A
Seminar report on

Text Mining

Submitted in partial fulfillment of the requirement for the award of degree
of MCA

SUBMITTED TO: **SUBMITTED BY:**

www.studymafia.org www.studymafia.org

Preface

I have made this report file on the topic **,Text mining,** I have tried my best to elucidate all the relevant detail to the topic to be included in the report. While in the beginning I have tried to give a general view about this topic.

My efforts and wholehearted co-corporation of each and everyone has ended on a successful note. I express my sincere gratitude to ……..who assisting me throughout the prepration of this topic. I thank him for providing me the reinforcement, confidence and most importantly the track for the topic whenever I needed it.

Content

Abstract of Text Mining

The volume of information circulating in a typical enterprise continues to increase. Knowledge hidden in the information however, is not fully utilized, as most of the information is described in textual form (as sentences). A large amount of text information can be analyzed objectively and efficiently with Text Mining.

The field of text mining has received a lot of attention due to the ever increasing need for managing the information that resides in the vast amount of available text documents. Text documents are characterized by their unstructured nature. Ever increasing sources of such unstructured information include the World Wide Web, biological databases, news articles, emails etc.

Text mining is defined as the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation.

As the amount of unstructured data increases, text-mining tools will be increasingly valuable. A future trend is integration of data mining and text mining into a single system, a combination known as duo-mining

Introduction

Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text mining is different from what are familiar with in web search.

In search, the user is typically looking for something that is already known and has been written by someone else. The problem is pushing aside all the material that currently is not relevant to your needs in order to find the relevant information. In text mining, the goal is to discover unknown information, something that no one yet knows and so could not have yet written down.

Machine intelligence is a problem for text mining. Natural language has developed to help humans communicate with one another and record information. Computers are a long way from comprehending natural language. Humans have the ability to distinguish and apply linguistic patterns to text and humans can easily overcome obstacles that computers cannot easily handle such as slang, spelling variations and contextual meaning. However, although our language capabilities allow us to comprehend unstructured data, we lack the computer's ability to process text in large volumes or at high speeds. Figure depicts a generic process model for a text mining application.

Starting with a collection of documents, a text mining tool would retrieve a particular document and preprocess it by checking format and character sets. Then it would go through a text analysis phase, sometimes repeating techniques until information is extracted.

Three text analysis techniques are shown in the example, but many other combinations of techniques could be used depending on the goals of the organization. The resulting information can be placed in a management information system, yielding an abundant amount of knowledge for the user of that system.

Meaning of Text mining

Text mining (TM) seeks to extract useful information from a collection of documents.
It is similar to data mining (DM), but the data sources are unstructured or semi-structured documents. The TM methods involve :
- Basic pre-processing / TM operations, such as identification / extraction of representative features (this can be done in several phases)
- Advanced text mining operations, involving identification of complex patterns (e.g. relationships between previously identified concepts)
TM exploits techniques / methodologies from data mining, machine learning, information retrieval, corpus-based computational linguistics

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 [Witten and Frank, 2000].

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 extractedis 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.

Applications of Text Mining

Unstructured text is very common, and in fact may represent the majority of information available to a particular research or data mining project.

Analyzing open-ended survey responses. In survey research (e.g., marketing), it is not uncommon to include various open-ended questions pertaining to the topic under investigation. The idea is to permit respondents to express their "views" or opinions without constraining them to particular dimensions or a particular response format.

This may yield insights into customers' views and opinions that might otherwise not be discovered when relying solely on structured questionnaires designed by "experts." For example, you may discover a certain set of words or terms that are commonly used by respondents to describe the pro's and con's of a product or service (under investigation), suggesting common misconceptions or confusion regarding the items in the study.

Automatic processing of messages, emails, etc. Another common application for text mining is to aid in the automatic classification of texts. For example, it is possible to "filter" out automatically most undesirable "junk email" based on certain terms or words that are not likely to appear in legitimate messages, but instead identify undesirable electronic mail.

In this manner, 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 or agency; e.g., email messages with complaints or petitions to a municipal authority are automatically routed to the appropriate departments; at the same time, the emails are screened for inappropriate or obscene messages, which are automatically returned to the sender with a request to remove the offending words or content.

Analyzing warranty or insurance claims, diagnostic interviews, etc. In some business domains, the majority of information is collected in open-ended, 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. This information can then be usefully exploited to, for example, identify common clusters of problems and complaints on certain automobiles, etc. Likewise, in the medical field, open-ended descriptions by patients of their own symptoms might yield useful clues for the actual medical diagnosis.

Investigating competitors by crawling their web sites. Another type of potentially very useful application is to automatically process the contents of Web pages in a particular

domain. For example, you could go to a Web page, and begin "crawling" the links you find there to process all Web pages that are referenced.

In this manner, you could automatically derive a list of terms and documents available at that site, and hence quickly determine the most important terms and features that are described. It is easy to see how these capabilities could efficiently deliver valuable business intelligence about the activities of competitors.

Text Mining Features

Data Access

- Accesses numerous forms of textual data such as PDF, extended ASCII text, HTML, Microsoft Word, and OpenDocument format
- Web crawling capabilities
- ETL textual data into an SAS data set for mining

Feature Extraction

- Vocabulary finder extracts technical terms, product and company names as well as common misspellings
- Phrase finder identify recurring phrases and expressions
- Normalizes and includes extracted entities in a matrix table
- Entity extraction is available for multiple languages

Text Processing Capabilities

- Content analysis on short alphanumeric variables (up to 255 characters) and longer ANSI, RTF, and other formats
- Captures and distills the most important underlying information within a document collection
- Default or customized stop lists for each language removes terms with little or no informational value
- Calls external text pre-processing to EXE or to DLL
- Integrated multilingual spell-checking
- Integrated thesaurus to assist the creation of taxonomies and comprehensive categorization schemas
- Case filtering on any numeric or alphanumeric field and on code occurrence (with AND, OR, and NOT Boolean operators)
- Excludes pronouns, conjunctions, etc. based on user-defined exclusion lists (or stop list)
- Categorizes words or phrases using existing or user-defined dictionaries
- Categorizes Word based on Boolean (AND, OR, NOT) and proximity rules (NEAR, AFTER, BEFORE)
- Substitutes and scores Word and phrase substitution using wildcards and weighing

Human text mining

All scientific researchers are expected to use the literature as a major source of information during the course of their work to provide new ideas and supplement their laboratory studies. However, some feel that this can be taken further: that new information, or at least new hypotheses, can be derived directly from the literature by researchers who are expert in information-seeking but not necessarily in the subject matter itself.

Subject-matter experts can only read a small part of what is published in their fields and are often unaware of developments in related fields. Information researchers can seek useful linkages between related literatures which may be previously unknown—particularly if there is little explicit cross-reference between the literatures.
We briefly sketch an example, to indicate what automatic text mining may eventually aspire to—but is nowhere near achieving yet.

By analyzing chains of causal implication within the medical literature, new hypotheses for causes of rare diseases have been discovered—some of which have received supporting experimental evidence [Swanson 1987; Swanson and Smalheiser, 1997]. While investigating causes of migraine headaches, Swanson extracted information from titles of articles in the biomedical literature, leading to clues like these:

- Stress is associated with migraines
- Stress can lead to loss of magnesium
- Calcium channel blockers prevent some migraines
- Magnesium is a natural calcium channel blocker
- Spreading cortical depression is implicated in some migraines
- High levels of magnesium inhibit spreading cortical depression
- Migraine patients have high platelet aggregability
- Magnesium can suppress platelet aggregability
- These clues suggest that magnesium deficiency may play a role in some kinds of migraine
- headache, a hypothesis that did not exist in the literature at the time Swanson found these links.

Approaches to Text Mining

To reiterate, text mining can be summarized as a process of "numericizing" text. At the simplest level, all words found in the input documents will be indexed and counted in order to compute a table of documents and words, i.e., a matrix of frequencies that enumerates the number of times that each word occurs in each document.

This basic process can be further refined to exclude certain common words such as "the" and "a" (stop word lists) and to combine different grammatical forms of the same words such as "traveling," "traveled," "travel," etc. (stemming). However, once a table of (unique) words (terms) by documents has been derived, all standard statistical and data mining techniques can be applied to derive dimensions or clusters of words or documents, or to identify "important" words or terms that best predict another outcome variable of interest.

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 those data including methods for clustering, factoring, or predictive data mining (see, for example, Manning and Schütze, 2002).

"Black-box" approaches to text mining and extraction of concepts. There are text mining applications which offer "black-box" methods to extract "deep meaning" from documents with little human effort (to first read and understand those documents). These text mining applications rely on proprietary algorithms for presumably extracting "concepts" from text, and may even claim to be able to summarize large numbers of text documents automatically, retaining the core and most important meaning of those documents.

While there are numerous algorithmic approaches to extracting "meaning from documents," this type of technology is very much still in its infancy, and the aspiration to provide meaningful automated summaries of large numbers of documents may forever remain elusive.

We urge skepticism when using such algorithms because 1) if it is not clear to the user how those algorithms work, it cannot possibly be clear how to interpret the results of those algorithms, and 2) the methods used in those programs are not open to scrutiny, for example by the academic community and peer review and, hence, we simply don't know how well they might perform in different domains.

As a final thought on this subject, you may consider this concrete example: Try the various automated translation services available via the Web that can translate entire paragraphs of text from one language into another. Then translate some text, even simple

text, from your native language to some other language and back, and review the results. Almost every time, the attempt to translate even short sentences to other languages and back while retaining the original meaning of the sentence produces humorous rather than accurate results. This illustrates the difficulty of automatically interpreting the meaning of text.

Text mining as document search. There is another type of application that is often described and referred to as "text mining" - the automatic search of large numbers of documents based on key words or key phrases.

This is the domain of, for example, the popular internet search engines that have been developed over the last decade to provide efficient access to Web pages with certain content.

While this is obviously an important type of application with many uses in any organization that needs to search very large document repositories based on varying criteria, it is very different from what has been described here.

Conclusion

Text mining is a burgeoning technology that is still, because of its newness and intrinsic difficulty, in a fluid state—akin, perhaps, to the state of machine learning in the mid-1980s. 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.

BIBLIOGRAPHY

R. Feldman, J. Sanger: The Text Mining Textbook: Advanced Approaches in Analyzing
Unstructured Data, Cambridge Univ. Press, 2007
S.Weiss, N.Indurkhya, T.Zhang, F. Damerau, Text Mining: Predictive Methods for Analysing
Unstructured Information, Springer, 2005.
R. Feldman, J. Sanger: The Text Mining Handbook: Advanced Approaches in Analyzing
Unstructured Data, Cambridge Univ. Press, 2007
F. Sebastiani: Machine Learning in Automated Text Classification, J. ACM Computing
Surveys, Vol. 34, No.1, 2002.
F. Colas, P. Brazdil: On the Behavior of SVM and Some Older Algorithms in Binary Text
Classification Tasks, in Text, Speech and Dialog, LNCS, Vol. 4188, pp. 45-52, 2006.
Cordeiro, J., Brazdil, P.: Learning Text Extraction Rules without Ignoring Stop Words. in
the 4th International Workshop on Pattern Recognition in Information Systems – PRIS
- 2004; pp. 128-138.
Patil, K. and Brazdil, P., SumGraph: Text Summarization using Centrality in the Pathfinder
Network, International Journal on Computer Science and Information Systems, 2(1),
pp. 18-32, 2007.
Ingo Feinerer: Introduction to the tm Package: Text Mining in R, http://cran.rproject.
org/web/ packages/tm/vignettes/tm.pdf
Luís Torgo: A Linguagem R, programação para a Análise de Dados,