A Review of Ontological-based Pattern Mining Techniques

Charushila Kadu, Praveen Bhanodia, Pritesh Jain

Abstract— in this paper, we present an overview of existing text mining methodologies. All these methodologies are described more or less on is much more their own. Text mining is a very popular and computationally expensive task. We also explain the fundamentals of text mining. We describe today's approaches for text mining. From the broad variety of efficient algorithms that have been developed we will compare the most important ones. We will systematize the algorithms and analyze their performance based on both their run time performance and theoretical considerations. Their strengths and weaknesses are also investigated. It turns out that the behavior of the algorithms similar as to be expected.

Index Terms— term based model, concept-based model, PTM, Pattern e

1 INTRODUCTION: TEXT MINING

Recently text mining has become an important research area. Text can be placed in newspaper articles, SMS, mails, online chats, journals, product reviews, and organization files. Text mining also known as text data mining, intelligent text analysis or knowledge discovery in text (KDT) refers to extracting the useful information from the natural language text. Text mining discovers new pieces of knowledge from textual data. Basically text mining is used to combine countless pages of plain-language digitized text to find useful information that has been hiding in plain sight.

Approximately 80% of the world's data is in unstructured format. Most of the industries, government sectors, organizations and institutions data are stored in electronic form. These data are stored in text database format. Text database is a combination of some structured and unstructured data which in results belong to semi-structured format. For example employees, have Employee. No, name, department are the structured fields and address, remarks are unstructured fields in an organization. Hence, Text mining becomes essential for any organization as most of the information in the organizations is in text format.

1.1 Steps in Text mining

Generally text mining involves several steps

(1) Conversion of unstructured text input into structured database.
(2) Identification of the patterns and trends from the structured database.
(3) Finally, Extraction of the useful information is done from the text. Analyze and interpret the patterns and trends are analyzed and interpreted.
(4)

Details steps of the text mining are depicted in fig 2.

1.1 Text mining vs. other mining Techniques

In text mining pattern are discovered from natural language whereas in data mining patterns are extracted from database. Free unstructured text is used in text mining for pattern discovery while web data is quite in structured form. In-

ccess where
stence of re-
e text collec-
ns which are
grams with-

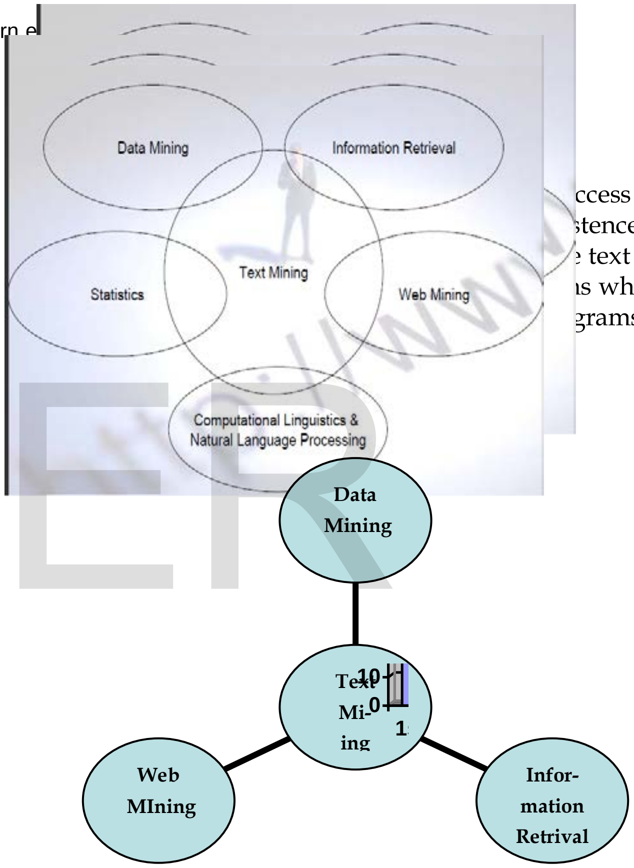

Fig. 1 Text Mining vs. Data mining, IR, Web mining

1.2 Evaluation Matrics

Your Precision = (Number of relevant documents retrieved) / (Total number of documents retrieved)
Recall = (Number of relevant documents retrieved) / (Total number of relevant documents)
Precision is the no. of retrieved documents that are relevant. Recall is the number of relevant documents that are retrieved. An example would be when a search engine returns 50 pages only 30 are relevant. Also it has failed to return 40 additional relevant pages. Precision and recall for the above is given be-

low. Precision = 30/50,
Recall = 30/70

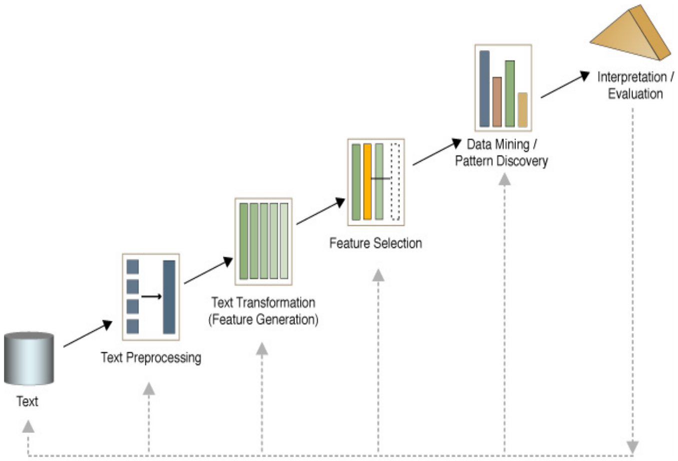
Text
Text Preprocessing
Text Transformation (Feature Generation)
Feature Selection
Data Mining / Pattern Discovery
Interpretation / Evaluation

Fig. 2 Text Mining Process

2 TEXT MINING TECHNIQUES

Th rocess can
hel ious Tech-
niq nformation
Ext gorization,
Clu age, Ques-
tio]. Some of
the

2.1

Be ey phrases
anc

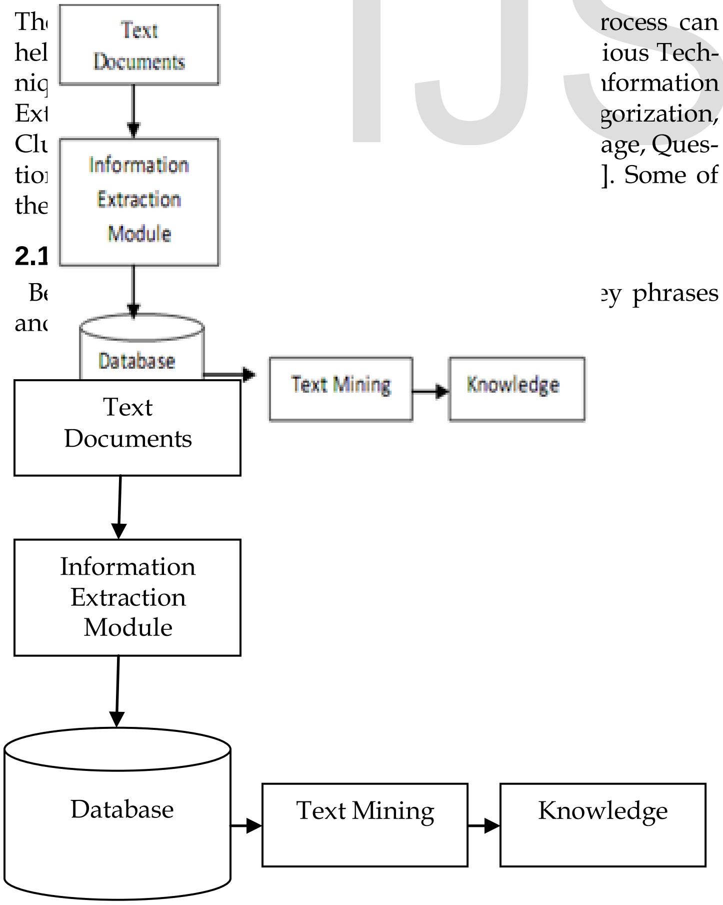

Fig. 3 Information Extraction

Pattern Matching is the base process that is used for Information Extraction which includes extracting the predefined sequences from the text.

Table. 1 Information Extracted

Ratan Tata	Indian Business
Tata Group	Chaieman

Information Extraction is highly valuable when dealing with large volume of text because data in today's world is mainly available in form of electronic documents which has large amount of text.

2.2 Topic Tracking [3][6]

Larger Topic Tracking is text mining technique in which documents of user's interest are presented to user, based on tracking all those documents that user views. This approach can be used in the companies for alert purposes. Various approaches have been used for Topic tracking like Vector Space Model, Hierarchical Clustering, Named Entity Recognition Model, Hidden Markov Model, and Based on Keyword Extraction System etc [10].

While doing Topic Tracking, Test Document needs to be represented and Similarity is calculated with the help of similarity function. This similarity is between topic and story and then threshold comparison is performed. If similarity is higher than threshold value then story is found to be related to topic otherwise not related to topic.

Topic tracking system can be implemented on any textual database for tracking the events. It helps one to keep updated with all the products in the market. It can be used in the medical industry to track the complete situation of patient and what procedure has been followed and what are new treatments. It can also be used for education purposes. In news industry, it is highly valuable technique to find which news articles tracks same events and helps to collect distributed information together.

2.3 Summerization [3][6]

Text Summarization is process of expressing large textual documents into reduced length documents while overall meaning remain same. Various techniques can be used for text summarization like sentence extraction i.e. extracting important sentences from a textual document by statistical calculation for sentences like weighting scheme, TF-ISF (Term Frequency- Inverse Sentence Frequency) further heuristics such as position weighting scheme can also be used for summarization. For example, those phrases which are followed by key phrases like "in conclusion", "at last", "finally", "in the end" etc. depicts the main points of document.

Various methods like statistical, linguistical, heuristic methods are used for text summarization where this system finds how often certain keywords are. Frequency of the keywords is calculated, in which sentence they are present, check for bold text tag etc. This information is helpful to generate summarized view of the original text.

An automatic summarization [11] process can be divided into three steps:

(1) Preprocessing step where a structured representation of the original text is obtained;

(2) Processing step where an algorithm only transforms original text into the summary structure of original document; and

(3) Generation step the final summary is generated.

2.4 Clustering [3][6]

The process in which similar documents are grouped together and dissimilar documents are placed in different cluster.

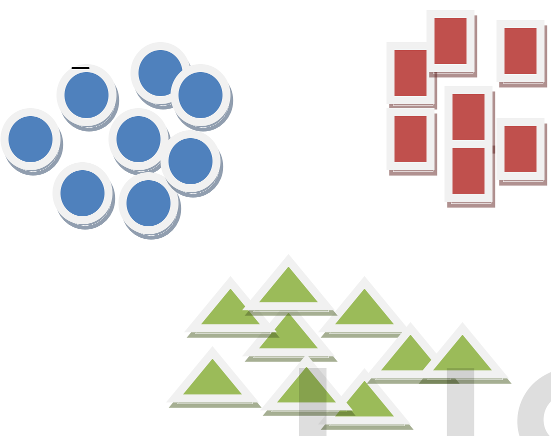

Fig. 4 Document Clustering

2.5 Categorization [3][6]

The Process of placing a document into predefined topic set is called as categorization. It can also be called as classification process. There are two types of document classification.

1. Content based classification
2. Request Oriented Classification (Indexing)

2.5 Assosiation Rule Mining

The Technique by which important association rules i.e. relationships are extracted from large databases is called as Association Rule Mining (ARM). It has been widely used in decision making process in business. E.g. these relationships are currently implemented in various supermarkets where items are placed on the basis of purchasing habits of customers i.e. those items are placed at minimum distance which are purchased frequently [8].

2.6 NLP Based Methods [3][6]

The Natural language processing (NLP) is a modern computational technology that can help people to understand the meaning of text documents. For a long time, NLP was struggling for dealing with uncertainties in human languages. Recently, a new concept-based model [13], [14] was presented to bridge the gap between NLP and text mining, which analyzed terms on the sentence and document levels.

2.7 PTM Based Methods

To overcome the disadvantages of phrase-based approaches, pattern mining-based approaches (or pattern taxonomy models (PTM) [15], [16] have been proposed, which adopted the concept of closed sequential patterns, and pruned non closed patterns. These pattern mining-based approaches have shown certain extent improvements on the ontological term based or purely phrase-based pattern discovering techniques but still it is not a perfect solution to match the user requirement.

3 CONCLUSION

In this paper, we surveyed the list of existing text or pattern mining techniques. We restricted ourselves to the classic text mining problem. Exact pattern from large pattern set which is most suitable for the user. When we analyze various techniques we come to know t large pattern but still the problem remains as it is that how to discover that each technique discover in a future, we pursue the development of a novel algorithm that efficiently discovers patterns from text.

ACKNOWLEDGEMENT

I am thankful very much thankful to my guideMr. Pritesh Jain for his guidance throughout the paper. I also thankful to my Head of Department, Mr. Praveen Bhanodia for encouraging us through out this reaserch paper.

REFERENCES

[1] W. Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smith,”Knowledge Discovery and Data Mining: Towards a Unifying Framework”, KDD-96 Proceedings, 1996.

[2] R Kang Sergio Bolasco, Alessio Canzonetti, Federico M. Capo, Francesca Della Ratta-Rinaldi, Bhupesh K. Singh, “Understanding Text Mining:a Pragmatic Approach”, Roam,Italy,2002.

[3] Chen, Weiguo Fan, Linda Wallace, Stephanie RichZhongju Zhang, “Tapping into the Power of Text Mining”, Journal of ACM, Blacksburg, 2005

[4] Weimin Seth Grimes, ”The Developing Text Mining Market”, white paper, Text Mining Summit05 Alta Plana Corporatopn,Boston,1-12, 2005. . 109.

[5] D Ananiadou, Sophia and McNaught, John(eds), Text Mining, March 2006.

[6] Peng Vishal Gupta, Gurpreet Singh Lehal, “A Survey of Text Mining Techniques and Applications”, Journal of Emerging Texhnologies in Web Intelligence, Vol. 1, No. 1, August 2009.

[7] D Panigrahi, “The Survey of Data Mining Applications And Feature Scope”, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012.

[8] Man Jiawei Han, Michelin Kamber, 2001, “Data Mining Concepts and Techniques”, Morgan Kaufmann publishers, USA, 70-181.

[9] *FLEXChip Kamaldeep kaur and Vishal Gupta, “Topic Tracking “*

[10] Zehua *Punjabi Language”, Computer Science and Engineering: An International Journal (CSEIJ), Vol. 1 No. 3, August 2011.*

[11] Zhang *Farshad Kyoomarsi ,Hamid Khosravi ,Esfandiar Eslami ,Pooya*

Khosravyan Dehkordy and Asghar Tajoddin (2008), "Optimizing Text Summarization Based on Fuzzy Logic", Seventh IEEE/ACIS International Conference on Computer and Information Science, IEEE computer *society, 347-352.*

[12] J. S.-T. Wu, Y. Li, and Y. Xu, "Deploying Approaches for Pattern Refinement in Text Mining," Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1157-1161, 2006.

[13] N. Jindal and B. Liu, "Identifying Comparative Sentences in Text Documents," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '06), pp. 244-251, 2006.

[14] Joachims, "A Probabilistic Analysis of the Rocchio Algorithm with tfidf for Text Categorization," Proc. 14th Int'l Conf. Machine Learning (ICML '97), pp. 143-151, 1997..

[15] J. S.-T. Wu, Y. Li, and Y. Xu, "Deploying Approaches for Pattern Refinement in Text Mining," Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1157-1161, 2006.

[16] S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, "Automatic PatternTaxonomy Extraction for Web Mining," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI '04), pp. 242-248, 2004.