Automatic Reference Tracking with On-Demand Relevance Filtering Based on User's Interest

G.S. Mahalakshmi, S. Sendhilkumar, and P.Karthik

Department of Computer Science and Engineering,
Anna University, Chennai, Tamilnadu, India
mahalakshmi@cs.annauniv.edu, thamaraikumar@cs.annauniv.edu

Abstract. Automatic tracking of references involves aggregating and synthesizing references through World Wide Web, thereby introducing greater efficiency and granularity to the task of finding publication information. This paper discusses the design and implementation of crawler-based reference tracking system, which has the advantage of online reference filtering. The system automatically analyses the semantic relevance of the reference article by harvesting keywords and their meanings, from title and abstract of the respective article. Indirectly this attempts to improve the performance of the reference database by reducing the articles that are actually being downloaded thereby improving the performance of the system. The number of levels for recursive downloads of reference articles are specified by the user. According to user's interest the system tracks up the references required for the understanding of the seed article, stores them in the databases and projects the information by threshold based view filtering.

Keywords: Clustering, Information filtering, Internet search, Relevance feedback, Retrieval models, Reference tracking, Citations.

1 Introduction

Everyday numerous publications are deposited in a highly distributed fashion across various sites. WWW's current navigation model of browsing from site to site does not facilitate retrieving and integrating data from multiple sites. Often budding researchers find themselves misled amidst conceptually ambiguous references while exploring a particular seed scholarly literature either aiming at a more clear understanding of the seed document or trying to find the crux of the concept behind the journal or conference article. Integration of bibliographical information of various publications available on the web by interconnecting the references of the seed literature is very much essential in order to satisfy the requirements of the research scholar.

An automatic online reference mining system [7] involves tracking down the reference section of the input seed publication thereby pulling down each reference entry and parsing them out to get metadata [2,6]. This metadata is populated into the database to promote further search in a recursive fashion. The representation of stored articles framed year wise, author wise or relevance wise, provides complete information about the references for the journal cited. The proposed system searches the

A. Ghosh, R.K. De, and S.K. Pal (Eds.): PReMI 2007, LNCS 4815, pp. 349 – 356, 2007.

entire web for referring to the scholarly literatures and retrieves the semantically relevant reference articles anywhere from the web even though the authorized copy available online is secured. More often, upon submitting the research findings to journals, the authors generally attach the draft article into their home page and of course, without violation of copyrights of the publisher involved. Our proposed system fetches these articles and lists to the user's purview thereby satisfying the only objective of aiding a researcher during the online survey of literatures. By this, we mean, there is no security breach over the automatic fetching of online articles and further, upon acquiring the key for protected sites, a thorough search and reference retrieval shall be achieved.

Though we take utmost care in filtering the cross-reference articles, the recursion can still grow infinite. Also, the harvested articles may be far from understanding the concept of seed paper. Therefore, filtering the reference articles by semantic relevance would be of much use. But filtering the harvested articles ends up with discarding the harvested articles at the cost of system performance. This paper discusses the crawler-based reference tracking system, which analyses the semantic relevance of the referred scholarly literature on demand i.e. at the time of downloading the reference article. Relevance analysis is done by extracting the semantic information from the abstract and keywords section of the particular article and by comparing it with that of the seed reference article thereby distinguishing closely relevant scholarly literatures. The relevant articles that qualify for downloading are submitted to any commercial search engine and the results are retrieved. The retrieved reference articles are filtered for cross references, and are stored in a database.

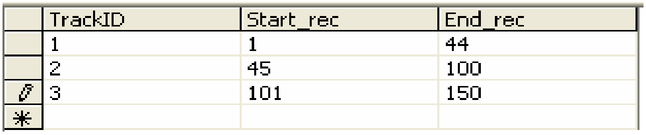
2 Related Work

Citation systems like CiteSeer [1] do not retrieve the entire contents of the file referred to and instead, only the hyperlink to the file in their respective database storage is displayed at the output. Therefore, such systems are not reliable since the successful retrieval is highly dependent on the cached scholarly articles in their own databases (the display 'Document not in database' in CiteSeer). The Google Scholar [4] also performs the search for a given query over scholarly articles but the search results are more unrealistic and deviating like CiteSeer when compared to the results of our system.

3 Background

3.1 Automatic Reference Tracker (ART)

The simple reference tracking system has two sections, namely, reference document retrieval and reference document representation. Of this, reference document representation is merely a hierarchical representation of the collected documents based on indexes like, author, title, and year of publication or relevance of key information. Reference document retrieval includes a) Reference Extraction b) Reference Parsing c) Reference filtering based on keyword relevance d) Link Extraction and Data retrieval.



	TrackID	Start_rec	End_rec
	1	1	44
	2	45	100
✎	3	101	150
*			

Fig. 1. Track details

The entire ART framework [6] is designed such that each reference document retrieved is stored as unique paper and can be tracked from the seed document. Assigning a unique paper id for each reference document facilitates the accessing of track details. Each document consists of a back reference id which points to the base document *pid*, which refers the document. This helps to track down the hierarchy of retrieved documents (refer Fig 1).

Reference documents are retrieved by a recursive algorithm, which inputs the seed scholarly literature and returns a hierarchy of reference documents that are relevant to the seed scholarly literature. The main tracker framework inputs the seed document and preprocess the seed document. Preprocessing involves converting the document file format to an appropriate one in order to extract text from the document.

From the preprocessed text of the seed article, the references are extracted; each individual references are parsed. All the references listed out in a publication's reference section may not be in the same reference format. Reference formatting for journals, books and conference proceedings (called as reference sub-formats) vary in their structure and style, and hence, identification of their individual formats and sub-formats is essential to parse them accordingly. We have limited our parsing to IEEE format. After eliminating duplicating and irrelevant links by comparing the cross references, the reference link extraction is initiated [3].

The references are first segregated into linkable and non-linkable formats. The linkable ones (http references) are extracted directly from the WWW. In non-linkable references, key words are identified and provided to a search engine (Yahoo, for our work) for locating the hyperlink of the reference document. The locations obtained through the search results are analyzed and the particular reference publication is downloaded. These documents retrieved are again pre-processed to sort out their content reference sections and they are again used for tracking up to a certain extent based on their relevance.

Reference Filtering. In reference filtering, initially keywords are harvested from the seed document. Two fundamental approaches of relevance calculation discussed in [6] are: title based analysis and abstract based analysis. Since, the title of every reference article is obtained at the stage of reference parsing, decisions regarding whether to include the reference article for reference extraction or not shall be made before submitting the query to the search engine. Therefore, title based keyword analysis is followed for filtering the references.

The metadata of each reference is compared with the set of keywords to obtain relevance count, a measure of relevance to the seed document. From the relevance count, the user can segregate documents based on a threshold limit. After the articles are retrieved, the relevance of every tracked reference article with respect to the abstract of the seed article is calculated at the time of projecting the results to the user.

Abstract based reference filtering is used as a second filter before the data is actually projected to the user. However, the tracked articles were found to be deviating from the seed scholarly literature on practical reference tracking scenario, which needed a serious analysis along the terms of semantics of the retrieved articles.

Data Representation. The representation can be channelised so that, using color legends and other layout schemes, the user is able to identify the exact publication that aids his/her literature search during research. The representation can be author based, relevance based or year based [6].

3.2 Need for Semantically Enhanced Reference Filtering

Automatic extraction of bibliographic data and reference linking information from the online literature has certain limitations. The primary limitation [6] was the performance bottleneck caused by recursively retrieving online reference documents. However, other publications of authors of seed reference article were maintained assuming that such articles will express the continuity of the author's previous work. Therefore, fixing a threshold in terms of relevance of documents retrieved for a seed document was an obvious solution [6], but poor selection of threshold resulted in reduction of quality of retrieved documents [7]. The reason is that, the number of scholarly articles retrieved will be actually more with a more popular research area. This happened to be the motivation for providing an alternate means of fixing dynamic thresholds by sense-based keyword analysis [5] to prioritize the reference entries in the track database, thereby improving the retrieval performance and clarity. Therefore, application of semantic based relevance-filtering algorithm over the track database of online reference tracker [6] would be of much use.

3.3 Semantically Enhanced Reference Tracking

In [6], seed document's title-based keyword analysis resulted in prioritizing the reference entries in a document so that least relevant reference entries may be discarded, resulting in limited articles getting projected to the user. Since the fixing of threshold in projecting the contents of the track database highly depends on reference filtering based on relevancy, new measures of relevance calculation and threshold fixing were explored. The modification of relevance based reference filtering is two-fold: 1. analysing the keywords collected not only from the title but also from the abstract of the seed document (Modified relevance based reference filtering) 2. analysing the semantic relevancy between keywords of the seed document and the retrieved reference document (Sense based reference filtering). In semantically enhanced reference tracking [7], the parsing of reference format has been extended to html and doc types of scholarly articles. After retrieval, the metadata of relevant references is stored in database (refer Figure 2. for details).

Modified relevance based reference filtering. In plain keyword based relevance filtering [6], elimination of reference articles was based only on the no. of keyword matches against the title of the reference article with that of the seed literature. Here,

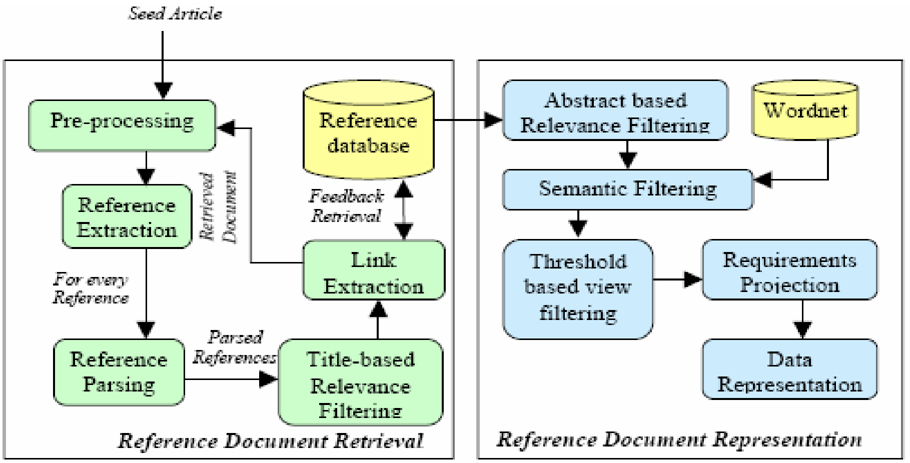
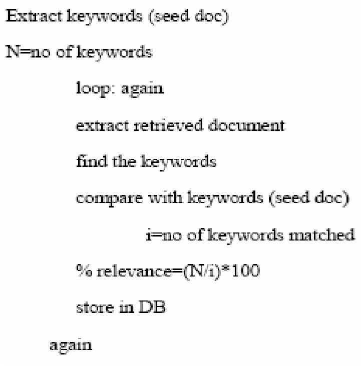


Fig. 2. Architecture of Semantically Enhanced Reference Tracker

that has been extended to keyword based relevance filtering of scholarly articles to involve extraction of words from the title, abstract and highlights in the seed article and the retrieved articles. Based on the frequency of word matches, the threshold to eliminate or include the article into the track database is calculated (refer Figure 3.)



```
Extract keywords (seed doc)
N=no of keywords
        loop: again
        extract retrieved document
        find the keywords
        compare with keywords (seed doc)
                i=no of keywords matched
        % relevance=(N/i)*100
        store in DB
    again
```

Fig. 3. Algorithm for Modified Relevance based Reference Filtering

The database column “percentage” indicates the percentage of keyword matches across the scholarly literatures. The above solution seemed to be obvious since the keyword region has been extended to abstract and all through the text of the reference article. However, the tracked articles were found to be deviating from the seed scholarly literature on practical reference tracking scenario, which needed a serious analysis along the terms of semantics of the retrieved articles.

Sense based Enhanced Reference Filtering This involved filtering relevant documents based on the semantics of the keywords obtained. Here, the keywords from the seed document are first extracted and then for each keyword, the synonymous words are further extracted using Wordnet [8]. Based on the priorities of senses as given by WordNet, the irrelevant and most deviating keywords are eliminated from the keyword set, thus refining the keyword based relevance filtering. During this process, the original keyword set obtained in modified relevance based reference filtering was found to be both populated by new senses and truncated from abstract keywords [7].

Dynamic Threshold Fixing. Dynamic threshold fixing [7] involved fixing adaptive thresholds dynamically as and when the semantic matches were obtained. Therefore, unlike in online reference tracker [6], the sense-based enhanced reference tracker [7] treated the scholarly articles in a more considerable manner.

4 Design of Crawler Based On-Demand Relevance Filtering

Semantic based relevance filtering [7] was efficient in reducing the articles in the reference projection. However, the performance of the reference database was still poor [6,7] due to populating varied reference articles, which actually have no meaning to lie only in the database storage without actually being projected to the user. Crawler based online relevance filtering involves the title-based, abstract-based and sense-based filtering of references in the pre-fetching phase of the crawler before every article is actually downloaded and stored into the reference database.

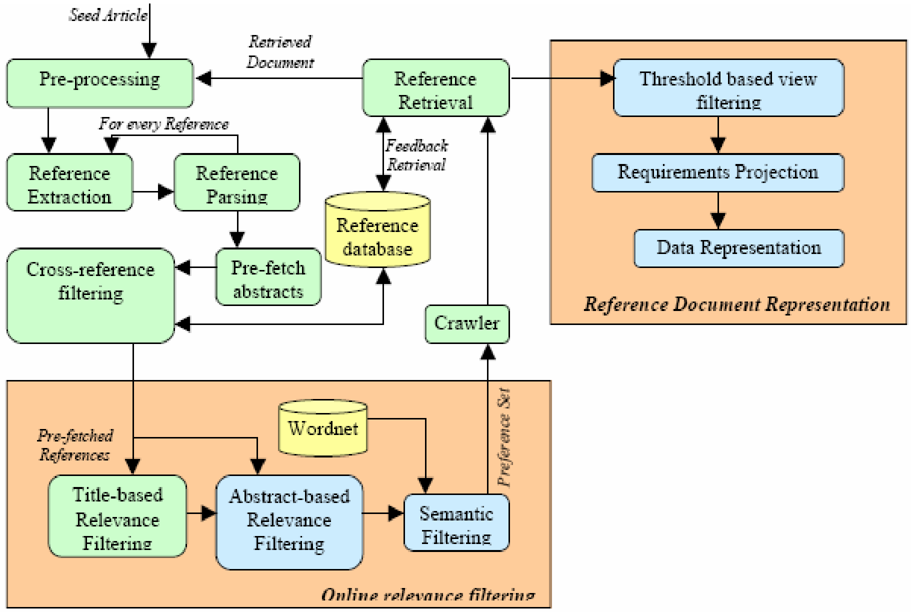


Fig. 4. Automatic Reference Tracking with On-demand Relevance Filtering

The architecture of online relevance filtering based reference tracker is shown in figure 4. The duplication of references are avoided during pre-fetching of articles by cross-reference filtering with respect to the reference database. Only after eliminating all cross-references, the relevance of the pre-fetched article is checked though online relevance filtering. Online relevance filtering basically performs the decision of inclusion / omission of the reference article for downloading which is actually carried out on–demand over the pre-fetched reference abstracts which results in a preference set. The links in the preference set form the seed urls of the crawler using which the reference articles are downloaded.

5 Results of On-Demand Reference Tracking

For implementation purposes, we have assumed the semantic relevance for any reference article as 50% of that of the seed document. The recursion levels are assumed as 2. We have conducted the experiment for 3 different seed research papers seed 1: Probabilistic models for focused web crawling, seed 2: An efficient adaptive focused crawler based on ontology learning, seed 3: Dominos: A new web crawlers' design. The observations are given below in figure 5. There are many reasons for discarding the retrieval of reference articles. The situations encountered in our retrieval process are tabulated in table 1.

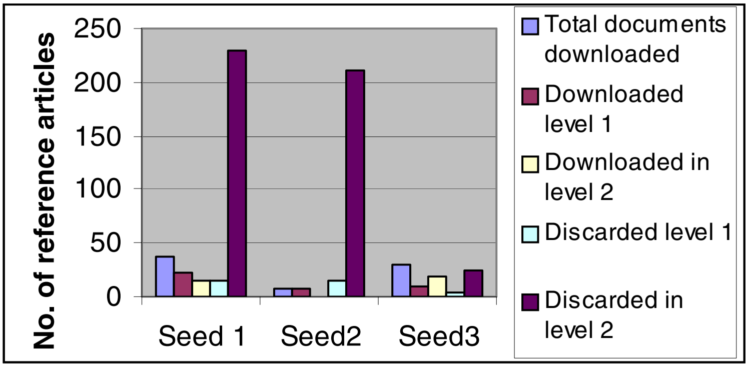


Fig. 5. Result Summary for selected seed research articles

Table 1. On-demand Reference Tracking

Levels	**1**			**2**		
Reasons for discarding	**Seed 1**	**Seed 2**	**Seed 3**	**Seed 1**	**Seed 2**	**Seed 3**
Database miss	8	0	1	30	20	5
Query miss	6	10	0	50	40	3
URL not found	0	1	0	0	0	0
Relevance miss (<50%)	1	4	2	150	150	17

6 Conclusion

This paper has explicitly proposed the integrated framework for on-demand retrieving and representation of online documents and listed the issues involved. The most relevant documents are filtered using separate semantic analysis. The system still has a serious performance bottleneck caused by fixing the semantic relevance and number of recursion levels. Since this scenario highly depends on reference filtering based on relevancy, new and adaptive measures of relevance calculation and threshold fixing should be explored. Sense based filtering based on WordNet is found to be more time consuming. Other techniques like ontology based semantic analysis shall be considered to improve the reference filtering of scholarly articles, which are the future directions of this work.

References

[1] Citeseer.: Scientific Literature Digital Library (2006), Retrieved from http://citeseer.ist.psu.edu/

[2] Day, M.Y., Tsai, T.-H., Sung, C.L., Lee, C.W., Wu, S.H., Ong, C.S., Hsu, W.L.: A Knowledge – based Approach to Citation Extraction. In: Proceedings of IEEE IRI-2005 (2005)

[3] Bergmark, D., Lagoze, C.: An Architecture for Automatic Reference linking. In: Constantopoulos, P., Sølvberg, I.T. (eds.) ECDL 2001. LNCS, vol. 2163, Springer, Heidelberg (2001)

[4] Google scholar (2006), http://scholar.google.com/intl/en/scholar/about.html

[5] Kushchu, I.: Web-based Evolutionary and Adaptive Information Retrieval. IEEE Transactions On Evolutionary Computation 9(2) (2005)

[6] Mahalakshmi, G.S., Sendhilkumar, S.: Design and Implementation of Online Reference Tracking System. In: Proceedings of First IEEE International Conference on Digital Information Management, Bangalore, India (2006)

[7] Mahalakshmi, G.S., Sendhilkumar, S.: Automatic Reference Tracking System. In: Song, M., Wu, Y.-F. (eds.) Handbook of Research on Text and Web Mining Technologies, Idea Group Inc, USA (2008)

[8] Snášel, V., Moravec, P., Pokorný, J.: WordNet Ontology based Model for Web Retrieval. In: Proceedings of WIRI 2005 Workshop, Tokyo, Japan, IEEE Press, Los Alamitos (2005)