Volume 5, Issue 9, September 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Overview of Text Mining

[1]R. Balakrishnan Msc, Mphil, [2]B. Kaniimozhi

[1]Assistant Professor, Department of Information Technology, Dr.N.G.P Arts and Science College, Coimbatore, India

[2]Research Scholar, Department of Computer Science, Dr.N.G.P Arts and Science, College, Coimbatore, India

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. 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. 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.

Keywords: Discovery, Hidden information, parsing, Quantitative methods

I. INTRODUCTION

Many organizations across the world have already realized the benefits of text mining by converting unstructured corpus of documents into structured data first and then applying data mining on the structured data to derive valuable insights. Firms with efficient algorithms for text mining have a competitive advantage over those who do not. Many researchers have published work related to applications of text mining in various domains. These applications mainly fall under the general categories of text categorization, information retrieval and measurement. In recent years text mining is also being used for discovering trends in textual data. Given a set of documents with a time stamp, text mining can be used to identify trends of different topics that exist in the text.

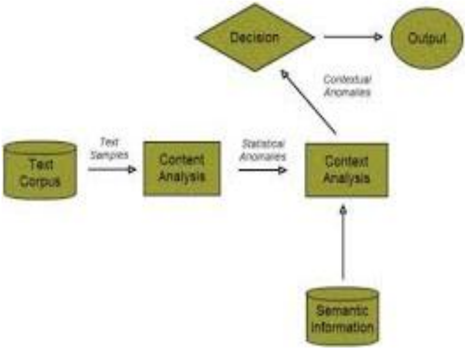

Fig 1. Basic structure of text mining

Text mining is a 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. The main Text Mining applications are most often used in the following sectors:

Publishing and media, Telecommunications, energy and other services industries, Information technology sector and Internet, Banks, insurance and financial markets, Political institutions, political analysts, public administration and legal documents, Pharmaceutical and research companies and healthcare.

Bioinformatics: In the bioinformatics domain, biomedical research literature has been a target for text mining. The goal of text mining in this area is to allow biomedical researchers to extract knowledge from the biomedical literature in facilitating new discovery in a more efficient manner.

Business Intelligence Of the major concerns in any business is to minimize the amount of guessing work involved in decision making. The risk of making wrong prediction should be reduced. Most of the data mining techniques are created to deal with prediction. The problem with data mining is that it can help only up to a certain point, since most of data are available in texts (reports, memos, emails, planning document, etc). Data mining and text mining techniques can complement each other. For example, data mining techniques may be used to reveal the occurrence of a particular event while text mining techniques may be used to look for an explanation of an event.

National Security: The use of text mining tool in national defence security domain has become an important issue. Government agencies are investing considerable resources in the surveillance of all kinds of communication, such as email, chats in chat rooms. Email is used in many legitimate activities such as messages and documents exchange.

Unfortunately, it can also be misused, for example in the distribution of unsolicited junk mail, mailing offensive or threatening materials. Since time is critical and given the scale of the problem, it is infeasible to monitor emails or chat rooms normally. Thus automatic text mining tools offer a considerable promise in this area. Although not much work has been conducted in this area (compared to bioinformatics), text mining technology is becoming an emergence technology for national security defence.

II. TEXT MINING STAGES

Text mining involves the application of techniques from areas such as Information Retrieval, Natural Language Processing, Information Extraction and Data Mining. These various stages of a text-mining process can be combined together into a single workflow.

Information Retrieval (IR) systems identify the documents in a collection which match a user's query. The most well known IR systems are search engines such as Google, which identify those documents on the World Wide Web that are relevant to a set of given words. IR systems are often used in libraries, where the documents are typically not the books themselves but digital records containing information about the books.

Natural Language Processing (NLP) is one of the oldest and most difficult problems in the field of artificial intelligence. It is the analysis of human language so that computers can understand natural languages as humans do.

Data Mining (DM) is the process of identifying patterns in large sets of data. The aim is to uncover previously unknown, useful knowledge. When used in text mining, DM is applied to the facts generated by the information extraction phase.

Information Extraction (IE) is the process of automatically obtaining structured data from an unstructured natural language document.

TASKS OF TEXT MINING ALGORITHMS

- Text categorization: assigning the documents with pre-defined categories (e.g decision trees induction).
- Text clustering: descriptive activity, which groups similar documents together (e.g.self-organizing maps).
- Concept mining: modelling and discovering of concepts, sometimes combines categorization and clustering approaches with concept/ logic based ideas in order to find concepts and their relations from text collections (e.g. formal concept analysis approach for building of concept hierarchy).
- Information retrieval: retrieving the documents relevant to the user's query.
- Information extraction: question answering.

III. METHODS OF MINING TEXT

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.

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.

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.

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.

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.

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.

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 thatmust 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.

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.

IV. TEXT MINING PROBLEMS & ISSUES

One main reason for applying data mining methods to text document collections is to structure them. A structure can significantly simplify the access to a document collection for a user. Well known access structures are library catalogues or book indexes. However, the problem of manual designed indexes is the time required to maintain them. Therefore, they are very often not up-to-date and thus not usable for recent publications or frequently changing information sources like the World Wide Web. The existing methods for structuring collections either try to assign keywords to documents based on a given keyword set (classification or categorization methods) or automatically structure document collections to find groups of similar documents (clustering methods).

The major challenging issue in text mining arise from the complexity of a natural language itself. The natural language is not free from the ambiguity problem. Ambiguity means the capability of being understood in two or more possible senses or ways. Ambiguity gives a natural language its flexibility and usability, and consequently, therefore it cannot be entirely eliminated from the natural language. One word may have multiple meanings. One phrase or sentence can be interpreted in various ways, thus various meanings can be obtained. Although a number of researches have been conducted in resolving the ambiguity problem, the work is still immature and the proposed approach has been dedicated for a specific domain.

V. 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.

REFERENCES

[1] Miller, W.T., (2005). Data and Text Mining A Business Applications Approach. Pearson Pentice Hall

[2] Battioui, C. (2008). A Text Miner analysis to compare internet and Medline information about allergy medications. SAS Regional Conference

[3] "Introduction to Text Miner." In "SAS Enterprise Miner Help." SAS Enterprise Miner 6.2 . SAS Institute Inc., Cary, NC

[4] Anwar M. Hossain, Mamunur M. Rashid, Chowdhury Mofizur Rahman, “A New Genetic Algorithm Based Text Classifier,” In Proceedings of International Conference on Computer and Information Technology,NSU, pp. 135-139, 2001.

[5] Canasai Kruengkrai , Chuleerat Jaruskulchai, “A Parallel Learning Algorithm for Text Classification,” The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002), Canada, July 2002.

[6] Jason D. M. Rennie, “Improving Multi-class Text Classification with Naive Bayes ,” 2001, Massachusetts Institute of Technology, http://citeseer.ist.psu.edu/cs.

[7] Jason Kroll, “Decision Tree Learning for Arbitrary Text Classification,” Sept 2003, www.cs.tufts.edu/~jkroll/dectree

BIOGRAPHY

Mr. R. Balakrishnan, working as a Assistant professor, Department of Information Technology, Dr N.G.P Arts and Science College, Coimbatore, Tamil Nadu, India

Ms. B. Kaniimozhi, Pursuing M.Phil Research Scholar, Department of Computer Science, Dr N.G.P Arts and Science College, Coimbatore, Tamil Nadu, India