International Journal of Education and Science Research Review
E-ISSN 2348-6457
March-April-2018, Volume-5, Issue-2 P-ISSN 2349-1817
www.ijesrr.org Email- editor@ijesrr.org

DIFFERENCE BETWEEN TEXT MINING AND DATA MINING APPLICATIONS

Vikas Jain
Assistant Professor
Dept. of Computer Application
C.C.S. University, Meerut

ABSTRACT
Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. In this paper, a Survey of Text Mining techniques and applications have been presented. Primary objective of the Feature Extraction operation is to identify facts and relations in text. Most of the times this includes distinguishing which noun phrase is a person, place, organization or other distinct object. Feature Extraction algorithms may use dictionaries to identify some terms and linguistic patterns to detect others. The Text-base navigation enables users to move about in a document collection by relating topics and significant terms. It helps to identify key concepts and additionally presents some of the relationships between key concepts.
KEYWORDS: Text Mining, document

INTRODUCTION

Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text mining is different from what are familiar with in web search. In search, the user is typically looking for something that is already known and has been written by someone else. The problem is pushing aside all the material that currently is not relevant to your needs in order to find the relevant information. In text mining, the goal is to discover unknown information, something that no one yet knows and so could not have yet written down. Text mining is a variation on a field called data mining, that tries to find interesting patterns from large databases. Text mining, also known as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT), refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Text mining is a young interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics. As most information (over 80%) is stored as text, text mining is believed to have a high commercial potential value. Knowledge may be discovered from many sources of information, yet, unstructured texts remain the largest readily available source of knowledge. The problem of Knowledge Discovery from Text (KDT) is to extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP) techniques. Its aim is to get insights into large quantities of text data. KDT, while deeply rooted in NLP, draws on methods from statistics, machine learning, reasoning, information extraction, knowledge management, and others for its discovery process. KDT plays an increasingly significant role in emerging applications, such as Text Understanding. Text mining is similar to data mining, except that data mining tools are designed to handle structured data from databases , but text mining can work with unstructured or semi-structured data sets such as emails, full-text documents and HTML files etc. As a result, text mining is a much better solution for companies. To date, however, most research and development efforts have centered on data mining efforts using structured data. The problem introduced by text mining is obvious: natural language was developed for humans to communicate with one another and to record information, and computers are a long way from comprehending natural language. Humans have the ability to distinguish and apply linguistic patterns to text and humans can easily overcome obstacles that computers cannot easily handle such as slang, spelling variations and contextual meaning. However, although our language capabilities allow us to comprehend unstructured data, we lack the

computer's ability to process text in large volumes or at high speeds. Figure 1 on next page, depicts a generic process model for a text mining application. Starting with a collection of documents, a text mining tool 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, yielding an abundant amount of knowledge for the user of that system.

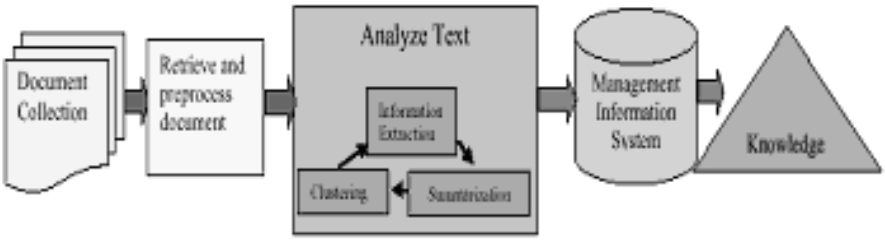

Figure 1. An example of Text Mining

TECHNOLOGY FOUNDATIONS

Although the differences in human and computer languages are expansive, there have been technological advances which have begun to close the gap. The field of natural language processing has produced technologies that teach computers natural language so that they may analyze, understand, and even generate text. Some of the technologies that have been developed and can be used in the text mining process are information extraction, topic tracking, summarization, categorization, clustering, concept linkage, information visualization, and question answering. In the following sections we will discuss each of these technologies and the role that they play in text mining. We will also illustrate the type of situations where each technology may be useful in order to help readers identify tools of interest to themselves or their organizations.

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. There are many areas where topic tracking can be applied in industry. It can be used to alert companies anytime a competitor is in the news. This allows them to keep up with competitive products or changes in the market. Similarly, businesses might want to track news on their own company and products. It could also 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. Individuals in the field of education could also use topic tracking to be sure they have the latest references for research in their area of interest. Keywords are a set of significant words in an article that gives high-level description of its contents to readers. Identifying keywords from a large amount of on-line news data is very useful in that it can produce a short summary of news articles. As on-line text documents rapidly increase in size with the growth of WWW, keyword extraction has become a basis of several text mining applications such as search engine, text categorization, summarization, and topic detection. Manual keyword extraction is an extremely difficult and time consuming task; in fact, it is almost impossible to extract keywords manually in case of news articles published in a single day due to their volume.

INFORMATION EXTRACTION

A starting point for computers to analyze unstructured text is to use information extraction. Information extraction software identifies key phrases and relationships within text. It does this by looking for predefined sequences in text, a process called pattern matching. The software infers the relationships between all the identified people, places, and time to provide the user with meaningful information. This technology can be very useful when dealing with large volumes of text. Traditional data mining assumes that the information to be "mined" is already in the form of a relational database. Unfortunately, for many applications, electronic information is only available in the form of free natural language documents rather than structured databases. Since IE addresses the problem of transforming a corpus of textual documents into a more structured database, the database constructed by an IE module can be provided to the KDD module for further mining of knowledge as illustrated in Figure 2.

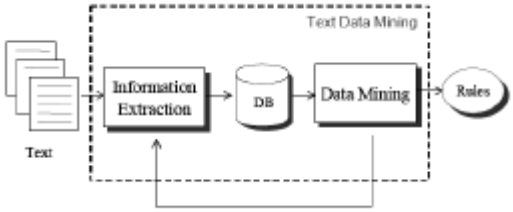

Figure 2. Overview of IE-based text mining framework

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. Generally, when humans summarize text, we read the entire selection to develop a full understanding, and then write a summary highlighting its main points. Since computers do not yet have the language capabilities of humans, alternative methods must be considered. 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. For example, summarization tools may extract the sentences which follow the key phrase "in conclusion", after which typically lie the main points of the document. Summarization tools may also search for headings and other markers of subtopics in order to identify the key points of a document. Microsoft Word's AutoSummarize function is a simple example of text summarization. Many text summarization tools allow the user to choose the percentage of the total text they want extracted as a summary. Summarization can work with topic tracking tools or categorization tools in order to summarize the documents that are retrieved on a particular topic. If organizations, medical personnel, or other researchers were given hundreds of documents that addressed their topic of interest, then summarization tools could be used to reduce the time spent sorting through the material. Individuals would be able to more quickly assess the relevance of the information to the topic they are interested in.

CATEGORIZATION

Categorization involves identifying the main themes of a document by placing the document into a pre-defined set of topics. When categorizing a document, a computer program will often treat the document as a "bag of words." It does not attempt to process the actual information as information extraction does. Rather, categorization only counts words that appear and, from the counts, identifies the main topics that the

document covers. Categorization often relies on a thesaurus for which topics are predefined, and relationships are identified by looking for broad terms, narrower terms, synonyms, and related terms. Categorization tools normally have a method for ranking the documents in order of which documents have the most content on a particular topic. As with summarization, categorization can be used with topic tracking to further specify the relevance of a document to a person seeking information on a topic. The documents returned from topic tracking could be ranked by content weights so that individuals could give priority to the most relevant documents first. Categorization can be used in a number of application domains. Many businesses and industries provide customer support or have to answer questions on a variety of topics from their customers. If they can use categorization schemes to classify the documents by topic, then customers or endusers will be able to access the information they seek much more readily. The goal of text categorization is to classify a set of documents into a fixed number of predefined categories. Each document may belong to more than one class.

CONCEPT LINKAGE

Concept linkage tools connect related documents by identifying their commonly-shared concepts and help users find information that they perhaps wouldn't have found using traditional searching methods. It promotes browsing for information rather than searching for it. Concept linkage is a valuable concept in text mining, especially in the biomedical fields where so much research has been done that it is impossible for researchers to read all the material and make associations to other research. Ideally, concept linking software can identify links between diseases and treatments when humans can not. For example, a text mining software solution may easily identify a link between topics X and Y, and Y and Z, which are well-known relations. But the text mining tool could also detect a potential link between X and Z, something that a human researcher has not come across yet because of the large volume of information s/he would have to sort through to make the connection.

QUESTION ANSWERING

Another application area of natural language processing is natural language queries, or question answering (Q&A), which deals with how to find the best answer to a given question. Many websites that are equipped with question answering technology, allow end users to "ask" the computer a question and be given an answer. Q&A can utilize multiple text mining techniques. For example, it can use information extraction to extract entities such as people, places, events; or question categorization to assign questions into known types (who, where, when, how, etc.). In addition to web applications, companies can use Q&A techniques internally for employees who are searching for answers to common questions. The education and medical areas may also find uses for Q&A in areas where there are frequently asked questions that people wish to search.

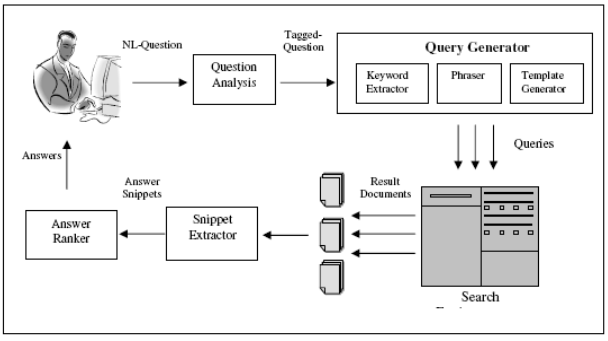

Figure 3. Architecture of Question answering system

The system takes in a natural language (NL) question in English from the user. This question is then passed to a Part-of-Speech (POS) tagger which parses the question and identifies POS of every word involved in the question. This tagged question is then used by the query generators which generate different types of queries,

which can be passed to a search engine. These queries are then executed by a search engine in parallel. The search engine provides the documents which are likely to have the answers we are looking for. These documents are checked for this by the answer extractor. Snippet Extractor extracts snippets which contain the query phrases/words from the documents. These snippets are passed to the ranker which sorts them according to the ranking algorithm.

ASSOCIATION RULE MINING

Association rule mining (ARM) is a technique used to discover relationships among a large set of variables in a data set. It has been applied to a variety of industry settings and disciplines but has, to date, not been widely used in the social sciences, especially in education, counseling, and associated disciplines. ARM refers to the discovery of relationships among a large set of variables, that is, given a database of records, each containing two or more variables and their respective values, ARM determines variable-value combinations that frequently occur. Similar to the idea of correlation analysis, in which relationships between two variables are uncovered, ARM is also used to discover variable relationships, but each relationship may contain two or more variables. ARM has been extensively employed in business decision-making processes. ARM discovers what items customers typically purchase together. These associations can then be used by a supermarket, for example, to place frequently co-purchased items in adjacent shelves to increase sales. Thus, if bread and cereal are often purchased together, placing these items in close proximity may encourage customers to buy them within single visits to the supermarket. ARM is a technique that is part of the field of data mining. Also known as knowledge discovery in databases. In Association Rules for Text Mining, The focus is to study the relationships and implications among topics, or descriptive concepts, that are used to characterize a corpus. The goal is to discover important association rules within a corpus such that the presence of a set of topics in an article implies the presence of another topic. For example, one might learn in headline news that whenever the words “Greenspan” and “inflation” occur, it is highly probably that the stock market is also mentioned. Figure15a shows a high-level system overview of the topic association mining system. A corpus of narrative text is fed into a text engine for topic extractions. The mining engine then reads the topics from the text engine and generates topic association rules. Finally, the resultant association rules are sent to the visualization system for further analysis.

DIFFERENCE BETWEEN TEXT MINING AND DATA MINING

The difference between regular data mining and text mining is that in text mining the patterns are extracted from natural language text rather than from structured databases of facts. One application of text mining is in, bioinformatics where details of experimental results can be automatically extracted from a large corpus of text and then processed computationally. Text-mining techniques have been used in information retrieval systems as a tool to help users narrow their queries and to help them explore other contextually related subjects.

Text Mining seems to be an extension of the better known Data Mining. Data Mining is a technique that analyses billions of numbers to extract the statistics and trends emerging from a company's data. This kind of analysis has been successfully applied in business situations as well as for military, social, government needs. But, only about 20% of the data on intranets and on the World Wide Web are numbers - the rest is text.

The information contained in the text (about 80% of the data) is invisible to the data mining programs that analyze the information flow in corporations. Text mining tries to apply these same techniques of Data mining to unstructured text databases. To do so, it relies heavily on technology from the sciences of Natural Language Processing (NLP), and Machine Learning to automatically collect statistics and infer structure and meaning in otherwise unstructured text. The usual approach involves identifying and extracting key features

from the text that can be used as the data and dimensions for analysis. This process is called feature extraction, is a crucial step in text mining.

Text mining is a comprehensive technique. It relates to data mining, computer language, information searching, natural language comprehension, and knowledge management. Text mining uses data mining techniques in text sets to find out connotative knowledge. Its object type is not only structural data but also semistructural data or non-structural data. The mining results are not only general situation of one text document but also classification and clustering of text sets.

CONCLUSIONS

To visualize text documents, we first need to convey to the analyst the underlying relationships between documents in a geometrically intuitive manner , we must preserve certain Visualization characteristics for the rendering to be meaningful. For example, documents that are close to each other in content should also be geometrically close to each other. Additionally, the analyst should be able to provide his or her own insight in determining what it means for documents to have close content or similar meaning. For users to provide their own insight, they have to access the meaning of the Visualization. That is, users must be able to interpret the rendered data so that the topic document relations are clearly defined.

REFERENCES

1. Alessio Canzonetti , Francesca Della Ratta-Rinald and Bhupesh K. Singh, (2009), “Understanding Text Mining:a Pragmatic Approach”, Roam, Italy.
2. Babis Theodoulidis Manchester, (2016), “Knowledge Discovery in Text and Text Mining Software”, Centre for Research in Information Management, UK
3. Berry Michael W., (2015), “Automatic Discovery of Similar Words”, in “Survey of Text Mining: Clustering, Classification and Retrieval”, Springer Verlag, New York, LLC, 24-43.
4. Fang Chen, Kesong Han and Guilin Chen (2018), “An approach to sentence selection based text summarization”, Proceedings of IEEE TENCON02, 489- 493.
5. Haiying Tu, Jianhui Luo and Krishna R. Pattipati (2012), “Experiments on Supervised Learning Algorithms for Text Categorization”, International Conference , IEEE computer society, 1-8.
6. Henzinger M.R. (2015), “Finding related pages in the world wide web”, Computer Networks, 31(11-16):1467-1479.
7. Jian-Suo Xu (2013), “TCBPLK: A new method of text categorization ”, Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong,, IEEE , 3889-3892.
8. Kleinberg J.M., (1999), “Authoritative sources in hyperlinked environment”, Journal of ACM, Vol.46, No.5, 604-632.
9. Lijun Jiang (2014), “Performing Text Categorization on Manifold”, 2014 IEEE International Conference on Systems, Man, and Cybernetics, Taipei, Taiwan , IEEE , 3872-3877.
10. Liritano S. and Ruffolo M., (2016), “ Managing the Knowledge Contained in Electronic Documents: a Clustering Method for Text Mining”, IEEE, 454-458 , Italy.
11. Liu Lizhen, and Chen Junjie, China (2009), “ Research of Web Mining”, Proceedings of the 4th World Congress on Intelligent Control and Automation, IEEE, 2333-2337.
12. Navathe, Shamkant B., and Elmasri Ramez, (2000), “Data Warehousing And Data Mining”, in “Fundamentals of Database Systems”, Pearson Education pvt Inc, Singapore, 841-872.
13. Page L.(2015), “ The anatomy of a largescale hyper textual Web search engine”, Computer Networks and ISDN Systems, 30(1-7): 107-117.
14. U. Ackermann, B. Angelini, F. Brugnara, M. Federico ,D. Giuliani, R. Gretter, G. Lazzari and H. Niemann (2015),“SpeeData: Multilingual Spoken Data Entry”, International Conference, IEEE, Trento, Italy., 2211 - 2214.