ISSN PRINT 2319 1775 Online 2320 7876

Research paper Volume 10, Iss 1, 2021

Advances and Applications of handling Text: IR, IE, TM with NLP Techniques

[1]M. Srinivasa Prasad
Department of Library Science, Government Degree College, Narsipatnam, Andhra Pradesh, India.

[2]Y. Jahnavi *
Department of Computer Science and Engineering, Audisankara College of Engineering & Technology (A), Nellore, Andhra Pradesh, India.

***Abstract*— Tremendous growth of text document collection has led to an increased interest in developing various approaches. Analysis and evaluation of useful information or patterns are must from the existing piles of raw data. To extract useful documents/ information/ patterns from the existing large amounts of unstructured retrospective corpora, various approaches such as Information Retrieval, Information Extraction and Text Mining were introduced. Different natural language processing techniques have been developed to improve the accuracy of extracting patterns. While several text handling approaches based on different features have been proposed in the past, there is no systematic study which discusses the similarities and distinctions of all these approaches. This paper discusses the overview of all these approaches, IR, IE, TM and NLP. We discuss about the methods and applications of existing approaches and also the theories in possible future research.**

***Keywords*— Information Retrieval, Information Extraction, Text Mining, Natural Language Processing.**

I. INTRODUCTION

Data Mining or knowledge Mining refers to extracting or mining knowledge from large amounts of data [5]. The data may be either structured or unstructured. Query processing techniques can able extract the data existing in the structured data bases [14]. But to extract the hidden patterns data mining algorithms are used. Data may be any type of data such as Relational, Transactional, object oriented, temporal, spatial, Text, Multimedia, web etc [5]. Text is unstructured, nebulous and convoluted to deal with. However, text is the mainly a frequent medium for the formal exchange of information in many applications such as news articles, research papers, books, digital libraries; email messages, blogs web pages etc. Many applications such as business management, publishing and media, Telecommunications, Information Technology Sector and Internet, Banks, Insurance, Financial markets, Political institutions, Public administration, Pharmaceutical and research companies and healthcare, Market analysis etc., can benefit by the use of relevant documents, information and patterns extracted from large amounts of unstructured raw data[2]. However searching for useful and relevant patterns is an open problem. Various approaches which handle with unstructured data are Information Retrieval, Information Extraction and Text Mining. Information Retrieval Systems are useful for extracting useful documents from the large collections of stored documents [3]. Unlike information retrieval which concerns how to recognize relevant documents from a document collection, information extraction generates structured data organized for further processing [6]. Text mining research area is useful for finding patterns in text collections. These patterns are the extraction of topics from texts or grouping of text or the identification of trends [1]. The main aim of the study of Natural Language Processing Techniques is creation and understanding of natural languages. NLP are useful in the context of disambiguating the ambiguous text by morphological and lexical processing, Syntactic-Semantic structures, and Context processing interpretation.

The rest of this paper is organized as follows. In section 2, 3, 4 some work related to Information Retrieval, Information Extraction and Text Mining are discussed. Similarities and Differences among them are discussed. In section 5, the system framework with the use of Natural Language Processing techniques are presented. Finally, section 6 concludes the entire paper.

II. INFORMATION RETRIEVAL

An information retrieval system as well called as information storage and retrieval system that is proficient of storage, repossession, and maintenance of information. Information in this context can be composed of text, images, audio, video and other multi-media objects. The success of Information Retrieval system is measured by the time needed for finding the user's needed information and this should be minimized. IR is a component of information system that should actively find out necessitate of the users.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper Volume 10, Iss 1, 2021

The crucial steps in the IR process are the document representation and query representation [8]. In the document representation the documents should be preprocessed and the terms in each document are weighted. To expedite the accessing these documents are indexed. The queries are represented by means of Boolean logic, proximity, continuous word phrases, fuzzy searches, term masking, ranking, canned queries etc [3]. These queries are compared with the documents by using different techniques wherein exact match the system finds the documents that fill all the conditions of a Boolean query (it predicts relevance as 1 or 0). To increase recall, the system can exploit synonym expansion and hierarchic expansion [7].

To improve the accuracy of IRS, relevance feedback was introduced by Rocchio in 1965. In this mechanism, the search process often goes through numerous iterations: knowledge of the features that distinguish relevant from irrelevant documents is used to improve the query or the indexing. The new query should be based on the old query modified to enhance the weight of terms in relevant items and lessen the weight of terms that are in non-relevant items [3].The system matches the stored documents with the queries and displays the documents similar to the query. The user selects relevant items based on their interest. The gauge of IRS is done by different evaluation measures such as precision, recall, F-measure etc [9].

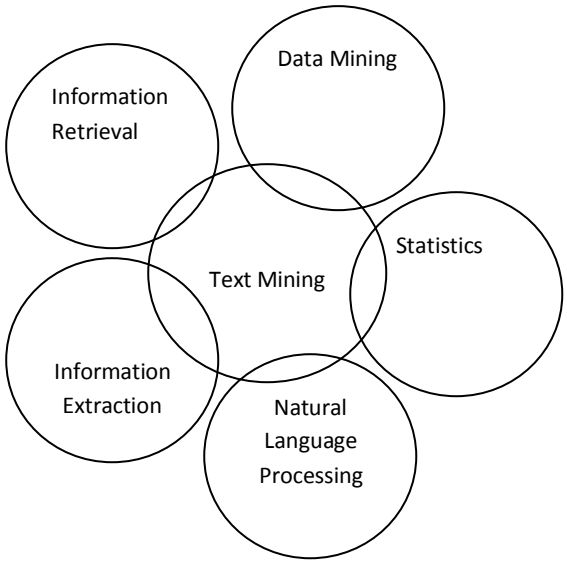
III. INFORMATION EXTRACTION

Information Extraction produces templates from texts. To automate the conversion of input pages into structured data, numerous efforts have been stanch in the area of information extraction. Unlike information retrieval which concerns how to identify relevant documents from a document collection, information extraction produces structured data prepared for post processing, which is vital to numerous applications [6].

In other words, Information extraction automatically converts from unstructured and semi structured text sources into the structured information such as entities, relationships between entities and attributes describing entities. This enables much richer forms of queries on the profuse unstructured and semi structured sources than possible with various keyword searches alone. The extraction of structure such as entities and its relationships from noisy semi structured and unstructured sources is an exigent task [10].

IV. TEXT MINING

These knowledge sources for Text mining are different from traditional data mining [11]. Text mining research area is useful for finding patterns in texts. These patterns are the extraction of topics from texts or grouping of text or the identification of trends [12]. Text classification and clustering are the most important parts of the text mining system. Text categorization automatically organizes the text documents into classes for subsequent analysis. Text clustering is an unsupervised learning process of grouping textual documents whose members are similar [13]. Text mining is opposite to Information Retrieval in the sense, it doesn't base on particular criteria where it will discover some unseen and unknown patterns which we don't know by exploring the corpus [16].



ISSN PRINT 2319 1775 Online 2320 7876

Research paper Volume 10, Iss 1, 2021

Most of the Text based techniques use vector space models in which the documents and the queries can be represented as vectors, which can be used to search their nearest neighbors in a document collection. However, for any nontrivial document database, the number of terms T and the number of documents D are usually relatively large. Such dimensionality leads to the difficulty of inefficient computation, since the consequential frequency table will have size T X D. Furthermore, the high dimensionality also leads to very sparse vectors and raises the difficulty in detecting and developing the relationships among them. To overcome these problems, dimensionality reduction techniques such as Latent Semantic Indexing, Probabilistic Latent Semantic Indexing and Locality Preserving indexing can be used.

A. Latent Semantic Indexing (LSI):

LSI which is based on Singular Value Decomposition (SVD), is one of the most popular algorithms for dimensionality reduction.

There exist different operations of Text Mining such as Automatic discovery of similar words, Simultaneous clustering and Dynamic Keyword Weighting, feature Selection and Document Clustering, Trend and Event Detection, Summarization and Question-Answering.

B. Automatic Discovery of similar words and clustering documents:

Automatic Discovery of analogous words such as synonyms and near-synonyms from different sources of corpora is a subarea of Text Mining. This is also very much useful in Information Retrieval. In this model documents are vectors in term space and terms are coordinators. Thus two terms are similar if their corresponding vectors are close to each other. The similarity between two vectors is computed using any similarity measure such as cosine, Jaccard, Dice, Pearson coefficient etc. Other ways of finding analogous words is using domain specific thesaurus or using the Dictionary.

C. Trend Detection and Tracking:

Topic detection and tracking (TDT) is a research inventiveness concerned with technique to organize news documents. In contrast to the more traditional information retrieval problems, the focus in TDT is on news events: In breaking the text into cohesive stories, spotting something previously unreported, tracking the progress of the event, and grouping together news that discuss the same event. The problem area has also been called event-based information organization.

D. Question Answering:

Question answering systems are designed to find answers to open domain questions in a large collection of documents.

E. Summarization:

Automatic summarization is the creation of shortened version of a text by a computer program .Summarization is a reductive transformation of source text to summary text through context reduction. Summarization techniques are used in search engines such as Google [17].

Types of Summarization techniques:

- Extractive summarization

  It copies the most important information by the system to the summary. i.e., it assigns scores to the sentences and paragraphs of the documents and extracting those with highest scores.
- Abstractive Summarization

  These methods built an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might generate. Such a summary might contain words not explicitly present in the original.

  Abstractive methods are quite weak, so most research has focused on extractive summarization techniques.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper Volume 10, Iss 1, 2021

V. NATURAL LANGUAGE PROCESSING

NLP is the endeavor to extract a fuller significance representation from free text. NLP imposes certain linguistic rules that extract a fuller meaning representation from free text. It naturally makes use of linguistic concepts such as Parts Of Speech tagging (Noun, Verb, and Adjective etc) and grammatical structure (either represented as phrase like noun phrase or prepositional phrase, or dependency relations like subject- of or object -of). NLP is started in the year of 1960 which is used for studying cohort and understanding of natural languages. There are different applications of NLP such as Information Retrieval, Information Extraction, Text Mining (Text grouping, Question-Answering, Language Translation, Opinion mining, Text Summarization et). [15]

There exists different functionalities of NLP such as Morphological and lexical processing, Syntactic Analysis, Semantic Analysis etc.

A. *Morphology:*

Morphology is the study of the way the words are building from smaller meaning bearing unit. Lexicon and Morphotactics are used to build a morphological parser.

Lexicon: Lexicon contains a list of stems and affixes, together
With basic information about them such as whether a stem is a noun stem or verb stem etc.

Morphotactics: The model of morpheme ordering that explains which classes of morphemes can follow other classes of morphemes within a word.

Probabilistic Models of Spelling: The detection and correction of spelling errors is an integral part of modern word processors. Single error misspellings can be corrected with insertion, deletion, substitution and transposition operations. Kernighan used the Bayesian model which is a statistical algorithm do candidate corrections bases on likelihood and prior probabilities. But Kernighan relied on the simplifying assumption that each word had only a single spelling error. Min Edit distance algorithm was introduced by Wagner in 1974 to correct multi error words. The min edit distance between two strings is the minimum number of editing operations (insertion, deletion, substitution) needed to transform one string into other.

Probability of a word depends only on the previous word is called a Markov assumption. A bigram is called a first-order Markov model, a trigram is a second-order Markov model and N-gram is an N-1 order Markov model [15].

B. Syntactic Analysis:

Parts of Speech Tagging, Context-Free grammars, parsing are used for checking the syntax of natural languages. Parts of Speech tagging is also called as word classes or morphological classes or lexical tags or POS tags is the larger class of tags designed for particular corpuses such as Pen Tree Bank or Brown Corpus etc. The significance of the POS for language processing is that it gives a significant amount of information about the word and its neighbors. This type of information is very much useful for the categories of Nouns and Verbs. POS can also be used in stemming for IR, since knowing word's POS can help us which morphological affixes it can take. It can also help an IR application by helping select out nouns or other important words from the document. It is also used in Word Sense disambiguating algorithms. POS are very much used for Partial Parsing texts and other phrases for the information extraction. It is used in linguistic research for example to help find instances or frequencies of particular construction in large corpora.

C. Semantic Analysis:

For checking the semantic structure of natural languages different techniques such as augmented Context Free Grammars, Lexical Semantics, and Thematic Roles etc are used [12].

ISSN PRINT 2319 1775 Online 2320 7876

Research paper Volume 10, Iss 1, 2021

VI. EVALUATION

The basic rationale of Information Retrieval System's (IRS) research area is for retrieving relevant information, where the information may be composed of text, images, audio, video etc. Natural language processing techniques are useful in IRS to improve the accuracy if the information is Text and the purpose of Information Extraction is exclusively different. Text Mining is the extraction of interesting and useful patterns in textual data. NLP techniques such as morphological, syntactic, semantic and pragmatic analysis support text classification, clustering algorithms, summarization, Question-Answering, anaphora resolution etc. There exists different text classification and clustering algorithms such as Naïve Bayes, Support Vector Machines, Hierarchical clustering, K-Means etc. All these NLP techniques are useful to remove indistinctness and ambiguity hidden in text documents. Various classification algorithms have been proposed on different kinds of data to extract useful patterns [18-25].

VII. CONCLUSION

In this study the discussion of Information Retrieval, Information Extraction, Text Mining, and Natural Language Processing techniques are discussed. Even though a large research has evolved into these areas, automatically extracting relevant documents and patterns is still an open area. As part of the ongoing and further research the preliminaries, techniques and evolution relationships of each interlinked disciplinary are illustrated. This paper can be used as a platform to impose new algorithms and ideas. Context based approaches bridges the gap between Text Mining, and Natural Language Processing. NLP resolves the semantic relationships between the terms within the corpus and thus improves the accuracy of relevant information and patterns.

REFERENCES

[1] S. M. Indurkhya**,** N. Zhang, T. Damerau, F. Weiss, "Text Mining Predictive Methods for Analyzing Unstructured Information", Springer, 2005.

[2] Sergio Bolasco, Alessio Canzonetti, Francesca Della Ratta-Rinald and Bhupesh K. Singh, "Understanding Text Mining: A Pragmatic Approach", Roam, Italy, 2002.

[3] J. Kowalski and Mark. T. Maybury, "Information Storage and Retrieval Systems Theory and Implementation", 2nd Edition, Springer publications.

[4] D. Sanchez, M.J. Martin-Bautista, I. Blanco, "Text Knowledge Mining: An Alternative to Text Data Mining", IEEE International Conference on Data Mining Workshops, pp. 664-672, 2008.

[5] H. Jiawei and K. Micheline,"Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers,Elsevier, vol.2., 2006.

[6] Chia-Hui Chang, Mohammed Kayed, Moheb Ramzy Girgis, Khaled Shaalan," A Survey of Web Information Extraction Systems", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, TKDE-0475-1104.R3.

[7] Amit Singhal, "Modern Information Retrieval: A Brief Overview", IEEE,2001.

[8] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, "An Introduction to Information Retrieval", Cambridge University Press, April 1, 2009.

[9] R Jizba, "Measuring Search Effectiveness", creghton.edu, 2007.

[10] Sunita Sarawagi," Information Extraction", Foundations and TrendsR in Databases Vol. 1, No. 3 (2007) 261–377.

[11] Weng, S.S. and Y.J. Lin, "A study on searching for similar documents based on multiple concepts and distribution of concepts", Expert Syst. Applications, 2003, 25:355-368.

[12] A. Kao, S. Potect, "Text Mining and Natural Language Processing Introduction for the Special Issue", SiGKDD Explorations, 2004, pp. 1-3.

[13] Wang, F., C. Zhang and T. Li, "Regularized clustering for documents. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval", Amsterdam, Netherlands, July 2007, 23-27, pp: 95-102.

[14] Y. Shiqun, W. Gang, Q. Yuhui and Z. Weiqun, "Research and implement of Classification Algorithm on Web Text Mining", IEEE, Third International Conference on Semantics Knowledge and Grid. 2007, pp. 446-449.

[15] Daniel Jurafsky, James H. Martin, "Speech and Language Processing", An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, University of Colorado, Boulder.

[16] Michael W. Berry, "Survey of Text Mining, Clustering, Classification and Retrieval", Springer, 2003.

[17] Shanmugasundaram Hariharan," Extraction Based Multi Document Summarization using Single Document Summary Cluster", Int. J. Advance. Soft Comput. Appl., Vol. 2, No. 1, pp. 1-16, March 2010.

[18] Jahnavi Yeturu, "Statistical data mining technique for salient feature extraction", Int. J. Intelligent Systems Technologies and Applications (Inderscience Publishers), Vol. 18, No. 4, 2019.

[19] Jahnavi Yeturu, "A Cogitate Study on Text Mining", International Journal of Engineering and Advanced Technology", Vol. 1, No. 6, pp. 189-196, 2012.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper Volume 10, Iss 1, 2021

[20] Jahnavi Yeturu, "FPST: a new term weighting algorithm for long running and short lived events", Int. J. Data Analysis Techniques and Strategies (Inderscience Publishers), Vol. 7, No. 4, 2015.

[21] Jahnavi Yeturu, "Analysis of weather data using various regression algorithms", Int. J. Data Science (Inderscience Publishers), Vol. 4, No. 2, 2019.

[22] Jahnavi, Y., and Y. Radhika. "Hot topic extraction based on frequency, position, scattering and topical weight for time sliced news documents." 15th International Conference on Advanced Computing Technologies (ICACT). IEEE, 2013.

[23] Sukanya, G et al., "Country Location Classification on Tweets." Indian Journal of Public Health Research & Development 10.5, 2019.

[24] Bhargav, Kanta et al., "An Extensive Study for the Development of Web Pages." Indian Journal of Public Health Research & Development 10.5, 2019.

[25] Lakshmi, Mutyala et al., "Security Health monitoring and Attestation of Virtual Machines in Cloud computing." Indian Journal of Public Health Research & Development 10.5, 2019.