International Journal of Modern Engineering Research (IJMER)
www.ijmer.com Vol.3, Issue.1, Jan-Feb. 2013 pp-564-567 ISSN: 2249-6645

Content Extraction with text mining using natural language processing for anatomy based topic summarization

K.FouziaSulthana, [1] N.Kanya[2]

[1]*Final Year Student, M.Tech CSE Department, Dr.M.G.R.Educational and Research Institute University, Tamil Nadu, India*
[2]*Associate Professor of CSE and IT Department, Dr.M.G.R.Educational And Research Institute University, Tamil Nadu, India*

Abstract: *Inorder to get exact content for the web browsers when searching for some information we have combined two methods together.The first method is text mining using natural language processing and parse tree query language, the second method is TSCAN (Topic summarization and Content Anatomy).*

In the first method, when a sentence is given for searching, the sentence is converted into a automatic query formation using natural language processing tools and information is retrieved using text mining methods.

In the second method a temporal similarity (TS) function is implemented to generate the event dependencies and context similarity to form an evolution graph of the topic search.
Both methods are integrated together to make the quality of the extraction of required content efficient and easier.

Keywords:*TSCAN, Natural Language Processing, TextMining, Information Extraction,PTQL*

I. Introduction

Information Extraction is a process which develops methods for fetching structured information from natural language text. The extraction of entities and relationships between entities is the best example of structured information.

INFORMATION EXTRACTION is typically seen as a one-time process for the extraction of a particular kind of relationships of interest from a document collection .The purpose of information extraction (IE) is to find desired pieces of information in natural language texts and stores them in a form that is suitable for automatic processing. IE is usually deployed as a pipeline of special-purpose programs, which include sentence splitters, tokenizes, named entity recognizers, shallow or deep syntactic parsers, and extraction based on a collection of patterns.

It provides automated query generation components so that casual users do not have to learn the query language in order to perform extraction.

We performed experiments to highlight two important aspects of an information extraction system: efficiency and quality of extraction. By applying our methods to a number of records in databases, the benefit of this method is efficient for real-time applications. The Text Processor in theis responsible for corpus processing and storage of the processed information in the Parse Tree Database (PTDB). The extraction patterns over parse trees can be expressed in our proposed parse tree query language. The Parse Tree Query Language(PTQL)query evaluator takes a PTQL query and transforms it into keyword-based queries and SQL queries, which are evaluated by the underlying RDBMS and information retrieval (IR) engine. To speed up query evaluation, the index builder creates an inverted index for the indexing of sentences according to words and the corresponding entity typesThe architectural diagram is shown in fig.1, clearly describes our approach.

ARCHITECTURAL DIAGRAM:

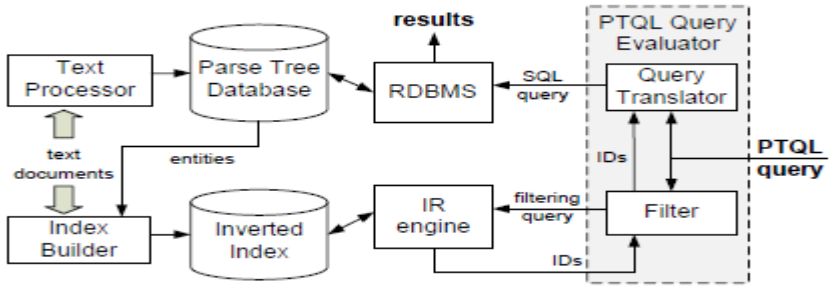

Fig.1

A typical IE setting involves a pipeline of text processing modules in order to perform relationship extraction.

These include:
Sentence splitting: identifies sentences from a paragraph of text
Tokenization: identifies word tokens from sentences
Named entity recognition: identifies mentions of entity types of interest
Syntactic parsing: identifies grammatical structures of sentences
Pattern matching: obtains relationships based on a set of extraction patterns that utilize lexical,syntactic and semantic features

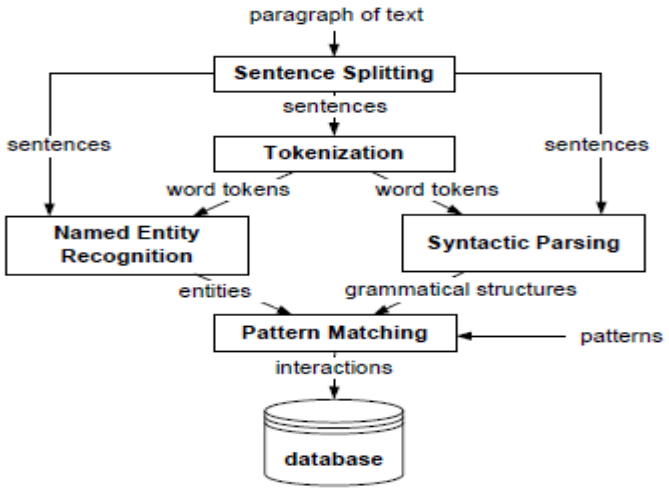

Fig. 2. A workflow of text processing modules that takes a paragraph of text as input to perform relationship extraction.

II. Tscan System

In this section, we present our model and the methods used in the proposed topic anatomy system.

2.1 Topic Model:

A topic is a real world incident that comprises one or more themes, which are related to a finer incident, a description,or a dialogue about a certain issue. During the lifespan of a topic, one theme may attract more attention than the others,and is thus reported by more documents. We define an event as a significant theme development that continues for a period of time. Naturally, all the events taken together form the storyline(s) of the topic. Although the events of a theme are temporally disjoint, they are considered semantically dependent in order to express the development of the theme. Moreover, events in different themes may be associated because of their temporal proximity and context similarity. The proposed method identifies themes and events from the topic's documents, and connects associated events to form the topic's evolution graph. In addition, the identified events are summarized to help readers better comprehend the storyline(s) of the topic. Fig. 3illustrates the relationships between the themes, events, and event dependencies of a topic in the proposed model.

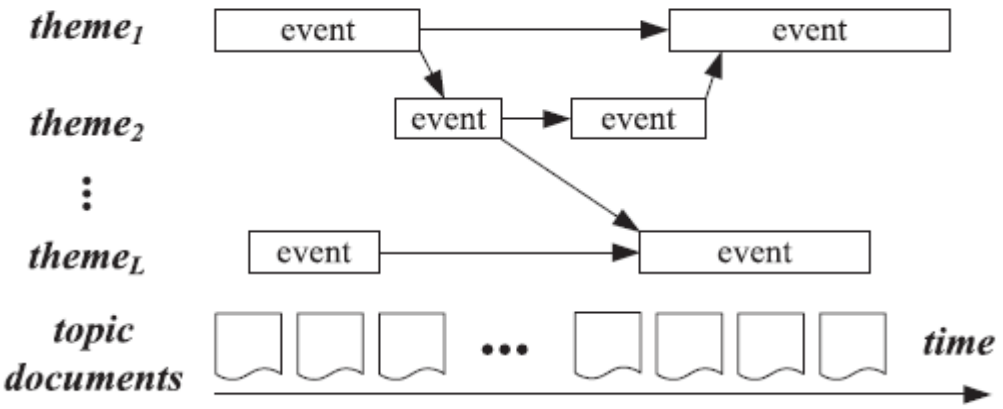

Fig. 3. The relationships between themes, events, and event dependencies.

2.2 Theme Generation:

A matrix $A=B^TB$ is called a block association matrix, is asymmetric matrix in which the (i,j)-entryis the inner product of columns i and j in matrix B. As a column of B is the term vector of a block, A represents the interblock association. Hence, entries with a large value imply a high correlation between the corresponding pair of blocks. A theme of

a topic is regarded as an aggregated semantic profile of a collection of blocks, and can be represented as a vector v of dimension n, where each entry denotes the degree of correlation of a block to the theme.

2.3 Event Segmentation and Summarization:

A theme $v_j$in is a normalized eigenvector of dimension n, where the (i,j) entry indicates the correlation between a block i and a theme j. As topic blocks are indexed chronologically, a sequence of entries in $v_j$with high values can be taken as a noteworthy event embedded in the theme, and valleys (i.e., a sequence of small values) in $v_j$may be event boundaries. However,according to the definition of eigenvectors, the signs of entries in an eigenvector are invertible. Both the positive and negative entries of an eigenvector contain meaningful semantics for describing a certain concept embedded in a document corpus and the amplitude of an entry determines the degree of its correlation to the concept. The tasks of event segmentation and speech endpoint detection are similar in that they both try to identify important segments of sequential data. In addition, it is the amplitude of sequential data that determines the data's importance.

2.3 Evolution Graph Construction:

Automatic induction of event dependencies is often difficult due to the lack of sufficient domain knowledge and effective knowledge induction mechanisms. However, as event dependencies usually involve similar contextual information, such as the same locations and person names, they can be identified through word usage analysis. Our approach, which is based on this rationale, involves two procedures. First, we link events segmented from the same theme sequentially to reflect the theme's development. Then, we use a temporal similarity function to capture the dependencies of events in different themes. For two events, ei and ej, belonging to different themes, we calculate their temporal similarity between these two events and providing the graph description from the result.

III. Integrating NLP and PTQL with Text Mining and Anatomy Based Topic Summarization

Natural Language Processing(NLP) is a field of computer science,artificial intelligence,and linguistics concerned with the interactions between computers and natural languages. It is used to transfer normal sentence to structured query text.

Since we are integrating both these methods,it will be easy for searching and retrieving documents.We can see the parse trees developed for the queries for the givensentences. Integrated architecture is as shown in the fig 4.

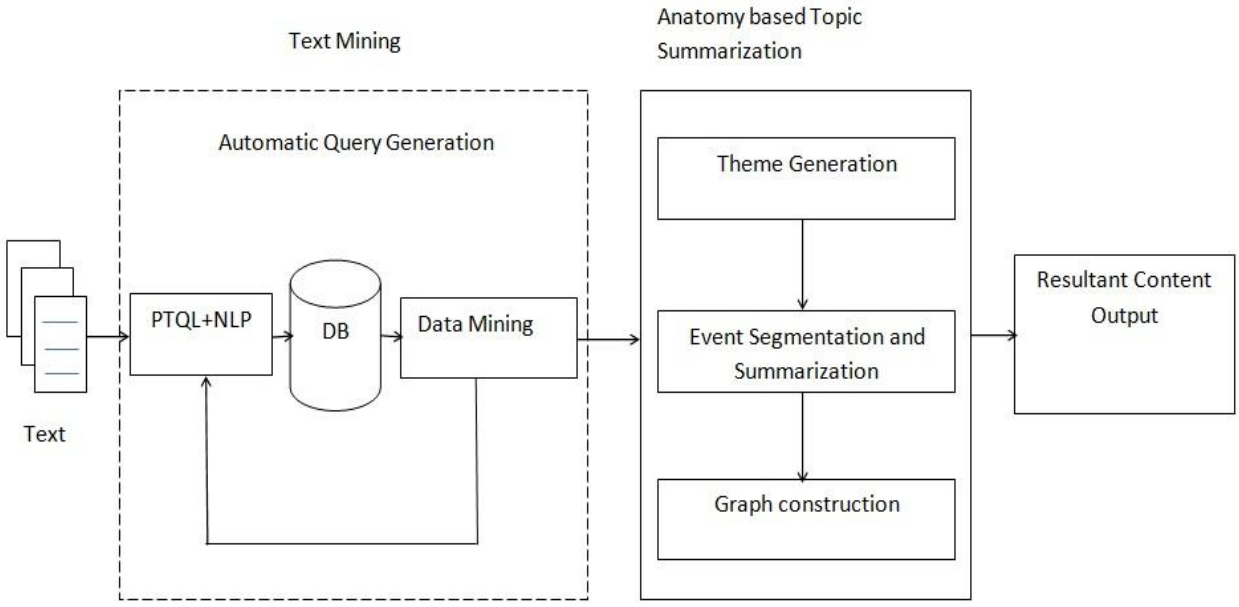

Fig 4 Integrated architectural diagram for content extraction

IV. Discussion and Conclusion

PTQL has the ability to perform a variety of information extraction tasks by taking advantage of parse trees unlike other query languages. Currently PTQL lacks the support of common features such as regular expression as frequently used by entity extraction task. PTQL also does not provide the ability to compute statistics across multiple extraction such as taking redundancy into account for boosting the confidence of an extracted fact.

Publishing activities on the Internet are now so prevalent that when a fresh news topic occurs, autonomous users may publish their opinions during the topic's life span. To helpInternet users grasp the gist of a topic covered by a large number of topic documents, text mining and topic summarization methods have been proposed to highlight the core information in the documents and also for the automatic query generation.

References

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang,"Topic Detection and Tracking Pilot Study: Final Report," Proc. USDefense Advanced Research Projects Agency (DARPA) Broadcast NewsTranscription and Understanding Workshop, pp. 194-218, 1998.
[2] V. Hatzivassiloglou, L. Gravano, and A. Maganti, "An Investigationof Linguistic Features and Clustering Algorithms for TopicalDocument Clustering," Proc. 23rd Ann. Int'l ACM SIGIR Conf.Research and Development in Information Retrieval, pp. 224-231, 2000.
[3] C.D. Manning, P. Raghavan, and H. Schutze, Introduction toInformation Retrieval. Cambridge Univ. Press, 2008.
[4] Y. Yang, T. Pierce, and J. Carbonell, "A Study on Retrospectiveand Online Event Detection," Proc. 21st Ann. Int'l ACM SIGIRConf. Research and Development in Information Retrieval, pp. 28-36,1998.
[5] C.C. Chen, M.C. Chen, and M.S. Chen, "An Adaptive ThresholdFramework for Event Detection Using HMM-Based Life Profiles,"ACM Trans. Information Systems, vol. 27, no. 2, pp. 1-35, 2009.
[6] D. D. Sleator and D. Temperley, "Parsing English with a Link Grammar," in Third Intl. Workshop on Parsing Technologies, 1993.
[7] R. Leaman and G. Gonzalez, "Banner: An executable surveyof advances in biomedical named entity recognition," in PacificSymposium on Bio computing 13, 2008, pp. 652–663.
[8] A. R. Aronson, "Effective mapping of biomedical text to the umls met thesaurus: the metamap program." in Proceedings of the AMIASymposium. American Medical Informatics Association, 2001,p. 17.
[9] M. J. Cafarella and O. Etzioni, "A search engine for naturallanguage applications," in WWW'05, 2005.
[10] T. Cheng and K. C.-C. Chang, "Entity search engine: Towardsagile best-effort information integration over the web," in CIDR,2007.
[11] H. Bast and I. Weber, "The CompleteSearch Engine: Interactive,efficient, and towards IR& DB integration," in CIDR, 2007, pp.88–95.
[12] S. Bird, Y. Chen, S. B. Davidson, H. Lee, and Y. Zheng, "Extendingxpath to support linguistic queries," in Workshop on ProgrammingLanguage Technologies for XML (PLAN-X), 2005.
[13]S.Geetha,G.S.Ananda Mala,N.Kanya,"A survey on information extraction using entity relation based methods,"SEISCON 2011.
[14] N.Kanya,S.Geetha,,"Information Extraction –A Text mining approach"ICTES 2007.