DISCOVERING LINKS INTO THE FUTURE ON THE WEB

Muhammad Tanvir Afzal
Institute for Information Systems and Computer Media, Graz University of Technology
Infelldgasse 16c, 8010 Graz, Austria

Keywords: Personalized search, Links into the future, Information supply, Digital journals.

Abstract: Current search engines require the explicit specification of queries in retrieving related materials. Based on personalized information acquired over time, such retrieval systems aggregate or approximate the intent of users. In this case, an aggregated user profile is often constructed, with minimal application of context-specific information. This paper describes the design and realization of the idea of 'Links into the Future' for discovering related documents from the Web, within the context of an electronic journal. The information captured based on an individual's current activity is applied for discovering relevant information along a temporal domain. This information is further pushed directly to the users' local contexts. This paper as such presents a framework for the characterization and discovery of highly relevant documents.

1 INTRODUCTION

As the number of electronic publications expands, acquiring relevant information to suit particular user needs becomes a major challenge. The searching for related or relevant material has to duly consider the task at hand. This research then focuses on the identification of relevant documents within a scholarly publishing environment. As opposed to the retrieval of a million hits as provided by generic search engines, we explore the direct provision of relevant information only.

A related document could mean different things to different people; e.g. it could be a paper written in the same area, one written by the same author, or merely taking about the same research problem. We use the term 'Links into the Future' in this paper to describe the discovery of related papers based on the information compiled from a user context while reading a paper within an electronic journal.

For a reader of an electronic journal, 'Links into the Future' refers to potentially useful papers to the one that he or she is reading. This idea was originally proposed by Hermann Maurer in a presentation entitled "Beyond Digital Libraries" at the "Global Digital Library Development in the New Millennium" in Beijing, China (Maurer, 2001) and was subsequently partially realized (Krottmaier, 2003). In this paper, we consider a paper written by the same author or team of authors at a later time (then the current paper) to be related and useful, particularly if it was also in the same area of expertise. At the same time, any paper that cited this paper is also considered related and relevant.

A person reading a paper published in 1995 may not directly know if the same author has published a newer paper in year 2000 in the same area or if anyone has cited this paper. Well, this reader can of course, search the Web to look for similar papers. In this case, the reader may then loose the original context, and also get distracted by the millions of hits from global search engines. There are also times when particular related items may not be found as the formulation of exact query terms can in itself be a challenge.

Alternatively, the reader has an option to explore citation indices such as CiteSeer[1] CiteULike[2] and DBLP[3] to search for related papers. This will also require a deliberate effort on the part of the reader. Additionally, results from multiple sources of information will then need to be compiled and consolidated.

The implementation of such a feature within the context of the Journal of Universal Computer

[1] http://citeseerx.ist.psu.edu/
[2] http://www.citeulike.org/
[3] http://www.informatik.uni-trier.de/~ley/db/

Afzal M.
DISCOVERING LINKS INTO THE FUTURE ON THE WEB.
DOI: 10.5220/0001821701230129
In *Proceedings of the Fifth International Conference on Web Information Systems and Technologies (WEBIST 2009)*, page
ISBN: 978-989-8111-81-4

Science (J.UCS)[4] was described in (Afzal, Kulathuramaiyer & Maurer, 2007). This paper then explores the extension of 'links into the future' for papers published beyond the closed database of J.UCS. The discovery of related links within a closed database has only partially addressed user concerns, in providing starting points for further exploration. By extending the notion of links into the future to also cover relevant papers from the Web, a more comprehensive solution can be obtained. The links discovered as such will also be more reflective of the state-of-the-art in this field.

J.UCS, an electronic journal covering all areas of Computer Science with more than 1,200 peer reviewed research contributions, serves as an excellent platform for the discovery of related papers on the Web. Papers within J.UCS are categorized and tagged according to the ACM categories[5]. A number of additional categories have been recently added to J.UCS to reflect the development in Computer Science discipline. These extended categories would be referred to as topics in the rest of the paper.

2 DEFINITION OF LINKS INTO THE FUTURE

A future link from paper "a" to paper "b" (Future_Link (a,b)) exists, if a semantic relationship can be established between them. For example: if paper "b" is written by the same team of authors of paper "a" and the topics of both papers are similar, then paper 'b' is considered to be related to paper 'a'. Alternatively if there exists a citation from paper b to a, there is a highly likely relationship. Current systems tend to perform similarity matches without considering semantic similarity, based on the task characteristics. The equation 1 describes the definition of "Links into the Future".

(Authors (b) € Authors (a) $\wedge$ Topics (b) € Topics (a)) $\vee$ Citation (b, a) $\longrightarrow$ Future_Link (a, b) (1)

3 RELATED WORK

Most related work explores the servicing of users within a present-context, making use of limited information, captured in-vivo. Our work, on the other hand, describes the augmentation and annotation of documents created in the past with information that became available later. In this way, a research paper is not seen as a static document, but rather one that is constantly updated and kept up-to-date with relevant links.

A number of past studies make use of user's context and activity to provide them the most relevant information. For example, typical search engines return relevant results based on the small amount of information from user queries and a measure of web site popularity, rather than considering individual user interest and context (Speretta & Gauch, 2005). Spretta & Gauch employed user profiles based on user queries, search activities and the snippets of each examined result to refine search result rankings. With this context specific ranking of search results, an improvement of 34% in the rank order has been obtained

Rhodes and Maes (Rhodes & Maes, 2000) described a new class of software agents called Just-in-Time Information Retrieval Agents (JITIRs), which has an ability to proactively present potentially valuable information based on a person's local context in a non-intrusive manner.

Another related work pushes the most relevant Web URLs based on the user activity and context. User context is determined by examining active personal desktop documents (Chirita, Firan & Nejdl, 2006). Similarly by observing user activity and context while reading a particular article, our notion of 'Links into the Future' presents the most related papers of the same team of authors within a local context. This paper discus how this concept can be extended to the WWW as a mechanism for contextual information supply for academic publications along a temporal dimension.

Existing approaches for finding related papers uses citation analysis, text similarity, bibliographic analysis and context based relatedness. For example, CiteSeer has employed three methods for finding related papers a) word vectors b) string distance c) and citations (Giles, Bollacker & Lawrence, 1998). PubMed (PubMed, 2008) on the other hand computes the relatedness between two papers using text-based similarities between paper titles, abstracts, and assigned MeSH terms ("Medical Subject Headings", 2008). For the focused paper, PubMed provides a list of related papers according to their relatedness scores.

Ratprasartporn (Ratprasartporn & Ozsoyoglu, 2007) have made use of context (topics) of research publications to determine the related papers. An ontology consisting of terms were utilized as a context of publications. A publication is assigned to

[4] http://www.jucs.org
[5] http://www.acm.org/class/1998/

one or more contexts with the context represents the publication topics.

Digital libraries are traditionally built largely by a massive human effort. Example of these include INSPEC for engineering, PubMed for medicine and DBLP for Computer Science. Alternatively automated approaches are being employed to construct large citation indices. Examples of these efforts include CiteSeer and Google Scholar. The limitations of these automatic approaches are that human effort is often required in verifying entries in the index. Fully automated techniques have problems in disambiguating entries while traditional constructed digital libraries are limited in their number of scientific publication.

Google Scholar has mistakenly identified name of places as authors of scientific publications (Postellon, 2008). Although Google Scholar has improved gradually, it continues to find citations backward in time (Jacsó, 2008). Their index covers a large collection of peer-reviewed papers (Google Scholar, 2008). It however also considers false positive citations like citations to press releases, resumes, and links to bibliographic records for cookbooks (Price, 2004). CiteSeer claims that 80% of the publications titles can be extracted accurately while their word and phrase matching algorithm further has an error margin of 7.7% (Giles, Bollacker & Lawrence, 1998).

Our system is fully automated in extracting papers from the Web and from citations. It also computes a conceptually enhanced similarity score between a source paper and candidate future papers.

4 IDENTIFYING FUTURE LINKS FROM WEB

The APIs for Google[6] , Yahoo[7] and MSN Live[8] were used for the experiments[9]. The identification of future links from the web includes the following steps: query formulation, removing duplicates, filtering papers only, similarity algorithm and determining future links. The description of each is shown in Figure 1.

4.1 Link Extraction

When querying a search engine, the formulation of query terms strongly affects the results. SOAP APIs have been used by our Web search service to seek Web documents. In performing a search it was found that the use of all available semantic information was able to narrow down search space significantly. The effects of query formulation and choice of query terms is shown in Table 1.

Table 1: Query Formulation.

Query	Google Hits	Yahoo Hits	MSN Live Hits
Hermann Maurer	1,68,000	1,260,000	4,48,000
“Hermann Maurer”	25,600	92,800	27,000
abstract references "Hermann Maurer”	918	1,720	446
abstract references "Hermann Maurer” filetype:PDF	193	775	114

4.2 Removing Duplicates

As a further pre-processing step, duplicates are filtered reducing the results by more than 50%. Documents are then downloaded in parallel into java threads. The importance of removing duplicates is shown in Table 2.

4.3 Identifying Research Papers

Even by specifying document formats to be either PDF, Doc or PS and also explicitly querying with formulated query terms, the retrieved documents also contains:

I. Theses supervised by the author.
II. Curriculum Vitae, Home page and Business cards of the author.
III. Conference programmes where the author’s name was mentioned.
IV. Documents edited by the author.
V. Presentation files
VI. The author’s publication list.
VII. The author may be listed in the reference entries or in the acknowledgement section of a research paper.

As we are only interested in actual research papers at this point, a further filtering step was performed. This process is important in potentially automating the discovery of Web-pages and publication lists.

[6]http://code.google.com/apis/soapsearch/reference.html
[7] http://www.programmableweb.com/api/yahoo-search
[8]http://msdn2.microsoft.com/enus/library/bb266180.aspx
[9] The SOAP based search API has been used since Dec 5, 2006 with permission.

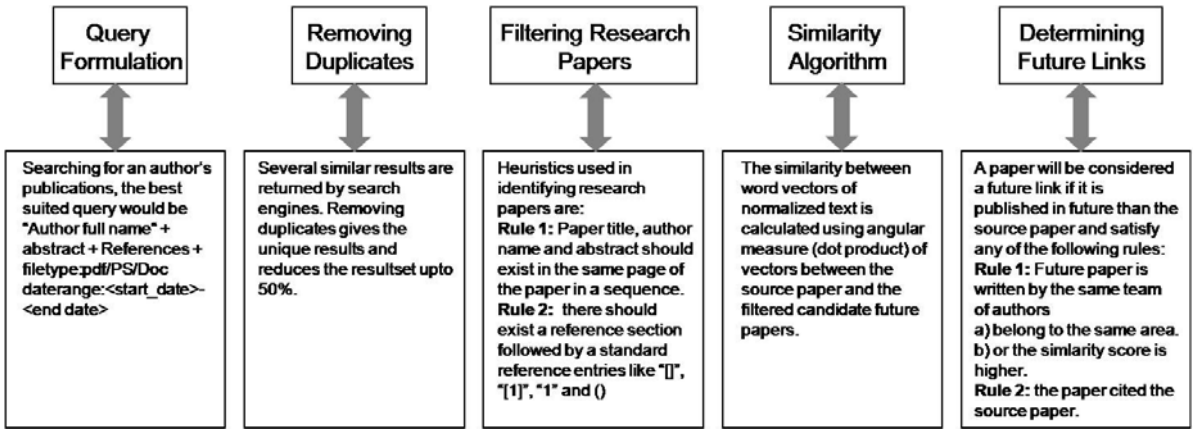


Figure 1: Rules for finding candidate future links from the Web.

Docments in PS and doc file formats are first converted to PDF using MiKTeX[10] and Openoffice tool[11] respectively. Then pdfbox[12], a java library is used to convert PDF to plain text for further analysis.

A heuristic approach is applied in the actual identification of research papers. The heuristic used are as follows:

1. Title of the paper followed by author name and abstract should exist in the same page. (need not be in the first page). Authors' full name is then searched to disambiguate author names.
2. Word "reference" or "references" is found followed by proper sequence starting with one of the them "[author]", "[1]", "1" and ().
Documents that were classified as research papers is shown in Table2 for selected authors.

4.4 Similarity Algorithm for Checking that Papers are in the Same Area

When we considered links into the Future for the articles published in Journal of Universal Computer Science only, the former implementation was straightforward. All articles are categorized according to ACM topics and metadata is already available. But when we locate papers from the Web, documents are not categorized according to the ACM topics and metadata cannot be expected to be found. We then performed similarity detection to automatically discover topics of documents.

We measured similarity by taking dot product of vectors from the source and the candidate paper. The results were, however, not satisfactory due to the following reasons: 1) Author's writing style was usually the same in his/her set of documents. A similar use of common terms produced impression of there being a larger similarity between documents 2) Paper's headers share the similar text such as author name, affiliation etc 3) The Reference List at the end of both documents make use of similar text.

To overcome these problems, we pre-processed the text removing the paper's header (section before abstract) and the reference section of the paper to focus only on the original text. We performed Yahoo Term Extraction[13] to extract key terms. This extraction scheme has been used in the number of past studies for extracting facet terms (Dakka, Dayal & Ipeirotis, 2006) (Dakka & Ipeirotis, 2008) and building expertise profile (Aleman-Meza, Decker, Cameron, & Arpinar, 2007). In our case, the results from Yahoo Term Extraction was seen to be not convincing until we removed the header and the references sections. The similarity measured on these terms was able to filter the most relevant papers as can be seen in Table 2 and Figure 4. For example, in Table 2, for the author "Vadim Bulitko", the relevant papers are only 3 out of 17 unique candidate papers found by the paper classification module. The manual inspection revealed that these three were the only papers in the same area.

4.5 Determining Future Links

A paper published later is considered a link into the future for a paper if it is also written by the same team of authors and is content wise in a similar area or cites the previously written paper.

[10] http://www.miktex.org/
[11] http://www.openoffice.org/
[12] http://www.pdfbox.org/

[13] http://developer.yahoo.com/search/content/V1/termExtraction.html

Table 2: Links into the Future results for selected authors.

Author	Focused paper in J.UCS	Search Engine	Formulated Query	After Duplicate Removal	Classified Papers	Unique paper	Actual Future Links
Maurer H.	Digital Libraries as Learning and Teaching Support vol. 1 Issue 11	Google	112	75	12	23	17
		Yahoo	495	86	19		
Abraham A.	A Novel Scheme for Secured Data Transfer Over Computer Networks Vol. 11 Issue 1	Google	148	62	13	33	22
		Yahoo	263	87	41		
Bulitko V	On Completeness of Pseudosimple Sets Vol.1 issue 2	Google	21	21	7	17	3
		Yahoo	45	28	13		
Shum S. B	Negotiating the Construction and Reconstruction of Organisational Memories Vol. 3 issue 8	Google	103	81	11	28	21
		Yahoo	546	104	26		
Abecker A.	Corporate Memories for Knowledge Management in Industrial Practice: Prospects and Challenges Vol. 3 Issue 8	Google	69	59	9	17	15
		Yahoo	335	65	14		

Papers determined to be relevant from multiple sources are then compiled and consolidated as annotations to each paper residing on the J.UCS server.

We have extracted all citations from papers published in J.UCS. These were 15,000 in number and links were created in the papers that have cited the J.UCS papers (Afzal, 2008). After extraction of future links for the focused paper, the future link ontology (Afzal & Abulaish, 2007) is updated (see Figure 2).

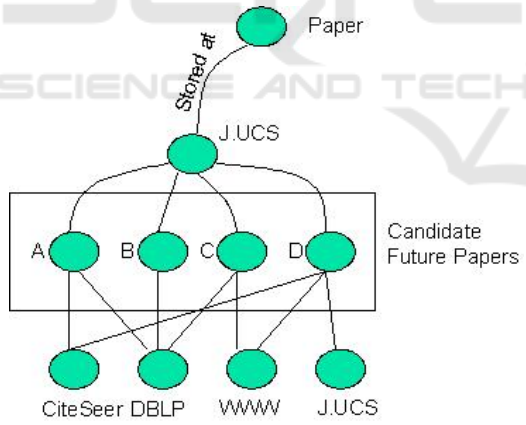


Figure 2: Future link ontology.

Figure 3 represents an example of a source paper and its candidate future papers. All of these candidates are acquired from Web by using SOAP APIs as discussed earlier. Candidates "C1, C2, C7, C11, C18" were published in J.UCS. The remaining 18 papers were published outside J.UCS. Figure 3 has been created by using "Graphviz" toolkit. The link distance between source "S_1" and candidate "C_n" node is inversely proportional to the term similarity. The figure is further annotated using key terms from the associated papers. Based on the visual representation it is possible to manually ascertain a threshold for candidate papers that belong to the same area. Threshold for this example has been represented by a dotted circle from source paper to the candidate future papers. In this way, it filters 17 papers out of 23. Here the source paper belongs to the topical areas of E-Learning, digital libraries and teaching support. It is obvious that the papers within the closed circle also belong to the topics of source paper. The threshold can be altered to refine the closeness of fit of target documents based on usage or application.

Figure 4 represents the user interface for this feature. The user viewing the source paper entitled "Digital Libraries as Learning and Teaching Support" at[14], clicked on "Links into Future" button and was shown the screen as in Figure 4. In Figure 4, the future links from J.UCS database (based on metadata similarity and citations) are consolidated with the future links extracted and filtered from web (as shown in figure 3). Readers are encouraged to explore this feature in Journal of Universal Computer Science (http://www.jucs.org).

This feature is currently fully implemented for the J.UCS papers and it suggests future related papers that are also published in J.UCS or cited in J.UCS papers. As we are also extending Links into the Future for documents published outside J.UCS, this prototype is being updated.

5 DISCUSSION

To evaluate the research, we have compared the citations extracted from our technique with citations extracted from CiteSeer. The dataset used for

[14]http://www.jucs.org/jucs_1_11/digital_libraries_as_learning

evaluation comprised of citations from J.UCS to J.UCS papers. We performed this experiment in the month of May, 2008. There were 92 unique papers that were cited by other J.UCS papers and numbers of citations for these 92 papers were 151. Citeseer indexed 67 papers (73%) out of 92 focused papers and citations found by CiteSeer were only 38 (25%) out of 151. It was due to citations that were in non-compliant formats in the original papers. Our local heuristics employed for J.UCS gave better results.

Our technique was able to disambiguate authors by looking for author's full name in the text of paper. This approach also avoids the mistaken identity of names of places as author of scientific publications as discussed earlier.

Table 2 highlights the retrieval results for Links into the Future that was determined from the Web for the paper "Digital Libraries as Learning and Teaching Support" published in J.UCS vol. 1 Issue 11.

When user performs a query on search engines, he/she is returned with millions of hits as can be seen in Table 1. The best formulated query for finding PDF/PS/Doc documents was applied to reduce the results to a few hundreds of documents.

The process of removing duplicates then reduced the number of documents by up to 50%. Our heuristic rules filtered 12 out of 75 as papers from Google's results and 19 out of 86 from Yahoo respectively. The similarity based on key terms was then further applied to select 17 out of 23.

Although a user has an option to explore citation indexes to search for related papers. But there are two issues 1) times when papers do not exist on these citation indexes like the source paper in our case study was not indexed by CiteSeer. While Google Scholar indexes it but suggests hundreds of related papers. 2) A deliberate effort is thus needed to find related papers outside the user's local context.

6 CONCLUSIONS

We have in this paper described the extension of the idea of links into the future to cover documents on the Web. The results are promising in providing candidates for future links. The key term similarity detection has further filtered the most relevant papers. As further works, we are currently developing a tool based on sentiment analysis of citations to evaluate the context of citations. We are also further exploring the discovery of future related papers from digital libraries like DBLP and CiteSeer.

REFERENCES

About Google Scholar, http://scholar.google.at/intl/en/scholar/about.html (accessed 23, May 2008).

Afzal, M.T. (2008). Citation Mining Technique for creating Links into the Future. submitted to International Journal on Digital Libraries.

Afzal, M. T., Abulaish, M. (2007). Ontological Representation for links into the Future. ICCIT Gyeongju-Korea, published by IEEE (CS).

Afzal. M. T., Kulathuramaiyer, N., & Maurer H. (2007). Creating Links into the Future. Journal of Universal Computer Science, vol. 13, issue 9.

Aleman-Meza, B., Decker, S.L., Cameron, D., & Arpinar, I.B. (2007). Association Analytics for Network Connectivity in a Bibliographic and Expertise Dataset. book chapter in Semantic Web Engineering in the Knowledge Society.

Chirita, Paul-A., Firan, C.S., & Nejdl, W. (2006). Pushing Task Relevant Web Links down to the Desktop. WIDM'06, November 10, 2006, Arlington, Virginia, USA.

Dakka, W., Dayal, R., & Ipeirotis, P. (2006). Automatic discovery of useful facet terms. ACM SIGIR Workshop on Faceted Search.

Dakka, W., Ipeirotis, P. (2008). Automatic extraction of useful facet hierarchies from text databases. Proceedings of ICDE.

Emamy, K., Cameron, R. (2007). CiteULike: A Researcher's Social Bookmarking Service. Ariadne, Issue 51.

Giles, C.L., Bollacker, K.D., & Lawrence, S. (1998).
CiteSeer: An Automatic Citation Indexing System. proceedings of Third ACM Conference on Digital Libraries, pp. 89-98.

Jacsó, P. (2008). Reference Reviews. http://www.gale.cengage.com/reference/peter/200708/SpringerLink.htm (accessed 23, May 2008)

Krottmaier, H. (2003). Links to the Future. Journal of Digital Information Management, Vol. 1, No. 1.

Maurer, H. (2001). Beyond Digital Libraries, Global Digtial Library Development in the New Millenium. Proceedings NIT Conference, 165-173.

Postellon, D. C. (2008). Hall and Keynes join Arbor in the citation indices. Nature, 452, 282.

Price, G. (2004). Google Scholar Documentation and Large PDF Files, http://blog.searchenginewatch.com/blog/041201-105511 (accessed 23, May 2008).

PubMed, http://www.ncbi.nih.gov/entrez/query.fcgi

Medical Subject Headings (MeSH), http://www.nlm.nih.gov/mesh/

Ratprasartporn, K., Ozsoyoglu, G. (2007). Finding Related Papers in Literature Digital Libraries. ECDL, LNCS 4675, pp. 271–284.

Rhodes, B.J., Maes, P. (2000). Just-in-time information retrieval agents. IBM Syst. J., 39(3-4):685–704.

Speretta, M., Gauch, S. (2005). Personilized Search Based on User Search Histories. IEEE/WIC/ACM International Conference on Web Intelligence.

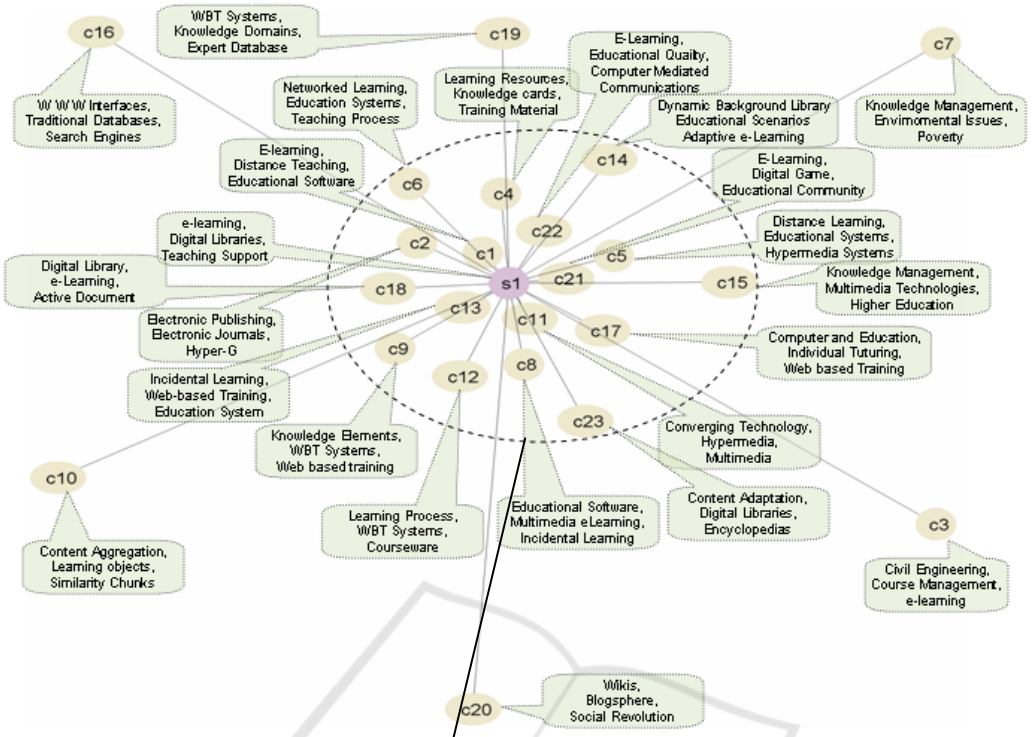
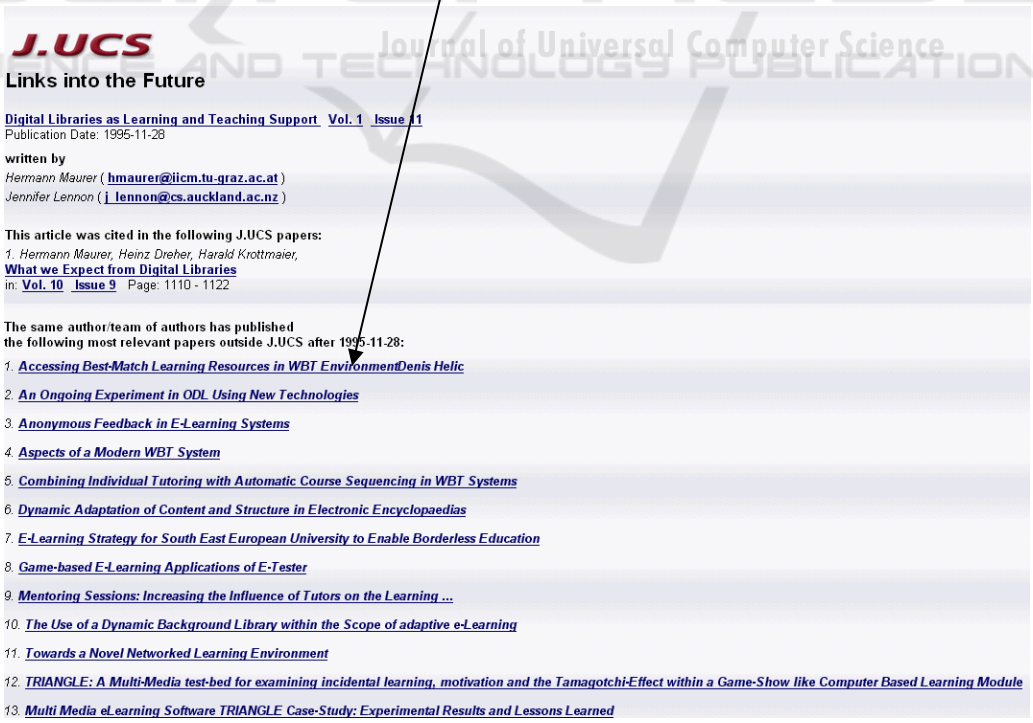


Figure 3: Similarity measure score for a source paper and its candidate future links.



J.UCS Journal of Universal Computer Science

Links into the Future

Digital Libraries as Learning and Teaching Support Vol. 1 Issue 11
Publication Date: 1995-11-28

written by

Hermann Maurer (hmaurer@iicm.tu-graz.ac.at)
Jennifer Lennon (j_lennon@cs.auckland.ac.nz)

This article was cited in the following J.UCS papers:

1. *Hermann Maurer, Heinz Dreher, Harald Krottmaier,*
What we Expect from Digital Libraries
in: Vol. 10 Issue 9 Page: 1110 - 1122

The same author/team of authors has published the following most relevant papers outside J.UCS after 1995-11-28:

1. *Accessing Best-Match Learning Resources in WBT EnvironmentDenis Helic*
2. *An Ongoing Experiment in ODL Using New Technologies*
3. *Anonymous Feedback in E-Learning Systems*
4. *Aspects of a Modern WBT System*
5. *Combining Individual Tutoring with Automatic Course Sequencing in WBT Systems*
6. *Dynamic Adaptation of Content and Structure in Electronic Encyclopaedias*
7. *E-Learning Strategy for South East European University to Enable Borderless Education*
8. *Game-based E-Learning Applications of E-Tester*
9. *Mentoring Sessions: Increasing the Influence of Tutors on the Learning ...*
10. *The Use of a Dynamic Background Library within the Scope of adaptive e-Learning*
11. *Towards a Novel Networked Learning Environment*
12. *TRIANGLE: A Multi-Media test-bed for examining incidental learning, motivation and the Tamagotchi-Effect within a Game-Show like Computer Based Learning Module*
13. *Multi Media eLearning Software TRIANGLE Case-Study: Experimental Results and Lessons Learned*

Figure 4: Links into the Future interface.