Journal of Computer Science and Applications.
ISSN 2231-1270 Volume 5, Number 1 (2013), pp. 21-30

http://www.irphouse.com

Exploiting Text Mining for Identifying Novel Information

Rajeev Tripathi

Research Scholar, CMJ University, Shillong Meghalaya, India.

Abstract

Data Mining is typically concerned with the detection of patterns in numeric data, but very often important (e.g., critical to business) information is stored in the form of text. Unlike numeric data, text is often amorphous, and difficult to deal with. Text Mining generally consists of the analysis of (multiple) text documents by extracting key phrases, concepts, etc. and the preparation of the text processed in that manner for further analyses with numeric Data Mining techniques (e.g., to determine co-occurrences of concepts, key phrases, names, addresses, product names, etc.). In this paper we have presented an overview of text mining and surveyed some of the techniques used to discover knowledge from text databases.

Keywords: Text Mining, Information Retrieval, Association Rule Mining, Metadata Mining.

1. Text Mining Techniques

Text Mining is an interdisciplinary field that utilizes techniques from the general field of Data Mining and additionally, combines methodologies from various other areas such as Information Extraction, Information Retrieval, Computational Linguistics, Categorization, Clustering, Summarization, Topic Tracking and Concept Linkage [1], [2], [3]. In the following sections, we will discuss each of these technologies and the role that they play in Text Mining.

1.1 Information Extraction

Information extraction (IE) is a process of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents, processing human language texts by means of NLP. The final output of the extraction process is some type of database obtained by looking for predefined sequences in text, a process called pattern matching [4].

Tasks performed by IE systems include:

- Term analysis, which identifies the terms appearing in a document. This is especially useful for documents that contain many complex multi-word terms, such as scientific research papers.
- Named-entity recognition, which identifies the names appearing in a document, such as names of people or organizations. Some systems are also able to recognize dates and expressions of time, quantities and associated units, percentages, and so on.
- Fact extraction, which identifies and extracts complex facts from documents. Such facts could be relationships between entities or events.

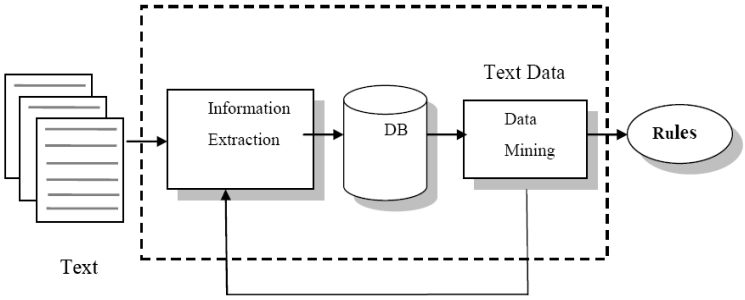

Figure 1: Overview of IE-based Text Mining framework.

IE transforms a corpus of textual documents into a more structured database, the database constructed by an IE module then can be provided to the KDD module for further mining of knowledge as illustrated in figure 1.

1.2 Information Retrieval

Retrieval of text-based information also termed Information Retrieval (IR) has become a topic of great interest with the advent of text search engines on the Internet. Text is considered to be composed of two fundamental units, namely the document (book, journal paper, chapters, sections, paragraphs, Web pages, computer source code, and so forth) and the term (word, word-pair, and phrase within a document). Traditionally in IR, text queries and documents both are represented in a unified manner, as sets of terms, to compute the distances between queries and documents thus providing a framework within to directly implement simple text retrieval algorithms.

1.3 Computational Linguistics/ Natural Language Processing

Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications. The goal of Natural Language Processing (NLP) is to design and build a computer system that will analyze, understand, and generate natural human- languages. Applications of NLP include machine translation of one human-language text to another; generation of human-language text such as fiction, manuals, and general descriptions; interfacing to other

systems such as databases and robotic systems thus enabling the use of human-language type commands and queries; and understanding human-language text to provide a summary or to draw conclusions.

NLP system provides the following tasks:

- Parse a sentence to determine its syntax.
- Determine the semantic meaning of a sentence.
- Analyze the text context to determine its true meaning for comparing it with other text.

1.4 Categorization

Categorization is the process of recognizing, differentiating and understanding the ideas and objects to group them into categories, for specific purpose. Ideally, a category illuminates a relationship between the subjects and objects of knowledge. Categorization is fundamental in language, prediction, inference, decision making and in all kinds of environmental interaction.

1.5 Topic Tracking

A topic tracking [64] system works by keeping user profiles and, based on the documents the user views, predicts other documents of interest to the user. Yahoo offers a free topic tracking tool (www.alerts.yahoo.com) that allows users to choose keywords and notifies them when news relating to those topics becomes available.

1.6 Clustering

Clustering [3] is a technique in which objects of logically similar properties are physically placed together in one class of objects and a single access to the disk makes the entire class available. There are many clustering methods available, and each of them may give a different grouping of a dataset. The choice of a particular method will depend on the type of output desired, the known performance of method with particular types of data, the hardware and software facilities available and the size of the dataset. In general, clustering methods may be divided into two categories based on the cluster structure which they produce. The non-hierarchical methods divide a dataset of N objects into M clusters, with or without overlap.

1.7 Concept Linkage

Concept linkage [5] identifies related documents based on commonly shared concepts and between them. The primary goal of concept linkage is to provide browsing for information rather than searching for it as in IR. For example, a Text Mining software solution may easily identify a link between topics X and Y, and Y and Z. Concept linkage is a valuable concept in Text Mining which could also detect a potential link between X and Z, something that a human researcher has not come across because of the large volume of information s/he would have to sort through to make the connection. Concept linkage is beneficial to identify links between diseases and treatments. In the near future, Text Mining tools with concept linkage capabilities will be beneficial in the biomedical field helping researchers to discover new treatments by associating treatments that have been used in related fields.

1.8 Information Visualization

Visual Text Mining [5], or information visualization, puts large textual sources in a visual hierarchy or map and provides browsing capabilities, in addition to simple searching e.g., Informatik V‘s DocMiner. The user can interact with the document map by zooming, scaling, and creating sub-maps. The government can use information visualization to identify terrorist networks or to find information about crimes that may have been previously thought unconnected. It could provide them with a map of possible relationships between suspicious activities so that they can investigate connections that they would not have come up with on their own. Text Mining with Information visualization has been shown to be useful in academic areas, where it can allow an author to easily identify and explore papers in which s/he is referenced. It is useful to user allowing them to narrow down a broad range of documents and explore related topics.

1.9 Summarization

A summary is a text that is produced from one or more texts, that contain a significant portion of the information in the original text(s), and that is no longer than half of the original text(s). Text‘ here includes multimedia documents, on-line documents, hypertexts, etc. Many types of summary that have been identified include indicative summaries (that provide an idea of what the text is about without giving any content) and informative ones (that do provide some shortened version of the content). Extracts are summaries created by reusing portions (words, sentences, etc.) of the input text verbatim, while abstracts are created by re-generating the extracted content. Generic summary is not related to specific topic while query-based summary generates a summary discussing the topic mentioned in the given query. Also summary can be created for single document or multi-documents.

2. Architecture of a Text Mining system

Text Mining system takes as an input a collection of documents and then preprocesses each document by checking its format and character sets [6]. Next, these preprocessed documents go through a text analysis phase, sometimes repeating the techniques, until the required information is extracted. Three text analysis techniques are shown in Figure 2, but many other combinations of techniques could be used depending on the goals of the organization. The resulting extracted information can be input to a management information system, yielding an abundant amount of knowledge for the user of that system. Figure 3 explores the detailed processing steps followed in Text Mining System.

Different steps of Text Mining process as shown above, are briefly discussed below:

a) Document files of different formats like PDF files, txt files or flat files are collected from different sources such as online chat, SMS, emails, message boards, newsgroups, blogs, wikis and web pages. This unstructured dataset of documents is pre-processed to perform following three tasks:

- Tokenize the file into individual tokens using space as the delimiter.
- Remove the stop words which do not convey any meaning.
- Use porter stemmer algorithm to stem the words with common root word.

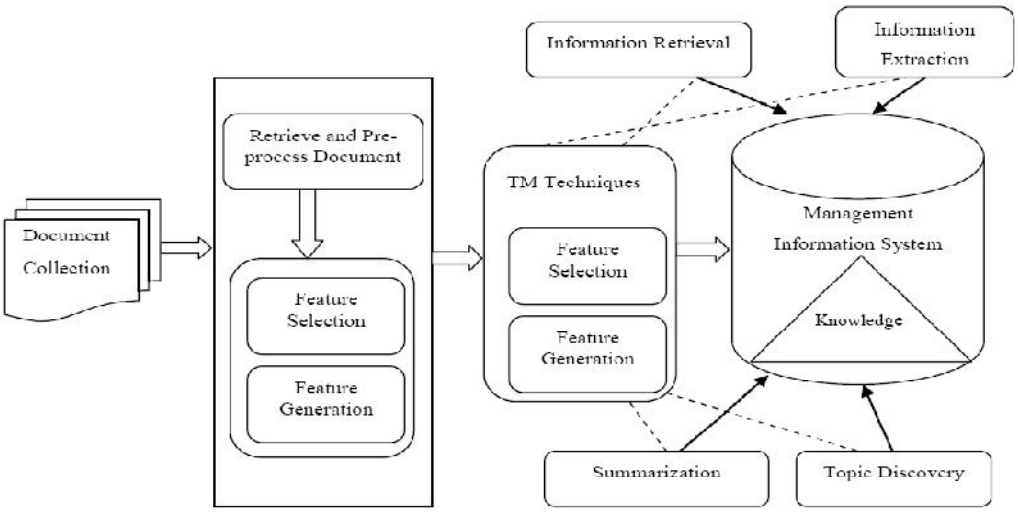

Figure 2: Text Mining Process.

b) Feature Generation and Feature Selection activities are performed on these retrieved and preprocessed documents to represent the unstructured text documents in a more structured spread sheet format. Feature Selection algorithms help to identify the important features which requires an exhaustive search of all subsets of features of chosen cardinality. If the large numbers are available this is impractical for supervised learning algorithms the search is for satisfactory set of features instead of optimal set.

c) After the appropriate selection of features the Text Mining techniques are incorporated for the applications like Information Retrieval, Information Extraction, Summarization and Topic Discovery for necessary knowledge discovery process**.**

3. Literatute Review

L.J. Haravu and A. Neelameghan[6] offers information about the advances in text and data-mining in knowledge-based software. After research and experience using these technologies, both Haravu and Neelameghan suggest two essential methods of creating these platforms.

Valitutti, Alessandro, and Carlo Strapparava[7] addresses lexicons and their connection to text mining. In linguistics, the lexicon of a language is its expressions, words, and vocabularies. In other words, lexicons, similar to text mining, are a language's inventory of lexemes, or combination pattern. Efficient computing is

consistently advancing as a field, and allows new forms of human-computer interactions, in addition to the use of a standardized natural language.

In 2003, Eiron and McCurley [8] studied the relation of an anchor text to a document and showed that the documents retrieved by anchor text techniques improve the quality of web text search than documents retrieved by content indexing. Their study revealed that anchor text is less ambiguous than other types of texts like title of document which are typically longer than individual anchor text and thus they resembles real-world queries in terms of its term distribution and length.

4. Approaches for identifying novel information

4.1 Association Rule Mining

In [9] Agrawal et al. introduce the notion of mining transaction data for association rules between sets of items in large databases, with a specified confidence level. An association rule is defined as an implication of the form $X \Rightarrow Y_i$ where X is a set of items and Y_i is an item not present in the set X. The items in the set X are termed the antecedents and Y_i is called the consequent. The support s of an association rule is defined as the percentage of the total number of transactions that contain both X and Y_i. The confidence c of an association rule is defined as the percentage of transactions of X that also contain Y_i. Association rules allow us to view implicit relationships between different entities and the confidence factor associated with each rule allows us to rank them. Thus one can execute queries such as "Give me the top 10 association rules for item A", where A is the antecedent.

Association rules are mined in two steps. In the first step those items sets that have support above a minimum support level specified are identified. Such an itemset is usually called a frequent itemset. This step can be computationally very intensive. In the second step, association rules are formed by finding all non-empty subsets a_i for each frequent itemset f and generating rules of the form $a_i \Rightarrow (f - a_i)$ if the ratio of the support of $(f - a_i)$ and the support of a_i is greater than a threshold. The Apriori algorithm[11] and the Direct Hashing and Pruning algorithm[12] are two well known algorithms used to mine association rules.

In the context of text mining, association rules have been used to discover potentially interesting relationships between concepts that co-occur in documents[10]. Association rules have also been used to establish rela- tionships between documents that do not share any terms. For example, consider an association rule of the form $B \Rightarrow C$, where B and C are words, whose confidence level is above the threshold. Then the document set re- trieved for C can be expanded to include B's document set. This allows us to find those documents that do not contain C but are still related to it because they have the related term B. Latent Semantic Indexing (LSI), a technique based on singular value decomposition, maps documents to a lower

dimension space where documents are considered close to each other if they share a sufficient number of term-based associations. This allows the original document set for a query to be expanded to include those documents that are closely located in the lower dimensional space.

The benefits of working with association rules are limited by the fact that mining association rules from text databases is a considerably intensive pro- cess. This is mainly because of the high dimensionality of the feature space. Hence the number of items (more realistically words) that need to be consid- ered when creating frequent itemsets is orders of magnitude larger than the number of items in a set of transactions (in a business setting). These factors reduce the effectiveness of the Apriori and Direct Hashing and Pruning Algorithms in text contexts.

4.2 Open and Closed Discovery

In [13] Swanson describes an approach that discovers relationships between concepts that are logically related but not bibliographically related, i.e., do not co-occur in any document. This approach forms the basis of the AR- ROWSMITH [14] system. The general idea is that two concepts A and C might be related if A co-occurs in some document with intermediate some concept B, and B co-occurs in some document with C . This implication based discovery process was successfully used by the authors to discover several novel relationships such as connections between Raynauds disease and fish oils [13], and migraine and magnesium [15].

Swanson and Smalheiser essentially designed two kinds of discovery processes that were later named 'open' and 'closed' discovery processes. The input to the open discovery process is a single concept (A) and the goal is to find related concepts (C) that do not co-occur with A in any document in the collection, i.e. the relationship between A and C has not been explored yet. Their process begins with a literature search for A and interesting phrases (B concepts) from titles of the retrieved documents (for A) are extracted. These B concepts are then used to initiate another round of literature search. Interesting phrases (C concepts) in the documents retrieved for the B concepts are then extracted. By analyzing (reading) the two sets of documents one can establish which B concepts connect the A and C concepts in a potentially interesting and novel way. In the closed discovery process one starts with both the A and C concepts and the goal is to establish potentially interesting B concepts that overlap with both A and C and connect them in a novel way. A big positive of both the open and closed discovery processes is that they have been successfully used to suggest novel connections between concepts. These processes, as implemented by Swanson [14] are however only semi-automatic as the B and C concepts have to manually selected and the connections between the concepts have to be manually inspected by a domain expert.

4.3 Metadata Mining

Metadata is defined as data that describes data. Instead of using free-text one may choose to use metadata where available for text mining. It is much easier to work with metadata as it is more structured that free text. In some cases the metadata for a text collection are manually created, as in MEDLINE, whereas in other cases metadata are automatically generated. Feature selection is an important part of text mining as it has a profound affect on the data model produced. With metadata a significant amount of feature selection is implicitly present. Using metadata can also significantly reduce the size of the feature space required to model the text collection[10]. This impacts suitability for large text collections.

There are several approaches in text mining that are built on metadata mining. In [15] the authors replicate Swanson and Smalheiser's experiments on Raynaud's disease and fish oils by limiting the interesting phrases extracted (Bs and Cs in the "open" discovery process) to metadata terms associated with the MEDLINE records. These metadata are known as MeSH (Medical Subject Heading) terms. In [47] Srinivasan replicates Swanson and Smalheiser's experiments using MeSH profiles for topics and also using IR- based term weighting schemes to identify interesting MeSH term connections between the topics. In [17] Feldman and Hirsh use the co-occurrence frequencies of metadata terms, which are taken from a hierarchically arranged vocabulary, to mine a text collection. As mentioned earlier, we were able to postulate a beneficial role for turmeric in retinal diseases. We were also able to postulate beneficial roles in the context of crohn's disease and problems related to the spinal cord.

Conclusion

This paper has been written with a two-fold aim. The first is to explored the concept of text mining and the second is to utilize text mining for identifying fresh information. We introduced thre approaches like Association Rule Mining, Open and Closed Discovery and meta data mining. In the context of text mining, association rules have been used to discover potentially interesting relationships between concepts that co-occur in documents. A big positive of open and closed discovery processes is that they have been successfully used to suggest novel connections between concepts. . Instead of using free-text one may choose to use metadata where available for text mining. It is much easier to work with metadata as it is more structured that free text.

References

[1] Fan W., Wallace L., Rich S., Zhang Z., Tapping into the power of text mining, Communications of ACM, vol. 49, no. 9, pp. 76-82, September 2006.

[2] Feinerer I., Hornik K., Meyer D., Text Mining Infrastructure in R, Journal of Statistical Software, vol. 25, no. 5, pp. 1-54, 2008.

[3] Stavrianou A., Andritsos P., Nicoloyannis N., Overview and Semantic Issues of Text Mining, ACM SIGMOD, vol. 36, no. 3, pp. 23-34, 2007.

[4] Gupta V., and Lehal G.S., A Survey of Text Mining Techniques and Applications, Journal of emerging technologies in Web Intelligence, vol. 1, no. 1, pp. 60-76, August 2009.

[5] Fan W., Wallace L., Rich S., Zhang Z., Tapping into the power of text mining, Communications of ACM, vol. 49, no. 9, pp. 76-82, September 2006.

[6] Haravu, L.J. and A. Neelameghan. "Text Mining and Data Mining in Knowledge Organization and Discovery: The Making of Knowledge-Based Products." Knowledge Organization and Classification in International Information Retrieval. Ed. Nancy J. Williamson and Clare Beghtol. Binghamton, NY: Haworth, 2003.

[7] Valitutti, Alessandro, and Carlo Strapparava. "Developing Affective Lexical Resources." Psychology Journal 2 (1): 2004. Pg. 61-83.

[8] Eiron N., and McCurley K.S., ―Analysis of Anchor Text for Web Search, Proc. 26th annual international ACM SIGIR conference on Research and development in IR, pp. 459 – 460, 2003.

[9] Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami. Mining association rules between sets of items in large databases. In Peter Buneman and Sushil Jajodia, editors, Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pages 207–216, Washington, D.C., May 1993.

[10] Catherine Blake and Wanda Pratt. Better rules, few features: A se- mantic approach to selecting features from text. In ICDM, pages 59–66, 2001.

[11] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In Jorge B. Bocca, Don R. Swanson and N. R. Smalheiser. Calcuim-independent phospho- lipase a2 and schizophrenia. Archives of General Psychiatry, 55(8):752– 753, 1998.

[12] Marc Weeber, Henny Klein, Lolkje T.W. de Jong-van den Berg, and Rein Vos. Using Concepts in Literature-Based Discovery: Simulating Swan- son's Raynaud-Fish Oil and Migraine-Magnesium Discoveries. Journal of the American Society for Information Science And Technology, 52(7):548–557, 2001.

[13] R. Feldman and H. Hirsh. Mining Text Using Keyword Distributions. Intelligent Information Systems, 10(3):281–300, 1998.

[14] Matthias Jarke, and Carlo Zan- iolo, editors, Proceedings of the 20th Int. Conf. Very Large Data Bases, VLDB, pages 487–499. Morgan Kaufmann, 1994.

[15] Jong Soo Park, Ming-Syan Chen, and Philip S. Yu. Using a hash- based method with transaction trimming for mining association rules. Knowledge and Data Engineering, 9(5):813–825, 1997.

[16] Don R. Swanson. Fish oil, raynaud's syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine, 30:7–18, 1986.

[17] Don R. Swanson and N. R. Smalheiser. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence, 91:183–203, 1997.