ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Knowledge Discovery in Textual Data Using Text Mining

Uppu Nirosha
Assistant Professor,
CSE Branch.

ABSTRACT:

Text Mining has become an important research area. Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. In this paper, Text Mining techniques and applications have been s presented. In the general context of Knowledge Discovery, specific techniques, called Text Mining techniques, are necessary to extract information from unstructured textual data. The extracted information can then be used for the classification of the content of large textual bases.

Keywords: Text mining, feature extraction, text categorization, clustering, customer relationship management

INTRODUCTION:

With the developing significance of electronic substance and electronic media for putting away and trading content reports, there is additionally a developing enthusiasm for apparatuses that can help finding and sorting data incorporated into the content archives. Content mining, otherwise called record mining, is a rising innovation for examining vast accumulations of unstructured reports for the reasons for separating intriguing and non - inconsequential examples or information. It can be conceived as a jump from information mining or learning revelation from (organized) databases. This short review endeavors to shed a few lights on this new and quick developing field. Initial, a content mining system is given accentuation on its most vital step: data extraction. The four noteworthy content mining methods: highlight extraction, topical indexing, grouping and synopsis will be investigated as well. The significance of content mining turns out to be clear, when its applications are presented. At that point, the condition of craftsmanship business devices that backing the procedure of content mining will be quickly talked about. At last, difficulties of content mining will be highlighted in the conclusion. Text mining is a burgeoning 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. Compared with the kind of data stored in databases, text is unstructured, amorphous, and difficult to deal with algorithmically. Nevertheless, in modern culture, text is the most common vehicle for the formal exchange of information. The field of text mining usually deals with texts whose function is the communication of factual information or opinions, and the motivation for trying to extract information from such text automatically is compelling—even if success is only partial.Text mining is an outgrowth of this “real text” mindset. Accepting that it is probably not much, what can be done with unrestricted input? Can the ability to process huge amounts of text compensate for relatively simple techniques? Natural language processing, dominated in its infancy by unrealistic ambitions and swinging in childhood to the other extreme of unrealistically artificial worlds and trivial amounts of text, has matured and now embraces both viewpoints: relatively shallow processing of unrestricted text and relatively deep processing of domain-specific material. It is interesting that data mining also evolved out of a history of difficult relations between disciplines, in this case machine learning—rooted in experimental computer science, with ad hoc evaluation methodologies—and statistics—well-grounded theoretically, but based on a tradition of testing explicitly-stated hypotheses rather than seeking new information. Early machine learning researchers knew or cared little of statistics; early researchers on structured statistical hypotheses remained ignorant of parallel work in machine learning. The result was that similar techniques (for example, decision-tree building and nearest-neighbor learners) arose in parallel from the two disciplines, and only later did a balanced rapprochement emerge.

MOTIVATION FOR TEXT MINING:

Approximately 90% of the World's data is held in unstructured formats

- Web pages
- Emails
- Technical documents
- Corporate documents
- Books
- Digital libraries
- Customer complaint
- letters

Growing rapidly in size and importance

Text mining and data mining:

Just as data mining can be loosely described as looking for patterns in data, text mining is about looking for patterns in text. However, the superficial similarity between the two conceals real differences. Data mining can be more fully characterized as the extraction of implicit, previously unknown, and potentially useful information from data. The information is implicit in the input data: it is hidden, unknown, and could hardly be extracted without recourse to automatic techniques of data mining. With text mining, however, the information to be extracted is clearly and explicitly stated in the text. It's not hidden at all—most authors go to great pains to make sure that they express themselves clearly and unambiguously—and, from a human point of view, the only sense in which it is "previously unknown" is that human resource restrictions make it infeasible for people to read the text themselves. The problem, of course, is that the information is not couched in a manner that is amenable to automatic processing.

Text mining strives to bring it out of the text in a form that is suitable for consumption by computers directly, with no need for a human intermediary. Though there is a clear difference philosophically, from the computer's point of view the problems are quite similar. Text is just as opaque as raw data when it comes to extracting information—probably more so. Another requirement that is common to both data and text mining is that the information extracted should be "potentially useful." In one sense, this means actionable—capable of providing a basis for actions to be taken automatically. In the case of data mining, this notion can be expressed in a relatively domain-independent way: actionable patterns are ones that allow non-trivial predictions to be made on new data from the same source. Performance can be measured by counting successes and failures, statistical techniques can be applied to compare different data mining methods on the same problem, and so on. However, in many text mining situations it is far harder to characterize what "actionable" means in a way that is independent of the particular domain at hand. This makes it difficult to find fair and objective measures of success. In many data mining applications, "potentially useful" is given a different interpretation: the key for success is that the information extracted must be comprehensible in that it helps to explain the data. This is necessary whenever the result is intended for human consumption rather than (or as well as) a basis for automatic action. This criterion is less applicable to text mining because, unlike data mining, the input itself is comprehensible. Text mining with comprehensible output is tantamount to summarizing salient features from a large body of text, which is a subfield in its own right: text summarization.

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.

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. Knowledge Discovery System's Concept Explorer is a visual search tool that helps to find precisely related content on the web. It "learns" relationships between words and phrases automatically from sample documents and visually guides you to construct searches. document retrieval tool with some text analysis and summarization capabilities.

IBM's Intelligent Miner is probably one of the most comprehensive text mining products around. It offers a set of text analysis tools, including a feature extraction tool, a set of clustering tools, a summarization tool, and a categorization tool. Also incorporated are the IBM's text search engine, Net Question Solution and the IBM web crawler package. TextWise, an R&D company based in Syracuse University, offers various text mining products. DR-LINK is an information retrieval system based on automatic concept expansion. CINDOR is its cross lingual version. CHESS is a text analysis and information extraction tool. Also an information extraction tool is the Data Junction's Cambio, which extracts data in the form of relational attributes from text. Megaputer's TextAnalyst uses a semantic net representation of documents and performs automated indexing, topic assignment, text abstraction, and semantic search.

TEXT MINING FRAMEWORK:

Text mining is firmly identified with the fields of data extraction , and data recovery, and to be sure can be thought to be worked from parts that perform these errands. The best perspective of a content mining framework would be one that takes after a grouping of steps, laid out underneath.

1. Information Retrieval (IR): Information retrieval might be regarded 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 could be followed by a text summarization stage that focuses on the query posed by the user, or an information extraction stage using techniques described below. In practice, however, standard textbooks use the term simply for plain document retrieval. Of course, the granularity 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.

2. Information Extraction (IE): "Information extraction" is used to refer to the task of filling templates from natural language input , one of the principal subfields of text mining. A commonly-cited domain is that of terrorist events, where the template may include slots for the perpetrator, the victim, type of event, where and when it occurred, etc.

3. Information Mining (IM): Information mining is the non-trivial process of identifying valid, novel, potentially useful, and understandable patterns in heterogeneous information sources.

4. Interpretation: The last step is to put an understanding on the examples recovered from the mining stage. Preferably the translation would itself be in a characteristic dialect.

TEXT MINING APPLICATIONS:

- Classification of news stories, web pages, … , according to their content
- Email and news filtering
- Organize repositories of document-related meta-information for search and retrieval (search engines)
- Clustering documents or web pages Gain insights about trends, relations between people, places and/or organizations.

CUSTOMER RELATIONSHIP:

Massive information is available in textual form in databases or online sources, and for many enterprise functions (marketing, maintenance, finance, etc.) represents a huge opportunity to improve business knowledge.

For example, text mining is starting to be used in marketing, more specifically in analytical customer relationship management, in order to achieve the holy 360° view of the customer (integrating elements from inbound mails, web comments, surveys, internal notes, etc.).

CHALLENGING ISSUE:

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. On the other hand, most of the IE systems that involve semantic analysis exploit the simplest part of the whole spectrum of domain and task knowledge, that is to say, named entities. IE does a more limited task than full text understanding. He pointed that in full text understanding, all the information in the text is presented, whereas in information extraction, the semantic range of the output, the relations will be presented are delimited. However, the growing need for IE application to domains such as functional genomics requires more text understanding.

CONCLUSION:

Generally accepted characterizations of what it covers do not yet exist. When the term is broadly interpreted, many different problems and techniques come under its ambit. In most cases it is difficult to provide general and meaningful evaluations because the task is highly sensitive to the particular text under consideration. Document classification, entity extraction, and filling templates that correspond to given relationships between entities, are all central text mining operations that have been extensively studied. Using structured data such as Web pages rather than plain text as the input opens up new possibilities for extracting information from individual pages and large networks of pages. Automatic text mining techniques have a long way to go before they rival the ability of people, even without any special domain knowledge, to glean information from large document collections.

At last we conclude that, Text mining is also known as Text Data Mining or Knowledge-Discovery in Text (KDT), refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Text mining is a young interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics. As most information (over 80%) 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.

REFERENCES:

[l] F. Marcotorchino, “Piece seriation issues: A bound together approach”, Applied Stochastic Models and Data Analysis, 3, 73-91 (1987).

[2] N. Wacholder and Y. Ravin, Disambiguation of Proper Names in Text, Proc. of the fifth Conference on Applied Natural Language Processing, April 1997, Washington, D.C.

[3] Danial Tkach, (1998), Text Mining Technology Turning Information Into Knowledge A white paper from IBM

[4] Helena Ahonen, Oskari Heinonen, Mika Klemettinen, A. Inkeri Verkamo, (1997), Applying Data Mining Techniques in Text Analysis, Report C-1997-23, Department of Computer Science, University of Helsinki, 1997 .

[5]Mark Dixon, (1997), An Overview of Document Mining Technology, http://www.geocities.com/Research-Triangle/Thinktank/1997/mark/compositions/dixm 97_ dm.ps .

[6] Ah-Hwee Tan, (1999), Text Mining: The condition of craftsmanship and the difficulties, In procedures, PAKDD’99 Workshop on Knowledge disclosure from Advanced Databases (KDAD’99), Beijing, pp. 71-76, April 1999.