WELCOME TO THE GATHERING PLACE

Computer Networks and ISDN Systems 30 (1998) 1–9

COMPUTER NETWORKS and ISDN SYSTEMS

The limits of Web metadata, and beyond

Massimo Marchiori [1]

The World Wide Web Consortium (W3C), MIT Laboratory for Computer Science, 545 Technology Square, Cambridge, MA 02139, USA

Abstract

The World Wide Web currently has a huge amount of data, with practically no classification information, and this makes it extremely difficult to handle effectively. It has been realized recently that the only feasible way to radically improve the situation is to add to Web objects a metadata classification, to help search engines and Web-based digital libraries to properly classify and structure the information present in the WWW.

However, having a few standard metadata sets is insufficient in order to have a fully classified World Wide Web. The first major problem is that it will take some time before a reasonable number of people start using metadata to provide a better Web classification. The second major problem is that no one can guarantee that a majority of the Web objects will be ever properly classified via metadata.

In this paper, we address the problem of how to cope with such intrinsic limits of Web metadata, proposing a method that is able to partially solve the above two problems, and showing concrete evidence of its effectiveness. In addition, we examine the important problem of what is the required "critical mass" in the World Wide Web for metadata in order for it to be really useful. © 1998 Published by Elsevier Science B.V. All rights reserved.

Keywords: Metadata; Automatic classification; Information retrieval; Search engines; Filtering

1. Introduction

The World Wide Web currently has a huge amount of data, with practically no classification information. It is well known that this makes it extremely difficult to effectively handle the enormous amount of information present in the WWW, as witnessed by everybody's personal experience, and by every recent study on the difficulties of information retrieval on the Web. It has been realized recently that the only feasible way to radically improve the situation is to add to Web objects a metadata classification, that is to say partially passing the task of classifying the content of Web objects from search engines and repositories to the users who are building and maintaining such objects. Currently, there are lots of preliminary proposals on how to build suitable metadata sets that can be effectively incorporated into HTML and that can help search engines and Web-based digital libraries to properly classify and structure the information present in the WWW (see e.g. [2–7,9,11]).

However, usage of metadata presents the big problem that it does not suffice to have some standard metadata sets in order to have a fully classified World Wide Web. The first major problem is that a long time will pass before a reasonable number of people will start using metadata to provide a better

[1] E-mail: max@lcs.mit.edu

0169-7552/98/$19.00 © 1998 Published by Elsevier Science B.V. All rights reserved.
PII S0169-7552(98)00085-3

Web classification. The second big problem is that none can guarantee that a majority of the Web objects will be ever properly classified via metadata, since by their nature metadata:

- are an optional feature,
- they make heavier the writing of Web objects,
- their usage cannot be imposed (this is necessary to ensure HTML backward compatibility).

Thus, even in the more optimistic view, there will be a long transitory period where very few people will employ metadata to improve the informative content of their Web objects, and after such period there will still be a relevant part of the WWW that will not make proper use of the metadata facilities.

In this paper, we address the problem of how to cope with such intrinsic limits of Web metadata. The main idea is to start from those Web objects that use some form of metadata classification, and extend this information thorough the whole Web, via suitable propagation of the metadata information. The proposed propagation technique relies only on the connectivity structure on the Web, and is fully automatic. This means that

(1) it acts *on top* of any specific metadata format, and
(2) it does not require any form of ad-hoc semantic analysis (which can nevertheless be smoothly added on top of the propagation mechanisms), therefore being completely *language independent*.

Another important feature is that the propagation is done by *fuzzifying* the metadata attributes. This means that the obtained metadata information can go beyond the capability of usual crisp metadata sets, providing more information about the strength of various metadata attributes, and also resolving eventual expressive deficiencies of the original metadata sets.

The method is applied to a variety of tests. Although, of course, they cannot provide a complete and thorough evaluation of the method (for which a greater number of tests would be needed), we think they give a meaningful indication of its potential, providing a reasonably good first study of the subject.

In the first general test, we show how the effectiveness of a metadata classification can be enhanced significantly. We also study the relationship between the use of metadata in the Web (how many Web objects have been classified) and the global usefulness of the corresponding classification. This enables also to answer questions like what is the "critical mass" for general WWW metadata (that is to say, how much metadata is needed in an area of the Web in order for it to be really useful).

Then, we present another practical application of the method, this time not in conjunction with a static metadata description, but with a dynamic metadata generator. We show how even a poorly performing automatic classifier can be enhanced significantly using the propagation method.

Finally, we apply the method to an important practical case study, namely the case of **PICS**[2]-compliant metadata, that allow parents to filter the Web from pages that can be offensive or dangerous for kids. The method is shown to provide a striking performance in this case, thus showing to be already directly applicable with success in real-life situations.

2. Fuzzy metadata

As stated in the introduction, existing WWW metadata sets are crisp, in the sense that they only have attributes, without any capability of expressing "fuzziness": either they assign an attribute to an object or they do not. Instead, we will *fuzzify* attributes, associating to each attribute a *fuzzy measure* of its relevance for the Web object, namely a number ranging from 0 to 1 (cf. [1]). An associated value of 0 means that the attribute is not pertinent at all for the Web object, while on the other hand a value of 1 implies that the attribute is fully pertinent. All the intermediate values in the interval 0–1 can be used to give a measure of how much the attribute is relevant to the Web object. For instance, one could have that an attribute is only relevant roughly for the 20% (and in this case the value would be 0.2), or is relevant for the 50% (value of 0.5), and so on. This allows greater flexibility, since a metadata categorization is by its very nature an approximation of a large and complex variety of cases into a summarization and simplification of concepts. This also has the beneficial effect that one can keep the complexity of the basic categorization quite small,

[2] http://www.w3.org/PICS

and need not artificially enlarge it significantly in order to cope with the various intermediate categorizations (as is well known, a crucial factor in the success of categorizations, especially of those that pretend to be mass-used, is simplicity). Indeed the number of basic concepts can be kept reasonably small, and then fuzzification can be employed to obtain a more complete and detailed variety of descriptions. Also, this allows us to cope with intimate deficiencies of already existing categorizations, in case they are not flexible enough.

3. Back-propagation

Suppose a certain Web object O has the associated metadatum **A**:v, indicating that the attribute **A** has fuzzy value v. If there is another Web object O′ with an hyperlink to O, then we can "back-propagate" this metadata information from O to O′. The intuition is that the information contained in O (classified as **A**:v) is also reachable from O′, since we can just activate the hyperlink. However, the situation is not like being already in O, since in order to reach the information we have to activate the hyperlink and then wait for O to be fetched. So, the relevance of O′ with respect to the attribute A is not the same as O′ (v), but is in a sense faded, since the information in O is only potentially reachable from O′, but not directly contained therein. The solution to this problem is to fade the value v of the attribute multiplying it by a "fading factor" f (with $0 < f < 1$). So, in the above example O′ could be classified as **A**:$v \cdot f$. The same reasonment is then applied recursively. So, if we have another Web object O″ with an hyperlink to O′, we can back-propagate the obtained metadatum **A**:$v \cdot f$ exactly in the same way, obtaining that O″ has the corresponding metadatum **A**:$v \cdot f \cdot f$.

3.1. Combining information

We have seen how it is possible to back-propagate metadata information along the Web structure. There is still another issue to consider, that is to say how to combine different metadata values that are pertinent to the same Web object. This can happen because a Web object can have several hyperlinks pointing to many other Web objects, and so it can be the case that several metadata information is back-propagated. Also, it can happen because the considered Web object has already some metadata information present in it.

The solution to these issues is not difficult. One case is easy: if the attributes are different, then we simply merge the information. For instance, if we calculate via back-propagation that a Web object has the metadatum **A**:v and also the metadatum **B**:w, then we can infer that the Web object has both the metadata, i.e. it can be classified as "**A**:v, **B**:w". The other case is when the attributes are the same. For instance, we may arrive to calculate via back-propagation that a Web object has attributes **A**:v (this metadatum back-propagated by a certain Web object) and **A**:w (this metadatum propagated by another Web object). The employed solution is to employ fuzzy arithmetic and use the so-called *fuzzy or* of the values v and w, that is to say we take the *maximum value* between the two. So, O would be classified as **A**:*max(v,w)* (here, *max* denotes as usual the maximum operator). This operation directly stems from classic **fuzzy logic**[3], and its intuition is that if we have several different choices for the same value (in this case, v and w), then we collect the most informative choice (that is, we take the maximum value), and discard the other less informative values. Above we only discussed the case of two metadata, but it is completely obvious how to extend the discussion to the case of several metadata, just by repeatedly combining them using the above operations, until the final metadatum is obtained.

As a technical aside, we remark how the above method can be seen as a "first-order" approximation of the back-propagation of the so-called hyper-information (see [8]); along the same line, more refined "higher-order" back-propagation techniques can be built, at the expense of computation speed.

4. Testing

In order to measure the effectiveness of the approach, we first need to "restrict" in some sense the Web to a more manageable size, that is to say to perform our studies in a reasonably sized region of

[3] http://www.cs.cmu.edu/Groups/AI/html/faqs/ai/fuzzy/part1/faq.html

it. In the following we explain how we coped with this problem.

Suppose to have a region of Web space (i.e. a collection of Web objects), say S; the following process augments it by considering the most proximal neighbourhoods:

Consider all the hyperlinks present in the set of Web objects S: retrieve all the corresponding Web objects and add them to the set S.

We can repeat this process as we want, until we reach a decently sized region of the Web S: we can impose as stop condition that the number of Web objects in S must be a certain number n.

As an aside, note we are overlooking for the sake of simplicity some implementation issues that we had to face, like for instance: checking that no duplications arise (that is to say, that the same Web object is not added to S more than once); checking that we have not reached a stale point (in some particular case, it may happen that the Web objects in S have only hyperlinks to themselves, and so the augmentation process does not increase the size of S); cutting the size of S exactly to the bound n (the augmentation process may in one step significantly exceed n).

So, in order to extract a region of the Web to work with, the following process is used:

(1) A random Web object is chosen (to this aim, we have employed **URouLette**[4]).
(2) It is augmented repeatedly using the aforementioned technique, until a Web region of the desired size n is reached.

Once a region S of the Web has been selected this way, we have to randomly select a certain percentage p, and manually classify them with metadata; then, we propagate the metadata using the back-propagation method.

4.1. Experimentation

For our general experimentation we have employed as metadata classification the well known ***Excite***[5] *Ontology*, also known as the "Channels Classification": it is a tree-like set of attributes (also known as categories), and is one of the best existing metadata sets for the general classification of World Wide Web objects.

We did several tests to verify the effectiveness of the back-propagation method, and also in order to decide what was the best choice for the fading factor f. Our outcome was that good performances can be obtained by using the following rule-of-thumb:

$$f = 1 - p$$

Intuitively, this means that when the percentage of already classified Web objects is low (p very low), then we need a correspondingly high value of f (f near to one), because we need to back-propagate more and we have to ensure that the fading does not become too low. On the other hand, when the percentage of already classified Web objects becomes higher, we need to back-propagate less, and so the fading can be kept smaller (in the limit case when $p = 1$, the rule gives indeed that $f = 0$, that is to say we do not need to back-propagate at all).

Even if our personal tests were quite good, our evaluations could be somehow dependent on our view of judging metadata. Moreover, besides confirmations on the effectiveness of the approach, we also needed more large-scale results analyzing in detail the situation with respect to the percentage p of already classified Web objects.

Therefore, we performed a more challenging series of tests, this time involving an *external population* of *34 users*. The initial part of the tests was exactly like before: first, a random Web object is selected, and then it is increased until a Web region of size n is reached. Then the testing phase occurs: we randomly choose a certain number m of Web objects that have not been manually classified, and let the users judge how well the calculated classifications describe the Web objects. The "judgment" of a certain set of metadata with respect to a certain Web object consists in a number ranging from 0 to 100, with the intended meaning that 0 stands for a completely wrong/useless metadata, while 100 for a careful and precise description of the considered Web object.

It is immediate to see that in the original situation, without back-propagation, the expected average judgment when $p = n\%$ is just n (for example, if $p = 7\%$, then the average judgment is 7%). So,

[4] http://www.iserv.net/links/roulette.html

[5] http://www.excite.com

we have to show how much back-propagation can improve on this original situation.

As said, for the tests we employed a population of 34 persons: they were only aware of the finalities of the experiment (to judge different methods of generating metadata), but they were not aware of our current method. Moreover, we tried to go further and to limit every possible "psychological pressure", clearly stating that we wanted to measure the qualitative differences between different metadata set, and so the absolute judgment was not a concern, while the major point was the judgment differences.

The first round of tests was done setting the size n of the Web region to 200, and the number m of Web objects that the users had to classify to 20 (which corresponds to the 10% of the total considered Web objects). Various tests were performed, in order to study the effectiveness of the method with the varying of the parameter p. We run the judgment test for each of the following percentages p of classified Web objects: $p = 2.5\%$, $p = 5\%$, $p = 7.5\%$, $p = 10\%$, $p = 15\%$, $p = 20\%$, $p = 30\%$, $p = 40\%$, $p = 50\%$, $p = 60{=}\%$, $p = 70\%$, $p = 80\%$ and $p = 90\%$. The final average results are shown in Table 1.

The data are graphically summarized in the following chart, where an interpolated spline is used to better emphasize the dependency between the percentage p of already classified Web objects and the obtained judgment j (Fig. 1; here, the diagonal magenta dotted line indicates the average judgment in the original case without back-propagation).

As we can see, there is an almost linear dependence roughly from the value of $p = 20\%$ on, until the value of $p = 70\%$ where the judgment starts its asymptotic deviation towards the upper bound. The behaviour is also rather clear in the region from $p = 5\%$ to $p = 20\%$: however, since in practice the lower percentages of p are also the more important to study, we run a second round of tests specifically devoted to this zone of p values. This time, the size of the Web region was five times bigger, setting $n = 1000$, and the value m was set to 20 (that is to say, the 2% of the Web objects). Again, various tests were performed, each one with a different value of p. This time, the granularity was increased, so that while in the previous phase we had the values of $p = 2.5\%$, $p = 5\%$, $p = 7.5\%$, $p = 10\%$, $p = 15\%$, $p = 20\%$, in this new phase we also run the tests for the extra cases $p = 12.5\%$ and $p = 17.5\%$, so to perform an even more precise analysis of this important zone. The final average results are shown in Table 2.

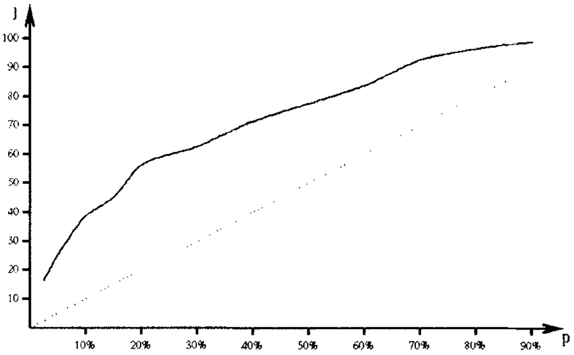


Fig. 1.

It can be seen as these new, more precise data essentially confirm the previous analysis performed with the more limited size $n = 200$. In the next chart, we again present an interpolated spline for this second round of tests (represented in green), and also overimpose the curve obtained in the first round (represented as a dashed curve, again in red), so to better show the connections (see Fig. 2).

Table 1
Average judgments in the $n = 200$ case

p:	2.5%	5%	7.5%	10%	15%	20%	30%	40%	50%	60%	70%	80%	90%
judgment:	17.1	24.2	31.8	39.7	44.9	56.4	62.9	72.1	78.3	83.8	92.2	95.8	98.1

Table 2
Average judgments in the $n = 1000$ case

p:	2.5%	5%	7.5%	10%	12.5%	15%	17.5%	20%
judgment:	14.7	20.6	30.0	38.3	40.7	47.3	54.8	58.5

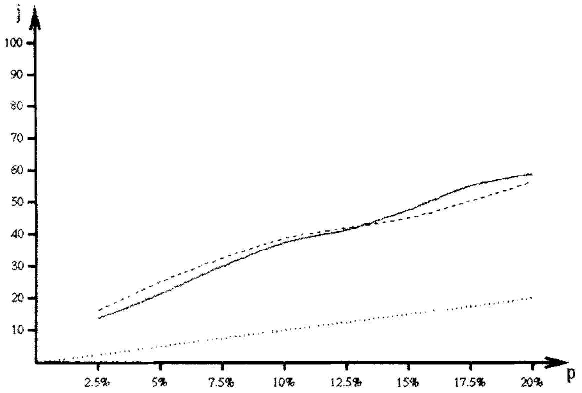


Fig. 2.

Therefore, the results we have obtained show that:

- When employing a particular metadata classification for the Web, even when the percentage of classified Web objects is relatively small, usage of the back-propagation method can significantly help the effectiveness of the classification, thus helping the metadata to get more and more widespread, especially in the initial crucial phases when the number of classified objects will be extremely limited.
- The experimentation can help to answer the fundamental question: what is the minimum "critical mass" that a general WWW metadata classification needs to reach in order to be really useful? If we fix the level of decency of metadata to a judgment level of 50, then by reverse interpolation we obtain that a reasonable prediction for the critical mass is roughly 16%. Thus, in order for a general WWW metadata classification to be really useful, at least 16% of the WWW (or of the local WWW zone we are considering) should be classified. Note that, without back-propagation, the critical mass would grow up to 50%.
- Analogously, we can predict what is the percentage metadata needs to have in order to reach an "excellence level": if we fix such a level at 80, then by reverse interpolation we obtain a value of 53%; so, using the back-propagation method we can reach an excellence level when slightly more than half of the WWW (or of the local WWW zone we are considering) has been classified.

The interpolations we have given also give an indication on how the situation is in the transitory phase before the "critical mass" is reached, and how the usefulness of metadata evolves after the critical mass is surpassed:

- Before the critical mass is reached, there is roughly a linear dependency between the *p*ercentage of metadata and its usefulness (as given by the judgment) with a linear factor of 3 (i.e., as soon as p increments, the usefulness increases three times faster);
- After the critical mass is reached, and until the value of $p = 70\%$, there is roughly a linear dependency between the *p*ercentage of metadata and its usefulness (as given by the judgment) with a linear factor of 0.6; so, this time usefulness increases slower (slightly more than at half speed) as the percentage of metadata increases;
- After the value of $p = 70\%$, we have an even slower linear dependency with the value of roughly 0.3.

5. Automatic classification

The strength of the propagation method lies in his independence on the specific metadata format, being in a sense a "meta-method" that enhances the effectiveness of existing metadata. This means it can be applied not only to already existing static collections of metadata, but also in connection with dynamic methods for the classification of Web objects. In order to show the potential of the method in this field, we built a naive automatic classifier, working this way. For each category **A** in the Excite ontology, some representative keywords are chosen. Then, when a document is processed, the classifier calculates the corresponding metadata **A**:v as follows; v is set to 1/2 of the following sum:

The ratio between the number of words in the document that are representative keywords of **A**, and the total number of words in the document.

plus

1/10 of the the number of words in the document that are representative keywords of **A**, with the provisos that:

(1) Each word contained in a Heading or Title header counts twice

(2) If the final number of words exceeds 10, then set it to 10.

The above classifier employs a mixture between percentage count (first component of the sum) and frequency count (second component of the sum). As it can be guessed, it is a rather rough metadata generator, like all the automatic classifiers that do not use semantic-based arguments. Nevertheless, many speed-demanding automatic classification systems, like for instance WWW search engines, currently employ classifiers that, although more sophisticated, are essentially based on the same evaluation criteria, and as such suffer of similar weaknesses (see for instance the **Search Engine Watch**[6]).

We have again conducted a test to measure the quality of the obtained metadata, employing the same population of 34 persons employed in the first experiment. The test was conducted much in the same way: again, we built various Web regions of size $n = 200$. Now, however, since we have an automatic classifier we are not limited any more to classify a minor part of the Web region: so, we run the automatic classifier on the whole Web region. Again, the users were asked to judge the quality of the obtained metadata, just like in the previous experiment, on a random sampler of $n = 20$ Web objects. The test was repeated 10 times, each time with a different Web region. The final average quality judgment was 43.2.

Then, we did the same tests, but this time we used the back-propagation method on the metadata obtained by the automatic classifier. The final outcome was: the average quality judgment jumped to 72.4 (in nice accordance with the interpolation results obtained in the previous experiment!).

6. PICS and parental control

As well-known, the **Platform for Internet Content Selection (PICS)**[7] is a metadata platform that was ideated in order to provide users with a standard metadata labeling services able to rate the WWW with respect to their "danger level" for kids, in various critical subjects like nudity, violence, language vulgarity etc. (see [10]). PICS-compliant metadata sets usually employ a crisp numeric scale to define the "level of safety" of a Web object. For instance, **SafeSurf**[8]'s **Internet Rating Standard**[9] employs a scale from 1 to 9 to denote the level of attributes like Age Range, Profanity, Violence, Sex and so on. **NetSheperd**[10]'s Community Rated Content (CRC) uses a five values scale from 1 to 5. **RSACi**[11], The Recreational Software Advisory Council on the Internet, employs a five values scale from 0 to 4 to rank attributes like violence, nudity, sex and language. All these crisp scales can be naturally mapped into a fuzzy attribute, just by normalizing them to stay in the range 0–1; for instance, in the RSACi case we would just divide by 4, obtaining that the corresponding fuzzy values are 0, e 0, 1/4, 1/2, 3/4, 1. As an aside, we note how all these cases also show how fuzzy metadata form a comprehensive unique platform encompassing many different ratings, each of them suffering from expressibility problems due to their crisp-valued character. Fuzzy metadata allow for wider expressibility, and are more elegant than ad-hoc scales too.

PICS-based metadata suffer even more from the wide applicability problem, since they essentially require a trustworthy third party (the *label bureau*) to produce the content rating (this is a de facto necessity, since user-defined local metadata are obviously too dangerous to be trusted in this context). And, indeed, already at the present moment there have been many criticisms on the capability of such third parties to survey a significant part of the Web (see e.g. [12]). The adopted solution is to limit Web navigation only to trusted refereed parts, but again this solution has raised heavy criticisms in view of the same argument as before, since this way a huge part of the Web would be excluded.

Therefore, we have employed the back-propagation method in order to enhance a publically accessible filtering system, namely the aforementioned NetSheperd. NetSheperd has acquired a somehow strategical importance in view of its recent agreement with **Altavista**[12] to provide filtered-based search. NetSheperd offers online its filtered search engine.

[6] http://searchenginewatch.com/

[7] http://www.w3.org/PICS

[8] http://www.safesurf.com/

[9] http://www.safesurf.com/ssplan.htm

[10] http://www.netshepherd.com

[11] http://www.rsac.org/homepage.asp

[12] http://www.altavista.digital.com

So, we have employed this tool as a way to access the corresponding metadata database, and generating metadata for a corresponding Web object.

The test was conducted as follows. First, we chose 10 Web objects that would be clearly be considered at a dangerous level by any rating system. The choice was done by taking the top 10 entries of the Altavista response to the query "nudity". Analogously to the previous experiments, we built from these Web objects the corresponding Web regions of size $n = 50$. Then, we used the NetSheperd public filtering system to provide suitable metadata for the Web objects. It turned out that 32% of the Web objects were present in the database as classified, while all the remaining 68% of Web objects were not. Soon afterwards, we run the back-propagation method on the Web regions. Finally, again, we asked the population of 34 users to judge (with $m = 20$) the obtained classifications in the perspective of a parental accessibility concern. The final average judgment was 96.4, and the minimum rating was 94: in a nutshell, all the generated metadata perfectly matched the expectations. It is interesting to note that the Web regions were not exclusively composed of inherently dangerous material: among the Web objects the users were asked to classify, a few were completely kids safe, and indeed they have been classified as such by the back-propagation.

The exceptionally good performance of the back-propagation in this case can be explained by the fact that we were essentially coping with an ontology composed by one category. In such situations, the performance of the back-propagation is very good, since the only uncertainty regards the fuzzy value of the attribute, and not the proper classification into different, possibly ambiguously overlapping, categories.

7. Conclusions

We have presented a new, easy method that is able to significantly enhance the utility of metadata, both static and dynamic. The proposed back-propagation technique relies only on the connectivity structure on the Web, and is fully automatic. So, it acts *on top* of any specific metadata format, and it does not require any form of ad-hoc semantic analysis, therefore being completely *language independent*. We have shown extensive testings of the technique, and also shed new light on the relationship between the percentage of metadata that need to be present in (a localized part of the) WWW, and the global usefulness of the metadata set, allowing to determine with a certain approximation the minimal *"critical mass"* that a metadata system needs to have in order to start being effective. We have also shown how the back-propagation method can be successfully employed to enhance automatic classification methods, and presented a real-life case study where the method has been applied to the important topic of PICS-based classification and parental control, obtaining striking performances. Finally, another advantage is the execution speed of the method. We have for space and clarity reasons passed over some technical implementation details that can be employed in order to make the back-propagation as fast as possible: anyway, by its same nature it is evident how back-propagation can be implemented employing massive parallelism, thus making this method utilizable with success also in time-critical and performance-demanding systems.

References

[1] Bezdek, J.C, Fuzzy models — what are they, and why, *IEEE Transactions on Fuzzy Systems*, 1(1): 1–6, 1993.

[2] Caplan, P., You call it Corn, we call it syntax-independent metadata for document-like objects (http://info.lib.uh.edu/pr/v6/n4/capl6n4.html), *The Public-Access Computer Systems Review* 6(4), 1995.

[3] Deatherage, M., HotSauce and meta-content format (http://www.ssrc.hku.hk/tb-issues/TidBITS-355.html#lnk3), *TidBITS*, 25 November 1996.

[4] Dublin Core Metadata Element Set, 1997, http://www.oclc.org:5046/research/dublin_core/

[5] Hardy, D., Resource Description Messages (RDM) (http://www.netscape.com/people/dhardy/rdm.html), July 1996.

[6] Lagoze, C., The Warwick framework (http://www.dlib.org/dlib/july96/lagoze/07lagoze.html), *D-Lib Magazine*, July/August 1996.

[7] Library of Congress, Machine-readable cataloging (MARC), http://lcweb.loc.gov/marc/

[8] Marchiori, M., The quest for correct information on the Web: hyper search engines, (http://proceedings.www6conf.org/HyperNews/get/PAPER222.html), in: *Proc. of the 6th International World Wide Web Conference (WWW6)* (http://proceedings.www6conf.org/, Santa Clara, California, U.S.A., 1997, pp. 265–276; also in: *Computer Networks*

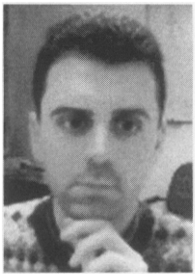
and ISDN Systems, 29: 1225–1235, 1997.

9] Miller, P., Metadata for the masses (http://www.ukoln.ac.uk/ariadne/issue5/metadata-masses/), *Ariadne*, 5, September 1996.

0] The World Wide Web Consortium (W3C), Platform for Internet content selection, http://www.w3.org/pub/WWW/PICS/

1] Web Developers Virtual Library, META tagging for search engines, http://WWW.Stars.com/Search/Meta/Tag.html

2] Weinberg, J., Rating the Net (http://www.msen.com/~weinberg/rating.htm), *Communications and Entertainment Law Journal*, 19, March 1997.



Massimo Marchiori received the M.S. in Mathematics with Highest Honors, and the Ph.D. in Computer Science with a thesis that won an EATCS (European Association for Theoretical Computer Science) best Ph.D. thesis award. He has worked at the **University of Padua**[13], at **CWI**[14], and at the **MIT Lab for Computer Science**[15] in the **Computation Structures Group**[16]. He is currently employed in **the World Wide Web Consortium (W3C)**[17], at **MIT**[18]. His research interests include World Wide Web and intranets (information retrieval, search engines, metadata, Web site engineering, Web advertisement, and digital libraries), programming languages (constraint programming, visual programming, functional programming, logic programming), visualization, genetic algorithms, and rewriting systems. He has published over thirty refereed papers on the above topics in various journals and proceedings of international conferences.

[13] http://www.unipd.it

[14] http://www.cwi.nl

[15] http://www.lcs.mit.edu

[16] http://www.csg.lcs.mit.edu

[17] http://www.w3.org

[18] http://web.mit.edu