Lexical Profiling of Existing Web Directories to Support Fine-grained Topic-Focused Web Crawling

Mark Greenwood, Goran Nenadic
School of Computer Science, University of Manchester, Manchester, UK
greenwom@cs.man.ac.uk, G.Nenadic@manchester.ac.uk

Topic-focused Web crawling aims to harness the potential of the Internet reliably and efficiently, producing topic specific indexes of pages within the Web. Previous work has focused on supplying suitably general descriptions of topics to generate large general indexes. In this paper we propose a method that uses lexical profiling of a corpus that consists of hierarchical structures in existing Web Directories to specify finer-grained topics on smaller training examples, while using the seemingly redundant information in related topics to make the process of gathering pages more efficient. We also suggest a link scoring formula that combines content, context and page lexical similarities to a given topic to prioritise the links for crawling. The initial experiments with the Open Directory Project show that the prioritised crawl provides significantly more pages than the breadth-first crawler. Also, the rate at which the number of relevant pages increases is much higher. Keeping the crawler close to the target subject allows "unproductive" periods to be reduced, by following links most likely to link to target pages.

Keywords: Topic Focused Web Crawling, World Wide Web, Information Retrieval, Lexical Profiling

1. INTRODUCTION

The ultimate aim of Internet search engines is to index the entire Web by utilising the links found within known pages. Topic-focused crawlers specialise this task by indexing only the subset of the Web which is relevant to some topic or information need (Chakrabarti et al. 1999b; Cho et al. 1998). Typically, the user can specify the topic and a set of start (seed) pages, and the topic-focused crawler will traverse the Web via hyperlinks, but will make relevance decisions at each step as to whether this page should be included in the resultant index. This sort of system may be useful as part of a larger Information Retrieval system, or to generate topic-specific indexes for search by users who are only interested in one area. This will typically mean that search results would contain less noise than those from a general index (because of polysemy) and will also make the task of keeping the index up to date less expensive as there will be fewer pages to re-visit (Dilligenti et al. 2000).

There are, however, two well-known issues with topic-focused Web crawling: making the quality and efficient relevance decisions at each stage, and collecting the most relevant documents while visiting the least amount of irrelevant ones. For the first issue, finding a suitable method of representation and assessment of the topic is a difficult and challenging task (Chakrabarti et al, 1999b), in particular when the size and diversity of the Web is considered. A further problem is estimating not only whether the next page to crawl is relevant, but whether it might link to a relevant page further down the crawling process. Some irrelevant pages i.e. links may have to be visited before finding a correct link, but to get towards the ideal crawl, methods to estimate whether an irrelevant page might contain a link to a relevant one before downloading it would be useful.

The problem of finding appropriate descriptions of target documents, including information on how they may be linked within the Web is the main focus of this paper. More precisely, we propose a method that uses the structure in existing Web directories (e.g. Yahoo Directory[1], Open Directory Project[2]) to lexically profile and specify finer-grained topics on smaller amounts of training examples, while using the seemingly redundant information in related topics to make the process of gathering pages more efficient. Each level in a Web directory is represented by a set of automatically recognised terms, which are then lexically profiled using all substring combinations they contain. These profiles are used for comparing crawl pages and their links to the target topic. We propose a link ranking algorithm that combines link's context, content and document relevance to prioritise the crawling of more relevant links.

[1] http://dir.yahoo.com
[2] http://www.dmoz.org

The paper is organised as follows. In Section 2 we overview the related work. Section 3 represents the methodology designed for the task. The results and discussions are presented in Section 4, while Section 5 concludes the paper and gives an outline of topics for future work.

2. RELATED WORK

Many different methods of representation and description of target documents have been used throughout work on topic-focused Web crawling. One of the first such approaches was FishSearch (De Bra and Post, 1994) which was driven by a keyword query. More comprehensive descriptions were used in later crawlers where the user usually had to specify a number of target documents, so the crawler could return ones which were similar (Chakrabarti et al. 1999a; Chakrabarti et al. 1999b; Pant and Srinivasan, 2005). This similarity comparison usually meant employing machine learning techniques to train a classifier to discriminate between relevant and irrelevant documents, using a large set of training documents, both positive and negative.

The ranking of individual links within a page has been approached in many ways for general Web crawling, based on various features related to them. Several popularity based methods have been used, such as PageRank (Page et al. 1998) and HITS (Kleinberg 1998). These measures look at the number of times pages are cited in different documents, and also at what sort of documents link to them (in terms of their own connectivity). This means that the links which are judged more popular (or more authoritative) are considered the better possible sources of information. These methods have been also used for topic-focused crawling: Chakrabarti et al. (1999a) applied the HITS based method, while PageRank was used by Cho et al. (1998).

Furthermore, anchor text which is used to display the link in the web page, and the context it appears in has also been utilised to judge its relevance to the driving topic (Cho et al. 1998). The anchor text will often be a word or phrase which describes the content of the page being linked to. It may also be a part of a larger phrase which describes the page. This means that the way the outgoing link is displayed, and its context may offer more clues as to whether the page being linked to will be relevant or irrelevant.

The neighbourhood that target examples occur in (i.e. pages that link to a certain page and pages which are linked to by it) has also been used to train classifiers to rank pages. For example, Diligenti et al. (2000) use a technique called backwards crawling, where an external search engine is used to find pages which link to target pages. This means that the crawler can build a picture of the usual/typical neighbourhoods the target pages will be found in, and then whether pages found during a crawl fit in to those profiles or not. The training phase here is expensive and relies on having a pre-compiled index of pages which link to the target page.

Collecting training data is always a problem for topic-focused crawlers. One source that has been used is the Open Directory Project (ODP). This is a publicly available, manually compiled Web directory, offering lists of pages on various topics. The topics are arranged in a taxonomy, where the topic being described becomes more specific the more levels there are (e.g. Arts/Movies/Awards/Academy-Awards). Under each heading there would be a list of pages pertaining to the topic described at that level. The correct heading will be decided by one of the editors who manually categorise pages under the most fitting category.

However, topic-focused Web crawlers will generally be employed to gather pages on very broad topics and therefore the top levels would be used. Typically, to generate a large enough training set, the levels below the target levels will be concatenated up as examples of the broader topic, meaning that the classifier can recognise pages on many (related) sub-topics but cannot differentiate between them. Negative examples will be typically extracted from other non-related topics within the directory.

Our proposal is to address finding suitable training data and methods to classify pages and links found during the crawl using lexical profiling of a fine-grained topic hierarchy. The fact that the taxonomy not only supplies example documents of specific well-defined topics, but also related topics (the levels above in the taxonomy) allows us to classify documents not only as relevant/irrelevant but also as “related” to the topic. This offers new opportunities to classify pages as part of a broader spectrum.

3. METHODOLOGY

In this section, the methodology proposed for profiling and modelling each level of a taxonomy will be described including a description of the profiling the taxonomical levels, page classification and link scoring. In all our experiments (Section 4), we have used the ODP as the underlying taxonomy.

Figure 1 shows the basic process that the topic-focused crawler follows. The example pages found in the taxonomy are provided to the system separated into hierarchical levels, which are then lexically profiled. Seed URLs are the pages the crawl will start with, and the crawl limit represents the total number of pages the crawler should visit (relevant and irrelevant). At the begging, the seed pages are placed in the queue. During each crawl iteration, the top link is selected and the corresponding page is assessed against the target level (“Classify page”). Also, all links found in the page are scored based on their lexical properties (“Score links” see Section 3.2) and added to the

queue. The process is continued until the queue is empty or the crawl limit has been met. In the following sub-sections we describe the steps in more details.

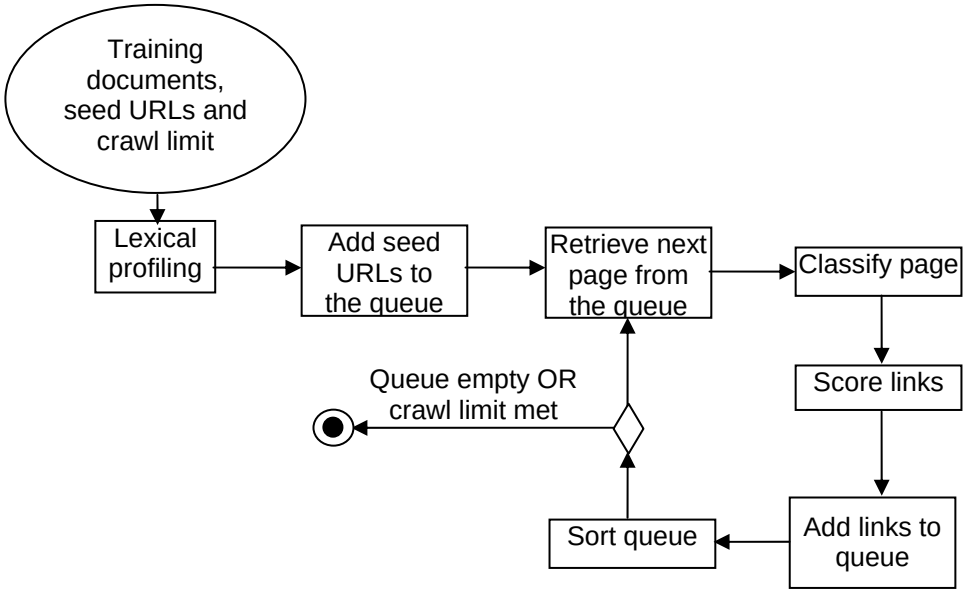

FIGURE 1: THE SYSTEM ARCHITECTURE

3.1. Lexical profiling

The main idea is to provide a profile of each level of the taxonomy so that all relevance assessments are based on it. To profile documents found within each level of the taxonomy, key terms (representing the main concepts found on a page) are extracted from the example documents. The pages are processed by an Automatic Term Recognition (ATR) service (TerMine[3], based on the C-value method (Frantzi et al. 2000)), in order to obtain the terms. The C-value method is a generic ATR approach that considers statistical information (frequency of occurrence, string nestedness, etc.) and recognises multi-word terms that are relevant for the collection. This resulting list includes common features of the documents from within the topic described at that particular level of the taxonomy.

Each term is assigned a lexical profile, represented by all possible linear combinations of the word-level sub-strings present in a term (Nenadic and Ananiadou, 2006). Examples of lexical profiles are in Table 1. Each level in the taxonomy is then represented as a set of such profiles.

Term	Lexical profile, *P*(term)
microarray core facility	{microarray, core, facility, microarray core, core facility, microarray core facility }
DNA microarray facility	{DNA, microarray, facility, DNA microarray, microarray facility, DNA microarray facility}
microarray database	{microarray, database, microarray database}
machine learning	{machine, learning, machine learning}

TABLE 1: EXAMPLES OF LEXICAL TERM PROFILING

The classification of pages found during the crawl will also use TerMine to generate a list of its most significant terms and then the level it will be assigned to will be the level with which it has the greatest lexical similarity (see

[3] http://www.nactem.ac.uk/software/termine/

below). The hypothesis here is that – since topic-content can be characterised by words and their collocations which appear more frequently in a given topic – we can use lexical correlations to associate the pages found within the crawl to the "nearest" taxonomical level.

3.2. Classifying pages

The main idea behind our document classification process is to measure the degree of lexical similarity between taxonomy levels and crawl pages. Since they are represented by terms, we need to provide a method to assess the lexical similarity between two terms. We use a weighted Dice-like coefficient to compare lexical profiles of two terms. We give more credit to pairs that share longer nested constituents, with an additional weight given to the similarity if the two terms have common heads. More precisely, lexical similarity (LS) between two terms is defined as:

$$LS(t_1,t_2)=\frac{|P(h_1)\cap P(h_2)}{|P(h_1)|+|P(h_2)|}+\frac{|P(t_1)\cap P(t_2)|}{|P(t_1)|+|P(t_2)|}$$

where h_1 and h_2 are terminological heads of terms t_1 and t_2 respectively, and P(s) refers to a lexical profile of s. Examples of lexical similarities are provided in Table 2.

$term_1$	$term_2$	$LS(term_1,term_2)$
microarray core facility	DNA microarray facility	0.67
microarray core facility	microarray database	0.11
DNA microarray facility	machine learning	0.00

TABLE 2: EXAMPLES OF LEXICAL SIMILARITY MEASURES

To apply the lexical similarity measure over two lists of terms, an average can be calculated. Therefore the formula used to classify the page is concerned with the average maximum lexical similarity of terms found within a visited page against those which make up the model of a given ODP level:

$$PageScore(Page,Level_n)=\frac{\sum \max\{LS(p,Level_n)\}}{|Page|}$$

Here, *Page* is the set of terms from the visited page; $Level_n$ is the set of terms which describe the n^{th} level of the taxonomy and *p* is an individual term within *Page*. The page will be assigned the level for which it has the highest *PageScore*. If the page has a 0 score for each level, it is judged irrelevant. The average *PageScore* is found for each example level using the other levels in the taxonomy, giving us a 'boundary score', and possible levels of overlap between topics within it giving us minimal acceptable scores for each level. Although only the pages judged as belonging to the most specific end of the taxonomy will be included in the resultant collection, it is important to assign the other pages a level for the overall link prioritisation functions discussed later.

3.3. Content-based link score

When classifying the links found within the page, a number of features of each one is taken into account, including the anchor text (usually a word or phrase which, when clicked on, will navigate the user to the target page). It is assumed that this text (or content of the link) is a brief description of the topic that the target page is concerned with. In our approach, the anchor text is assessed using its lexical similarity against the modelled levels of the taxonomy:

$$LinkContentScore(text,Level_n)=\frac{\sum LS(text,Level_n)}{|Level_n|}$$

As with page classification, the link is assigned the level for which its score is the greatest, allowing a step towards an estimate of the topic to which the target page is concerned.

3.4. Context-based link score

We also consider the context within which the link is found in the page. This text should typically contain a detailed and more specific idea of the topic being discussed when the citation i.e. hyperlink was given, and therefore should give an idea of what topic should be expected on the page. When a link is found within the page, words up to a certain length either side of the link are extracted and lexically profiled as previously explained. This is then subjected to the lexical similarity measures against each level model, and the maximum value for each combination is found:

$$LinkContextScore(P(context), Level_n) = \max\{\max\{LS(p, Level_n)\}\}$$

Here *P(context)* is the lexical profile of the words found within the context of the link, p ∈ *P(context)* and $Level_n$ is the set of terms which describes the n^{th} level of the taxonomy. Again, the level assigned to the text is the one for which this score is the greatest.

3.5. Overall link score

Combining level scores, the overall link score is calculated, which can be then used to assign a priority to links yet to be visited, with the URLs expected to yield documents on topics at the more specific end of the taxonomy being visited first, as these would be assumed to be closer to a target page. The score is calculated as:

$$LinkScore = LinkContextScore \times LinkContentScore \times PageScore$$

where the PageScore is the estimated level of the page the link was found in. Combining the scores like this will not only allow the crawler to prioritise links which seem more likely to yield relevant documents, but when no such links are available will allow the crawler to at least stay near to the target topic, selecting links from at least related topics. This should hopefully reduce the amount of wasted page visits where no expected relevant results exist.

4. EXPERIMENTS AND RESULTS

We have performed an initial experiment to build separate Web-crawlers focused on two topics: bioinformatics companies and genetically modified food. We have used two branches within the ODP taxonomy: Science/Biology/Bioinformatics/Companies and Society/Issues/Science_and_Technology/Biotechnology/Genetics/Genetically_Modified_Food). At each level there is a set of pages which falls under the title topic, and we have used them to lexically profile the Web directory. Tables 3 and 4 show the number of terms used to represent each level which have been extracted using TerMine. Throughout this section, the lowest level (e.g. 'Companies') will be referred to as Level 1, the level above (e.g. 'Bioinformatics') Level 2 and so on. The negative or irrelevant level will be level 5.

Taxonomy Level	Documents	Terms
Companies	4	49
Bioinformatics	5	103
Biology	2	84
Science	0	0

TABLE 3: THE NUMBER OF DOCUMENTS AND TERMS IN THE TAXONOMY LEVELS USED FOR LEXICAL PROFILING (BIOINFORMATICS COMPANIES). [NOTE THAT THE TOP LEVEL DID NOT HAVE DOCUMENTS, AND CONSEQUENTLY NO TERMS WERE EXTRACTED.]

Taxonomy Level	Documents	Terms
GM Foods	4	34
Genetics	4	9
BioTechnology	4	62
Science_and_ Technology	4	136
Issues	0	0
Society	0	0

TABLE 4: THE NUMBER OF DOCUMENTS AND TERMS IN THE TAXONOMY LEVELS USED FOR LEXICAL PROFILING (GM FOODS).

To test the suitability of the lexical profiling and the associated scoring method, the result of the crawl has been compared to a baseline crawl performed on a breadth-first basis. This means that no link was prioritised, and they

were merely visited in the order they were found. In both cases, the same number of the crawl iterations was performed, and both methods started from the same seed pages. What we wanted to test is whether the resultant prioritised crawl would yield more relevant pages earlier on in the crawl, and would it stay at least close to the target topic when no promising links are available.

To evaluate the results, the cumulative number of documents the crawler classified as belonging to Level 1 was plotted against the total number of documents retrieved for each test crawl, giving the harvest rate for each.

Figures 2 and 3 show that the prioritised crawl provides more pages than the baseline in terms of the number of relevant pages found, which is maintained during the rest of the crawl. For example, at the end of the first 100 visits in the bioinformatics company crawl, the prioritised crawl has collected more than twice the number of relevant pages of the breadth-first crawl. The rate at which the number of pages increases is much higher in the prioritised crawl as well, meaning that it is better at seeking out sub-webs of relevant documents. Similar results are observed for the GM Foods branch.

Figures 4 and 5 show the number of pages which were categorised as level 1 or 2, showing the rate at which each crawler was able to stay near the topic, even when no links expected to yield relevant links could be found. Both methods started off increasing fairly constantly. With both topics, the prioritised crawl manages to gain a small lead during most of the crawl despite this, and towards the end (once the links within the start pages had been exhausted) the prioritised crawl manages to maintain a constant rate of related harvesting where the breadth-first crawl flattens off for a largely unproductive period.

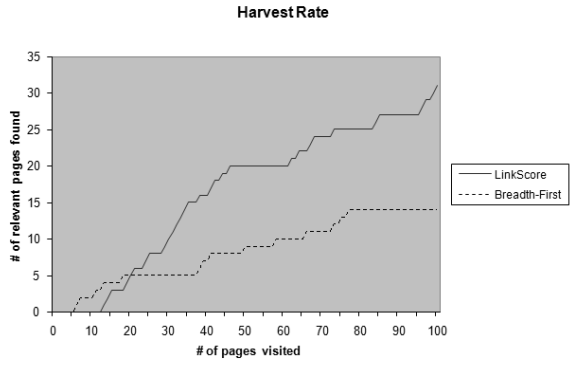

FIGURE 2: HARVEST RATE COMPARISON BETWEEN THE LINKSCORE AND THE BASELINE (BIOINFORMATICS COMPANIES).

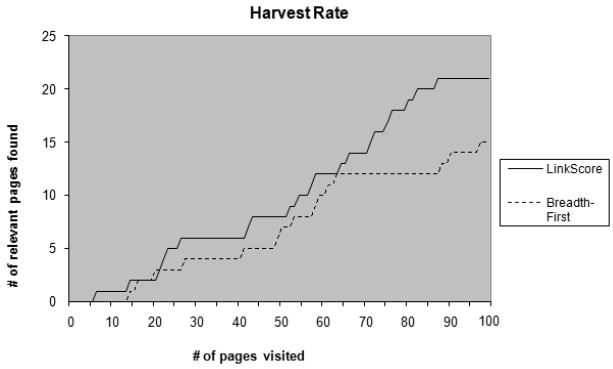

FIGURE 3: HARVEST RATE COMPARISON BETWEEN THE LINKSCORE AND THE BASELINE (GM FOODS).

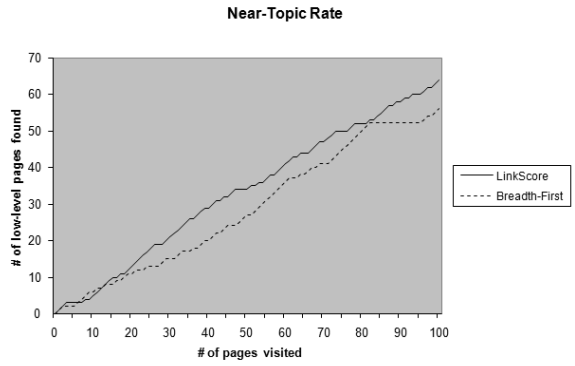

FIGURE 4: NEAR-TOPIC RATE COMPARISON BETWEEN THE LINKSCORE AND THE BASELINE (BIOINFORMATICS COMPANIES).

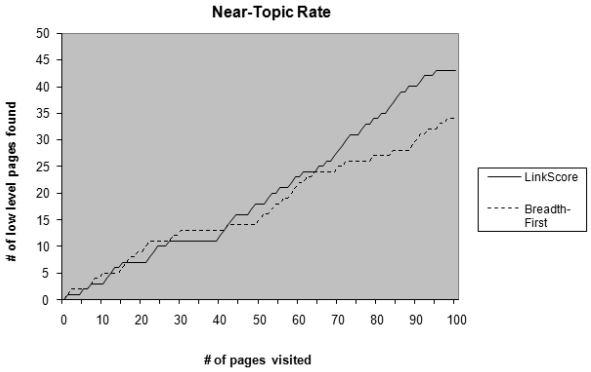

FIGURE 5: NEAR-TOPIC RATE COMPARISON BETWEEN THE LINKSCORE AND THE BASELINE (GM FOODS).

Topic taxonomies like those of the ODP supplying training examples for well-known classification methods have proved effective in the past (Chakrabarti et al. 1999a; Chakrabarti et al. 1999b; Pant and Srinivasan, 2005). But using them to describe more specific subjects and closely related topics enables document collections to be harvested from the Web with not only a finer-granularity of subject matter, but also with greater efficiency as the crawl can be kept close to the topic when no relevant documents are available (see Figure 4 and 5). Keeping the crawler close to the target subject allows these unproductive periods to be reduced, by following links most likely to link to target pages meaning a substantial collection can be produced quickly.

5. CONCLUSIONS

In this paper we have presented a methodology to lexically profile a taxonomy of an existing Web Directory using automatically extracted terms and their substrings. These profiles are then used to lexically compare a page visited during a topic-focused crawl to the fine-granularity classification schema. In addition to that, we have used the same lexical profiling approach to compare content (anchor text) and context of a given link to prioritise the links for the crawl. The three relevance scores (page, content and context) are combined to assess the rank of a given link not only for the topic being crawled, but to related topics as well, facilitating a ranking method that can keep the crawler close to (or not so far from) the target when no relevant pages (with higher scores) are available.

The initial experiments in the domain of bioinformatics and GM Foods have shown that the proposed profiling and scoring method is effective in (a) providing a higher harvest rate than the breadth-first crawl (twice as many relevant documents); (b) seeking out sub-webs of relevant documents; (c) keeping the crawler close to the target subject, allowing unproductive (i.e. non relevant) visits to be reduced.

For our future work, we plan to test the proposed solution on even more domains and using other Web directories with more example documents. We would also like to test the optimal size of example documents needed to lexically profile a topic. As for the scoring mechanism, different weights could be used for content and context scores, and an optimal topic-specific weighting could further improve the results. Incorporating the C-value score of the matching terms (their termhood), to allow weighting towards more significant terms will also be implemented. Finally, we will test how an automatically built controlled vocabulary and an ontology (if available) for a given domain can be used to improve lexical profiling of the hierarchy (Afzal et al. 2008).

ACKNOWLEDGEMENTS

This work was partially supported by the Bio-MITA project (“Mining Term Associations from Literature to Support Knowledge Discovery in Biology”) funded by the UK Biotechnology and Biological Science Research Council.

REFERENCES

H. Afzal, R. Stevens and G. Nenadic (2008): *Towards Semantic Annotation of Bioinformatics Services: Building a Controlled Vocabulary*, Proc. of 3rd Int. Symposium on Semantic Mining in Biomedicine (SMBM 2008), p. 5-12

S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson and J. Kleinberg (1998): *Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text*, Computer Networks and ISDN Systems Vol. 30

S. Chakrabarti, M.H. van den Berg and B.E. Dom (1999a): *Distributed Hypertext resource Discovery through Examples*, Proc. of VLDB 1999 conference, p. 375-386

S. Chakrabarti, M.H. van den Berg and B.E. Dom (1999b): *Focused Crawling: a new approach to topic-specific Web resource discovery*, Computer Networks 31.

J. Cho, H. Garcia-Molina and L. Page (1998): *Efficient Crawling through URL Ordering*, Computer Networks and ISDN Systems Vol. 30, pp. 161-172

P.M.E. De Bra and R.D.J. Post (1994): *Information Retrieval in the World-Wide Web: Making client-based searching feasible*, Computer Networks and ISDN Systems 27, p. 183 - 192

M. Diligenti, F.M. Coetzee, S. Lawrence, C.L. Giles and M. Gori (2000): Focused Crawling Using Context Graphs, Proc. of VLDB 2000 conference, p. 537-534

K. Frantzi, S Ananiadou and H. Mima (2000): *Automatic Recognition of Multi-Word Terms: the C-Value/NC-Calue Method*, International Journal on Digital Libraries Vol. 3 (2000), p. 115-130

J.M. Kleinberg (1999): *Authoritative Sources in a Hyperlinked Environment*, Journal of the ACM, Vol. 46 No. 5

G. Nenadic and S. Ananiadou (2006): *Mining Semantically Related Terms from Biomedical Literature,* ACM Transactions on ALIP Vol. 5 – 2006, p. 22-43

L Page, S Brin, R Motwani, T Winograd (1998): *PageRank: Bringing Order to the Web,* Stanford Digital Libraries Working Paper – 1998

G. Pant and P. Srinivasan (2005): *Learning to Crawl: Comparing Classification Schemes*, ACM Transactions on Information Systems 23 – 2005, p. 430 - 462

Mark Greenwood is an MPhil student in the School of Computer Science, University of Manchester, UK. He has received a BSc in computing science from the same University. His main interests are in Information Retrieval and topic-focused Web crawling, and in particular in prioritisation of links based on their contextual and lexical characteristics.

Goran Nenadic is a lecturer in text mining in the School of Computer Science, University of Manchester, UK. He holds a BSc and MPhil from the University of Belgrade, and PhD from the University of Salford. His main research interests are in text mining, automated terminology acquisition and automated sentiment analysis from textual resources. He also investigates machine-learning approaches to language modelling, in particular in the domains of biomedicine and health-care.