The approach proposed in [Melucci, 2008] aimed at modeling term relationship based on local co-occurrence data gathered from the top retrieved documents and exploit the obtained model for re-ranking. That approach can be interpreted as a Pseudo-

Relevance Feedback (PRF) technique. PRF is based on the assumption that the top ranked documents are relevant or that can provide useful evidence to enhance the information need representation of the user. In this case, they are supposed to be good sources for term relationship. When the user explicitly assesses some documents as relevant, a possible question is if relationships modeled from the content of these documents can be effective to support re-ranking. Thus, the research question is:

What is the effect of document re-ranking based on term relationship extracted from document judged as relevant by the user?

5.1.1.2 Effect of Relevant Feedback Sets on Document Re-ranking

Let us consider that the user provides judgments on the set of top n retrieved documents. The most straightforward approach is providing the user with the top n documents obtained by the first stage prediction. But the top ranked results are not necessarily the best source for term relationships. The system could adopt alternative strategies to provide documents to judge and that could be good sources for feedback. Thus, the research question is:

What is the effect of the selection of the source for term relationships on the effectiveness of the considered methodology application?

5.1.1.3 Effect of Document Representation on Document Re-ranking

The document representation proposed in [Melucci, 2008] aims at modeling correlation among terms, where the correlation is based only on the presence/absence of the terms in the document. This representation does not take into account statistics of terms occurrence in the document or in the collection as done by most effective weighting schemes. A possible approach is to not consider correlation information on the document side and exploit a vector representation where statistics on the terms are explicitly taken into consideration. Thus, the research question is:

What is the effect of the document representation on the effectiveness of the considered methodology application?

5.1.1.4 Effect of Term Selection Strategy on Document Re-ranking

The objective of the methodology application described in Section 4.1 is to model relationship among terms. Terms are those used for the information need representation, i.e. those constituting the query. But queries are usually short and the information

need representation can benefit from expansion based on feedback documents. Diverse term selection strategies can be adopted to select a subset of terms appearing in the documents judged as relevant; then these terms, together with those appearing in the original query, are the input for the dimension modeling step. A possible approach is to rely on strategies that extract terms on the basis on their discriminative power in the feedback set and in the collection, e.g. rTF·IDF. But alternative strategies can be adopted, e.g. LCA where terms are scored also taking into account the co-occurrence with query terms in the feedback set. Thus, the research question is:

What is the effect of the term selection strategy on the effectiveness of term relationship-based re-ranking?

5.1.1.5 Effect of Properties of a Single Feedback Document on Re-ranking

The research question reported in Section 5.1.1.2 concerns with the investigation of the criteria to select an effective feedback sets then used as input for modeling the dimension. But the effectiveness of the methodology application to term relationship could be also affected by specific properties of documents when considered as individual sources for feedback, i.e. when considered in isolation and not as part of a feedback set. Thus, the research question is:
What properties make a single relevant document an effective input for term relationship-based re-ranking?

5.1.2 Document Re-ranking through User Behavior Dimension

5.1.2.1 Effect of Feedback Source for Interaction Features on Document Re-ranking

The features distilled from the behavior of individual users are often unavailable, insufficient or unnecessary. The behavior of groups of interrelated users, e.g. those searching for the same topic, can be considered as another possible source for features. Since both a representation for the user behavior and for the documents is required by the adopted methodology and there are two possible feature granularities, four possible combinations `X/Y` can be defined, where `X` denotes the granularity of the user model and `Y` the granularity for document representation — `X` or `Y` is either `P` (personal) or `G` (group). With regard to the possible combinations, the question is if `P/P`, i.e. using only personal data, outperforms those combinations using group data, which can be adopted when personal data is not available for one or both the required representations. Thus, the research question is:

What is the effect of the group data on document re-ranking when modeling user behavior and representing documents instead of personal data?

5.1.2.2 Effect of the Number of Relevant Documents Used for Feedback on Document Re-ranking

Since the users are reluctant to provide relevance assessments, recent research activity has been devoted to designing methods which minimize user effort, for example, by collecting implicit indicators of relevance. In particular, the impact of the number of relevant documents on relevance feedback has been thoroughly investigated for designing feedback algorithms which perform well also when fed with no or little evidence extracted from the content of the documents. A similar issue arises when the evidence is extracted from the behavior of the user examining the documents. Thus, the question is:

What is the effect of the number of relevant documents among those used for user behavior dimension modeling on the effectiveness of document re-ranking?

5.1.2.3 Effect of User Behavior Dimension-based Document Re-ranking on Query Expansion

Pseudo-Relevance Feedback exploits the top-ranked documents to refine the initial information need representation, e.g. extracting terms to expand the query. The underlying assumption is that the top-ranked documents are relevant and their properties can be adopted as evidence for feedback. But this assumption is not always valid. An alternative solution is to adopt the documents visited by the user, even if previous works [Agichtein et al., 2006b, Joachims et al., 2007] showed that his decision can be affected by the trust in the IR system capability. If re-ranking based on the user behavior dimension is able to increase the number of good documents for feedback in the top-ranked, feedback techniques, e.g. query expansion, can benefit from re-ranking. Thus, the question is:

What is the effect of using the documents re-ranked by user behavior as a source for query expansion instead of the retrieved top-ranked?

5.1.2.4 Effect of the Number of Relevant Documents Used for Feedback on Query Expansion

Let us assume that documents have been re-ranked by user behavior dimension. When considering the top re-ranked documents as a source for query expansion, a further

question is if user behavior-based query expansion is less sensitive to the number of relevant documents among the top-ranked than PRF. Thus, the research question is:

What is the effect of the number of relevant documents among the top-ranked on user behavior-based query expansion? Is it less sensitive than PRF?

5.2 Experimental Methodology

5.2.1 Experimental Methodology for Term Relationship Dimension

5.2.1.1 Stage 1: First retrieval run

The weighting scheme adopted for prediction at the first stage is the BM25, particularly exploiting the implementation where the normalization constant $(k_1 + 1)$ is not adopted — a brief description is reported in Section 4.1.4. In the experiments the parameter k_1 is heuristically set to $k_1 = 2$ and the value of b adopted was $b = 0.75$.

A subsequent re-ranking of the top ten documents obtained by BM25 is performed on the basis of the number of query terms in the url of the document. If the same number of query keywords is present in the URL of two documents, they are ranked by BM25 weights; if they both have the same BM25 weight and the same number of keywords in the URL, they are ranked by document identifier.

BM25 combined with URL re-ranking is adopted as baseline for the experiments concerning with term relationship dimension.

5.2.1.2 Stage 2: Re-ranking Exploiting Term Relationship Dimension using Relevant Documents as Source for Feedback

The experimental methodology adopted for investigating the research questions described in Sections 5.1.1.1–5.1.1.4 can be summarized by the following steps performed for each topic:

1. **Source Selection**: Select the set of relevant documents from which the term relationship dimension will be modeled. If no relevant documents are present in the feedback set, results are not re-ranked, that is results obtained from the first stage prediction are returned.

2. **Evidence Collection**: Extract the top k terms among those appearing in the documents with highest weight w_{term} and expand the query, constituted by the topic keywords, with the selected terms. The expanded query has $k + h$ terms

where h is the number of terms in the initial query. w_{term}'s investigated are rTF·IDF and LCA.

3. **Dimension Modeling**:

   a– Computation of the local co-occurrence matrix C by windows of text in the considered relevant document. In particular a window of text of size 7 is centered around each occurrence of a keyword $t_i \in \mathcal{T}$. If a keyword $t_j \in \mathcal{T}$ appears in the window of text centered around t_i, the TF·IDF weight of t_j is added to the elements c_{ij} and c_{ji} of C.

   b– Decomposition of the matrix C by SVD.

   c– Selection of the first s eigenvectors $\{\mathbf{b_1} \dots \mathbf{b}_s\}$; each of selected eigenvector $\mathbf{b}_i$ is multiplied by $\lambda_i / \sum_{j=1}^{|\mathcal{T}|} \lambda_j$, where λ_i is the eigenvalue corresponding to $\mathbf{b}_i$; adoption of the subspace $L(\mathcal{R}_F)$ spanned by those vectors as model of the dimension.

4. **Document Representation**: Represent each document as a vector $\mathbf{y} \in \mathbb{R}^{k+h}$, where y_i is the weight of the term t_i in the considered document.

5. **Prediction**: Re-ranking of the top m results retrieved by the baseline according to the distance between the vector representation of the document and the computed subspace; the ranking function adopted is that described by Eq. 3.8: $m_{\mathcal{R}_F}(\mathbf{y}) = \mathbf{y}^T \cdot \mathrm{P}_{L(R_F)} \cdot \mathbf{y}$, where $\mathbf{y}$ is the document vector and $\mathrm{P}_{L(\mathcal{R}_F)} = \omega_1 \mathbf{b}_1 \mathbf{b}_1^T + \dots + \omega_s \mathbf{b}_s \mathbf{b}_s^T$ the projector onto the subspace $L(\mathcal{R}_F)$, where $w_i = \lambda_i / \sum_{j=1}^{|\mathcal{T}|} \lambda_j$.

As mentioned in the first step, when no documents judged as relevant were among the top five retrieved, the baseline (and stage one) results were returned. The reason for the latter choice is due to the difference between the "subspace of irrelevance" and the subspace spanned by non relevant documents. Indeed, as stated in [Melucci, 2008], if orthogonality is chosen to model mutual exclusion and $L(\mathcal{R}_F)$ denotes the subspace of relevance, $L(\mathcal{R}_F)^\perp$ may denote irrelevance. While the subspace of irrelevance is orthogonal to $L(\mathcal{R}_F)$, $L(\overline{\mathcal{R}}_F)$ is in general oblique — $L(\overline{\mathcal{R}}_F)$ denotes the subspace spanned by non relevant documents. When considering a probabilistic interpretation of the ranking function described by Equation 3.8 [van Rijsbergen, 2004, Melucci, 2008], ranking according to $1 - \Pr[L(\mathcal{R}_F)^\perp | L(\{\mathbf{y}\})]$ is in general different than ranking by $1 - \Pr[L(\overline{\mathcal{R}}_F) | L(\{\mathbf{y}\})]$. If all the feedback documents are judged by the searcher as non relevant, $L(\overline{\mathcal{R}}_F)$ can be computed but not $L(\mathcal{R}_F)^\perp$.

The research question reported in Section 5.1.1.5 is investigated using the above methodology evaluation, where each document set is constituted by a single relevant document. Therefore re-ranking can be performed for each topic. In the event of the

investigation of this research question the term selection strategy adopted scores terms by IDF and only the first eigenvector is selected to model the dimension.

5.2.2 Experimental Methodology for User Behavior Dimension

5.2.2.1 Stage 1: Indri Search Engine

The prediction process at the first stage is performed by the Indri search engine[1]. The retrieval model adopted by Indri [Metzler and Croft, 2004] is based on a combination of the Inference Network Framework [Turtle and Croft, 1991] and the Language Modeling Framework [Ponte and Croft, 1998]. It is able to handle complex queries, namely structured queries, but in this work query likelihood retrieval paradigm from language modeling has been adopted. Each document in the collection is modeled as a collection of samples from a multiple-Bernoulli distribution, a sample for each word in the document. Details are provided in [Metzler et al., 2004]. Indri constitutes the baseline for the user behavior experiments. The reason for this choice is that it was shown to be a reasonable baseline in past evaluation performed on the TREC evaluation campaign [Metzler et al., 2005]. This collection is that used to obtain the user behavior test collection adopted in the experiments for the user behavior dimension. A description of the test collection is reported in Section 5.2.3.2. When indexing and performing retrieval, stopwords are removed and Porter Stemmer is adopted. Retrieval is performed using default parameters, i.e. Dirichlet smoothing and smoothing parameter $\mu = 2500$; this configuration was shown to be effective when experimented in the TREC 2001 test collection without query expansion [Collins-thompson et al., 2005].

5.2.2.2 Stage 2-1: Re-ranking Exploiting User Behavior Dimension

The evaluation methodology adopted to investigate questions 5.1.2.1–5.1.2.2 assumes that a user visited n documents among the ten displayed in the result page returned by Indri in response to a query. Then for each query q and for each user u who searched using that query the following steps are performed:

1. **Source selection**: Selection of the combination of the source for features, that is, either `P/P`, `P/G`, `G/P`, `G/G` or `Gd/G`.

2. **Evidence collection**: Collection of the features from the first $n_B = 3$ visited documents. The collected features are prepared in a $n_B \times k$ matrix where k is the number of features collected from the n_B visited documents. The reason for adopting the top visited documents and not the top ranked documents is to simulate a scenario similar to that introduced in Section 3.2.1, where first

[1] http://www.lemurproject.org/indri/

obtained interaction data is adopted for feedback. In the adopted dataset – see Section 5.2.3.2 – the first visited results in general differ from the top ranked.

3. **Dimension modeling**: Modeling the user behavior and the documents by extracting possible behavioral patterns by applying PCA on the matrix. The result of the application of this technique is an orthonormal basis — one basis vector $\mathbf{b}$ for each pattern. Patterns, namely eigenvectors associated with non-null eigenvalues are tested one at a time.

4. **Document modeling**: Representation of the documents in terms of features gathered from the source selected at step 1. Each document is represented as a vector $\mathbf{y}$ of k features.

5. **Prediction**: Re-ranking of the top $m = 10$ results of the baseline list according to the measure $m_{\mathbf{b}}(\mathbf{y}) = \mathbf{y}^T \cdot \mathbf{P}_{L(\{\mathbf{b}\})} \cdot \mathbf{y}$, where $\mathbf{P}_{L(\{\mathbf{b}\})} = \mathbf{b} \cdot \mathbf{b}^T$ is the projector onto the subspace spanned by $\mathbf{b}$.

As described in Section 4.2 when the *group* is adopted as source for features, namely in the `G/-` or `-/G` combinations, the value $f^G_{i,u',d,q}$ of a feature i for a specific user-query-document triple (u', q, d) is computed as

$$f^G_{i,u',d,q} = \frac{1}{|G| - 1} \sum_{u \in G \text{ and } u \neq u'} f^I_{i,u,d,q}$$

where G denotes the group constituted by all the users which visited the document d with regard to the query q and $f^I_{i,u,d,q}$ the feature value observed for a specific individual u with regard to (d, q).

With regard to the number of feedback documents, $n_B = 3$ is selected because we are interested in investigating the adoption a small number of visited results — when $n_B = 2$ for a large part of the combinations, the number of possible behavioral patterns extracted by PCA was one and this seems not to provide an effective model for document re-ranking.

5.2.2.3 Stage 2-2: Query Expansion Based on Top Documents Re-ranked by User Behavior Dimension

Besides the impact on document re-ranking, the effectiveness of user behavior to support query expansion is investigated. It is supposed that a first stage prediction has been performed based on the baseline described in Section 5.2.2.1. For each query q the following steps are performed:

1-PRF. Consider the top $n_F = 5$ documents retrieved by the baseline.

1-IRF. Perform step 1–5 described in Section 5.2.2.2. Dimension modeling is based on the top $n_B = 3$ documents; the combination of sources of features adopted is `G/G`, that is the tests are performed in a non personalized scenario. Interaction features of all the users who searched using query q are adopted for dimension and document modeling for user behavior-based re-ranking. The dimension is automatically obtained using the first eigenvector among those extracted. Consider the top $n_F = 5$ documents re-ranked by user behavior dimension.

2. Re-ranking of the top $m = 50$ documents returned by the baseline by using the Indri Pseudo-Relevance Feedback algorithm with $k = 10$ expansion terms. When considering 1-IRF as the first step the strategy is not actually PRF since we are using the top re-ranked by user behavior dimension. The Indri Pseudo-Relevance Feedback mechanism[2] is an adaptation of that introduced in [Lavrenko and Croft, 2001].

The relevance judgments adopted to measure the effectiveness of PRF and IRF are those provided by the TREC assessors. The underlying idea is that the TREC assessor is considered as a new user, not among those in the group, who will be supported using group evidence. In other words this experiment aims at investigating if the pattern extracted by PCA could be useful for non-personalized re-ranking.

A remark should be made on the Indri Pseudo-Relevance Feedback mechanism. This technique, even if extracts terms from the feedback documents then adopted to expand the query, computes also weights (actually probabilities) for the extracted terms adopted both for term selection and for weighting selected terms in the expanded query. Therefore, the query modification involves both expansion and term weighting.

5.2.3 Test Collections and Measures

The research questions reported in Section 5.1 are experimentally investigated using the test collection based evaluation [Sanderson, 2010]. In the test collection based evaluation of IR systems, the three classic components constituting a test collection are:

- a collection of documents, or corpus; each document is uniquely identified by a doc_id;
- a set of topics, each of them uniquely identified by a $topic\_id$;
- a set of relevance judgments, often referred to as *qrels*, which consists of a list of (doc_id, $topic\_id$, rel), where rel is the relevance judgment expressed by the

[2] See http://ciir.cs.umass.edu/~metzler/indriretmodel.html

assessor on the document identified by *doc_id* with regard to the topic identified by *topic_id*.

The feedback strategies based on the methodology applications described in Chapter 4 involve two consecutive stages. The prediction in the first stage does not exploit any evidence other than the initial textual query formulated by the user. The classic components in a test collection allows first stage prediction to be performed. Indeed, the title of the topics in the test collection is adopted to simulate queries submitted by the user; available relevance judgments can be then adopted to measure the effectiveness of the first stage prediction. The adopted measures are discussed in Section 5.2.3.3.

Differently, the second stage requires additional information, specifically the evidence to refine the initial information need representation, then adopted to perform feedback. In the event of the term-relationship dimension the methodology is applied to a scenario where explicit judgments provided by the user are available, e.g. in an Explicit RF scenario. In this work two different test collections based on the same corpus are adopted. The first test collection considers the scenario where a user submits a query, obtains a list of results and provides judgments on the top five documents returned. The judged documents are then adopted as source for feedback for a second stage. The second test collection considers a different scenario: the user submits a query, obtains a list of results and indicates a single relevant document. Only this document can be adopted to perform feedback at a subsequent stage. The two test collections are those adopted respectively in the TREC 2009 and TREC 2010 RF Track; those test collections are briefly described in Section 5.2.3.1.

Unlike the former source of evidence, the evaluation of the methodology implementation for the source user behavior requires information not available in TREC and other standard test collections made available to the IR community. In particular the information required is the evidence gathered by monitoring the behavior of the user when interacting with the first visited results. For this reason, a user study was designed and carried out with the specific aim of gathering post-search navigation features to describe the user behavior and use these features as evidence to model the user behavior dimension to support IRF; the test collection used in the TREC 2001 Web Track was adopted thus allowing us to extend the information already available with interaction data. The user study and the obtained dataset are described in Section 5.2.3.2.

5.2.3.1 Test Collections for Term Relationship Dimension

The research questions concerning the implementation of the methodology for the term relationship dimension are addressed by adopting the two test collections used in the

TREC 2009 and the TREC 2010 RF Track. Those collections are briefly described in the remainder of this section, together with the objective of the RF track in the two considered TREC editions. Before the description of the test collections, the ClueWeb09 dataset will be described since it constitutes the corpus adopted in the RF track both in 2009 and 2010.

ClueWeb09 Dataset. The ClueWeb09 dataset is a corpus of one billion pages collected by the Language Technologies Institute at Carnegie Mellon in January and February 2009. The corpus is constituted by web pages in ten different languages. The entire collection is split into twenty-four segments, ten of which constitute the subset of English documents in the corpus. Each segment is constituted by a number of gzipped files, each of them storing web pages in Web ARChive (WARC) format. This dataset was initially adopted as corpus for a test collection in TREC 2009, specifically in the Web, the Entity, the Million Query, and the RF track. The actual corpus adopted in the diverse tracks was a subset of the entire corpus, specifically the ten segments which constitute the English portion of the ClueWeb09 dataset; this subset is also known as *TREC 2009 Category A* (in the remainder of this thesis referred to as *Category A*). Category A consists of 503,903,810 pages (2.08 TB compressed, 13.4 TB uncompressed). Another subset of the corpus adopted in TREC 2009 was the first of the ten English segments, named *TREC 2009 Category B* (in the remainder of this thesis referred to as *Category B*). This subset, which is actually the corpus adopted in this thesis, is constituted by 50,220,423 pages (246.9 GB compressed, 1.53 TB uncompressed)[3]. Category B includes a complete snapshot of Wikipedia.

TREC 2009 RF Track Test Collection. The objective of the RF in TREC 2009 was to evaluate the capability of the systems to retrieve good documents to be judged. Each participant was asked to provide one or two small sets of documents to be judged, specifically five documents per set. Each set of documents was then judged by TREC assessors. This first stage was named *Phase-1*. The evidence gathered from Phase-1 was then adopted as input for a second stage, named *Phase-2*. Then a set of Phase-1 runs was assigned to each participant — the information provided for each Phase-1 run was the identifier of the five documents and the relevance judgments of the assessor on them. The participants were then asked to use the assigned Phase-1 runs as input for their RF algorithms. The test collection resulting from the participation in the TREC

[3] Further information on the ClueWeb09 dataset is available at http://boston.lti.cs.cmu.edu/Data/clueweb09/

2009 RF track is constituted by the following components[4]:

- a corpus of documents, namely Category A or Category B;
- a set of fifty topics;
- thirty Phase-1 runs; each run is constituted by a set of five documents and relevance assessments on those documents for the fifty topics;
- relevance judgments for the Phase-2 runs on Category A and Category B documents with regard to the fifty topics.

In the following a brief description is reported of the Phase-1 runs assigned to the Information Management System (IMS) research group during the participation to the TREC 2009 RF track, and that will be used in the experiments in this thesis:

UPD.1 - [Di Buccio and Melucci, 2009b]. The top ten documents were retrieved by BM25 and then re-ranked according to the number of query terms in the URL of the pages. The top five documents were returned as results — see Section 5.2.1.1.

ilps.2 - [Meij et al., 2009]. The query was transformed in a full dependency query model using MRF [Metzler and Croft, 2005] and the top ranked document retrieved by this model were adopted as input to generate Relevance Model (RM)s [Lavrenko and Croft, 2001]. The top 50 terms with highest probability were adopted to expand the query and retrieve the phase 1 set.

PRIS.1 - [Li et al., 2009a]. K-means clustering was applied and the five documents in the center of the clusters were returned as phase 1 results. No information is provided by the author on the specific parameters adopted.

UMas.1 - [Cartright et al., 2009]. Two runs were performed: the first using Query Likelihood (QL) unigram model [Ponte and Croft, 1998] and the second using MRF (weights: 0.80 for term, 0.015 for ordered components and 0.05 for unordered component). Then

1. The best five documents from MRF were chosen when they satisfied both the following criteria: (i) the document does not appear in the QL run or (ii) the document was ranked higher by QL than MRF.
2. If less than five documents were provided at step one, then any document provided by MRF that satisfied the following criteria was selected: (i) the document was ranked worse than rank five, and (ii) both QL and MRF assigned it to the same rank.

[4]TREC 2009 RF track data is available at http://trec.nist.gov/data/relevance.feedback09.html

3. If less than five documents were provided at step two, then any document provided by MRF that satisfied the following criteria was selected: (i) the document was ranked worse than rank five, and (ii) was not already in the phase 1 list.

The baseline was MRF and PRF performed by RM.

QUT.1 - [Li et al., 2009b]. Feedback documents were obtained using query expansion and term weighting based on an ontology encoded from a library catalog system, specifically the Library of Congress Subject Heading (LCSH). Each catalog entry is characterized by a set of subjects (e.g. "Consumption (Economics)– Germany (East)"). The first step consists in the identification of a positive subject set $\mathcal{S}^+$, i.e. subjects related to the TREC topic. If a subject shares terms with the TREC topic, then it is considered positive with a certain degree *pos* that is obtained by a combination of different measures, e.g. specificity of the subject determined by hierarchical relations *is-a* and *part-of* in the ontology, or the belief, a score inversely proportional to the position of the subject in the catalog entry subject list and the frequency of the subject in the entry. The preliminary set identified according to these measures is then expanded including other subjects related to those in $\mathcal{S}^+$ because of the hierarchical relations determined by the ontology, but not containing topic terms. A *pos* score is then assigned also to these subjects. A set of documents is associated to each subject, specifically those document where the subject appears in the subject list. Those documents are adopted as source for expansion terms. The weight w_t of a term t is computed as a weighting sum of *support* scores (sum of the *pos* obtained for the document subjects) for all the documents where the term t appears, where the *support* score of a document was weighted by the normalized frequency of the term in the document. The top weighted 150 terms were selected to expand the query and the documents were ranked according to the sum of w_t's appearing in the document titles. The top five documents constitute the feedback set.

CMU.1 - Document selection was performed exploiting a fuzzy clustering algorithm in order to diversify documents to be judged. The distance between two documents was computed on the basis of a vector representation of document pairs. Each document pair was represented as a vector of features: specifically document, URL, Webgraph and query derived features were adopted. The distance was obtained as output of a logistic regression classifier trained on a set of known relevant documents. Once obtained the clusters, documents in each cluster were ranked by Indri and

the top document per cluster was included as part of the Phase 1 feedback set. Indri was adopted as baseline, using a full dependence query model [Metzler and Croft, 2005] obtained from query text.

Details on the selection of the fifty topics and the procedure to perform relevance judgments are reported in [Carterette et al., 2009].

TREC 2010 RF Track Test Collection. The objective of the RF Track in TREC 2010 was to focus on a single relevant document and understand which properties make a document good or bad for feedback. Given a query, namely the topic title, and a single document judged as relevant with regard to the query, the participants were asked to perform feedback on the basis of this evidence. More specifically ten documents known to be relevant and with different properties were assigned to the participants; they were asked to use only one document at a time to perform feedback. The documents were selected with diverse criteria, more specifically:

1. a randomly chosen document from among the topic's known relevant documents;
2. the most commonly returned relevant document in TREC 2009;
3. the least commonly returned relevant document in TREC 2009;
4. the longest relevant document;
5. the shortest relevant document;
6. another random relevant document;
7. the most spammy relevant document, determined by the approach proposed in [Cormack et al., 2010];
8. the least spammy relevant document;
9. a random highly relevant document;
10. the most commonly returned non-relevant document.

The test collection resulting from the participation in the TREC 2010 RF track is constituted by the following components:

- a corpus of documents, namely Category A or Category B;
- a set of one hundred topics selected from those run in the TREC 2009 Million Query Track topics; this set of topics includes the fifty topics adopted in TREC 2009 RF Track;
- ten documents with different properties for performing explicit relevance feedback, knowing that those documents are relevant;

- relevance judgments on the documents with regard to the one hundred topics.

The IMS Research Group participated in the Relevance Feedback Track both in TREC 2009 [Di Buccio et al., 2009b] and TREC 2010; the submission was performed using Category B. In [Lin et al., 2009] the author showed that the quality of the pages and the level of spam in Category A and Category B is different: Category B contains less spam and higher quality pages than Category A. In [Clarke et al., 2009] the authors suggested that one of the factors that cause this difference between the two corpora can be the crawl order. Since the adoption of Category A would require focusing on other issues outside the scope of this work, e.g. addressing spam issues, in the experiments reported in the remainder of this thesis Category B is adopted.

5.2.3.2 Test Collection for User-behavior Dimension

Standard test collections, e.g. those adopted in TREC, are not sufficient for evaluating the methodology implementation for the user behavior dimension. For this reason a user study was carried out to gather interaction data by monitoring the behavior of the users when examining the top ten retrieved results in response to assigned topics and assessing their relevance with a four-graded scale. This section describes the corpus used to perform this study, the procedure adopted to gather interaction data, and the resulting dataset.

TREC 2001 Web Track Test Collection. The test collection adopted in the user study was the TREC 2001 Web Track Test Collection. The corpus in this test collection is the WT10g web corpus, which is constituted by 1,692,096 documents (2.7 GB compressed, 11 GB uncompressed). The test collection includes fifty Ad-hoc topics and 145 Homepage-Finding topics together with the corresponding relevance judgments[5]. A subset of the Ad-hoc topics was adopted in the user study described in the remainder of this section.

The WT10g corpus was indexed by the Indri Search Engine component of the Lemur Toolkit[6]; english stop-words were removed and the Porter stemmer was adopted. The documents of the WT10g test collection were ranked by query likelihood as described in Section 5.2.2.1 and the top ten documents were considered for each topic.

In order to address the research questions reported in Section 5.1.2 information is required on the behavior of diverse users when assessing the same topic. However, fifty topics were too many to be judged for each user. Hence, only a subset of the

[5] TREC 2001 Web Track data is available at http://trec.nist.gov/data/t10.web.html

[6] http://www.lemurproject.org/lemur/ — The version adopted was Lemur 4.9

Difficulty	Number of relevant documents	Topics
High	1/2	506 - 517 - 518 - 543 - 546
Medium	3/4/5	501 - 502 - 504 - 536 - 550
Low	6/7/8/9/10	509 - 510 - 511 - 544 - 549

Table 5.1: TREC 2001 Ad-Hoc web track topics divided according to the number of relevant documents in the top ten retrieved.

	Difficulty		
Set	High (1-2)	Medium (3-5)	Low (6-10)
A	506 - 517 - 518	501 - 502 - 504	509 - 510 - 511
B	517 - 518 - 543	502 - 504 - 536	510 - 511 - 544
C	518 - 543 - 546	504 - 536 - 550	511 - 544 - 549

Table 5.2: Topic sets, each of them constituted by three topics for each set in Table 5.1.

ad-hoc topics were considered. The number of documents judged as relevant among the top ten retrieved was considered as an indicator of topic difficulty — at this stage a document was considered relevant if it was assessed as relevant by the TREC assessors, i.e. according to the judgment reported in the qrels. The topics with no relevant documents, namely 534, 542, 513, 516, and 531, were removed. The remaining topics were divided in three groups. The first group was constituted by those topics with one or two relevant documents among the top ten retrieved (highly difficult topics). The second group was constituted by the topics with a number of relevant documents ranging from three to five. The third group included the topics with six to ten relevant documents among the top ten retrieved — there was actually one topic with ten relevant documents (topic 544). Then we randomly selected five topics from each group, thus finally obtaining the fifteen topics reported in Table 5.1.

Then three distinct groups of nine queries were built, each group thus being composed by three topics for each of the three groups — see Table 5.2. The topics were distributed so that at least one topic from each group (518, 504, and 511) would be assessed by all the users involved in the study. The topics were also assigned so that the average topic difficulty per user was uniform.

In order to collect the information about user interaction behavior, we developed a web application. The first web page presented allowed the user to login into the system: the authentication required previously assigned user-name and password. After authentication, the displayed web page provided the list of topics available, specifically their TREC identifiers. Once the user selected one of the assigned topics, a new web page divided in three frames was presented — see Figure 5.1. On the upper right

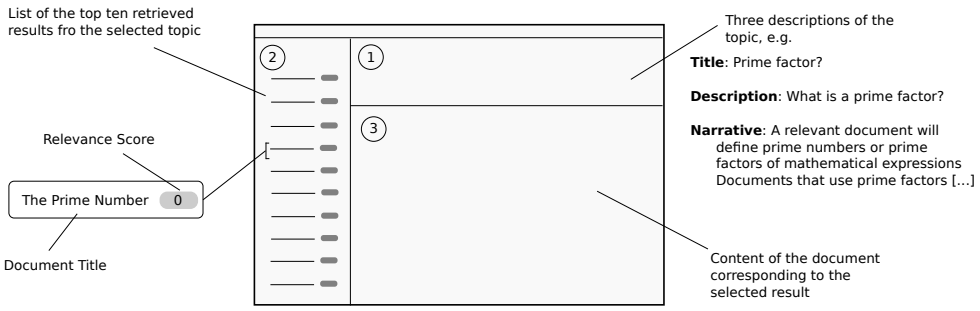


Figure 5.1: Structure of the web application used by the users to examine and assess the results.

Option	Relevance grade	Description
no flag	*not relevant*	the document is not about the subject of the request
flag+1	*marginally relevant*	the topic of the request is mentioned, but only in passing
flag+2	*fairly relevant*	the topic of the request is discussed briefly
flag+3	*highly relevant*	the topic is the main theme of the article

Table 5.3: Four-graded relevance scale description and corresponding option on the drop-down menu to specify it in the web application.

frame, namely frame 1, the diverse topic descriptions were reported, i.e. title, narrative, and description; the left frame, namely frame 2, reported the title of the top ten retrieved documents, ranked by the baseline score. When a user clicked on one of the titles, the content of the corresponding web page was displayed on the bottom right frame, namely frame 3. Beside each title in frame 2 a flag was available to assess if the corresponding document was perceived as relevant or not with regard to the considered topic. Moreover a drop down menu was available to select the relevance degree of the document corresponding to that title. In particular, the four graded relevance scale proposed in [Järvelin and Kekäläinen, 2002] was adopted. The possible choices in the drop down menu and the corresponding meaning in terms of the considered relevance scale are reported in Table 5.3.

Each action concerning the selection of the topic, the selection of the results, and the relevance assessments was stored in the server where the web application was running; both information about the type of action and when the action was performed was collected. The user-name provided to each participant was adopted as user identifier to distinguish the diverse users and their entries. For each user the software application generated three files in the server: `userX1.txt`, `userX2.txt` and `userX3.txt`, where `userX` is the identifier, namely the user-name, assigned to the user in the user study. In the file `userX1.txt` a set of entries was stored, each entry referring to the time the

user started the evaluation of a query. An example is the following:

```
user3 518 OG 11:34:32 29\7\2008
```

where

- `user3` is the identifier of the user;
- `518` is the identifier of the query;
- `OG` is the type of query (not actually used);
- `11:34:32` is the time the user started the evaluation for the considered query;
- `29\7\2008` is the date the user did the evaluation.

The file `userX2.txt` contains information about the time and the date when the user `userX` accessed the documents. An instance of entry is:

```
user3 518 ..\docs\WTX012-B04-132.html 11:35:28 29\7\2008
```

The meaning of this entry is that `user3`, during the evaluation of topic `518`, accessed document `WTX012-B04-132` at `11:35:28` of the `29\7\2008`.

Lastly, in file `userX3.txt` information was stored about relevance judgments of the user for a specified query. An instance of the entries in the file is the following:

```
user3 518 OG ..\docs\WTX012-B04-132.html##2
    ..\docs\WTX088-B43-238.html##1 12:5:13 29\7\2008
```

The meaning of this entry is that the user `user3` submitted their relevance judgments at `12:05:13` of the `29\7\2008`. As mentioned above, the information about the type of query (`OG`) is not used. In particular, the documents assessed as relevant in the example are `WTX012-B04-132` with relevance degree 2 and `WTX088-B43-238` with relevance degree 1.

Other interaction features were stored locally in the browser cookies. In particular the following features were stored for each visited document: the number of scrolling up and scrolling down actions performed by the mouse scroll wheel or by page up/down keys, the depth and the width of the window as displayed, and maximum depth and width achieved when examining the page, e.g. by scrolling.

Fifteen volunteers were recruited: three undergraduate students and twelve PhD students or postdoctoral researchers. One of the three groups of topics, namely $\mathcal{A}$, $\mathcal{B}$, or $\mathcal{C}$, was assigned to each user. The users were instructed on how to use the application by being providing a document containing a brief description of the application and a description of the activity they had to perform — the document reported the description of the different degrees of relevance, as presented in [Järvelin and Kekäläinen, 2002]

User Study Instructions: Summary

1. configuration of the browser
2. the interface for the evaluation is available at the url `http://kmi-web09`
3. login with the assigned user-name and password
4. click on Task Evaluation link
5. the user will select the query number in the main page, then a new page will appear. On the left there are the titles of the documents the user will assess — 10 documents per query
6. when the user clicks on a title, the content of the corresponding page will appear on the right
7. the user can assess if the document is relevant or not by using the flag above the title in the left frame, and it can specify the degree of relevance by a four-graded relevance scale.
8. the assessment will be submitted when the user clicks on the "submit" button. After the submission the user will be returned back to the main page and will select the next query to be evaluated — see point 4.
9. copy the file that stores the cookies to save user behavior activities

Figure 5.2: Summary instructions provided to the participants of the user study.

and reported above. The results could be visited in an arbitrary order, not necessarily according to the ranked list order. Users were asked to provide explicit relevance judgments for each of the visited results. At the end of the evaluation session, the file with the cookies stored by the browser where the interaction data were stored was returned by each participant. The steps each user was asked to perform during the user study, as reported in the documents provided to the users, are summarized in Figure 5.2. The first step, namely the configuration of the browser, was required to preserve the information stored in the cookies.

Some users did not assess all the documents in the result list for some topics: only thirteen of the fifteen users assessed all the documents. The interaction data gathered from those users resulted in a total of 79 (user, topic) pairs and 790 entries where each entry refers to the visit of a specific user to a particular document with regard to a topic. Table 5.4 reports the resulting (user, topic) pairs after the removal of the pairs where part of the top ten documents were not assessed. The bold topic identifiers are those for which the user did not visit the results according to the order they were presented. Topic 549 is not reported since only evidence about `user3` was available after the removal of the entries.

User	Topics													
	Low Difficulty				Medium Difficulty					High Difficulty				
1	509	510	-	-	501	502	504	-	-	506	517	518	-	-
2	-	510	**511**	**544**	-	-	-	-	-	-	**517**	518	543	-
3	-	-	-	544	-	-	-	**536**	-	-	-	**518**	-	**546**
5	-	510	511	-	-	502	504	536	-	-	-	518	-	-
7	-	-	-	-	501	-	504	-	-	-	517	518	-	-
8	-	510	-	-	-	502	504	536	-	-	**517**	518	543	-
9	-	-	511	-	-	-	504	536	550	-	-	518	-	546
10	509	-	-	-	-	502	504	-	-	506	**517**	518	-	-
11	-	510	-	544	-	-	-	-	-	-	-	-	**543**	-
12	-	-	511	544	-	-	504	-	550	-	-	518	-	546
13	509	510	-	-	501	502	-	-	-	**506**	**517**	518	-	-
15	-	-	511	544	-	-	504	536	550	-	-	518	543	-
16	-	510	-	-	501	502	504	-	-	506	517	518	-	-

Table 5.4: Topics assigned for each user involved in the user study after the removal of the pairs without information about implicit features for all the top ten documents.

Dataset of post-search navigation features

The features gathered from the above study are reported in Table 5.5. They can be divided in two groups: features concerning the results or the displayed document, specifically the way in which they were presented, and those concerning user behavior. These features were considered because the hypothesis underlying the methodology implementation described in Section 4.2 is that a factor that explains the user behavior when interacting with the results could be modeled by extracting the relationship between diverse features; none of the features is considered as an individual implicit indicator of relevance. Document length was considered together with the display-time since a greater display-time on a short document can have a different meaning than a display-time on a long document. The dimensions of the browser window were considered together with the scrolling actions because different styles of scrolling interactions observed for diverse users can be also due to the different size of the browser window when visiting the same document with regard to the same query. The adopted matrix transformation technique, i.e. PCA, allows the relationship (e.g. correlation) between the diverse features to be captured.

Some additional remarks are required to clarify the procedure adopted to compute display-time values. The display-time value for a document was computed as the difference between the time when the user accessed that document and the time when the user accessed the next document. For instance, when considering the two following entries in the file `user32.txt`:

Feature	Description
Features observed from document/browser window	
query terms	number of topic terms displayed in the title of the corresponding result
ddepth	depth of the browser window when examining the document
dwidth	width of the browser window when examining the document
doc-length	length of the document (number of terms)
Features observed from the user behavior	
display-time	time the user spent on the page in their first visit
total display-time	time the user spent on the page in their first visit
scroll-down	number of actions to scroll down the document performed both by page-down and mouse scroll
scroll-up	number of actions to scroll up the document performed both by page-up and mouse scroll
sdepth	maximum depth of the page achieved by scrolling down, starting from the ddepth value

Table 5.5: Features adopted to model the user behavior dimension and to represent documents.

```
user3 518 ..\docs\WTX047-B31-168.html 11:41:43 29\7\2008
user3 518 ..\docs\WTX100-B36-4.html 11:42:8 29\7\2008
```

the observed display-time was 25 seconds. In contrast, when the considered document was accessed immediately before the judgment submission, i.e. the last document accessed for a topic, the display-time was computed as the difference between the time the user accessed the document and the time at which the submission of the judgment was performed. This information could be obtained by exploiting the last entry for the topic in file `userX2.txt` and the entry for the topic in file `userX3.txt`, e.g.

```
user3 518 ..\docs\WTX100-B36-4.html 12:4:57 29\7\2008
```

and

```
user3 518 OG ..\docs\WTX012-B04-132.html##2
     ..\docs\WTX088-B43-238.html##1 12:5:13 29\7\2008
```

where the observed display-time is 26 seconds. Since a user could have visited a document multiple times, we considered two distinct features for display-time: the time spent by the user on the document during the first visit when assessing a topic, and the total time spent on that document during the assessment of that topic. The total display-time is not adopted in the experiments.

Summary on the information in the user-behavior test collection

Summarizing the test collection obtained to evaluate the user behavior dimension is constituted by the following components:

- a document corpus, i.e. the WT10g web corpus;
- fourteen distinct topics selected from among the fifty of the TREC 2001 Web Track;
- ten feedback documents per topic;
- explicit relevance judgments in a four-graded relevance scale provided by the users involved in the user study;
- post-search interaction features monitored during the assessment performed by the user on the feedback documents, specifically those reported in Table 5.5.

5.2.3.3 Adopted Retrieval Measures

In this thesis different measures of retrieval effectiveness are adopted for the two considered sources.

Term Relationship Dimension Measure: The measure of retrieval effectiveness adopted for the term relationship dimension is the statMAP. This measure provide an estimation of the MAP when the North-eastern University (NEU) [Aslam et al., 2006, Allan et al., 2009] evaluation method is adopted. This measure is one of those actually adopted in the TREC 2009 RF Track to judge the runs performed on Category B, as those submitted by IMS Research Group of the University of Padua.

User-behavior Dimension Measures: The first research question on the user-behavior dimension concerns with the impact of the source for interaction features on document re-ranking. The effect on retrieval effectiveness is measured with regard to each user individually, thus investigating the effect on result personalization; this is possible since in the user behavior dimension test collection individual gains provided by diverse users are available for the same topic. The measure adopted to investigate this question is the Normalized Discounted Cumulative Gain (NDCG) [Järvelin and Kekäläinen, 2002]. The NDCG was adopted because it can handle usefulness scores ranging in a non binary scale and "systematically combine document rank and degree of relevance" [Järvelin and Kekäläinen, 2002]. Since in the dataset for the user behavior dimension four graded relevance assessments are available, the NDCG allows us

to investigate how well the methodology implementation presents highly relevant documents at high rank positions.

NDCG is defined as the ratio between DCG and Ideal Discounted Cumulative Gain (IDCG), where IDCG is the DCG of the perfect ranking — i.e. the documents are ranked in decreasing order of the gains provided by the assessor. In this thesis, for the user behavior dimension the DCG is computed according to the alternative formulation reported in [Croft et al., 2009], namely

$$DCG = \sum_{i} (2^{r(i)} - 1) / \log(i + 1), \tag{5.1}$$

where $r(i)$ is the relevance of the document at position i. For a specific cut-off n, the sum is computed over the first n ranked documents.

The reason for the choice of this alternative formulation is that when no binary judgments are adopted, it puts a strong emphasis on retrieving highly relevant documents [Croft et al., 2009].

An additional remark concerns the gains adopted to compute the NDCG. When measuring the effectiveness with regard to individual users and a particular topic, the gains adopted were those provided by the users when assessing the documents for the considered topic. When measuring the effectiveness with regard to a group of users, the approach proposed in [Teevan et al., 2010] to compute group gains was adopted. When computing the group gain for a document with regard to a topic, the gain is obtained as the sum of the individual gains provided by the users in the group. As shown in [Teevan et al., 2010] the group gain is suboptimal for the users when considered as individuals, since the ranking needs to satisfy more than one person. An example inspired by [Teevan et al., 2010] but using gains from the collected dataset is reported in Table 5.6.

5.2.4 Experimental System

In order to investigate the research question posed in Section 5.1, an experimental system was developed. The aim of this section is twofold. The first objective is to provide a description of the experimental system, focusing on the implemented modules and the libraries adopted thus making the experiments repeatable. The second objective of this section is to show how the abstraction defined in Section 3.1 can support the design and development of a system able to handle informative resources at diverse resource levels, test collection of different media and described in terms of diverse sources of evidence, and to support different search tasks — e.g. similarity search and cover song identification. In particular, Section 5.2.4.1 will describe the modules developed to provide functionalities to index and retrieve test collections spread over

Best Ranking user1		Best Ranking user13		Best Ranking Group			
Document ID	G	Document ID	G	Document ID	1	13	1+13
WTX004-B45-113	3	WTX040-B06-53	3	WTX076-B42-85	3	2	5
WTX076-B42-85	3	WTX002-B06-114	2	WTX040-B06-53	1	3	4
WTX060-B35-144	2	WTX076-B42-85	2	WTX082-B29-38	2	2	4
WTX082-B29-38	2	WTX082-B29-38	2	WTX092-B14-265	2	2	4
WTX092-B14-265	2	WTX092-B14-265	2	WTX093-B18-263	2	2	4
WTX093-B18-263	2	WTX093-B18-263	2	WTX002-B06-114	1	2	3
WTX002-B06-114	1	WTX004-B45-113	0	WTX004-B45-113	3	0	3
WTX040-B06-53	1	WTX060-B35-144	0	WTX060-B35-144	2	0	2
WTX068-B03-276	0	WTX068-B03-276	0	WTX068-B03-276	0	0	0
WTX080-B01-223	0	WTX080-B01-223	0	WTX080-B01-223	0	0	0
NDCG	1.00	NDCG	1.00	NDCG	0.891	0.894	1.00

Table 5.6: Example of group gain computation for topic 509.

a distributed architecture. Section 5.2.4.2 will describe the modules that are responsible for dealing with test collection of a different medium, specifically music, when considering two search tasks: content-based music similarity search and content-based cover song identification. Finally, Section 5.2.4.3 will focus on the modules developed to implement and evaluate the two methodology applications.

5.2.4.1 SPINA Software Architecture: Exploiting Informative Resources Distributed across a Peer-to-Peer Network

The core of the experimental system adopted in this work was developed within the SAPIR Project[7] and it is constituted by the SPINA (Superimposed Peer Infrastructure for iNformation Access) software architecture [Di Buccio et al., 2008]. This architecture was developed to provide functionalities to index and retrieve documents spread across a distributed architecture. The current implementation of SPINA exploits unstructured, hierarchical and hybrid P2P networks. In this network topology each peer is responsible for a document collection, stored locally and is provided by indexing and retrieval functionalities to support the user when searching his local collection. Those functionalities was developed using the Open Source Library Apache Lucene[8]. A peer has also functionalities to perform a P2P search. Indeed, the collections is distributed across a network of peers, each of which has both client and server functionalities. In the adopted network topology peers are divided in groups, each group led by a particular peer, named super-peer, that is responsible for: (i) propagating P2P search queries to other peers in the group, and (ii) propagating P2P search queries to other

[7] http://www.sapir.eu

[8] The version currently adopted in Lucene 2.4.1. http://lucene.apache.org/java/2_4_1/

groups. Not all the peers in a group and not all the groups are contacted: in order to minimize the communication load the query is forwarded only to the most promising ones. The contacted peers perform a local search and provide the most promising results in their local collection to the super-peer leading their group. The super-peers return the results to the super-peer leading the group of the requesting peer. Also the most promising peers in the group to which the requesting peer belongs, return the results to their super-peer. The final merged list of results is returned to the requesting peer. In this P2P mechanism we can note that: (i) three different prediction process are performed at three diverse informative resource levels – document, peer, and super-peer level – and (ii) the effectiveness of the P2P search depends also on the local predicting capabilities of the peers since the most promising documents per peer constitute the final result list.

With regard to the prediction at the diverse resource levels, each informative resource is described in terms of a set of descriptors, that are actually document descriptors — i.e. both peers and super-peers are characterized by document descriptors. Each descriptor is characterized by a set of features according to the considered abstraction — specifically they are weights, computed according to the weighting framework proposed in [Melucci and Poggiani, 2007], and statistical information on the descriptor in the specific resource level — i.e. the number of document with the descriptor in the peer and the number of peers characterized by that descriptor in the group. These information can be stored in an inverted index by threating informative resources at higher levels – e.g. peers and super-peers – as documents — the posting list associated to a descriptor is a set of (resource, feature) pairs instead of (document, term) pairs; a weight, computed by the adopted weighting scheme, is associated to each pair. While Lucene provides functionalities to directly deal with documents, higher resource levels required the indexing part to be adapted and the weighting scheme to be implemented both for peer and super-peer level. The basic rationale was to map a peer or a super-peer in a Lucene `Document`[9] and each descriptor in a Lucene `Term`. The architecture is depicted in Figure 5.3 where small boxes with thin border in the box `TEXT` are functionalities already provided by Lucene, while boxes in the `HIGHER RESOURCE LEVELS` box are those implemented to perform prediction for informative resources at higher levels. The functionalities to deal with document collections has been enhanced thus dealing with standard test collection — this functionalities has been implemented in the analysis sub-module depicted as a box with thick border in the `TEXT` box. Indeed, SPINA was adopted as experimental system to evaluate a query piggybacking technique to improve super-peer selection [Di Buccio et al., 2009c]; the

[9] "A `Document` represents a collection of fields [...] Each field corresponds to a piece of data that is either queried against or retrieved from the index during search" [Hatcher and Gospodnetic, 2004]

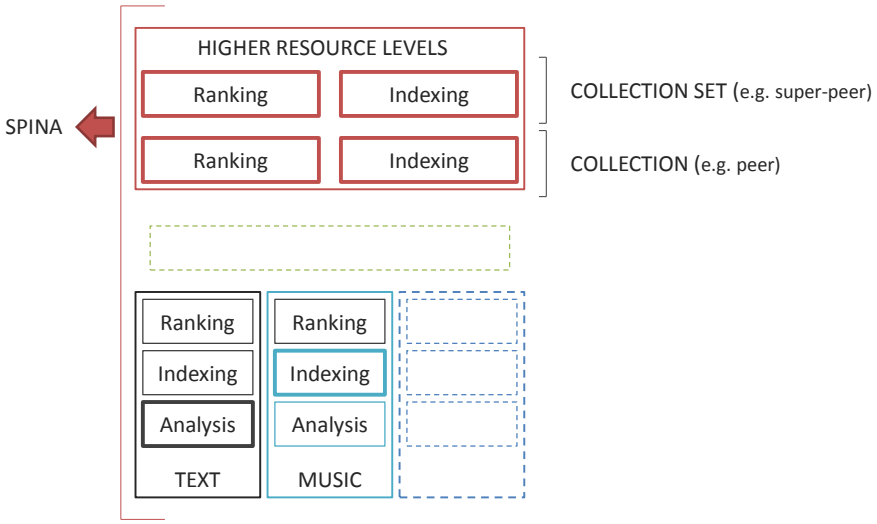


Figure 5.3: Module of the experimental system developed to implement the abstraction for diverse levels of informative resource.

test collection adopted was the DLLC (Digital Libraries Lu and Callan) based on the WT10g web corpus.

With regard to the predicting capabilities of peers when dealing with their local collections, further extensions of the system were focused on two aspects: search functionalities for content-based retrieval of documents of media other than text, and exploiting diverse sources to perform prediction. Section 5.2.4.2 will focus on the former aspect, namely diverse media, while Section 5.2.4.3 will focus on the extension to investigate the two methodology applications considered in this thesis.

5.2.4.2 Abstraction Applications to Different Resource Media and Test Collections

The first extension to the core architecture concerned the functionalities to deal with media diverse than text, specifically music files. Exploiting the segmentation and indexing technique proposed in [Neve and Orio, 2004] a music file can be represented according the abstraction described in Section 3.1. A music file is segmented in a set of rhythmic and melodic patterns, which are adopted as descriptors. Each pattern is in its turn characterized by a set of statistical features, e.g. pattern frequency in the music document or inverse pattern frequency. Prediction is then performed through a weighting scheme, e.g. TF·IDF. When searching for music files, peer and super-peer selection exploit music pattern as descriptors for resources at higher levels [Di Buccio et al., 2009a]. These functionalities has been integrated in SPINA by

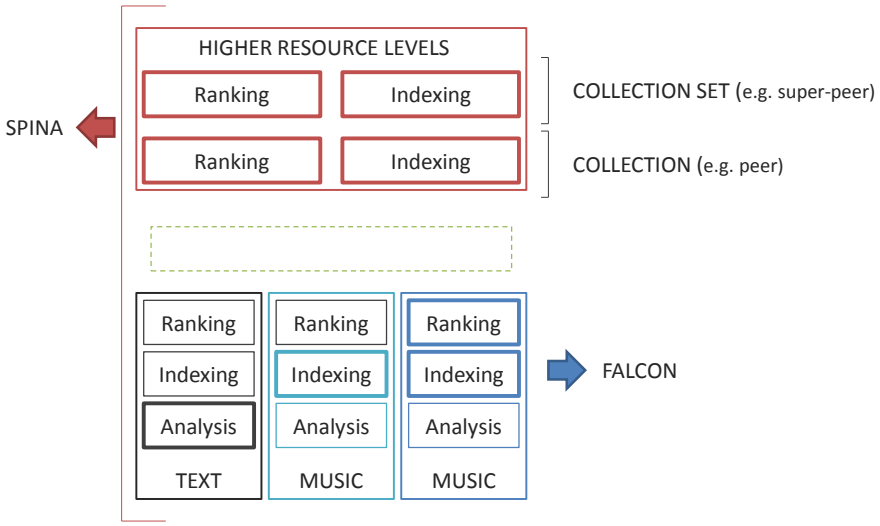


Figure 5.4: Module of the experimental system developed to implement the abstraction for content-based cover song identification.

implementing a new functionalities at the document level, depicted in Figure 5.3 in the box `MUSIC`. The functionalities implemented mainly concerned with indexing.

The type of search supported by this module is content-based similarity search. A further extension based on the same IR abstraction has been introduced in order to support content-based cover song identification. This module is constituted by FALCON [Di Buccio et al., 2010a] – see Figure 5.2.4.2. As described in Section 3.4.2 each song is divided in excerpts, each of them represented by a sequence of chroma vectors. Each sequence is mapped into an hash that is threated as a descriptor. Finally, a song excerpt is interpreted as a passage in a textual document and each hash as an index term in a passage. Each segment is mapped onto a Lucene `Document` and each hash, namely descriptor, is mapped into a Lucene `Term`. Cover identification is performed by identifying the best segment per song and then rank songs by their best segment score. The results reported in [Di Buccio et al., 2010b] showed that this system can be used to retrieve a small number of candidate cover songs from a large collection because of the good trade-off achieved for scalability and effectiveness. Then more sophisticated music alignment techniques can be adopted to refine the list of candidate; these techniques are indeed more effective but hardly scalable for large collections.

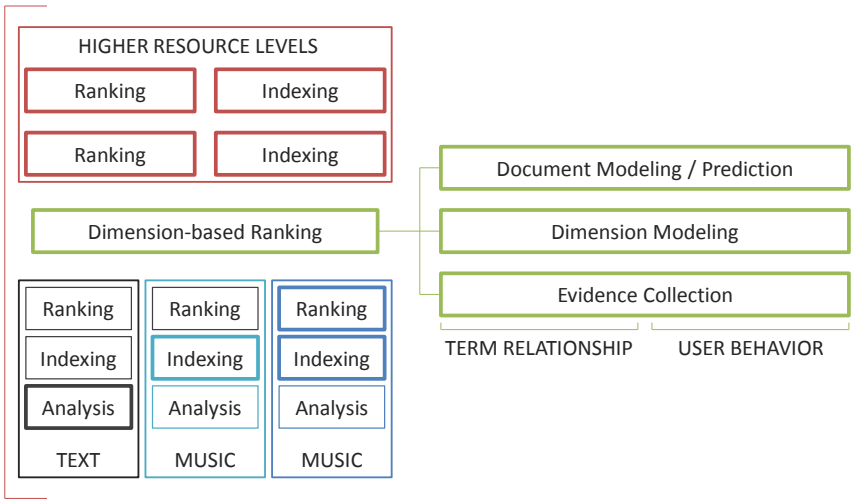


Figure 5.5: Module of the experimental system developed to implement the abstraction for source-based prediction.

5.2.4.3 From Exploiting Diverse Resources to Exploiting Diverse Sources of Evidence

Previous sections aimed at describing how the abstraction introduced in Section 3.1 can be adopted to support the design and the development of an IR system able to handle informative resources at diverse resource levels, of different media and to support different search task.

This section is focused on a further extension of the architecture that provides functionalities to exploit diverse sources to support prediction. The module implements the diverse methodology steps for each of the considered source. The packages of the architecture concerning the methodology application were developed thus being modular, in order to investigate the effect of diverse variants for each methodology step implementation. A specific package was implemented for each source considered. In particular, in the context of this thesis the functionalities developed aimed at supporting the evaluation of the two considered methodology applications — term relationship and user behavior source. All the parameters for indexing, ranking adopted for the experiments can be specified through an eXtensible Markup Language (XML) file.

Term Relationship

Parsing and Indexing. Parsing and indexing functionalities of the software architecture have been extended to handle the ClueWeb09 document corpus. The specific choices adopted for the experiments are described in the following. Each web-

page of the TREC 2009 "Category B" dataset was parsed, particularly the following information was extracted from each record in WARC format: the TREC-ID, the URI and the content. Each of them was stored in a distinct `Field` of a Lucene `Document`. All the content of the document was processed during indexing except for the text contained inside the `<script></script>` and the `<style></style>` tags. Moreover an additional field was stored, which contained the keywords extracted from the URL of the document. In particular during the extraction of the terms from the full content of the documents the presence of each term was checked in the URL; the obtained keywords were then indexed in a separate field, which was used to re-rank the top ten retrieved documents as describe in 5.2.1.1. Stop words were removed during indexing[10]. No stemming was adopted. During indexing not only statistical information about the occurrence of the terms in the documents, namely their frequency, was stored, but also information about the positions where terms occurred and offset information[11]. The information on position of the terms was used to implement the methodology described in Section 4.1.

The wall-clock time to index the 1492 records of the TREC 2009 "Category B" dataset was 45 hours, 46 minutes and 45 seconds, while the CPU time was 38 hours, 3 minutes and 39 seconds (36:29:08 user time and 01:34:30 system time).

Retrieval (Stage 1). The implementation for the Stage 1 prediction described in Section 5.2.1.1 is based on the BM25 implementation for Apache Lucene described in [Pérez-Iglesias et al., 2009][12]. This package was adopted to extend the retrieval functionalities of the `Ranking` block for textual files. Information stored in the URL field is adopted for re-ranking based on the number of query keywords in the URL.

Retrieval (Stage 2): Methodology step implementation. The functionalities currently implemented for the *evidence collection* step consists of the extraction of feature to characterize descriptors in the feedback document, then adopted to support the descriptor selection strategy. The current strategies implemented are based on a single feature, e.g. TF or IDF, or derived feature, e.g. LCA. The *dimension modeling* step has been implemented by exploiting the JAMA Library[13] to handle matrix operations and decompositions. The matrix operation are not directly handled by JAMA,

[10] The stop words list is that available at the url http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words

[11] In Lucene information about the unique terms in a field, their counts, their positions and their offsets can be stored at indexing time and then accessed by using `TermVectors`. The specific `TermVector` option chosen for the Lucene `Field` used for the "content" was `TermVector.WITH_POSITIONS_OFFSETS`. When storing positions, the actual position of terms before stopword removal was considered.

[12] The library is available at the url http://nlp.uned.es/~jperezi/Lucene-BM25/

[13] JAMA : A Java Matrix Package: http://math.nist.gov/javanumerics/jama/

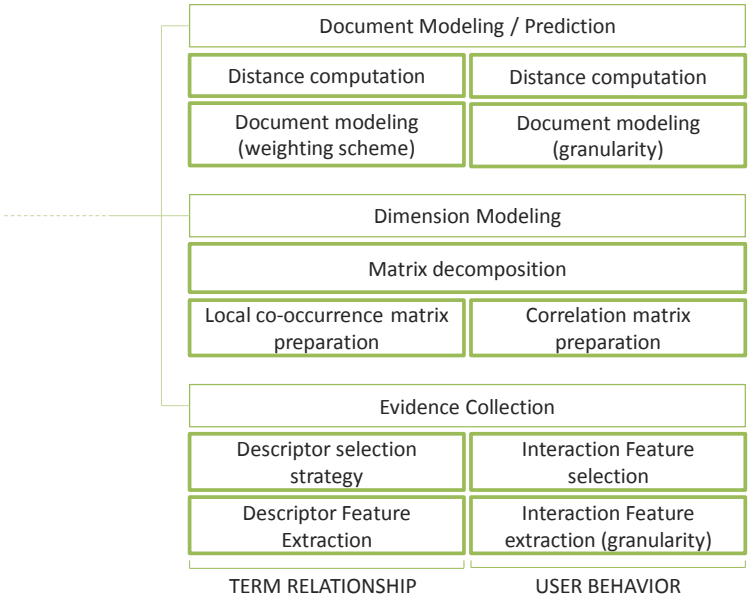


Figure 5.6: Functionalities implemented for source-based prediction.

but through an intermediate interface thus allowing alternative packages offering matrix utilities to be adopted. Entries of the basis vector spanning the dimension are adopted as boost for term weights, to perform *prediction* by Equation 3.8.

User Interaction Behavior

As mentioned in Section 5.2.2 the Indri Search Engine was adopted to perform content-based prediction for the user behavior dimension. But the software architecture was extended to handle user behavior based re-ranking. The user interaction data were parsed and stored in a SQLite database[14]. Functionalities were developed to access user interaction data and handle them, e.g. obtaining feature values at different granularities, for different topics, and perform feature selection. That constitutes the implementation for the evidence collection step. With regard to the dimension modeling step, functionalities to handle matrix operations were extended in order to compute the correlation matrix, adopted for the computation of PCA. Prediction is performed through an implementation of the ranking function described by Equation 3.8 that allows selection of the diverse possible eigenvectors obtained by PCA; it provides also functionalities to select the best performing projector in terms of NDCG among all the possible projectors.

[14] http://www.sqlite.org/

CHAPTER

SIX

EXPERIMENTAL RESULTS AND DESCRIPTION

Chapter 5 introduced some research questions concerning the application of the methodology to the two sources considered in this thesis; moreover the experimental methodology for investigating these questions was presented. This chapter describes the results obtained from the experimental investigation. Findings on term relationship dimension mainly concern its effectiveness to support document re-ranking and the effect of different implementations of the methodology steps. With regard to the user behavior dimension, besides its effectiveness for re-ranking, findings reported in this chapter concern the effect of the source for interaction features and the capability of the dimension to support query expansion.

6.1 Document Re-ranking through Term Relationship Dimension

6.1.1 Effect of Term Relationship in Relevant Documents on Re-ranking

Previous work [Melucci, 2008] showed that the technique adopted in this thesis to obtain the term relationship dimension when applied to the top ranked documents, i.e. PRF, is effective for document re-ranking. A further question is investigate its effectiveness when relevance data is available and therefore when term relationship can be modeled directly from document known to be relevant.

Let us assume that a user submits a query and obtains a list of m results. He assesses the top n documents in the result list. The question is if the user can benefit from re-ranking by term relationship dimension of the remaining $m - n$ documents or

the ranking induced by the first stage prediction is more beneficial. Results reported in Table 6.1 show that dimension-based re-ranking using the `UPD.1` feedback set was not able to outperform the stage one run: actually, the dimension-based re-ranking seems to provide an high decrement in terms of performance. When considering the variation in terms of performance with regard to diverse values of m, the obtained results show that the decrement is lower when the number of re-ranked documents becomes lower.

We carried out the same experiments using the TREC 2001 Web Track test collection. We measure the difference in terms of NDCG@10 respect to the baseline in order to gain some insights on the capability of the approach to rank high relevant documents at high rank position. Here the feedback set was constituted by the relevant documents among the top five retrieved by BM25, the latter being the baseline. As shown in Table 6.2 and Table 6.3 also in this case an high decrement both in terms of MAP and NDCG@10 was observed.

In order to investigate if the negative results were due to dimensionality reduction performed in the dimension modeling step, we performed the same experiments increasing the number of eigenvectors, specifically considering all the eigenvectors that saturated the 90% variance. Considering an higher number of eigenvectors provides results similar to those obtained when the first two eigenvectors are adopted for dimension modeling.

The above results suggest that, the specific implementation adopted for the methodology application to model term relationship is not effective for document re-ranking, but actually can provide a large negative effect in terms of statMAP, MAP and NDCG@10.

The lack of effectiveness observed from the obtained results can be due to the specific implementation of some of the steps constituting the methodology. In order to investigate this issue in the following sections we will adopt the methodology steps to support the evaluation of the considered implementation: each of the constituting steps will considered as a possible cause of the lack of effectiveness and alternative step implementations will be investigated in order to improve the predicting capability of dimension based re-ranking. Next section will investigate the selection of the source for term relationship in order to investigate the effect of using alternative feedback sets as input for the dimension modeling step.

6.1.2 Effect of Relevance Feedback Set on Document Re-ranking

One of the possible causes for poor results obtained in the previous section could have been the source for term relationship adopted, i.e. the specific feedback set. Since the TREC 2009 RF Track test collection provides several feedback sets, the effect of the

source for expansion terms and term relationship can be investigated. The feedback sets adopted are those assigned to the IMS Research Group during the participation to the track and briefly described in Section 5.2.3.1. Also in this case residual collection is adopted, but the actual number of document to re-rank can be higher than $m-n$ since the criteria for feedback set selection does not necessarily select documents present in the top m returned by the baseline described in Section 5.2.1.1.

Results reported in Table 6.4 show that also when exploiting diverse sources, re-ranking based on the term relationship dimension provides a large negative impact on the effectiveness.

The basic rationale of the above approach is to investigate if providing a feedback set with criteria other than that adopted in our baseline would allow to improve re-ranking. The documents re-ranked was those provided by our baseline. An alternative approach is to investigate if the feedback set is coupled with its baseline. For instance, when considering the `CMU.1` feedback set, instead of performing re-ranking of the our baseline the top m documents of the `CMU` baseline list are adopted. The effectiveness was measured also in this case on the residual collection. Results reported in Table 6.5 show that also in this case dimension-based re-ranking does not provide a positive contribution.

An analysis of the weights assigned to the document to re-rank by the adopted document representation show that many documents are represented by the same vectors and therefore many ties are present in the result list. Moreover, in the adopted document representation, when all the terms of the expanded query are present, the document is a vector of all zeros. Indeed a document is represented by a vector $\mathbf{d} = \mathbf{d}' - \overline{\mathbf{d}}$ where the ith element is $d'_i = 1$ if the term t_i is present, $d'_i = 0$ otherwise. The vector $\overline{\mathbf{d}}$ is a vector whose constituting elements are $\overline{d} = \frac{1}{|\mathcal{T}|}\sum_{i=1}^{|\mathcal{T}|} d'_i$. When all terms occur in the document, both $\mathbf{d}$ and $\mathbf{d}'$ are one vectors of dimension $\mathcal{T}$; therefore $\mathbf{d}$ is a vector of all zeros. That may be a cause poor performance observed for dimension-based re-ranking. Another reason could be that the document representation basically exploits only the presence or the absence of the terms in considered document. Next section will explicitly investigate if these aspects are causes of the poor results obtained by dimension-based re-ranking.

6.1.3 Effect of Document Representation on Document Re-ranking

The basic rationale of the document representation adopted in the previous section was to capture the correlation among the occurrence of query terms in a document. When considering this representation two remarks can be made: (i) when all the terms are present a document is represented as a vector of all zeros; (ii) the correlation

among terms is based only on the presence or the absence of a term in the document is considered.

In order to investigate if the first property of the representation can affect the effectiveness of re-ranking, we exploit a simple variant. We assume the presence in the query of an additional term that never occurs in the documents to re-rank; in this way the $\overline{d}$ is always less than 1. That corresponds to $\overline{d} = \frac{1}{|\mathcal{T}|+1} \sum_{i=1}^{|\mathcal{T}|} d'_i$. Experiments carried out with the same set of parameters show no difference between the two representation in terms of effectiveness.

In order to investigate the second question, namely the effect of the document representation, we perform re-ranking on the basis of three representations: TF·IDF, BM25, and satTF that does not consider the IDF component in the BM25 weighting scheme. Results reported in Tables 6.6b–6.8b show that dimension-based re-ranking can benefit from document representation when additional information besides term occurrence is considered. Indeed, re-ranking based on representations that consider statistical information on term occurrence outperform the original document representation. Same results are observed for other values of s and k. Obtained results show that re-ranking based on our feedback set provides the largest improvement ($m = 100$) in terms of effectiveness when BM25 and satTF are adopted as representation. In other words re-ranking of the residual top m in our baseline seems to benefit from the related feedback set `UPD.1`. Tables 6.6a–6.8a report the comparison with the baseline. None of the three representation is able to significantly outperform the baseline. For high values of m, all the three representation negatively affect the effectiveness.

An analysis of the values of the eigenvectors shows that in some cases the dimension modeling technique assigned zero values to terms in the original query. This is the case, for instance, of topic 7 "air travel information" where a zero weight is assigned to the term "information". A slight modification of the matrix preparation procedure is adopted to investigate if boosting entries corresponding to the correlation of terms in the original query with themselves (self-correlation) can affect the effectiveness of dimension-based re-ranking. Let us consider a term t_i part of the original query submitted by the user in the first stage. Let us consider a window of text centered around t_i and a generic term t_j among those constituting the expanded query and occurring within the text window. The procedure described in Section 4.1.3, propagates the t_j weight to both the entries of the correlation matrix corresponding to the pair (i, j). We test a modified version of this strategy where the term weight is propagated also to the entry (i, i) of the matrix, namely that corresponding to the self-correlation of the term t_i. This is done only when the term t_i is part of the original query. Results are reported in Table 6.9. Also in this case the dimension-based re-ranking is not able to significantly outperform the baseline. But the modified procedure for the preparation

of the correlation matrix provides a slight improvement in terms of effectiveness. This results suggest investigate alternative modeling procedures, e.g. with more effective procedure to prepare the correlation matrix or diverse matrix decompositions.

6.1.4 Effect of Term Selection Strategy on Document Re-ranking

The term selection strategy, as part of the evidence collection step, is crucial since provides terms then adopted for dimension modeling. The measure adopted in the experiments discussed in the previous sections was rTF·IDF. This measure does not take into account information on co-occurrence with query terms in order to extract expansion terms. Since the methodology application for modeling term relationship is based on co-occurrence data, we can investigate if exploiting a term selection strategy that exploit this information can be beneficial. In this thesis we considered LCA. Table 6.10 compared the effectiveness of document re-ranking when rTF·IDF and LCA are adopted for term selection. LCA provides a positive contribution in 7/18 cases, mainly for high values of m, but only in one of these cases the difference is significant. These results confirm previous findings that LCA is not effective in relevance feedback setting, even if the small number of documents used for feedback could have affected LCA effectiveness. Indeed, for most of the runs the average number of relevant document per topic is approximately two.

6.1.5 Effect of Properties of a Single Feedback Document on Re-ranking

The last research question concerns with the properties that can make a relevant document an effective input for dimension modeling. This investigation was carried out in the TREC 2010 Relevance Feedback track. At the moment of writing results have not be released yet for the RF track. But a preliminary evaluation based on the relevance judgments gathered in TREC 2009 indicates that, when BM25 weights are adopted for document representation, dimension-based re-ranking:

- is more effective when a random relevant document is adopted as input for feedback than when exploiting the shortest or the longest relevant document;
- is more effective when the most common relevant documents, e.g. a the document retrieved by most IR systems in TREC 2009, is adopted as input than when exploiting the least common relevant documents;
- is more effective when a random highly relevant document is adopted as input than when exploiting the most spammy relevant document.

	$m = 2500$			$m = 1000$			$m = 100$		
	statMAP		Δ (%)	**statMAP**		Δ (%)	**statMAP**		Δ (%)
k	B	UPD.1		B	UPD.1		B	UPD.1	
5	0.182	0.070	-61.54	0.153	0.084	-45.26	0.073	0.067	-8.34
10	0.182	0.074	-59.50	0.153	0.084	-45.02	0.073	0.067	-8.41
20	0.182	0.074	-59.57	0.153	0.083	-45.47	0.073	0.067	-7.98
30	0.182	0.076	-58.07	0.153	0.083	-46.54	0.073	0.066	-9.63

(a) s=1

	$m = 2500$			$m = 1000$			$m = 100$		
	statMAP		Δ (%)	**statMAP**		Δ (%)	**statMAP**		Δ (%)
k	B	UPD.1		B	UPD.1		B	UPD.1	
5	0.182	0.070	-61.32	0.153	0.084	-45.26	0.073	0.067	-8.34
10	0.182	0.070	-61.29	0.153	0.084	-45.06	0.073	0.067	-8.14
20	0.182	0.070	-61.46	0.153	0.083	-45.55	0.073	0.067	-7.71
30	0.182	0.067	-63.09	0.153	0.082	-46.73	0.073	0.066	-9.78

(b) s=2

Table 6.1: statMAP for the baseline (B) and the dimension-based re-ranking (UPD.1) computed on the residual results list, after the removal of the feedback documents. m denotes the number of documents obtained in the first stage prediction. The actual number of document re-ranked is therefore $m-n$ with $n = 5$ number of documents judged by the user. The number of relevant documents R for each topic can vary from $0 \leq R \leq n$. Results refers to the UPD.1 feedback set that consists on the top 5 documents of the baseline run B. The number of topics in UPD.1 with at least one relevant document among the top 5 was 37. For each table, results are reported for different values of the (i) number of documents obtained by the first stage prediction, $m \in \{2500, 1000, 100\}$, and (ii) the number of terms adopted to expand the query, $k \in \{5, 10, 20, 30\}$. Table 6.1a and Table 6.1b report the results respectively when the first eigenvector and the first two eigenvectors are adopted for dimension modeling.

	$m = 2500$			$m = 1000$			$m = 100$		
	MAP		Δ (%)	**MAP**		Δ (%)	**MAP**		Δ (%)
k	B	Prj		B	Prj		B	Prj	
5	0.077	0.018	-76.99	0.074	0.025	-66.26	0.050	0.035	-28.77
10	0.077	0.014	-81.31	0.074	0.022	-70.20	0.050	0.030	-39.44
20	0.077	0.014	-81.18	0.074	0.022	-70.07	0.050	0.030	-40.44
30	0.077	0.014	-81.44	0.074	0.022	-70.48	0.050	0.028	-43.66

(a) s=1

	$m = 2500$			$m = 1000$			$m = 100$		
	MAP		Δ (%)	**MAP**		Δ (%)	**MAP**		Δ (%)
k	B	Prj		B	Prj		B	Prj	
5	0.077	0.018	-76.73	0.074	0.025	-66.12	0.050	0.036	-28.17
10	0.077	0.015	-80.78	0.074	0.022	-69.93	0.050	0.031	-38.63
20	0.077	0.014	-81.18	0.074	0.022	-70.07	0.050	0.030	-40.44
30	0.077	0.014	-81.44	0.074	0.022	-70.34	0.050	0.028	-43.66

(b) s=2

Table 6.2: MAP for the baseline (B) and the dimension-based re-ranking (Prj) computed on the residual results list, after the removal of the feedback documents. Experiments are carried out on the TREC 2001 Web Track test collection. Results refers to the feedback set constituted by the relevant documents among the top 5 documents retrieved by BM25. For each table, results are reported for different values of the (i) number of documents obtained by the first stage prediction, $m \in \{2500, 1000, 100\}$, and (ii) the number of terms adopted to expand the query, $k \in \{5, 10, 20, 30\}$. Table 6.2a and Table 6.2b report the results respectively when the first eigenvector and the first two eigenvectors are adopted for dimension modeling.

	$m = 2500$			$m = 1000$			$m = 100$		
	NDCG@10		Δ (%)	**NDCG@10**		Δ (%)	**NDCG@10**		Δ (%)
k	B	Prj		B	Prj		B	Prj	
5	0.169	0.043	-74.62	0.169	0.052	-69.53	0.169	0.121	-28.46
10	0.169	0.029	-82.78	0.169	0.042	-74.91	0.169	0.098	-41.78
20	0.169	0.025	-85.50	0.169	0.043	-74.85	0.169	0.093	-44.91
30	0.169	0.026	-84.38	0.169	0.049	-71.24	0.169	0.086	-49.29

(a) s=1

	$m = 2500$			$m = 1000$			$m = 100$		
	NDCG@10		Δ (%)	**NDCG@10**		Δ (%)	**NDCG@10**		Δ (%)
k	B	Prj		B	Prj		B	Prj	
5	0.169	0.044	-73.96	0.169	0.052	-69.11	0.169	0.123	-27.46
10	0.169	0.031	-81.48	0.169	0.044	-74.14	0.169	0.100	-40.59
20	0.169	0.025	-85.50	0.169	0.043	-74.85	0.169	0.093	-44.91
30	0.169	0.026	-84.38	0.169	0.049	-71.24	0.169	0.086	-49.29

(b) s=2

Table 6.3: NDCG@10 for the baseline (B) and the dimension-based re-ranking (Prj) computed on the residual results list, after the removal of the feedback documents. Experiments are carried out on the TREC 2001 Web Track test collection. Results refers to the feedback set constituted by the relevant documents among the top 5 documents retrieved by BM25. For each table, results are reported for different values of the (i) number of documents obtained by the first stage prediction, $m \in \{2500, 1000, 100\}$, and (ii) the number of terms adopted to expand the query, $k \in \{5, 10, 20, 30\}$. Table 6.3a and Table 6.3b report the results respectively when the first eigenvector and the first two eigenvectors are adopted for dimension modeling.

		UPD.1			CMU.1			ilps.2		
m	k	B	Prj	Δ (%)	B	Prj	Δ (%)	B	Prj	Δ (%)
2500	5	0.182	0.070	-61.61**	0.190	0.071	-62.55**	0.193	0.072	-62.82**
	10	0.182	0.074	-59.57**	0.190	0.070	-63.25**	0.193	0.071	-63.46**
	20	0.182	0.074	-59.64**	0.190	0.068	-64.36**	0.193	0.068	-64.62**
	30	0.182	0.076	-58.14**	0.190	0.066	-65.38**	0.193	0.066	-65.57**
1000	5	0.153	0.084	-45.20**	0.161	0.084	-47.66**	0.163	0.083	-49.02**
	10	0.153	0.084	-45.01**	0.161	0.084	-48.04**	0.163	0.083	-49.20**
	20	0.153	0.083	-45.46**	0.161	0.084	-47.58**	0.163	0.083	-49.24**
	30	0.153	0.083	-45.81**	0.161	0.084	-47.91**	0.163	0.083	-48.97**
100	5	0.073	0.067	-8.37	0.080	0.071	-11.32	0.083	0.066	-19.85*
	10	0.073	0.067	-8.45	0.080	0.070	-12.97	0.083	0.066	-20.56*
	20	0.073	0.067	-8.02	0.080	0.068	-15.60	0.083	0.064	-22.81*
	30	0.073	0.066	-9.66	0.080	0.066	-18.03	0.083	0.069	-16.34

(a)

		PRIS.1			QUT.1			UMas.1		
m	k	B	Prj	Δ (%)	B	Prj	Δ (%)	B	Prj	Δ (%)
2500	5	0.194	0.070	-64.12**	0.195	0.063	-67.86**	0.191	0.070	-63.30**
	10	0.194	0.069	-64.31**	0.195	0.063	-67.84**	0.191	0.069	-63.94**
	20	0.194	0.072	-62.96**	0.195	0.064	-67.16**	0.191	0.069	-63.97**
	30	0.194	0.071	-63.23**	0.195	0.061	-68.90**	0.191	0.069	-63.99**
1000	5	0.165	0.073	-55.95**	0.165	0.067	-59.35**	0.162	0.083	-48.99**
	10	0.165	0.073	-55.93**	0.165	0.065	-60.80**	0.162	0.082	-49.00**
	20	0.165	0.073	-55.58**	0.165	0.064	-61.47**	0.162	0.082	-49.34**
	30	0.165	0.072	-56.09**	0.165	0.063	-62.05**	0.162	0.082	-49.37**
100	5	0.084	0.072	-14.80	0.085	0.074	-12.51	0.082	0.074	-9.58
	10	0.084	0.072	-14.69	0.085	0.074	-12.91	0.082	0.075	-8.38
	20	0.084	0.073	-13.06	0.085	0.073	-13.46	0.082	0.074	-10.07
	30	0.084	0.071	-15.33	0.085	0.073	-13.58	0.082	0.074	-10.19

(b)

Table 6.4: statMAP computed on the residual result list for baseline (B) and dimension-based re-ranking (Prj) using diverse feedback sets. Results are reported varying the number of documents re-ranked m and the number of expansion terms k. Differences (Δ's) statistically significant are marked by one asterisk (paired t-test, $p < 0.05$) or two asterisks (paired t-test, $p < 0.01$). Dimension modeling is performed using the first eigenvector ($s = 1$).

	statMAP					
	CMU			PRIS		
m	B	Prj	Δ (%)	B	Prj	Δ (%)
2500	0.222	0.068	-69.22**	0.209	0.059	-71.93**
1000	0.194	0.096	-50.70**	0.179	0.075	-58.00**
100	0.083	0.071	-14.24**	0.088	0.064	-27.30**

(a)

	statMAP					
	QUT			UMas		
m	B	Prj	Δ (%)	B	Prj	Δ (%)
2500	0.051	0.013	-74.07**	0.214	0.143	-33.21
1000	0.044	0.021	-52.04**	0.214	0.143	-33.21
100	0.013	0.012	-9.38**	0.084	0.078	-7.47

(b)

Table 6.5: statMAP computed on the residual result list for baseline (B) and dimension-based re-ranking (Prj) using diverse feedback sets. The baseline adopted for a specific feedback set was that submitted by the TREC participant that provided that feedback set. Results are reported varying the number of documents re-ranked m. The number of expansion terms adopted is $k = 5$ and dimension modeling is performed using the first eigenvector ($s = 1$). Differences (Δ's) statistically significant are marked by one asterisk (paired t-test, $p < 0.05$) or two asterisks (paired t-test, $p < 0.01$). Dimension modeling is performed using the first eigenvector ($s = 1$).

	UPD.1			CMU.1			ilps.2		
m	B	TFIDF	Δ (%)	B	TFIDF	Δ (%)	B	TFIDF	Δ (%)
2500	0.182	0.138	-24.39*	0.190	0.152	-19.98	0.193	0.144	-25.31*
1000	0.153	0.129	-15.60	0.161	0.139	-13.59	0.163	0.125	-23.12
100	0.073	0.074	1.07	0.080	0.079	-2.18	0.083	0.069	-16.93*
	PRIS.1			QUT.1			UMas.1		
m	B	TFIDF	Δ (%)	B	TFIDF	Δ (%)	B	TFIDF	Δ (%)
2500	0.194	0.157	-19.32	0.195	0.099	-49.11**	0.191	0.146	-23.86*
1000	0.165	0.126	-23.32	0.165	0.085	-48.71**	0.162	0.126	-21.79
100	0.084	0.091	8.41	0.085	0.076	-10.53	0.082	0.084	2.33

(a)

	UPD.1			CMU.1			ilps.2		
m	Prj	TFIDF	Δ (%)	Prj	TFIDF	Δ (%)	Prj	TFIDF	Δ (%)
2500	0.070	0.138	96.93**	0.071	0.152	113.70**	0.072	0.144	100.87**
1000	0.084	0.129	54.03**	0.084	0.139	65.10**	0.083	0.125	50.82*
100	0.067	0.074	10.31	0.068	0.079	15.32	0.066	0.069	3.64
	PRIS.1			QUT.1			UMas.1		
m	Prj	TFIDF	Δ (%)	Prj	TFIDF	Δ (%)	Prj	TFIDF	Δ (%)
2500	0.070	0.157	124.87**	0.063	0.099	58.35**	0.070	0.146	107.44**
1000	0.073	0.126	74.09**	0.067	0.085	26.18	0.083	0.126	53.33*
100	0.072	0.091	27.24	0.074	0.076	2.27*	0.074	0.084	13.18*

(b)

Table 6.6: Table 6.6a reports the statMAP computed on the residual result list for the baseline (B) and dimension-based document re-ranking using TF·IDF document representation (TFIDF). Table 6.6b reports the comparison between dimension-based re-ranking using the document representation based on correlation (Prj) and on TF·IDF weights (TFIDF). Results are reported varying the number of documents re-ranked m. The number of expansion terms is $k = 5$. Differences (Δ's) statistically significant are marked by one asterisk (paired t-test, $p < 0.05$) or two asterisks (paired t-test, $p < 0.01$). Dimension modeling is performed using the first eigenvector ($s = 1$).

	UPD.1			CMU.1			ilps.2		
m	B	BM25	Δ (%)	B	BM25	Δ (%)	B	BM25	Δ (%)
2500	0.182	0.160	-12.07	0.190	0.170	-10.61	0.193	0.141	-26.72**
1000	0.153	0.149	-2.07	0.161	0.151	-5.86	0.163	0.122	-25.15**
100	0.073	0.089	21.88	0.080	0.089	10.18	0.083	0.083	-0.18
	PRIS.1			QUT.1			UMas.1		
m	B	BM25	Δ (%)	B	BM25	Δ (%)	B	BM25	Δ (%)
2500	0.194	0.122	-37.15**	0.195	0.082	-57.93**	0.191	0.147	-23.32**
1000	0.165	0.105	-36.35**	0.165	0.076	-53.80**	0.162	0.129	-20.20*
100	0.084	0.098	17.06	0.085	0.078	-8.20	0.082	0.085	2.99

(a)

	UPD.1			CMU.1			ilps.2		
m	Prj	BM25	Δ (%)	Prj	BM25	Δ (%)	Prj	BM25	Δ (%)
2500	0.070	0.160	129.01**	0.071	0.170	138.72**	0.072	0.141	97.08**
1000	0.084	0.149	78.72*	0.084	0.151	79.87**	0.083	0.122	46.84*
100	0.067	0.089	33.02**	0.068	0.089	29.89**	0.066	0.083	24.54**
	PRIS.1			QUT.1			UMas.1		
m	Prj	BM25	Δ (%)	Prj	BM25	Δ (%)	Prj	BM25	Δ (%)
2500	0.070	0.122	75.16**	0.063	0.082	30.91*	0.070	0.147	108.92**
1000	0.073	0.105	44.50**	0.067	0.076	13.67	0.083	0.129	56.44**
100	0.072	0.098	37.39	0.074	0.078	4.93*	0.074	0.085	13.91*

(b)

Table 6.7: Table 6.7a reports the statMAP computed on the residual result list for the baseline (B) and dimension-based re-ranking using BM25 document representation (BM25). Table 6.7b reports the comparison between dimension-based re-ranking using the document representation based on correlation (Prj) and on BM25 weights (BM25). Results are reported varying the number of documents re-ranked m. The number of expansion terms is $k = 5$. Differences (Δ's) statistically significant are marked by one asterisk (paired t-test, $p < 0.05$) or two asterisks (paired t-test, $p < 0.01$). Dimension modeling is performed using the first eigenvector ($s = 1$).

	UPD.1			CMU.1			ilps.2		
m	B	satTF	Δ (%)	B	satTF	Δ (%)	B	satTF	Δ (%)
2500	0.182	0.160	-12.00	0.190	0.173	-9.42	0.193	0.145	-24.63**
1000	0.153	0.153	0.09	0.161	0.152	-5.46	0.163	0.125	-23.37**
100	0.073	0.090	23.90	0.080	0.086	7.15	0.083	0.081	-2.25
	PRIS.1			QUT.1			UMas.1		
m	B	satTF	Δ (%)	B	satTF	Δ (%)	B	satTF	Δ (%)
2500	0.194	0.123	-36.54*	0.195	0.084	-57.11**	0.191	0.152	-20.37**
1000	0.165	0.103	-37.70**	0.165	0.077	-53.24**	0.162	0.133	-17.84*
100	0.084	0.098	17.05	0.085	0.078	-7.39	0.082	0.085	3.55

(a)

	UPD.1			CMU.1			ilps.2		
m	Prj	satTF	Δ (%)	Prj	satTF	Δ (%)	Prj	satTF	Δ (%)
2500	0.070	0.160	129.20**	0.071	0.173	141.89**	0.072	0.145	102.70**
1000	0.084	0.153	82.66**	0.084	0.152	80.63**	0.083	0.125	50.32**
100	0.067	0.090	35.22**	0.068	0.086	26.32**	0.066	0.081	21.96**
	PRIS.1			QUT.1			UMas.1		
m	Prj	satTF	Δ (%)	Prj	satTF	Δ (%)	Prj	satTF	Δ (%)
2500	0.070	0.123	76.87**	0.063	0.084	33.44*	0.070	0.152	116.96**
1000	0.073	0.103	41.45**	0.067	0.077	15.05	0.083	0.133	61.06**
100	0.072	0.098	37.38	0.074	0.078	5.85	0.074	0.085	14.53*

(b)

Table 6.8: Table 6.8a reports the statMAP computed on the residual result list for the baseline (B) and dimension-based re-ranking using a document representation based on saturated and normalized term frequency (satTF). Table 6.8b reports the comparison between dimension-based re-ranking using the document representation based on correlation (Prj) and satTF. Results are reported varying the number of documents re-ranked m. The number of expansion terms is $k = 5$. Differences (Δ's) statistically significant are marked by one asterisk (paired t-test, $p < 0.05$) or two asterisks (paired t-test, $p < 0.01$). Dimension modeling is performed using the first eigenvector ($s = 1$).

Set	m	B	BM25	$\Delta_{\text{BM25}-\text{B}}$ (%)	boost	$\Delta_{\text{boost}-\text{B}}$ (%)	$\Delta_{\text{boost}-\text{BM25}}$ (%)
UPD.1	2500	0.182	0.160	-12.07	0.166	-9.03	3.46**
	1000	0.153	0.149	-2.07	0.155	1.50	3.65**
	100	0.073	0.089	21.88	0.089	22.47	0.49
CMU.1	2500	0.190	0.170	-10.61	0.172	-9.90	0.80
	1000	0.161	0.151	-5.86	0.152	-5.25	0.64
	100	0.080	0.089	10.18	0.087	8.56	-1.48
PRIS.1	2500	0.194	0.122	-37.15**	0.124	-35.99**	1.85
	1000	0.165	0.105	-36.35**	0.109	-33.91*	3.85*
	100	0.084	0.098	17.06	0.100	19.32	1.93
QUT.1	2500	0.195	0.082	-57.93**	0.086	-56.05**	4.46*
	1000	0.165	0.076	-53.80**	0.080	-51.37**	5.25
	100	0.085	0.078	-8.20	0.078	-8.11	0.10
Ilps.2	2500	0.193	0.141	-26.72**	0.148	-23.50**	4.40**
	1000	0.163	0.122	-25.15**	0.127	-22.34**	3.75*
	100	0.083	0.083	-0.18	0.081	-2.52	-2.34
UMas.1	2500	0.191	0.147	-23.32**	0.149	-22.10**	1.59
	1000	0.162	0.129	-20.20*	0.132	-18.50*	2.13
	100	0.082	0.085	2.99	0.085	3.46	0.46

Table 6.9: statMAP computed on the residual result list for the baseline and dimension-based re-ranking propagating term weights to query term entries in the correlation matrix for dimension modeling. BM25 refers to dimension-based re-ranking using BM25 weights for document representation and the original procedure to prepare the correlation matrix. boost refers to the modified version of the procedure to prepare the correlation matrix. Differences (Δ's) statistically significant are marked by one asterisk (paired t-test, $p < 0.05$) or two asterisks (paired t-test, $p < 0.01$). Dimension modeling is performed using the first eigenvector ($s = 1$).

Set	m	B	rTFIDF	$\Delta_{\text{rTFIDF}-\text{B}}$ (%)	LCA	$\Delta_{\text{LCA}-\text{B}}$ (%)	$\Delta_{\text{LCA}-\text{rTFIDF}}$ (%)
UPD.1	2500	0.182	0.160	-12.07	0.144	-20.94**	-10.08*
	1000	0.153	0.149	-2.07	0.135	-11.68	-9.81
	100	0.073	0.089	21.88	0.083	13.66	-6.75
CMU.1	2500	0.190	0.170	-10.61	0.164	-14.05	-3.85
	1000	0.161	0.151	-5.86	0.141	-12.08	-6.61
	100	0.080	0.089	10.18	0.081	1.30	-8.06*
PRIS.1	2500	0.194	0.122	-37.15**	0.131	-32.75*	7.00
	1000	0.165	0.105	-36.35**	0.111	-32.53*	6.02
	100	0.084	0.098	17.06	0.078	-7.38	-20.88
QUT.1	2500	0.195	0.082	-57.93**	0.082	-58.00**	-0.16
	1000	0.165	0.076	-53.80**	0.079	-52.41**	3.00
	100	0.085	0.078	-8.20	0.077	-8.78	-0.64
Ilps.2	2500	0.193	0.141	-26.72**	0.150	-22.30*	6.04*
	1000	0.163	0.122	-25.15**	0.125	-23.67**	1.97
	100	0.083	0.083	-0.18	0.075	-9.57	-9.41*
UMas.1	2500	0.191	0.147	-23.32**	0.146	-23.52**	-0.26
	1000	0.162	0.129	-20.20*	0.133	-17.93	2.84
	100	0.082	0.085	2.99	0.087	5.36	2.30

Table 6.10: statMAP computed on the residual result list for baseline (B) and dimension-based re-ranking using diverse term selection strategies: rTFIDF and LCA. Results are reported varying the number of documents re-ranked m. The number of expansion terms adopted is $k = 5$ and dimension modeling is performed using the first eigenvector ($s = 1$). Differences (Δ's) statistically significant are marked by one asterisk (paired t-test, $p < 0.05$) or two asterisks (paired t-test, $p < 0.01$).

6.2 Document Re-ranking through User Behavior Dimension

6.2.1 Effect of group data on document re-ranking.

6.2.1.1 Source selection

The research question reported in Section 5.1.2.1 concerns with the impact of the selection of the source combination on document re-ranking. In particular, we are interested in understanding if using group data both for modeling the user behavior and for representing documents negatively affects document re-ranking respect to exploit data distilled from the individual. If re-ranking effectiveness is comparable, group data can be adopted when personal data is not available for one or both the required representations. In terms of the combinations discussed in Section 4.2.1 the goal is to compare the effectiveness of the `P/P` combination with the `G/-` and `-/G` combinations.

Table 6.11 reports the obtained results. The effectiveness of the diverse combination is comparable. The only significant differences are observed for the `G/P` case. The `G/P` performance could be due both to the fact the a non-personalized dimension is adopted and the comparison is performed between a dimension and a document representation obtained from diverse sources for interaction features. Because of its lack of effectiveness, in the remainder of this section this combination will be no longer considered. Looking at the mean and the median NDCG@10, results show that `P/P` and `G/G` benefit from more evidence; yet NDCG@10 increased monotonically with the number of documents used as evidence for none of the combinations. Even if the results are comparable, the adoption of group data can negatively affect re-ranking. When exploiting $n_B = 3$ group-based combinations provide a little improvement on `P/P`; for instance, the `Gd/G` combination provide an improvement of 6%. But there is basically no improvement when increasing the number of feedback. Moreover, even if the `G/G` and `Gd/G` combinations seem to be promising for $n_B = 3$, when considering the results per topic and per user — see Table 6.12 — they show that also in this case the adoption of group data can affect effectiveness of re-ranking. This is the case, for instance, of the topics 536, 543 and 550 where all the three combinations involving group data perform worse than the `P/P` case.

The fact the `-/G` combinations perform similarly to `P/P` could be due to the level of agreement between what the individual user and the group perceived as relevant. In order to investigate this hypothesis, the NDCG@10's for these combinations are plotted against the AP correlation coefficient [Yilmaz et al., 2008], τ_{AP}, computed between the ideal individual ranking and the ideal group ranking for each (user,topic) pair. The

ideal ranking for a user is obtained by ranking documents by the gain he provided. For the group ideal ranking, the gain of each document is obtained as the sum of the gains provided for that document by the users in the group as discussed in Section 5.2.3.3. The basic rationale underlying this choice is to investigate if there is a relationship between the effectiveness of the group combination to support personalized document re-ranking and the agreement between the perception of relevance of the group and that of the individual. Here the correlation among the ideal ranking of the individual and the ideal ranking of the group is adopted as measure of agreement. τ_{AP} is adopted since it has been shown to give more weight to the errors at high rank positions than the Kendall Tau correlation [Yilmaz et al., 2008]. The scatter-plots in Figure 6.1 do not indicate a strong correlation between agreement and performance.

One of the questions introduced in Section 4.2.3 concerned the adoption of average feature values for dimension modeling. In particular the question was if using average feature values on group-based combinations could affect re-ranking effectiveness. This question can be addressed by comparing the `G/G` and `Gd/G` combinations. Results are reported in Table 6.13. For $n_B = 3$ exploiting average values in the adopted approach can reduce the power of PCA of extracting effective behavioral patterns. But when increasing the number of documents used for modeling the user behavior dimension, the difference between the two rapidly decreases both in terms of mean and median NDCG@10 — see Table 6.13b. Since `G/G` is more robust than `Gd/G` when varying the number of documents for dimension modeling, the former combination will be adopted when using user behavior dimension to address the research questions reported in Section 5.1.2.3 and Section 5.1.2.4.

With regard to the research question posed in Section 5.1.2.1 the findings reported in this section show that in the considered dataset:

- the adoption of group data instead that personal data can negatively affect re-ranking, even if we observed a positive contribution of group-based combination when a small number of feedback documents is adopted for dimension modeling.

Moreover, we observed that

- in the event of a small number of feedback documents, when group data provide a positive contribution, the re-ranking effectiveness is not strongly correlated with the agreement on document relevance among the users constituting the groups;

- the adoption of average feature values for the `G/G` combination does not significantly affect re-ranking effectiveness.

6.2.1.2 Comparison with the Baseline

The above remarks concerned with the comparison among the diverse combinations thus investigating the effect of using group data instead of personal data for personalized user behavior-based re-ranking. But no comparison was performed with the baseline `B`. Results reported in Table 6.14 show that none of the combinations significantly outperformed the baseline ranking; the same result holds when considering NDCG@10 per topic and per user — see Table 6.15.

The relatively small number of experimental records is definitely a limitation since small numbers make the detection of significant differences harder than the detection based on large datasets. But, the small size of the dataset allows us to note that the two sources of features (i.e. `P` and `G`) adopted seem to provide diverse contributions. For instance, for topics 518, 536, and 546 only one of the two combinations performs better than the baseline. This suggests to investigate combinations of the diverse feature granularities.

6.2.2 Effect of the number of relevant documents on document re-ranking.

The representation of the user behavior exploits the data gathered from the first visited documents by the users, extracts possible patterns (i.e. eigenvectors of the correlation matrix) from those data and uses the most effective pattern for re-ranking. If the visited documents are relevant, it is necessary to investigate whether the improvement in terms of effectiveness can mainly be due to the ability of the user to select relevant documents. To this end, we investigated the relationship between the number of relevant documents among the top $n_B = 3$ visited and NDCG@10 across the diverse combinations. In Figure 6.2b the results are depicted. The NDCG@10's measured for the baseline when considering all the users and all the topics is plotted against the number of relevant documents in the top three visited — the regression lines are reported for providing an idea of the trend. For the diverse combinations the dependence with the number of relevant documents among the top three visited is still linear, but the slope decreases and the intercept increases. The average and the median NDCG@10 are higher than for the baseline when only one relevant document is present among those used for feedback, but this increment decreases when increasing the number of relevant documents; the same results are observed for $n_B \in \{4, 5\}$. The main limitation is the robustness of the adopted approach. Indeed, when considering the variance of NDCG@10 values, in the event of one relevant documents the variance is smaller than that obtained for the baseline; differently, when the number of relevant documents increase, the baseline has smaller variance, thus suggesting that even if the

user behavior based re-ranking can provide some improvement, the latter is less robust than the baseline. The same result is observed when the first principal eigenvector is adopted for dimension modeling — see Figure 6.2c.

Therefore, with regard to the research question posed in Section 5.1.2.2 findings reported in this section indicate that re-ranking based on the user behavior dimension when only one relevant document is present among the top visited, is able to increase the NDCG@10; but when the number of relevant documents increase, also the variance increase, thus suggesting a lack of robustness of the approach respect to the baseline.

6.2.3 Effect of User Behavior-based Document Re-ranking on Query Expansion.

The investigation of the research question reported in Section 5.1.2.1 and Section 5.1.2.2 showed that the improvement in terms of retrieval effectiveness provided by user behavior based re-ranking is not consistent throughout all the topics or all the users, but the effectiveness of the top ten document re-ranking seems to be not strictly dependent from the relevance of the documents used to model the user behavior. For this reason we investigate if the top n_F documents re-ranked by the user behavior are a more effective evidence for query expansion than the top n_F retrieved by the baseline. The basic idea is to investigate if use behavior-based re-ranking is able to bring at high rank positions good sources for query expansion, thus improving the effectiveness respect to PRF where the top ranked document retrieved by the baseline are supposed to be good sources for feedback. In the remainder of this chapter we will denote with PRF the application of the Indri Pseudo-Relevance Feedback algorithm to the top n_F documents retrieved by the baseline. IRF will denote the application of the Indri Pseudo-Relevance Feedback algorithm to the top n_F documents re-ranked by user behavior.

Table 6.16 reports the results when $n_B = 3$ documents are used for dimension modeling. Results show that query expansion can benefit from user behavior based re-ranking, even if the improvement is modest. We investigated the effect of the number n_F of feedback documents adopted for query expansion and of the number of expansion terms, k. Results show that the adopted approach, IRF, benefits from a small number of feedback documents, i.e. $n_F = \{1, 2\}$, and an increment of the number of terms used for query expansion, i.e. $k = \{10, 20\}$. For most of the parameters pairs IRF can improve PRF: for 39/45 cases IRF perform equal of better than PRF, and for 26/45 the increment in terms of NDCG is higher than 5%.

However, this specific methodology implementation should be improved since it is not robust. Let us consider, for instance, the case for $k = 10$ and $n_F = 2$, where PRF does not improve the baseline ($\Delta_{\text{PRF-B}} = 0.24\%$) differently from IRF ($\Delta_{\text{IRF-B}} =$

8.67%), and the difference in terms of NDCG@10 is greater than 5%. Table 6.18 reports the results for each topic and show that also in this case, IRF is not able to outperform PRF for all the topics.

Therefore, with regard to the research question posed in Section 5.1.2.3, findings reported in this Section indicate that user behavior-based re-ranking is able to provide an improvement respect to PRF, increasing the number of good sources for feedback at high rank positions and supporting feedback when a small number of documents is adopted as input, e.g. one or two documents. But an analysis of the effectiveness per topic show that a more robust methodology implementation is required since for some topics PRF performed better than IRF.

6.2.4 Effect of the Number of Relevant Documents Used for Dimension Modeling on Query Expansion

The objective of the research question discussed in Section 6.2.3 was to investigated the capability of the user behavior dimension to increase the number of good sources, namely documents, for query expansion at high rank position, thus increasing the effectiveness of Pseudo-Relevance Feedback. The results showed that the Indri Pseudo-Relevance Feedback algorithm can benefit from a preliminary user behavior based re-ranking. In order to gain more insights in the user behavior dimension capability to support query expansion, as done in Section 6.2.2, the effect of the number of relevant documents in the top n_B can be investigated. The objective is to understand if, also when there is a small number of relevant documents in the n_B adopted to model the dimension, user behavior-based re-ranking is able to improve the Indri Pseudo-Relevance Feedback algorithm capability of ranking highly relevant documents at high rank positions.

Results are reported in Table 6.19. When there are no relevant documents among the top 3 of the baseline, the effectiveness of feedback is low and PRF performs better. Differently when only one or two relevant documents are present in the top 3 used for pseudo-feedback (PRF) or dimension modeling (IRF), IRF outperforms PRF thus suggesting the is able to improve the number of good sources for content-based feedback in the top 3. The fact the PRF performs better than IRF suggests that when no relevant documents are adopted for dimension modeling the effectiveness of the model is negatively affected. When the number of relevant documents is three, namely all the feedback documents are relevant, the two approaches perform equally.

When compared with the baseline — see Table 6.20 — IRF is able to provide a positive contribution when one or two relevant documents are present in the top 3, but both the feedback techniques hurt the initial ranking when the top three documents are relevant.

Therefore, with regard to the research question posed in Section 5.1.2.4, results reported in this section indicate that the number of relevant documents among those used for content-based feedback (PRF) or for dimension modeling (IRF) does not affect the effectiveness of query expansion. The main reason is that relevant documents are not necessarily good sources for query expansion; with regard to the specific algorithm for feedback adopted, relevant documents are not necessarily good sources to extract the factors that, through the Relevance Model, allow a better representation of the information need to be obtained.

	Personal	Group-based combinations				Increment (%)			
n_B	P/P	P/G	G/P	G/G	Gd/G	$\Delta_{P/G}$	$\Delta_{G/P}$	$\Delta_{G/G}$	$\Delta_{Gd/G}$
3	0.765	0.791	0.759	0.777	0.797	3.328	-0.823	1.509	4.230
4	0.772	0.781	0.761	0.787	0.799	1.154	-1.406	1.915	3.471
5	0.775	0.789	0.773	0.792	0.786	1.845	-0.205	2.200	1.504
6	0.784	0.796	0.774	0.803	0.784	1.583	-1.186	2.508	0.029
7	0.785	0.801	0.775	0.811	0.786	1.995	-1.245	3.303	0.064
8	0.787	0.788	0.774	0.802	0.789	0.072	-1.718	1.863	0.295
9	0.784	0.791	0.773	0.803	0.795	0.966	-1.340	2.481	1.466
10	0.789	0.790	0.773	0.789	0.792	0.136	-2.086	0.011	0.372

(a)

	Personal	Group-based combinations				Increment (%)			
n_B	P/P	P/G	G/P	G/G	Gd/G	$\Delta_{P/G}$	$\Delta_{G/P}$	$\Delta_{G/G}$	$\Delta_{Gd/G}$
3	0.817	0.832	0.799	0.825	0.869	1.748	-2.238	0.922	6.313
4	0.839	0.825	0.805*	0.827	0.848	-1.615	-4.056	-1.324	1.133
5	0.833	0.835	0.826	0.835	0.832	0.288	-0.817	0.288	-0.179
6	0.839	0.843	0.833	0.839	0.825	0.524	-0.656	0.045	-1.615
7	0.847	0.840	0.831	0.848	0.833	-0.798	-1.829	0.137	-1.662
8	0.841	0.832	0.839*	0.835	0.833	-1.164	-0.333	-0.701	-1.014
9	0.847	0.835	0.839*	0.838	0.833	-1.351	-0.985	-1.109	-1.662
10	0.853	0.835	0.839**	0.832	0.832	-2.049	-1.686	-2.506	-2.506

(b)

Table 6.11: Comparison among the mean (Table 6.14a) median (Table 6.14b) NDCG@10 of the diverse source combinations when varying n_B, i.e. the number of documents used to obtain the user behavior dimension. Mean and median NDCG@10 are computed over all the (topic,user) pairs for all the combinations when $3 \leq n_B \leq 10$ and the eigenvector adopted for dimension modeling for each pair is the one that maximizes the NDCG@10 among all the possible eigenvectors extracted by PCA. $n_B < 3$ is not considered because the number of patterns for the diverse combinations is usually one and this pattern is not effective. Significant differences are marked by one asterisk (one-tailed Wilcoxon signed ranked test, $p < 0.05$) or two asterisks (one-tailed Wilcoxon signed ranked test, $p < 0.01$). One-tailed test was adopted in order to investigate if P/P combination performed better than the group-based combinations.

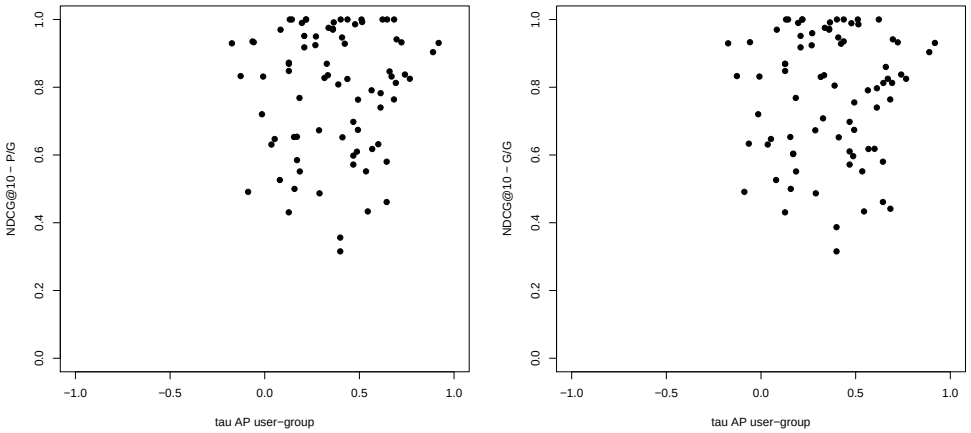
	Personal	Group-based			Increment (%)		
Topic	P/P	P/G	G/G	Gd/G	$\Delta_{P/G-P/P}$	$\Delta_{G/G-P/P}$	$\Delta_{Gd/G-P/P}$
501	0.702	**0.854**	0.667	0.710	21.658	-4.888	1.133
502	0.551	0.613	**0.618**	0.616	11.162	12.089	11.703
504	0.768	0.725	0.725	**0.802**	-5.648	-5.554	4.451
506	0.772	**0.834**	**0.834**	**0.834**	7.986	7.986	7.986
509	0.863	0.883	0.883	**0.900**	2.406	2.406	4.349
510	0.829	0.838	0.840	**0.849**	1.005	1.335	2.393
511	0.844	0.879	0.879	**0.897**	4.185	4.095	6.205
517	0.770	**0.894**	0.850	0.872	16.101	10.441	13.255
518	0.786	0.902	0.902	**0.903**	14.798	14.756	14.840
536	**0.760**	0.661	0.663	0.663	-13.024	-12.697	-12.697
543	**0.816**	0.593	0.566	0.592	-27.272	-30.567	-27.388
544	0.793	0.814	0.813	**0.850**	2.664	2.534	7.212
546	0.593	0.830	0.834	**0.835**	40.004	40.544	40.765
550	**0.820**	0.696	0.712	0.742	-15.075	-13.185	-9.442

(a)

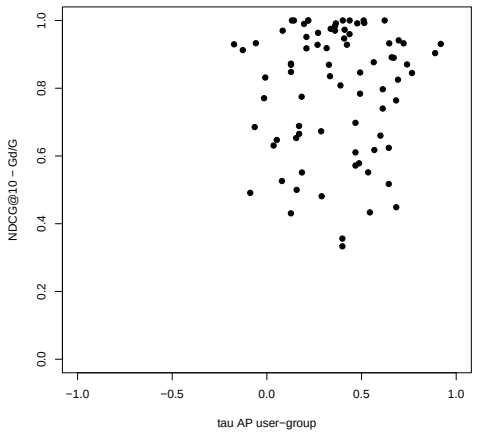
	Personal	Group-based			Increment (%)		
Topic	P/P	P/G	G/G	Gd/G	$\Delta_{P/G-P/P}$	$\Delta_{G/G-P/P}$	$\Delta_{Gd/G-P/P}$
user1	0.774	0.766	0.764	**0.810**	-1.021	-1.239	4.613
user2	0.796	0.844	0.820	**0.885**	5.983	3.073	11.238
user3	0.701	0.729	0.728	**0.747**	4.053	3.991	6.647
user5	0.703	0.798	**0.804**	0.803	13.523	14.313	14.223
user7	**0.862**	0.823	0.684	0.699	-4.516	-20.717	-18.875
user8	0.760	0.758	0.759	**0.766**	-0.341	-0.171	0.715
user9	**0.856**	0.759	0.757	0.770	-11.292	-11.515	-9.988
user10	0.754	**0.886**	0.836	0.847	17.508	10.843	12.265
user11	**0.880**	0.776	0.776	0.776	-11.812	-11.812	-11.812
user12	0.712	0.756	0.747	**0.767**	6.188	4.991	7.769
user13	0.813	0.839	0.812	**0.849**	3.215	-0.078	4.420
user15	0.826	0.820	0.834	**0.849**	-0.663	1.021	2.798
user16	0.584	0.733	0.737	**0.745**	25.358	26.037	27.495

(b)

Table 6.12: NDCG@10 per topic and per user for the diverse combinations and percentage of increment with regard to the P/P combination.



(a) `P/G` (b) `G/G`



(c) `Gd/G`

Figure 6.1: NDCG@10 for the `P/G` (Fig. 6.3b), the `G/G` (Fig. 6.3c) and the `Gd/G` (Fig. 6.3d) combination plotted against the τ_{AP} between user and group (not including the user) gains. NDCG@10's refer to the case when observations gathered from the first $n_B = 3$ documents visited were adopted for dimension modeling.

	Combinations		Increment (%)		Combinations		Increment (%)
Topic	G/G	Gd/G	$\Delta_{\texttt{Gd/G}-\texttt{G/G}}$	User	G/G	Gd/G	$\Delta_{\texttt{Gd/G}-\texttt{G/G}}$
501	0.667	0.710	6.33	user1	0.764	0.810	5.92
502	0.618	0.616	-0.34	user2	0.820	0.885	7.92
504	0.725	0.802	10.59	user3	0.728	0.747	2.55
506	0.834	0.834	0.00	user5	0.804	0.803	-0.08
509	0.883	0.900	1.90	user7	0.684	0.699	2.32
510	0.840	0.849	1.04	user8	0.759	0.766	0.89
511	0.879	0.897	2.03	user9	0.757	0.770	1.73
517	0.850	0.872	2.55	user10	0.836	0.847	1.28
518	0.902	0.903	0.07	user11	0.776	0.776	0.00
536	0.663	0.663	0.00	user12	0.747	0.767	2.65
543	0.566	0.592	4.58	user13	0.812	0.849	4.50
544	0.813	0.850	4.56	user15	0.834	0.849	1.76
546	0.834	0.835	0.16	user16	0.737	0.745	1.16
550	0.712	0.742	4.31				

(a)

	Mean				Median		
	Combinations		Increment (%)		Combinations		Increment (%)
n_B	G/G	Gd/G	$\Delta_{\texttt{Gd/G}-\texttt{G/G}}$	n_B	G/G	Gd/G	$\Delta_{\texttt{Gd/G}-\texttt{G/G}}$
3	0.777	0.797	2.681	3	0.825	0.869	5.342**
4	0.787	0.799	1.527	4	0.827	0.848	2.491
5	0.792	0.786	-0.681	5	0.835	0.832	-0.466
6	0.803	0.784	-2.419	6	0.839	0.825	-1.659
7	0.811	0.786	-3.136	7	0.848	0.833	-1.797*
8	0.802	0.789	-1.539	8	0.835	0.833	-0.316
9	0.803	0.795	-0.991	9	0.838	0.833	-0.560
10	0.789	0.792	0.361	10	0.832	0.832	0.000

(b)

Table 6.13: NDCG@10 obtained by user behavior-based re-ranking using G/G and Gd/G combinations. Table 6.13a reports mean NDCG@10 per topic and per user when $n_B = 3$ documents are used for dimension modeling. Table 6.13b reports mean and median NDCG@10 for the G/G and Gd/G combination for different values of n_B. When considering the median values in Table 6.13b, significant differences are marked by one asterisk (two-tailed Wilcoxon signed ranked test, $p < 0.05$) or two asterisks (two-tailed Wilcoxon signed ranked test, $p < 0.01$).

	Baseline	Source combinations				Increment (%)			
n_B	B	P/P	P/G	G/G	Gd/G	$\Delta_{\text{P/P}-\text{B}}$	$\Delta_{\text{P/G}-\text{B}}$	$\Delta_{\text{G/G}-\text{B}}$	$\Delta_{\text{Gd/G}-\text{B}}$
3	0.765	0.765	0.791	0.777	0.797	0.061	3.391	1.571	4.294
4	0.765	0.772	0.781	0.787	0.799	0.969	2.133	2.903	4.474
5	0.765	0.775	0.789	0.792	0.786	1.308	3.177	3.537	2.832
6	0.765	0.784	0.796	0.803	0.784	2.485	4.107	5.056	2.514
7	0.765	0.785	0.801	0.811	0.786	2.692	4.741	6.085	2.758
8	0.765	0.787	0.788	0.802	0.789	2.944	3.019	4.862	3.248
9	0.765	0.784	0.791	0.803	0.795	2.516	3.506	5.060	4.019
10	0.765	0.789	0.790	0.789	0.792	3.214	3.354	3.226	3.598

(a)

	Baseline	Source combinations				Increment (%)			
n_B	B	P/P	P/G	G/G	Gd/G	$\Delta_{\text{P/P}-\text{B}}$	$\Delta_{\text{P/G}-\text{B}}$	$\Delta_{\text{G/G}-\text{B}}$	$\Delta_{\text{Gd/G}-\text{B}}$
3	0.838	0.817	0.832	0.825	0.869	-2.462	-0.757	-1.563	3.696
4	0.838	0.839	0.825	0.827	0.848	0.053	-1.563	-1.272	1.187
5	0.838	0.833	0.835	0.835	0.832	-0.604	-0.317	-0.317	-0.781
6	0.838	0.839	0.843	0.839	0.825	0.053	0.577	0.098	-1.563
7	0.838	0.847	0.840	0.848	0.833	1.048	0.242	1.187	-0.632
8	0.838	0.841	0.832	0.835	0.833	0.387	-0.781	-0.317	-0.632
9	0.838	0.847	0.835	0.838	0.833	1.048	-0.317	-0.072	-0.632
10	0.838	0.853	0.835	0.832	0.832	1.769	-0.317	-0.781	-0.781

(b)

Table 6.14: Table 6.14a and Table 6.14b report respectively mean and median NDCG@10 for the baseline and the diverse source combinations when varying the number of documents adopted for dimension modeling, i.e. n_B.

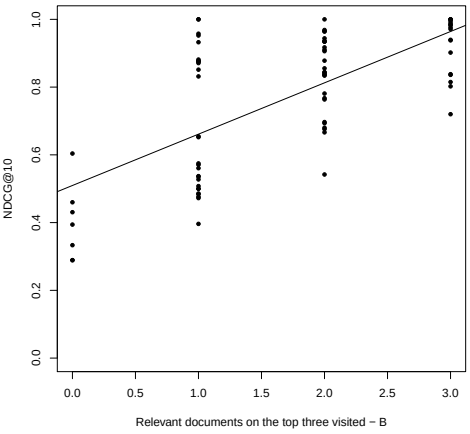
	Baseline	Source combinations				Increment (%)			
Topic	B	P/P	P/G	G/G	Gd/G	$\Delta_{P/P-B}$	$\Delta_{P/G-B}$	$\Delta_{G/G-B}$	$\Delta_{Gd/G-B}$
501	0.715	0.702	**0.854**	0.667	0.710	-1.831	19.431	-6.630	-0.719
502	0.448	0.551	0.613	0.618	**0.616**	23.074	36.812	37.952	37.477
504	**0.946**	0.768	0.725	0.725	0.802	-18.858	-23.441	-23.364	-15.246
506	0.827	0.772	**0.834**	**0.834**	**0.834**	-6.560	0.901	0.901	0.901
509	**0.900**	0.863	0.883	0.883	**0.900**	-4.200	-1.896	-1.896	-0.034
510	0.801	0.829	0.838	0.840	**0.849**	3.579	4.620	4.961	6.058
511	**0.904**	0.844	0.879	0.879	0.897	-6.610	-2.701	-2.785	-0.815
517	0.555	0.770	**0.894**	0.850	0.872	38.665	60.992	53.143	57.045
518	0.875	0.786	0.902	0.902	**0.903**	-10.184	3.107	3.069	3.144
536	0.694	**0.760**	0.661	0.663	0.663	9.424	-4.827	-4.470	-4.470
543	0.494	**0.816**	0.593	0.566	0.592	64.966	19.977	14.542	19.786
544	**0.868**	0.793	0.814	0.813	0.850	-8.626	-6.192	-6.311	-2.036
546	0.764	0.593	0.830	0.834	**0.835**	-22.415	8.623	9.042	9.213
550	**0.918**	0.820	0.696	0.712	0.742	-10.697	-24.159	-22.471	-19.128

(a)

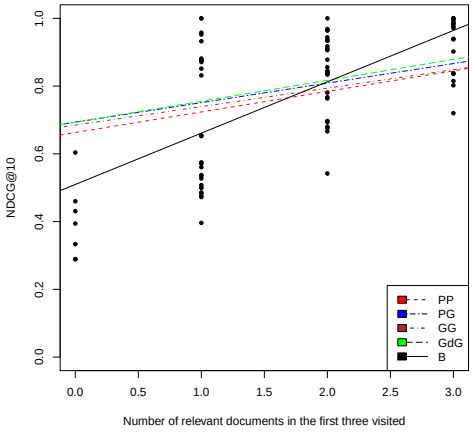
	Baseline	Source combinations				Increment (%)			
User	B	P/P	P/G	G/G	Gd/G	$\Delta_{P/P-B}$	$\Delta_{P/G-B}$	$\Delta_{G/G-B}$	$\Delta_{Gd/G-B}$
user1	0.760	0.774	0.766	0.764	**0.810**	1.823	0.784	0.562	6.520
user2	0.688	0.796	0.844	0.820	**0.885**	15.737	22.662	19.293	28.744
user3	0.726	0.701	0.729	0.728	**0.747**	-3.468	0.445	0.385	2.949
user5	0.798	0.703	0.798	**0.804**	0.803	-11.865	0.054	0.750	0.671
user7	0.775	**0.862**	0.823	0.684	0.699	11.310	6.283	-11.750	-9.700
user8	0.737	**0.760**	0.758	0.759	**0.766**	3.205	2.853	3.029	3.942
user9	0.792	**0.856**	0.759	0.757	0.770	8.069	-4.134	-4.375	-2.724
user10	0.850	0.754	**0.886**	0.836	0.847	-11.229	4.314	-1.603	-0.341
user11	0.799	**0.880**	0.776	0.776	0.776	10.190	-2.825	-2.825	-2.825
user12	0.**866**	0.712	0.756	0.747	0.767	-17.769	-12.681	-13.665	-11.381
user13	0.676	0.813	0.839	0.812	**0.849**	20.237	24.103	20.143	25.552
user15	0.839	0.826	0.820	0.834	**0.849**	-1.651	-2.302	-0.647	1.102
user16	0.670	0.584	0.733	0.737	**0.745**	-12.823	9.284	9.876	11.147

(b)

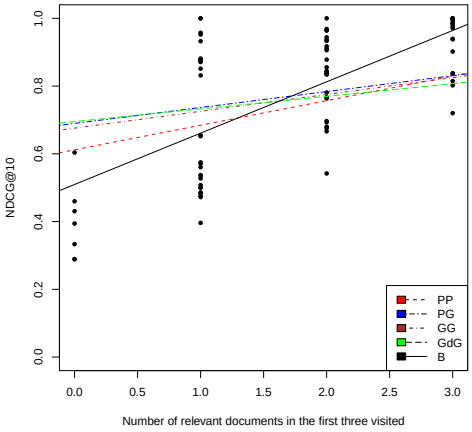
Table 6.15: Mean and Median NDCG@10 per topic and per user for the baseline and the diverse source combinations. The highest value of NDCG@10 for each topic and for each user is marked in bold.



(a) `B` (Baseline)

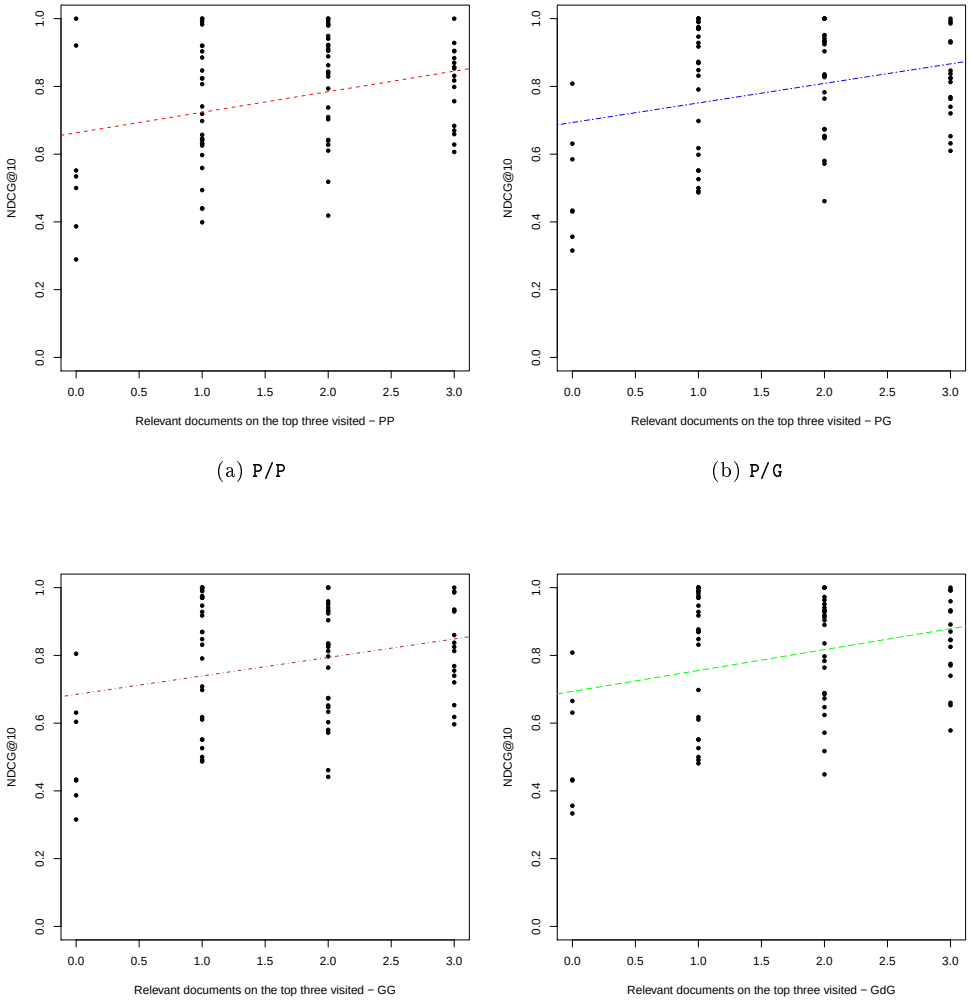


(b) Comparison - Best Eigenvector



(c) Comparison - First Eigenvector

Figure 6.2: NDCG@10's of the baseline (Figure 6.2a) plotted against the number of relevant documents among the top three visited by the users. Comparison among the regression line of the baseline and those of the diverse combinations both in the case when the most effective eigenvector to model the dimension is selected manually (Figure 6.2b) and when the first principal eigenvector is selected (Figure 6.2c). Points depicted in the figures refer to value obtained for the baseline. Values adopted to obtained the regression line of the combinations are reported in Figure 6.3 when the manual selection of the eigenvector is adopted.



(a) `P/P` (b) `P/G`

(c) `G/G` (d) `Gd/G`

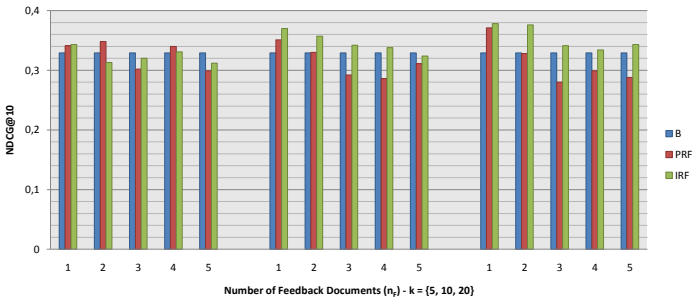
Figure 6.3: NDCG@10 for the diverse source combinations plotted against the number of relevant documents among the top three visited by the users. Values refer to the case when the manual selection of the most effective eigenvector is adopted.

	NDCG@10			Δ (%)		NDCG@20			Δ (%)	
	B	PRF	IRF	$\Delta_{\text{PRF-B}}$	$\Delta_{\text{IRF-B}}$	B	PRF	IRF	$\Delta_{\text{PRF-B}}$	$\Delta_{\text{IRF-B}}$
median	0.324	0.319	0.341	-1.31	5.44	0.295	0.286	0.311	-3.15	5.23
mean	0.329	0.318	0.350	-3.45	6.38	0.295	0.290	0.312	-1.69	5.60
	NDCG@30			Δ (%)		NDCG@50			Δ (%)	
	B	PRF	IRF	$\Delta_{\text{PRF-B}}$	$\Delta_{\text{IRF-B}}$	B	PRF	IRF	$\Delta_{\text{PRF-B}}$	$\Delta_{\text{IRF-B}}$
median	0.249	0.281	0.293	12.86	17.78	0.208	0.207	0.220	-0.48	5.80
mean	0.288	0.285	0.303	-0.95	5.05	0.228	0.217	0.225	-5.10	-1.35

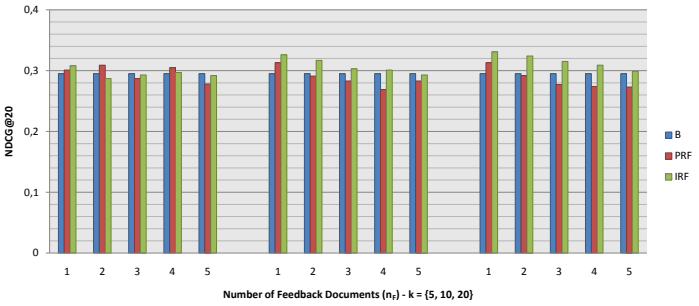
Table 6.16: Mean and Median NDCG@n, with $n \in \{10, 20, 30, 50\}$, computed over all the values of the parameters k and n_F. The number of documents used to model the dimension is $n_B = 3$.

		NDCG@10			NDCG@20			NDCG@30		
k	n_F	B	PRF	IRF	B	PRF	IRF	B	PRF	IRF
5	1	0.329	0.341	0.343	0.295	0.301	0.308	0.288	0.290	0.296
	2	0.329	**0.348**	0.313	0.295	0.309	0.287	0.288	0.296	0.282
	3	0.329	0.302	0.320*	0.295	0.287	0.293	0.288	0.283	0.289
	4	0.329	0.340	0.331	0.295	0.305	0.297	0.288	0.295	0.293
	5	0.329	0.299	0.312	0.295	0.278	0.292*	0.288	0.277	0.284
10	1	0.329	**0.351**	**0.370***	0.295	**0.313**	**0.326**	0.288	0.298	**0.309**
	2	0.329	0.330	**0.357****	0.295	0.291	**0.317***	0.288	0.285	**0.305***
	3	0.329	0.292	0.342**	0.295	0.283	0.303*	0.288	0.280	0.300*
	4	0.329	0.286	0.338*	0.295	0.269	0.301**	0.288	0.273	0.300*
	5	0.329	0.311	0.324	0.295	0.283	0.293	0.288	0.278	0.295*
20	1	0.329	**0.371**	**0.378**	0.295	**0.313**	**0.331***	0.288	**0.305**	**0.319**
	2	0.329	0.328	**0.376****	0.295	0.292	**0.324****	0.288	0.286	**0.311***
	3	0.329	0.280	0.341**	0.295	0.277	**0.315****	0.288	0.278	0.299*
	4	0.329	0.299	0.334**	0.295	0.274	0.309**	0.288	0.280	0.298*
	5	0.329	0.288	0.343*	0.295	0.273	0.299*	0.288	0.277	0.298*

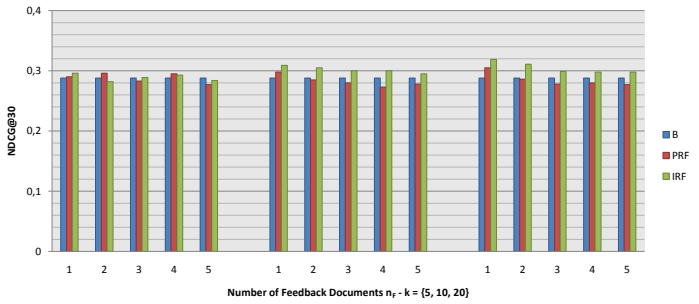
Table 6.17: NDCG@n's for the baseline (B), Pseudo-relevance Feedback (PRF) and Implicit Relevance Feedback (IRF) for different values of n, number of expansion terms, k, and number of document used for query expansion, n_F. The results marked by one star ($\star$) are those for which the increment of IRF respect to PRF in terms of NDCG@n, $\Delta_{\text{IRF}-\text{PRF}}$, is greater than 5%; those marked by two stars ($\star\star$) are those for which $\Delta_{\text{IRF}-\text{PRF}} > 10\%$. Values in bold are those that increased the baseline more than 5%. Moreover, Figure 6.4a-6.4d reports a pictorial description of the same results, considering also NDCG@m, where $m = 50$ is the number of top ranked document re-ranked by the Indri feedback strategy.



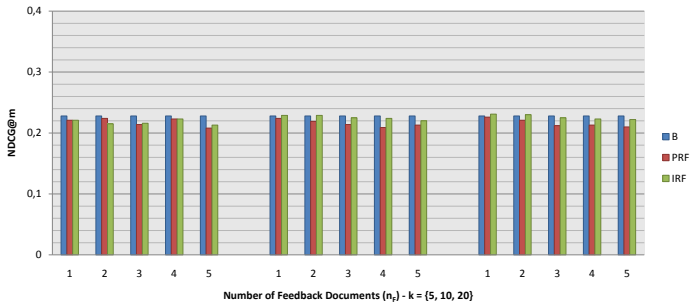
(a)



(b)



(c)



(d)

Figure 6.4: Comparison among the NDCG@n's of the baseline (B), Pseudo Relevance Feedback (PRF) and Implicit Relevance Feedback (IRF) for different values of n, number of expansion terms, k, and number of document used for query expansion, n_F. For each $n \in \{10, 20, 30, 50\}$ three sets of five columns are depicted. Each set corresponds to a specific number of expansion terms k, while each column corresponds to the NDCG@n for specific pair (n_F,k).

	NDCG@10			NDCG@20			NDCG@30		
Topic	PRF	IRF	Δ	PRF	IRF	Δ	PRF	IRF	Δ
501	0.386	0.432	12.02	0.400	0.497	24.22	0.408	0.505	23.82
502	0.300	0.454	51.62	0.219	0.318	45.68	0.199	0.276	38.62
504	0.393	0.333	-15.10	0.316	0.289	-8.52	0.372	0.308	-17.11
506	0.125	0.108	-13.84	0.125	0.108	-13.84	0.125	0.108	-13.84
509	0.518	0.518	0.00	0.478	0.478	0.00	0.466	0.466	0.00
510	0.697	0.697	0.00	0.450	0.450	0.00	0.393	0.393	0.00
511	0.310	0.323	4.36	0.337	0.383	13.57	0.370	0.395	6.62
517	0.140	0.155	10.81	0.154	0.121	-21.40	0.157	0.139	-11.43
518	0.000	0.000	-	0.102	0.070	-31.96	0.101	0.100	-1.19
536	0.355	0.426	20.12	0.269	0.323	20.13	0.291	0.345	18.68
543	0.064	0.078	23.27	0.041	0.051	23.11	0.037	0.045	23.29
544	0.673	0.700	4.10	0.670	0.737	9.97	0.633	0.675	6.67
546	0.169	0.240	42.03	0.188	0.266	41.48	0.190	0.242	27.37
550	0.489	0.539	10.25	0.328	0.348	6.04	0.252	0.267	6.03
all	0.330	0.357	8.40	0.291	0.317	8.86	0.285	0.305	6.80

Table 6.18: NDCG@$\{10, 20, 30\}$'s per topic for Pseudo Relevance Feedback (PRF) and Implicit Relevance Feedback (IRF) when the number of expansion term adopted was $k = 10$ and the number of document for PRF was $n_F = 2$.

	NDCG@10			NDCG@20			NDCG@30		
n_R	PRF	IRF	Δ	PRF	IRF	Δ	PRF	IRF	Δ
0	0.074	0.069	-6.09	0.067	0.045	-33.13	0.075	0.076	1.27
1	0.221	0.252	13.95	0.251	0.288	14.84	0.240	0.256	6.72
2	0.342	0.454	32.71	0.367	0.387	5.28	0.341	0.406	19.00
3	0.514	0.514	0.11	0.475	0.475	-0.09	0.472	0.497	5.24

(a)

	NDCG@10			NDCG@20			NDCG@30		
n_R	PRF	IRF	Δ	PRF	IRF	Δ	PRF	IRF	Δ
0	0.071	0.066	-7.01	0.080	0.063	-20.88	0.083	0.076	-8.14
1	0.259	0.279	7.45	0.229	0.257	12.34	0.223	0.243	9.08
2	0.339	0.489	44.31	0.331	0.418	26.25	0.324	0.384	18.64
3	0.500	0.511	2.09	0.487	0.489	0.36	0.487	0.489	0.54

(b)

Table 6.19: Table 6.19a and Table 6.19b report respectively the median and the mean NDCG@n's for different values of n and different numbers, n_R, of relevant documents among the top three documents of the baseline. $n_F = 3$ documents are provided as input to the Indri Pseudo-Relevance Feedback algorithm. In the event of PRF, n_R is the number of relevant documents among those used for feedback. In the event of IRF, n_R is the number of relevant documents among those used for modeling the user behavior dimension.

	Median			Δ (%)		Mean			Δ (%)	
n_R	B	PRF	IRF	$\Delta_{\text{PRF-B}}$	$\Delta_{\text{IRF-B}}$	B	PRF	IRF	$\Delta_{\text{PRF-B}}$	$\Delta_{\text{IRF-B}}$
0	0.069	0.074	0.069	7.25	0.00	0.081	0.071	0.066	-12.35	-18.52
1	0.166	0.221	0.252	33.13	51.81	0.184	0.259	0.279	40.76	51.63
2	0.437	0.342	0.454	-21.74	3.89	0.449	0.339	0.489	-24.50	8.91
3	0.625	0.514	0.514	-17.76	-17.76	0.610	0.500	0.511	-18.03	-16.23

(a) NDCG@10

	Median			Δ (%)		Mean			Δ (%)	
n_R	B	PRF	IRF	$\Delta_{\text{PRF-B}}$	$\Delta_{\text{IRF-B}}$	B	PRF	IRF	$\Delta_{\text{PRF-B}}$	$\Delta_{\text{IRF-B}}$
0	0.065	0.08	0.063	23.08	-3.08	0.045	0.067	0.045	48.89	0.00
1	0.193	0.229	0.257	18.65	33.16	0.187	0.251	0.288	34.22	54.01
2	0.376	0.331	0.418	-11.97	11.17	0.382	0.367	0.387	-3.93	1.31
3	0.555	0.487	0.489	-12.25	-11.89	0.596	0.475	0.475	-20.30	-20.30

(b) NDCG@20

	Median			Δ (%)		Mean			Δ (%)	
n_R	B	PRF	IRF	$\Delta_{\text{PRF-B}}$	$\Delta_{\text{IRF-B}}$	B	PRF	IRF	$\Delta_{\text{PRF-B}}$	$\Delta_{\text{IRF-B}}$
0	0.040	0.075	0.076	87.50	90.00	0.06	0.083	0.076	38.33	26.67
1	0.162	0.240	0.256	48.15	58.02	0.176	0.223	0.243	26.70	38.07
2	0.381	0.341	0.406	-10.50	6.56	0.364	0.324	0.384	-10.99	5.49
3	0.591	0.472	0.497	-20.14	-15.91	0.564	0.487	0.489	-13.65	-13.30

(c) NDCG@30

Table 6.20: Median and mean NDCG@n for different numbers n_R of relevant documents among the top three documents of the baseline, when considering $n_F = 3$. Results are reported for different values of n for the baseline (B), the Indri Pseudo-Relevance Feedback algorithm applied to the top $n_F = 3$ returned by the baseline (PRF) or the top $n_F = 3$ re-ranked by user behavior dimension (IRF).

CHAPTER

SEVEN

CONCLUSION

7.1 Conclusion

The thesis addresses the problem of uniformly modeling heterogeneous forms of user interaction that are selected as sources for feedback. The problem of uniform source modeling is addressed by way of a complete methodology. The methodology aims at designing, implementing and evaluating a system that validates an experimental hypothesis. The hypothesis being validated regards the possible factors that can explain the user perception of relevance through the evidence gathered from the user interaction. The objective is to obtain and exploit a usable representation of the factors that can explain the user perception of relevance in the role of a new dimension of the information need representation.

The definition of the methodology is supported by an abstraction of the IR problem. We show that the abstraction of the IR problem can cope with heterogeneous informative resources in terms of the organization and media involved. Two IR systems named SPINA (Superimposed Peer-to-peer INformation Access) and FALCON (FAst Lucene-based Cover sOng identificatioN) have been developed to address, respectively, the problem of Distributed IR and music identification by using the abstraction considered in the thesis.

The methodology aims at being general and not tailored to a specific source or purpose. The methodology defines the set of steps needed for obtaining a vector subspace-based representation of the information need dimensions to further exploit this representation for relevance prediction purposes. The set of steps identified are source selection, evidence collection, dimension modeling, document modeling and prediction.

We showed how the methodology can be used for modeling two sources of evidence:

term relationship in documents judged as relevant and the relationship between interaction features gathered from the behavior of the user when interacting with a set of documents.

As for the term relationship dimension, we showed that the current implementation of term relationship is feasible with a very large text collection delivered within the 2009 and 2010 Relevance Feedback tracks of the Text Retrieval Conference (TREC) initiative. The methodology supported the evaluation of term relationship for document re-ranking. We showed that a document representation based on normalized term frequency is more effective than a document representation based on term relationships estimated without considering the term frequency within a document. The propagation of the weights to the entries along the diagonal of the term correlation matrix (also known as term self-correlation entries) and corresponding to terms in the original query can provide a slight improvement, yet it is still not significantly better than the baseline. Overall, the results are inconclusive, thus suggesting the need to investigate diverse implementations of term relationship and of document modeling.

As for interaction feature relationships, we investigated the adoption of the user behavior dimension for document re-ranking both without query expansion and with query expansion. To this end, we investigated whether the use of interaction features gathered from a group of users searching the same collection for information relevant to the same topic can surrogate interaction features gathered at an individual user basis in the event the latter are not available.

We showed that, with the dataset developed for the purposes of our study, using group data can negatively affect re-ranking when effectiveness is measured on a per user basis (i.e. using gains provided by individual users). Even if re-ranking based on the user behavior dimension did not provide significant improvement with respect to the baseline, its effectiveness was observed to be not strictly dependent on the number of relevant documents adopted for dimension modeling.

We investigated whether group based behavior could allow to bring at high rank position good sources for query expansion, even though it is less effective than individual behavior and a small number of relevant documents are present among those used for dimension modeling.

The results show that the top-ranked results after re-ranking by group behavior are comparable with the highest ranking results in the baseline list when used for query expansion. Provided the former is an instance of Implicit Relevance Feedback (IRF) and the latter is an instance of Pseudo-Relevance Feedback (PRF), in some cases IRF was able to provide an improvement although PRF was less effective than the baseline, thus suggesting the need to investigate combinations both of content and user behavior as evidence to support query expansion at a larger scale than the study of this thesis

to see whether they can effectively complement each other or whether they instead tend to cancel each other out.

Our results suggest that the correlation between interaction features is not sufficient to obtain an effective information need representation. One cause is perhaps the features used as evidence for dimension and document modeling. It may be that the use of retention features, when available, could improve the effectiveness of the models since they have been shown to be good implicit indicators of user interest.

Alternative implementations of the methodology steps should be investigated for both dimension and document modeling because the methodology applications investigated in the thesis are not robust enough to support document re-ranking. However, the methodology provides a principled approach for performing this investigation in a single framework and for evaluating the effectiveness of alternative implementations. Indeed the modularity of the methodology allows us to formulate the problem of uniformly modeling heterogeneous sources in subproblems, thus avoiding the re-design of the entire methodology or the entire IR approach which is rather necessary in the event of heuristic approaches developed ad-hoc.

The modularity of the methodology has allowed us to extend SPINA with additional modules that provide the functionalities for performing document re-ranking on the basis of term relationship and user interaction. Hence, another outcome of the thesis is the experimental system which is able to perform re-ranking on the basis of both term-relationship in the feedback documents and user behavior when interacting with a set of results.

Another contribution of the thesis is the test dataset used to support the evaluation of the methodology implementation for the user behavior dimension. The creation of a test collection was necessary because no test collection with interaction data is publicly available and appropriate for the evaluation of a methodology application. For instance, the dataset obtained in [Claypool et al., 2001] provides both interaction features and explicit judgments, but the dataset was gathered from generic browsing activity, not in the context of a specific search task. Therefore, we have carried out a user study for gathering the interaction data necessary for our investigation. Besides supporting the investigation reported in this thesis, this test collection has been adopted to perform some preliminary investigation of the dependency between time intervals and user relevance assessments [Di Buccio et al., 2011].

7.2 Future Work

There are a number of directions for the future work suggested by the research presented in this thesis. They are concerned with other methodology implementations,

the evaluation by considering other variables or additional sources of evidence. In the remainder of this section some of these direction will be discussed.

Relationship among Sources

Going back to the IR problem, the two crucial questions Robertson identified are [Robertson, 1977]:

- On the basis of what kinds of information can the system make the prediction?
- How should the system utilize and combine these various kinds of information?

As for the first question, the methodology introduced in the thesis allows the diverse sources of evidence involved during the search process, e.g. diverse forms of interaction, to be uniformly modeled. In particular, the proposed methodology allows us to investigate the user perception of relevance.

As for the second question, the main point is the way to combine diverse sources. A solution is the combination of the scores. If appropriate representations are considered for dimensions and informative resources, the measure provided by the projector-based function can be interpreted as a probability – more specifically according to the Quantum Mechanics (QM) interpretation of probability. Had the QM interpretation been accepted, we could combine probabilities in a principled way. A solution is to compute a new subspace representation obtained from the two distinct dimension subspaces thanks to the uniform vector subspace-based model of the sources.

The are a number of questions when investigating approaches for computing a subspace representation to exploit several sources simultaneously. A question is about the investigation of the relationships between source contributions and the effect of these relationships on retrieval effectiveness. Do we need to remove the contribution of the relationship, thus obtaining uncorrelated sources? Being this relationship in the overlap among source contributions, can this relationship be used as a new source?

Previous works investigated possible measures of distance among semantic subspaces [Zuccon et al., 2009]. Let us consider the term relationship dimension obtained from a document explicitly judged as relevant and from top ranked documents, i.e. pseudo feedback. These sources can be characterized on the basis of homogeneous features. A possible approach for combination is to obtain a mixture of dimensions on the basis of the two distinct models obtained by the two distinct sources. A possible question here is if this mixture can be automatically tuned on the basis of the distance between the subspaces, e.g. on the basis on the relationship among the sources.

User Behavior-based Representation

A possible extension to the work reported in the thesis is to investigate additional variables and features for dimension modeling within more naturalistic settings of the user study as done in [Kelly, 2004]. Exploiting additional interaction features and variables could result in more effective models. For instance, previous works have shown that retention features can be useful indicators of user interests [Fox et al., 2005]. Additional features can be gathered through a new user study. Recently, tools have been made available that offer support to capture some of these features. An example is the Lemur Query Log Toolbar that besides search results page information captures some user interaction data; its functionalities can be extended to gather other features, e.g. retention features.

As for the variables, task information is not considered in the thesis, but it could be another criterion for aggregating interaction data, e.g. as done for a single interaction feature in [White and Kelly, 2006]. Grouping criteria are indeed one of the possible research directions worth investigating. In [Teevan et al., 2009] the authors explicitly investigated the impact of diverse grouping criteria to aggregate scores obtained by personalization algorithms mainly based on content-based features. A research question is to investigate how the grouping criteria affect the effectiveness of the methodology application discussed in this thesis. Another question is to investigate if we can exploit user behavior model as a grouping criterion. Each group and each user could be uniformly modeled as a subspace on the basis of specific dimensions, e.g. their behavior. The effectiveness of subspace distances for capturing relationship between groups of users or individual users and groups can be investigated. These representations suggest that the framework would allow us to rank not only documents but also users.

The main reason for investigating user behavior as a possible source for evidence is to obtain a representation of the information need able to capture different information from that based on content based features. A document model could be obtained from a set of observations concerning diverse interaction features gathered from diverse users when examining the document. In the current implementation of the methodology a document is a vector that represents an observation for that document with regard to a query and an individual user, or can be a vector of average feature values when considered with regard to a group of users. But a document can be a subspace. A subspace representation of document has already been proposed in the literature [Piwowarski et al., 2010b, Piwowarski et al., 2010a] but it exploits only content-based features, possibly ratings [Frommholz et al., 2010], but not interaction features.

A further opportunity to evaluate the methodology adopted in this thesis is participation in the TREC 2011 Session Track, the objective of which is to evaluate the

predictive capability of an IR approach "over a sequence of queries and user interactions, rather than for a single "one-shot" query"[1]. Considering multiple queries in a session will make additional content-based and user interaction features available. Participation in the track will allow us to perform the evaluation using a common protocol, as done in the thesis as for the term relationship dimension, thus comparing our approach with those of the other participants. Exploring the methodology for a session can yield new challenges on how to model dimension and eventually combine different dimensions across a session, e.g., explicitly considering time.

Methodology Applications to other Sources of Evidence

The methodology aims at being general, not tailored to a specific source. During the search process, besides the user and the document, other units are involved and their properties can be exploited to support prediction. A challenging direction is to investigate methodology applications for other units. One unit of interest is the task. The task could be useful for several reason. Learning display time thresholds using task information was shown to be successful [White and Kelly, 2006] and the task was shown to affect the predicting capability of interaction features [Kelly and Belkin, 2004]. Explicitly modeling and exploiting task properties, e.g. task type, could be beneficial for obtaining a better characterization of the user information need. Therefore, a further research question is to investigate if a vector subspace representation of the task could be obtained from the available evidence and its contribution in terms of retrieval effectiveness.

[1] http://trec.nist.gov/pubs/call2011.html

BIBLIOGRAPHY

[Agichtein et al., 2006a] Agichtein, E., Brill, E., and Dumais, S. (2006a). Improving web search ranking by incorporating user behavior information. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–26, New York, NY, USA. ACM.

[Agichtein et al., 2006b] Agichtein, E., Brill, E., Dumais, S., and Ragno, R. (2006b). Learning user interaction models for predicting web search result preferences. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–10, New York, NY, USA. ACM.

[Agosti et al., 2010] Agosti, M., Cisco, D., Di Nunzio, G. M., Masiero, I., and Melucci, M. (2010). i-TEL-u: a query suggestion tool for integrating heterogeneous contexts in a digital library. In *Proceedings of the 14th European conference on Research and advanced technology for digital libraries*, ECDL'10, pages 397–400, Berlin, Heidelberg. Springer-Verlag.

[Allan et al., 2009] Allan, J., Carterette, B., Aslam, J. A., Pavlu, V., Dachev, B., and Kanoulas, E. (2009). Million query track 2007 overview. In Ellen M. Voorhees, L. P. B., editor, *Proceedings of The Sixteenth Text REtrieval Conference (TREC 2007)*, NIST Special Publication: SP 500-278, Gaithersburg, Maryland, USA.

[Almeida and Almeida, 2004] Almeida, R. B. and Almeida, V. A. F. (2004). A community-aware search engine. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 413–421, New York, NY, USA. ACM.

[Aslam et al., 2006] Aslam, J. A., Pavlu, V., and Yilmaz, E. (2006). A statistical method for system evaluation using incomplete judgments. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 541–548, New York, NY, USA. ACM.

[Bai and Nie, 2008] Bai, J. and Nie, J.-Y. (2008). Adapting information retrieval to query contexts. *Information Processing and Management*, 44(6):1901–1922.

[Bai et al., 2005] Bai, J., Song, D., Bruza, P., Nie, J.-Y., and Cao, G. (2005). Query expansion using term relationships in language models for information retrieval. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, pages 688–695, New York, NY, USA. ACM.

[Bates, 1989] Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5):407–424.

[Belkin et al., 1982] Belkin, N. J., Oddy, R. N., and Brooks, H. M. (1982). Ask for information retrieval: Part 1. background and theory. *Journal of Documentation*, 38(2):61–71.

[Bilenko and White, 2008] Bilenko, M. and White, R. W. (2008). Mining the search trails of surfing crowds: identifying relevant websites from user activity. In *Proceeding of the 17th international conference on World Wide Web*, WWW '08, pages 51–60, New York, NY, USA. ACM.

[Bishop, 2006] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

[Bodoff, 2004] Bodoff, D. (2004). Relevance models to help estimate document and query parameters. *ACM Transactions on Information Systems (TOIS)*, 22(3):357–380.

[Bodoff and Robertson, 2004] Bodoff, D. and Robertson, S. E. (2004). A new unified probabilistic model. *Journal of the American Society for Information Science and Technology*, 55(6):471–487.

[Broder, 2002] Broder, A. (2002). A taxonomy of web search. *SIGIR Forum*, 36(2):3–10.

[Buckley et al., 1994] Buckley, C., Salton, G., and Allan, J. (1994). The effect of adding relevance information in a relevance feedback environment. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 292–300, New York, NY, USA. Springer-Verlag New York, Inc.

[Burges et al., 2005] Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., and Hullender, G. (2005). Learning to rank using gradient descent. In

ICML '05: Proceedings of the 22nd international conference on Machine learning, pages 89–96, New York, NY, USA. ACM.

[Callan, 2002] Callan, J. (2002). Distributed information retrieval. In Croft, W. B., editor, *Advances in Information Retrieval*, volume 7 of *The Information Retrieval Series*, pages 127–150. Springer US.

[Callan et al., 1995] Callan, J. P., Lu, Z., and Croft, W. B. (1995). Searching distributed collections with inference networks. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '95, pages 21–28, New York, NY, USA. ACM.

[Campbell and van Rijsbergen, 1996] Campbell, I. and van Rijsbergen, C. J. (1996). The ostensive model of developing information needs. In *Proceedings of the 3rd international conference on conceptions of library and information science*, pages 251–268.

[Cao et al., 2008] Cao, G., Nie, J.-Y., Gao, J., and Robertson, S. (2008). Selecting good expansion terms for pseudo-relevance feedback. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 243–250, New York, NY, USA. ACM.

[Carterette et al., 2009] Carterette, B., Pavlu, V., Fang, H., and Kanoulas, E. (2009). Million query track 2009 overview. In Ellen M. Voorhees, L. P. B., editor, *Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009)*, NIST Special Publication: SP 500-278, Gaithersburg, Maryland, USA.

[Cartright et al., 2009] Cartright, M.-A., Seo, J., and Lease, M. (2009). UMass Amherst and UT Austin @ The TREC 2009 Relevance Feedback Track. In Ellen M. Voorhees, L. P. B., editor, *Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009)*, NIST Special Publication: SP 500-278, Gaithersburg, Maryland, USA.

[Clarke et al., 2007] Clarke, C. L. A., Agichtein, E., Dumais, S., and White, R. W. (2007). The influence of caption features on clickthrough patterns in web search. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 135–142, New York, NY, USA. ACM.

[Clarke et al., 2009] Clarke, C. L. A., Craswell, N., and Soboroff, I. (2009). Overview of the trec 2009 web track. In Ellen M. Voorhees, L. P. B., editor, *Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009)*, NIST Special Publication: SP 500-278, Gaithersburg, Maryland, USA.

[Claypool et al., 2001] Claypool, M., Le, P., Wased, M., and Brown, D. (2001). Implicit interest indicators. In *IUI '01: Proceedings of the 6th international conference on Intelligent user interfaces*, pages 33–40, New York, NY, USA. ACM.

[Collins-thompson et al., 2005] Collins-thompson, K., Ogilvie, P., and Callan, J. (2005). Initial results with structured queries and laguage models on half a terabyte of text. text retrieval conference. In Ellen M. Voorhees, L. P. B., editor, *Proceedings of The Fourteenth Text REtrieval Conference (TREC 2005)*, NIST Special Publication: SP 500-266, Gaithersburg, Maryland, USA.

[Cooper, 1991] Cooper, W. S. (1991). Some inconsistencies and misnomers in probabilistic information retrieval. In *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '91, pages 57–61, New York, NY, USA. ACM.

[Cooper and Maron, 1978] Cooper, W. S. and Maron, M. E. (1978). Foundations of probabilistic and utility-theoretic indexing. *Journal of the ACM*, 25:67–80.

[Cormack et al., 2010] Cormack, G. V., Smucker, M. D., and Clarke, C. L. A. (2010). Efficient and effective spam filtering and re-ranking for large web datasets. *CoRR*, abs/1004.5168.

[Croft, 1999] Croft, W. B. (1999). *Advances in Information Retrieval*, chapter Combining Approaches to Information Retrieval, pages 1–36. Springer.

[Croft et al., 2009] Croft, W. B., Metzler, D., and Strohman, T. (2009). *Search Engines: Information Retrieval in Practice*. Addison Wesley.

[Cutrell and Guan, 2007] Cutrell, E. and Guan, Z. (2007). What are you looking for?: an eye-tracking study of information usage in web search. pages 407–416.

[Deerwester et al., 1990] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407.

[Di Buccio et al., 2008] Di Buccio, E., Ferro, N., and Melucci, M. (2008). Content-based information retrieval in spina. In *IRCDL: Post-proceedings of the Forth Italian Research Conference on Digital Library Systems*, pages 89–92, Padova, Italy.

[Di Buccio et al., 2009a] Di Buccio, E., Ferro, N., Melucci, M., Miotto, R., and Orio, N. (2009a). Design of a music retrieval system based on a peer-to-peer paradigm. In *IRCDL: Post-proceedings of the Fifth Italian Research Conference on Digital Libraries*, Padova, Italy.

[Di Buccio et al., 2009b] Di Buccio, E., Masiero, I., Mass, Y., Melucci, M., Miotto, R., Orio, N., and Sznajder, B. (2009b). Towards an integrated approach to music retrieval. In *IRCDL: Post-proceedings of the Fifth Italian Research Conference on Digital Libraries*, Padova, Italy.

[Di Buccio et al., 2009c] Di Buccio, E., Masiero, I., and Melucci, M. (2009c). Improving information retrieval effectiveness in peer-to-peer networks through query piggybacking. In *ECDL: Research and Advanced Technology for Digital Libraries, 13th European Conference, ECDL 2009*, pages 420–424, Corfu, Greece.

[Di Buccio and Melucci, 2009a] Di Buccio, E. and Melucci, M. (2009a). Exploiting individual users and user groups interaction features: methodology and infrastructure design. In *Proceedings of the Second Workshop on Very Large Digital Libraries*, Corfu, Greece.

[Di Buccio and Melucci, 2009b] Di Buccio, E. and Melucci, M. (2009b). University of Padua at TREC 2009: Relevance Feedback Track. In Ellen M. Voorhees, L. P. B., editor, *Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009)*, NIST Special Publication: SP 500-278, Gaithersburg, Maryland, USA.

[Di Buccio et al., 2011] Di Buccio, E., Melucci, M., and Song, D. (2011). Towards Predicting Relevance Using a Quantum-Like Framework. In *Proceedings of the 33rd European Conference on Information Retrieval*, Dublin, Ireland. To appear.

[Di Buccio et al., 2010a] Di Buccio, E., Montecchio, N., and Orio, N. (2010a). FALCON: FAst Lucene-based Cover sOng identification. In Bimbo, A. D., Chang, S.-F., and Smeulders, A. W. M., editors, *ACM Multimedia: Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010*, pages 1477–1480. ACM.

[Di Buccio et al., 2010b] Di Buccio, E., Montecchio, N., and Orio, N. (2010b). A scalable cover identification engine. In Bimbo, A. D., Chang, S.-F., and Smeulders, A. W. M., editors, *ACM Multimedia: Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010*, pages 1143–1146. ACM.

[Dou et al., 2007] Dou, Z., Song, R., and Wen, J.-R. (2007). A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 581–590, New York, NY, USA. ACM.

[Dragunov et al., 2005] Dragunov, A. N., Dietterich, T. G., Johnsrude, K., McLaughlin, M., Li, L., and Herlocker, J. L. (2005). Tasktracer: a desktop environment to

support multi-tasking knowledge workers. In *Proceedings of the 10th international conference on Intelligent user interfaces*, IUI '05, pages 75–82, New York, NY, USA. ACM.

[Dumais, 1991] Dumais, S. T. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers*, 23(2):229–236.

[Dumais, 1996] Dumais, S. T. (1996). Combining evidence for effective information filtering. *AAAI Spring Symposium on Machine Learning and Information Retrieval, Tech Report SS-96-07*.

[Efron, 2008] Efron, M. (2008). Query expansion and dimensionality reduction: Notions of optimality in rocchio relevance feedback and latent semantic indexing. *Information Processing and Management: an International Journal*, 44(1):163–180.

[Fox et al., 2005] Fox, S., Karnawat, K., Mydland, M., Dumais, S., and White, T. (2005). Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems*, 23(2):147–168.

[Frommholz et al., 2010] Frommholz, I., Larsen, B., Piwowarski, B., Lalmas, M., Ingwersen, P., and van Rijsbergen, K. (2010). Supporting polyrepresentation in a quantum-inspired geometrical retrieval framework. In *Proceeding of the third symposium on Information interaction in context*, IIiX '10, pages 115–124, New York, NY, USA. ACM.

[Gao et al., 2004] Gao, J., Nie, J.-Y., Wu, G., and Cao, G. (2004). Dependence language model for information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 170–177, New York, NY, USA. ACM.

[Göker and Myrhaug, 2008] Göker, A. and Myrhaug, H. (2008). Evaluation of a mobile information system in context. *Information Processing and Management*, 44(1):39–65.

[Harman, 1992] Harman, D. (1992). Relevance feedback revisited. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '92, pages 1–10, New York, NY, USA. ACM.

[Hatcher and Gospodnetic, 2004] Hatcher, E. and Gospodnetic, O. (2004). *Lucene in Action (In Action series)*. Manning Publications Co., Greenwich, CT, USA.

[Hull, 1994] Hull, D. (1994). Improving text retrieval for the routing problem using latent semantic indexing. In *Proceedings of the 17th annual international ACM*

SIGIR conference on Research and development in information retrieval, SIGIR '94, pages 282–291, New York, NY, USA. Springer-Verlag New York, Inc.

[Ingwersen, 1996] Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: elements of a cognitive ir theory. *Journal of Documentation*, 52:3–50.

[Järvelin and Kekäläinen, 2002] Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

[Joachims et al., 2005] Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 154–161, New York, NY, USA. ACM.

[Joachims et al., 2007] Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., and Gay, G. (2007). Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems*, 25(2):7.

[Kelly and Belkin, 2001] Kelly, D. and Belkin, N. J. (2001). Reading ti scrolling and interaction: exploring implicit sources of user preferences for relevance feedback. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 408–409, New York, NY, USA. ACM.

[Kelly and Belkin, 2004] Kelly, D. and Belkin, N. J. (2004). Display time as implicit feedback: understanding task effects. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 377–384, New York, NY, USA. ACM.

[Kelly and Teevan, 2003] Kelly, D. and Teevan, J. (2003). Implicit feedback for inferring user preference: a bibliography. *SIGIR Forum*, 37(2):18–28.

[Kelly, 2004] Kelly, J. D. (2004). *Understanding implicit feedback and document preference: a naturalistic user study*. PhD thesis, New Brunswick, NJ, USA.

[Kleinberg, 1999] Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46:604–632.

[Konstan et al., 1997] Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., and Riedl, J. (1997). Grouplens: applying collaborative filtering to usenet news. *Commun. ACM*, 40:77–87.

[Kontostathis and Pottenger, 2006] Kontostathis, A. and Pottenger, W. M. (2006). A framework for understanding latent semantic indexing (lsi) performance. *Information Processing and Management*, 42(1):56–73.

[Lavrenko and Croft, 2001] Lavrenko, V. and Croft, W. B. (2001). Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 120–127, New York, NY, USA. ACM.

[Lease, 2008] Lease, M. (2008). Incorporating relevance and pseudo-relevance feedback in the markov random field model. In Ellen M. Voorhees, L. P. B., editor, *Proceedings of the Seventeenth Text REtrieval Conference (TREC 2008)*, NIST Special Publication: SP 500-277, Gaithersburg, Maryland, USA.

[Li et al., 2009a] Li, S., Li, X., Zhang, H., Gao, S., Chen, G., and Guo, J. (2009a). PRIS at 2009 Relevance Feedback track: Experiments in Language Model for Relevance Feedback. In Ellen M. Voorhees, L. P. B., editor, *Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009)*, NIST Special Publication: SP 500-278, Gaithersburg, Maryland, USA.

[Li et al., 2009b] Li, Y., Tao, X., Algarni, A., and Wu, S.-T. (2009b). Mining Specific and General Features in Both Positive and Negative Relevance Feedback - QUT E-Discovery Lab at the TREC'09 Relevance Feedback Track. In Ellen M. Voorhees, L. P. B., editor, *Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009)*, NIST Special Publication: SP 500-278, Gaithersburg, Maryland, USA.

[Lin et al., 2009] Lin, J., Metzler, D., Elsayed, T., , and Wang, L. (2009). Of Ivory and Smurfs: Loxodontan MapReduce Experiments for Web Search. In Ellen M. Voorhees, L. P. B., editor, *Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009)*, NIST Special Publication: SP 500-278, Gaithersburg, Maryland, USA.

[Lu and Callan, 2006] Lu, J. and Callan, J. (2006). Full-text federated search of text-based digital libraries in peer-to-peer networks. *Information Retrieval*, 9:477–498.

[Lund and Burgess, 1996] Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers*, 28(2):203–208.

[Meij et al., 2009] Meij, E., He, J., Weerkamp, W., and de Rijke, M. (2009). Topical Diversity and Relevance Feedback. In Ellen M. Voorhees, L. P. B., editor, *Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009)*, NIST Special Publication: SP 500-278, Gaithersburg, Maryland, USA.

[Melucci, 2005] Melucci, M. (2005). Context modeling and discovery using vector space bases. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 808–815, New York, NY, USA. ACM.

[Melucci, 2008] Melucci, M. (2008). A basis for information retrieval in context. *ACM Transaction on Information Systems*, 26(3):1–41.

[Melucci and Poggiani, 2007] Melucci, M. and Poggiani, A. (2007). A Study of a Weighting Scheme for Information Retrieval in Hierarchical Peer-to-Peer Networks. In Amati, G., Carpineto, C., and Romano, G., editors, *ECIR: Advances in Information Retrieval, 29th European Conference on IR Research, ECIR 2007, Rome, Italy, April 2-5, 2007, Proceedings*, volume 4425 of *Lecture Notes in Computer Science*, pages 136–147. Springer.

[Melucci and White, 2007a] Melucci, M. and White, R. W. (2007a). Discovering hidden contextual factors for implicit feedback. In *CIR: Proceedings of the CIR'07 Workshop on Context-Based Information Retrieval*, Roskilde, Denmark.

[Melucci and White, 2007b] Melucci, M. and White, R. W. (2007b). Utilizing a geometry of context for enhanced implicit feedback. In *Proceedings of the ACM Sixteenth Conference on Information and Knowledge Management (CIKM 2007)*.

[Metzler and Croft, 2004] Metzler, D. and Croft, W. B. (2004). Combining the language model and inference network approaches to retrieval. *Information Processing and Management: an International Journal*, 40:735–750.

[Metzler and Croft, 2005] Metzler, D. and Croft, W. B. (2005). A markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 472–479, New York, NY, USA. ACM.

[Metzler et al., 2004] Metzler, D., Lavrenko, V., and Croft, W. B. (2004). Formal multiple-bernoulli models for language modeling. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 540–541, New York, NY, USA. ACM.

[Metzler et al., 2005] Metzler, D., Strohman, T., and Yun Zhou, W. B. C. (2005). Indri at TREC 2005: Terabyte Track. In Ellen M. Voorhees, L. P. B., editor, *Proceedings of The Fourteenth Text REtrieval Conference (TREC 2005)*, NIST Special Publication: SP 500-266, Gaithersburg, Maryland, USA.

[Meyer, 2000] Meyer, C. D., editor (2000). *Matrix analysis and applied linear algebra*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

[Miotto and Orio, 2008] Miotto, R. and Orio, N. (2008). A music identification system based on chroma indexing and statistical modeling. In Bello, J. P., Chew, E., and Turnbull, D., editors, *ISMIR: Proceedings of the 9th International Conference on Music Information Retrieval*, pages 301–306, Drexel University, Philadelphia, PA, USA.

[Morita and Shinoda, 1994] Morita, M. and Shinoda, Y. (1994). Information filtering based on user behavior analysis and best match text retrieval. In *Proceedings of SIGIR '94*, pages 272–281, New York, NY, USA. Springer-Verlag New York, Inc.

[Nallapati and Allan, 2002] Nallapati, R. and Allan, J. (2002). Capturing term dependencies using a language model based on sentence trees. In *Proceedings of the eleventh international conference on Information and knowledge management*, CIKM '02, pages 383–390, New York, NY, USA. ACM.

[Neve and Orio, 2004] Neve, G. and Orio, N. (2004). Indexing and retrieval of music documents through pattern analysis and data fusion techniques. In *ISMIR: Proceedings of the 5th International Conference on Music Information Retrieval*, Barcelona, Spain.

[Nichols, 1997] Nichols, D. M. (1997). Implicit rating and filtering. In *Proceedings of 5th DELOS Workshop on Filtering and Collaborative Filtering*, pages 31–36. Citeseer.

[Nottelmann and Fuhr, 2006] Nottelmann, H. and Fuhr, N. (2006). Comparing different architectures for query routing in peer-to-peer networks. In Lalmas, M., MacFarlane, A., Rüger, S. M., Tombros, A., Tsikrika, T., and Yavlinsky, A., editors, *ECIR: Advances in Information Retrieval, 28th European Conference on IR Research, ECIR 2006, London, UK, April 10-12, 2006, Proceedings*, volume 3936 of *Lecture Notes in Computer Science*, pages 253–264. Springer.

[Oard and Kim, 2001] Oard, D. and Kim, J. (2001). Modeling information content using observable behavior. In *Proceedings of the Annual meeting of the American Society for Information Science*, volume 38, pages 481–488. Citeseer.

[Pearson, 1901] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572.

[Pérez-Iglesias et al., 2009] Pérez-Iglesias, J., Pérez-Agüera, J. R., Fresno, V., and Feinstein, Y. Z. (2009). Integrating the Probabilistic Models BM25/BM25F into Lucene. *CoRR*, abs/0911.5046.

[Piwowarski et al., 2010a] Piwowarski, B., Frommholz, I., Lalmas, M., and van Rijsbergen, K. (2010a). What can quantum theory bring to IR? In Huang, J., Koudas, N., Jones, G., Wu, X., Collins-Thompson, K., and An, A., editors, *CIKM'10: Proceedings of the nineteenth ACM conference on Conference on information and knowledge management*. ACM.

[Piwowarski et al., 2010b] Piwowarski, B., Frommholz, I., Moshfeghi, Y., Lalmas, M., and van Rijsbergen, K. (2010b). Filtering documents with subspaces. In Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S., and van Rijsbergen, K., editors, *Advances in Information Retrieval*, volume 5993 of *Lecture Notes in Computer Science*. Springer.

[Ponte and Croft, 1998] Ponte, J. M. and Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 275–281, New York, NY, USA. ACM.

[Qi Guo, 2010] Qi Guo, Ryen W. White, S. T. D. J. W. B. A. (2010). Predicting query performance using query, result, and user interaction features. In *Proceedings of the 9th International Conference on Adaptivity, Personalization and Fusion of Heterogeneous Information (RIAO 2010)*, Paris, France.

[Rafter and Smyth, 2001] Rafter, R. and Smyth, B. (2001). Passive profiling from server logs in an online recruitment environment. In *Proceedings of the IJCAI Workshop on Intelligent Techniques for Web Personalization (ITWP 2001)*, pages 35–41.

[Richardson et al., 2006] Richardson, M., Prakash, A., and Brill, E. (2006). Beyond pagerank: machine learning for static ranking. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 707–715, New York, NY, USA. ACM.

[Robertson, 1977] Robertson, S. E. (1977). The Probability Ranking Principle in IR. *Journal of Documentation*, 33(4):294–304.

[Robertson, 1990] Robertson, S. E. (1990). On term selection for query expansion. *Journal of Documentation*, 46(4):359–364.

[Robertson et al., 1982] Robertson, S. E., Maron, M. E., and Cooper, W. S. (1982). Probability of relevance: A unification of two competing models for document retrieval. *Information Technology: Research and Development*, 1(1):1–21.

[Robertson and Sparck Jones, 1976] Robertson, S. E. and Sparck Jones, K. (1976). *Relevance weighting of search terms*, volume 27, pages 129–146.

[Robertson and Zaragoza, 2009] Robertson, S. E. and Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

[Rocchio, 1966] Rocchio, J. J. (1966). *Document Retrieval System - Optimization and Evaluation*. PhD thesis, Harvard University, Cambridge, Massachusetts.

[Rocchio, 1971] Rocchio, J. J. (1971). *Relevance Feedback in Information Retrieval*, pages 313–323.

[Ruthven and Lalmas, 2003] Ruthven, I. and Lalmas, M. (2003). A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review*, 18(2):95–145.

[Ruthven et al., 2003] Ruthven, I., Lalmas, M., and Van Rijsbergen, K. (2003). Incorporating user search behavior into relevance feedback. *Journal of the American Society for Information Science and Technology*, 54:529–549.

[Salton, 1968] Salton, G. (1968). *Automatic Information Organization and Retrieval.* McGraw Hill Text.

[Salton and Buckley, 1990] Salton, G. and Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American society for information science*, 41(4):288–297.

[Sanderson, 2010] Sanderson, M. (2010). Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375.

[Schütze et al., 1995] Schütze, H., Hull, D. A., and Pedersen, J. O. (1995). A comparison of classifiers and document representations for the routing problem. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '95, pages 229–237, New York, NY, USA. ACM.

[Shen et al., 2007] Shen, X., Tan, B., and Zhai, C. (2007). Privacy protection in personalized search. *SIGIR Forum*, 41(1):4–17.

[Smyth and Balfe, 2006] Smyth, B. and Balfe, E. (2006). Anonymous personalization in collaborative web search. *Information Retrieval*, 9:165–190. 10.1007/s10791-006-7148-z.

[Song and Croft, 1999] Song, F. and Croft, W. B. (1999). A general language model for information retrieval. In *Proceedings of the eighth international conference on*

Information and knowledge management, CIKM '99, pages 316–321, New York, NY, USA. ACM.

[Spark Jones, 1972] Spark Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.

[Srikanth and Srihari, 2002] Srikanth, M. and Srihari, R. (2002). Biterm language models for document retrieval. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '02, pages 425–426, New York, NY, USA. ACM.

[Story, 1996] Story, R. E. (1996). An explanation of the effectiveness of latent semantic indexing by means of a bayesian regression model. *Information Processing and Management: an International Journal*, 32(3):329–344.

[Taylor, 1968] Taylor, R. S. (1968). Question-Negotiation and Information Seeking in Libraries. *College and Research Libraries*, (29):178–194.

[Teevan et al., 2004] Teevan, J., Alvarado, C., Ackerman, M. S., and Karger, D. R. (2004). The perfect search engine is not enough: a study of orienteering behavior in directed search. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '04, pages 415–422, New York, NY, USA. ACM.

[Teevan et al., 2005] Teevan, J., Dumais, S. T., and Horvitz, E. (2005). Personalizing search via automated analysis of interests and activities. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 449–456, New York, NY, USA. ACM.

[Teevan et al., 2010] Teevan, J., Dumais, S. T., and Horvitz, E. (2010). Potential for personalization. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 17(1):1–31.

[Teevan et al., 2008] Teevan, J., Dumais, S. T., and Liebling, D. J. (2008). To personalize or not to personalize: modeling queries with variation in user intent. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 163–170, New York, NY, USA. ACM.

[Teevan et al., 2009] Teevan, J., Morris, M. R., and Bush, S. (2009). Discovering and using groups to improve personalized search. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining (WSDM '09)*, Barcelona, Spain.

[Turtle and Croft, 1991] Turtle, H. and Croft, W. B. (1991). Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9:187–222.

[van Rijsbergen, 1979] van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworth.

[van Rijsbergen, 1986] van Rijsbergen, C. J. (1986). A new theoretical framework for information retrieval. *SIGIR Forum*, 21:23–29.

[van Rijsbergen, 2004] van Rijsbergen, C. J. (2004). *The Geometry of Information Retrieval*. Cambridge University Press, New York, NY, USA.

[Vassilvitskii and Brill, 2006] Vassilvitskii, S. and Brill, E. (2006). Using web-graph distance for relevance feedback in web search. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 147–153, New York, NY, USA. ACM.

[White et al., 2009] White, R. W., Bailey, P., and Chen, L. (2009). Predicting user interests from contextual information. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 363–370, New York, NY, USA. ACM.

[White et al., 2003] White, R. W., Jose, J. M., and Ruthven, I. (2003). A task-oriented study on the influencing effects of query-biased summarisation in web searching. *Information Processing and Management*, 39:707–733.

[White and Kelly, 2006] White, R. W. and Kelly, J. D. (2006). A study on the effects of personalization and task information on implicit feedback performance. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 297–306, New York, NY, USA. ACM.

[White and Roth, 2009] White, R. W. and Roth, R. A. (2009). *Exploratory Search: Beyond the Query-Response Paradigm*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers.

[White et al., 2005] White, R. W., Ruthven, I., Jose, J. M., and Rijsbergen, C. J. V. (2005). Evaluating implicit feedback models using searcher simulations. *ACM Transactions on Information Systems*, 23:325–361.

[Wong et al., 2008] Wong, W. S., Luk, R. W. P., Leong, H. V., Ho, K. S., and Lee, D. L. (2008). Re-examining the effects of adding relevance information in a relevance feedback environment. *Information Processing and Management*, 44:1086–1116.

[Xu and Croft, 2000] Xu, J. and Croft, W. B. (2000). Improving the effectiveness of information retrieval with local context analysis. *ACM Transaction on Information Systems*, 18(1):79–112.

[Yilmaz et al., 2008] Yilmaz, E., Aslam, J. A., and Robertson, S. (2008). A new rank correlation coefficient for information retrieval. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 587–594, New York, NY, USA. ACM.

[Zhai, 2008] Zhai, C. (2008). *Statistical Language Models for Information Retrieval*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

[Zhai and Lafferty, 2001] Zhai, C. and Lafferty, J. (2001). Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the tenth international conference on Information and knowledge management*, CIKM '01, pages 403–410, New York, NY, USA. ACM.

[Zuccon et al., 2009] Zuccon, G., Azzopardi, L., and van Rijsbergen, C. J. (2009). Semantic spaces: Measuring the distance between different subspaces. In Bruza, P., Sofge, D., Lawless, W., van Rijsbergen, K., and Klusch, M., editors, *Quantum Interaction*, volume 5494 of *Lecture Notes in Computer Science*, pages 225–236. Springer Berlin / Heidelberg.

ACRONYMS

AP Average Precision

BIR Binary Independence Retrieval

BLUE Best Linear Unbiased Estimator

BPP Best Performing Projector

DCG Discounted Cumulative Gain

DF document frequency

DIR Distributed IR

EMIM Expected Mutual Information Measure

FA Factor Analysis

GVSM Generalized Vector Space Model

HAL Hyperspace Analogue to Language

IDF Inverse Document Frequency

IMS Information Management System

IDCG Ideal Discounted Cumulative Gain

IF Information Flow

IR Information Retrieval

IRF Implicit Relevance Feedback

LCA Local Context Analysis

LM Language Modeling

LLSI Local Latent Semantic Indexing

LSI Latent Semantic Indexing

MAP Mean Average Precision

MRR Mean Reciprocal Rank

MRF Markov Random Field

MST Maximum Spanning Tree

NDCG Normalized Discounted Cumulative Gain

NEU North-eastern University

P2P Peer-To-Peer

PCA Principal Component Analysis

PRF Pseudo-Relevance Feedback

PRP Probability Ranking Principle

P2P Peer-To-Peer

QL Query Likelihood

QM Quantum Mechanics

QT Quantum Theory

RF Relevance Feedback

RM Relevance Model

SNR Signal-to-Noise Ratio

SPINA Superimposed Peer Infrastructure for iNformation Access

SVD Singular Value Decomposition

TF term frequency

TREC Text REtrieval Conference

VSM Vector Space Model

WARC Web ARChive

XML eXtensible Markup Language

ACKNOWLEDGMENTS

Firstly, I would like to thank my supervisor, Professor Massimo Melucci, without whose guidance, knowledge, and encouragement this thesis would never have been possible.

A great deal of gratitude is due to the members of the Information Management Systems Research Group of the University of Padua for the fruitful discussions during these years and for letting me be part of challenging and intriguing projects.

I was lucky enough to be able to do two research stays. In 2008 I spent two months at the Knowledge Media Institute (KMi) of The Open University, in Milton Keynes. I would like to thank Professor Dawei Song for inviting me and giving me the opportunity to carry out the user study that allowed part of the investigation reported in this thesis. I would like to thank Dr. Qiang Huang for his support in the design of the user study and Srikanth Reddy Chilumula for helping me with the development of the system to gather interaction data. Moreover, I would like to thank all people involved in the user study because their *interaction* was crucial for part of the work reported in this thesis.

In 2010 I spent six months at the Department IRO of the University of Montreal (UdeM). I would like to thank all the members of the RALI Laboratory. I am grateful to Professor Jian-Yun Nie for giving me the opportunity to visit UdeM, for providing me with invaluable insights and for the best Chinese food ever.

I had the opportunity to meet many people during these three years. I would like to thank all of them for the fruitful discussions, especially the members of the Information Retrieval Group of the University of Glasgow, Professor Ian Ruthven for mentoring me during the IIiX2010 doctoral consortium, and all TREC people for their work and their insights in evaluation in IR.

I would especially like to acknowledge my parents and my sister Paola for persistent encouragement and support. Thanks to the Cogo family for their trust in me.

Finally, I would like to thank Lucia for being so patient and for her constant support during this journey.

If we knew what it was we were doing, it
would not be called research, would it?
- Albert Einstein (1879 - 1955)