Information and Knowledge Management
ISSN 2224-5758 (Paper) ISSN 2224-896X (Online)
Vol.10, No.1, 2020

Text Mining Technique for Driving Potentially Valuable Information from Text

Fantaye Ayele
School of Informatics, Wolaita Sodo University, PO box 138, Wolaita Sodo, Ethiopia

Abstract
With the growing number of digitized documents and having large text databases, text mining will become increasingly important. Text mining can be a huge benefit for finding relevant and desired text data from unstructured data sources. Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. It is an important step of Knowledge Discovery process. The aim of the paper is to study the concept of Text Mining and various techniques with a particular focus on text mining process. In the text mining community have been trying to apply many methods such as rule-based, knowledge based, statistical and machine-learning-based approaches. Finally, the paper discusses issues towards the techniques for driving potentially valuable information from text and also, discuss on integration data mining. The paper ends with conclusion and the future line of works in the combining text mining and data mining techniques into a single system, a combination known as duo-mining, and also be more effective text mining techniques for contextual extraction.
Keywords: Data mining, Information Extraction, Information Retrieval, Text Mining
DOI: 10.7176/IKM/10-1-01
Publication date: January 31st 2020

1. Introduction

Text mining is a burgeoning technology that is still, because of its newness and intrinsic difficulty, in a fluid state-akin, perhaps, to the state of machine learning in the mid- 1980s. It is also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. It can be used as a preprocessing technique to harvest data and get an initial understanding of the patterns that exist in the data (Malik, 2006).Today, large amount of information is available in textual form in databases and online sources. Approximately 90% of the world's data is held in unstructured formats (source: Oracle Corporation). However, large amounts of data such as textual data are unstructured, and challenge simple attempts to make sense of it. In this context, manual analysis and effective extraction of useful information are not possible. As a result, text mining techniques are being developed to automate the process of analyzing large textual collections (Raymond and Un Yong, 2003). Because of the internet there has been a huge growth of easily available textual information over the last decade in the form of documents, news, blogs, forums, emails, and etc. This increased amount of available textual information has led to a research field devoted to knowledge discovery in unstructured data (textual data) known as text mining. Text Mining is to exploit information contained in textual documents in various ways, including discovery of patterns and trends in data, associations among entities and predictive rules (Konchady, 2006). The results can be important both for: the analysis of the collection, and providing intelligent navigation and browsing methods (Ananiadou and McNaught, 2006).

2. The aim of the paper

- to study the concept of text mining
- to identify some of techniques used in text mining
- to identify effective technique for driving potentially valuable information from text

3. Theoretical backgrounds

3.1 Text mining

Text mining is the use of automated methods for exploiting the enormous amount of knowledge available in text documents. Its represents a step forward from text retrieval. Text mining is a relatively new and vibrant research area which is changing the emphasis in text-based information technologies from the level of retrieval to the level of analysis and exploration. The prime aim of the text mining is to identify the useful information without duplication from various documents with synonymous understanding. Text Mining is an empirical tool that has a capacity of identifying new information that is not apparent from a document collection (Vidhya and Aghila, 2010).

3.2 Text mining process

Text Mining starts with a collection of documents; which would retrieve a particular document and preprocess it by checking format and character sets. Then it would go through a text analysis phase, sometimes repeating techniques until information is extracted. Three text analysis techniques are shown in the example, but many other combinations of techniques could be used depending on the goals of the organization. The resulting information can be placed in a management information system (Ning, 2008), yielding an abundant amount of knowledge for the user of that system. The following figure 1 explores the detail processing methods in Text Mining. The document collection from figure 1 is set of files might be with any extension like PDF, txt or even flat file extension which are normally collected and named as noisy unstructured text data found in informal settings such as online chat, SMS, emails, message boards, newsgroups, blogs, wikis and web pages. Also, text data set is created by processing spontaneous speech, printed text and handwritten text contains processing noise. 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. The process of using Knowledge Discovery in Database, which is the fundamental step in Text Mining, knowledge experts can obtain important strategic information for their business. KDD has more intensive transformation methods to cross examine traditional databases, where data are in structured form, by automatically finding new and unknown patterns in huge quantity of data. Mostly, structured data represent only a little part of the overall organization knowledge and the knowledge is incorporated in textual documents (Vidhya and Aghila, 2010).

4. Text Mining Techniques

The process of extracting information and knowledge from unstructured text led to the need for various mining techniques for useful pattern discovery. Text mining techniques are incorporated for the applications like Information retrieval, Information Extraction, Summarization and Topic Discovery for necessary knowledge discovery process. Text mining involves the application of techniques from areas such as information retrieval, natural language processing, information extraction and data mining.

4.1. Information Extraction

A starting point for computers to analyze unstructured text is to use information extraction. Information Extraction is the process of automatically obtaining structured data from an unstructured natural language document. Often this involves defining the general form of the information that we are interested in as one or more templates, which are then used to guide the extraction process. Tasks that IE systems can perform include: Term analysis, which identifies the terms in a document, where a term may consist of one or more words. 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 in a document, such as the 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 (Jusoh and Alfawareh, 2012). Therefore the IE task is defined by its input and its extraction target. The input can be unstructured documents like free texts that are written in natural language or the semi-structured documents that are pervasive on the Web, such as tables or itemized and enumerated lists. Using IE approach, events, facts and entities are extracted and stored into a structured database. Then data mining techniques can be applied to the data for discovering new knowledge.
Information extraction is based on understanding of the structure and meaning of the natural language in which documents are written, and the goal of information extraction is to accumulate semantic information from text. Technically, extracting information from texts requires two pieces of knowledge: lexical knowledge and linguistic grammars. Using the knowledge we are able to describe the syntax and semantic of the text (Nazarenko, 2005).

4.2. Information Retrieval

Information retrieval is concerned with identifying documents that are most relevant to a user's need within a very large set of documents. More precisely, given a large database of documents, and a specific information need—usually expressed as a query by the user—the goal of information retrieval methods is to find the documents in the database that satisfy the information need. Naturally, the task has to be performed accurately and efficiently.IR systems are often used in libraries, where the documents are typically not the books themselves but digital records containing information about the books. This is however changing with the advent of digital libraries, where the documents being retrieved are digital versions of books and journals (Gupta and Lehal, 2009).

4.3 Topic Tracking

A topic tracking 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. Topic tracking technology does have limitations, however. For example, if a user sets up an alert for "text mining", s/he will receive several news stories on mining for minerals, and very few that are actually on text mining. Some of the better text mining tools let users select particular categories of interest or the software automatically can even infer the user's interests based on his/her reading history and click-through information. It could be used in the medical industry by doctors and other people looking for new treatments for illnesses and who wish to keep up on the latest advancements (Gupta and Lehal, 2009).

4.4 Text summarization

Text summarization is immensely helpful for trying to figure out whether or not a lengthy document meets the user's needs and is worth reading for further information. With large texts, text summarization software processes and summarizes the document in the time it would take the user to read the first paragraph. The key to summarization is to reduce the length and detail of a document while retaining its main points and overall meaning. The challenge is that, although computers are able to identify people, places, and time, it is still difficult to teach software to analyze semantics and to interpret meaning (Jisha, 2010). One of the strategies most widely used by text summarization tools, sentence extraction, extracts important sentences from an article by statistically weighting the sentences. Further heuristics such as position information are also used for summarization (Gupta and Lehal, 2009).

4.5 Text categorization

Text categorization (or text classification) is the assignment of natural language documents to predefined categories according to their content (Sebastiani, 2002). The set of categories is often called a "controlled vocabulary." Document categorization is a long-standing traditional technique for information retrieval in libraries, where subjects rival authors as the predominant gateway to library contents—although they are far harder to assign objectively than authorship. The Library of Congress Subject Headings (LCSH) is a comprehensive and widely used controlled vocabulary for assigning subject descriptors. They occupy five large printed volumes of 6,000 pages each—perhaps two million descriptors in all. The aim is to provide a standardized vocabulary for all categories of knowledge, descending to quite a specific level, so that books –on any subject, in any language—can be described in a way that helps librarians retrieve all books on a given subject (Witten and Bainbridge, 2003).

4.6. Clustering

Clustering is a technique used to group similar documents, but it differs from categorization in that documents are clustered on the fly instead of through the use of predefined topics. Another benefit of clustering is that documents can appear in multiple subtopics, thus ensuring that a useful document will not be omitted from search results. A basic clustering algorithm creates a vector of topics for each document and measures the weights of how well the document fits into each cluster. Clustering technology can be useful in the organization of management information systems, which may contain thousands of documents (Jisha, 2010).

The first step in text clustering is to transform documents, which typically are strings of characters into a suitable representation for the clustering task.

1. Remove stop-words: The stop-words are high frequent words that carry no information (i.e. pronouns, prepositions, conjunctions etc.). Remove stop-words can improve clustering results.
2. Stemming: By word stemming it means the process of suffix removal to generate word stems. This is done to group words that have the same conceptual meaning, such as work,worker, worked and working.
3. Filtering: Domain vocabulary V in ontology is used for filtering. By filtering, document is considered with related domain words (term). It can reduce the documents dimensions. A central problem in statistical text clustering is the high dimensionality of the feature space(Veni, 2013).

Standard clustering techniques cannot deal with such a large feature set, since processing is extremely costly in computational terms. We can represent documents with some domain vocabulary in order to solving the high dimensionality problem. In the beginning of word clustering, one word randomly is chosen to form initial cluster. The other words are added to this cluster or new cluster, until all words are belong to m clusters. This method allow one word belong to many clusters and accord with the fact. This method implements word clustering by calculating word relativity and then implements text classification (Gupta and Lehal, 2009).

5. Conclusion

Text Mining also known as Text Data Mining or KDT refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. 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 corpora can be developed. In particular, the growing interaction between computational linguistics and machine learning is critical to the development of effective text-mining systems. Some of text mining techniques are: information retrieval, Information Retrieval, Topic Tracking, Text summarization Text categorization clustering. This mostly deals with finding documents that satisfy a particular information need within a large database of documents, information extraction (IE), a subfield of NLP, centered on finding explicit entities and facts in unstructured text. Finally, text mining is the combined, automated process of analyzing unstructured, natural language text in order to discover information and knowledge that are typically difficult to retrieve.

However, form the above methods of text mining an information extraction technique is effective technique for driving valuable information from text. This technique focuses on extracting information from actual texts. The goal of text mining is to discover knowledge in unstructured text. The related task of IE concerns transforming unstructured text into a structured database by locating desired pieces of information. Although handmade Information Extraction systems have existed for a while, automatic construction of information extraction systems using machine learning is more recent.

Hence, the following recommendations are identified for further work are integration of data mining and text mining techniques into a single system, a combination known as duo-mining, and also be more effective text mining techniques for contextual extraction, at the same time increasing the amount of annotated review data for better classifier performance through actively learning.

References

Ananiadou, S. and McNaught, J. (2006). Text Mining for Biology and Biomedicine. Artech House Publishers, ISBN 1-58053-984-X, 302pp

Gupta, V. and Lehal, G. (2009), A Survey of Text Mining Techniques and Applications, Panjab University Chandigarh and Patiala: India.

Jusoh, S.andAlfawareh, H. (2012), Techniques, Applications and Challenging Issue in Text Mining, International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, ISSN (Online): 1694-0814

Malik, R. (2006). Conan: Text mining in biomedical domain, PhD thesis: Utrecht University, Austria.

Nazarenko, D, (2005). A Ontologies and information extraction: A necessary symbiosis, In Ontology Learning from Text: Methods, Evaluation and Applications: IOS Press Publication.

Ning, Z. et al., (2008), A Visualization Model for Information Resources Management, 12th International Conference Information Visualization, China, IEEE, 57- 62.

Raymond. J and Yong.N, (2003). Text Mining with Information Extraction: vol 10, p 821- 855.

Sebastiani, F. (2002).Machine learning in automated text categorization. ACM Computing Surveys: Vol. 34, No. 1, pp. 1–47

Veni, M., Praveena, M. and GanaPriya, V., (2013).A Review on Duo Mining Techniques, International Journal of Science and Research (IJSR), India Online ISSN: 2319-7064Volume 2 Issue 3

Vidhya, K. and Aghila, G. (2010). Text Mining Process, Techniques and Tools: an Overview International Journal of Information Technology and Knowledge Management, Volume 2, No. 2, pp. 613-622

Witten, I.H. and Bainbridge, D. (2003), How to build a digital library. Morgan Kaufmann, San Francisco, CA.

.

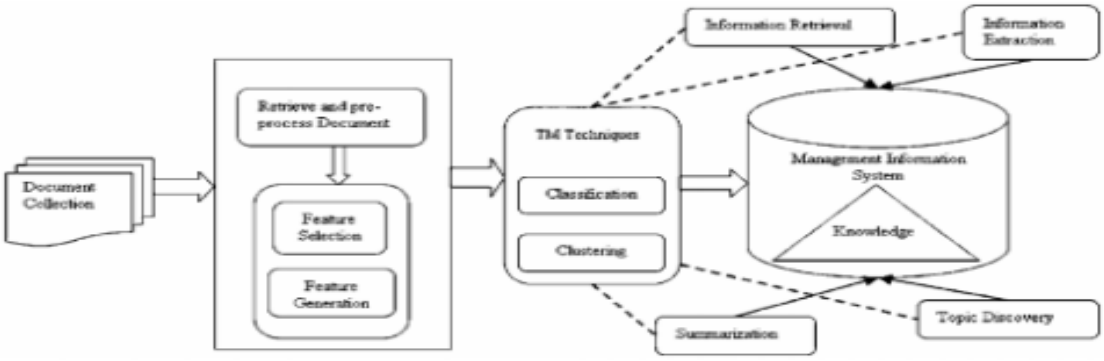

Figure 1: Text Mining Process