Information Extraction Using Metadata andSolving Polysemy Problems

Iswarya R.J[1], Bharathi.N[2]
School of Computing, SASTRA University, Thanjavur-613401, TamilNadu, India.
[1] iswaryajeyaraman@gmail.com
[2]bharathi_n@cse.sastra.edu

***Abstract*—Data mining is the exploration and evaluation of large quantity of data to discover substantial, novel, useful and effectively understandable data. Hence determining the knowledge of a document becomes a necessary task in data mining. There are three approaches of metadata in general. They are stylistic, machine learning and knowledge bases. Sometimes the problem occurs when mining a document that contains polysemic words which leads to irrelevant extraction and increased processing time. Polysemy refers to coexistence of many possible meaning for a word or phrase. In order to extract exact information, polysemy like issue should be solved. This work uses knowledge based metadata to extract information using Domain-based Information Extraction technique (DIE). Hence this work targets in solving polysemy which can increases the accuracy of information extraction and reduce processing time. By applying this method to a enormous amount of Engineering domains contains fields like computer science, biomedical, nanotechnology, physics, this work shows that the information extraction is efficient for day-to-day applications with reduced processing time.**

Keyword-Data mining, Information extraction, Metadata, Polysemy, Domain-based extraction.

I. INTRODUCTION

The amount of information increases as the technology grows. The number of the databases over the engineering domain increases even faster. It is important to generate metadata and file tagging for the documents, in order to discover knowledge from a large collection of engineering information. Obtaining information depending upon the predefined pattern, so called “data about data” is the major function of Information Extraction (IE) system [10]. The cooperative work on IE and metadata will achieves both sides [18],[15].

Metadata is a keynote that ensures property will persists, maintained and handled to use into the future needs. Metadata is the systematized information that describes file name, size, and type, locates path [12]. Generated metadata can be stored in the repository or relational database (RDBMS). The major four functions of metadata include Identifying content which is a descript metadata, Managing content which is an administrative or structural metadata, Retrieving content which is a descriptive metadata and track usage of content like user rating, link data, downloads and forwards data[1],[19].

Metadata makes the documents easier to recover, use or handle information. Hence from [2], Metadata narrates elements including data, articles, humans etc... Also from [2], a super class metadata can efficiently characterize element and easily categories element from element.

Sometimes problem occurs when reader attempt to retrieve a document that contains polysemic words which leads to irrelevant extraction and increased processing time. This type of problem can lead to ambiguity. One such type of ambiguity is called polysemy. Polysemy refers to coexistence of many possible meaning for a word or phrase [22]. For example, the word "mole" has several meanings. Mole: the animal; a spy; skin mark; chemistry unit; an unofficial holiday; popular dessert in Brazil; record label; video game etc. When concentrating on the engineering databases, mole with meaning “chemistry unit” is relevant. But all other meanings like animal; a spy; skin mark etc to the same word ‘mole’ is irrelevant to the engineering domain based documents.

Hence efforts are needed to extract exact information when concentrating on the engineering databases alone. In order to extract exact information, polysemy like issue should be solved[1],[4],[5]. This can be achieved by proposed technique called Domain-based Information Extraction (DIE). Hence this work targets in solving polysemy which can increases the accuracy of information extraction and reduces the time taken for processing.

In order to understand this proposed extraction methodology, choosing database management systems is the right choice to address the run time extraction requirements. The proposed information extraction methodology is composed of two stages:

Text processing stage:This phase performs processing of text document with uploading file, metadata generation and tagging of the entire document and the result is stored in the RDBMS.

Domain-based Extraction stage: Extraction can be obtained by sending queries to the databases. The extraction pattern over solved polysemy problem can be expressed as Domain-based Information Extraction (DIE) technique.

In this extraction framework, intermediary result of each data processing module is stored. Thus only the upgraded component is adapted to the whole collection. Then the Knowledge discovery is done on the already processed data from the unaffected components and modified data produced by the enhanced component. Hence this technique reduces the reprocessing time of the whole corpus of data in the current extraction goals and deployment of enhanced processing components. Since this knowledge discovery is specified as queries, a reader does not require to code and run the programs for each specific extraction.

II. RELATED WORKS

A. *Metadata and text processing*

Kuang-hua Chen narrates the link among metadata and data extraction. Then developed a fast parsing technique for information processing. The proposed parsing algorithm has the ability of quick analyzing documents. The main advantage of proposed algorithm is only a Part of speech information required to build a parsing algorithm which can be easily constructed Kuang-hua Chen not only concentrate on the relationship among metadata, IE, and IR, but also narrates how to apply NLP technology to automating the IE tasks. Kuang-hua Chen also evaluated the performance with respect to two different parameters include parseval and accuracy [13].

B. Classification of polysemy

Pustejovsky[8], proposes a classification of polysemy, distinguishing two types. The first type is the complementary polysemy. This involves cases where the senses of a word are overlapping, dependent or shared. An example of complementary polysemy can be seen with the word hammer. It can refer to a physical object and to an action. The sense difference is accompanied with a change in category, the first sense associated with usage as a noun, and the second as a verb. The second type is the non-complementary polysemy. This type of polysemy is independent. A more specific type of complementary polysemy [25] is logical polysemywhich is constrained to cases where there is no change in lexical category. The noun door can refer to an opening and to a physical object. The senses are related since one can refer to both senses within a single sentence without any problem: He walked through the red door. The phrase walked through evokes the opening sense, while the adjective red evokes the physical object sense.

C. Distinguish Polysemy From Vagueness

Kilgarriff, Lewandowska-Tomaszczyk [21],[24], distinguished polysemy from vagueness(or generality). Lewandowska-Tomaszczyk exemplifies with the noun *student*. It can be used equally well to describe a man or a woman. That does not necessarily mean that *student* has two senses, one for ‘male student’ and another for ‘female student’. Instead, this verb is vague regarding gender; it is unmarked for this characteristic. In that spirit, many different types of linguistic tests have been proposed in order to distinguish cases of polysemy from.

D. Verdikt framework For Metadata Extraction

Christian Schonberg and Burkhard Freitag[12],[10], proposed a framework called Verdikt model which shows different methods for metadata extraction from text documents. The Document Model is the package comprises of abstract classes and interfaces which indicates the document model. The Metadata Model is the interface that indicates the document model as a entire, managing all data and provides access to its statements. The VerdiktObject is the base interface base for the entities that make up the collection of sentences and data of the document model. The Data Object is the interface that indicates simple data entries.

III. PROPOSED SYSTEM DESIGN

The proposed system consist of two stages: *text processing stage* for processing of text document and *domain based extraction stage* make use of relational database queries to achieve exact information extraction by solving polysemy problem. The text processing phase is responsible for uploading files to the server, generating metadata and file tagging of the uploaded file, and storage of the generated information in the *relational databases* (RDBMS). The extraction pattern over solved polysemy problem can be expressed as Domain-based Information Extraction (DIE) technique.

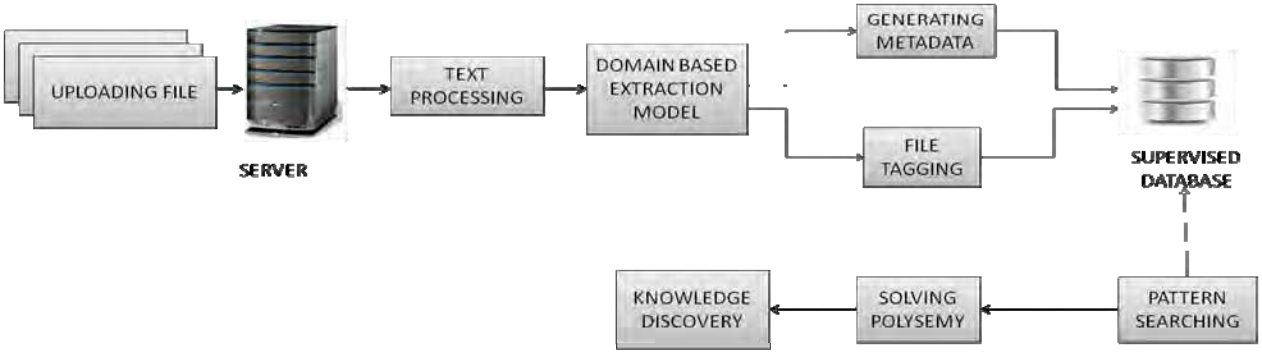

Fig.1. DOMAIN-BASED INFORMATION EXTRACTION (DIE) MODEL

A. Uploading Files To The Server

Multiplefiles can be uploaded to the server. The files that were uploaded successfully are moved to a given server directory. Authorization to Upload files only for Authors who are all registered. Other users or readers can view or retrieve the document. Authors can upload any kind of document to the server. If the author needs to upload document irrelevant to engineering domain, the author should mention their domain as general.On the other hand if the author needs to upload documents relevant to the engineering domain, the authors should mention the domains which their documents comes under like computer science, biomedical, nanotechnology, physics etc., The Administrator maintains these uploaded files in server Directory. Once the authors uploaded file successfully, metadata and file tagging has been generated and stored in the *relational databases*(RDBMS)[16].

B. Generating Metadata And File Tagging

The three types of metadata in general [6],[7], include *informative metadata* to generate resources for discover properties like file name, size, type, keyword, tagging, subject and path. *Structural metadata* is used to denote how compound objects are composed together. *Organizational metadata* to organize information to manage resources like cause of making the data, goal of the information, duration of making, path of the data, creator or author of the data. This work makes use of descriptive metadata to extract file name, size, type, keyword, tags, subject, path and administrative metadata to extract creation of the data, goal of the information, duration of making, author of the data, and path of the documents. File tagging generates the keywords or pronouns of each document and stored in the relational databases.

C. Supervised Databases

The assortments of voluminous engineering documents were collected and it is supervised by means of different domains like computer engineering, nanotechnology, etc., and is stored in the database. It is used to collect all the types of words, their real, associated meaning, and high level meaning to the particular word from the dictionary and to train all the types of words in the database. In this proposed framework, it will match with any kind of unsupervised extracted keywords based on domain to the database and then it solves the problem automatically. In Existing framework, the problem was solved only from trained extracted keywords. Using this proposed supervised database, it is possible to solve problems from any type of supervised/unsupervised occurrence of keywords.

D. Domain-Based Extraction

In domain based extraction, the supervised database is utilized for solving the polysemy problem by comparing the user data with supervised data. Supervised database consists of trained collection of documents isolated according to the different domains. Keywords were generated for each document based on the domain and maintained in the supervised database. When the user attempts to retrieve information, user needs to specify the content of document they want to find using domain-based information extraction technique. Once the user has specified their domain and searches using one of the keyword, the supervised database matches with its keywords stored in it. Once the user's input matches with the keyword stored in the supervised database, then it allows the users to view or retrieve the exact information.

If the user does not specify their domain, then the user cannot retrieve the exact information. The supervised database will match the keyword in general and not according to the domain-based information extraction. Hence the user can get the information in general which will not be relevant to the particular domain. This can lead to ambiguity and irrelevant information extraction.

E. Knowledge Discovery

Knowledge is then extracted by sending database queries to Relational Database. Before issuing query, the users or readers have to select domain to indicate which type of data they want to find. To denote

Knowledge patterns, this work structured and implemented a pattern searching using Domain-based Information Extraction (DIE) that helps for generic knowledge discovery. In the event of a variation to the knowledge extraction goals (For example, the reader gets involved in recent kind of interactions among entities), the reputed modules is developed for the whole data collection and the progressed data are gathered. The queries are reported for classifying the collection of words with recently identified mentions. This kind of knowledge extraction can be achieved only on such affected sentences instead of whole collection. Hence it is possible to achieve exact discovery of knowledge, which repudiate the need to retrogress the whole collection of text compared to the file-based pipeline approaches.

IV. EXPERIMENTS AND RESULTS

From this approach, keywords have been extracted from the supervised database which describe about some topic. Hence, Polysemy problems were solved by using supervised database. In this framework also can solve the polysemy problem from untrained occurring of keyword. Finally Domain based technique was supervised using conditions and most probable keywords are extracted and exact knowledge discovery has been achieved.

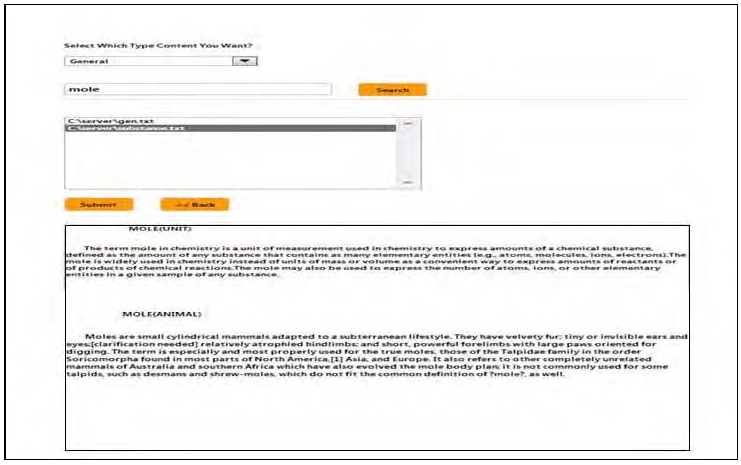

FIG.2. BEFORE SOLVING POLYSEMY

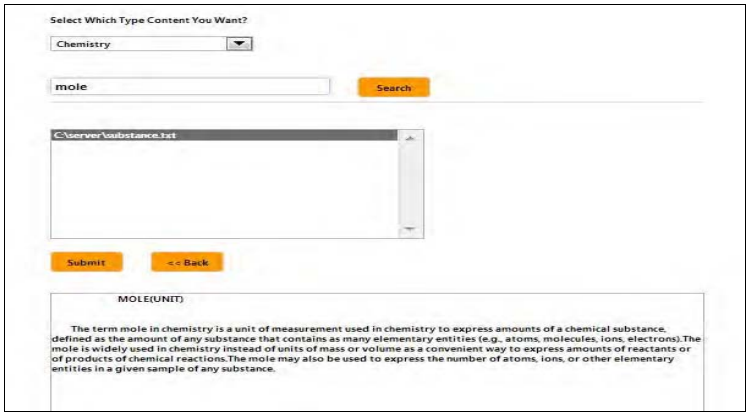

FIG.3. AFTER SOLVING POLYSEMY

Fig.2 shows the result before solving polysemy problem which leads to irrelevant information extraction. On the other hand, Fig.3 shows the result after solving polysemy problem using domain-based information extraction. Hence this technique has showed the exact information extraction on the user needs.

V. CONCLUSION

This study is primarily an investigation of solving polysemy problem when extracting information in the engineering domain. The proposed domain-based extraction have showed how to extract exact information in engineering domain by isolating keywords according to the particular field with the help of supervised database. Supervised database manages the collection of keywords. Here keywords have been extracted from the supervised database which describe about some topic. Polysemy problems were solved by using supervised database and this method extract the information very exactly than the existing techniques. Also the other

existing techniques do not have the capacity for managing intermediate processed data which can be computationally expensive. Hence the proposed technique reduces reprocessing of the entire text collection and the computation speed. As indicated in this experiment, the exact extraction approach saves much more time compared to performing extraction by unsupervised database and then other components.

REFERENCES

[1] Fleur Mougin, Olivier Bodenreider and Anita Burgun, "Analyzing polysemous concepts from a clinical perspective: Application to auditing concept categorization in the UMLS," Journal of Biomedical Informatics 42 (2009) 440–451.

[2] Kuang-hua Chen and Taipei, "Digital Libraries, Metadata and Text Processing", Proceedings of the First Asia Digital Library Workshop 6-7 August 1998, Hong Kong, pp. 123-135.

[3] Kristina M. Irvin, "comparing information retrieval effectiveness of different metadata generation methods," A Master's paper for the M.S. in I.S. degree. April, 2003.

[4] Emma Skalmann, "The Interplay of Synonymy and Polysemy," Master's Thesis in Theoretical Linguistics (LIN-3990), Springer 2012.

[5] GergelyPetho, "what is polysemy? —a survey of current research and results," OTKA (national scientific foundation), grant no. T 030295, 1999.

[6] Paul Clough,"Extracting Metadata for Spatially-Aware Information Retrieval on the Internet," University of Sheffield, Western Bank, *GIR'05*, November 4, 2005, Bremen, Germany. Copyright 2005 ACM 1-59593-165-1/05/0011.

[7] Kevin Yao, "Header Metadata Extraction from Scientific Documents ".

[8] Robert Krovetz, "Homonymy and Polysemy in Information Retrieval," NEC Research Institute,"4 Independence Way Princeton, NJ. 08540 Department of Commerce raider cooperative agreement number EEC-9209623.

[9] Jun Araki, "Text Classification with a Polysemy Considered Feature Set," A Thesis Presented to Graduate School of Engineering at the University of Tokyo for the Degree of Master in 2003.

[10] Charles H. Heenan, "A Review of Academic Research on Information Retrieval," Engineering Informatics Group Department of Civil and Environmental Engineering Stanford University, August 6, 2002.

[11] Kevin Yao, "Header Metadata Extraction from Scientific Documents".

[12] Christian Schonberg and Burkhard Freitag, "Extracting and Storing Document Metadata," Technical report, Number MIP-0907, University of Passau.

[13] Krovetz R and W B Croft, "Lexical Ambiguity and Information Retrieval", A CM Transactions on Information Systems, pp. 145-161, 1992.

[14] Krovetz R, "Word Sense Disambiguation for Large Text Databases", PhD dissertation, University of Massachusetts. 1995.

[15] Braschler, M. and Schauble, P. (2000). Using corpus-based approaches in a system for multilingual information retrieval. Information Retrieval, 3, PP. 273–84.

[16] Hui Han, C. Lee Giles, ErenManavoglu, HongyuanZha, Zhenyue Zhang, and Edward A. Fox. 2003. Automatic document metadata extraction using support vector machines. In Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries (JCDL'03). IEEE Computer Society, Washington, DC, USA, 37-48.

[17] Erik Hetzner. 2008. A simple method for citation metadata extraction using hidden markov models. In Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries (JCDL '08). ACM, New York, NY, USA, 280-284. DOI=10.1145/1378889.1378937 http://doi.acm.org/10.1145/1378889.1378937.

[18] Cowie, J. and Lehnert, W. (1996) Information Extraction. Communications of the ACM, 39(1), 80-91.

[19] Lam, Wai; Lai, Kwok-Yin. "A Meta-Learning Approach for Text Categorization" Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pages 303 – 309. September 2001.

[20] Belkin, Nicholas J.; Croft, Bruce W. "Information Filtering and Information Retrieval: Two Sides of the Same Coin?" Communications of the ACM, Volume 35, Issue 12, Pages 29 – 38. December 1992.

[21] Yarowsky, D. (1992). Word Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. Proceedings of 15th International Conference on Computational Linguistics, pp.454–60.

[22] Cowie, J. and Lehnert, W. (1996) Information Extraction.*Communications of the ACM*, 39(1), 80-91.

[23] FumiyoFukumoto, "Toward Optimal Feature Selection for Word Sense Disambiguation (in Japanese)", *Information Processing Society of Japan, Special Interest Group on Natural Language Processing (IPSJ-SIGNL)*, 2001-NL-141-12, pp.69-76, 2001.

[24] David Aumüller. 2009. Retrieving metadata for your local scholarly papers. *BTW 2009*.