International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-7,October 2015
ISSN: 2395-3470
www.ijseas.com

Text Mining and Information Management

Murat YAZICI, M.Sc.

In recent years, we have seen a tremendous growth in the amount of available textual information, both on the World Wide Web and in institutional document repositories. Increasing the amount of available textual information has led to the need of obtaining meaningful information. In this context, text mining has become very prevalent because vast amounts of textual information can be accessed and analyzed. The development of new technologies about solution of some problems such as topic detection, tracking, and trending, where a machine automatically identifies topics in a text, has enabled wide application in the future. Some of text mining tasks are text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling etc. Indeed, there are numerous applications of text mining, including cuttingedge research in the analysis and classification of news reports, email and spam filtering, hierarchical topic extraction from web pages, automated ontology extraction and management, and competitive intelligence. The focus of this article is to explain how text mining can be used by information management professionals, the text mining process and machine learning, finding patterns in a text with kernel methods, statistical learning methods and their use in text mining, and classification and clustering in text mining.

Information management professionals are naturally associated with the text mining because of their existing skill sets. They are knowledgeable about available products and information-retrieval techniques. Expert information management professionals have analytical and creativity skills, have developed the ability to adapt and try different approaches to problems. Specific roles for information management professionals in text mining projects can include the following:

- *Facilitating of conversations between internal teams and vendors*. Scientists and vendors may speak very different languages. Someone needs to negotiate and articulate the various needs and desires and get everyone in agreement.

- *Place the text-mining tool in context of other information sources.* It can be difficult to understand what a text-mining tool offers versus a familiar search tool. Customers need help understanding what they should expect and how it will be different from other search outputs.

- *Advise vendors and customers on source selection.* From the customer viewpoint, it can be difficult to understand why certain commercial database information cannot be included in the text-mining effort. Licensing and copyright issues are best addressed by an information

professional. Vendors likely won't have the expertise in sources for your specific area of interest. It can make a big difference in the output based on the sources used for the input.

- *Advise on search strategies to retrieve the content set.* Even if the vendor is going to use a content source that is familiar to all, such as PubMed, the search strategies used to retrieve the corpus are of critical importance. We have either required documentation of the exact search strategies used by the vendor or we have provided the search strategies to be used.

- *Consult on appropriate taxonomies and ontologies.* Again, the vendor may not be familiar with taxonomies specific to your area of interest. The categorization and the organization of the text can be useful (or not) in manipulating results: Be sure the taxonomies will be useful for your data. In some very specialized areas of focus, it may be necessary to create and provide the vendor with some or all of a taxonomy. In one case, although we were using the MeSH taxonomy, we built out a specific area of interest in much greater detail, as that was the focus of our research.

- *Help customers evaluate and manipulate results.* Most scientists are already so overloaded with job responsibilities that they don't have the time or the inclination to invest in learning to use a new tool. It is important for the information professional to facilitate the usability and to help them gain value from the output. It may be that the information professional will have to act as an intermediary—using the tool and producing output to which the scientist can then react.

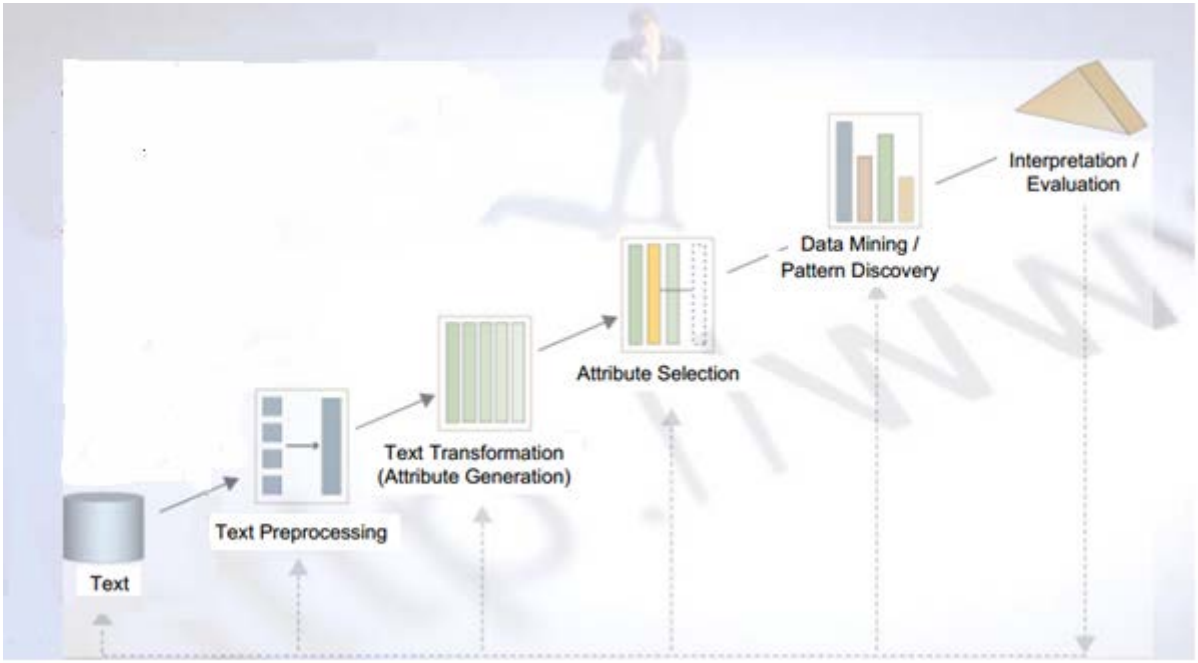

Figure 1. Text Mining Process [1]

Text mining is a process comprising several steps as shown figure 1. In the first step, what documents suit the application is determined in large volume of textual data. Document clustering methods are used to solve this problem. These methods are unsupervised learning methods. Most popular document clustering methods are k-means clustering and agglomerative hierarchical clustering. In the second step, text is cleared: e.g. remove from ads from web pages, normalize text converted from binary formats, deal with tables, figures and formulas etc. Then, the process of marking up the words in a text with their corresponding parts of speech begins. There are two approaches on marking up the words. One of them is rule-based approach. The other one is statistically-based approach. The rule-based approach depends on grammatical rules. The statistically-based approach relies on different word order probabilities and needs a manually target corpus for machine learning. Then, in which sense a word having a number of distinct senses is used in a given sentence is determined. Finally, semantic structures are determined. There are two approaches for determining semantic structures. One of them is full parsing: produces a parse tree for a sentence. The other one is chunking with partial parsing: produces syntactic constructs like Noun Phrases and Verb Groups for a sentence. Producing a full parse tree often fails due to grammatical inaccuracies, novel words, bad tokenization, wrong sentence splits, errors in POS tagging etc. Hence, chunking and partial parsing is more commonly used. In the third step, the words (features) is determined for text representation. There are two main approaches for document representation: Bag of words and Vector space. These approaches aims to determine which features best characterize a document. In the fourth step, features' dimension is reduced. For this, irrelevant attributes is removed. In the fifth step, Text Mining process merges with the traditional Data Mining process. Classic Data Mining techniques such as clustering, classification, desicion trees, regression analysis, neural networks, nearest neighbor are used on the structured database that resulted from the previous stages. This is a purely application-dependent stage. In the final step, if the results are not satisfactory, they are used as part of the input for one or more earlier stages. Text mining takes advantage of machine learning especially in determining features, reduce dimensionality and remove irrelevant attributes. Some of machine learning algorithms are desicion tree learning, association rule learning, artificial neural learning, inductive logic programming, support vector machines, bayesian networks, genetic algorithms, and sparse dictionary learning.

One of finding pattern methods in a text is kernel methods. The kernel approach offers a very general framework for performing pattern analysis on many types of data and it can be used in a wide variety of tasks and application areas. The kernel technique also enables us to use feature spaces whose dimensionality is more than polynomial in the relevant parameters of the systems even though the computational cost of the pattern analysis algorithm remains polynomial. Kernel methods owe their name to the use of kernel functions, which enable them to operate in a high-dimensional, implicit feature space without ever computing the coordinates of the data in that space, but rather by simply computing the inner products between the images of all pairs of data in the feature space. This operation is often computationally cheaper than the explicit computation of the coordinates. This approach is called the kernel trick. Kernel functions have been introduced for sequence data, graphs, text, images, as well as vectors. Kernel Methods and Pattern

Analysis can be considered two of the most important topics in machine learning in the last few years. Their adaptability and modularity has produced a variety of kernels and algorithms in a number of different topic areas. In particular, well known algorithms have been modified into a kernel version. Some of kernel methods are Fisher Kernel, Graph Kernel, Polynomial Kernel, Radial Basis Function (RBF) Kernel, and String Kernel.

Text mining takes advantage of statistical learning methods especially in machine learning stage. Some of them are support vector machines, bayesian networks, desicion trees, and association rules. Support Vector Machines (SVMs) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. Bayesian networks (BNs), also known as belief networks belong to the family of probabilistic graphical models (GMs). These graphical structures are used to represent knowledge about an uncertain domain. In particular, each node in the graph represents a random variable, while the edges between the nodes represent probabilistic dependencies among the corresponding random variables. These conditional dependencies in the graph are often estimated by using known statistical and computational methods. Decision tree learning is one of the most successful techniques for supervised classification learning. A decision tree is a tree in which each internal (non-leaf) node is labeled with an input feature. The arcs coming from a node labeled with a feature are labeled with each of the possible values of the feature. Each leaf of the tree is labeled with a class or a probability distribution over the classes. Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness.

The classification of text documents according to categories or topics, is an important component of any text processing system. There is a large body of work which makes use of content – the words appearing in the documents, the structure of the documents – and external sources to build accurate document classifiers. In addition, there is a growing body of literature on methods which attempt to make use of the link structure among the documents in order to improve document classification performance. Text documents can be connected together in a variety of ways. The most common link structure is the citation graph: e.g., papers cite other papers and webpages link to other webpages. But links among papers can be constructed from other relationships such as co-author, co-citation, appearance at a conference venue, and others. All of these can be combined together to create a interlinked collection of text documents. In these cases, we are often not interested in determining the topic of just a single document, but we have a collection of unlabeled (or partially labeled) documents, and we want to correctly infer values for all of the missing labels. Collective classification methods range from simple local influence propagation algorithms to more complex global optimization algorithms. At their heart, they try to model the combined correlations among labels of neighboring documents. Some models assume that neighboring labels are likely to be the same or similar (homophily, or autocorrelation), while others are capable of learning more complex dependencies. Some of algorithms used

in classification are Iterative Classification, Gibbs Sampling, Loopy Belief Propagation, and Mean Field Relaxation Labeling.

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. An advantage 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. Partitioning algorithms are widely used in the database literature in order to efficiently create clusters of objects. The most widely used distance-based partitioning algorithm is the k-means clustering algorithm. The k-means clustering algorithm uses a set of k representatives around which the clusters are built. The simplest form of the k-means approach is to start off with a set of k seeds from the original corpus, and assign documents to these seeds on the basis of closest similarity. In the next iteration, the centroid of the assigned points to each seed is used to replace the seed in the last iteration. In other words, the new seed is defined, so that it is a better central point for this cluster. This approach is continued until convergence. One of the advantages of the k-means method is that it requires an extremely small number of iterations in order to converge.

Text mining is one of the most powerful techniques about obtaining meaningful information in a text data. With the development of new technologies on data storage and data access, it seems to find more application areas in the future. In this article, I tried to explain text mining and its process, kernel methods, statistical learning methods, clustering, classification, and its application to information management. For further reading on the topic text mining, I recommend reading Srivastava and Sahami's book called *Text mining classification, clustering, and applications* to information management professionals. It has a profound knowledge about text mining. I aslo recommend examining Duda, Hart, and Stork's book called *Pattern classification*. It is one of the main books about finding patterns.

References

[1] Ben-Gal I., *Bayesian Networks*, in Ruggeri F., Faltin F., and Kenett R., *Encyclopedia of Statistics in Quality & Reliability*, Wiley & Sons 2007.

[2] Chiwara M., Al-Ayyoub M., Hossain M. S., and Gupta R., *Text mining*, A presentation, 2014. http://www3.cs.stonybrook.edu/~cse634/presentations/TextMining.pdf

[3] Duda R. O., Hart P. E., and Stork D. G., *Pattern classification*, John Wiley & Sons, Second Edition 2012.

[4] Hearst M., *What is text mining?*, Essay written, Oct. 17, 2003. http://www.sims.berkeley.edu/~hearst/text-mining.html

[5] Hofmann T., Scholkopf B., and Smola A., *Kernel methods in machine learning*, The Annals of Statistics 2008, Vol. 36 (3) p: 1171-1220.

[6] Lavengood K. A., Kiser P., *Information professionals in the text mine*, Online Magazine 2007, Vol.31 (3) p: 16.

[7] Liritano S., Ruffolo M., *Managing the knowledge contained in electronic documents: a clustering method for text mining*, IEEE 2001, 454-458.

[8] Srivastava A., Sahami M., *Text mining classification, clustering, and applications*, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series 2009.
[9] Taylor J. S., Cristianini N., Kernel methods for pattern analysis, Cambridge University Press 2004.
[10] Zhao Y., *R and data mining: Examples and case studies*, Elsevier 2012.