Wuhan University of technology

School of computer science and technology

Modern software Engineering

Report on " TEXT DATA MINING: OVERVIEW"

Submitted by: OUMOUSS EL MEHDI

Student No: 58944

TEL No: 13125029535

Email: oumoussmehdi@hotmail.com

Submitted to: Prof. Dr. Jianwen Xiang

TEXT DATA MINING: OVERVIEW

Abstract

Text mining (TM) is used to extract hidden valuable information from semi-structured or unstructured data. The amount of information available today is increasing at an exponential rate. This reality lead us to investigate various text mining techniques. These techniques and processes discover and present knowledge, facts and relationships. Which are otherwise locked in textual form, impenetrable to automated processing. All in all TM helps industries and researchers innovate and become more efficient [1].

In this paper, our focus is to review the basic concept of text mining and its applications. We will shed the light on the basic differences between text mining and relative terminologies. Additionally we will present an overview of text mining sources, limitations and process.

Keywords: Text Mining (process), Text Analytics, Data mining, Free Text, Biomedical Text.

1. Introduction

In the contemporary world the text is the most common means for exchanging information. Data take three form either it is structured (databases), or semi structured (html, xml ...) or unstructured (free text). Huge amount of data today are stored in unstructured format, approximately 80% percent of the corporate data.

Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of discovering hidden, useful and interesting High-quality pattern and information from unstructured text documents. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness.

Typical text mining tasks include text categorization, clustering, concept/entity extraction, sentiment analysis, document summarization, and entity relation modeling.

The overarching goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP) and analytical methods. And, thus make the information contained in the text accessible to the various data mining techniques.

Information can be extracted to derive summaries for the words contained in the documents or to compute summaries for the documents based on the words contained in them. Hence, you can analyze words, clusters of words used in documents, etc., or you could analyze documents and determine similarities between them or how they are related to other variables of interest in the data mining project [2].

2. Text Mining Characteristics

2.1 Interdisciplinary

Text mining is an interdisciplinary field that draws on information retrieval, data mining, machine learning, statistics, and computational linguistics and other neighboring fields.

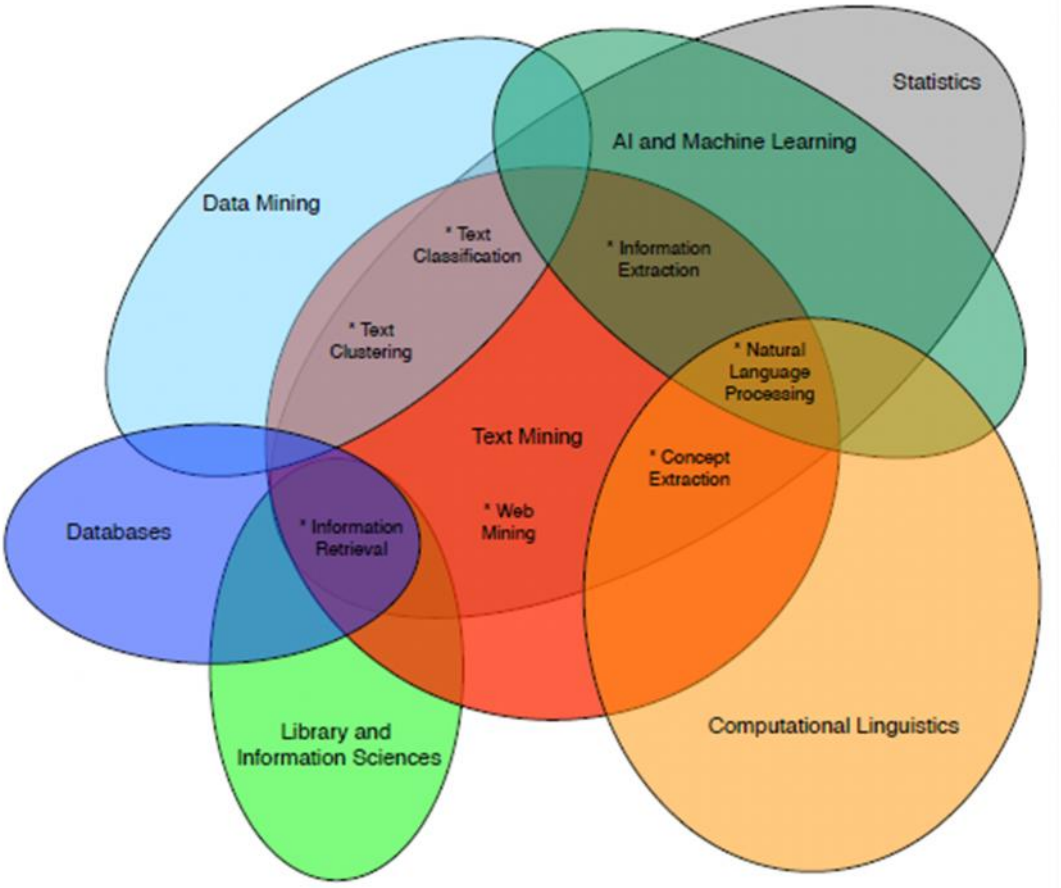

Figure1: Text Mining interdisciplinary [3]

As we can see from the figure1, text mining share boundaries if to say with so many other fields and technologies, which make it sometime hard to draw the difference between all these disciplines.

Martin Hearst from University of California, in her essay *what is text mining?* tries to shed the light on the differences between text mining and the other disciplines, and that the main goal of a text miner is to discover heretofore unknown information, something that no one yet knows and so could not have yet written down. And she state a beautiful analogy, where she explain that discovering new knowledge vs. showing trends is like the difference between a detective following clues to find the criminal vs. analysts looking at crime statistics to assess overall trends in car theft [4].

The main 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. Databases are designed for programs to process automatically; text is written for people to read.

Text mining is also different from web mining. One major difference is the web sources which represent a wider range of data types, structured, semi-structured (xml, Jason…), unstructured and multimedia (videos, audio …). Additionally In web 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 isn't relevant to your needs in order to find the relevant information.

As for Information Retrieval (IR) the problem consist just to locate the desired information which is already know, in other word there is no novelty when we deal with IR.

Another field worth mentioning is Computation Linguistics (CPL) (also known as natural language processing) where The main challenges is natural language understanding, in other words, enabling computers to derive meaning from human or natural language input [4][5][7].

Here below is a table that summarize the ideas stated above.

*finding new, relevant information between otherwise worthless data

	Finding Patterns	Finding Nuggets*	
		Novel	**Non-novel**
non-textual data	standard data mining	no comparable form	Databases
Textual Data	Computational linguistics	Real text data mining	Information retrieval

Table1: A classification of mining applications [5]

2.2 Sources of Text

As for any analysis project before we start we need data, and for text mining it is obvious we need text. So here we are going to state some the possible resources. There is numerous possible sources of text that can hold valuable information. Waiting to be discovered using text mining, the sources we are going to present are by no means a complete list:
Customer feedback: costumer are always talking so are we listening. Ways to get feedback from customers include, marketing survey, online product reviews and chatters from message board and social media. For example with twitter it is easy to search for key terms or hashtags then the result twits can be fed to a text miner to find overall sentiments and patterns.
Emails: is a very famous example, and everybody can see clearly the improvement of spam detection due to text mining techniques.
Comment fields for example represent a good source for text mining where we can get for example some collection that may contain description of issue and resolution for warranty claims, or maybe we may have accident reports, where we have a description of a certain event or its surrounding conditions.
Published articles are a rich source of information that can be accessed in electronic form text mining this documents can represent trends in current research or find new patterns that form new hypothesis to explore. Last and not least, website content. The internet is full of information hence, it is very important to periodically mine for example the web site of some company's competitor, in order to see what are their new products and services, another example which come to mind is what the bloggers talk about which can give us insight about our future decision makings. [6]

2.3 Challenges

Text mining is a very challenging task, because when dealing with text we can face so many problems or inconveniences. For instance, natural language text is often inconsistent. It contains ambiguities caused by inconsistent syntax and semantics. The text may include slangs, language specific to some specific domain (eg. Industry, medicine, etc), double intenders and sarcasm [9]. In addition, the text data is not always well-organized (semi-structured or unstructured data). Some learning techniques may need annotated training examples, etc.

2.4 Limitations

As any other tool or technology, text mining suffers also from several limitation. Fundamental ones are first, that we still are not able to write programs that fully interpret text mining especially at high level as the human brain do [4]. Second, that the information that we may need sometimes is just not available in text format; maybe it is in any other format such us video or audio. Third, all the text miner are application or task dependent which means even when they perform good for some collection of files of a certain domain let's say news they usually lose in performance once applied to another collection or file data set of some other domain. And since we are talking about limitations, it is worth to mention that there is other types of limitations, for example due to the lack of flexibilities in some countries law concerning copyright, the mining of in-copyright works without the permission of the copyright owner is illegal. Add to this, that we may encounter a problem of accessibility when it comes to get a full paper content in a publisher web site.

3. Text Mining Process

Text mining is surely not an easy task, thus it is not weird if the text mining process seems complicated, especially that this process go through several stages in order to get to the final result which is mining and extracting valuable information.
As it is shown in the Figure bellow, text mining process goes through six different stages, starting from our primary material which is the free text. The collection of documents first should get preprocessed, then comes the attribute generation and selection, and right after we apply data mining techniques and finally evaluation and interpretation of the results. Each stage within the process deals with different sub-tasks and its output result represent the input of the stage that comes after.

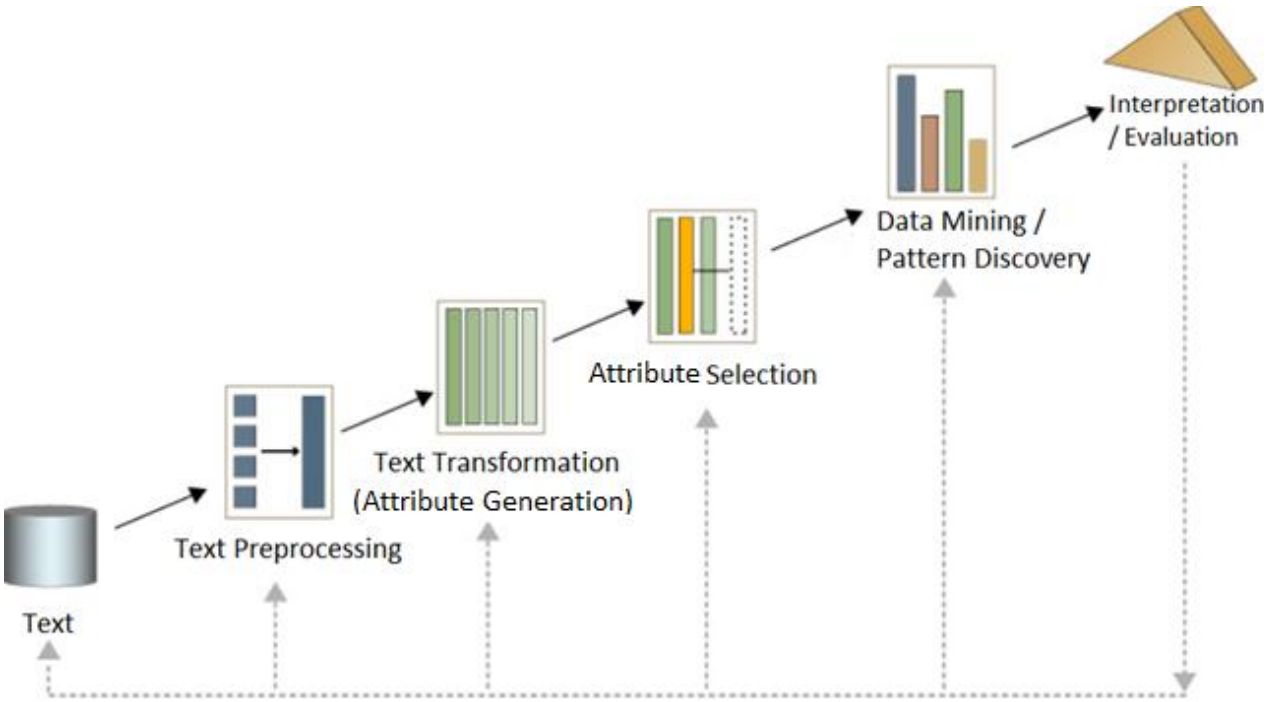

Figure2: the different stages of Text Mining Process [12]

3.1 Text:

Dealing with free text represent a hard challenge. First, because we are dealing here with large file data bases (High dimensionality). Second, due to the different characteristics that text may entail. We should understand that Text may come in several input modes. Text is intended for different consumers, i.e. different languages (human consumers) and different formats (automated consumers). Add to this Words and phrases create context for each other which we design as Rependency. also text contains Ambiguity wither in word level or sentence level, it may contains also noisy data (Erroneous or Misleading data). We should not forget to deal with synonyms too. finally, maybe we will have to Excluding certain characters, short words or numbers which does not represent any interest for our application.

3.2 Text Preprocessing

Text preprocessing depends on a numerous of techniques that prepare the text we have for the next step which is attribute generation. Here below is not in case a complete list of the methods used:
Text cleanup*:* e.g., remove ads from web pages, normalize text converted from binary formats, deal with tables, figures and formulas, etc.

Tokenization: Splitting up a string of characters into a set of tokens.
Need to deal with issues like: Apostrophes, e.g., “John’s sick”, is it 1 or 2 tokens? Hyphens, e.g., database vs. data-base vs. data base. How should we deal with term such us “C++”, “A/C”, “:-)”, “…”?, Is the amount of white spaces significant? And other question we may ask depending on the text available and also the application.

Parts Of Speech tagging: The process of marking up the words in a text with their corresponding parts of speech (noun, verb, complement ...). the POS tagging take two

form a rule based one which depends on grammatical rules , or a statistical based form which relies on words order probabilities.

Word Sense Disambiguation: Determining in which sense a word having a number of distinct senses is used in a given sentence. For instance the following sentence:
“The king saw the rabbit with his glasses” How many meanings?

Semantic Structures: there exists two methods for semantic structures. First, the full parsing where a parse tree for each sentence is produced. Second, chunking with partial parsing. Which produces syntactic constructs like Noun Phrases and Verb Groups for a sentence.

3.3 Attribute Generation

Text document is represented by the words (features) it contains and their occurrences. Two main approaches of document representation are “Bag of words” and Vector Space model.In “Bag of words” representation, each word is represented as a separate variable having numeric weight. While vector space model is an algebraic model for representing text documents as vectors of identifiers, such as, for example, index terms. In both cited model, we affecte to words a weigth, there is many methods to calculate some word’s weight but the most popular and widely used schema is the normalized word frequency tf-idf (term frequency – inverse document frequency)[8].
We use then a classifier to automatically generate labels (attributes) from the features we feed into it.

3.4 Attribute Selection

Feature Selection: select just a subset of the features to represent a document.
Can be viewed as creating an improved text representation. And the reason by doing this is that, there many features have little information content: e.g. stop words.
Some features are misleading or redundant.

Stop words removal: The most common words are unlikely to help text mining, e.g., “the”, “a”, “an”, “you”…

Stemming: Identifies a word by its root. This technique helps reduce dimensionality (number of features). e.g. flying, flew →fly. The two common algorithms are Porter’s Algorithm and KSTEM Algorithm.

> Example
> Original Text: Document will describe marketing strategies carried out by U.S. companies for their agricultural chemicals, report predictions for market share of such chemicals, or report market statistics for agrochemicals.
>
> Porter Stemmer (stop words removed): market strategy carr company agriculture chemic report predict market share chemic report market statist agrochem

Actual Attribute Generation: Attributes generated are merely labels of the classes automatically produced by a classifier on the features that passed the feature selection process.

The next step is to populate the database that results from above.

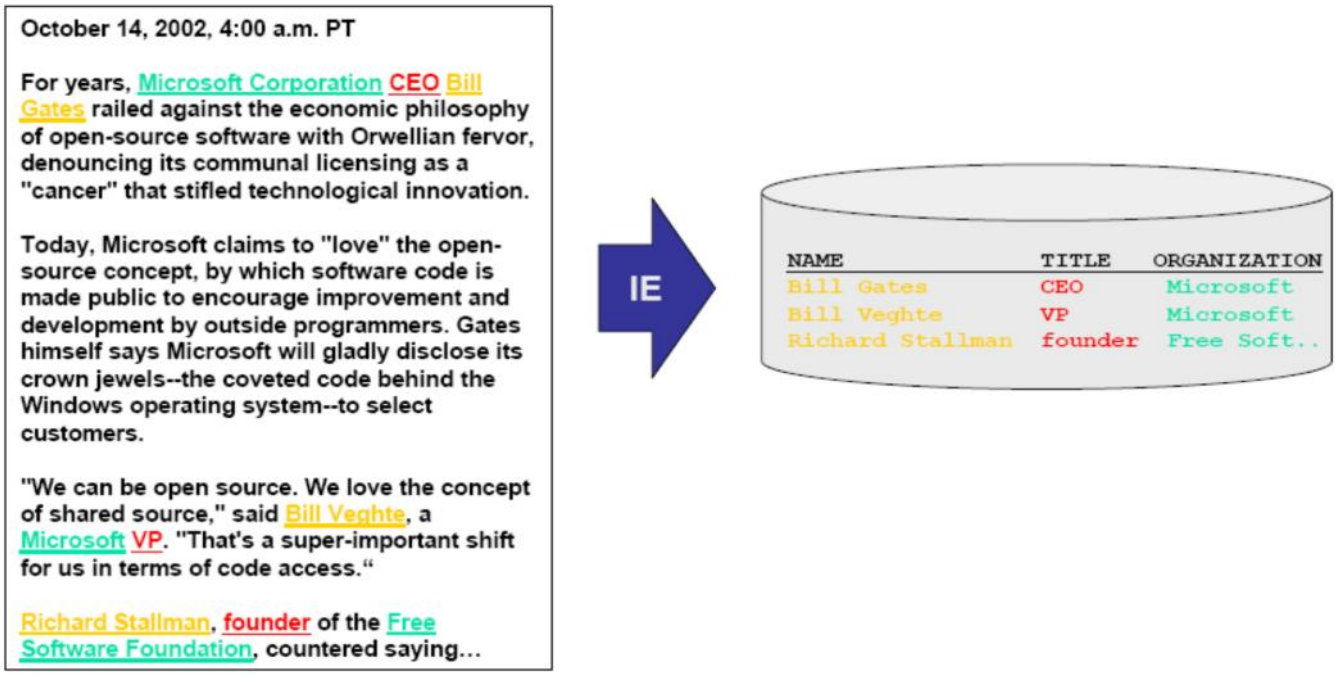

Figure 3: database population [12]

Attribute Selection: Further reduction of dimensionality. Learners have difficulty addressing tasks with high dimensionality. Scarcity of resources and feasibility issues also call for a further cutback of attributes. Add to this that sometimes we have Irrelevant features: Not all features help: e.g., the existence of a noun in a news article is unlikely to help classify it as “politics” or “sport”.

3.5 Data Mining

At this point the Text mining process merges with the traditional Data Mining process which is a purely application-dependent. Classic Data Mining techniques such as clustering, classification and predictive methods are used on the structured database that resulted from the previous stages.
For further reading on data mining techniques, I recommend the book *Data Mining Concepts and Techniques* from his author Jiawei Han et al. the book is a good overview of all the different aspect within the field of data mining.

3.6 Interpretation & Evaluation

If the results are well-suited for the application at hand then we terminate the whole process. Otherwise we iterate by using the results generated as part of the input for one or more earlier stages. There exist many tools for searching the data and visualizing results to provide researchers or customers with insight into the data [11].

4. Application

In recent years, text mining technology has known a fast evolution due of course to its importance. And now it is broadly applied for a wide variety of domains. Applications include: Search engines, email spam filters, Fraud detection, customer relationship management, social media analysis, marketing surveys, analyzing web content, financial services, research and development, national security, etc [1][6].

Study Case: Biomedical text mining

The most active, and I think promising, application area for text mining is in the biosciences. In the past several years, A range of text mining applications has been developed.

One common example is PubGene that combines biomedical text mining with network visualization as an Internet service. GoPubMed is a knowledge-based search engine for biomedical texts.

What makes biomedical domain a fertile ground for text mining applications, is the sheer volume of the biomedical publications, which afford an immense source of text ready to mine. The current text mining task that represent an interest to biomedical community are:

Information extraction: which refers to facts extraction right from free text. It has a major subtasks, named entity recognition, relation extraction and event extraction.

Summarization: that has as main object to summarize the content not necessiraly of one document but several ones.

Question Answering (QA): QA systems provide direct and precise answers to natural language questions. And this systems are regarded as the next generation of search engines.

Finaly, **Literature-Based Discovery:** refers to the task of using scientific literature to discover hidden and previously unknown or neglected relationships between existing knowledge.

Using text in Medical Hypothesis Discovery

The best known example is Don Swanson's work on hypothesizing causes of rare diseases by looking for indirect links in different subsets of the bioscience literature [13].

For example, when investigating causes of migraine headaches, Don Swanson extracted various pieces of evidence from titles of articles in the biomedical literature. Some of these clues can be paraphrased as follows:

- Stress is associated with migraines.
- Stress an lead to loss of magnesium.
- Calcium channel blockers prevent some migraines.
- magnesium is a natural calcium channel blocker.
- Spreading Cortical Depression (SCD)is implicated in some migraines.
- High levels of magnesium inhibit SCD.
- Migraine patients have high platelet aggregability.
- Magnesium can suppress platelet aggregability.

These clues suggest that magnesium deficiency may play a role in some kinds of migraine headache; a hypothesis which did not exist in the literature at the time Swanson found these links.
The hypothesis has to be tested via non-textual means, but the important point is that a new, potentially plausible medical hypothesis was derived from a combination of text fragments and the explorer's medical expertise.
According to [Swanson 1991], subsequent study found support for the magnesium-migraine hypothesis [Ramadan 1989].

5. Conclusion

To conclude, Text mining is a prominent and fast growing technology. It does follow a defined process in order to get high quality information and patterns. Text mining has different applications in different domains. Researchers are steal working on text mining tools and algorithms so to have better results. It may seem a far and hard task to get computers mimic our ability – we humans – to deal with text. But who knows maybe we will be able to see a revolution in the few years to come.

References

[1] Text Mining, Wikipedia, the free encyclopedia.
http://en.wikipedia.org/wiki/Text_Mining [accessed 2015-06-12]

[2] text mining vs data mining.
http://www.researchgate.net/post/Text_mining_and_Data_mining [accessed 2015-06-14]

[3] Gary Miner et al. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications – January 25, 2012

[4] Marti Hearst, What Is Text Mining?
http://people.ischool.berkeley.edu/~hearst/text-mining.html [accessed 2015-06-13]

[5] Marti Hearst, Untangling text data mining
http://people.ischool.berkeley.edu/~hearst/papers/acl99/acl99-tdm.html [accessed 2015-06-21]

[6] StatSoft, Introduction to Text Mining video series
https://www.youtube.com/channel/UCSvbQoORLzvHyPMGBYsZEuQ [accessed 2015-06-1]

[7] J.H. Kroeze et al. Differentiating between data-mining and text-mining terminology, South African Journal of Information Management, Vol.6(4) December 2004

[8] Text Mining (Big Data, Unstructured Data)
http://documents.software.dell.com/Statistics/Textbook/text-mining [accessed 2015-06-22]

[9] text mining (text analytics)
http://searchbusinessanalytics.techtarget.com/definition/text-mining [accessed 2015-06-24]

[10] Grobelnik, M. and Mladenic, D. Text-Mining Tutorial. In the Proceeding of Learning Methods for Text Understanding and Mining, Grenoble, France, January 26 –29, 2004

[11] Text Mining Solution.
http://www.textminingsolutions.co.uk/our-process.html [accessed 2015-06-15]

[12] José María Gómez Hidalgo, Text Mining and Internet Content Filtering
http://www.esi.uem.es/~jmgomez/tutorials/ecmlpkdd02/slides.pdf [accessed 2015-06-13]

[13] S. Dang, P.H. A Review of Text Mining Techniques Associated with Various Application Areas International Journal of Science and Research (IJSR) ISSN (Online), 2013