International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-3, Issue-3, May-2015

Text Mining: Process and Techniques

Lata Gohil

***Abstract*— Massive amount of digital data is available in form of unstructured text. It is highly required to extract useful information from this textual data. The process to discover non-trivial information and knowledge which are previously unknown is knows as Text Mining. This paper discusses process and techniques with respect to Text Mining.**

***Index Terms*—Text Mining, Text Preprocessing, Text Transformation, Text Mining Techniques.**

I. Introduction

With the advancement of technologies and its immense use generates massive amount of digital data in form of unstructured text. This free form text contains valuable information and knowledge. To extract knowledge out of unstructured text requires performing mining techniques to the textual data. Text Mining is termed as "the non-trivial extraction of hidden, previously unknown, and potentially useful information from (large amount of) textual data"[1]. Text mining is also known as Knowledge Discovery in Text (KDT), Text Data Mining or Text Analysis. Text mining applies on unstructured or semi-structured data such Text file, pdf file, email, on-line chat, SMS, product review, html file, xml file etc. Text mining applies to unstructured or semi-structured data while Data mining applies to structured data [2].

II. Fields Involved In Text Mining

Text Mining involves various fields such as Computational Linguistic and Natural Language Processing (NLP), Information Extraction, Information Retrieval, Machine Learning and Data Mining.

NLP enables computer system to understand, analyze and generate natural languages. It has various applications such as classify text into categories, index and search large texts, automatic translation, speech understanding, information extraction, automatic summarization, question answering, knowledge acquisition, text generations etc. NLP techniques such as part-of-speech (POS) tagging, word stemming and lemmatization, word sense disambiguation are frequently used by text mining.

Information Retrieval finds document or material from large collection of unstructured/semi-structured documents that satisfies an information need. It also ranks the documents to give best suitable answer to the user's query. Google search engine is an example of application of information retrieval system. Evaluation Metrics of Information Retrieval task are (i) Precision (ii) Recall. Precision is (number of relevant information instances retrieved) / (total number of information instances retrieved). Recall is (number of relevant information instances retrieved) / (total number of information instances). Harmonic mean of Precision and Recall is known as F-measure or F-score. F-score is $2 \cdot (\text{Precision} \cdot \text{Recall}) / (\text{Precision} + \text{Recall})$.

Information Extraction (IE) involves identifying terms or words from the text file [3]. It extracts structured information from unstructured or semi-structured documents. It identifies the named entities, attributes and relationship between entities. It concern with extraction of semantic information from the text. It is used to find related terms from the documents.

Machine Learning is a study of method for programming computer to learn from data and build a model to infer from data. It builds models of two categories: Predictive (for prediction in future) and Descriptive (gain knowledge from data). Machine learning does prediction based on known properties learned from the trained data.

III. Text Mining Process

A. *Text Preprocessing*

Text pre-processing is applied on collection of documents containing unstructured or semi-structure data. Text pre-processing task converts a raw text file into a well-defined sequence of linguistically-meaningful units. It involves following kind of processing.

Text Cleanup

It performs tasks such as removal of advertisement from web pages, removal of tables, figures etc.

Tokenization

It divides sentences into words by removing spaces, commas etc.

Filtering (Stop word Removal)

It removes words that bear little or no content information such as articles, conjunctions, prepositions etc. Words which occur extremely often are also removed. Stemming It is a process of transforming word to its stem (normalize form). It builds basic form of words to identify word by its root. E.g. go is stem of gone, goes, going. Porter's Stemmer is the most popular algorithm and researchers make

Manuscript received May 23, 2015.

Lata Gohil, Research Scholar, CHARUSAT, Changa, India, Assistant Professor, School of Computer Studies, Ahmedabad University, Ahmedabad, India.

changes in the basic algorithm to cater to their requirements [4].

Lemmatization
It converts word to its linguistic correct root, that is, base form of the verb. The process, first understands the context, then determines the POS of a word in a sentence and then finally finds the 'lemma'. Its implementation is difficult because it is related to the semantics and the POS of a sentence. E.g. go is lemma of goes, gone, going, went.

Linguistic processing
It involves Part-of-speech tagging (POS), Word Sense Disambiguation (WSD) and Semantic structure.

Part-of-speech tagging
It determines linguistic category of word. It assigns word class to each token. In English language, there are eight lexeme classes: noun, pronoun, adjective, verb, adverb, preposition, conjunction and interjection. The techniques for POS-tagging are Hidden Markov Model based approaches and Rule-based approaches.

Word Sense Disambiguous (WSD)
It is an activity of finding that given word in a text is ambiguous. e. g. resolving ambiguity in words "bank" and "financial institution". It is the task of automatically assigning the most appropriate meaning to a polysemous word (same form related meaning e.g. blood bank, financial institution) in a given context.

Semantic structure
There are two methods for building semantic structure: full parsing and partial parsing. Full parsing produces full parse tree for a sentence. It often fails due to bad tokenizing, error in POS tagging, novel word, wrong sentence splitting, grammatical inaccuracy etc. Partial parsing or word chunking produces syntactic constructs like Noun Phrases and Verb Groups and it is more commonly used parsing.

B. Text Transformation

It performs feature generation followed by feature selection task. Feature generation represents documents by the words they contain and their occurrences where order of word is not significant. It uses bag-of-words or vector space model. Feature selection is a process of selecting a subset of important features in order to use in model creation. It reduces dimensionality by removing redundant and irrelevant features.

C. Text Mining Methods

There are different text mining methods such as classification, clustering, summarization, topic modeling.

IV. Text Mining Techniques

A. Categorization

Text categorization [5] is the problem of automatically assigning predefined categories to free format text documents. The major difficulty of text categorization is high dimensionality of feature space. The applications of text categorization include document organization, spam filtering, SMS categorization, hierarchical categorization of web pages. Text classification [6] task can be of type supervised, unsupervised or semi-supervised where in supervised document classification, external mechanism provides information on the correct classification for documents, in unsupervised document classification, classification must be done without any external information while semi-supervised document classification, some documents are labeled by the external mechanism.

B. Clustering

Clustering technique uses similarity measures between different objects; it put most similar objects in one class and dissimilar objects in another class. If differs from categorization in that objects are clustered without prior knowledge of classes [7]. The key advantage of clustering technique is objects can pertain to multiple classes. The quality of a clustering result highly depends on similarity measures used by the method and its implementation. A good clustering method produces high quality of clusters with high intra-cluster similarity and low inter-cluster similarity.

C. Summarization

Text summarization is condensing the text into shorter form with retaining its information and overall meaning. It can be classified into abstractive and extractive summarization. An abstractive summarization [8][9] attempts to develop an understanding of the key concepts in the text and then represent those concepts in natural language. It uses linguistic methods to understand, interpret and describe the text in shorter version. Extractive summarization [10] are performed by extracting key text segments based on statistical analysis of features of text such as word/phrase frequency, location or cue words to locate the sentences to be extracted. There are two evaluation measures for text summarization: extrinsic and intrinsic. Extrinsic method measures summarization using task-based[11] performance measure while intrinsic method relies on human evaluation.

V. Conclusion

Text mining is the process of extracting non-trivial information from unstructured text. It is an interdisciplinary field involving computational linguistic and natural language processing, information extraction, information retrieval, machine learning and data mining. Text mining process generates features from unstructured text followed by applies mining techniques to discover knowledge. As most information is stored in form of unstructured text, text mining becomes essential to generate hidden useful information and knowledge.

References

[1] Daniel Waegel. "The Development of Text-Mining Tools and Algorithms", Ursinus College, 2006.
[2] Navathe, Shamkant B. and Elmasri Ramez. "Data Warehousing and Data Mining". In "Fundamentals of Database Systems", Pearson Education pvt Inc, Singapore, 841-872, 2000.
[3] R. Sagayam, S. Srinivasan, S. Roshini, "A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques". International Journal of Computational Engineering Research (ijceronline.com) Vol.2 Issue.5
[4] A. Jivani et al, International Journal of Computer Technology and Applications, Vol 2 (6), 1930-1938.

[5] Yimming Yang and Jan O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization", CiteSeerX, 1997.
[6] Nidhi and Vishal Gupta, "Recent Trends in Text Classification Techniques", International Journal of Computer Applications, Vol.35, No.6,2011.
[7] Liritano S. and Ruffolo M., (2001), "Managing the knowledge Contained in Electronic Documents: a Clustering Method for Text Mining", IEEE, 454-459, Italy.
[8] G Erkan and Dragomir R. Radev, "LexRank,: Graph-based Centrality as Salience in Text Summarization", Journal of Artificial Intelligence Research, Re-search, Vol. 22, pp. 457-479 2004.
[9] Udo Hahn and Martin Romacker, "The SYNDIKATE text Knowledge base generator", Proceedings of the first International conference on Human language technology research, Association for Computational Linguistics, ACM, Morristown, NJ, USA, 2001.
[10] Farshad Kyoomarsi, Hamid Khosravi, Esfandiar Eslami and Pooya Khosravyan Dehkordy, "Optimizing Text Summarization Based on Fuzzy Logic", In proceedings of Seventh IEEE/ACIS International Conference on Computer and Information Science, IEEE, University of Shahid Bahonar Kerman, UK, 347-352, 2008.
[11] Katheleen Mackeown, Ani Nenkova, David Elson, Rebecca Passonneau, and Julia Hirschberge, "A task based evaluation of multidocument system", in SIGIR'05, ACM, 2005.