International Journal of Information Technology and Management

Vol. IV, Issue No. I, February-2013, ISSN 2249-4510

A STUDY OF TEXT MINING FOR WEB INFORMATION RETRIEVAL SYSTEM

AN INTERNATIONALLY INDEXED PEER REVIEWED & REFEREED JOURNAL

www.ignited.in

International Journal of Information Technology and Management
Vol. IV, Issue No. I, February-2013, ISSN 2249-4510

A Study of Text Mining for Web Information Retrieval System

A. Sindhu[1]* Dr. C. A. Kanabar[2]

[1]Laxmi Institute of Commerce and Computer Application (BBA-BCA) Sarigam

[2]Assistant Professor (Computer Science) Saurastra University, Rajkot

Abstract – Text Information Retrieval is also known as text data mining or knowledge discovery from textual databases refers to the method of extracting interesting and non-trivial form or information from text documents. Consider by many as the next signal of knowledge discovery, text Information Retrieval has very elevated commercial principles. Previous calculate reveals that there are many modern companies present products for text Information Retrieval. Has text Information Retrieval developed so quickly to become a grown-up field? This paper attempts to discard some lights to the query. We first present a text Information Retrieval framework consisting of two components: Text cleansing that converts shapeless text documents into a middle form; and knowledge refinement that deduces patterns or knowledge from the middle form. We then study the state of the art text Information Retrieval products or applications and align them based on the text and knowledge cleansing functions as well as the middle form that they accept. In ending, we show up the upcoming challenges of text Information Retrieval.

Keywords: Text Mining, Web Information, Retrieval System, Data Mining, Knowledge Discovery, Database, etc.

----------♦----------

INTRODUCTION

This paper presents a general framework for text Information Retrieval consisting of two components: Text refining that transforms free form of text documents into a middle form; and knowledge cleansing that deduces patterns or knowledge from the middle form. Then we use the proposed framework to study and align the state of the art text Information Retrieval products and applications based on the text refining and knowledge cleansing functions as well as the middle form that they adopt. The rest of this paper is planned as follows. The first part presents the proposed text Information Retrieval framework, which bridges the gap between text Information Retrieval and data Information Retrieval (Li and Zhong, 2006). The second part gives an overview of the current text Information Retrieval products and applications in the light of the proposed framework. The final part discusses open problems and research directions. The normal form of accumulated information is text, text Information Retrieval is believed to include a viable potential higher than that of data Information Retrieval (Li and Zhong, 2006). In truth, a recent study indicated that 85% of a company's information is contained in text documents. Text Information Retrieval is also a much more complex task than data Information Retrieval as it involves dealing with text data that are essentially unstructured and unclear. Text Information Retrieval is a multidisciplinary field, involving information retrieval, information extraction, text analysis, clustering, visualization, categorization, machine learning, database technology and data Information Retrieval (Li *et. al.,* 2003).

REVIEW OF LITERATURE:

Data Information Retrieval operations, such as predictive system and associative discovery, drop into this category. A document based MF can be transformed into a concept based MF by realigning or extracting the related information according to the objects of interests in a specific domain. It follows that document based MF is usually domain independent and concept based IF is domain dependent. Text Information Retrieval can be visualized as consisting of two parts: Text refining that transforms free form text documents into a chosen middle form, and information distillation that deduces patterns or knowledge from the middle form (MF). Middle form can be semi structured such as the conceptual graph demonstration or structured such as the relational data representation. Middle form can be document-based wherein each entity represents a document, or concept based wherein each entity represents an object or concept of interests in a specific domain.

Information Retrieval a document based IF deduces patterns and relationship crosswise documents (Li and Zhong, 2004). Document clustering, visualization and categorization are examples of Information Retrieval from a document related MF. Information Retrieval a concept based MF derives system and relationship across objects or concepts.

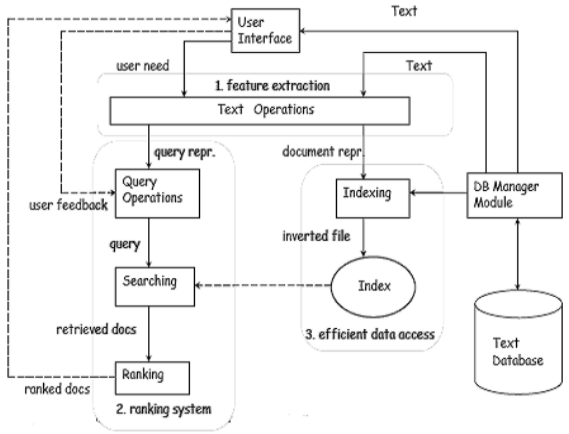

Fig-1- Structure of Text Information Retrieval

A text Information Retrieval framework: Text refining converts unstructured text documents into a middle form. MF can be document based or concept based. Knowledge distillation from a document based MF deduces patterns or knowledge across documents. A document based MF can be projected onto a concept based MF by extracting object information relevant to a domain. Knowledge distillation from a concept based MF deduces patterns or knowledge across objects or concepts. For example, given a set of news bulletin articles, text refining first converts each document into a document based MF. One can then perform knowledge distillation on the document based MF for the purpose of organizing the article, according to their content, for visualization and navigation purposes for knowledge discovery in an exact domain, document based MF of the news articles can be projected onto a concept based MF depending on the task obligation (Lau *et. al.,* 2004). For example, one can take out information related to company from the document based MF and form a company database. Information distillation can then be performed on the company database (company based MF) to derive company related information.

Text Information Retrieval Products: An additional group focuses on text analysis function, information retrieval, information extraction, categorization, and summarization. While we see that most text Information Retrieval systems are based on usual words processing none of the products has integrated data Information Retrieval functions for information distillation across theory or objects (Hammouda and Kamel, 2002). The text Information Retrieval products and applications based on the text refining and knowledge distillation functions as well as the middle form adopted. One group of products focuses on document group, image, and map reading.

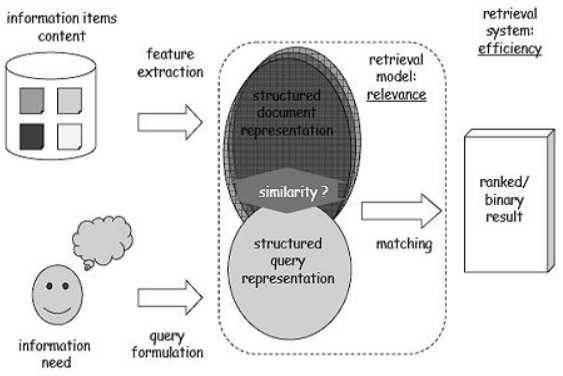

Fig- 2- Consequently the information retrieval system

Information Retrieval: Generating structured representations of information needs: often these tasks are solved by providing users with a query language and leave the formulation of structured queries to them. This is the case for example of simple keyword based query languages as used in Web search engine. Some information retrieval systems also maintain the user in the query formulation, e.g. through image interface. Similar of information needs with information substance: this is the algorithmic duty of computing parallel of information items and information need of constitutes the heart of the information retrieval system. Similarity of the structured representations is used to system relevance of information for users (Feldma & Dagan, 1995) This imposes fundamental constraints on the retrieval system. Retrieval systems that would capture relevance very well, but are computationally prohibitively expensive are not suitable for an information retrieval system.

Text Information Retrieval: At present most popular information retrieval methods are Web search engine. To a huge amount, they are text retrieval system, since they exploit only the textual content of Web documents for retrieval.

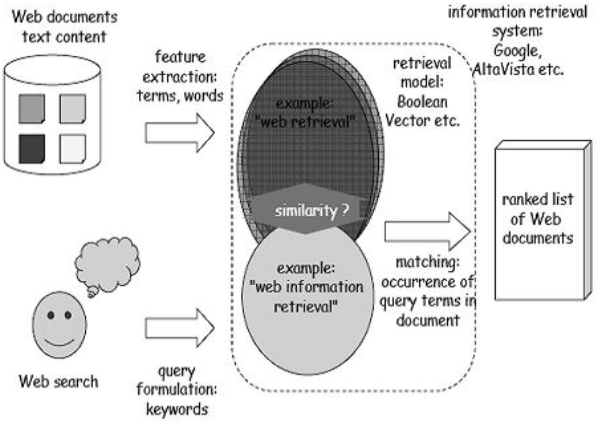

Fig-3- Text Information Retrieval

However, more recently Web search engines also create to exploit link information and even image

information. The three tasks of a Web search engine for retrieval are:

1. Computing the similarity of documents with the query and producing from that a grade result. Here Web search engines are used for standard text a retrieval method that is such as Boolean retrieval and vector space retrieval. We will introduce these methods in detail subsequently.

2. Extracting the textual features, which are the words or terms that occur in the document. We guess that the web search engine has already composed the documents from the Web using a Web crawler.

3. Support the formulation of textual queries. This is usually done by allowing the entry of keywords through Web forms.

Text analysis and understanding: The second group of the text Information Retrieval products is mainly based on natural language processing techniques, as well as text analysis, text categorization, information removal and summarization. Knowledge Discovery System's Concept Explorer is a visual search tool that helps to find precisely related content on the web. It "learns" relationships between words and phrases automatically from model documents and visually guides you to create searches. Inxight's LinguistX is another document retrieval tool with some text analysis and summarization capabilities (Fayyad *et. al.,* 1996). IBM's intellectual Miner is almost certainly one of the majority comprehensive text Information Retrieval goods approximately. It offers a set of text analysis tools, including a quality removal tool, clustering tools, summarization tool, and categorization tool. Also incorporated are the IBM's text search engines, Net Question Solution the IBM web crawler package. Text Wise, an R&D company based in Syracuse University, offers various text Information Retrieval products. DR LINK is an information retrieval method based on regular concept development. CINDOR is its cross lingual edition. CHESS is a text analysis and information mining tool. Also an information extraction tool is the Data Junction's Cambia, which extracts information in the structure of relational attributes from text. Mega computer's Text Analyst uses a semantic net representation of documents and performs computerized indexing, topic assignment, text abstraction, and semantic search.

Document visualization: There are a good number of text Information Retrieval products that fall into this category. The general approach is to organize the documents based on their similarities and present the groups or clusters of the documents in certain graphical representation. The subsequent register is by no means exhaustive but is enough to demonstrate the variety of the representation schemes available. Cartia's them escape is an enterprise information mapping application that presents clusters of documents in landscape demonstration. Canis'scMap is a file clustering and visualization tool based on Self Organizing Map. IBM's Technology looks at, developed together with Synthema in Italy, and is a text Information Retrieval request in the scientific field. It performs text clustering plus visualization in the shape of maps for patent databases and technical publications. Inxight's also offers a visualization device, known as VizControls, which performs value added post processing of search results by clustering the documents into groups and displaying based on a hyperbolic tree representation. Semio Corp's Semio Map employs a 3D graphical interface that maps the links between concepts in the document collection. Note that SemioMap is concept based in the sense that it explores the relationships between concepts whereas most other visualization tools are document based.

CONCLUSION:

Information retrieval was concerned over the last 20 years with the problem of retrieving information from large bodies of documents with mostly textual content, as they were typically found in library and document management systems. The problems addressed were classification and categorization of documents, systems and languages for recovery, client interfaces and image. The area was perceived as being one of narrow interest for highly specialized applications and users. The advent of the WWW altered this opinion totally, as the web is a worldwide warehouse of documents with universal access. Since at the present time the majority of the information content is at rest existing in textual appearance, text is an important basis for information recovery. Usual language text carries a set of meaning, which still cannot completely be captured computationally. Consequently information retrieval methods are based on powerfully simplified methods of text, ignoring the majority of the grammatical formation of text and reducing texts fundamentally to the terms they include. This approach is called full text retrieval and is an oversimplification that has verified to be very successful.

REFERENCES:

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From data Information Retrieval to knowledge discovery: An Overview. In Advances in Knowledge Discovery and Data Information Retrieval , U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R.

Uthurusamy, eds., MIT Press, Cambridge, Mass., pp. 1-36.

Feldman, R. & Dagan, I. (1995). Knowledge discovery in textual databases (KDT). In proceedings of the First International Conference on Knowledge Discovery and Data Information Retrieval (KDD-95), Montreal, Canada, August 20-21, AAAI Press, pp. 112-117.

K. Hammouda and M. Kamel (2002). Sphrase-based document similarity based on an index graph model. In ICDM02, pages 203–210.

R. Y. K. Lau, P. Bruza, and D. Song (2004). Belief revision for adaptive information retrieval. In SIGIR, pages 130–137.

Y. Li and N. Zhong (2004). Web mining model and its applications for information gathering. Knowledge-Based Systems, 17: pp. 207–217.

Y. Li and N. Zhong (2006). Mining rough association from text documents. In RSCTC, pages 368–377.

Y. Li and N. Zhong (2006). Rough association rule mining in text documents for acquiring web user information needs. In IEEE/WIC/ACM International Conference on Web Intelligence, WI06, pp. 226 – 232.

Y. Li, C. Zhang, and S. Zhang (2003). Cooperative strategy for web data mining and cleaning. Applied Artificial Intelligence, 17(5-6): pp. 443–460.

Corresponding Author

A. Sindhu*

Laxmi Institute of Commerce and Computer Application (BBA-BCA) Sarigam

E-Mail – sindhubhilai@yahoo.com