Journal of Theoretical and Applied Information Technology
30th April 2012. Vol. 38 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A PERSONALIZED ONTOLOGY MODEL FOR WEB INFORMATION GATHERING USING LOCAL INSTANCE REPOSITORY

[1]SURYA NATARAJAN, [2]Mr.J.SETHURAMAN

[1]M.tech (ADC), School of Computing, Sastra University, Thanjavur,

[2]Assistant Professor, School Computing, Sastra University, Thanjavur,

E-mail: suryanatarajan30@gmail.com, lalgudisethu@yahoo.com

ABSTRACT

Ontology describes a standardized representation of knowledge as a set of concepts within the domain, and the relationship between those concepts. Ontology is also used to represent user profiles in personalized web information gathering. For representing user profiles many models have been developed, these models provide knowledge from either a global or local knowledge base. The global analysis uses existing global knowledge bases and to produce effective performance. The local analysis observes user behavior in user profiles. The user background knowledge can be better discovered and represented if we integrate global and local analysis. Our model proposes to bring out data from global resources depending on the user whose profiles match the global content. Compared with other models ontology model has an edge. The LGSM(Local Global Search Methodology) is used for calculating the hit/miss ratio.

Keywords: *Broader-term(BF), Library of Congress Subject Headings(LCSH),Local instance Repository(LIR), Ontology, Related-to(RT), semantic relations, Used-for(UF).*

1. INTRODUCTION

Over the last decade, we have witnessed an explosive growth in the information available on the Web gathering useful information from the web has become a challenging issue for users. The Web users expect more intelligent systems (or agents) to gather the useful information from the huge size of Web related data sources to meet their information needs. The user profiles are created for user background knowledge description [1],[2],[3].

User profiles represent the concept models possessed by users when gathering web information. A concept model is implicitly global analysis method is effective method for gathering the global knowledge.

Local analysis is used for analyzing the user behavior in user profiles. In some works, users provided with set of documents from that background knowledge can be discovered. The user background knowledge can be better discovered if we integrate both global and local information. It can be better improved by using ontological user profiles. A multidimensional ontology mining method, Specificity and Exhaustivity, for analyzing the concept specified machine-readable documents. The goal of ontology learning is to semi-automatically possessed by users and is generated from their background knowledge. This knowledge is used to gather relevant information about a user's preference and choices. World knowledge is a common sense knowledge acquired by people from experience and education.[4].

For representing the user profiles, the user background knowledge must be gathered by using global or local analysis. Global analysis uses worldwide knowledge base for background knowledge representation. The commonly used knowledge bases include generic ontologies e.g. Word net, Thesauruses, digital libraries. The in ontologies. Compared with other benchmark models ontology model is successful.

2. RELATED WORK

2.1Ontology Learning

Ontology learning is also known as ontology extraction and is a subtask of information extraction. Information extraction (IE) is a type of information retrieval whose goal is to automatically extract structured information from unstructured and/or semi-structured

Journal of Theoretical and Applied Information Technology
30th April 2012. Vol. 38 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

extract relevant concepts and relations from a given corpus or other kinds of data sets to form ontology.

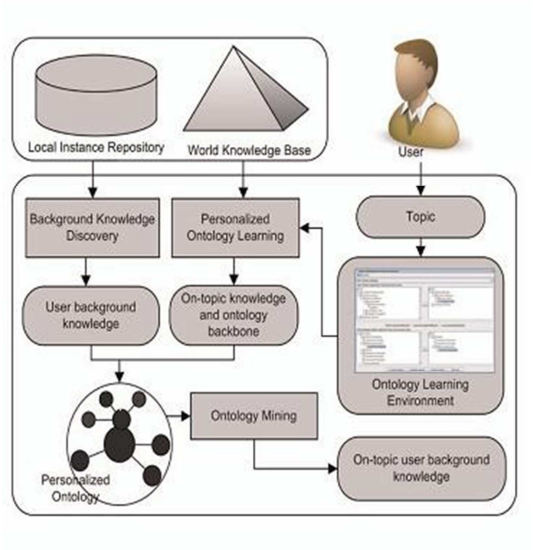

Fig1.Architecture of the ontology model

In this Fig1. It retrieves data based on the local or global information. i.e. it considers only local or global information. It does not consider about the primary key it gathers data based on the user name in local profile. So if two persons having the same name it will retrieve both the information.

2.2Local Profiles

For capturing the user information needs User Profiles were used in web Information gathering. A user profile is a collection of personal data associated to a specific user. A profile refers therefore to the explicit digital representation of a person's identity [11]. A user profile can also be considered as the computer representation of a user model. A profile can be used to store the description of the characteristics of person.

User profiles are categorized into three groups: Interviewing, semi-interviewing, and non-interviewing. Interviewing user profiles are considered to be perfect user profiles. They are acquired by using manual techniques, such as questionnaires, interviewing users, and analyzing user classified training sets. One typical example is the TREC Filtering Track training sets, which were generated manually [4]. The users read each document and gave a positive or negative judgment to the document against a given topic.

Semi-interviewing user profiles are acquired by semi automated techniques with limited user involvement. These techniques usually provide users with a list of categories and ask users for interesting or non interesting categories. One typical example is the web training set acquisition model introduced by Tao et al. [5], which extracts training sets from the web based on user fed back categories. Non interviewing techniques do not involve users at all, but ascertain user interests instead. They acquire user profiles by observing user activity and behavior and discovering user background knowledge [6].

A typical model is OBIWAN, proposed by Gauch et al. [1], which acquires user profiles based on users' online browsing history. The interviewing, semi-interviewing, and non interviewing user profiles can also be viewed as manual, semiautomatic, and automatic profiles, respectively.

3. PERSONALIZED ONTOLOGY CONSTRUCTION

Personalized ontologies that formally describe and specifies user background knowledge. For example a user searching for a word might have different expectations, for searching the same query. For example if we are searching for the term "New Jersey", business travelers may expect different search from leisure travelers. A user may become a business traveler when planning for a business trip, or a leisure traveler when planning for a family holiday. A user's concept model may change according to different information needs.

3.1Global Knowledge Representation

World Knowledge representation research involves analysis of how to accurately and effectively reason and how best to use a set of symbols to represent a set of facts within a knowledge domain. In this model user background knowledge is extracted from a world knowledge base encoded from the Library of Congress Subject Headings (LCSH).

First, step is the construction of world knowledge base. The world knowledge base must cover the wide range of topics, since users expect different results for searching a single word query. The LCSH was developed for organizing and retrieving information from a

Journal of Theoretical and Applied Information Technology
30th April 2012. Vol. 38 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

large volume of library collections. For over a hundred years, the knowledge contained in the LCSH has undergone continuous revision and enrichment and new words are updated. The LCSH represents the natural growth and covers the wide range of exhaustive topics.

The LCSH is considered to be superior and best when compared with other world knowledge taxonomies; Table 1 presents a comparison of the LCSH with the Library of Congress Classification (LCC) used by Frank and Paynter [7], the Dewey Decimal classification (DDC) used by Wang and Lee [8] and King et al. [9], and the reference categorization (RC) developed by Gauch et al. [1] using online categorizations. The LCSH covers large number of topics, has a more specific structure, and specifies more semantic relations. The LCSH descriptors are classified by professionals, and the classification quality is guaranteed. These features make the LCSH an ideal world knowledge base.

The LCSH system contains three types of references: Broader term (BT), Used-for (UF),and Related-Term(RT)[10]. The BT references are for two subjects describing the same topic, but at different levels of abstraction. In our model, they are encoded as the is-a relations in the world knowledge base. The UF references in the LCSH are used for many semantic situations, including broadening the semantic extent of a subject and describing compound subjects and subjects subdivided by other topics.

When object A is used for an action, A becomes a part of that action (e.g., “a fork is used for dining”); when A is used for another object, B, A becomes a part of B(e.g., “a wheel is used for a car”). These cases can be encoded as the part-of relations.

TABLE 1
Comparison of Different World Taxonomies

	LCSH	LCC	DDC	RC
# of Topics	394,070	4,214	18,462	100,000
Structure	Directed Acyclic Graph	Tree	Tree	Directed Acyclic Graph
Depth	37	7	23	10
Semantic Relations	Broader, Used-for, Related-to	Super- and Sub-class	Super- and Sub-class	Super- and Sub-class

3.2 Global knowledge retrieval based on Local profiles

During the global search[18] the data are retrieved based on the information given in the local profile. Depending on the users interest in the local profile data are classified and retrieved from the global search[20].

4. PROPOSED MODEL

In the proposed Ontology model there are two type of search operations are performed[13]. The two type of search operations are local search and global search. For local search it considers only about the local information. For global search it considers about the world knowledge base.

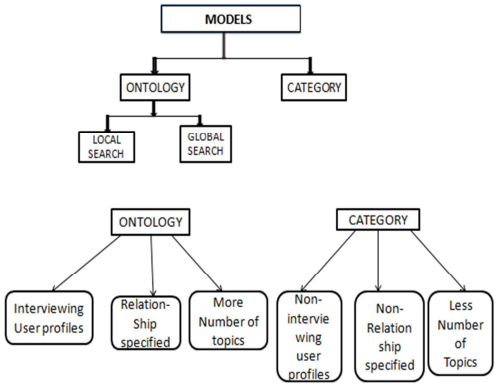

Fig2.Comparision between ontology and category model.

In this Fig.2 it retrieves[16] global information based on the local database because of this time consumption for execution is very less and it gives accurate results, cost is also reduced[20]. It considers mainly the primary key for retrieving the datas so if user is having same name also it fetches and gives the absolute results. It covers wide range of topics in ontology model.

Specificity describes about the percentage of finding how much people satisfying is specificity. Exhaustivity is considering about all the things without omitting anything. In is-a relationships, a parent subject is the abstract description of its child subjects. If a subject has part-of child subjects, the spea(s) of all part-of child subjects takes part of their parent subject’s semantic specificity. As a part-of relation, the concepts referred to by a parent subject are the combination of its part-of child

Journal of Theoretical and Applied Information Technology
30th April 2012. Vol. 38 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

subjects. For performing local search interviewing user profiles is considered.

In interviewing user profiles, is considered to be perfect user profiles because user read each document and gave positive or negative judgments, because users only can judge their interests accurately.

First considering about the local search, for example we have already users information in the database such as name, address, age etc. In this we are classifying the database into four categories such as new user, education, occupation, personal. In this data are classified already fed into the database.

By using the primary key we are linking up all the different category of database[11].When performing local search, it will spilt up and give the results as positive nodes, negative nodes and neutral nodes. For example when we searching the result for the name "raj". In the positive result it will give the correct results matched with raj. In the neutral results it will give the results related with raj, rajkumar. In the negative results it will display even if the middle name consists of raj for eg.muthurajkumar.

In the global search, it consists of three types of knowledge positive subjects, negative subjects, and neutral subjects. In this three nodes[15] are classified based on the relationship such as Is-a, Part-of and Related-to[14]. i.e., Broader-term, Used-for and Related-to. In this especially positive nodes re classified based upon some more[17] relationships such as THINGS: Defined as, Has a, Has Property, Is a, Made of, Part of; SPATIAL: At Location, Located near, Obstructed by; EVENTS: Created by, Has First Sub event, Has Last Sub event; CAUSAL: Causes, Causes Desire; AFFECTIVE: Desires, Motivated by goal; FUNCTIONAL: Receives Action, symbol of, Used-for; AGENTS: Capable-of.

5. ALGORITHM: ANALYZING THE SEMANTIC RELATIONS:

Input: a personalized ontology $\partial(\Gamma) := <tax^S, rel>$;a coefficient Θ between (0,1).

Output: $spea_a(s)$ applied to specificity.

1 set k=1, get the set of leaves S_0 from tax^S, for($s_0 \epsilon S_0$) assign $spe_a(s_0) = k$;
2 get S' which is the set of leaves in case we remove the nodes S_0 and the related edges from tax^S;
3 if (S' == Ø) **then** return;// the terminal condition;
4 **for each** s' ϵ S' **do**
5 **if** (is A(s') == Ø) **then** $spea_a^1(s')=k$;
6 **else** $spea_a^1(s') = \Theta * \min\{spea_a(s)|s\epsilon isA(s')\}$;
7 **if** (part of(s') == Ø) **then** $spea_a^1(s')=k$;
8 **else** $spe_a^2(s')=\sum_{s\epsilon} partOf_{(s')} {}^{spe}{}_a(s)/|partOf(s')|$;
9 $spe_a(s')=\min(spe_a^1(s'),spe_a^2(s'))$;
10 **End**
11 $k=k*\Theta, S_0=S_0 \cup S'$, go to step 2.

Semantic specificity is also called absolute specificity and denoted by $spe_a(s)$[12]. $\partial(\Gamma)$ is a world knowledge base. The semantic specificity is measured based on the hierarchical semantic relations(is-a and part-of) held by a subject and its neighbors in tax^s. The A(s') and part of(s') are two functions in the algorithm satisfying $isA(s') \cap part\ of(s') = \emptyset$.As the tax^s of $\partial(\Gamma)$ is a graphic taxonomy, the leaf subjects have no descendants. If a subject has direct child subjects mixed with is-a and part-of relationships, a $spea_a^1$ and $spea_a^2$ are addressed separately with respect to the is-a and part-of child subjects.

6. METHODOLOGY

The LGSM (Local Global search methodology) it is used to calculate the hit/miss rate. For calculating hit ratio,

$$\text{Hit Ratio} = \frac{\text{Number of Hits}}{(\text{Number of Hits} + \text{Number of Miss})}$$

The performance of memory is frequency measured in terms of quantity is called hit ratio. When cpu needs to find the word in cache, if word is found in cache then it produces a hit. If the word is not found in the cache, it is in main memory it is counted as miss. If it retrieves information from the local repository it is considered as hit. If it retrieves data directly from global it is considered as miss[19].

7. MODELS

7.1Ontology Model

This model was the implementation of the proposed ontology model[13]. The input to this model was a topic and the output was a retrieval of data for the particular search term based on the user profile. The global and local database it can

Journal of Theoretical and Applied Information Technology
30th April 2012. Vol. 38 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

be linked by using a id[15]. For example in the local database if a person has given his/her interest under cricket as Indian team when that particular person searches in the global database it will retrieve the accurate content about the Indian teams instead of retrieving about the contents about the cricket as a common search. In this proposed model query execution time and navigation cost is reduced.

7.2 Category Model

This model is demonstrated by using non-interviewing user profiles that are it do not involve user at all. In particular OBIWAN[12] model. In the OBIWAN model, a user's interests and preferences are described by a set of weighted subjects learned from the user's browsing history. The Category model differed from the Ontology model in that there were *no is-a, part-of, and related-to* knowledge considered and no ontology mining performed in the model[15]. The positive subjects are weighted as one because there were no evidence that the user might prefer some positive subjects more than others.

8. RESULTS

The experiments were designed to compare the results generated by ontology model and the baseline (category) model. The comparison is modeled as an graph in Fig.9. In Ontology model the local profile is used as an aid to search which can bring out precise information based on the user's profile.

8.1 Performance Analysis

The Ontology model has been implemented based on local, global database and based on the semantic relations in .NET framework. The results are,

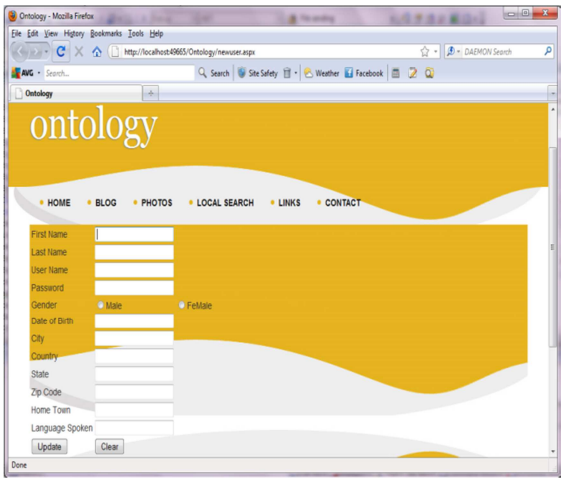

Fig 3. New user signing up

In this Fig.3 describes about new user signing up by creating primary key for each user and for storing up users personal information.
Fig.4 the local search it categorizes data as positive, negative and neutral results.

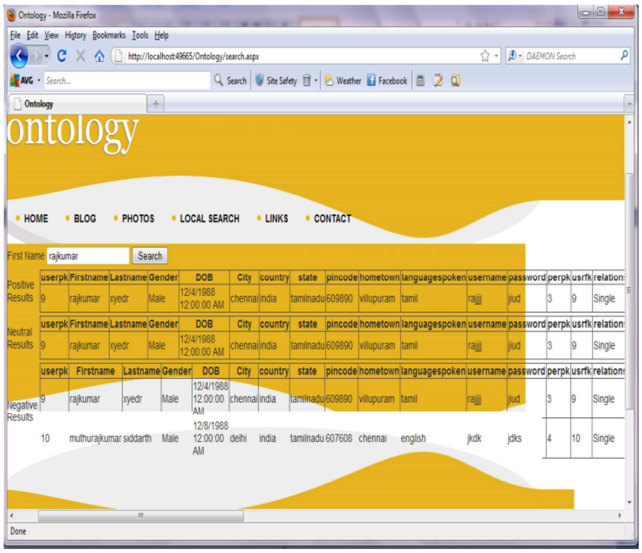

Fig. 4. Local search

In Fig.5 it clusters and classifies the data based on the semantic relations between the objects.

Journal of Theoretical and Applied Information Technology
30th April 2012. Vol. 38 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

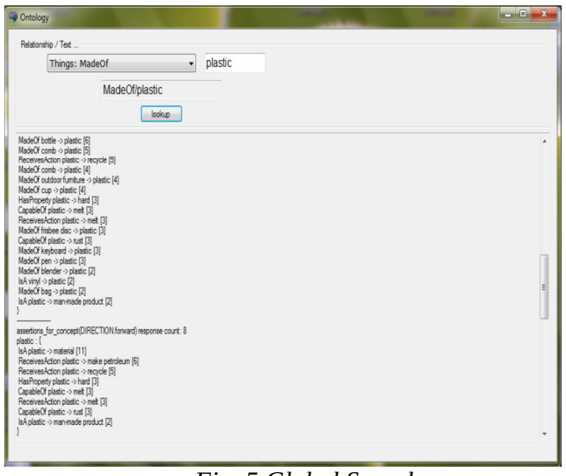

Fig. 5.Global Search

The performance of the models was measured:

$(\sum^{N}_{i=1}\text{precision}\lambda)/N;$

$\lambda=\{0.0,0.1,0.2,\ldots.1.0\}$

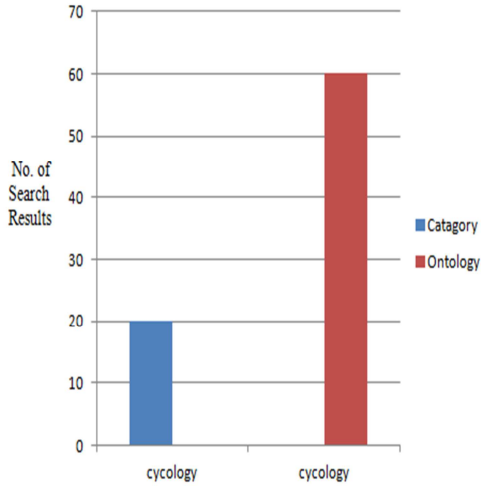

Fig.6. Comparison between ontology and category model

9. CONCLUSION AND FUTURE WORK

In this paper, an Ontology model is proposed for representing user background knowledge for personalized web information gathering. This model constructs the global search from the world knowledge base and local search from local instance repository. This model is compared with the baseline model. In this we found that the combination of local and global works in a better way. In addition, Ontology model using both is-a and part-of is an advantage. In this ontology model, performing both local and global search provides a better solution.

In our future work, we will investigate the methods that generate user local repositories to match the global base. In the present work it has content-based descriptors, a large volume of documents on the global base do not have content-based descriptors for this strategy like ontology mapping and text classification/clustering were suggested.

10.ACKNOWLEDGEMENTS

Author wishes to thank Dr.Sairam. N, Professor and Sethuraman.J, Assistant Professor, Sastra University for their time, linguistic and technical support.

REFERENCES

[1]. S. Gauch, J. Chaffee, and A. Pretschner,"Ontology-Based Personalized Search and Browsing," Web Intelligence and Agent Systems, vol. 1, nos. 3/4, pp. 219-234, 2003.

[2]. Y. Li and N. Zhong, "Web Mining Model and Its Applications for Information Gathering," Knowledge-Based Systems, vol. 17, pp. 207-217, 2004.

[3]. Y. Li and N. Zhong, "Mining Ontology for Automatically Acquiring Web User Information Needs,"IEEE Trans. Knowledge and Data Eng., vol. 18, no. 4, pp. 554-568, Apr. 2006.

[4]. S.E. Robertson and I. Soboroff, "The TREC 2002 Filtering Track Report," Proc. Text Retrieval Conf., 2002.

[5]. X. Tao, Y. Li, N. Zhong, and R. Nayak, "Automatic Acquiring Training Sets for Web Information Gathering," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence, pp. 532-535, 2006.

[6]. J. Trajkova and S. Gauch, "Improving Ontology-Based User Profiles," Proc. Conf. Research 'Information Assistee par Ordinateur (RIAO '04), pp. 380-389, 2004.

[7]. E. Frank and G.W. Paynter, "Predicting Library of Congress Classifications from Library of Congress Subject Headings," J. Am. Soc. Information Science and Technology, vol. 55, no. 3, pp. 214-227, 2004.

[8]. J. Wang and M.C. Lee, "Reconstructing DDC for Interactive Classification," Proc. 16th ACM Conf. Information and Knowledge Management (CIKM '07), pp. 137-146, 2007.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[9]. J.D. King, Y. Li, X. Tao, and R. Nayak, "Mining World Knowledge for Analysis of Search Engine Content," Web Intelligence and Agent Systems, vol. 5, no. 3, pp. 233-253, 2007.

[10]. L.M. Chan, Library of Congress Subject Headings: Principle and Application. Libraries Unlimited, 2005.

[11].Claudia Marinica and Fabrice Guillet,"Knowledge based interactive post-mining of association rules using Ontology" IEEE Transactions on Knowledge and data engineering, Vol. 22, NO. 6, June 2010.

[12]. T. Tran , P. Cimiano, S. Rudolph, and R. Studer,"Ontology-Based Interpretation of KeywordsforSemantic Search, " Proc. Sixth Int'l Semantic Web and Second Asian Semanti Web Conf.(ISWC '07/ASWC '07), pp.523-536, 2007.

[13] Z. Cai, D.S. McNamara, M. Louwerse, X. Hu, M.Rowe, and A.C. Graesser, "NLS: A Non-Latent Similarity Algorithm," Proc . 26th Ann. Meeting of the Cognitive Science Soc. (CogSci '04), pp. 180-185,2004.

[14]. P.A. Chirita, C.S. Firan, and W. Nejdl, Personalized Query Expansion for the Web," Proc ACM SIGIR ('07),pp. 7-14, 2007.

[15]. Xiaohui Tao, Yuefeng Li, and Ning Zhong, Senior Member, IEEE,"A Personalized ontology model for Information gathering" IEEE Transactions on Knowledge and data engineering.

[16]. M.D. Smucker, J. Allan, and B. Carterette, "A Comparison of Statistical Significance Tests for Information Retrieval Evaluation,"Proc. 16th ACM Conf. Information and Knowledge Management (CIKM '07),pp. 623-632,2007.

[17]. D. Downey, S. Dumais, D. Liebling, and E. Horvitz, "Understanding the Relationship between Searchers' Queries and Information Goals," Proc. 17th ACM Conf. Information and Knowledge Management (CIKM '08), pp. 449-458, 2008.

[18]. D.N. Milne, I.H. Witten, and D.M. Nichols, "A Knowledge-Based Search Engine Powered by Wikipedia," Proc. 16th ACM Conf. Information and Knowledge Management (CIKM '07), pp. 445-454, 2007.

[19] X. Tao, Y. Li, N. Zhong, and R. Nayak, "Automatic Acquiring Training Sets for Web Information Gathering," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence, pp. 532-535, 2006.

[20]. W. Jin, R.K. Srihari, H.H. Ho, and X. Wu, "Improving Knowledge Discovery in Document Collections through Combining Text Retrieval and Link Analysis Techniques," Proc. Seventh IEEE Int'l Conf. Data Mining (ICDM '07), pp. 193-202, 2007.