Volume 4, Issue 8, August 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Approaches to Text Mining and Incorporating Text Mining Results in Data Mining Project

S. Mahalakshmi
Research Scholar, Bharathiar University, India

M. Suriyakala
Asst. Professor, Dr. Ambedkar Govt. Arts College, Chennai, India

Abstract—The purpose of Text Mining is to process unstructured (textual) information, extract meaningful numeric indices from the text and thus make the information 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, we can analyze words, clusters of words used in documents, etc., or we could analyze documents and determine similarities between them or how they are related to other variables of interest in the data mining contained in the text accessible to the various data mining (statistical and machine learning) algorithms. Information can be extracted project. In the most general terms, text mining will "turn text into numbers" (meaningful indices), which can then be incorporated in other analyses such as predictive data mining projects, the application of unsupervised learning methods (clustering), etc. This paper explains about the approaches to text mining and incorporating text mining results in data mining projects.

Key words—Unstructured, statistical, machine learning, Information, unsupervised learning, predictive data mining.

I. INTRODUCTION

Text mining is a new field that attempts to glean meaningful information from natural language text. It may be loosely characterized as the process of analyzing text to extract information that is useful for particular purposes. Text mining strives to bringit out of the text in a form that is suitable for consumption by computers directly, with no need for a human intermediary.

Research in text mining has been carried out since the mid 80s .According to [6] there are so many changing issues in text mining . Researchers like [8] J. Cowie and Y. Wilks, pointed about Information extraction.

This paper is organized as follows.Section 2 presents Text Mining Process.Section 3 explains about Issues and Considerations for Numericizing Text.Section 4 presents about Transforming Word Frequencies.Section 5 explains about Latent Semantic Indexing via Singular Value Decomposition.Section 6 presents about Incorporating Text Mining Results in Data Mining Projects.Section 7 explains about summary of the paper.

II. TEXT MINING PROCESS

Given a corpus of documents and a user's information need expressed as some sort of query, document retrieval is the task of identifying and returning the most relevant documents. Traditional libraries provide catalogues (whether physical card catalogues or computerized information systems) that allow users to identify documents based on surrogates consisting of metadata—salient features of the document such as author, title, subject classification, subject headings, keywords. Metadata is a kind of highly structured (and therefore actionable) document summary, and successful methodologies have been developed for manually extracting metadata for identifying relevant documents based on it, methodologies that are widely taught in library school .

A. Text Categorization

Many text mining problems involve assessing the similarity between different documents; for example, assigning documents to pre-defined categories and grouping documents into natural clusters. These are standard problems in data mining too, and have been a popular focus for research in text mining, perhaps because the success of different techniques can be evaluated and compared using standard, objective, measures of success.

When classifying a document, no information is used except for the document's content itself. Some tasks constrain documents to a single category, whereas in others each document may have many categories.Sometimes category labeling is probabilistic rather than deterministic, or the objective is to rank the categories by their estimated relevance to a particular document. Sometimes documents are processed one by one, with a given set of classes; alternatively there may be a single class—perhaps a new one that has been added to the set and the task is to determine which documents it contains.

Using words as features perhaps a small number of well-chosen words, or perhaps all words that appear in the document except stop words and word occurrence counts as feature values, a model is built for each category. The documents in that category are positive examples and the remaining documents negative ones. The model predicts whether or not that category is assigned to a new document based on the words in it, and their occurrence counts. Given a new document, each model is applied to determine which categories to assign. Alternatively, the learning method may produce a likelihood of the category being assigned, and if, say, five categories were sought for the new document, those with the highest likelihoods could be chosen. the features are words, documents are represented using the "bag of words" for document retrieval. Sometimes word counts are discarded and the "bag" is treated merely as a set. Bag (or) of words models neglect word order and contextual effects

B. Key Word Based Mining

Key word based mining is used to extract a value associated with a keyword. If a key word is specified ,we first locate the keyword in the pages.Once it is located then the following heuristic rules will be applied to mine the information.

If a String is a hyperlink
Else
if a String is inside the title <h1> .. <h6>
Else
if a string is in the item <li> /<ui>
Else
if a string is in the table <td> ,<th>
Else
if a string is in between //<u></u>
Else
if the string appears after the verbs is ,are
Then the String is a keyword

Key-phrase extraction works as follows. Given a document, rudimentary lexical techniques based on punctuation and common words are used to extract a set of candidate phrases. Then, features are computed for each phrase, like how often it appears in the document (normalized by how often that phrase appears in other documents in the corpus), how often it has been used as a key-phrase in other training documents, whether it occurs in the title, abstract, or section headings, whether it occurs in the title of papers cited in the reference list, and so on. The training data is used to form a model that takes these features and predicts whether or not a candidate phrase will actually appear as a key-phrase—this information is known for the training documents. Then the model is applied to extract likely key-phrases from new documents. Such models have been built and used to assign key-phrases to technical papers; simple machine learning schemes (e.g. Naïve Bayes) seem adequate for this task.

C. Extracting Structured Information

An important form of text mining takes the form of a search for structured data inside documents. Documents are full of structured information: phone numbers, fax numbers, street addresses, email addresses, email signatures, abstracts, tables of contents, lists of references, tables, figures, captions, meeting announcements, Web addresses, and more. In addition, there are countless domain-specific structures, such as ISBN numbers, stock symbols, chemical structures, and mathematical equations. Many short documents describe a particular kind of object or event, and in this case elementary structures are combined into a higher-level composite that represent the document's entire content. In constrained situations, the composite structure can be represented as a "template" with slots that are filled by individual pieces of structured information. From a large set of documents describing similar objects or events it may even be possible to infer rules that represent particular patterns of slot-fillers.

D. Entity Extraction

Many practical tasks involve identifying linguistic constructions that stand for objects or "entities" in the world. Often consisting of more than one word, these terms act as single vocabulary items, and many document processing tasks can be significantly improved if they are identified as such. They can aid searching, interlinking and cross-referencing between documents, the construction of browsing indexes, and can comprise machine-processable "metadata" which, for certain operations, act as a surrogate for the document contents.

Examples of such entities are
Name of people, places, things
E-mail address, URLs
Dates, numbers, money
Acronyms etc…..

E. Pattern Based Matching

A central area of library science is devoted to the creation and use of standard names for authors and other bibliographic entities (called "authority control"). In most applications, novel names appear. Sometimes these are composed of parts that have been encountered before, say John and Smith, but not in that particular combination. Others are recognizable by their capitalization and punctuation pattern (e.g. Randall B. Caldwell). Still others particularly certain foreign names will be recognizable because of peculiar language statistics (e.g. Kung-Kui Lau). Others will not be recognizable except by capitalization, which is an unreliable guide particularly when only one name is present. Names that begin a sentence cannot be distinguished on this basis from other words. It is not always completely clear what to "begin a sentence" means: in some typographic conventions, itemized points have initial capitals but no terminating punctuation. Of course, words that are not names are sometimes capitalized (e.g. important words in titles and, in German, all nouns). And a small minority of names are conventionally written unpunctuated and in lower case (e.g. some English names starting with ff, the poet e e cummins the singer k d lang). Full personal name recognition conventions are surprisingly complex, involving baronial prefixes in different languages (e.g. von, van, d e), suffixes (Snr, Jnr), and titles (Mr, Ms, Rep., Prof., General). It is generally impossible to distinguish personal names from other kinds of names in the absence of context or domain knowledge. Consider places like Berkeley, Lincoln, Washington; companies like du Pont, Ford, even General Motors; product names like Mr Whippy and Dr Pepper; book titles like David Copperfield or Moby Dick. Names of organizations present special difficulties because they can contain linguistic constructs, as in The Food and Drug Administration (contrast Lincoln and Washington, which conjoins two separate names) or the League of Nations (contrast General Motors of Detroit, which qualifies one name with a different one). Some artificial entities like e-mail addresses and URLs are easy to recognize because they are specially designed for machine processing. They can be unambiguously detected by a simple grammar, usually encoded in a regular expression, for the appropriate pattern. Of course, this is exceptional: these items are not part of "natural" language.

Dates include standard textual forms for absolute and relative dates. Numbers include both absolute numbers andpercentages, and can be written in numerals or spelled out as words. Sums of money can be expressed in various currencies.

F. Acronym Identification

The dictionary definition of "acronym" is

A word formed from the first (or first few) letters of a series of words, as radar, from radio detecting and ranging.

Acronyms are often defined by following (or preceding) their first use with a textual explanation—as in this example. Heuristics can be developed to detect situations where a word is spelled out by the initial letters of an accompanying phrase. Three simplifying assumptions that vastly reduce the computational complexity of the task while sacrificing the ability to detect just a few acronyms are to consider

If a word is made up of three or more letters
and
only the first letter of each word for inclusion in the acronym,
and
written either fully- or mostly-capitalized.
Then that word is **acronym.**

However, the vast majority of acronyms that pervade today's technical, business, and political literature satisfy these assumptions, and are relatively easy to detect.

III. ISSUES AND CONSIDERATIONS FOR NUMERICIZING TEXT

To reiterate, text mining can be summarized as a process of "numericizing" text. At the simplest level, all words found in the input documents will be indexed and counted in order to compute a table of documents and words, i.e., a matrix of frequencies that enumerates the number of times that each word occurs in each document. This basic process can be further refined to exclude certain common words such as "the" and "a" (stop word lists) and to combine different grammatical forms of the same words such as "traveling," "traveled," "travel," etc. (stemming). However, once a table of (unique) words (terms) by documents has been derived, all standard statistical and data mining techniques can be applied to derive dimensions or clusters of words or documents, or to identify "important" words or terms that best predict another outcome variable of interest.

A. Excluding Certain Characters, Short Words, Numbers, Etc.

Excluding numbers, certain characters, or sequences of characters, or words that are shorter or longer than a certain number of letters can be done before the indexing of the input documents starts. You may also want to exclude "rare words," defined as those that only occur in a small percentage of the processed documents.

B. Include Lists, Exclude Lists (Stop-Words)

Specific list of words to be indexed can be defined; this is useful when you want to search explicitly for particular words, and classify the input documents based on the frequencies with which those words occur. Also, "stop-words," i.e.,

terms that are to be excluded from the indexing can be defined. Typically, a default list of English stop words includes "the", "a", "of", "since," etc, i.e., words that are used in the respective language very frequently, but communicate very little unique information about the contents of the document.

C. Synonyms and Phrases

Synonyms, such as "sick" or "ill", or words that are used in particular phrases where they denote unique meaning can be combined for indexing. For example, "Microsoft Windows" might be such a phrase, which is a specific reference to the computer operating system, but has nothing to do with the common use of the term "Windows" as it might, for example, be used in descriptions of home improvement projects.

D. Stemming Algorithms

An important pre-processing step before indexing of input documents begins is the stemming of words. The term "stemming" refers to the reduction of words to their roots so that, for example, different grammatical forms or declinations of verbs are identified and indexed (counted) as the same word. For example, stemming will ensure that both "traveling" and "traveled" will be recognized by the text mining program as the same word.

IV. TRANSFORMING WORD FREQUENCIES

Once the input documents have been indexed and the initial word frequencies (by document) computed, a number of additional transformations can be performed to summarize and aggregate the information that was extracted.

A. Log-Frequencies.

First, various transformations of the frequency counts can be performed. The raw word or term frequencies generally reflect on how salient or important a word is in each document. Specifically, words that occur with greater frequency in a document are better descriptors of the contents of that document. However, it is not reasonable to assume that the word counts themselves are proportional to their importance as descriptors of the documents. For example, if a word occurs 1 time in document A, but 3 times in document B, then it is not necessarily reasonable to conclude that this word is 3 times as important a descriptor of document B as compared to document A. Thus, a common transformation of the raw word frequency counts (wf) is to compute:

f(wf) = 1+ log(wf), for wf > 0

This transformation will "dampen" the raw frequencies and how they will affect the results of subsequent computations.

B. Binary Frequencies.

Likewise, an even simpler transformation can be used that enumerates whether a term is used in a document; i.e.:

f(wf) = 1, for wf > 0

The resulting documents-by-words matrix will contain only 1s and 0s to indicate the presence or absence of the respective words. Again, this transformation will dampen the effect of the raw frequency counts on subsequent computations and analyses.

C. Inverse Document Frequencies.

Another issue that you may want to consider more carefully and reflect in the indices used in further analyses are the relative document frequencies (df) of different words. For example, a term such as "guess" may occur frequently in all documents, while another term such as "software" may only occur in a few. The reason is that we might make "guesses" in various contexts, regardless of the specific topic, while "software" is a more semantically focused term that is only likely to occur in documents that deal with computer software. A common and very useful transformation that reflects both the specificity of words (document frequencies) as well as the overall frequencies of their occurrences (word frequencies) is the so-called inverse document frequency (for the i'th word and j'th document):

$$idf(i,j) = \begin{cases} 0 & \text{if } wf_{i,j} = 0 \\ (1+\log(wf_{i,j}))\log\dfrac{N}{df_i} & \text{if } wf_{i,j} \geq 1 \end{cases}$$

In this formula , N is the total number of documents, and dfi is the document frequency for the i'th word (the number of documents that include this word). Hence, it can be seen that this formula includes both the dampening of the simple word frequencies via the log function (described above), and also includes a weighting factor that evaluates to 0 if the word occurs in all documents (log(N/N=1)=0), and to the maximum value when a word only occurs in a single document (log(N/1)=log(N)). It can easily be seen how this transformation will create indices that both reflect the relative frequencies of occurrences of words, as well as their semantic specificities over the documents included in the analysis.

V. LATENT SEMANTIC INDEXING VIA SINGULAR VALUE DECOMPOSITION

As described above, the most basic result of the initial indexing of words found in the input documents is a frequency table with simple counts, i.e., the number of times that different words occur in each input document. Usually, we would transform those raw counts to indices that better reflect the (relative) "importance" of words and/or their semantic specificity in the context of the set of input documents (see the discussion of inverse document frequencies, above).

A common analytic tool for interpreting the "meaning" or "semantic space" described by the words that were extracted, and hence by the documents that were analyzed, is to create a mapping of the word and documents into a common space, computed from the word frequencies or transformed word frequencies (e.g., inverse document frequencies). In general, here is how it works:

Suppose we indexed a collection of customer reviews of their new automobiles (e.g., for different makes and models). We may find that every time a review includes the word "gas-mileage," it also includes the term "economy." Further, when reports include the word "reliability" they also include the term "defects" (e.g., make reference to "no defects"). However, there is no consistent pattern regarding the use of the terms "economy" and "reliability," i.e., some documents include either one or both. In other words, these four words "gas-mileage" and "economy," and "reliability" and "defects," describe two independent dimensions - the first having to do with the overall operating cost of the vehicle, the other with the quality and workmanship. The idea of latent semantic indexing is to identify such underlying dimensions (of "meaning"), into which the words and documents can be mapped. As a result, we may identify the underlying (latent) themes described or discussed in the input documents, and also identify the documents that mostly deal with economy, reliability, or both. Hence, we want to map the extracted words or terms and input documents into a common latent semantic space.

A. Singular Value Decomposition.

The use of singular value decomposition in order to extract a common space for the variables and cases (observations) is used in various statistical techniques, most notably in Correspondence Analysis. The technique is also closely related to Principal Components Analysis and Factor Analysis. In general, the purpose of this technique is to reduce the overall dimensionality of the input matrix (number of input documents by number of extracted words) to a lower-dimensional space, where each consecutive dimension represents the largest degree of variability (between words and documents) possible. Ideally, you might identify the two or three most salient dimensions, accounting for most of the variability (differences) between the words and documents and, hence, identify the latent semantic space that organizes the words and documents in the analysis. In some way, once such dimensions can be identified, you have extracted the underlying "meaning" of what is contained (discussed, described) in the documents.

VI. INCORPORATING TEXT MINING RESULTS IN DATA MINING PROJECTS

After significant (e.g., frequent) words have been extracted from a set of input documents, and/or after singular value decomposition has been applied to extract salient semantic dimensions, typically the next and most important step is to use the extracted information in a data mining project.

A. Graphics (Visual Data Mining Methods).

Depending on the purpose of the analyses, in some instances the extraction of semantic dimensions alone can be a useful outcome if it clarifies the underlying structure of what is contained in the input documents. For example, a study of new car owners' comments about their vehicles may uncover the salient dimensions in the minds of those drivers when they think about or consider their automobile (or how they "feel" about it). For marketing research purposes, that in itself can be a useful and significant result. You can use the graphics (e.g., 2D scatterplots or 3D scatterplots) to help you visualize and identify the semantic space extracted from the input documents.

B. Clustering And Factoring

We can use cluster analysis methods to identify groups of documents (e.g., vehicle owners who described their new cars), to identify groups of similar input texts. This type of analysis also could be extremely useful in the context of market research studies, for example of new car owners. You can also use Factor Analysis and Principal Components and Classification Analysis (to factor analyze words or documents).

C. Predictive Data Mining

Another possibility is to use the raw or transformed word counts as predictor variables in predictive data mining projects.

VII. CONCLUSION

This paper explains about the complete detail of text mining process ,Transforming word frequencies,single value decomposition and incorporating text mining results in data mining.The purpose of this section is to give overall view of text mining and it's process which will be more helpful tor scholars and students.

REFERENCES

[1] Witten, I.H. and Frank, E. (2000) Data mining: "Practical machine learning tools and techniques with Java implementations", Morgan Kaufmann, San Francisco, CA.

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

[3] Nahm, U.Y. and Mooney, R.J. (2002) "Text mining with information extraction" ,Proc AAAI-2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases. Stanford, CA.

[4] Nahm, U.Y. and Mooney, R.J. (2002) "Text mining with information extraction.", Proc AAAI-2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases. Stanford, CA.

[5] Freitag, D. (2002) "Machine learning for information extraction in informal domains." Machine Learning, Vol. 39, No. 2/3, pp. 169-202.

[6] Shaidah Jusoh and Hejab Alfare, "Techniques,Applications and Challenging Issue in Text Mining", International Journal of Compuet Science Issues ,Vol9 ,Issue 6,No 2,Nov 2012

[7] Franklin, D. (2002) "New software instantly connects key bits of data that once eluded teams of researchers." Time, December 23.

[8] J.Cowie and Y.Wilks, "Information extraction", New York,2000.

