SIMILARITY-BASED MEASURE FROM TEXT- DOCUMENT

Nway Htet Htet Aung, Soe Hay Mar
Computer University (Monywa)
nhhaung24@gmail.com

Abstract

In this paper, there is describing a similarity-based retrieval framework that addresses the challenges associated with the relational database text documents. This system proposed to automatically classify documents based on the meanings of words and the relationships between groups of meanings or concepts. There may be found similar documents based on a set of common keywords and retrieved these documents based on the degree of relevance which is measured on the relative frequency of the keywords. So, this system uses measure similarity between new and old thesis title description to detect duplicate system in case study.

1. Introduction

The information that interests users come from a variety of sources, including text documents, photographic images, sensor data, Web pages, and biological sources. Accessing this data requires that information meaningful to human be extracted from weakly structured or totally unstructured sources, in addition to conventional structured sources. Data mining and knowledge discovery work focuses on finding unexpected but interesting patterns within any of the varied types of information. Patterns might be found in relationships between individual pieces of information, in recurring sensor events over time, or in collections of strongly related text documents. Finally, to ensure information is valuable to users, there may be investigated techniques to assess the quality, reliability, and authenticity of information. To ensure information is handled safely, there are investigating techniques for protecting against unexpected disclosures that can threaten privacy.Text retrieval has been traditionally called information retrieval. Information retrieval includes retrieval of all kinds of non-textual information (e.g., images, video) as well, even though in old days, we mostly had textual information. Thus we use text retrieval to refer to finding relevant information in a text collection. The retrieval of textual information such as text retrieval is especially important because the most frequently wanted is often textual and techniques for retrieving textual information can be useful for retrieving other media information. A document collection, the task of text retrieval can be defined as using a user query to identify a suhset of documents that can satisfy the user's information need. As a result, most existing research in information retrieval has assumed that the goal is to develop a good ranking function.

2. Related work

Paul Thompson et.al discussed a different application: improving information retrieval through name recognition. It investigates name recognition accuracy, and the effect on retrieval performance of indexing and searching personal names differently from non-name terms in the context of ranked retrieval. The main conclusions are: the name recognition in text can be effective; the names occur frequently enough in a variety of domains, including those of legal documents and news databases, to make recognition worthwhile; and that retrieval performance can be improved using name searching [2].

Andreas H et.al presented text mining as a young and interdisciplinary field in the intersection of the related areas information retrieval, machine learning, statistics, computational linguistics and especially data mining. They describe the main analysis tasks preprocessing, classification, clustering, information extraction and visualization. In addition, they briefly discuss a number of successful applications of text mining [7].

Tynor T et.al discussed one of the main subjects of data mining is data analysis. With the increasing amount of data, or more precisely, text available on the Internet, choosing what will be transmitted to the user is a great challenge. This article uses different similarity measure methods available in order to analyze sets of summarized texts. These were generated through the usage of different summarization algorithms [8].

3. Background Theory

3.1 Information retrieval

Data Mining is the task of discovering interesting patterns from large amount of data where the data can be stored in database, data warehouses Data mining refers to extracting or mining knowledge from large amount of data. Data mining should have been more appropriately named knowledge mingling from data which is unfortunately somewhat long. Data mining is the process of extracting hidden patterns from data. As more data is gathered, with the amount of data doubling every three years, data mining is becoming an increasingly important tool to transform this data into information. It is commonly used in a wide range of applications, such as marketing, fraud detection and scientific discovery. Data mining can be applied to data sets of any size. However, while it can be used to uncover hidden patterns in data that has been collected, obviously it can neither uncover patterns which are not already present in the data, nor can it uncover patterns in data that has not been collected. Information retrieval (IR) is the science of searching for documents, for information within documents and for metadata about documents, text classification is an important task of data mining. Information retrieval concerned with the retrieval of information from a large number of text-based documents [3].

3.2 Text mining

A reader will make connections between seemingly unrelated facts to generate new ideas or hypotheses. Text mining offers a solution to this problem by replacing or supplementing the human reader with automatic systems undeterred by the text explosion. It involves analysing a large collection of documents to discover previously unknown information. The information might be relationships or patterns that are buried in the document collection and which would otherwise be extremely difficult, if not impossible, to discover.

Text mining can be used to analyse natural language documents about any subject, although much of the interest at present is coming from the biological sciences. Text mining can not only extract information on interactions from documents, but it can also go one step further to discover patterns in the extracted interactions. Information may be discovered that would have been extremely difficult to find, even if it had been possible to read all the documents [5].

3.3 Keyword-based and similarity- based retrieval

In keyword-based information retrieval, a document is represented by a string, which can be identified by a set of keywords. A user provides a keyword or an expression formed out of a set keywords, such as “car and repair shops”, “tea or coffee”, or “database systems but not Oracle”. A good information retrieval system should consider synonyms answering such queries.. Keyword-based retrieval is a simple model that can encounter two major difficulties. The first is the synonyms problem: a keyword, such as software product, may not appear anywhere in the document, even though the document is closely related to software product. The second is the polysemy problem: the same keyword, such as mining, many mean different things in different contexts.

Similarity-based retrieval finds similar documents based on asset of common keywords. The output of such retrieval should be based on the degree of relevance, where relevance is measured based on the closeness of the keywords, the relative frequency of the keywords, and so on. It is difficult to provide a precise measure of the degree of relevance between a set of keywords, such as the distance between data mining and data analysis. [4]

A text retrieval system often associates a stop list with a set of documents. A stop list is a set of words that are deemed “irrelevant”. For example, a, the, of, for, with, and so on are stop words even though they may appear frequently. Stop lists may vary when the documents set vary. For example, database systems could be an important keyword in a newspaper. However, it may be considered as a stop word in a set of research papers presented in a database systems conference.

3.4 Similarity measure

Texts may be an extract or an abstract of the original text. An extract is obtained by the selection of the sentences which have the main idea of the text. In the automatic approach this technique is more complicated as it involves artificial intelligence. In this article, three algorithms were used to summarize a set of texts. All of them give an extract of the original text. Three algorithms were used to measure the similarity between the automatic extracts and the manual abstracts.

The extraction algorithms are mainly based on the fact that texts have always a central idea behind it. That idea can be identified through one or more sentences of the original text. To identify such central sentences statistics methods can be used to value each one of the sentences in the text. After that

the sentences with the highest values are selected to be in the final text. [6]

Three simple and well - known similarity measures to calculate the similarity between sentences: the Dice, Jaccard, Cosine Coefficients. These measures all take the words in two sentences and calculate the similarity on how many words they have in common. [1]

Jaccard similarity coefficient measure is used to handle asymmetric binary attributes as only non-zero values are relevant for the calculation. Cosine similarity measure is used to represent objects with different frequencies of its attributes. The system uses Die similarity measure. When the system compares two sentences using Die similarity measure receives more exact similarity result.

3.4.1 Dice

In similarity context, Dice is a measure based on the number of matched attributes between two objetcs divided by the number attributes of one of them. When the number of possible attributes is too large, words for example, only a set of relevant attributes may be considered. Dice is defined by the following formula:

$$S_{A,B}^{Dice} = \frac{2\,|\{words_A\} \cap \{words_B\}|}{|\{words_A\} + \{words_B\}|}$$

3.5.4 Example for Die calculation

$$S_{A,B}^{Dice} = \frac{2\,|\{words_A\} \cap \{words_B\}|}{|\{words_A\} + \{words_B\}|}$$

$$S_{A,B}^{Dice} = \frac{2\,|\{TextMining\} \cap \{TextMining\}|}{|\{TextMining\} + \{TextMining\}|}$$

$$S_{A,B}^{Dice} = \frac{2\,|\{10\}|}{|\{10\} + \{10\}|} = 1$$

4. System Design

The detail design of system must enter the thesis title that user wants to request to text box. When system search input title, it will parse and remove disused stop words. The system calculates term frequency weighting (Counting data from the input title) depending on document number in database. And then the system is find relevant title from database using Die similarity measure. Finally, the system display title and abstract approximately similarity with input title as a result.

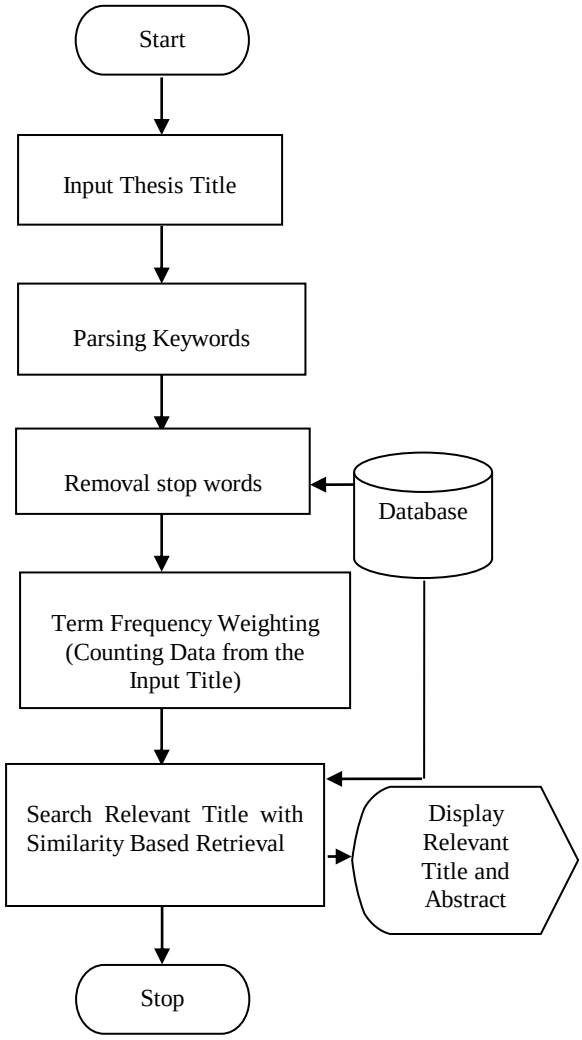

Figure 1. System flow diagram of the system

5. Implementation

In this system, user must enter the requirement title or sentence in to the text box and then press " Search". The system will parse the input title or sentence and removal stop words and similar words from database and then term frequency weighting (counting keywords form the input title) and then search relevant title and abstract with the similarity based retrieval.

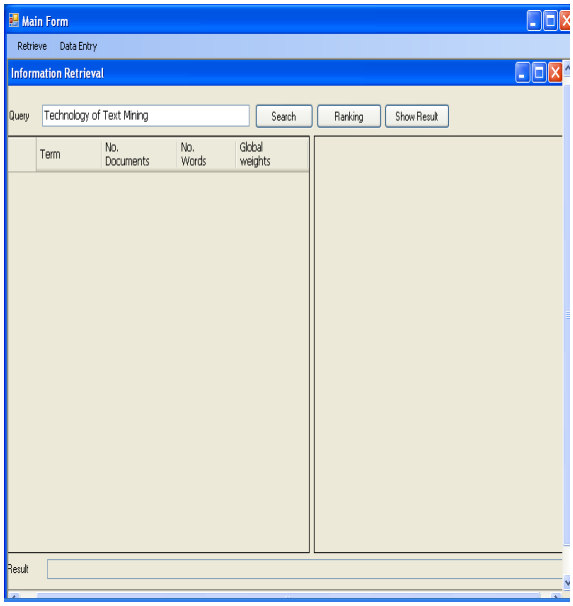

Figure 2. User input title form

During the process, the system showing the elapsed time on the tool bar and also the progress bar beside the elapsed time will show the status of the process, it will be running or not. It is important because the process is sometimes overloaded the processing range of the host PC and it may hang the host operating system. After processing, the system will be displayed similarity measure values of each document and the associate document ranking in to the result box.

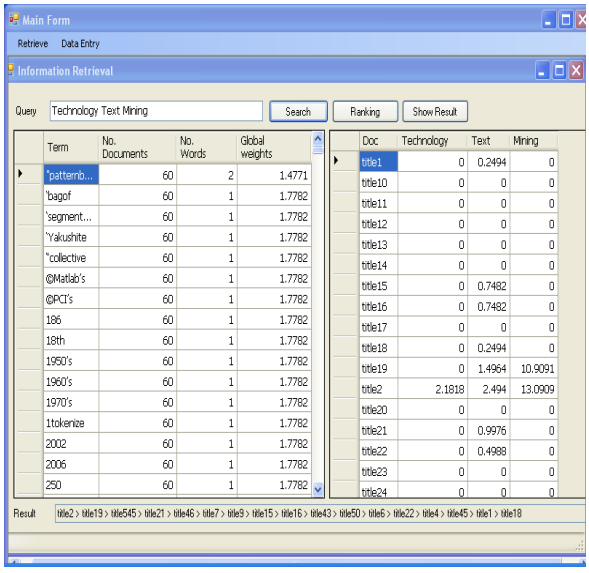

Figure 3. Display ranking document form

6. Conclusion

In this system, we have presented an approach that uses an automatically learned information retrieval system to extract structured databases from a text document, and then mines this database with existing knowledge discovery in databases tools. The experimental results demonstrate that information extraction and data mining can be integrated for the mutual benefit of both tasks. Information retrieval enables the application of knowledge discovery in databases to unstructured text and knowledge discovery in databases can discover predictive rules useful for improving information retrieval performance.

This system has presented initial results on integrating information retrieval and knowledge discovery in databases that demonstrate both of these advantages. Text mining is a relatively new research area at the intersection of natural-language processing, machine learning, data mining, and information retrieval. By appropriately integrating techniques from each of these disciplines, useful new methods for discovering knowledge from large text documents can be developed.

In this system, data mining have focused on structure data. However, in reality, a portion of the available information is stored in text databases, which consist of large collection of documents. Data stored in most text database are semi-structured data. Information retrieval is concerned with the retrieval of information from a large number of text base documents. This system is implemented the system to locate relevant documents on user input such as title of thesis.

7. References

[1] E. Brill, “Extracting semantically related queries by exploiting user session information”, Technical Report, Microsoft Research, 2005.

[2] B. Christine “Name Searching and Information Retrieval” Getty's Synon~e and its cousins: A survey of applications of personal name-matching algorithms Journal of the American Society of lnjbrrnation Science, 43: 459-476 1992.

[3] T.Feldman and K. Dagan, “Improving Similarity Measures for Short Segments of Text”, Knowledge discovery in textual databases (KDT). In proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), Montreal, Canada, August 20-21, AAAI Press, 112-117,(1995).

[4] J.Han and M. Kamber, “Data Mining: Concepts and Techniques”, Morgan Kaufmann, San Francisco, 2000.

[5] A. T. Kahn, "Text Mining at the Term Level", Opportunity Explorer: Navigating Large Databases Using Knowledge Discovery Templates. In: Proceedings of the 1993 workshop on Knowledge Discovery in Databases, (1993).

[6] T. Mandl, "Learning Similarity Functions in Information Retrieval", Social Science Information Centre, Lennéstraße 30 - 53113 Bonn – Germany

[7] P. N. Tan, M. Steinbach and V.Kumar, "A Brief Survey of Text Mining", Modern Information Retrieval Addison Wesley Longman, 1999.

[8] T. Tan, "Text Mining: The state of the art and the challenges" Kent Ridge Digital Labs, 21 Heng Mui Keng Terrace, Singapore 119613.