Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 3, Issue 4; April-June, 2016, pp. 288-293
© Krishi Sanskriti Publications
http://www.krishisanskriti.org/Publication.html

The Study of Text Mining and Knowledge Extraction (TAKE) for Textual Database

Anu Yadav

University School of Information & Communication Technology, GGSIPU, New Delhi
E-mail: noor4anisha@gmail.com

Abstract—*In recent decades, Internet technology becoming a world of ICT (information and communication technology), it is universally adapting technology & expanding continuously into private, and public sector for better future life changes. As internet expanding, the amount of information increasing and get doubles in every 24 months. In shortest time, the user demand for reliable and useful information, but there are difficultly to fulfil such user demands due to language barrier, unclear specification for information and diversified users. It need to perform the searching and extraction process for exact knowledgeable information from the collection of natural language texts. Hence text mining becoming an important and hot research area. The text mining is a process of discovering the knowledge and patterns of data from previously unknown and unstructured data or new information from different resources of information by using technology of computers. In this paper, we are going to discuss Text mining And Knowledge Extraction (TAKE) in detail and its processes and application areas. The text mining is a type of data mining and it is a discovery of knowledge from textual database that extract interesting or non-retrieval knowledge from unstructured texts. The text mining techniques categorizes into the Information extraction, Clustering, Categorisation, Summarisation, Classification, and Visualization. Text mining Applications including marketing, business intelligence, research, media, publishing, healthcare, communications.*

Keywords*: Text mining and knowledge extraction.*

1. INTRODUCTION

Every day the innovation of computer technology expanding the amount of information. Everyone need useful, reliable exact and access of information. The retrieving useful knowledge from different kinds of huge number of documents like unstructured, semi-structured & structured is very tedious task. To discover information from an unstructured document we used TAKE, it is an inspiring and useful research area. Mining text from unstructured data is challenging with statistical modeling, natural language processing (NLP), and machine learning techniques because natural language text have ambiguities, inconsistent syntax & semantics and double meaning. A hot research area of computer science is TAKE it solve the issues that occur in the area of IE (information extraction), IR (information retrieval), data mining, machine learning, Computational Linguistics, NPL, Artificial Intelligence, Statistics and classification of knowledge discovery and management. TM (text mining) also known as Text Data Mining (TDM) and KDTD (Knowledge Discovery in Textual Database). TAKE includes 13 different types of tasks, to handle different sources of information, and discover useful knowledge from them. These tasks are IE, categorization of text, topic detection, summarization of documents, NLP, web mining, concept extraction, text-based navigation, search & retrieval, document clustering, trends analysis and association analysis, and visualizations.

2. TAKE

TAKE (Text mining And Knowledge Extraction) is a process of identifying, extracting and analyzing novel as well non trivial information and finding interesting patterns or trends from a huge collection of words or texts present in different types of documents for analysis purpose. It discovers information from new or previously unknown data and extracting information from huge collection of different unstructured (naturally occurring text) textual resources by using computer technology.

2.1 TAKE VS others

The difference between TAKE and DM is, DM (Data Mining) extract interesting information, patterns and trends from structured data, but 90% of data is a form of unstructured (implicit structure e.g. grammatical structure). In TAKE the patterns are extracted from natural language texts rather than from structured databases (structuring of the data into predefined fields and facts). DM is a standard data mining and TAKE is a real text data mining. The difference between WM (web mining) and TAKE is WM mine web sources that are mostly in structured data formats whereas in TAKE input is free unstructured data.

The difference between information access /IR and TAKE, IR don’t extract new information, whereas TAKE discover information from new or previously unknown data. The fig 1 showing the difference between DR, IR, DM and TAKE.

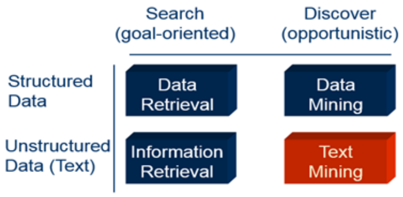

Fig. 1: The difference between DR,DM,IR and TM

2.2 Need of TAKE

Drastic increase in data via the web and 90% data is in an unstructured format. Hence for systematic literature review, discovering knowledge, research in computational linguistics and patterns enrich of relevant content we use TAKE. The Real E.g.: the bio industry, 85% of biological knowledge are only in a research paper (unstructured) if a scientist manually read 50 research paper/week and only 10% of those data are more useful than he/she manages only 5 research paper/week. But online database like Medline adds more than 10000 abstract per month using TAKE. Thus the performance of gathering relevant data is increased dramatically when we use TAKE.

2.3 TAKE as Interdisciplinary field

A hot research area of computer science is TAKE, it is an interdisciplinary field which incorporates and solve the issues that occur in the area of IE (information extraction), IR (information retrieval), data mining, machine learning, NPL (natural language processing), Computational Linguistics, Artificial Intelligence, Statistics and classification of knowledge discovery and management. TAKE uses data mining techniques to discover trends/ patterns or information from unstructured textual data or we can say by the help of TAKE DM mine all kinds of data. The fig 2 shows that TAKE also requires techniques from other fields like DM for Text clustering and classification, DB (database) for IR, Computational linguistics for NLP, artificial intelligence & statistics for NLP and IE and web mining is a part of TM.

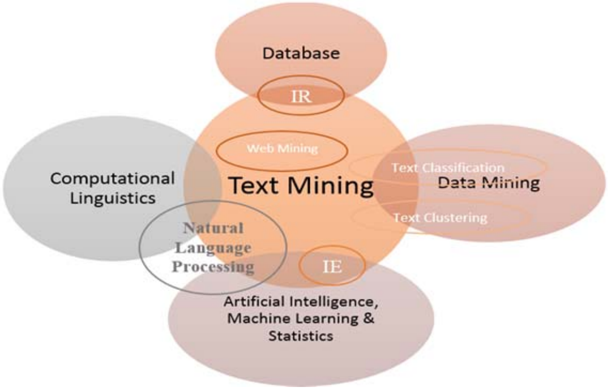
Database
IR
Web Mining
Computational Linguistics
Text Mining
Text Classification
Data Mining
Text Clustering
Natural Language Processing
IE
Artificial Intelligence, Machine Learning & Statistics

Fig. 2. TAKE is an interdisciplinary field

3. FRAMEWORK OF TAKE

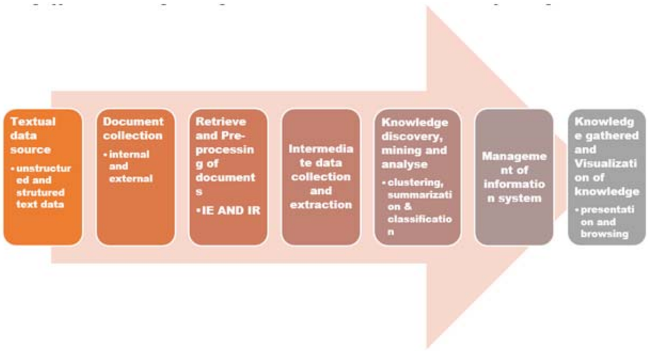
Textual data source
• unstructured and strutured text data
Document collection
• internal and external
Retrieve and Pre-processing of documents
• IE AND IR
Intermediate data collection and extraction
Knowledge discovery, mining and analyse
• clustering, summarization & classification
Management of information system
Knowledge gathered and Visualization of knowledge
• presentation and browsing

Fig. 3. The stepwise framework for TAKE

The different components/stages in TAKE framework are shown in Fig. 3. The first step is textual data sources, it is providing all kinds of data such as, unstructured (natural language data), semi structured (a data that is a combination of structured data and unstructured data) and structured data (database, not free text) that available by using next step that is document collection which provide textual data available in different types of internal and external data source such as research papers, emails, corporate documents, web, social media, application form, electronic text, etc. The third step is retrieving and preprocessing of documents textual data, it is a process that collects data from multiple data sources and extract stop words, stemming words, tokenization of words, etc. the preprocessing is discussed in detailed in section 3.2. The preprocessing use different techniques and transform unstructured/semi structured, raw, original data into intermediate structured data that used in the fourth step. The next step is intermediate data collection and extraction that used to extract noisy text (spelling error, missing punctuation, repetitions, abbreviations, misleading/false information) by text cleansing method, tagging of POS (Part-of-speech) or grammatical tagging used in NLP applications (speech recognition and IR), syntactical parsing (analysis words according to rule of grammar such as constituency or dependency), IE (key words and relationship in different textual data). The relevant and meaningful information collected from this step is stored in data repositories for KDMA. The extraction process that applied are discussed in section 3.3 task of TAKE, these processes are topic tracking (track information of (by government) and for user (by public)), summarization (summary of detailed text), feature or term selection from data, Name entity extraction (specific subject), categorization (text classification), concept extraction, theme extraction, clustering (object organizing and grouping into one based on specific similarity), etc. The next step is KDMA (Knowledge Discovery, Mining and Analyze) in this process operations applied on collected structured intermediate data. The intermediate data extracted information based on relationships, but KDMA again apply techniques or algorithms to extract useful information, trends and patterns from extracted data. The algorithms used in KDMA are SVM, Genetic algorithms, neural, decision trees, associations, K-mean, KNN etc. The pattern mining techniques are applied for finding the text patterns like frequent term sets, & co-occurring terms. The next step is management of information system, in this step we have to manage all information gathered from early steps. The different types of algorithm (genetic, neural, KNN) applied on information extracted, they produce different results. The management of these result is conducted in this step. The last step is Knowledge gathered and visualization or presentation of knowledge, here we have to merge all information collected till now from raw data. The different algorithms applied on information are grouped at one place and combine it into a single a pieces of relevant information. Visualization such as browsing functionality (content based or dynamic) and facilities provided to user which represent graphical representation of pattern discovered and their concepts in a hierarchical form and help in analyzing the information or investigating about something. Presentation provides drill down and merging options. Visualization tools clear the concepts by reducing complexity of information and relationship between information. Here a user can't see what algorithm applied, but result presented for the user request in interactive graphical representation.

3.1 TAKE Process

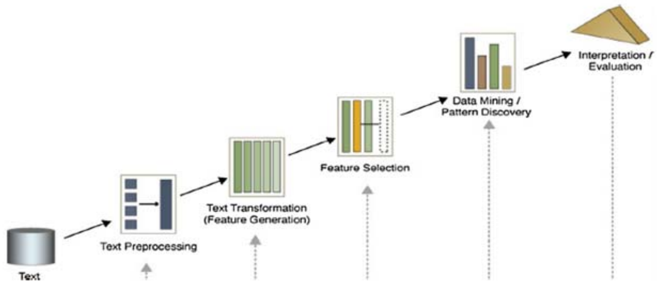
Interpretation / Evaluation
Data Mining / Pattern Discovery
Feature Selection
Text Transformation (Feature Generation)
Text Preprocessing
Text

Fig. 4. The stepwise process of TAKE

There are 6 steps in TAKE process to collect, extract and produce an outcome, these are shown in fig 4.

The first step is "Text" hierarchy structure shown in fig 5, in this we have to describe the characteristics of text and document clustering. The characteristics of text are six types. The first type is "Many Input Modes ", in this text collected from different medium such as text collected from different languages by using human, and in different formats (structured/ semi structured/ unstructured / html/ web). The second type is a dependency of text (phrases/words produce context for each other). The third type is ambiguity, it is subdivided into two parts word ambiguity and sentence ambiguity. The fourth type of text characteristic is noisy data it is important task because it can produce wrong or fake knowledge, it is subdivided into two parts misleading text and erroneous data. The fifth type of text characteristics is unstructured text that produced by natural language example, such as social media, chat room, and normal speech. The last part of text characteristics is high dimensionality or sparse input. The document clustering (unsupervised learning) is the second type of text characteristics. As we known the amount of data increases every day and TAKE is a process of extraction from large volumes of document. In this step we used to combine different and a huge number of documents volumes in an efficient manner as we don't known at this point which document is useful and which useless for our requirement. We use document clustering by applying K-Means clustering and hierarchical clustering.

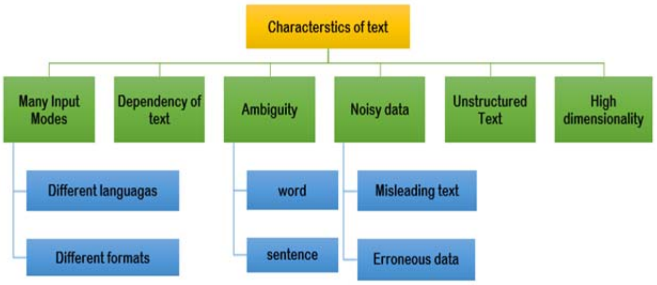

Fig. 5. The hierarchy of characteristics of text in TAKE.

The second step is text preprocessing which included text cleanup, tokenization, semantic, disambiguation, stopword removal, stemming, and POS tagging. In section 3.2, it is described in detailed there.

The third step is Text Transformation or attribute/feature generation, it is divided into three sub parts Text Representation, Feature Selection and actual attribute generation. The representation of text is a text representation using document by the use of words / phrases or features they contain and according to their occurrences. It is represented by approaches, these are vector space and "bag of words". The feature selection is a selection of subsets of features that represent the different text documents. In other words, it describes and represent which features are best features to characterize the particular document. It also used to improve text representations by showing that this feature contains less information. We used feature selection in TAKE because it is used to represent document containing little or no information, stop words, misleading words, redundant features, and assumptions. It is showing number of features and dimensionality of documents by using two types of algorithms Porter's Algorithm and KSTEM Algorithm. The methods that used by feature selection these are before using classifier select features (require feature ranking method) and how well it works in classifier based on its select features. The feature selection by simple counting or by using statistic. The third sub part is actual attribute generation it uses classifier to generate attributes/labels automatically from the feature selection process.

The fourth step is attribute selection this is sub divided into two parts these are Reduce Dimensionality and Remove irrelevant attributes. In this step we further reduce the dimensionality because high dimensionality produces difficulty for users. Here we deduct and remove irrelevant features and resolve feasibility and Scarcity of resources issues of attributes.

The next step is pattern discovery and DM (data mining), in this we apply traditional DM techniques (with TM process) on structured data that collected from the above steps. This step is only an application dependent stage.

The last step is Interpretation and Evaluation, it terminates or iterate. Here the result is generated into 3 parts: outcome is well suitable for application using for TAKE so iterate it, outcome is not satisfactory, but significant and the last is outcome generated are used as input for earlier stages.

3.2 Preprocessing of TAKE

The TAKE from a preprocessed text is quite easy rather than directly from documents in natural languages. Hence, before applying TM techniques, preprocessing of multiple documents from different sources is required. The Text preprocessing in TAKE basically include 9 steps as shown in Fig. 6 but mainly 3 steps are used for preprocessing these are tokenization/ extraction, Stopword removal and stemming, rest all steps are for helping in the preparation of text preprocessing such as different types of documents collecting data from multiple sources for preprocessing.

Extraction of unrequired things is done in text normalization/tokenization. The tokenization is a method of splitting sentences/statements present in documents or sequence of strings into pieces like words by removing spaces/commas, keywords, symbols, and phrases these called as tokens (individual words/phrase). White spaces & punctuation marks are discarded in tokenization process.

The removal/ elimination of Stopwords is required because 75% words occurring in natural language documents are useless and don’t carry information for knowledge discovery such as a, after, again, all, also, always, an, and, any, are, as, at, be, because, become, before, between, both, but, by, can, cannot, could, do, due, each, else, every, except, find, for, from, give, has, hence, how, if, in, is, it, last, less, many, more, most, must, no, not, nothing, now, of, on, only, or, other, our, please, rather, same, see, so, some, such, take, than, that, the, their, then, there, these, they, this, to, too, un, was, we, what, when, where, which, while, who, why, will, with, without, you. It also eliminates tags such as HTML/XML from web pages.

The indexing & searching system support word stemming, it is used in TAKE application, NLP & IR. The stemming is a process that finds root/stem of the word and reduce word into root/stem, in other words, it is a process of eliminating prefixes and suffixes from root/stem word. Example disagrees, disagree, agreed, agrees, agreeing, disagreement, & agreement converted in agree (root word).

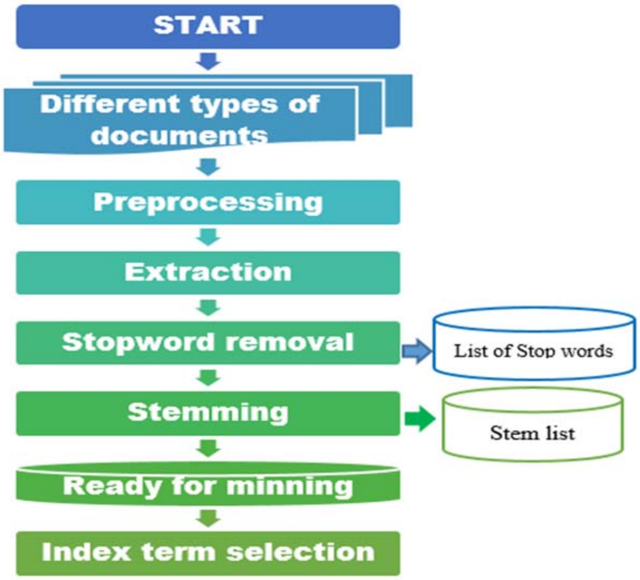

Fig. 6: Stepwise showing preprocessing activities

3.3 Tasks of TAKE

TAKE includes 13 different types of tasks to handle different sources of information, find non trivial information from huge collection of documents and discover useful knowledge from them. These tasks are:

The IE (information extraction) is in the process of identification (in terms of key phrases) and extracts relevant information (facts) and relationships from unstructured text and converting unstructured/semi structured data into structured text. It used to find relevant information from documents. The Summarization of documents is a process of collecting different kinds of documents from different sources (external or internal) and summarizing their knowledge in a mannered way. The Topic detection is a process which search for a particular topic or area such as IEEE papers based on topic extraction. The Concept extraction is a process which group phrases and words into semantically similar groups in an unstructured data. The Search and retrieval is the process that used to storage and retrieve text documents (search engines and specific keywords search). The strategy used to answer the questions for the specific needs of users and providing relevant information. The Web mining task used the huge amount of data available on internet for mining with the help of DM and TM with a specific focused scale. The NLP (natural language processing) is a low level language and understanding speech, it often used with artificial intelligence, machine learning, statistics and computational linguistics. The Text-based navigation systems it used to find the related terms discussed in some specific documents and finding the relationships between them. The clustering of documents is the process of grouping or clustering the snippets, paragraphs or documents using DM clustering methods. It used to find knowledge or hidden information from documents by evaluating the similarities between the documents. The Categorization or classification of text is a process of grouping based on similarities and then classify into category of snippets, paragraphs or documents using data classification methods such as decision tree, genetic algorithm, and neural networks. The Summarization is a process that reduces the content of documents and providing readable material to user without changing the meaning of content. The Trend Analysis & Association Analysis are used to find trends or predict patterns that can used in future, based on data that dependent on time and merge/associate these patterns into another pattern. The Visualization is a task that defined as representation of extracted knowledge or features and help in identifying relevant according to their importance on the representation. It is also used for discovering the location of the content in graphical representation.

4. APPLICATIONS OF TAKE

There is a range of applications in medical like Biomedical, bioinformatics, pharmaceutical, healthcare and research companies used Text mining And Knowledge Extraction. TAKE application in biomedical literature such as PUBGENE is an online application, for search and navigation tool used TPX that run on PUBMED to analyses biomedical literatures, knowledge search tool from biomedical texts is GOPUBMED, and MEDMINER used to extract detail of diseases. Internet increases user participation in web, user search doctor and hospital services nearby their location, information regarding specialist field, location, and availability, e-consultations and for understanding of specific diseases with precautions, new therapies, medical advice, and home remedy. Health24x7 site on the web used for, such as medical facilities, medicines, & consultants. In Bioinformatics, increase research work for IE (information extraction), journal and articles in biomedical motivate biologists to work in their field due to expand in a number of publications and they help in updating the relevant recent literature. The 90% drug derived from genomics literature. The aim of TAKE is to allow biomedical researchers to extract knowledge from the biomedical literature regarding new discovery. MEDLINE and MEDMINNER tools used to records such as title, an abstract, and metadata terms. The main concern in business is to reduce the amount of guessing work and extracting the content for decision making. The TAKE used to reduce risk of wrong predictions. The data mining techniques designed to deal with prediction but here exist problems and they predict up to a certain point, hence we use TAKE. Organization and information of company grow with time and changes occurred in company, hence in business intelligence, corporate knowledge used to find value in historical records, to identify trends in datasets and new innovation used disruptive technology, for regular update from growing large dataset used systematic literature review and harvesting of information such as industry sector, location, contact details and recent activities of company extracting for users used by market analysis. TAKE in Market Analysis is used mainly to monitor customer’s opinion, where the data sources are present for analysis, analyzing about competitors, telemarketing by email to acquire new customers and determine the image of organizations by analyzing press reviews, and customer’s feedback. It used to model customer purchasing pattern over time and customer cluster who share the same characteristics like expenditure, discount, coupons and interest. Cross-market analysis also used TAKE for Associations or co-relations between different product sales in different locations. The business analytics applications concentrate on customers and areas like CRM (customer relationship management), customer service quality, customer reviews and customer experience management. CRM used to handle user request for specific product service or supply queries (contact / call center or by email). IBM offer Customer Relationship Intelligence application based on the Intelligent Miner for Text product, it is designed to help companies for better understanding about what their customer needs and their review about the company. The use of TAKE in national security applications (defense) domain is compulsory for monitoring, study of text encryption/decryption and analyzing activities on internet like news, blogs, emails, and chats for national security purpose. The government agencies are also investing in security and considerable resources. The other applications of TAKE are online media/ social media monitoring and publishing, telecommunications and energy industries, Information Technology sector and Internet (web mining), Banking, insurance and financial markets, Political institutions & analysis, public administration and legal documents, monitor database and restore/backup, search engine and information access, Call Center Software, Anti-Spam, exploring text, extract entities, index files, question answer, email archive, Wordnet dictionary, parts of speech tagger, E-Discovery, Records Management, Scientific discovery, Sentiment Analysis Tools, Listening Platforms, Natural Language and Semantic analysis toolkit (Wordnet), Automated ad placement, Software applications for indexing/tracking, Academic applications, Spam filtering, Creating suggestion/recommendations (like amazon), Automatic labeling of documents in business libraries, Measuring customer preferences by analyzing, Fraud detection by investigating, Fighting cybercrime, resume filtering from thousands of job applicant resume, Customer profile analysis, Information filtering and routing, Event tracks and analyze billions of documents in different languages.

5. MAIN CHALLENGES IN TAKE

The information available is always not in a well-organized way, it is in an unstructured textual form and data is a collection of words in Natural Language so it is a free text with complexities. Hence, the major challenging issue in text mining and knowledge extraction is the Word ambiguity and context sensitivity of alphabets, words and sentences. Word ambiguity refers as capability of understanding same thing in more than one possible way by human beings or one word has multiple meanings or definitions. In text data ambiguity means same sentence in different way interpreted for different meanings. Example of ambiguity bike, vehicle, automobile and pulsar all words are used for the same thing and the example for sensitivity is apple is Cell Phone Company whereas apple is eatable fruit. Researchers try to solve it, but unsuccessful due to usability, consequently, easy understandability and flexibility in natural language we can't eliminate ambiguity from natural languages. Information is expanding every day due to the internet and computer technology, there exist a large textual database and collections of documents. Dealing with such a large data for text mining there is a need to collect and extract all data that is another issue in TAKE. Noisy data, such as spelling mistakes in already written documents can't be corrected, but the knowledge extracted from such data also containing mistakes. There is not a readily access and uniform accessible for all sources because every source has separate storage like email, web and applications. TAKE required skilled, experienced and trained specialists of a particular knowledge domain that required for selection, mining and analysis of data to get output. Hence, TAKE products, tools and applications are designed for trained knowledge expert. The TAKE required more time and Cost for creating, maintaining and updating tools and product of TAKE. Methods for semantic analysis are computationally more expensive (the order cost for a few words per second).

6. TOOL USED FOR TAKE

The two types of tools used in TAKE one is general commercial and open source TM tools. The tools provide high level of overview of SOLD (strengths, output, limitation and source) TM capabilities with strengths, data source that are applicable and relevant, limitations, and result in terms of output.

Few examples of General commercial tools are Angoss (categorization, summarization, sentiment analysis, & theme extraction), Autonomy (clustering, categorization & pattern recognition), Basis (uses artificial intelligence techniques), Cogito, Inxight(SAP), DiscoverText, IBM SPSS (text analytics tool) and SAS Text Miner.

Few examples of Open source tools are GATE, R language, OpenNLP, carrot2 and RapidMiner

7. CONCLUSION

To discover information from an unstructured document we used TAKE, it is an inspiring and useful research area. There is a range of TAKE applications in medical, bioinformatics, research, telecommunications and energy industries, Information Technology sector and Internet (web mining), Banking, online media/social media monitor and publishing, insurance and financial markets, Political institutions & analysis, public administration and legal documents, monitor database and restore/backup, search engine and information access, etc. The major challenging issue in TAKE is the ambiguity and context sensitivity of alphabets, words and sentences. It is still emerging field.

REFERENCES

[1] Vijay Kumar, Manish R., Priyanka M.: "Text mining and information professionals: Role, issues and challenges", *Emerging Trends and Technologies in Libraries and Information Services (ETTLIS),* 2015, Pp. 133-137.

[2] Yu Zhang; Mengdong Chen; Lianzhong Liu, "A review on text mining", *International Conference on Software Engineering and Service Science (ICSESS),* 2015, Pp. 681 – 685.

[3] D. Sánchez, M. J. Martín, I. Blanco; C. Justicia de la Torre, "Text Knowledge Mining: An Alternative to Text Data Mining", *IEEE International Conference on Data Mining Workshops,* 2008, Pp. 664 – 672.

[4] A. Akilan, "Text mining: Challenges and future directions", *International Conference on Electronics and Communication Systems (ICECS),* 2015, Pp. 1679 – 1684.

[5] Gupta, V., "A survey of text mining techniques and applications", *Journal of Emerging Technologies in Web Intelligence, 1(1) (August),* 2009.