New and Fast Emerging Advance Structure of Text Mining from Unstructured Data

Tae-Jin Oh and Anthony

Abstract--- At present, text mining has developed into an significant research topic. Text Mining is the discovery using computer of new, previously unidentified information, by automatically extracting information from several written resources. Data mining and text mining have been of massive value in several business sectors. In order to make data mining (discovering “facts” and “insights” from structured data) and text mining (discovering “facts” and “insights” from unstructured data), important resources are necessary. While considering industries like pharmaceuticals, these massive efforts which can cost tens of millions of dollars comprise proprietary information and extremely larger quantities of open source data.

Here, advance structure of text mining with its framework, techniques and applications areas have been completely presented. A new and fast emerging field a text mining framework is formulated by highlighting on its most important step: information extraction, retrieval, mining and interpretation. And four major text mining approaches: feature extraction, thematic indexing, clustering and summarization will be explored clearly too. And also represented the significance of text mining becomes extremely clear, when its applications are introduced.

Keywords--- Text Mining (TM), Information Extraction, Information Retrieval, Information Mining, Interpretation, Summarization, Clustering, Fact Integration and Fact Extraction etc.

I. INTRODUCTION

TEXT MINING (TM) is the process of extracting novel, undetected and unstructured knowledge “hidden” in a large collection of unstructured text documents, using advanced technology. It enables the knowledge worker to uncover relationships in a text collection and to explore them in order to discover new knowledge.

The phrase “text mining” is generally used to denote any system that analyzes large quantities of natural language text and detects lexical or linguistic usage patterns in an attempt to extract probably useful (although only probably correct) information. In discussing a topic that lacks a generally accepted definition in a practical Handbook such as this, I have chosen to cast the net widely and take a liberal viewpoint of what should be included, rather than attempting a clear-cut characterization that will inevitably restrict the scope of what is covered.

Tae-Jin Oh, Kyoto University, Japan.

Anthony, UCL (University College London), UK.

DOI : 10.9756/BIJIEMS.8325

Advance Structure of Text Mining

As the most natural form of storing and exchanging information is written words, text mining (TM) has a very high commercial potential. In fact, a recent study indicated that 80% of a company's information is contained in text documents, such as emails, memos, customer correspondence, and reports. Traditional document and text management tools are inadequate to meet the utilities. Document management systems work well with homogeneous collections of documents but not with the heterogeneous mix that knowledge workers face every day. Even the best Internet search tools suffer from poor precision and recall. (Precision is a measure of how many documents returned from a search actually meet the intended query criterion. Recall measures the percentage of documents returned versus how many should have been returned.). The ability to distil this untapped source of information, free text document, provides substantial competitive advantages for a company to succeed in the era of a knowledge-based economy.

II. TEXT MINING FRAMEWORK

Text mining is closely related to the fields of information extraction (IE), and information retrieval, and indeed can be considered to be built from components that perform these tasks.

The best view of a text mining system would be one that follows a sequence of steps, outlined below:

A. Information Retrieval (IR)

The first step is to locate and retrieve documents that can be considered relevant to the task at hand. Typically the users of the system can specify document sets manually, but we still need a system that filters out irrelevant documents. For

example, in an IR search of the Web, a person would like to find relevant documents among all the other documents available online.

B. Information Extraction (IE)

The next stage is to extract information from the selected documents. This extraction is typically a process of filling out user specified templates of expected information. IE is arguably the most important component of the text mining process, in fact, most papers treat the two as the same thing, though they are not equivalent. IE is the process concerned with scanning a text or set of texts with the goal of extracting facts present in the text(s). Information extraction steps will be detailed shortly.

C. Information Mining (IM)

Once a template entry has been filled out for each document, we are in a stage where we effectively have a database-like structure(s) which is compatible with standard data mining techniques. Here we are seeking to discover patterns within the data.

D. Interpretation

The final step is to place an interpretation on the patterns retrieved from the mining phase. Ideally the interpretation would itself be in a natural language.

III. STEPS OF INFORMATION EXTRACTION

There are three basic steps involved in the IE process, or more specifically-the subset of IE concerned with extracting a particular class of actions or events:

A. Fact Extraction

At this stage we are concerned with searching for individual facts contained within the document. Here the domain specific knowledge is crucial, because we can encode pattern recognition for particular facts that we know to expect in the document. Some techniques in fact extraction follows:

Pattern Matching

Pattern matching is the process of using common regular expressions to form the lowest level of extraction, effectively constructing a bottom-up parsing of the text. Pattern matching is also used on a larger scale when processing tokens that have been assigned a syntactic value.

Lexical Analysis

Many phases go into building up a lexical based approach to fact extraction. At the lowest level, we are concerned with breaking the text up into tokens (as in programming language parsing) and from there identifying sentences.

Within sentences we look to determine likely context of words and phrases, using various dictionaries and domain specific lexicons. By this stage we should have recognized tokens that are proper names.

Syntactic and Semantic Structure

The next layer of processing endeavors to assign a syntactic component to words and/or phrases in each sentence. Looking for noun or verb groups can be performed while

looking at the local text (e.g., the current sentence), and provide further clues to the context of word occurrences in a sentence. Having built up the basic syntactic meaning of a sentence, an IE system can then start to look for semantic patterns that it has been trained to look for. For instance, having recognized the presence of the structure "person retires as position" in a sentence, the semantic pattern matcher can recognize this as an event.

B. Fact Integration

In fact integration, we look at all individual facts in the document, and see how the facts fit together to form a bigger picture. One of the earlier phases of integration is to resolve indirect references, for example a reference to "he" must be resolved to determine just which person is being referenced. On a higher level of integration, we can look at the concept of event merging. Here is an example:

"Samy was president. He was succeeded by Hany"

We can merge the two separate facts to establish that Hany was president of the company. Merging earlier facts would tell us which company Hany is president of, and possibly why Samy left. Processes such as this may use production systems, and are often hard coded for a particular problem set.

C. Knowledge Representation

This is a trivial phase of the IE process, but can be important for the end use of the information extracted. One option is filling template entries that can be entered into a database. Many other styles of output could also be useful; there is some early research into graphical visualization of results. Of the two key stages outlined above (fact extraction and integration), there are many techniques that can be used to achieve their goals. These approaches cover natural language processing techniques, statistical techniques, even neural networks.

IV. TEXT MINING TECHNIQUES

The key techniques of text mining include:

- Feature extraction
- Thematic indexing
- Clustering
- Summarization

These four techniques are essential because they solve two key problems with applying text mining: they make textual information accessible, and they reduce the volume of text that must be read by end users before information is found.

A. Feature Extraction

It deals with finding particular pieces of information within a text. The target information can be of a general form such as type descriptions or business of the former, while be pattern-driven. For example, applications analyzing merger and acquisition stories may extract names of the companies involved, cost, funding mechanisms and whether or not regulatory approval is required.

B. Thematic Indexing

It uses knowledge about the meaning of words in a text to identify broad topics covered in a document. For example, documents about aspirin and ketofan might be both classified under pain relievers or analgesics. Thematic indexing such as this is often implemented using multidimensional taxonomies. Taxonomy, in the text mining sense, is a hierarchical knowledge representation scheme. This construct, sometimes called ontology to distinguish it from navigational taxonomies such as Yahoo's, provides the means to search for documents about a topic instead of documents with particular keywords. For example, an analyst researching mobile communications should be able to search for documents about wireless protocols without having to know key phrases such as wireless application protocol (WAP).

C. Clustering

It is another text mining technique with applications in business intelligence. Clustering groups similar documents according to dominant features. In text mining and information retrieval, a weighted feature vector is frequently used to describe a document. These feature vectors contain a list of the main themes or keywords along with a numeric weight indicating the relative importance of the theme or term to the document as a whole. Unlike data mining applications which use a fixed set of features for all analyzed items (e.g. age, income, gender, etc.), documents are described with a small number of terms or themes chosen from potentially thousands of possible dimensions.

There is no single, best way to deal with document clustering; but three approaches are often used: hierarchical clusters, binary clusters and self-organizing maps. Hierarchical clusters use a set-based approach. The root of the hierarchy is the set of all documents in a collection, and the leaf nodes are sets with individual documents. Intervening layers in the leaf nodes have progressively larger sets of documents, grouped by similarity. In binary clusters each document is in one and only one cluster, and clusters are created to maximize the similarity measures between documents in a cluster and minimize the similarity measure between documents in different clusters. Self-organizing maps

(SOMs) use neural networks to map documents from sparse high-dimensional spaces into two-dimensional maps. Similar documents tend to same position in two dimensional grids.

D. Summarization

The last text mining technique is summarization. The purpose of summarization is to describe the content of a document while reducing the amount of text a user must read. The main ideas of most documents can be described with as little as 20 percent of the original text. Little is lost by summarizing. Like clustering, there is no single summarization algorithm. Most use morphological analysis of words to identify the most frequently used terms while eliminating words that carry little meaning, such as the articles the, an and a. Some algorithms weight terms used in opening or closing sentences more heavily than other terms, while some approaches look for key phrases that identify.

V. APPLICATION AREAS OF TEXT MINING

From government and legislative organizations, to corporations and universities, and to journalists, writers and college students, we all create, store, retrieve, and analyze texts.

Hence, numerous organizations are faced with various document management and text analysis tasks. Consider a few simple examples:

- Internet search engines could deliver much better quality results by accepting and making sense of natural language queries. If documents found in response to a query were analyzed semantically for their relevance in the context of the original query, it could significantly increase the precision of the search: instead of finding a knockout amount of more than 10,000 documents in response to your query, the system could provide you with a short list of the most relevant documents.
- Call center specialists have to understand customer support questions, quickly select relevant documents among available manuals, frequently asked questions lists, and engineering notes, and retrieve those bits of knowledge that help answer the question. An automated system for categorizing available materials and retrieving the most relevant fragments matching natural language questions could save hundreds of thousands of man-hours and dramatically reduce response time. Identifying the best fragments through thesauruses and ontologism could significantly improve recall, or the thoroughness of the search.
- Lawyers, insurers and venture capitalists often have to

quickly grasp the meaning of cases, claims and proposals, correspondingly. They need to improve the quality of querying the Web and diverse databases to find and retrieve relevant documents. Their practice could benefit tremendously from automated summarization of texts and feature extraction, when key points from the text are or gained in a database holding meta-information to improve future access to knowledge contained in documents.

- Intelligent Email Routing, Automatic Chat Rooms Monitoring, Web pages monitoring are all important applications which depend largely on Text Mining.
- Researching medical journals for new hypotheses of cause and effect for a disease is an ideal case of what text mining ought to be able to do.

VI. CONCLUSION

Unlike imposing the researcher's preconceived ideas on the phenomenon under study by using force-option survey responses or pre-determined hypotheses, in text mining the researchers let the themes emerge freely from the subjects. Text mining for discovering common threads fits naturally into the sequential exploratory design in the mixed-method paradigm (Creswell & Plano Clark, 2007). According to Creswell and Plano Clark, when the research problem is ill-defined and the variables are unknown, the researcher could employ qualitative methods to understand the problem rather than imposing a theory on the phenomenon. This design is particularly useful when the researcher needs to develop an instrument to collect data, but is not sure about what constructs should be measured and what questions should be asked.

In this short paper-view, the notion of text mining have been introduced, a rough framework for text mining, its techniques and its applications areas available have been presented.

REFERENCES

- [1] W. Berry Michael, "Automatic Discovery of Similar Words", in "Survey of Text Mining: Clustering, Classification and Retrieval", Springer Verlag, New York, Pp. 24-43, 2004.
- [2] N. Kanya and S. Geetha, "Information Extraction: A Text Mining Approach", IET-UK International Conference on Information and Communication Technology in Electrical Sciences, Pp. 1111-1118, 2007.
- [3] R.N. Kostoff, "Science and technology text mining: global technology watch", Office of Naval Research (ONR).
- [4] Seth Grimes, "The developing text mining market", white paper, Text Mining Summit05 Alta Plana Corporation, Boston, 2005.
- [5] U. Nahm and R. Mooney, "Text mining with information extraction", Proceedings of AAAI2002 Spring Symposium on Mining Ans from Texts & Knowledge Bases, 2002.
- [6] I. Mani and M.T. Maybury, "Advances in Automatic Text Summarization", MIT, 1999.
- [7] M. Hearst, "Untangling text data mining", Proc. of ACL'99 the 37th Annual Meeting of the Association for Computational Linguistics, 1999.
- [8] J. M. G. Hidalgo, "Tutorial on text mining and internet content filtering", Tutorial Notes, 2002.
- [9] D. Tkach, "Text mining technology: turning information into knowledge", IBM, 1998.
- [10] D.K. Harman, "Overview of the third text retrieval conference", Proc Text RetrievalConference TREC-3. National Institute of Standards, Gaithersburg, Pp. 1-20, 1995.

- [11] W. Fan, L. Wallace, S. Rich and Z. Zhang, "Tapping the power of text mining", Communications of the ACM, Vol. 49, No. 9, Pp. 76-82, 2006.
- [12] J. Dorre, P. Gersti and R. Seiffert, "Text Mining: Finding Nuggets in Mountains of Textual Data", Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 1999.
- [13] R. Bilisoly, "Practical Text Mining with Perl", New York: John Wiley & Sons, 2008.
- [14] A. Hotho, A. N'urnberger and G. Paaß, "A Brief Survey of Text Mining", Ldv Forum, Vol. 20, No. 1, Pp. 19-62, 2005.
- [15] Ah-Hwee Tan, "Text Mining: The state of art and the challenges", 1999.
- [16] G. Salton, "The state of retrieval systems evaluation", Information Processing and Management, Vol. 28, Pp. 441-449, 1992.
- [17] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann, San Francisco, 2000.