An Information Visualization Tool for the Interactive Component-Based Evaluation of Search Engines

Giacomo Rocco and Gianmaria Silvello(✉)

Department of Information Engineering, University of Padua, Padua, Italy
gianmaria.silvello@unipd.it

Abstract. In this paper, we present an InfoVis tool based on SanKey diagrams for the exploration of large combinatorial combinations of IR components – the *Grid of Points (GoP)*.

The goal of this tool is to ease the comprehension of the behavior of single IR components within fully functioning off-the-shelf IR systems without recurring to complex statistical tools. In order to assess the quality of the proposed SanKey-based InfoVis tool we conduceted an initial user study that led to interesting conclusions, yet to be validated in a future and more comprehensive study.

Keywords: Information Retrieval · Evaluation · Grid of Points · Information visualization · Sankey

1 Motivations

Information Retrieval (IR) systems are constituted of "pipelines" of components such as stop lists, stemmers and IR models, which are stacked together in order to process both documents and user queries and to match them returning a ranked result list of documents in decreasing order of estimated relevance. The performance of IR systems are evaluated in terms of *effectiveness* that can be determined only after that the system has been built; indeed, no effectiveness prediction about a specific component can be done before it has been tested within a fully functioning IR system.

Currently, the only viable means to determine the contribution to the system effectiveness of single components is to measure their impact on the overall performances by testing all the different combinations of such components. This leads to a very high number of cases to be considered, making the space of system combinations large and complex to explore.

Besides requiring a great deal of effort and resources to be produced, these combinatorial compositions constitute a challenge when it comes to explore, analyze, and make sense of the experimental results with the goal of understanding how different components contribute to the overall performances and interact together. Indeed, it is typically needed to resort to rather complex statistical

© Springer Nature Switzerland AG 2020
M. Ceci et al. (Eds.): IRCDL 2020, CCIS 1177, pp. 15–25, 2020.
https://doi.org/10.1007/978-3-030-39905-4_3

tools (e.g. multi-way *ANalysis Of VAriance (ANOVA)* models) requiring a careful experimental design and producing results which call for a considerable extent of expertise to be interpreted [6]. To this end, we developed an extensive set of $612 \times 6 = 3,672$ systems – i.e. the *Grid of Points (GoP)*[1] – arising from the combinatorial composition of several open-source publicly available components such as stop lists, stemmers, and IR models, and run against 6 different public test collections shared by the *Text REtrieval Conference (TREC)* international evaluation initiative. Thanks to this GoP, in [8] we presented the deep statistical analyses we run and the insights we gathered about the individual contributions of single IR components to the overall performances of fully working IR systems.

In this paper we present an InfoVis system based on SanKey diagrams – often used in physics to represent energy inputs, useful output, and wasted output – to allow the exploration of the GoP to quickly understand which combinations perform best under specific criteria, how components behave across a wide range of cases, and how they interact together. Our main goal is to give IR researchers and practitioners a fast and easy way to understand and analyze the GoP without recurring to demanding and complex statistical tools.

Hence, the InfoVis tool we present enables the analysis and comparison of a complex set of measures associated with a large combinatorial space of IR systems and the intuitive exploration and understanding of many component configurations. It is thought to be simple to use and to favor interaction, thus it provides functionalities as component filtering, measure selection and tooltips presenting statistical information easy to interpret. We present a user study to validate the presented tool.

The rest of the paper is organized as follows: Sect. 2 presents the related works, Sect. 3 describes the experimental setup and the Grid of Points we are considering for the visual tool, Sect. 4 describes the visual tool based on the Sankey visualization detailing the main components and its use, Sect. 5 reports the results of the user study which compared the present tool with another state-of-the-art visual tool though for the same task and Sect. 6 draws some final remarks.

2 Related Work

InfoVis techniques are typically exploited for the presentation and exploration of the *documents* managed by an IR system [16]. Typical examples are: identification of the objects and their attributes to be displayed [9]; different ways of presenting the data [13]; the definition of visual spaces and visual semantic frameworks [15]. The development of interactive means for IR is an active field which focuses on search user interfaces [10], displaying of results [4] and browsing capabilities [11].

Less attention has been dedicated to the application of InfoVis techniques to the analysis of experimental evaluation results. One example of a system

[1] http://gridofpoints.dei.unipd.it/.

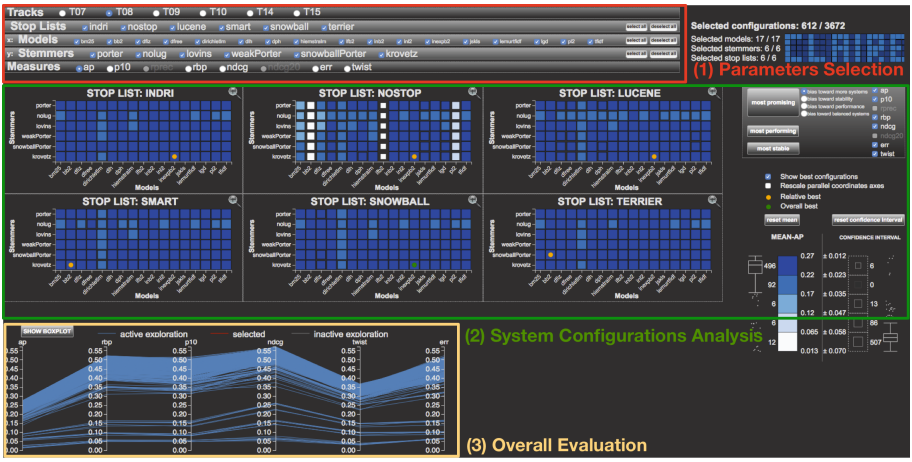

Fig. 1. The overall view of the CLAIRE visual analytics tool [1].

applying visualization to IR is *Visual Information Retrieval Tool for Upfront Evaluation (VIRTUE)*, a visual analytics tool supporting performance and failure analysis [2]. In the same vein, [3] presents an analytical framework trying to learn the behavior of a system just from its outputs for obtaining a rough estimation of the possible effects of a modification to the system. More recently, [12] presented an InfoVis tool to explore pooling strategies.

However, to the best of our knowledge only one solution – i.e. the CLAIRE tool [1], see Fig. 1 – exists for dealing with large sets of IR systems – the GoP [6,7] – generated by many IR components which allows the inspection of both configurations and measures. CLAIRE is based on a totally different visual paradigm since it uses tiles, parallel coordinates and boxplots to explore system configurations. *Combinatorial visuaL Analytics system for Information Retrieval Evaluation (CLAIRE)* is composed of three main areas: (i) the *Parameters Selection* area, dealing with the exploration coordinates; (ii) the *System Configurations Analysis* area, enabling the performance analysis of the system configurations; and, (iii) the *Overall Evaluation* area, where the system configurations performances are evaluated.

The visual tool we present in this paper follows the same overall organization, but it relies on a different visual paradigm allowing for an intuitive, yet less deep comprehension of the evaluation results over the considered Grid of Points. Indeed, CLAIRE has a strong focus on Visual Analytics, whereas the SanKey-based InfoVis tool we present here is specifically tailored to Information Visualization. The main difference is that visual analytics aims at exploiting visual clues to actually inform or modify analytical or algorithmic tools working over some data; on the other hand, information visualization aims at providing visual tools to better understand complex and possibly high-dimensional data.

Overall, CLAIRE is a more complex system than the SanKey-based InfoVis tool presented here, even though they are comparable for the information visualization part since they both allow the user to select different evaluation collections and measures. Both the systems aim at intuitively visualize multi-dimensional data from different perspectives. Moreover, they both allow the user to select different IR system components and understand how they interact with one another also grasping the overall contribution of a single component over the whole search pipeline.

3 Experimental Setting

The GoP data adopted by our InfoVis tool is based on three main components of an IR system: stop list, stemmer, and IR model. We selected a set of alternative implementations of each component and, by using the Terrier v.4.0[2] open source system, we created a run for each system defined by combining the available components in all possible ways. The selected components are:

- *Stop list*: `nostop`, `indri`, `lucene`, `snowball`, `smart`,`terrier`;
- *Stemmer*: `nolug`, `weakPorter`, `porter`, `snowballPorter`, `krovetz`, `lovins`;
- *Model*: `bb2`, `bm25`, `dfiz`, `dfree`, `dirichletlm`, `dlh`, `dph`, `hiemstralm`, `ifb2`, `inb2`, `inl2`, `inexpb2`, `jskls`, `lemurtfidf`, `lgd`, `pl2`, `tfidf`.

Overall, these components define a $6 \times 6 \times 17 = 612$ runs. The stop lists differ from each other by the number of terms composing them; specifically, `indri` has 418 terms, `lucene` has 33 terms, `snowball` has 174 terms, `smart` has 571 terms and `terrier` 733 terms. Stemmers can be classified into aggressive (e.g. `lovins`) and weaker stemmers (e.g. `porter`).

The models we employ are classified into the three main approaches currently adopted by search engines: (1) the vector space model – e.g. `tfidf` and `lemurtfidf`; (2) the probabilistic model – e.g. `bm25` and the *Divergence From Randomness (DFR)* models; and, (3) the language models – e.g. `dirichletlm`, `hiemstralm` and `lgd`. We considered 6 standard and shared collections with 50 different topics each: *TREC Adhoc tracks* `T07` *and* `T08`; *TREC Web tracks* `T09` *and* `T10`; and, *TREC Terabyte tracks* `T14` *and* `T15`. We evaluate the GoPs by employing 8 evaluation measures: AP, P@10, Rprec, RBP, nDCG, nDCG@20, ERR, and Twist.

Summarizing, the GoP we visualize with the proposed InfoVis tool consists of 612 runs over 6 collections with 50 topics each and evaluated with 8 measures, which amounts to almost 1.5M data points.

[2] http://www.terrier.org/.

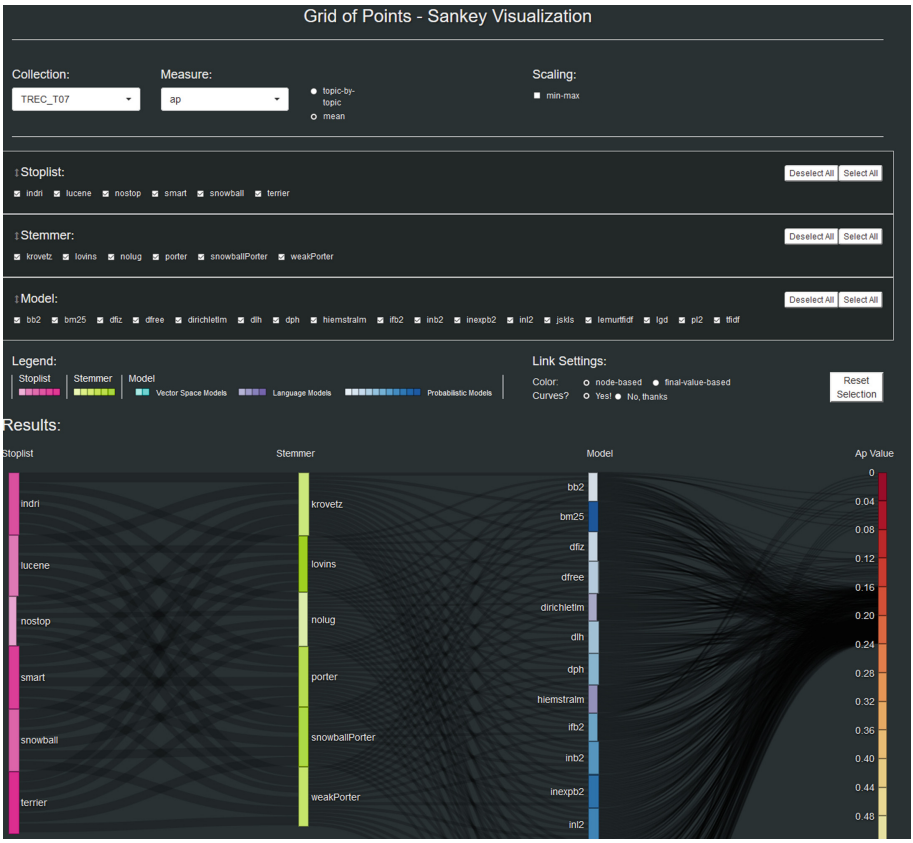

Fig. 2. The overall InfoVis system; on the top there is the parameter selection area and on the bottom the dynamic SanKey diagram.

4 The InfoVis Tool

The InfoVis tool we realized, see Fig. 2 is composed of two main areas:

Parameters selection area: (top of Fig. 2) it allows the user to load the runs relative to the desired experimental collection, to select the components s/he wants to consider and the evaluation measure to be used.

System analysis area: (bottom of Fig. 2) it allows the actual analysis and exploration of the various components and their evaluation on the basis of the parameters selected above.

4.1 Parameter Selection Area

In Fig. 3 we can see a detailed view of the parameter selection area.

The first two parameters that can be selected (in the green box) are the experimental collection and the evaluation measure of interest. On the left of

Fig. 3. A detailed view of the *p*arameter selection area (Color figure online)

these two drop down menus we can choose to visualize the system performances topic-by-topic (if this option is selected a new drop-down menu appears allowing the user to select the topic of interest) or on average (e.g. MAP). The "scaling" option enables a normalized visualization of the SanKey diagram (only actual min-max values or the whole range such as $[0, 1]$ for AP). The blue box in Fig. 3 shows the control panel enabling the dynamic selection of component families to be visualized in the SanKey diagram.

The three component families (stoplists, stemmers and IR models) can be re-ordered by a simple drag-and-drop action, leading to a dynamic re-ordering of the axes of the Sankey diagram; this is particularly useful when during the data analysis phase we want to highlight the components interaction. The default axes order better shows the interaction between stoplists and stemmers and between stemmers and IR models, but by re-ordering the axes we can highlight, for instance, the stoplists-models interaction.

Below the blue box we can see the legend of the SanKey diagram where three chromatic variations are used to differentiate between the components of each family and sub-family of components: fuchsia for stoplists, green for stemmers, light blue for vector space models, purple for language models and dark blue for probabilistic models. The fuchsia box highlights the link settings where we can choose the shape of the SanKey curves and their color schema – i.e. based on component selection or based on evaluation measure value selection.

Every single interaction with the parameter selection area produces an effect on the SanKey diagram which is rendered dynamically and in real-time; this is intended to ease the interaction with the system and the data analyses to be performed.

4.2 System Analysis Area

On the bottom of Fig. 2 we can see the entire analysis space where all the available components are displayed by the SanKey diagram, whereas in Fig. 4 we

can see a restricted analysis area where only some specific components have been selected and highlighted for an in-depth analysis of their performances and interactions.

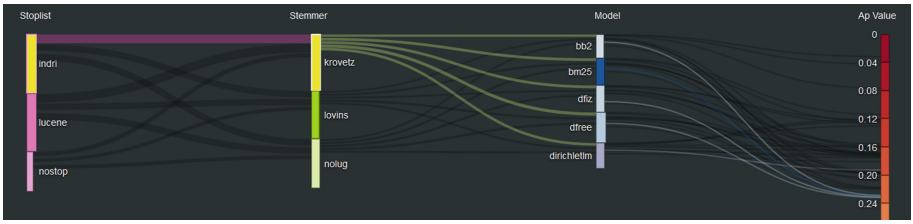

Fig. 4. A detailed view of the *s*ystem analysis area where some components have been filtered out and some other are highlighted for an in-depth analysis of the interactions. (Color figure online)

The rightmost column presents the evaluation measure values divided into 25 rectangles of equal size, each one representing a 0.04 value interval. The color of each rectangle follows the red-yellow-green schema where reddish rectangles are assigned to lower values and the greenish ones to higher values. By the means of a drag-and-drop mouse action it is also possible to re-order the rectangles representing family components. Each single link insisting on these rectangles represents one of the 612 systems and their overall performance values.

A single system is represented by a path, i.e. a series of links connecting one component with the next one. The user can select a set of components (left click on one or more rectangles) to highlight the paths of interest as shown in Fig. 4 where we selected the `indri` stoplist and the `krovetz` stemmer.

The component columns present a number of rectangles equal to the components selected in the *p*arameter selection area and the size of the rectangle gives a visual idea of the performances of the component it represents. This is done by calculating the marginal arithmetic mean of the performance values obtained by the systems using a specific component; the means are dynamically re-calculated every time a component is filtered out or added to the visualization. In Fig. 4, we can see that `krovetz` has a bigger rectangle than `lovins` and `nolug` (meaning no stemmer) showing the positive effect of the `krovetz` stemmer when interacting with the `indri` stoplist and the selected models.

The same idea is applied to the link size: the thicker the line the better the interaction between the components it connects. For instance, in Fig. 4 we can see that the stoplist-stemmer pair `indri-krovetz` has higher performances than the pair `lucene-krovetz`.

With a mouse-over action on a rectangle or a link, a tooltip reporting the top 5 systems using the selected component (rectangle) or the selected components pair (link) is visualized to the user. The InfoVis system also runs the Dunnett [5] statistical test to determine if the reported means are statistically different one

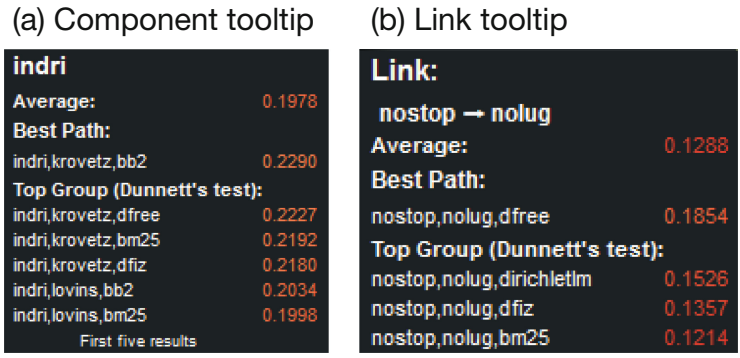

Fig. 5. (a) The tooltip visualized with a mouse-over action on the `indri` stoplist component and (b) the tooltip visualized with a mouse-over the `nostop-nolug` link.

from the other. In Fig. 5(a) we can see the tooltip visualized when the `indri` stoplist is selected: we show the average measure (AP in this case) of all the system using this stoplist, the best system adopting the stoplist and the top group of system adopting the `indri` stoplist that are not statistically different one from the other. In Fig. 5(b) we see the tooltip reporting the statistical information related to the `nostop-nolug` link.

5 User Evaluation

We did an initial user study with nine users (i.e., master degree students in Information Engineering) with a basic knowledge and previous experience with IR systems and experimental evaluation in the field; the study had a twofold goal, to compare the SanKey-based InfoVis tool with the CLAIRE system and to conduct an in-depth analysis of the newly proposed SanKey-based InfoVis tool. Of course, CLAIRE is a more complex system providing a wide range of functionalities, but we focused on the common features which regards the exploration of the combinatorial space of IR system pipelines.

The test was organized in three phases: (i) in-depth description of the two visual tools and hands-on phase to get to know them; (ii) *comparative study:* execution of three tasks with both CLAIRE and the SanKey-based InfoVis tool (in this phase we divided the users into two groups where one group used firstly CLAIRE and then the SanKey-based InfoVis tool and the second group did the opposite); (iii) *in-depth analysis:* execution of five tasks by using only the SanKey-based InfoVis tool. The tasks were centered around core activities enabled by the two visual tools such as the ability to determine the best IR system, the best combination of components, the comparison between two or more alternative components and so on. After the resolution of the first group of tasks the users were required to fill closed questionnaire. After the resolution of the second group of tasks the users were required to fill in an open questionnaire.

The questionnaire relative to the first set of tasks required to get a preference between the SanKey-based InfoVis tool and CLAIRE, was composed of two sets of questions; the first set with three questions: (Q1) How intuitive was the SanKey-based InfoVis tool (CLAIRE) tool? (Q2) In your opinion how much useful is the SanKey (CLAIRE) tool to understand the performances of IR systems? (Q3) How much effective was the SanKey (CLAIRE) tool to solve the given tasks? Each question of the questionnaire had to be answered by using an interval Likert scale ranging from 1 to 5 in which each numerical score was labeled with a description: {1: not at all, 2: a little, 3: enough, 4: a lot, 5: quite a lot}.

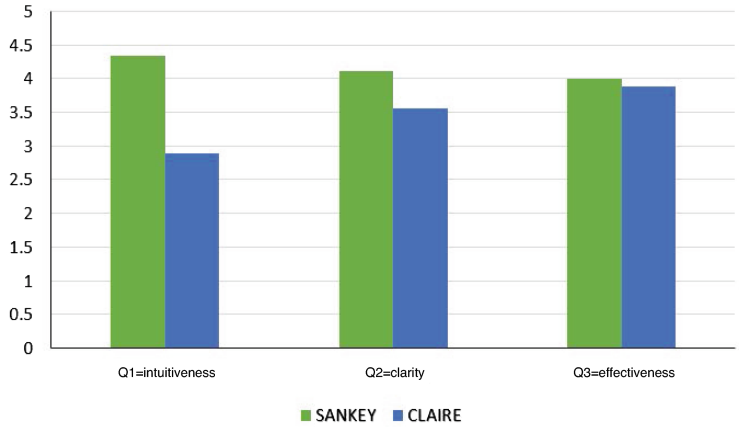

Fig. 6. Average answers for the first set of questions of the comparative study

In Fig. 6 we can see that both systems were evaluated as clear to use (Q2) and effective (Q3) in both cases with a slight preference for SanKey; but SanKey was considered more intuitive than CLAIRE (Q1).

The second set of questions for the comparative study was: (Q1) Which system does represent better the experimental data? (Q2) Which system does offer the most intuitive interface to interact with the data? (Q3) Which system is more complete to solve the assigned tasks? (Q4) Which system did you prefer to use? Each question of the questionnaires had to be answered by indicating a strong preference (a "2" in our interval scale) or a mild preference (a "1") for CLAIRE or SanKey where a "0" value indicated equality between the systems.

In Fig. 7 we can see that on average SanKey was preferred by the users with the only exception of Q3 where the systems were judged equivalent.

6 Final Remarks

The InfoVis tool we presented has the goal to ease the exploration and analysis of large experimental GoP enabling IR researchers and practitioners to better

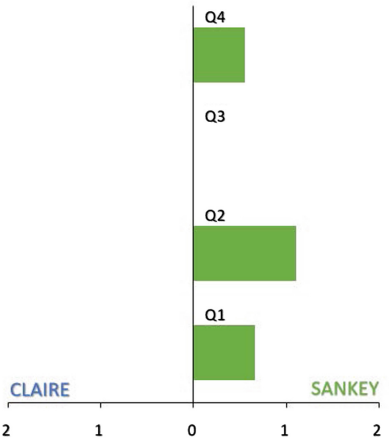

Fig. 7. Average answers for the second set of questions of the comparative study

understand the performances of single components, their interactions and their impact on off-the-shelf IR systems. The InfoVis tool we propose is highly interactive and remarkably simple as shown by the user study we conducted, yet offering advanced statistical information and analytics functionalities. Note that the user study has to be improved, thus the quality assessment of the SanKey-based InfoVis tool is initial and has to be further investigated to lead to more solid conclusions.

The presented tool is available on-line at the URL:

http://gridofpoints.dei.unipd.it/sankey/

and the source code is openly shared at the URL:

https://github.com/giansilv/sankey_eval

References

1. Angelini, M., Fazzini, V., Ferro, N., Santucci, G., Silvello, G.: CLAIRE: A combinatorial visual analytics system for information retrieval evaluation. Inf. Process. Manag. (2018). https://doi.org/10.1016/j.jvlc.2013.12.003. in print
2. Angelini, M., Ferro, N., Santucci, G., Silvello, G.: VIRTUE: a visual tool for information retrieval performance evaluation and failure analysis. J. Vis. Lang. Comput. (JVLC) **25**(4), 394–413 (2014)
3. Angelini, M., Ferro, N., Santucci, G., Silvello, G.: A visual analytics approach for what-if analysis of information retrieval systems. In: Perego et al. [14], pp. 1081–1084 (2016)
4. Crestani, F., Vegas, J., de la Fuente, P.: A graphical user interface for the retrieval of hierarchically structured documents. Inf. Process. Manag. **40**(2), 269–289 (2004)
5. Dunnett, C.W.: A multiple comparison procedure for comparing several treatments with a control. J. Am. Stat. Assoc. **50**(272), 1096–1121 (1955)

6. Ferro, N., Silvello, G.: A general linear mixed models approach to study system component effects. In: Perego et al. [14], pp. 25–34
7. Ferro, N., Silvello, G.: 3.5K runs, 5K topics, 3M assessments and 70M measures: what trends in 10 years of Adhoc-ish CLEF?.Inf. Process. Manag. 53(1), 175–202 (2017)
8. Ferro, N., Silvello, G.: Towards an anatomy of IR system component performances. J. Am. Soc. Inf. Sci. Technol. (JASIST) **69**(2), 187–200 (2017)
9. Fowler, R.H., Lawrence-Fowler, W.A., Wilson, B.A.: Integrating query, thesaurus, and documents through a common visual representation. In: Fox, E.A. (ed.) Proc. 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1991). ACM Press, New York, USA (1991)
10. Hearst, M.A.: "Natural" search user interfaces. Commun. ACM (CACM) **54**(11), 60–67 (2011)
11. Koshman, S.: Testing user interaction with a prototype visualization-based information retrieval system. J. Am. Soc. Inf. Sci. Technol. (JASIST) **56**(8), 824–833 (2005). https://doi.org/10.1002/asi.20175
12. Lipani, A., Lupu, M., Hanbury, A.: Visual pool: a tool to visualize and interact with the pooling method. In: Kando, N., Sakai, T., Joho, H., Li, H., de Vries, A.P., White, R.W. (eds.) Proceedings of 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2017). ACM Press, New York (2017)
13. Morse, E.L., Lewis, M., Olsen, K.A.: Testing visual information retrieval methodologies case study: comparative analysis of textual, icon, graphical, and spring displays. J. Am. Soc. Inf. Sci. Technol. (JASIST) **53**(1), 28–40 (2002)
14. Perego, R., Sebastiani, F., Aslam, J., Ruthven, I., Zobel, J. (eds.): Proceedings of 39th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2016). ACM Press, New York (2016)
15. Zhang, J.: TOFIR: a tool of facilitating information retrieval - introduce a visual retrieval model. Inf. Process. Manag. **37**(4), 639–657 (2001)
16. Zhang, J.: Visualization for Information Retrieval. Springer, Heidelberg, Germany (2008)