Journal of Universal Computer Science, vol. 13, no. 9 (2007), 1234-1245
submitted: 6/7/06, accepted: 26/9/07, appeared: 28/9/07 © J.UCS

Creating Links into the Future

Muhammad Tanvir Afzal
(Institute for Information Systems and Computer Media (IICM)
Graz University of Technology, Graz, Austria
mafzal@iicm.edu)

Narayanan Kulathuramaiyer
(Institute for Information Systems and Computer Media (IICM)
Graz University of Technology, Graz, Austria and
Faculty of Computer Science and Information Technology
University Malaysia Sarawak, Kota Samarahan, Malaysia
nara@iicm.edu)

Hermann Maurer
(Institute for Information Systems and Computer Media (IICM)
Graz University of Technology, Graz, Austria
hmaurer@iicm.edu)

Abstract: We are approaching an era where research materials will be stored more and more as digital resources on the World Wide Web. This of course will enable easier access to online publications. As the number of electronic publications expands, it will, however, become a challenge for individuals to find related or relevant papers. Related papers could be papers written by the same team of authors or by one of the authors, or even papers that deal with the same topic but were written by other authors. This, of course, raises the issue of linking to papers forward in time, or as we call it "links into the future". To be concrete, while reading a paper written in the year 1980, it would be nice to know if the same author has written another related paper in 1990's or if the same author has written a paper earlier, all this without making an explicit search. Based on the ascertained interest of a person reading a particular paper from a digital repository, an auto-suggestion facility could be useful to indicate papers in the same area, category and subject that might potentially be of interest to the reader. One is typically interested in finding related papers by the same author or by one of the authors of a paper. This feature can be implemented in two ways. The first is by creating links from this paper to all the relevant papers and updating it periodically for new papers appearing on the World Wide Web. Another way is by going through the references of all papers appearing on the WWW. Based on the references, one can create mutual links to the papers that are referred to.
In this paper, we focus on offering personalised services beyond standard global access. We explore means of identifying the relevance (or relatedness) of papers. A related paper can mean different things to different people as explained above. Ideally, related papers are found and made accessible using links into the future that could be customised to suit the needs of individual users. In this paper, we will focus on a subset of the problem. We explore links into the future in the context of a particular journal which has existed for the past 13 years with over 1500 published papers. We discuss problems that arise in this restricted context while providing details of partial implementations. We plan to pursue our ideas in a more general setting in future implementations.

Keywords: typed-link, annotations, citation, citation index, links into the future, similarity
Categories: H.1, H.3, H.4, H.m, K.6, K.m

1 Introduction

The number of digital publications is growing exponentially. Users expect to instantly get access to information relevant to them. The management of such digital publications has to take into consideration the anticipated needs of users in providing highly customized services. It is currently possible to locate a paper in a previously published journal for a relevant research area, if a citation has been made explicitly. However, the user will not be able to view future relevant works from within the same paper based on a citation list alone. In order to achieve this, a user has to go through a citation index [Citeulike, 2006] [Citeseer 2006] [DBLP 2006] and then find a future paper that has cited the former. We explore the possibility of producing a shortcut for the user, to enable links into the future to be accessed from within the paper itself. This has been achieved by employing typed-linking technology in the context of digital journals [Maurer 1996]. We also explore additional futuristic and innovative features in the Journal of Universal Computer Science (J.UCS). Readers can use some of those features since they are available online with a button "links into future" at http://www.jucs.org.

In the past, an initial attempt has demonstrated the idea of links into the future to some extent [Krottmaier 2003]. This concept has, however, only been partially realized so far. In this paper, we shall extend the idea further by providing details on its full realization and its implementation. We explore the discovery of links into the future to be incorporated within a previously published paper in the Journal of Universal Computer Science [J.UCS 2006]. J.UCS will be unique in being the first electronic journal to implement this idea for enhancing a user's ability to gather future information on published papers in the same area. A published contribution is typically a static document; an author is not allowed to add or edit the published work. By implementing this idea, published papers will not need to remain as static documents, as it becomes able to record new related developments as they get published. The previously published paper in the same area will get a link to the new paper as well. What we have just mentioned is part of our vision of dynamic publications in a modern digital library [Dreher, Krottmaier and Maurer 2004]. Note, however, that the body of a published paper is never changed; only notes in the sense of links to other publications are made available, additionally.

The Journal of Universal Computer Science (J.UCS) is a high-quality electronic publication that deals with all aspects of Computer Science. J.UCS has been appearing monthly since 1995 with uninterrupted publications [J.UCS 2006]. According to the survey paper on electronic journals [Liew, and Foo, 2001], J.UCS has incorporated innovative features such as the enabling of semantic and extended search and its annotative and collaborative features. It was one of the first electronic published journals to have implemented features such as personal and public-annotations, multi-format publications, multi categorization, etc. These features have made J.UCS a rather unique electronic journal. Readers of such high-quality electronic journals expect and anticipate highly sophisticated features, such as automatic reference analysis, similarity search between documents and other features using knowledge management technology [Krottmaier, 2003]. A team is currently working on including many novel ideas into J.UCS. We will report on one such idea here. See [http://www.jucs.org] for more details.

While writing a research publication, researchers can cite any previously published paper. But of course they can not cite future contributions in the same area. While reading a paper, a reader may overlook papers published later by the same authors dealing with the similar topic. The readers would then need to make a deliberate effort to access the citation index [Citeulike 2006] [Citeseer 2006] to access later publications. But there are other limitations to the user of such a citation index: We illustrate the potential problem with one example: An author has published a paper "A" in the past, and subsequently published two other papers "B" and "C" in the same area. In paper "B", the author cited the paper "A" but in "C" the author did not cite the paper "A". While reading a paper "A", if one wants to see future relevant publications, one goes to the citation index where one would be able to find paper "B", but would not find the paper "C" which may be more relevant to paper "A". Thus, by using the citation index, a reader is able to find relevant papers only to a certain extent depending on how many references have been provided by authors. This has led us to explore the idea of incorporating links into the future within a paper itself. This idea creates two benefits: the user will have a shortcut to access future work from within the paper and it will make sure that the reader gains access to the all research papers published in the future in the same area. This will make the papers dynamic in the sense that readers may be able to see all potentially similar publications in the same area for a particular publication.

In this paper, we consider papers to be related when one of the authors is similar and papers have appeared in the same journal. Of course it would be nice to also find papers that are written by different authors on the same topic in other journals. However, this a way of managing knowledge available in dispersed form is clearly much more difficult to handle, hence the current restriction..

As illustrated by [Dreher, Krottmaier and Maurer 2004], another limitation of typical citation systems is that they are constrained by the implementation of unidirectional links. In such systems when a document "A" cites document "B", there is a link between A to B. There is however no link specified between "B" to "A". Thereby the reader of article "B" may not know of "A". One can claim that it is possible to create automatically two links: from article "A" to article "B" and vice-versa. However, this is very often not possible as 'write access' would be needed at the system where article "B" was published to create a new link. As J.UCS has been put on a Hyperwave Information Server [Hyperwave 2006] as its publishing system, it can facilitate the implementation of this idea of bidirectional links. A link-database is used in Hyperwave, i.e. links are stored separately from contents. No write-access restriction is imposed to create links to a document [Dreher, Krottmaier and Maurer 2004].

2 Related work

The third author of this paper has given a presentation on "Beyond Digital Libraries" [Maurer 2001] at Beijing. It describes new innovative ideas that should exist in modern digital libraries. It conceptualizes the idea of "links to the future" and describes how this feature can benefit existing digital libraries.

User expectations for a news aggregator have been surveyed in [Chowdhury and Landoni 2006]. News aggregator is a client which provides information to a user by

using user profiles. It has been found in this study that a user wants advanced search facilities for acquiring some task oriented information. Users typically require automated feedback without the need for a deliberate effort. In the context of links into future, when users are reading a particular article, they would then be furnished with links to future related papers written by the same team of authors in the same area. Context-specific task oriented information is thus directly made available to the user.

Andrei Broder [Broder 2006] has given an IEEE talk at Stanford University about the future of search. He described how future search will evolve from information retrieval to information supply. He explained that in the future, users would not need to perform search by typing keywords in a text box. Instead, the user would be given information based on his task profile. Our idea of links into future implements this concept in a way that enables a user to see related contributions that may even be created after the original document was published. Links into the future are pushed to the user without an explicit request made by the user.

Important aspects of modern digital libraries have been discussed in [Maurer, Krottmaier and Dreher 2006]. This paper describes intelligent and conceptual search including results visualization, white lists, and adaptive user interfaces. "Links into the Future" can serve as intelligent context aware conceptual search. Users can get the most relevant take oriented papers by using this novel idea.

3 The process of creating links into the future

Each J.UCS paper comes with a set of topics which describe the area of the paper. The topics are basically the categories of the ACM classification [ACM Classifications 1998] (J.UCS has explicit permissions to use them) with minor extensions reacting to the growth of the field. A paper may belong to more than one topics like: D.2.1, H.5, K.2.

The process of creating links into the future can be summarised as follows:

1. For a particular paper which one is interested in, identify target items to be considered as potential links into the future. This will require the retrieval of all other papers of authors and co-authors. In order to validate the relevance of a future link, we check the mutual relatedness of topics and date of the publications.
2. Each of the above target items (potentially related papers) will then be verified and validated to establish links into the future.
3. Incorporate the links into the journal and make them accessible to readers. This step involves the determination of an effective way to incorporate the links into the system.

This process will be described in more detail in section 4.1.2.

4 Establishing links into the future

In this paper we have restricted our focus to J.UCS to demonstrate the realisation of links into the future. In future efforts, we will explore the possibility of applying this concept for a wider range of scholarly publications on the WWW.

We will describe two different techniques to establish links into the future. One possibility (we will refer to it as citation mining technique) would be to look at the references of a paper, as proposed by [Krottmaier 2003] and check whether the list of papers contains publications in J.UCS. For a paper found in J.UCS, there is a great likelihood that the cited work (by the author) is of the same topic. As such, links are created from the former to these future papers. Citation indices [Citeulike, 2006], [Citeseer 2006] may also perform a similar operation. The citation mining technique [Krottmaier 2003] will also place future relevant links within the paper for future usage. This, however, cannot be achieved via a citation index alone.

Another possible approach (we will refer to it as metadata extracting technique) is to examine a particular paper and its authors and check for other papers in J.UCS with the same author or one of the authors. In this case, we need to explicitly ensure that both papers are in the same field or same topic in J.UCS. After validation, links to these future papers are created. This is not done by citation indices [Citeulike 2006, [Citeseer 2006] in a direct way.

4.1 Citation mining technique

To identify links into the future for a paper "A", we search all the papers that have cited the paper "A" and have been published after paper "A" in J.UCS. These papers are considered to have a link with paper "A", and we identify and locate the related documents for linking.

This approach explores a paper from the J.UCS server and checks the reference section of a paper. Each entry in the reference section begins in a new line with [name, years], authors names and then title of the paper in double quotes.

This effort to associate links to documents to make them more readable may be compared to digitisation (and linking) efforts of Citation Indices in providing links directly from their citation list to cited publications. This feature is extremely useful in helping researchers finding relevant papers. As such our efforts to incorporate such a feature within papers in a journal would also benefit readers to enable them to become aware of new development and updates.

The J.UCS server maintains the publications in multiple formats like postscript, PDF and HTML. In our experiments, we employ the PDF version of a document and a conversion from PDF to text is performed to extract the text. The citation section of the paper is then parsed to identify links. Our approach does not change the original publications. It merely augments the meta-data of the publication to provide additional information about other published papers.

4.1.1Problems associated with this technique

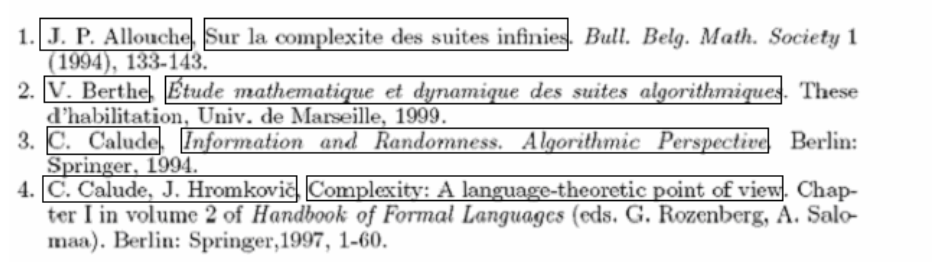
The style guide following the one used in Springer Journals [Springer 2006] for publication describes the standard form of stating references as explained below:

[Authors_last_ name year_ of_ publication] Last_name, first_ name. "Title of the paper"; publisher, page no.

Although a majority of published work has complied with this reference specification format, there are authors who have not followed the rules specified in the style guide, and in some cases this was not detected in the editorial process. This also seems to imply that the editorial process should be improved to detect such cases.

Because of this, the extraction of author's name and title of the publication from a reference section of a paper may not be simple at all. Figure 1 illustrates an example of a citation that may not be deterministically parsed to extract publication details. (see Figure 1)

There are other examples where the text parser may not be able to extract the precise authors' names due to the missing standard delimiter after the names of authors, the period.



1. J. P. Allouche, Sur la complexite des suites infinies. *Bull. Belg. Math. Society* 1 (1994), 133-143.
2. V. Berthe, *Étude mathematique et dynamique des suites algorithmiques*. These d'habilitation, Univ. de Marseille, 1999.
3. C. Calude, *Information and Randomness. Algorithmic Perspective*. Berlin: Springer, 1994.
4. C. Calude, J. Hromkovič, Complexity: A language-theoretic point of view. Chapter I in volume 2 of *Handbook of Formal Languages* (eds. G. Rozenberg, A. Salomaa). Berlin: Springer,1997, 1-60.

Figure 1: Sample Non compliant references

As problems can be encountered while extracting the metadata with this technique due to non-compliant referencing formats, we face difficulties in locating the right resources for linking in all citation lists. This approach has thus been reported to have produced discouraging results in [Krottmaier 2003]. For this reason an improved citation mining technique will be required. This is quite a challenging task. For the time being we describe in the next section the metadata extraction approach.

4.2 Metadata extracting technique

When authors submit papers to J.UCS for publication, metadata (in XML) files are created as they upload a paper. This file is maintained in a hierarchical representation of Volumes and Issues. This file stores the information on names of authors, submission-date, acceptance-date, title of the paper etc. For exploring links into the future, the attributes title, authors, date and topic need to be examined. These metadata or attributes are shown in Figure 2. We have written an XML parser that parses these XML files to populate our database in proper order. We then search all the papers of the authors in the same topic (including later published papers) and create links from a paper to the selected papers.

Figure 2: Metadata XML File

The metadata extracting technique can be described as follows:

Select Candidates as potential links (into the future)

a) Select a paper to be considered for creating links into the future.
b) Find references to authors and co-author's names from the entire list of publication in the metadata file. Extract the entries that contain their names.

Links verification and validation

c) Validate an author's publication (as relevant and from the future) by examining metadata such as date and topic of each entry extracted in (b)
A publication is considered a link into the future if:
   i. The age of publication is less than original document of source and
   ii. The document has the same topic.

We suggest that the use of document similarity checking as a means of finding relevant documents should also be investigated. A user profile will be maintained for all users, to be able to allow the visualisation of types of links (to the future) the user wants to see.

Realisation and incorporation of links

d) Construct an internal representation to highlight all discovered information about the author. We have developed a publication ontology (see Figure 3) which will represent currently known information about authors and their publications, together with information about discovered links. As new issues are published, these ontologies are examined and updated accordingly, instead of repeating the metadata extraction all over.

e) Perform visualization of the discovered links to be incorporated into the system.

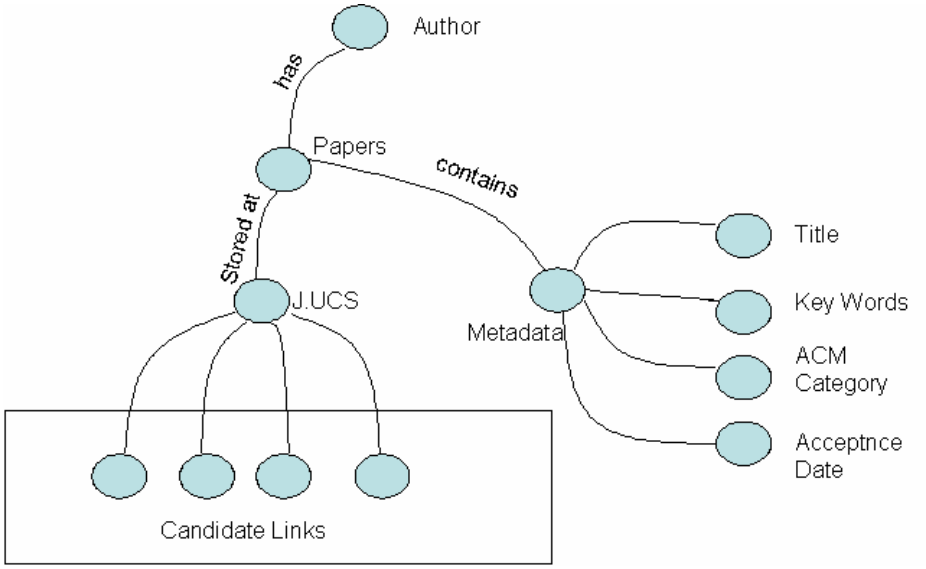


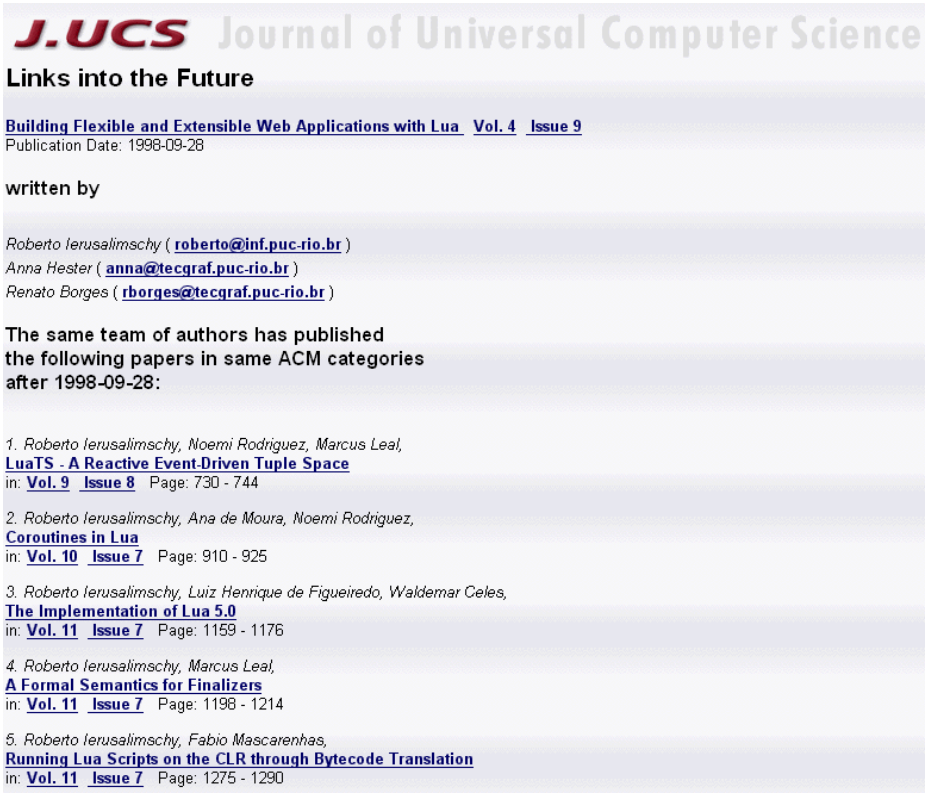
Figure 3: Publication Ontology

4.2.1 Update problems

The execution of the metadata extracting technique has to be performed incrementally to ensure that all future links are discovered. Since this is not a static repository, either a periodic bulk update or a regular update when new papers comes in, has to be performed. The current implementation of the technique has created future links for all papers published until volume 13 issue 5 and monitors every new paper that comes into the system and creates future links into the existing relevant papers pointing to the new paper as soon as new paper is published.

5 Realisation and incorporation of future links

We have discussed in the previous section the identification and validation of the candidate documents to be linked. In this section, we are going to talk about realisation and incorporation of the concept of links into the future. Here we will discuss some of the results produced by our system. On the first page of a paper in J.UCS, we have introduced a button titled "Links into Future". When a user is viewing some particular paper and wants to see related future papers, the user simply clicks the button and all related future papers for the same team of authors in the same topic are shown.


J.UCS Journal of Universal Computer Science

Links into the Future

Building Flexible and Extensible Web Applications with Lua Vol. 4 Issue 9
Publication Date: 1998-09-28

written by

Roberto Ierusalimschy (roberto@inf.puc-rio.br)
Anna Hester (anna@tecgraf.puc-rio.br)
Renato Borges (rborges@tecgraf.puc-rio.br)

The same team of authors has published the following papers in same ACM categories after 1998-09-28:

1. *Roberto Ierusalimschy, Noemi Rodriguez, Marcus Leal,*
LuaTS - A Reactive Event-Driven Tuple Space
in: Vol. 9 Issue 8 Page: 730 - 744

2. *Roberto Ierusalimschy, Ana de Moura, Noemi Rodriguez,*
Coroutines in Lua
in: Vol. 10 Issue 7 Page: 910 - 925

3. *Roberto Ierusalimschy, Luiz Henrique de Figueiredo, Waldemar Celes,*
The Implementation of Lua 5.0
in: Vol. 11 Issue 7 Page: 1159 - 1176

4. *Roberto Ierusalimschy, Marcus Leal,*
A Formal Semantics for Finalizers
in: Vol. 11 Issue 7 Page: 1198 - 1214

5. *Roberto Ierusalimschy, Fabio Mascarenhas,*
Running Lua Scripts on the CLR through Bytecode Translation
in: Vol. 11 Issue 7 Page: 1275 - 1290

Figure 4: Discovered Future Links

For example, a reader is viewing a paper in Volume 4, Issue 9 and the paper title is "Building Flexible and Extensible Web Applications with Lua". The user clicks the button "Links into Future" and is shown all the discovered future links in a screen similar to Figure 4. This paper was written by three authors. It was published on September 28, 1998. This paper was considered by the authors as belonging to topics D.2 and H.5. If any of the three authors has written a paper in the same topics (like D and H (we have taken only first level of the topics to get a maximum number of

related papers) after or on September 28, 1998, those papers are shown to the user as Figure 4.

6 Discussion

The citation mining technique is able to present valuable information by showing links considered to be relevant by the author while publishing. It may be used to provide directions for further exploration because citations may contain links to the other journals. The metadata extracting technique, however, does not take into consideration the information stored in the cited papers. The metadata extracting technique may in future be enhanced by incorporating user specified references.

The advantage of the metadata extracting technique is that it does not depend on the correct formatting of reference section as compared to the citation mining technique. Efforts in enforcing compliance need to be strengthened to further enhance the citation mining technique in the future.

In extracting the information from the reference section, we are not able to extract the first name, middle name and last name of an author accurately. This information will be useful in validating authors in finding their potential future papers. Using the metadata extracting technique, we are able to do that in a better way because we have tags of first name, middle name and last name of the authors in the XML file. There may also be situations where semantic information of publication may be required to validate an author.

By using the citation mining technique, we were able to extract information from the reference section. The author's decision not to cite a relevant paper in the past will lead to a paper being not represented in the future links section of some other paper. The Metadata extracting technique overcomes this problem.

In an XML file, standard tags provide us with additional information to acquire information and to validate the context. However, a number of variations have to be explored in concatenating the first and last name of an author in order to be able to discover all potential publications of an individual. It is possible that a query that does not specify the exact form of the author's name and hence may miss a publication. This problem will become more difficult when we are searching for an author's publication across the WWW. We propose the integration of a measure of document similarity checking to address this issue by being able to identify the publications of individuals, despite slight variations in author name specification.

7 Conclusions

We have introduced and implemented a useful new feature within the context of a particular journal. Links to the past already exist in the form of citations. But the concept of "Links into the Future" is a new idea which opens more horizons for digital resources. We have illustrated this concept to animate static published contributions to automatically be linked to the previously or later published papers of the same team of authors in a related area. We will explore the expansion of this feature to also find papers for the same area that are written by other authors. The metadata extracting technique for J.UCS is able to support the realisation of links into

the future. Users are encouraged to browse J.UCS e.g. “Software patents and the Internet” to see some of the links into the future that have been created by our system.

8 Future Work

We have restricted the scope of this research to the Journal of Universal Computer Science (J.UCS). Within J.UCS, we are also incorporating additional functionalities such as similarity checking tools to enhance the performance of the proposed technique in checking for similarity between publications.

A future enhancement to this work would be made to have “Links into the Future” for other journals. We are currently exploring the use of metadata from citation indices to achieve this. We can then minimize the effort of a user to search from the citation index to become aware of possible relevant works. We have also not discussed having links into the future for printed versions of the document. Some suggestions have been made by Krottmaier which can be further extended to create links for the printed version.

References

[ACM Classification 1998] http://www.acm.org/class/1998/

[Broder, 2006] Broder A.: (2006) IEEE Presentation, "The next generation Web Search: From Information Retrieval to Information Supply", Talk on Nov 16, 2006 at Stanford University, see http://www.cs.sjsu.edu/~tylin/ieeesilicon/broder.html

[Citeseer 2006] http://citeseer.ist.psu.edu/

[Citeulike 2006] http://www.citeulike.org/

[Chowdhury and Landoni 2006] Chowdhury, S. and Landoni, M.: “News aggregator services: user expectations and experience” in Online Information Review, vol. 30, No. 2 (2006), pp. 100-115.

[DBLP 2006]Computer Science digital repository. http://dblp.uni-trier.de/.

[Dreher, Krottmaier, and Maurer 2004] Dreher, H., Krottmaier, H., and Maurer, H.: “What we Expect from Digital Libraries“, Journal of Universal Computer Science Vol. 10, Issue 9 (2004), pp. 1110-1122.

[Hyperwave 2006] Enterprise content management solution, http://hyperwave.com/e/

[J.UCS 2006] Journal of Universal Computer Science, http://www.jucs.org

[Krottmaier 2003] Krottmaier, H.: “Links to the Future”, Journal of Digital Information Management Vol. 1, No. 1 (2003), pp. 3-8

[Liew, C.L. and Foo, S. 2001] Liew, C. L., Foo, S. (2001): ”Electronic Documents: What Lies Ahead?”, Proc. 4th International Conference on Asian Digital Libraries (ICADL 2001), Banglore, India, December 10-12 (2001), pp 88-105.

[Maurer 1996] Maurer, H.: "Hyperwave- The Next Generation Web Solution", Addison Wesley Pub. Co. (1996)

[Maurer 2001] Maurer, H. “Beyond Digital Libraries”, Global Digtial Library Development in the New Millenium (Proceedings NIT Conference), Beijing, Tsinghua University Press (2001), pp.165-173.

[Maurer, Krottmaier and Dreher 2006] Maurer, H., Krottmaier, H., and Dreher, H.: “Important Aspects of Modern Digital Libraries“, Proc. ICDL, New Delhi, India (2006), pp. 843-855.

[Springer 2006] http://www.springer.com/