TEXT PREPROCESSING FOR TEXT MINING USING SIDE INFORMATION

Ms. Nikita P.Katariya[1], Prof. M. S. Chaudhari[2]
[1]Dept. of Computer Science & Engg, P.B.C.E., Nagpur, India, nikitakatariya@yahoo.com
[2]Dept. of Computer Science & Engg, P.B.C.E., Nagpur, India, manojchaudhry2@gmail.com

Abstract

Text mining is the analysis of data contained in natural language text. Text Databases are rapidly growing due to the increasing amount of information available in various electronic forms. User need to access relevant information across multiple documents. In many text mining applications, side-information is available along with the text documents. Side-information may be document origin information, the links in the document, user-access behavior from web logs, or other non-textual attributes which are embedded into the text document. Such attributes may contain a tremendous amount of information for mining purposes. Initial process in Text Mining system is preprocessing. Thus this paper presents different steps involved in text preprocessing.

Keywords: text mining, side information, preprocessing

1. Introduction

Text mining is a new area of computer science which strong connections with natural language processing, data mining, machine learning, information retrieval and knowledge management. Text mining seeks to extract useful information from unstructured textual data through the identification and exploration of interesting patterns. Text mining can help an organization derive potentially valuable business insights from text-based content such as word documents, email and postings on social media streams like Facebook, Twitter and LinkedIn. Mining unstructured data with natural language processing (NLP), statistical modeling and machine learning techniques can be challenging, however, because natural language text is often inconsistent. It contains ambiguities caused by inconsistent syntax and semantics, including slang, language specific to vertical industries and age groups, double entendres and sarcasm.

Text analytics software can help by transposing words and phrases in unstructured data into numerical values which can then be linked with structured data in a database and analyzed with traditional data mining techniques. With an iterative approach, an organization can successfully use text analytics to gain insight into content-specific values such as sentiment, emotion, intensity and relevance. Because text analytics technology is still considered to be an emerging technology, however, results and depth of analysis can vary wildly from vendor to vendor.

2. Process of text mining

Text Mining is the process of extracting interesting information or knowledge or patterns from the unstructured text that are from different sources. As the text is in unstructured form, it is quite difficult to deal with it. Finding chunk of interesting information from the natural language text is the purpose of text mining. The Text Mining Processing shown in Fig. 1

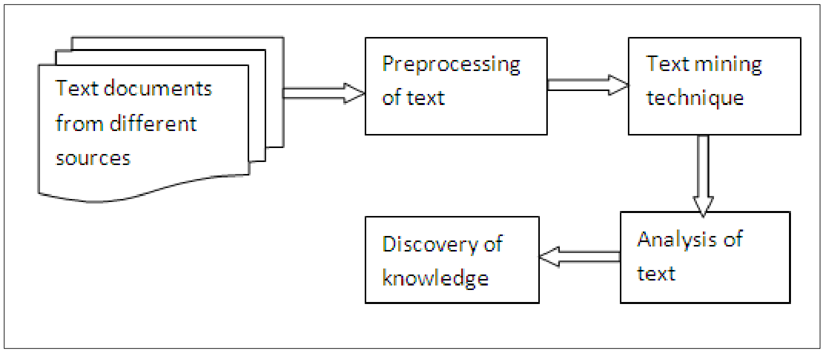

Fig 1. Text mining process

3. Side Information [1]

The problem of text mining arises in the context of many application domains such as the web, social networks, and other digital collections. A tremendous amount of work has been done in recent years on the problem of text collections in the database and information retrieval communities. However, this work is primarily designed for the problem of pure text collection, in the absence of other kinds of attributes. In many application domains, a tremendous amount of side information is also associated along with the documents. This is because text documents typically occur in the context of a variety of applications in which there may be a large amount of other kinds of database attributes or meta-information which may be useful to the mining process. Some examples of such side-information are as follows

3.1. Web logs

In an application in which we track user access behavior of web documents, the user-access behavior may be captured in the form of web logs. For each document, the meta-information may correspond to the browsing behavior of the different users. Such logs can be used to enhance the quality of the mining process in a way which is more meaningful to the user, and also application-sensitive. This is because the logs can often pick up subtle correlations in content, which cannot be picked up by the raw text alone.

3.2. Links present in Text Document

Text documents, which can also be treated as attributes. Such links contain a lot of useful information for mining purposes. As in the previous case, such attributes may often provide insights about the correlations among documents in a way which may not be easily accessible from raw content.

3.3. Meta-data

Many web documents have meta-data associated with them which correspond to different kinds of attributes such as the provenance or other information about the origin of the document. In other cases, data such as ownership, location, or even temporal information may be informative for mining purposes. In a number of network and user-sharing applications, documents may be associated with user-tags, which may also be quite informative.

4. Text preprocessing

Mining from a preprocessed text is easy as compare to natural languages documents. So, preprocessing of documents that are from different sources is an important task during text mining process before applying any text mining technique. As Text documents can be represented as bag of words on which different text mining methods are based. Let Ω be the set of documents & W= {w1, w2, ----wm} be the different words from the document set. In order to reduce the dimensionally of the documents words, special methods such as filtering and stemming are applied. Filtering methods remove those words from the set of all words, which do not provide relevant information; stop word filtering is a standard filtering method. Words like prepositions, articles, conjunctions etc. are removed that contain no informatics as such stemming methods: are used to produce the root from the plural or the verbs. For

e.g. Doing, Done, Did may be represented as Do. After this method is applied, every word is represented by its root word.
Preprocessing text is called tokenization or text normalization. Preprocessing is a procedure which can be divided mainly into five text operations (or transformations):

- Lexical Analysis of the Text
- Stemming
- Elimination of Stopwords
- Index Terms Selection
- Thesauri

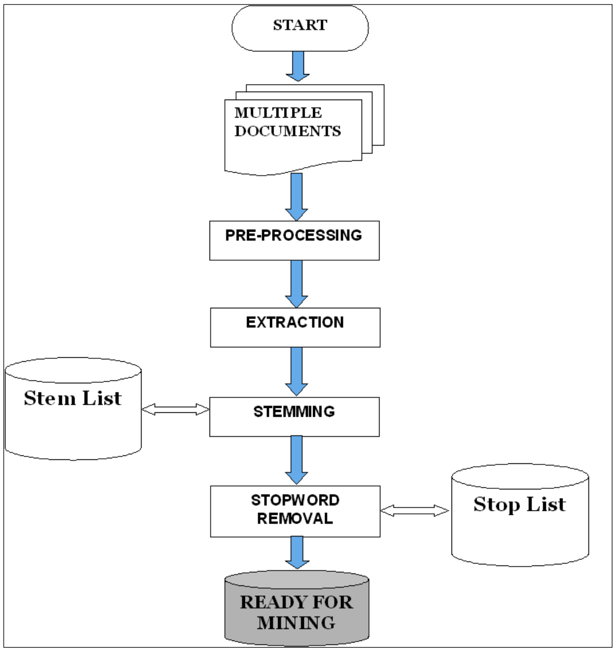

Fig 2. Steps in preprocessing

4.1. Lexical Analysis of the Text

Lexical analysis is the process of converting a stream of characters into a stream of words. Thus, one of the major objectives of the lexical analysis phase is the identification of the words in the text. However, there is more to it than this. For instance, the following four particular cases have to be considered with care: digits, hyphens, punctuation marks, and the case of the letters. Numbers are usually not good index terms because, without a surrounding context, they are inherently vague. The problem is that numbers by themselves are just too vague. Normally, punctuation marks are removed entirely in the process of lexical analysis. The case of letters is usually not important for the identification of index terms. As a result, the lexical analyzer normally converts all the text to either lower or upper case.

4.2. Elimination of Stopwords

In fact, a word which occurs in 80% of the documents in the collection is useless for purposes of retrieval. Such words are frequently referred to as *stopwords* and are normally filtered out as potential index terms. Articles, prepositions, and conjunctions are natural candidates for a list of stopwords. Elimination of stopwords has an additional important benefit. It reduces the size of the indexing structure considerably.

In fact, it is typical to obtain a compression in the size of the indexing structure of 40% or more solely with the elimination of stopwords. A list of 425 stopwords is illustrated. Programs in C for lexical analysis are also provided. Despite these benefit, elimination of stopwords might reduce recall. For instance, consider a user who is looking for documents containing the phrase '*to be or not to be*.' Elimination of stopwords might leave only the term *be* making it almost impossible to properly recognize the documents which contain the phrase specified.

4.3. Stemming

Frequently, the user specifies a word in a query but only a variant of this word is present in a relevant document. This problem can be partially overcome with the substitution of the words by their respective stems. A stem is the portion of a word which is left after the removal of its affixes (i.e., prefixes and suffixes). Stems are thought to be useful for improving retrieval performance because they reduce variants of the same root word to a common concept. Furthermore, stemming has the secondary effect of reducing the size of the indexing structure because the number of distinct index terms is reduced. Many Web search engines do not adopt any stemming algorithm whatsoever. Frakes distinguishes four types of stemming strategies: affix removal, table lookup, successor variety, and n-grams. Table lookup consists simply of looking for the stem of a word in a table. Since such data is not readily available and might require considerable storage space, this type of stemming algorithm might not be practical. Successor variety stemming is based on the determination of morpheme boundaries, uses knowledge from structural linguistics, and more complex than affix removal stemming algorithm.

4.4.Index Terms Selection

Distinct automatic approaches for selecting index terms can be used. A good approach is the identification of noun groups. Since it is common to combine two or three nouns in a single component (e.g., computer science), it makes sense to cluster nouns which appear nearby in the text into a single indexing component (or concept). A noun group is a set of nouns whose syntactic distance in the text does not exceed a predefined threshold.

4.5. Thesauri

The word thesaurus has Greek and Latin origins and is used as a reference to a treasury of words. In its simplest form, this treasury consists of (1) a precompiled list of important words in a given domain of knowledge and (2) for each word in this list, a set of related words. To the adjective cowardly, Roget's thesaurus associates several synonyms which compose a thesaurus class. While Roget's thesaurus is of a generic nature, a thesaurus can be specific to a certain domain of knowledge. According to Foskett, the main purposes of a thesaurus are basically: (a) to provide a standard vocabulary for indexing and searching; (b) to assist users with locating terms for proper query formulation; and (c) to provide a classified hierarchies that allow the broadening and narrowing of the current query request according to the needs of the user. The motivation for building a thesaurus is based on the fundamental idea of using a *controlled vocabulary* for the indexing and searching.

5. Conclusion

Text Mining can be defined as a technique which is used to extract interesting information or knowledge from the text documents which are usually in the unstructured form. This paper presents methods for mining text data with the use of side-information. Many forms of text databases contain a large amount of side-information or meta information, which may be used in order to improve the mining process. Pre-processing activities plays a vital role in the various applications. Therefore it is concluded that the domain specific applications are more proper for text mining. The paper present three important pre-processing techniques namely stop word removal, stemming and indexing.

References

[1] Charu C. Aggarwal, Fellow, IEEE Yuchen Zhao, and Philip S. Yu, Fellow, IEEE, "On the Use of Side Information for Mining Text Data", *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, VOL. 26, NO. 6, JUNE 2014. pp. 1415-1429

[2] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining", *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERIN*G, VOL. 24, NO. 1, JANUARY 2012, pp. 30-44.

[3] C. C. Aggarwal and C.-X. Zhai, "A survey of text classification algorithms," in Mining Text Data. New York, NY, USA: Springer, 2012.

[4] R. Sagayam, S.Srinivasan, S. Roshni, "A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques", *International Journal of Computational Engineering Research (ijceronline.com)*, Vol. 2 Issue. 5, September 2012,pp.1443-1446

[5] C. C. Aggarwal and C.-X. Zhai, "Mining Text Data", New York, NY, USA: Springer, 2012.

[6] S. Ceri et al.," The Information Retrieval Process", *Web Information Retrieval, Data-Centric Systems and Applications*, DOI 10.1007/978-3-642-39314-3_2, © Springer-Verlag Berlin Heidelberg 2013

[7] Zdenek Ceska and Chris Fox, "The Influence of Text Pre-processing on Plagiarism Detection", International Conference RANLP 2009 - Borovets, Bulgaria, pages 55–59

[8] C.Ramasubramanian and R.Ramya, "Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm", *International Journal of Advanced Research in Computer and Communication Engi*neering, Vol. 2, Issue 12, December 2013