ISSN 2349-0292
Impact Factor 2.675

GLOBAL JOURNAL OF ADVANCED ENGINEERING TECHNOLOGIES AND SCIENCES

A STUDY ON COMBINING INFORMATION EXTRACTION AND NATURAL LANGUAGE PROCESSING WITH TEXT MINING

Neha. E. Prince*, Sowmya. S. V, Ms. Mintu Movi

* BCA Students, Department of Computer Science, Christ College of Science and Management, Bangalore University

Under the guidance of Coordinator, Department of Computer Science, Christ College of Science and Management, Bangalore University

ABSTRACT

The problem of text mining, is discovering useful knowledge from unstructured text, is attracting increasing attention. The overarching goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP) and analytical methods. This technology is now broadly applied for a wide variety of government, research and business needs like security, biomedical, software, online media, marketing, sentiment analysis and academic applications. There are many software's available in the market that has made text mining easier. There are still researches going on to make new advancements in this field. An important approach to text mining involves the use of natural-language information extraction. Information extraction (IE) distills structured data or knowledge from unstructured text by identifying references to named entities as well as stated relationships between such entities. Information extraction systems can be used to directly extricate abstract knowledge from a text corpus, or to extract concrete data from a set of documents which can then be further analyzed with traditional data-mining techniques to discover more general patterns.

KEYWORDS: text mining, Information Extraction, Natural Language Processing, data analytics.

INTRODUCTION

Text mining is defined as "the process of finding useful or interesting patterns, models, directions, trends or rules from unstructured text"[1]. It is also referred to as text data mining and is roughly equivalent to text analytics which refers to the process of deriving high quality information from text. High quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text, deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty and interestingness. Text mining seeks to extract useful information through the identification and exploration of interesting patterns.

Information Extraction locates specific pieces of data from a corpus of natural-language texts. Although constructing an IE system is a difficult task, there has been significant recent progress in using machine learning methods to help automate the construction of IE systems. Manually annotating a small number of documents with the information to be extracted, a fairly accurate IE system can be induced from this labeled corpus and then applied to a large corpus of text to construct a database. However, the accuracy of current IE systems is limited and therefore an automatically extracted database will inevitably contain significant numbers of errors. An important question is whether the knowledge discovered from this "noisy" database is significantly less reliable than knowledge discovered from a cleaner database.

LITERATURE REVIEW

Text mining usually involves the process of structuring the input text, deriving patterns within the structured data, and finally evaluation and interpretation of the output. Text mining is defined as "the process of finding useful or interesting patterns, models, directions, trends or rules from unstructured text" [1].

An important part of our information-gathering behavior has always been to find out what other people think. With the growing availability and popularity of opinion-rich resources such as online review sites and personal blogs, new opportunities and challenges arise as people now can, and do, actively use information technologies to seek out and understand the opinions of others. The sudden eruption of activity in the area of opinion mining and sentiment analysis, which deals with the computational treatment of opinion, sentiment, and subjectivity in text, has thus occurred at least in part as a direct response to the surge of interest in new systems that deal directly with opinions as a first-class object. There are many techniques and approaches that promise to directly enable opinion-oriented information-seeking systems. We should focus on methods that seek to address the new challenges raised by sentiment-aware applications, as compared to those that are already present in more traditional fact-based analysis [2].

Text mining concerns look for patterns in unstructured text. The related task of Information Extraction, is about locating specific items in natural-language documents. The initial version of DiscoTEX integrates an IE module acquired by an IE learning system, and a standard rule induction module. However, this approach has problems when the same extracted entity or feature is represented by similar but not identical strings in different documents [3].

OBJECTIVES

- To show the evolving trend of information extraction.
- Increasing use of information extraction in text mining.

INFORMATION EXTRACTION

Information Extraction is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents. In most of the cases this activity concerns processing human language texts by means of natural language processing. Recent activities in multimedia document processing like automatic annotation and context extraction out of images/audio/video could be seen as information extraction. Due to the difficulty of the problems, current approaches to IE focus on narrowly restricted domains.

A broad goal of IE is to allow computation to be done on the previously unstructured data. A more specific goal is to allow logical reasoning to draw inferences based on the logical context of the input data. Structured data is semantically well-defined data from a chosen target domain, interpreted with respect to category and context.

Information Extraction is the part of a greater puzzle which deals with the problem of devising automatic methods for text management, beyond its transmission, storage and display. The discipline of information retrieval (IR) has developed automatic methods, typically of a statistical flavour, for indexing large document collections and classifying documents. Another complementary approach is that of natural language processing which has solved the problem of modelling human language processing with considerable success when taking into account the magnitude of the task. In terms of both difficulty and emphasis, IE deals with tasks in between both IR and NLP. In terms of input, IE assumes the existence of a set of documents in which each document follows a template, i.e. describes one or more entities or events in a manner that is similar to those in other documents but differing in the details. An example, consider a group of newswire articles on Latin American terrorism with each article is presumed to be based upon one or more terroristic acts. We also define for any given IE task a template, which is a (or a set of) case frame(s) to hold the information contained in a single document. For the terrorism example, a template would have slots corresponding to the perpetrator, victim, and weapon of the terroristic act, and the date on which the event happened. An IE system for this problem is required to “understand” an attack article only enough to find data corresponding to the slots in this template.

IE can be useful in a variety of applications, e.g. seminar announcements, course homepages, job postings and apartment rental ads. In particular, Califf (1998) suggested using machine learning techniques for extracting information from text documents in order to create easily searchable databases from the information, thus making the online text more easily accessible. For instance, information extracted from job postings in USENET newsgroup misc. Jobs offered can he used to build a searchable database of jobs. DIscoTEX is concerned with this aspect of IE, transforming unstructured texts to structured databases. Although most information extraction systems have been built entirely by hand until recently, automatic construction of complex IE systems began to be considered lately by many researchers. Recent proliferation of research on information extraction implies the possibility of using a successfully-built IE component for a larger text-mining system. The below example shows, a paired (shortened) document and template from an information extraction task in the resume domain. This template includes only slots that are filled by strings taken directly from the document. Several slots may have multiple fillers for the resume domain as in (programming} languages, platforms, and areas.

Example:
Document: I am a software engineer seeking a permanent position in Bangalore City. I have over twenty years of experience in all aspects of development of application software, with recent focus on design and implementation of systems involving multithreading and client/server architecture. For the past five years, I have implemented services in C and C++. I also have designed and implemented multithreaded applications in Java. Before that, I was working in programmed in OpenVMS for 5 years. I am pretty comfortable to develop software’s in Windows machines, also good in UNIX flavours.
F
illed Template: Title: software engineer Location: Bangalore Language: C, C++, Java Platform: OpenVMS, Windows, UNIX Area: multi-threading, client/server Years of Experience: twenty years.

ISSN 2349-0292
Impact Factor 2.675

DIFFERENCE BETWEEN TEXT MINING AND DATA MINING

The difference between regular data mining and text mining is that in text mining the patterns are extracted from natural language text rather than from structured database of facts. One application of text mining is in, bio informatics where details of experimental results can be automatically extracted from a large corpus of text and then processed computationally. Text mining techniques have been used in information retrieval systems as a tool to help users narrow their queries and to help them explore other contextually related subjects. Text Mining seems to be an extension of the better known Data Mining. Data Mining is a technique that analyses billions of numbers to extract the statistics and trends emerging from a company's data. This kind of analysis has been successfully applied in business situations as well as for military, social, government needs. But, only about 20% of the data on intranets and on the World Wide Web are numbers - the rest is text. The information contained in the text (about 80% of the data) is invisible to the data mining programs that analyze the information flow in corporations.

DISCOTEX

Text mining could be either summarizing the document or extracting keywords in the form of rules. DiscoTEX system is used to integrate information extraction module with KDD (knowledge discovery from database) module. DiscoTEX stands for "Discovery from text EXtraction". Initially a template is prepared with the slot and the fillers which are to be extracted. Using this template a structured database is created. DiscoTEX uses an automatically learned IE system (Rapier) to extract a structured database from a text corpus, and then mines this database with existing KDD tools. Before extracting rules similar words are assigned to a unique slot. KDD module intern helps extraction of the keywords with the help of the discovered rules. Rapier a learned machine system is used to perform information extraction and it constitutes an IE module for DiscoTEX system.

TEXT ANALYTICS SOFTWARE

- ActivePoint: offering natural language processing and smart online catalogues based contextual search and ActivePoint's TX5 discovery engine.
- Aiaioo Labs: offering API's for intentional analysis, sentiment analysis and event analysis.
- Alceste: a software for the automatic analysis of textual data.
- AlchemeAPI: the world's leading text analysis service, processing billions of documents every month.
- Anderson Analytics OdinText: complete text analytics software platform for consumer insights and customer service professionals.
- Angoss Text Analytics: part of knowledge studio, allows users to merge the output of unstructured, text-based analytics with structured data to perform data mining and predictive analytics.
- Ascribe: offering a unique hybrid technology approach, blending natural language processing, machine learning and semi-automated coding tools, since 1999.
- Attensity: offers a complete suite of text analytic applications, including the ability to extract "who", "where", "when" and "why" facts and then drill down to understand people, places and events and how they are related.
- Basis Technology: provides natural language processing technology for the analysis of unstructured multilingual text.
- ClearForest: tools for analysis and visualization of your document collection.

FINDINGS

- The technology of information extraction has become an important part of text mining.
- It is being used widely all over the world because of its various advantages.
- In future text mining can be applied to larger text corpora and DiscoTEX can be applied to other domains like medical, infrastructure and business.
- Also some other techniques for evaluating its performance will also be developed.

CONCLUSION

Text mining is a young, inter disciplinary field which draws on information retrieval, data mining, machine learning and statistics. As most information is stored as text, text mining is believed to have a high commercial potential value. Knowledge may be discovered from many sources of information; yet, unstructured texts remain the largest readily available source of knowledge. Text-mining systems can be developed relatively rapidly and evaluated easily on existing IE corpora by utilizing existing Information Extraction and data mining technology.

REFERENCES

[1] Q Mei, knowledge discovery in data mining,2005.
[2] B pang, Foundations and trends in information retrieval,2008.
[3] UY Nahm, symposium on mining,2002.

ISSN 2349-0292
Impact Factor 2.675

[4] Raymond J. mooney , Text mining with information extraction,2008.
[5] A don,Discovering interesting usage patterns,2007.
[6] W Duan, Decision support system,2011.
[7] A Huang, Similarity measures for text document,2010.
[8] N uramoto, Text mining system for knowledge discovery,2004.
[9] D Demner-Fushman, Briefings in text mining,2007.
[10] Chi min ho, Using DiscoTEX in Text mining, 2001.