A Study of Text Mining For Web Information Retrieval System From Textual Databases

K. Sankar¹, Dr. G. N. K. Suresh babu²

¹Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore, India.
²Assistant Professor, Department of Computer Applications, Bharath University, Chennai, India.

Abstract—Text Information Retrieval is also known as text data mining or knowledge discovery from text 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 information, Textual database, User interface, searching, ranking, DB manager, Text Database, etc.

I. INTRODUCTION

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. 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.

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. 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.

II. STRUCTURE OF TEXT INFORMATION RETRIEVAL

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. 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.

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.
2.1. 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. 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.

III. A Survey Of Text Information Retrieval Products

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.

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.

3.1. Information Retrieval

Consequently the information retrieval system has to deal with the following tasks:

Generating structured representations of information items: this process is called feature extraction and can include simple tasks, such as extracting words from a text as well as complex system, e.g. for image or video analysis.

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.

As a result a selection of relevant information items or a ranked result can be presented to the user. Since information retrieval systems deal usually with large information collections and large user communities the efficiency of an information retrieval system is crucial.
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.

3.2. Text Information Retrieval

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 Syntehma 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 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 SemioMap 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.

3.4. 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. 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.

3.3. 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.

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. 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. 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.
2. Support the formulation of textual queries. This is usually done by allowing the entry of keywords through Web forms.
3. 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.
A synopsis of illustrative text Information Retrieval goods and applications based on the text refinement and knowledge distillation functions as well as the intermediate form adopted.

IV. OPEN PROBLEMS AND FUTURE DIRECTIONS

4.1. Middle form

Middle forms with varying degrees of complexity are suitable for different Information Retrieval purposes. For a fine-grain domain specific knowledge discovery task, it is necessary to perform semantic analysis to derive a sufficiently rich representation to capture the correlation between the objects or concepts described in the papers. However, semantic analysis methods are computationally luxurious and often function in the order of a small number of words for every second. It remains a challenge to observe how semantic analysis is able to be made much more efficient and scalable for extremely huge text corpora.

4.2. Multilingual text refining

Whereas data Information Retrieval is basically language self-governing, text Information Retrieval involves an important language module. It is necessary to expand text refining algorithms, which progress multilingual text documents and produce language-independent middle forms. While most text Information Retrieval tools focus on processing English documents, Information Retrieval from documents in other languages allows access to previously untapped information and offers a new host of opportunities.

4.3. Domain knowledge integration

Domain knowledge, not catered for by any current text Information Retrieval tools, could play an important role in text Information Retrieval. Specifically, domain knowledge can be used as early as in the text cleansing phase. It is attractive to discover how one can obtain advantage of domain information to develop parsing efficiency and receive a more compact middle form. Domain knowledge could also play a part in knowledge distillation. In a classification or predictive system task, domain knowledge helps to improve learning/Information retrieval efficiency as well as the quality of the learned system (or mined knowledge) [5]. It is also interesting to explore how a user’s knowledge can be used to initialize a knowledge structure and make the discovered knowledge more interpretable.

4.5. Personalized autonomous Information Retrieval

Current text Information Retrieval products and applications are still tools intended for trained knowledge specialists. Future text Information Retrieval tools, as part of the knowledge management systems, should be readily usable by technical users as well as management executives. There have been some efforts in developing systems that interpret natural language queries and automatically perform the appropriate Information Retrieval operations. Text Information Retrieval tools could also appear in the form of intelligent personal assistants [6]. Under the agent paradigm, a personal miner would learn a user’s profile, conduct text information Retrieval process repeatedly, and forward information not including requiring an explicit request from the client.

V. 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 textural 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


About the Author

K. Sankar
K. Sankar, graduated the Faculty of Science, Bharath University at Chennai in India. Working as an Assistant Professor in the Department of Computer Applications from the same University. His research and interest fields are Computer graphics, Data Mining and Computer Network.

Dr. G. N. K. Suresh Babu
Dr. G. N. K. Suresh Babu Working as Associate Professor in GKM College of Engineering and Technology, Chennai. He is having 22+ Years of Experience in Industry / Teaching. His research areas are Data Mining, Fuzzy Logics, Neural Networks, Computer Networks and Software Engineering.