Text Mining For Information Retrieval by Cosine Similarity Measure

Hay Man Oo, Khin Lay Thwin
Computer University (Kalay)
haymanoo99@gmail.com, khinlaythwin@gmail.com

Abstract

This paper can describe a similarity-based retrieval framework that addresses the challenges associated with the relational database text documents. This system proposes to automatically classify documents based on the meanings of words and the relationships between groups of meanings or concepts. There may be find 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 will measure similarity between new and old thesis document description to detect duplicate system that is used in case study.

1. Introduction

The information that makes users interests comes from a variety of sources, including text documents, photographic images, sensor data, Web pages, and biological sources. Accessing this data that require information the meaningful to humans is extracted from weakly structured or totally unstructured sources, in addition to conventional structured sources. Data mining and knowledge discovery work focuses on finding unexpected disclosures 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, certain 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.

The rest of this paper is organized as follows: Section 2, represent related words. In section 3 explain theory of the system. The next section 4 also presents the system design. And then, implementation of the system is section 5. In the section 6, experimental results of the system. Finally, the conclusion of this paper is presented in section 7.

2. Related Works

Paul Thompson., etal 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: that name recognition in text can be effective; that 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. [3]

Andreas H., etal 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. [8]

Ronen Feldman1., etal presented an approach to performing text mining at the term level. The mining process starts by preprocessing the document collection and extracting terms from the documents. Each document is then represented by a set of terms and annotations characterizing the document. Terms and additional higher-level entities are then organized in a hierarchical taxonomy. In this paper they will describe the Term Extraction module of the Document Explorer system, and provide experimental evaluation performed on a set of 52,000 documents published by Reuters in the years 1995-1996. [2]

Text mining, also known as text data mining or knowledge discovery from textual databases, refers to the process of extracting interesting and non-trivial patterns or knowledge from text documents. Regarded by many as the next wave of knowledge

discovery, text mining has very high commercial values. Last count reveals that there are more than ten high-tech companies offering products for text mining. Has text mining evolved so rapidly to become a mature field? This article attempts to shed some lights to the question. *Ah-Hwee Tan* first present a text mining framework consisting of two components: *Text refining* that transforms unstructured text documents into an *intermediate form;* and *knowledge distillation* that deduces patterns or knowledge from the *intermediate form*. They then survey the state-of-the-art text mining products/applications and align them based on the text refining and knowledge distillation functions as well as the intermediate form that they adopt. In conclusion, they highlight the upcoming challenges of text mining and the opportunities it offers. [5]

Fernando Mendonc Ana Ozaki 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. [7]

3. Background Theory

3.1. Information Retrieval

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. 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 concerned with the retrieval of information from a large number of text-based documents. A typical information retrieval problem is to locate relevant documents based on user input. Calculation of similarity between corresponding documents becomes a major task in information retrieval from a textual database. [5]

Information retrieval in the broader sense deals with the entire range of information processing, from data retrieval to knowledge retrieval. Although information retrieval is a relatively old research area where first attempts for automatic indexing, it gained increased attention with the rise of the World Wide Web and the need for sophisticated search engines. Even though the definition of information retrieval is based on the idea of questions and answers, systems that retrieve documents based on keywords, i.e. systems that perform *document retrieval* like most search engines are frequently also called information retrieval systems.

3.2. Knowledge Discovery in Database (KDD)

Knowledge discovery in databases is a process that is defined by several processing steps that have to be applied for a data set of interest in order to extract useful patterns. These steps have to be performed iteratively and several steps usually require interactive feedback from a user.

The analysis of data in KDD aims at finding hidden patterns and connections in these data. A quantity of facts, which can be, for instance, data in a database, but also data in a simple text file can be understood by data.

3.3. Text Mining

Text mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. Text mining applies the same analytical functions of data mining to the domain of textual information. Text Mining can be performed by a collection of methods from various technological areas

A reader will make connections between seemingly unrelated facts to generate new ideas or hypotheses. However, the burgeoning growth of published text means that even the most avid reader cannot hope to keep up with all the reading in a field, let alone adjacent fields. Nuggets of insight or new knowledge are at risk of languishing undiscovered in the literature. 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

analyzing 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 analyze 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. [1]

3.4. Text Mining Works

Text mining involves the application of techniques from areas such as information retrieval, natural language processing, information extraction and data mining. These various stages of a text-mining process can be combined together into a single workflow. [1]

In this system, 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 find 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 will measure similarity between new and old thesis title description to detect duplicate system is used in case study.

3.5. 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. For example, given the keyword car, synonyms such as automobile and vehicle should be considered in the search as well. 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. In many cases, 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. [6]

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.

A group of different words may share the same word stem. A text retrieval system needs to identify groups of words where the words in a group are small syntactic variants if one another, and collect only the common word stem per group. For example, the group of words drug, drugged, and drugs, share a common word stem, drug, and can be viewed as different occurrences of the same word. [6]

3.6. Similarity Measures

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. [9]

3.7. Measure of Similarity and Dissimilarity

Three simple and well - known similarity measures to calculate the similarity between sentences: the Dice, Jaccard and Cosine Coefficients. These measures all take the words in two sentences and calculate the similarity on how many words they have in common. [4]

3.7.1. Jaccard Similarity Coefficient. Some attributes are present in just a few objects of a data set. As they assume zero values in most of the cases, they are called asymmetric. Jaccard Similarity Coefficient measure is used to handle asymmetric binary attributes as only non-zero values are relevant for the calculation.

Its formula is defined by:

$$S_{A,B}^{Jaccard} = \frac{|\{words_A\} \bigcap \{words_B\}|}{|\{words_A\} + \{words_B\}| - |\{words_A\} \bigcap \{words_B\}|} \quad (3.1)$$

3.7.2. Cosine Similarity. Cosine Similarity measure is used to represent objects with different frequencies of its attributes. Documents are an example of objects that may have different frequencies of its attributes, words. Like jaccard coefficient, cosine measure only considers attributes that are present at least in one of the two objects being analyzed.

In the documents example all the words would be the set of attributes, but for each document most of then would be zero valued. If the 0-0 matches were considered then documents in general would be highly similar. The cosine similarity is defined by:

$$S_{A,B}^{Cosine} = \frac{|\{words_A\} \cap \{words_B\}|}{|\sqrt{\{words_A\}} \| \{words_B\}|} \quad (3.2)$$

3.7.3. Dice. In similarity context, Dice is a measure based on the number of matched attributes between two objects 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:

$$S_{A,B}^{Dice} = \frac{2|\{words_A\} \cap \{words_B\}|}{|\{words_A\} + \{words_B\}|} \quad (3.3)$$

For example, Dice

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

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

$$S_{A,B}^{Dice} = \frac{2|\{20\} \cap \{20\}|}{|\{20\} + \{20\}|} = \frac{40}{40} = 1$$

Among these three similarity measures, this system is used cosine similarity to measure frequency between sentences.

4. System design

In this system design, when user can enter thesis documents, the system will parse keywords and remove stop words in the database. When user search the required title or sentence, the system will parse the user's query and remove stop words and similar words. And then, term frequency will weighted to count keyword from the documents from the user's input and search relevance documents with similarity based retrieval. Finally, the system will display the relevance documents.

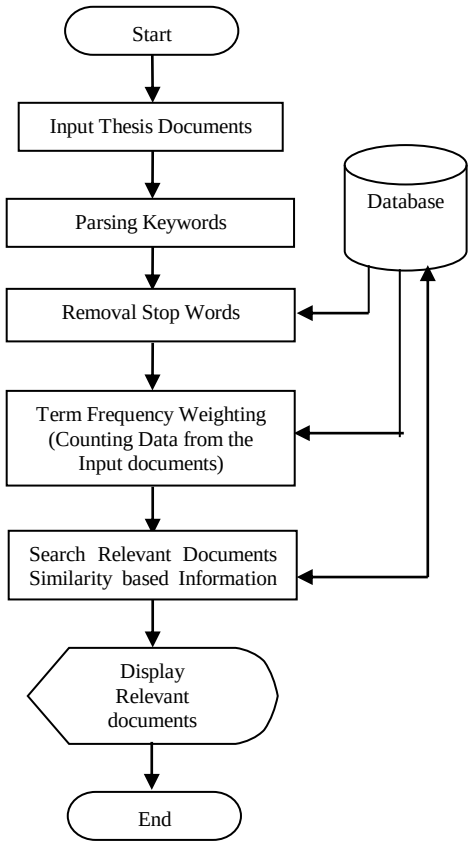

Figure 1. System Flow Diagram of the System

5. Implementation

In this system, user must enter the required title or sentence in to the text box and then press "Search". By using Cosine similarity, the system parses the user's query and then it removes stop words, such as a, an, the, etc, but it uses the other words, for example image, process, similar, etc, as

keywords. After that, it substitutes common or stem words with keywords. If the keyword is “process”, the common or stem words may be ‘processes, processed, processing, etc. These stop words, keywords, and common words can be chosen by user. User query calculates how many same keywords are included in documents based on the keywords. Finally, the system display the documents in which user’s queries are most included and show the accuracy results on retrieving documents to user’s input.

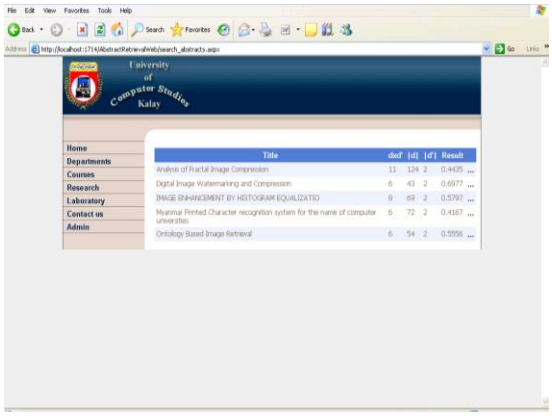

Figure 2. Information Retrieval Form

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 word by each document and the associate document ranking.

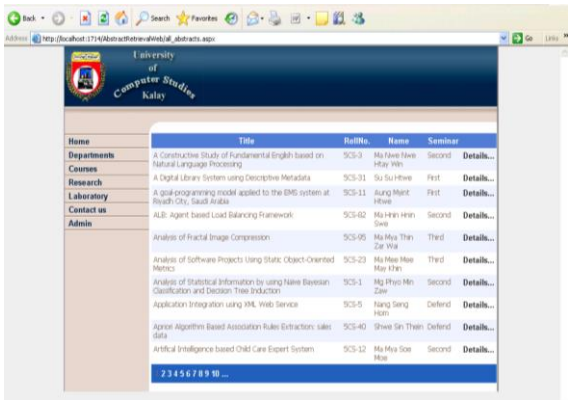

Figure 3. Display Ranking Document Form

6. Experimental results

In the experimental results, the system will calculate the frequency to get the accuracy measure, that is to display how many the user’s query are included in the documents as shown in table 1.

Table 1 The experimental results of accuracy measure

No	User's query	d'	d	d*d'	|d|	|d'|	Accuracy Results
1	image	1	6	6	43	1	0.1395
2	image processing	1 1	6 0	6	43	2	0.0698
3	digital image processing	1 1 1	2 6 0	8	43	3	0.0620

Accuracy Result= $\frac{d * d'}{|d| * |d'|}$

where, d' = frequency of user's query
d = time of the frequency of word in documents
|d| = documents that is clear stop words
|d'| = user's query that is clear stop words

In figure 4, it shows the experimental results for the most relevant to the user’s query in accordance with user’s input. When the user searches the word such as "image", four documents concerned with "image" will be displayed. Among them, document 3 or D3 and document 4 or D4 are the most relevant to user’s input. When the user searches the title "image processing", the most relevant document with "image processing" or user’s input is document 4 or D4. If the user searches the title such as "digital image processing", document 2 or D2 is the most relevant to "digital image processing" or user’s input.

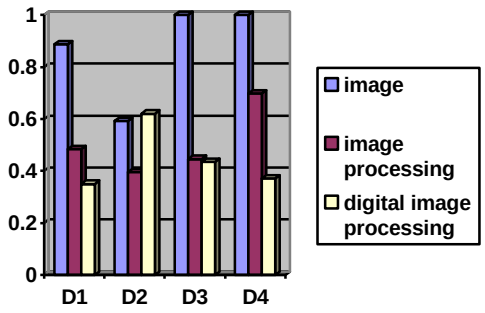

Figure 4. Comparison of the accuracy results on retrieving documents to user's input

7. Conclusion

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

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 to locate relevant documents on user input such as title of thesis.

8. References

[1]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).

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

[3]B.Christine and S. L. Siegfried. "*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 lnformation Science, 43: 459-476 1992.

[4]Cucerzan and E.Brill, "*Extracting semantically related queries by exploiting user session information*", Technical Report, Microsoft Research, 2005.

[5]Feldman and 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)

[6]J.Han and M. Kamber, "*Data Mining: Concepts and Techniques*", Morgan Kaufmann, San, Francisco, 2000.

[7]P. N. Tan ,M. Steinbach and V.Kumar," *Comparison of text sets using Data Mining and Similarity Measure Methods*" Introduction to Data Mining , Michael Steinbach and Vipin Kumar.

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

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