IJCSC International Journal of Computer Science and Communication Vol. 3, No. 1, January-June 2012, pp. 237-241

RANK BASED CONTENT MINING

Anshu Khurana[1], Ankita Gupta[2], Deepika Vatsa[3] and Prerana Mukherjee[4]

[1,2,3,4] Department of Information Technology, Delhi Technological University, Delhi, India, [1]*E-mail: anshukhurana@dce.ac.in*, [2]*E-mail: ankita.gupta.mail@gmail.com* [3]*E-mail: vatsa.deepika@gmail.com* [4]*E-mail: mukherjee.prerana@gmail.com*

ABSTRACT

The rapid growth of information on the web leads to difficulty of extracting potentially useful knowledge. So, web content mining is a probable solution for tackling this problem. Web content mining (WCM) is the scanning and mining of text, pictures and graphs of a Web page to determine the relevance of the content to the search query. WCM is very effective when used in relation to a content database dealing with specific topics. For a new comer in the field of research it's very difficult to reach to the required and relevant results. While searching for research papers on internet, one requires either the knowledge of authors of the paper or the exact title of the paper. It's not possible to get appropriate results if a person has only basic knowledge of the topic or does not exactly know the author name. For this, a number of ranking algorithms are there namely, page rank, weighted page rank, HITS algorithm, distance rank algorithm etc. In this paper, we have proposed another page ranking algorithm which is easy to apply on a large database of papers. This algorithm uses a certain number of parameters that are easy to retrieve as compared to other ranking algorithms.

Keywords: Stemming, ranking, ATM, TAM

1. INTRODUCTION

With the advent of the Web and various specialized digital libraries, the automatic extraction of useful information from text has become an increasingly important research area in web mining.

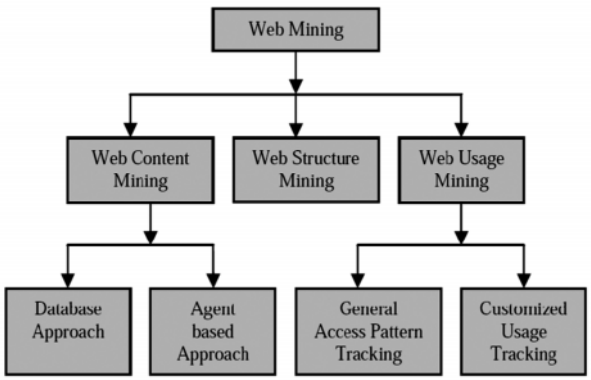

Figure 1: Areas of Web Mining

There are three areas of Web mining according to the usage of the Web data used as input in the data mining process, namely, Web Content Mining (WCM), Web Usage Mining(WUM) and Web Structure Mining (WSM).

Web Content Mining (WCM)

Web Content Mining is the process of extracting useful information from the contents of web documents. The web documents may consists of text, images, audio, video or structured records like tables and lists. Mining can be applied on the web documents as well the results pages produced from a search engine.

Libraries are now facing a crucial transitional state, which necessitates adopting modern methods of knowledge organization. The advent of Information Technology offers a powerful means of translating intellectual contributions into value-added information products, which satisfies the specific requirements of each end user. The implications of this phenomenon have reflected a gradual transformation of the physical format of information resources in Libraries. The concept of multimedia Library was introduced when audio-visual materials were added along with the printed media. The next stage has been marked by the digitalization of information and the digital library came into vogue. Digital library does not mean merely converting the information into digital form but it is an asset, which facilitates free flow of exchange of information at global level through Internet. This led to the metamorphosis of library into virtual library.

1.1 Motivation

The inspiration for this project comes particularly from CiteSeerX, a digital library and repository for scientific and academic papers with a focus on computer and information science. However, there are also other interesting online academic literature repositories, such as the local Highwire Press, an offshoot of the Stanford libraries and Google Scholar.

The proposed system will be very useful to the researchers as initially they are not aware of the precise research areas to search upon and the associated papers where they can work on, and also the sequencing of the papers of an author on the same field is not known, so this system assists the researcher in searching the papers that they are interested to search.

1.2 Problem Statement

Our project aims at describing the various relations between topic and author-name within the documents. We are looking for a concise answer to the questions of "what do I need to know about topic and author". Our proposed method gives a relationship between author-name and topic-name and retrieve the papers for a specific author and thus retrieve the related research field (may be a topic name) and vice versa. It also supports the ranking scheme which is able to prioritize the papers based on certain parameters. Author-Name is the name of the author who wrote the papers on various topics of his field either individually or with some co-authors. The papers are the author's papers which are written by the author or with collaboration of a co-author. Research field indicates the core topics of research of an author and we have assumed that an author always repeats his researcher topic related words in his papers. The task is to recognize the mentions of each paper and research fields using author-name while papers of various authors using topic-name and prioritize them according to the proposed ranking system.

1.3 ASSUMPTIONS

1. Our project is doing best work when it is being installed for a web search engine, where there is a fixed document format of a research paper as shown below:

   TOPIC

   AUTHORNAME, COAUTHORS

   <ENTER>

   ABSTRACT

   <ENTER>

   INTRODUCTION

   (*a*) The Paper should strictly follow the above mentioned format.

   (*b*) The names of all the author and co-authors (if any) is in the second line separated by a comma (,).

2. The algorithm does not take sentence structure into account.

3. Author will use his research field word more number of times compared to the words which is not related to his research field.

2. PROPOSED METHOD

STEMMING

Stemming refers to identifying the root of a certain word. There are basically two types of stemming techniques, one is inflectional and other is derivational. Derivational stemming can create a new word from an existing word, sometimes by simply changing grammatical category (for example, changing a noun to a verb). A commonly used algorithms is the 'Porter's Algorithm' for stemming. When the normalization is confined to regularizing grammatical variants such as singular/plural or past/present, it is referred to inflectional stemming. To minimize the effects of inflection and morphological variations of Words (stemming), our approach has pre-processed each word using a provided version of the Porter stemming algorithm with a few changes towards the end in which we have omitted some cases.

e.g. apply – applied – applies

print – printing – prints – printed

In both the cases, all words of the first example will be treated as 'apply' and all words of the second example will be treated as 'print'.

The 2 models used in this approach are as follows:

(*a*) ATM (Author-Topic Model)

In this model, user needs to enter the author name. On the basis of the query the scanning of the paper dataset is done. On finding the appropriate set of papers, inflectional stemming and removal of stop words, punctuation marks, adjectives and verbs occurs.

Based on the frequency count of words, the top 5 research fields are selected. Now, weighted ranking factor (WRF) is calculated for all the selected papers and they are sorted according to the WRF of papers i.e. higher WRF is given more priority. Table containing paper name, authors, co-authors and research field is displayed.

(b) TAM (Topic-Author Model)

In this model, user needs to enter the topic. On the basis of the query the scanning of the paper dataset is done. On finding the appropriate set of papers, inflectional stemming and removal of stop words, punctuation marks, adjectives and verbs occurs. Now, weighted ranking factor (WRF) is calculated for all the selected papers and sorted as explained in above model. Table containing paper name and author name is displayed.

3. RANKING SYSTEM

As the number of Web pages, documents, papers and users on the Web are increasing; the number of queries submitted to the search engines are also increasing rapidly. Therefore, it is required that the search engines give best and efficient results. The search engines become very successful and popular if they use efficient Ranking mechanism. Some page ranking algorithms are Page Rank,

Weighted Page Rank, Hypertext Induced Topic Search HITS, DistanceRank and DirichletRank algorithms.

4. OUR RANKING ALGORITHM

Basically our aim is to rank the papers belonging to particular category or research field based on certain parameters so that when a user queries the system on some research field or author name, a list of most relevant papers are listed rank wise, i.e., the best paper in that category or of that author is listed first and others later.

Document Ranking used the approach of a term weighting system. Generally, a typical term-weighting formula is defined as being composed of two component pair: *tfcq and cfcq* which represents the weight of a term in a user query *q*, and *tfcd and cfcd* which represents the weight of a term in a document *d*. The term frequency component (*tfc*) represents how many times a term occurs in a document or query. The collection frequency component (*cfc*) considers the number of documents in which a term appears. Low frequencies indicate that a term is unusual and thus more important to distinguish documents.

Typical term-weighting formulas combine these two components. We can define $wtd = tfcd \times cfcd$, and $wtq = tfcq \times cfcq$, where *wtd* is the weight of term *t* in document *d* and *wtq* is the weight of term *t* in query *q*.

We can express a ranking function based on such a term weighting system as follows:

$sim(q, d) = (\text{summation})\ t \in q\ wtd \times wtq$

But we propose a different approach and a formula for ranking of research papers based on a weighted rating formula referred to in actuarial science as a *credibility formula*. This label arises because a statistic is taken to be more credible the greater the number of individual pieces of information and frequency of words in a document or just the citations.

If a query is based on the author name then all his papers are listed according to the WRF (weighted rank factor).

The Parameters:

1. F = Aggregated counts of the research field (and sub fields) of the papers. This plays an important role because these counts indicate the degree of association of the topic or research fields with the paper.
2. N = Number of Citations. This gives the number of times a paper has been cited by other research papers of the same field. So this count indicates the reputability of a paper.
3. NV = Number of views of a paper.

   Our proposed system also counts the number of times a paper gets viewed by regular users of the system. This again directly indicates the ranking of a paper.
4. n = Number of papers which are displayed as a result corresponding to a particular search by author name or topic name.

The Ranking Formula:

$$WRF = (F + N + NV)/n$$

5. RESULTS

To implement these two mapping models we had used the following tools:

FRONT END

Visual Studio 2010 with C#

BACK END

Here we had assumed some predefined datasets namely:

Dataset1: punctuation marks (., ?, /, %, &)

Dataset2: stop words (a, an, the, across)

Dataset3: verbs (apply, bring, push)

Dataset4: adjective (bad, good, alive)

Dataset5: Papers

We have maintained a database in SQL Server Management Studio 2008 for the records of all the papers that we have taken in our dataset of papers. We have stored the fields namely authorname, papername, pathname, noofcitations, noofviews for each paper. These fields are taken into account for calculation of the rank factor by the formula given in Section III. But the system can be extended and make use of the dataset of any digital library like CiteSeer.

We need to select the path of the folder that contains all the papers in the required format as mentioned in the algorithm in the field DataSet. We can make the search by author name which gives all the papers with that author name showing 3 tables namely papername, author and co-author name table, research fields of that paper.

Figure 2: Search Results for Author Name

We can make the search by topic name which gives all the papers with that topic as the research field showing a table indicating the papername and authorname of that paper.

Figure 3: Search Results for a Topic

When the button for the papers displayed is clicked the details of the respective paper are shown as below.

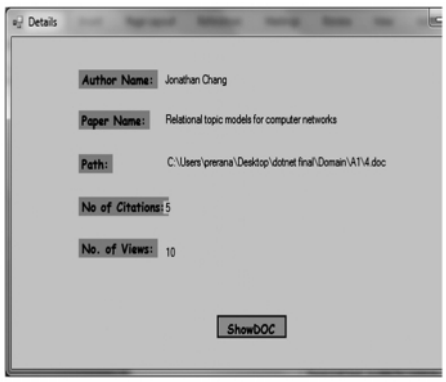

Figure 5: Details for a Paper

6. CONCLUSION

We have proposed a ranking model which will rank the papers based on query and the model applied. By this method, ambiguity of author name and topic name is removed.

We can view the details of the papers which are shown as a result of search by author name or topic name. When the search was done by author name for the author Jonathan Chang all the papers by this author were displayed to get the details of every respective paper.

Figure 4: View Details for a Search

One of the main problems in the current search engines is "rich-get-richer" that causes young high quality pages receive less popularity. In other words, popular high rank pages have more chances to be browsed by users and, consequently, young high quality pages can be the victims. The proposed method solves this problem as it takes into consideration not only the no. of views/ no. of clicks but also the no. of citations and relevance of the topic. The given algorithm depend on the link structure of the documents i.e. their popularity scores as well as the actual content in the documents.

Moreover, this algorithm does not require the parameters like hubs, authorities, forward and back links as in algorithms like HITS and Page Rank algorithm making it easier to comprehend.

The complexity for this is $O(n \log n)$. Though the complexity of the proposed ranking system is high but it is quite easy to implement with the digital libraries.

The comparison between the various ranking scheme is given below.

Table 1
Comparisons of Page Rank Based Algorithms

Algorithm Criteria	*Page Rank PageRank*	*Weighted*	*HITS Rank*	*Distance*	*Dirichlet*	*Our Proposed Ranking System*
Mining technique used	*WSM*	*WSM*	*WSM & WCM*	*WSM*	*WSM*	*WCM*
Working	Computes cores at index time. Results are sorted on the importance of pages.	Computes scores at index time. Results are sorted on the Page importance.	Computes scores of n highly relevant pages on the fly.	Computes scores by calculating the minimum average distance between pages	Works same as PageRank but computes transition Probabilities using Bayesian estimation	Compares various papers based on the ranking factor and sorts them according to WRF

Table Cont'd

I/P Parameters	*Backlinks*	*Backlinks, Forward links*	*Backlinks, Forward Links & content*	*Backlinks*	*Backlinks*	*Database and content of the papers*
Complexity	O(log N)	<O(log N)	<O(log N)	O(log N)	O(log N)	O (N lon N)
Limitations	Query independent	Query independent	Topic drift and efficiency problem	Needs to work along with PageRank	Needs to work along with PageRank	Query dependent(based on the search)
Search Engine	Google	Research model	Clever	Research model	Research model	None

7. FUTURE SCOPE

This proposed content model is easily pluggable and extendible in real-time datasets like CiteSeer digital library etc. As library and information science becomes more and more personcentred, and not just document-centred, we will be expecting to see ripples that will affect the semantic web, world of publishing, the indexing of data collections, and the design of search engines.

REFERENCES

[1] Deepak Gupta, "Content Modelling Paradigm:An Interplay of Relationship Between Author, Document, Topic and Words", IJCA 2010.

[2] Ashutosh Kumar Singh, Ravi Kumar P., "A Comparative Study of Page Ranking Algorithms for Information Retrieval", *IJECE*, 2009.

[3] Jonathan Chang, David M. Blei, "Relational Topic Models for Document Networks", *AISTATS*, 2009.

[4] Yi Guo, Zhiqing Shao, Hua Nan, "Content Oriented Automatic Text Categorization with the Cognitive Situation Models", *ISCSCT*, 2008.

[5] Ian H. Witten, "Text Mining".

[6] Mccallum, A., Corrada-Emmanuel, Andres & Wang, X., "Topic and Role Discovery in Social Networks", *in Proceeding of IJCAI*. 2005.

[7] Blei, D.M. and McAuliffe, J., "Supervised Topic Models". *In Advanced In NIPS*, 2007.

[8] Newman, D., Chemudugunta, C. & Smyth, P., Statistical Entity-topic Models". *In: 10th ACM SigKDD Conference Knowledge Discovery and Data Mining*, 2006.

[9] Minka, T. & Lafferty, J., 2002. "Expectation Propagation for the Generative Aspect Model". *In Proceedings of UAI*.

[10] Branavan, S., Chen, H., Eisenstein, J. and Barzilay, R., 2008. "Learning Document-Level Semantic Properties from Free Text Annotations". *In proceedings of ACL*.

[11] N. Duhan, A.K. Sharma and K.K. Bhatia, "Page Ranking Algorithms: A Survey", *Proceedings of the IEEE International Conference on advance Computing*, 2009.

[12] E. Garfield, "Citation Analysis as a Tool in Journal Evaluation", *Science 178*, pp. 471-479, 1972.

[13] J. Cho, S. Roy and R.E. Adams, "Page Quality: In Search of an Unbiased Web Ranking". *Proc. of ACM International Conference on Management of Data*. pp. 551-562. 2005.

[14] Golwater, S., Griffiths, T.L. and Johnson, M., 2006. "Contextual Dependencies in Unsupervised Word Segmentation". *In Proceedings of Coling/ACL*.

[15] Griffiths, T.L. and Steyyers, M., 2004. "Finding Scientific Topics. *In Proc Natl acad Sci U.S.A Griffiths,T.L. Steyyers, M., Blei, D.M. & Tenenbaumm, J.B.*, 2005.

[16] Integrating Topics and Syntax. In advances in NIPS 17.

[17] Blei, D. and Lafferty, J., 2006. "Dynamic Topic Models". *In Proceedings of the 23rd International Conference on Machine Learning*.

[18] Blei, D. and Lafferty, J., 2007. "A Correlated Topic Model of Science". *In Annals of Applied Statistics*.

[19] Baeza-Yates, R. and Ribieri-Neto, B. (1999). "Modern Information Retrieval". *Addison Wesley Longman*, Essex, England.

[20] Cavnar, W.B. and Trenkle, J.M. (1994). "N-Gram-based Text Categorization". *Proc Symposium on Document Analysis and Information Retrieval*, Las Vegas, NV, pp. 161-175.

[21] X. Wang, N. Mohanty, and A. McCallum. "Group and Topic Discovery from Relations and Text". *Proceedings of the 3rd International workshop on Link Discovery*, 2005.