International Journal of Soft Computing and Engineering (IJSCE)
ISSN: 2231-2307, Volume-2, Issue-6, January 2013

A Detailed Study on Text Mining Techniques

Rashmi Agrawal, Mridula Batra

Abstract - Text Mining is an important step of Knowledge Discovery process. It is used to extract hidden information from not-structured or semi-structured data. This aspect is fundamental because most of the Web information is semi-structured due to the nested structure of HTML code, is linked and is redundant. Web Text Mining helps whole knowledge mining process in mining, extraction and integration of useful data, information and knowledge from Web page contents. Web Text Mining process able to discover knowledge in a distributed and heterogeneous multi-organization environment. In this paper, our basic focus is to study the concept of Text Mining and various techniques. Here, we are able to determine how to mine the Plain as well as Structured Text. It also describes the major ways in which text is mined when the input is plain natural language, rather than partially-structured Web documents.

Keywords: Plain, Structured, Text Mining, Web Documents.

I. INTRODUCTION TO TEXT MINING

The Text mining processes unstructured information, extracts meaningful numeric indices from the text, and makes the information contained in the text accessible to the various data mining (statistical and machine learning) algorithms. Information can be extracted from the summarized words of the documents, so the words can be analyzed and also the similarities between words and documents can be determined or how they are related to other variables in the data-mining project. Basically, text mining converts text into numbers which can then be included in other analyses such as predictive data mining projects, clustering etc. Text mining is also known as text data mining, which refers the process of deriving high-quality information from text. High-quality information is derived through the statistical pattern learning. Text mining includes the process of structuring the input text like parsing and other successive insertion into a database. TM derives patterns within the structured data, evaluates them and finally produces the output. Text mining takes account of text categorization, text clustering, sentiment analysis, document summarization, and entity relation modeling. Text mining is a process that employs a set of algorithms for converting unstructured text into structured data objects and the quantitative methods used to analyze these data objects.

Manuscript received on January, 2013.

Rashmi Agrawal, Department of Computer Applications, Manav Rachna International University,Faridabad,India.

Mridula Batra, Department of Computer Applications, Manav Rachna International University,Faridabad,India.

II. APPLICATIONS OF TEXT MINING

There are various applications of Text mining like automatic processing of messages and emails. For example, it is possible to "filter" out automatically "junk email" based on certain terms, such messages can automatically be discarded. Such automatic systems for classifying electronic messages can also be useful in applications where messages need to be routed automatically to the most appropriate department. Another application is Analyzing warranty or insurance claims, diagnostic interviews**.** In some business domains, the majority of information is collected in textual form. For example, warranty claims or initial medical (patient) interviews can be summarized in brief narratives, or when you take your automobile to a service station for repairs, typically, the attendant will write some notes about the problems that you report and what you believe needs to be fixed. Increasingly, those notes are collected electronically, so those types of narratives are readily available for input into text mining algorithms.

Analyzing open-ended survey responses. Survey questionnaires typically contain two broad types of questions: open-ended and closed-ended. Closed-ended questions present a discrete set of responses from which to choose. Such types of responses are easily quantified and analyzed while open-ended questions allow the respondent to answer a question in his own words. Such types of unstructured responses often provide richer and more valued information than closed-ended questions and are an important source of insight since they can generate information that was not anticipated. Despite their added value, researchers often prefer to avoid including open-ended questions in their surveys because of the tedious task of reading and coding responses, a time-consuming and expensive task especially when one has more than a few hundred written responses.

III. VARIOUS TERMINOLOGIES OF TEXT MINING

A. Text Mining Vs. Data Mining

In Text Mining, patterns are extracted from natural language text but in Data Mining patters are extracted from databases.

B. Text Mining Vs. Web Mining

In Text Mining, the input is free unstructured text, but in Web Mining web sources are structured.

IV. WHY TEXT MINING?

Text mining is data mining which is applied to textual data. Text is "unstructured, vague and difficult to deal with but it is the most common method for formal exchange of information. Whereas data mining belongs in the corporate world because that's where most databases are, text mining promises to move machine learning technology out of the companies and into the

home" as an increasingly necessary Internet adjunct i.e., as "web data mining" provide a current review of web data extraction tools.

Text mining is nothing but "nontraditional information retrieval strategies." The goal of these strategies is to reduce the effort required of users to obtain useful information from large computerized text data sources. Traditional information retrieval strategies simultaneously retrieve both less and much information from the text. The nontraditional strategies represent a useful system that must go beyond simple retrieval.

A. How does Mining Work

- Traditional keyword search retrieves documents containing pre-defined keywords. Text mining extracts precise information based on much more than just keywords, such as entities or concepts, relationships, phrases, sentences and even numerical information in context.
- Text mining software tools often use computational algorithms based on Natural Language Processing, or NLP, to enable a computer to read and analyze textual information. It interprets the meaning of the text and identifies extracts, synthesizes and analyzes relevant facts and relationships that directly answer the question.
- Text can be mined in a systematic, comprehensive and reproducible way, and business critical information can be captured automatically.
- Powerful NLP-based queries can be run in real time across millions of documents. These can be pre-written queries.
- Using wildcards, one can ask questions without even having to know the keywords for which he is looking for and still get back high quality, structured results.
- One can switch in any vocabularies or thesauri to take advantage of terminology used in its own specific domain.

V. METHODS OF MINING TEXT

A. Mining Plain Text

This section describes the major ways in which text is mined when the input is plain natural language, rather than partially-structured Web documents. We begin with problems that involve extracting information for human consumption. Here are the various techniques which mine the plain text like text summarization, document retrieval, Information retrieval, Assessing document similarity and Text categorization.

A1. Text summarization

A text summarizer produces a compressed representation of its input, which specifies human Consumption. It also contains individual documents or groups of documents. Text Compression is a related area but the output of text summarization is specific to be human-readable. The output of text compression algorithms is definitely not human-readable and it is also not actionable, It only supports decompression, that is, automatic reconstruction of the original text. Summarization differs from many other forms of text mining in that there are people, namely professional abstractors, who are skilled in the art of producing summaries and carry out the task as part of their professional life.

A2. Document Retrieval

Document retrieval is the task of identifying and returning the most relevant documents. Traditional libraries provide catalogues that allow users to identify documents based on resources which consist of metadata. Metadata is a highly structured document for summary, and successful methodologies have been developed for manually extracting metadata and for identifying relevant documents based on it, methodologies that are widely taught in library school. Automatic extraction of metadata (e.g. subjects, language, author, key-phrases) is a prime application of text mining techniques. The idea is to index every individual word in the document collection. It specifies many effective and popular document retrieval techniques.

A3. Information retrieval

Information retrieval is considered as an extension to document retrieval where the documents that are returned are processed to condense or extract the particular information sought by the user. Thus document retrieval is followed by a text summarization stage that focuses on the query posed by the user, or an information extraction stage. The modularity of documents may be adjusted so that each individual subsection or paragraph comprises a unit in its own right, in an attempt to focus results on individual nuggets of information rather than lengthy documents.

A4. Assessing document similarity

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 the basic problems in data mining too, and have been a focus for research in text mining, perhaps because the success of different techniques can be evaluated and compared using standard, objective, measures of success.

A5. Text categorization

Text categorization is the assignment of natural language documents to predefined categories according to their content. The set of categories is often called a “controlled vocabulary.” Document categorization is a long-standing traditional technique for information retrieval in libraries, where subjects rival authors as the predominant gateway to library contents—although they are far harder to assign objectively than authorship. Automatic text categorization has many practical applications, including indexing for document retrieval, automatically extracting metadata, word sense disambiguation by detecting the topics a document covers, and organizing and maintaining large catalogues of Web resources. As in other areas of text mining, until the 1990s text categorization was dominated by *ad hoc* techniques of “knowledge engineering” that sought to elicit categorization rules from human experts and code them into a system that could apply them automatically to new documents. Since then—and particularly in the research community—the dominant approach has been to use techniques of machine learning to infer categories automatically from a training set of pre-classified documents. Indeed, text categorization is a hot topic in machine learning today. The pre-defined categories are symbolic labels with no

additional semantics. 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. Many machine learning techniques have been used for text categorization.

B. Mining structured text

Much of the text that we have on the Internet contains explicit structural markup and differs from traditional plain text. Some markup is internal and indicates document structure or format; some is external and gives explicit hypertext links between documents. These information sources give additional benefits for mining Web documents. Both sources of information are extremely noisy: they involve arbitrary and unpredictable choices by individual page designers. However, these disadvantages are offset by the total amount of data that is available, which is relatively unbiased because it is aggregated over many different information providers. Thus "Web mining" is emerging as a new subfield, similar to text mining but taking advantage of the extra information available in Web documents, particularly hyperlinks—and even capitalizing on the existence of topic directories in the Web itself to improve results. We briefly review three techniques for mining structured text. The first, wrapper induction, uses internal markup information to increase the effectiveness of text mining in marked-up documents. The remaining two, document clustering and determining the "authority" of Web documents, capitalize on the external markup information that is present in hypertext in the form of explicit links to other documents.

B1. Wrapper Induction

Internet resources that contain relational data—telephone directories, product catalogs, etc.—use Formatting markup to clearly present the information they contain to users. However, with standard HTML, it is quite difficult to extract data from such resources in an automatic way. The XML markup language is designed to overcome these problems by encouraging page authors to mark their content in a way that reflects document structure at a detailed level; but it is not clear to what extent users will be prepared to share the structure of their documents fully in XML, and even if they do, huge numbers of legacy pages abound. Many software systems use external online resources by hand-coding simple parsing modules, commonly called "wrappers," to analyze the page structure and extract the requisite information. This is a kind of text mining, but one that depends on the input having a fixed, predetermined structure from which information can be extracted algorithmically. Given that this assumption is satisfied, the information extraction problem is relatively trivial. But this is rarely the case. Page structures vary; errors that are insignificant to human readers throw automatic extraction procedures off completely; Web sites evolve. There is a strong case for automatic induction of wrappers to reduce these problems when small changes occur, and to make it easier to produce new sets of extraction rules when structures change completely.

B2. Document clustering with links

Document clustering techniques are based on the documents' textual similarity. However, the hyperlink structure of Web documents, encapsulated in the "link graph" in which nodes are Web pages and links are hyperlinks between them, can be used as a different basis for clustering. Many standard graph clustering and partitioning techniques are applicable. Link-based clustering schemes typically use factors such as:□□ The number of hyperlinks that must be followed to travel in the Web from one document to the other; □ The number of common ancestors of the two documents, weighted by their ancestry distance and The number of common descendents of the documents, similarly weighted. These can be combined into an overall similarity measure between documents. In practice, a textual similarity measure is usually incorporated as well, to yield a hybrid clustering scheme that takes account of both the documents' content and their linkage structure. The overall similarity may then be determined as the weighted sum of four factors. Such a measure will be sensitive to the characteristics of the documents and their linkage structure, and given the number of parameters involved there is considerable scope for tuning to maximize performance on particular data sets.

B3. Determining "authority" of Web documents

The Web's linkage structure is a valuable source of information that reflects the popularity, sometimes interpreted as "importance," "authority" or "status," of Web pages. For each page, a numeric rank is computed. The basic premise is that highly-ranked pages are ones that are cited, or pointed to, by many other pages. Consideration is also given to (a) the rank of the citing page, to reflect the fact that a citation by a highly-ranked page is a better indication of quality than one from a lesser page, and (b) the number of out-links from the citing page, to prevent a highly ranked page from artificially magnifying its influence simply by containing a large number of pointers. This leads to a simple algebraic equation to determine the rank of each member of a set of hyperlinked pages. Complications arise from the fact that some links are "broken" in that they lead to nonexistent pages, and from the fact that the Web is not fully connected; these are easily overcome. Such techniques are widely used by search engines (e.g. Google) to determine how to sort the hits associated with any given query. They provide a social measure of status that relates to standard techniques developed by social scientists for measuring and analyzing social networks.

VI. APPROACHES TO TEXT MINING

Using well-tested methods and understanding the results of text mining:- Once a data matrix has been computed from the input documents and words found in those documents, various well-known analytic techniques can be used for further processing which includes methods for clustering, factoring, or predictive data mining

Black-box approaches to text mining and extraction of concepts. There are text mining applications which use black-box methods to take out detailed meaning from documents with less human effort. These text-mining applications summarize large numbers of text documents automatically, retaining the core and most important meaning of those documents.

Text mining as document search. The another approach of text mining is the automatic search of large numbers of documents based on key words or key phrases. This provides efficient access to Web pages with certain content. It searches very large document repositories based on varying criteria.

VII. CONCLUSION

In this paper our major focus is on how text is mined whether it is plain text or structured text. In structured text we have discussed how internal documents structure and external structure is mined which gives explicit hypertext links between documents. We have also discussed the functioning of text mining like one can switch in any vocabularies or thesauri to take advantage of terminology used in its own specific domain and NLP-based queries can be run in real time across millions of documents.

REFERENCES

[1] Agrawal, R. and Srikant, R. (1994) "Fast algorithms for mining association rules." *Proc Int Conf on Very Large Databases VLDB-94*, Santiago, Chile, pp. 487-499.

[2] Aone, C., Bennett, S.W., and Gorlinsky, J. (1996) "Multi-media fusion through application of machine learning and NLP." *Proc AAAI Symposium on Machine Learning in Information Access*. Stanford, CA.

[3] Appelt, D.E. (1999) "Introduction to information extraction technology." *Tutorial, Int Joint Conf on Artificial Intelligence IJCAI'99*. Morgan Kaufmann, San Mateo. Tutorial notes available at w w w.ai.sri.com/~appelt/ie-tutorial.

[4] Apte, C., Damerau, F.J. and Weiss, S.M. (1994) "Automated learning of decision rules for text categorization." *ACM Trans Information Systems*, Vol. 12, No. 3, pp. 233-251.

[5] Baeza-Yates, R. and Ribiero-Neto, B. (1999), *Modern information retrieval*. Addison Wesley Longman, Essex, England.

[6] Blum, A. and Mitchell, T. (1998) "Combining labeled and unlabeled data with co-training." *Proc Conf on Computational Learning Theory COLT-98*. Madison, Wisconsin, pp. 92-100.

[7] Borko, H. and Bernier, C.L. (1975) *Abstracting concepts and methods*. Academic Press, San Diego, California.

[8] Brill, E. (1992) "A simple rule-based part of speech tagger." *Proc Conf on Applied NaturalLanguage Processing ANLP-92*. Trento, Italy, pp. 152-155.

[9] Brin, S. and Page, L. (1998) "The anatomy of a large-scale hypertextual Web search engine." *ProcWorld Wide Web Conference WWW-7*. In *Computer Networks and ISDN Systems*, Vol. 30, No. 1-7, pp. 107-117.

Rashmi Agrawal is working as Associate Professor in Department of Computer Applications, Manav Rachna International University, Faridabad. Her qualifications are M.Tech, MBA, M.Phil (Computer Science) and having more than 11 years of teaching experience. She is pursuing Ph D in Computer Science from Manav Rachna International University, Faridabad. Her research area is Artificial Intelligence. She has published 5 papers in National/ International Journals and 9 papers in National/ International Conferences. Her area of interests is Data Structures, Artificial Intelligence and Software Testing. She has also written a book on Artificial Intelligence for Manav Rachna International Publication House. She is an active member of Computer Society of India. She is a reviewer/ member of IJRPES.

Mridula Batra is working as Assistant Professor in Department of Computer Applications, Manav Rachna International University, Faridabad. Her qualifications are MCA, M.Phil (Computer Science) and having more than 9 years of teaching experience. Her research area is Data Mining. Her area of interests is Data Base Systems, Computer Networks.