Improving Search Accuracy by Combination of Keyword based Search with Semantic Information Search

G.Narender, Dr. Meda Srinivasa Rao

Abstract— Digital libraries today are expected to store millions of articles or items of interest. Intelligent and effective information retrieval becomes critical for the success of such digital libraries. When a large number of mostly irrelevant results are returned in response to a user search, it poses a problem for the user who then has to sort through the search results looking for articles of interest. This can be time consuming and inefficient. This can also keep users away from utilizing digital libraries. Articles in digital libraries are often indexed with a set of keywords. Computer programs can be highly effective at performing keyword based searches and coming up with a candidate set of articles of interest. In this paper, we propose a method for annotating articles in the database with a variety of semantic information. User queries are pre-parsed to find the real intent of a user search. The information from this pre-parsing can be used to search through the annotations of articles in the resulting candidate set from a keyword search. Articles in this candidate set can then be ranked in terms of number of annotations that match and the results displayed in decreasing order of number of annotations that match. This increases the probability that the initial search results that get displayed are useful to the user thereby improving the experience of using digital libraries. In this paper we present the results of our work combining results from a keyword based search using semantic information from added annotations.

Index Terms—Digital Library, Semantic Search, Information Retrieval, Semantic Annotations, Search Ranking

1 Introduction

A digital library is a library in which articles of interest are stored in digital format as opposed to print, or other physical media. The collections are made easily accessible via computers [1]. As mentioned before, a digital library can potentially store millions of articles and efficient indexing and retrieval methods become critical for the success of digital libraries. Semantic search algorithms seek to improve searches by returning more relevant results. In order to do this, these algorithms try to understand user intent and the contextual meaning of words in the knowledge domain that is being searched. Very often the goal of semantic search is to return a set of results that have a high probability of being useful and relevant to the user. This saves the user the effort of sorting through potentially a large number of results that have been returned based on results of a simple keyword based search. As an example, consider the following query from a user – *Computer Science faculty at Hyderabad Central University*. The user is specifically seeking information about teaching faculty that work at a specific university whose specialization is Computer Science. A simple keyword based search may resort to returning documents that contain each of the keywords – computer, science, faculty, Hyderabad, central and university. The search is very likely to return a large number of results most of which will not be relevant or useful for what the user is looking for. On the other hand, consider a semantic search system which really understands the intent of the user query and also organizes its database in an intelligent manner. An example would be annotations for the articles in its collection with semantic information such as area of specialization and name of the university which enables a semantic search algorithm to make use of this information to return search results that have a high probability of being useful to the user. In our past research [2], we have shown that fast brute force search made possible by ever increasing computing power can be a useful tool. In this paper, we show the results of combining brute force keyword search with a semantic search to order or cut down the number of successful results. The ordering of the results increases the probability of finding relevant articles upfront. Cutting down on the number of results reduces the number of articles that the user has to browse through thereby saving them time. Both aspects improve the experience of using digital libraries.

This paper presents results from our work combining a keyword based brute force search of a database of bibliography of research papers with a search for several semantic annotations. The rest of the paper is organized as follows. In section 2, we give a survey of semantic search algorithms. Section 3 gives the implementation details of our novel approach. We present our results in section 4. Our conclusions and scope for future work is presented in section 5.

2 Survey of Semantic Search Approaches

We are seeing a tremendous growth in the amount of information that is readily available at one's finger tips with the explosion of the World Wide Web or the internet. Internet removes the physical barriers that were once associated with

- *Research Scholar, Associate Professor, Department of CSE,Keshav Memorial institute of technology,Hyderabad,INDIA. E-mail:guggillanarender@gmail.com*

Professor and Director ,School of Information technology, JNTUH Hyderabad.INDIA. E-mail:srmeda@gmail.com

knowledge dissemination making a person's physical location largely irrelevant. Current research and scholarly articles that were once available to only a select few can now be made available for everyone on the internet. Once made available on the internet, any student in any part of the world with access to the internet can access these articles to enhance their knowledge. The tremendous amount of information available on the internet comes with its own set of challenges. One of these challenges is the challenge of information retrieval. The vast amount of information available on the internet can only be used effectively with efficient information retrieval. Given a user search in the form of a query, information retrieval tries to return links to a set of articles that are likely useful to satisfy the user need for information. A variety of the search algorithms resort to simple keyword based searches. The user query is split into a set of keywords and a brute force search of the article database is made to get a set of matching results. The number of results that are returned can often vary a large amount depending on the degree of specialization of the user specified keywords. To improve user experience search engines resort to a variety of ranking techniques such as PageRank, Citation Indexing to present the search results [3]. The World Wide Web as it exists today can be considered a large semi structured database. The international standards body has recently taken on the task of a collaborative movement towards a Semantic Web. The standards promote the use of standardized data formats for information on the web. The main aim of the semantic web is to allow users to find and share information more easily. As a parallel, semantic search techniques try to understand the actual user intent of a query so that they can return better search results to the user.

A survey of semantic web search engines is provided in [4]. It also lists four different approaches to perform a semantic search. The first approach uses context information to interpret the intent of the user query. As an example, consider the word *bark*. Depending on the context of the user query it can refer to two different words. When used in the context of *'dog bark'*, it means the short loud cry of a dog. When used in the context of trees, it refers to the tough exterior cover of a root or stem. Contextual analysis attempts to improve the search results by making use of the context information of a query. The second approach uses reasoning. A search system that makes use of reasoning to improve its results knows about relationships among objects and how to infer new relationships from existing relationships. It uses this knowledge to give better search results. The third approach uses natural language processing. Search engines using this approach attempt to parse a user query and deduce information from the same to identify things such as people and places. Such search engines capture things like what was the subject, the object and relationships among the words. When the user inputs a query, it attempts to match this semantic information with the semantic information of the web article. An example of this would be queries like 'who conquered Ravana' vs `Ravana conquered who' where the intent of the two queries is totally different and a simple keyword based search will not be fully accurate as both queries will likely return the same set of results. However, a natural language processing system would be able to interpret the intent of such queries accurately to arrive at better search results. The last approach makes use of ontology for domain specific searches. An ontology formally represents knowledge as a set of concepts in a domain. An ontology can be used to model the domain - the type of objects that exist and their properties or relations. As an example an ontology that is used to model vehicles knows that a car is a type of a vehicle and it can use this relation to broaden or generalize the search.

Several semantic search engines have been implemented. One such search engine is the hakia search engine. Hakia calls itself the 'meaning based semantic search engine'. For an overview of the semantic search technology in the hakia search engine, we refer the reader to [5]. Hakia search engine uses an approach named QDEX which allows a full-fledged meaning based analysis similar to ontological semantics while overcoming the extreme scalability requirements for indexing conventional ontological systems. Data storage in this model does not grow linearly when a new article is added to the digital library and only grows when new knowledge gets added.

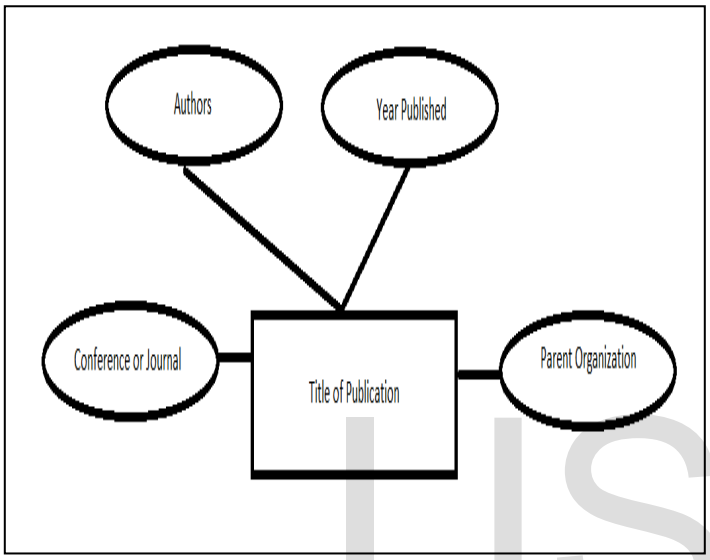
Another semantic search engine is the kngine search engine. It stands for Knowledge Engine. It tries to predict what a user really means when a user types a query. For an overview of how kngine works, we refer the reader to [6]. Kngine uses web crawlers to organize unstructured data over the web. In order to organize the data, it makes use of natural language processing, data mining and machine learning. The organized data becomes the foundation for the search engine's knowledge graph. Once a user types in their query, kngine again utilizes natural language processing to understand the user intent and follows it up with a search of the knowledge graph to arrive at the search results.

Powerset a search engine that has now been acquired by Microsoft is a search engine that is focused on natural language processing. It tries to understand the nature of the user question to search for relevant results. Several other semantic search engines such as Kosmix, Sensebot, and DuckDuckGo are becoming popular. We refer the reader to [4] for an overview of these search engines. For more examples of different kinds of search engines and to also get an understanding of the common issues associated with semantic search approaches we refer the reader to [7].

3 Combining Keyword based Search with Semantic Search

As previously mentioned, semantic searches typically suffer from scalability issues. Keyword based searches on the other hand scale well as the database size grows. It is also possible to parallelize keyword based searches to get further improvements in performance [2] in a straight forward and easy manner using C++ language extensions such as Intel Cilk Plus [8]. It would be nice if we can combine the two searches so that we can take advantage of scalability of keyword based searches along with the possibility of getting more accurate results from the semantic search approach. In this section, we present

the details of our work to implement a combination of the two searches. In order to demonstrate the suitability of our approach, we consider a database of publications. The database can be considered as having details about each publication along with a link to the actual publication. Each publication has a title and is annotated with several pieces of semantic information such as whether it was published in a conference or a journal, the year it was published in, the authors of the publication, and the parent body of the journal or the conference such as IEEE or ACM. An entry in the database can be shown pictorially as follows.



The search algorithm starts out by doing a pre-parsing of the user query to attempt to understand the user intent. Consider the user query: *IEEE journal publication on semantic search by Joe Someone in 2009.* To keep our implementation simple, we look for certain keywords to guess the user intent. Certain words such as IEEE, ACM are treated as special and are considered as specifying the parent organization. The word conference or journal is used to determine whether the user is interested in conference or journal publications respectively. The words *'on, by and in'* are treated as special keywords and are used to determine the search keywords, author, and year of publication respectively. On encountering one of these keywords, the rest of the words in the query are considered as specifying the corresponding query information until we encounter another special keyword. In the query example shown before, the words semantic and search which follow *on* are treated as search keywords. We stop treating words as search keywords as soon as we encounter the word *by* which we recognize as one of the special keywords. The search process starts once the query has been pre-parsed to extract different pieces of semantic information. A quick keyword based search is made through the database to search for publication titles that contain the words *semantic* and *search*. The keyword based search is used to determine the potential candidates for improving the search using the semantic information. Each candidate item from the keyword based search already has associated semantic information as annotations. These semantic annotations are compared against the semantic information that we obtained from our pre-parsing of the query. The candidates with the most number of matching semantic pieces are considered higher probability candidates for satisfying the user request. When the query results are presented to the user, we have the option of only presenting the results that match all the semantic search criteria or the option of presenting the search results in decreasing order of number of semantic information matches. The pseudo code for the algorithm looks as shown below.

```
Combination_keyword_semantic_search(user_query_string)
        parse_output = pre_parse_query(user_query_string)

        for each article in the database
                if parse_output->search_keywords are found
                   in article->title
                        add article to candidate_set
                end if
        end for

        for each article in the candidate_set
                num_semantic_matches = 0

                if article->author matches
                  parse_output->author
                        increment num_semantic_matches
                end if

                if article->type matches
                  parse_output->type
                        increment num_semantic_matches
                end if

                if article->publication_year matches
                   parse_output->publication_year
                        increment num_semantic_matches
                end if

                if article->parent_organization matches
                  parse_output->parent_organization
                        increment num_semantic_matches
                end if

                if num_semantic_matches == 0
                        continue
                else
                        add article to a vector corresponding
                        to num_semantic_matches
                end if

        end for

        if only displaying all_match search results
                display articles in vector corresponding
                to max_allowed_semantic_pieces
        else
                for( num_match = max_allowed_semantic_pieces;
                    num_match--;
```

```
                    num_match >= 1)
                        display articles in vector corresponding
                        to num_match
                end for
        end if
End
```

4 Results

In order to demonstrate the usefulness of our combined approach, we started out with a small database that had information for about 30 publications in the specified format. Since our database is really small, we did not resort to parallelization of the keyword based search. We have shown from our earlier work in [2] that significant speedups can be obtained using the Intel Cilk Plus extensions if the database size is significantly larger. Our sample database is shown in Appendix A. The results of the keyword based search and the combined keyword/semantic search are shown in Appendix B. As can be seen from the results in Appendix B, we can get more accurate results or cut down on the number of search results with the combination search approach. A simple keyword based search returns nine articles in the publication database as a possible match. A combined keyword and semantic search however only returns two articles as possibly interesting to the user as only these two articles match the semantic information from the query. The combined search algorithm also gives the same nine results from the simple keyword based search in decreasing order of number of semantic matches. This order increases the likely hood that the user finds most relevant articles of interest upfront in the search results. The results thus clearly demonstrate the benefit of the combination search approach for the user.

5 Conclusion and Future work

Our work has demonstrated the benefit of a combined keyword based and semantic search approach. However, there are still potential issues with our approach. The initial keyword based search to determine the candidate set does not take any semantic information into account. Because of this it is possible that the candidate set is empty or a subset of what the user may be interested in. One way to fix this will be to use some meaning based or ontological approach during the keyword search to start with a better candidate set. This however will not come free and will incur the cost of some additional search time to build the candidate set. There can also be potential issues during the pre-parsing of the user query. The pre-parsing considers certain words as special. Any use of these keywords other than in the intended manner is likely to confuse the pre-parser. The pre-parser can be made smarter using natural language processing techniques. For future work we may consider looking into some of these aspects.

References

[1] Greenstein, Daniel I., Thorin, Suzanne Elizabeth. The Digital Library: A Biography. Digital Library Federation (2002)

[2] G.Narender, Dr. Meda Srinivasa Rao. Parallelizing Digital Library Search. International Journal of Scientific & Engineering Research, Volume 4, Issue 4, April-2013 1671

[3] Wang Wei, Payam M. Barnaghi, Andrzej Bargiela. Search with Meanings: An Overview of Semantic Search Systems. International Journal of Communications of SIWN, Vol.3, pp.76-82. April 2008

[4] G. Sudeepthi, G. Anuradha, Prof. M. Surendra Prasad Babu. A Survey on Semantic Web Search Engine. IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012.

[5] http://company.hakia.com/new/documents/White%20Paper_Semantic_Search_Technology.pdf

[6] http://www.kngine.com/Technology.html

[7] G.Madhu, Dr.A.Govardhan, Dr.T.V.Rajinikanth. Intelligent Semantic Web Search Engines: A Brief Survey. International journal of Web and Semantic Technology, Vol. 2, No. 1, January 2011.

[8] http://software.intel.com/en-us/articles/intel-cilk-plus/

Appendix A: Sample Publication Database

Parent Org	Type of Publication	Authors	Publication Title	Year Published
IJCSIT	Journal	G. Narender, Dr. Meda Srinivasa Rao	Parallel OCR Error Correction	2012
IJSER	Journal	G. Narender, Dr. Meda Srinivasa Rao	Parallelizing Digital Library Search	2013
ACM	Journal	Brahim Medjahed, Athman Bouguettaya, Ahmed K. Elmagarmid	Composing Web Services on the Semantic Web	2003
ACM	Journal	Nigel Shadbolt, Tim Berners-Lee, Wendy Hall	The Semantic Web Revisited	2006
ACM	Journal	Enrico Motta	Knowledge Publishing and Access on the Semantic Web: A Sociotechnological Analysis	2006
ACM	Conference	Mohammad Ali H. Eljinini	Health: related information structuring for the semantic web	2011
Springer-Verlag	Conference	Wolf Siberski, Jeff Z. Pan, Uwe Thaden	Querying the semantic web with preferences	2006
ACM	Workshop	Guoqian Jiang, Harold R. Solbrig, Christopher G. Chute	Using semantic web technology to support ICD-11 textual definitions authoring	2011
ACM	Workshop	Stuart N. Wrigley, Dorothee Reinhard, Khadija Elbedweihy, Abraham Bernstein, Fabio Ciravegna	Methodology and campaign design for the evaluation of semantic search tools	2010
ACM	Workshop	Peter Mika	Distributed indexing for semantic search	2010
ACM	Workshop	Hannah Bast, Florian Bäurle, Björn Buchhold, Elmar Haussmann	A case for semantic full-text search	2012
ACM	Transactions	Kareem Darwish, Walid Magdy	Error correction vs. query garbling for Arabic OCR document retrieval	2007
IEEE	Symposium	B. Kruatrachue, K. Somguntar, K. Siriboon	Thai OCR Error Correction Using Genetic Algorithm	2002
IEEE	Conference	Ahmad Abdulkader, Mathew R. Casey	Low Cost Correction of OCR Errors Using Learning in a Multi-Engine Environment	2009
IEEE	Conference	M. Wick, M. Ross, E. Learned-Miller	Context-Sensitive Error Correction: Using Topic Models to Improve OCR	2007
IEEE	Conference	B. B. Chaudhuri, U. Pal	OCR Error Detection and Correction of an Inflectional Indian Language Script	1996
IEEE	Workshop	Dar-Shyang Lee, Ray Smith	Improving Book OCR by Adaptive Language and Image Models	2012
ACM	Workshop	Eugene Borovikov, Ilya Zavorin, Mark Turner	A filter based post-OCR accuracy boost system	2004
ACM	Conference	Ann Blandford, Suzette Keith, Iain Connell, Helen Edwards	Analytical usability evaluation for digital libraries: a case study	2004
IEEE	Conference	Gregory Crane, Clifford Wulfman	Towards a cultural heritage digital library	2003
ACM	Conference	Dion Goh, John Leggett	Patron-augmented digital libraries	2000
ACM	Conference	Tamara Sumner, Melissa Dawe	Looking at digital library usability from a reuse perspective	2001
ACM	Conference	Ann Blandford, Hanna Stelmaszewska, Nick Bryan-Kinns	Use of multiple digital libraries: a case study	2001
IEEE	Conference	Tamara Sumner, Michael Khoo, Mimi Recker, Mary Marlino	Understanding educator perceptions of "quality" in digital libraries	2003
IEEE	Conference	David Bainbridge, John Thompson, Ian H. Witten	Assembling and enriching digital library collections	2003
ACM	Conference	George Buchanan, Jeremy Gow, Ann Blandford, Jon Rimmer, Claire Warwick	Representing aggregate works in the digital library	2007
IEEE	Conference	Ahu Sieg, Bamshad Mobasher, Robin Burke	Ontological User Profiles for Representing Context in Web Search	2007
ACM	Conference	Ahu Sieg, Bamshad Mobasher, Robin Burke	Web search personalization with ontological user profiles	2007
ACM	Conference	Kapil Goenka, I. Budak Arpinar, Mustafa Nural	Mobile web search personalization using ontological user profile	2010
ACM	Symposium	Chaitali Gupta, Rajdeep Bhowmik, Madhusudhan Govindaraju	Ontological framework for a free-form query based grid search engine	2008

Appendix B: Sample Results

Search results for query: acm conference publications on digital library in 2001

Results of simple keyword search:

1. G. Narender, Dr. Meda Srinivasa Rao ; Parallelizing Digital Library Search ; IJSER Journal ; 2013
2. Ann Blandford, Suzette Keith, Iain Connell, Helen Edwards ; Analytical usability evaluation for digital libraries: a case study ; ACM Conference ; 2004
3. Gregory Crane, Clifford Wulfman ; Towards a cultural heritage digital library ; IEEE Conference ; 2003
4. Dion Goh, John Leggett ; Patron-augmented digital libraries ; ACM Conference ; 2000
5. Tamara Sumner, Melissa Dawe ; Looking at digital library usability from a reuse perspective ; ACM Conference ; 2001
6. Ann Blandford, Hanna Stelmaszewska, Nick Bryan-Kinns ; Use of multiple digital libraries: a case study ; ACM Conference ; 2001
7. Tamara Sumner, Michael Khoo, Mimi Recker, Mary Marlino ; Understanding educator perceptions of "quality" in digital libraries ; IEEE Conference ; 2003
8. David Bainbridge, John Thompson, Ian H. Witten ; Assembling and enriching digital library collections ; IEEE Conference ; 2003
9. George Buchanan, Jeremy Gow, Ann Blandford, Jon Rimmer, Claire Warwick ; Representing aggregate works in the digital library ; ACM Conference ; 2007

Results from combined search with all semantic match :

1. Tamara Sumner, Melissa Dawe ; Looking at digital library usability from a reuse perspective ; ACM Conference ; 2001
2. Ann Blandford, Hanna Stelmaszewska, Nick Bryan-Kinns ; Use of multiple digital libraries: a case study ; ACM Conference ; 2001

Results from combined search in decreasing order of semantic matches:

1. Tamara Sumner, Melissa Dawe ; Looking at digital library usability from a reuse perspective ; ACM Conference ; 2001
2. Ann Blandford, Hanna Stelmaszewska, Nick Bryan-Kinns ; Use of multiple digital libraries: a case study ; ACM Conference ; 2001
3. Ann Blandford, Suzette Keith, Iain Connell, Helen Edwards ; Analytical usability evaluation for digital libraries: a case study ; ACM Conference ; 2004
4. Dion Goh, John Leggett ; Patron-augmented digital libraries ; ACM Conference ; 2000
5. George Buchanan, Jeremy Gow, Ann Blandford, Jon Rimmer, Claire Warwick ; Representing aggregate works in the digital library ; ACM Conference ; 2007
6. Gregory Crane, Clifford Wulfman ; Towards a cultural heritage digital library ; IEEE Conference ; 2003
7. Tamara Sumner, Michael Khoo, Mimi Recker, Mary Marlino ; Understanding educator perceptions of "quality" in digital libraries ; IEEE Conference ; 2003
8. David Bainbridge, John Thompson, Ian H. Witten ; Assembling and enriching digital library collections ; IEEE Conference ; 2003
9. G. Narender, Dr. Meda Srinivasa Rao ; Parallelizing Digital Library Search ; IJSER Journal ; 2013