REVIEW ON TEXT MINING TECHNIQUES, TOOLS AND DATA BASE

[1]R. Nandhini M.Phil.
Department Of Computer Science,
Government Arts College (Autonomous),
Coimbatore-18.
Mail: nandhudpm92@gmail.com

[2] Dr. R. Roseline M.Phil, Ph.D.,
Head of the Department,
Department of Computer Applications,
Government Arts College (Autonomous),
Coimbatore-18
Mail: roselineera@yahoo.com

Abstract — In this paper survey on text mining techniques such as Natural Language Processing (NLP) Based Techniques, Property function based Techniques, Rule based techniques, Semantic analysis based techniques, Neural Network based techniques, and Visualization techniques. Further text mining tools used for various applications are elaborated. The databases used for text mining are Bio Rat, EFIP, FACTA+, Hit Predict, Nagatome, Open DMAP, Poly search, PP Look, STRING, IMID, PPI Extractor, and Gene Ways also reviewed. Text mining techniques include categorization of Text, Summarization, Topic detection, Concept Extraction, Search, Non-Retrieval, Document Clustering, etc. [1]. Each of these techniques can be used in finding some non-trivial information from collection of documents.

Keywords- *Techniques, Tools, databases, Data mining, Knowledge Discovery, NLP*

1 INTRODUCTION

Text mining can be employed to detect a document's main topic/theme which is useful in creating taxonomy from the documents collection. Areas of applications for text mining included Publishing, Media, Telecommunications, Marketing Research, Healthcare, Medicine, etc. Text mining has also been applied on many applications on the world-wide web for developing recommendation system. Knowledge about data or text mining from the important and relatively large database recognized by numerous scholars and researcher. Data Mining or knowledge discovery, work well on data stored in a structured manner. Often, data that has not been well structured yet still contain a lot of hidden information. Text mining entails automatically analysing a corpus of text documents and discovering previously hidden information. The result might be another piece of text or any visual representation. Text mining generally includes categorization of information or text, clustering the text, extraction of Entity or Concept development and formulation of text taxonomies. Text mining deals with unstructured or textual information for the extraction of meaningful information and knowledge from huge amount of text. They required for the efficient analysis and exploration of information available on text form. Text mining is required to convert the text into data which then pass through other data mining techniques for analysis. Most of the times data gathered from different sources are so large that cannot read it and analyse by manually, so text mining techniques to deal with such data. Text mining includes statistical, linguistic and machine learning techniques that are needed for studying and examining textual information required for further data analysis, research and investigation. Text mining in different domains such as

- Web Document Based Text Clustering [2].
- Information Retrieval [3].
- Knowledge Transfer and Integration [4].
- Topic Tracking [5].
- Summarization, Categorization, Clustering [6].

2. TECHNIQUES

2.1 Natural Language Processing (NLP) Based Techniques

The NLP is a text mining methodology that habits computational means to explore besides symbolize on the way to textual data existent in present ongoing brochures. Now patent analysis, the NLP consumes likewise remained recycled on behalf of the makeover of the high-tech data keen on guileless morphological buildings by take out the structural buildings from as of the documented data and developing the structural correlation in the middle of the modules [13]. The NLP established text mining attitudes stand fundamentally classified into:

- Keyboard Based Approaches.
- Subject – Action – Object (SAO) Based Approaches [14**].**

2.1.1 Keyword Based Techniques [KBT]

Keyboard text mining methodologies are humble to appliance, they surplus in representation of important technological concepts and relationships. The keyword based approach involves predefining keywords and key phrases that require expert knowledge [15]. The SAO support text mining techniques are able to evaluate formless information by the associations in the middle of key technological mechanism.

2.1.2 Subject Action Object [SAO]

It agrees to the illustration of the notion in the arrangement of trouble – answer in addition to be regularly support lying on TRIZ [16]. The NLP based come close to experience since the issue of lexical and grammatical and also lack in representing the semantic relational ship among the grammatical structures. The text mining engine uses Pipelines, such as Title Pipeline, Abstract Pipeline, Patent Claim Pipeline and Detailed Description Pipeline. The final result is attain by manipulative the Confidence Index (CI) of both patent record and the final score of similarity is designed using weighted parameters given that by the person manner search. The technique of assemble maps dynamically by examine the SAO support contents to identify the technological opposition improvement.

2.2 Property Function Based Techniques

The property –function analysis move toward take out assets and utility from patent documents as improvement models from beginning to end grammatical investigation. A property conveys an exact feature of a system. Whereas a function symbolize a fitting action of system [17]. In spite of their value, property-function based techniques also have subject parallel to additional text mining and based techniques. Presented a system called Trend Preceptors for detect the technological development from make use of. Trend Preceptors meant for make use of a property-function based method for supplementary specialist in the resourceful perception and performing the progress trend analysis for tools forecasting. The method was practical in silicon support thin film solar cells and the outcome were establish hopeful in meeting the investigate objectives.

2.3 Rule Based Techniques

Rule based techniques for text mining frequently use several sort of inference rules and association rules. Such variety of techniques are successful helpful for construct meaningful associations in the middle of the structures haul out from large datasets**.** The rules are generally IF-Then rules that aid in haul out the suitable data commencing the patents, element the numeral of rules enlarge [18]. The PTCM come close to consists of apparatus such as

- Patent Fetcher.
- Patent Transformer.
- Patent Indicator Calculator.
- Change Detection Module the Patent Indicator

Calculator modules determine the patent values. The change detection modules are the key module to agree on the patent change trends. A strategy planning method that joins together the patent analysis procedure with IF_THEN rules based Fuzzy Inference system (FIS). The proposed move toward is context sensitive and get hold of knowledge from a global patent database instead of domain specialist. The important element of FIS:

- Patent Quality.
- Revealed Patent Advantage.
- Patent Activity.
- Relativity Citation Index.

2.4 Semantic Analysis Based Techniques

Semantic based text mining techniques pass on domain knowledge and create associations surrounded by domain specific concepts [19]. The types of techniques are useful in discover the similarities between patents and formative the future hi-tech trends via logically connecting parsed grammatical structures. [20] An approach for semantic analysis of the claims made in patent documents to discover the infringement. The hierarchical keyword vector utilizes similarity indicators to identify the relationship among the claim elements. The SIPMS has capabilities of semantic analysis and uses text mining techniques to process and analyse the patent documents. It comprise of mainly three processes such as a) Pre-Processing b) Patent Analysis c) Invention Support. The task of three segmentation agent is to divide the selected patent in a semi-structured format based on filing date, assignee, IPC codes, titles, abstracts, claims and description of invention. [20] A facts planned based software frame effort to enable the reclamation of patent related information from numerous, various and un-coordinated information birthplaces widespread in the U.S Patent System.

2.5 Neural Network Based Techniques

Neural network based methods have also been second hand for patent classification and technology forecasting. Further exactly the back propagation neural network algorithm partakes be present recycled to train a patent quality of patents. The neural network based methods consume too stayed in aggregation through rule based approaches constructed on artificial neural network. The research [21] absorbed on reducing the energies then he time essential to pursuit aimed at to control the patent quality to the R&D action specific to a modernization. The cold start problem arises when the system to begin with has less data for building recommendations that in the end may result in indefinite recommendations of patents.

2.6 Visualization Techniques

Another major approach for contemporary patent analysis is the use of visualization tools to represent patent information and result analysis [22] Presented framework to identify the technological trends by analysing patents for Carbon Nanotube Field Emission Display (CNT-FED). Technology cycle time represents the technological progress between two time intervals. TCT refers to the time interval between a previously field patent and a target patent. Contrary to the existing topic models, the authors proposed a model called Inventor-Company-Topic model that incorporates information about the inventors and companies. The patent terms consists of process such as

- Content Retrieval.
- Content Ranking.
- Content Selection.

3. Text Mining Tools

3.1 Bio Rat

Bio Rat (Bio logical Research Assistant for Text mining) [23] is a separate PPI's are present. Bio Rat effort to discover download full paper or, uncertainty not conceivable abstracts, preliminary from Pub Med then existence following next connections, cross out from and to web pages. Informative relations, such as proteins and genes, are vital in the together corpus. Presented on table beside through the textual with textual information (sentences) that point to their proof of identity. Bio RAT was evaluated using DIP [25] subsets. It extracts information by user-defined semantic templates such as "Interaction of (PROTEIN-1) and PROTEIN". It proven aim for information extraction (IE) system, the Gate Tool Box [24]. It found bio entities such as per proteins, among their names resemble common English Words using a "Part of Speech" tagger and dictionaries called (Gazetteers).

3.2 EFIP

Extracting Functional Impact of Phosphorylation [26] is a text mining system concentrated going on mining protein interactions networks of phosphorroyalted proteins. The user may idea a protein and regain the results a list of correlated toward to phosphorylation and protein communications.

3.3 FACTA+

Finding Associated Concepts with Text Analysis is the up-to-date version of the FACTA= [27] tool. FACTA= can use between others one or more proteins as query and retrieve Pub Med abstracts via before specific word/ concepts indexes. Concepts establish in the documents are then graded. FACTA+ achieved its benchmarking with data of the BionLP'09 shared task.

3.4 HIT Predict

It is a database of high assurance [28] PPI's integrated as of more PPI database such as Intact [29] and BIOGRID [30]. Interactions from countless databases are mutual and allotted a self-reliance score. User can search Hit Predict for a protein and see putative PPI's interactions lengthwise through the type of supporting proof.

3.5 Negatome

It shadows a distinctly different methods beginning entirely added system that are termed [31] in this analysis. The associations remain allocated a self-assurance score grounded on five simple features of the sentence such as the word that signposts the negations. Negatome be located manually evaluated

3.6 Open DMAP

It is a view of analysis and information extraction system that attentions [32] in the middle of others on protein transport and protein interactions. Biomedical corpus as input, open DMAP can make available a list of imaginable protein transformation events or PPI's open DMAP was evaluated exploiting Bio-creative datasets.

3.7 Poly SEARCH

It is web tool [33] which exploits a variation of techniques in text mining and information retrieval to isolate, best part and rank informative abstracts, paragraph or sentences merging text from Pub Med nevertheless too supplementary data bases. It grants to the user relationships in the middle of human dis-ways, tissues, organs and sub cellular localizations.

3.8 PP Look

It is a standalone tool[34] which extracts PPI's starting by a query protein and going on to discovery sentences they enclose the protein of interest. PP look uses the GENIA tagger [35] form. Outcomes be located envisioned in a 3d graph based open GL.

3.9 String

[36] It is an online database that holds experimentally tested and putative PPI's. They are kind of non-text mining methods such as genomic context, experiments, co-expression, literature and public repositories.

3.10 IMID

INTEGRATED MOLECULAR INTERACTION DATABASES gather PPI's from created databases but also [37] of interaction molecules and the type of interaction. The rules are based on 12 linguistic qualities narrowly associated to language rules those terms PPI relations.

3.11 PPI Extractor

The tool receives a list of Pub med synopses as idea and creates a PPI network by inferring PPI's. [38] The tags at that moment are normalized and subsequent that extract associated features centered, convolution tree and kernels to extract PPI's which are sub-sequent recycled towards establish visualizing PPI's network in the graph where edges represent the reliabilities of the PPI's.

3.12 GENE Ways

It analyzes interactions between molecular substances regarding signal transduction pathway performing NER and by disambiguating [39]. Handler can pursuit by thru by means of a protein name as effort and recover the result in a table of interactions along with a self-assurance score. It stood physically assessed by skillful in molecular technology.

3.13 PPI Finder

The guidelines are practical on sentences holding more than a protein and exam line in the relative position [30] and kind of further words surrounded in those sentences. PPI Finder stood evaluated using AIMED, HPRD50, Intact corpora.

4. DATA BASE.

4.1 PPI Data Base

Protein interaction data deposited on public databases which different in size. Repositories are often species-specific and hold information about by hand justified by computational predicted PPIs. The link on database attention high validated communications by manually deleted out low-quality and erroneous. Some databases are:

- The biological general repository for interaction datasets [40].
- The molecular interaction database [41].
- The high quality intra comes database [42].
- The in act molecular interaction database [43].
- The human protein reference database.

A most important concern of these repositories remains that their simulations demonstrationns varies expressively commencing database to database and interchange and direct comparison of not straight forward. Broadly recycled to automate the representing among dissimilar protein identifiers between databases are the Protein Resource Information (PIR) Identifier Mapping Tool (IMT), the mapping tool of different [44] and I Reflexed.

4.2 PPI Benchmark Datasets

PPI Benchmark Datasets are to calculated the PPI declaration in instruction of the methodology they behind. The estimate of the PPI "gold standard" datasets is not a nearby point task as information and foregoing ability unlike from organism to organism. MIPS database was started used to evaluate yeast PPI predictions to any further extent probable database designed. Functional associations, KEGG pathways gene ontology [45] and panther [46] library are comprehensively used. Idea of literature based negative supervision database developed. Y2h results can potentially benchmarking.

5 CONCLUSIONS

Most of the techniques are based on different methodologies such as Clustering, Classification, Relationship Mining and Pattern Matching. Those methodologies consume stayed recycled in discovering, detecting and removing relevant information and data from unstructured and disorganized textual resources. All those approaches have individual importance in designing and implementing effective data warehouse that would be used for different purposes. Primarily data warehouses are hand-me-down by Scholars, Researchers, development centres.

6. REFERENCES

[1] “Selection Criteria for Text Mining Approaches” by Hussein Hashini, Alaaeldin Hafez, Hassan Math our College of computer and information sciences, king Saud University. Riyadh, Saudi. www.elsevier.com/locata/comphumbeh

[2] Ahmad & Khanum 2010; Bhushan, Pushkar, Shivaji & Nikhil 2014; Navanethakumar & Chandrasekhar, 2012.

[3] Rath Jena, Nayak 7 Biosoyee 2011; Senellart & Blondel, 2008, vashishta & Jain 2011.

[4] Achtert et al., 2006; Kriegel Kroger, Zimek 2009; Silwattananusan & Tuamsuk, 2012.

[5] Krause, LeskWovec & guestrin 2006; Patel & Sharma, 2014.

[6] Ropers, Mat wish, & Sebastian, 2009; Krieger et al., 2009; Lehman 2010; Lincy Liptha, Raja & Tholkappia Arasu; Navanethakumar & Chandrasekhar, 2012. Patel 7 Sharma, 2014; Senellart 7 Blondel, 2008.

[7] Tseng YH, Lin YL, "Text Mining techniques for Patents Analyses". Inform process manag 2007; 43(5):1216_47.

[8] S.Ghazinoory, F.Ameri, S.Farnoodi. "Applicants, techno forecast soc changes (2012: 80: 918-31).

[9] Coussement K, Poel DVD, "Integrating a voice of Customer through Call Center E-mails into Decision Support System for Prediction". Inform Manag 45(3):164. http://dx>doi.org/10.1016/j.im.2008.01.005.

[10] A.Zanasi, "Virtual weapons for real wars and text mining for national security" In: proceedings international workshop on computational intelligence in security for information systemCISIS'08", Springer Berlin Heidelberg: (2009 pp.53-60).

[11] Lu.B, Zhang "A survey of Opinion Mining and Sentiment Analysis". In Mining Text Data 2012. pp.415.63

[12] KB.Cohen, L.Hunter. "Getting Started In Text Mining". PLoS Comput Biol4 (1):e20. http://dx.doi.org/10.1371/journal.pcbi.0040020.PMC 2217579. PMD 18225946

[13] Masiakowski.P, Wang.S "Integration of Software Tools in Patent Analysis" World Pat INF 2013:35(2):97-104.

[14] Park H, Ree|| Kim. K "Identification of Promising Patents for Technology Transfers Using TRIZ Evolution Trends", Expert Syst Appl (2013:736-43).

[15] Yean Kim k. "Detecting Signals of New Technological Opportunities Using Semantic Patent Analysis And Outlier Detection", Scientometrics (2011:1-17).

[16] Dewulf S. "Directed Variation of Properties New or Improved Function Product DNA Base or Connect and Develop". Procedia Eng. (2011; 9: 646-52).

[17] Han J.Kambeer, “Mining Association Rules in Larger Databases" in Data Mining Concepts and Techniques". San Francisco. CA, Morgan Kaufmann: 2001.

[18] Shill M.J, LU. Dr.Hsu "Discovering Competitive Intelligence by Mining Changes In Patent Trends". Expert Syst Appl (2010:37(4):2882-90sss).

[19] Bonimo.; D, Claramella. A, C.orona. F, "Review of the State of the Art in Patent Information and Forth Coming Evolutions Intelligent Patent Informatics". World Pat. In (2010: 32 (1): 30-8).

[20] Lee C, Song Barky. "How to asses patent infringement risks", "A semantic patent Claim Analysis Using Dependency Relationships". Technol anal start manang (2013:25910:23-28).

[21] Trappey A|C, Trappey CV, Wu C-YW, Fan CY, Lin Y_-l "Intelligent Patent Recommendation System for Innovative Design Collaboration". Newt comput Appl (2013:1441-50).

[22] Chang Pl, Wu Cc, Leu Hj.: "Using Patent Analysis to Monitor the Technological trends in an emerging Field Technology",” A case of carbon nanotube field emission display". Scienometrics 2010:82(1): 5-19.

[23] D.P.A.Corney, B.F Buxton, W.B.Langdon, D.T Jones, Bioinformatics (Oxford, England) (2004) 1699-1706.

[24] H. Cunningham, D. Maynard, K. Bontcheva, V.Tablan, in: Proc. 40th Annual Meeting Assoc. Comput. Linguist, "Association for computational linguist", Stroudsburg, PA, USA, 2002, pp. "168-175” (ACL'02').

[25] L.Salwinskt, Nucleic Acids Res .32 (2004) "D449-D451" (Database Issue).

[26] C.O. Tudor, C.N. Arighi, Q, Wang, C.H. WU, Vijay-Shankar, Database 2012(2012) bas044.

[27] Y.Tsuruoka, J.Tsujiis, S.Ananiadou, Bio Informatics (Oxford England) 27 (2011-119.

[28] A Patil, K. Nakal, H.Nakamura, Nucleic Acids res 39(2001) "D744-D749" (Database Issue).

[29] S.Kerrien, B.DU.Mousseau, BAranda, A Bridge, F Broackers- Carter, C. Chen, M.Duesbury, M.Feuraman , U.Hinz, C.Jandaris, RC Jimenez, J. Khadake, U.Mahadevan, M.Tyres, Nucleic acids Res 40(2012) " D841-D846" (Database Issue).

[30]A.Chart_Arymontn, B-j Bretkreutz, S.Heniciker, L.Boucher, D Chen, J Rust, M. Livestone, S.A Winter, C.Sark A Tyres nucleic acids Res 41 (2013)" D816-D823" (Databases).

[31] P.Blohmn, G.Frishman, P.Smailowski, F.Geoberts, B.Wachinger, A.Ruepo frishman, nucleic acids Res 42 (2014) "D369-400".

[32] L.Hunter, Z.Lu J.Firby, W.A BaurnGartment Jr.BMC Bio-Informatics 9 (2009, 79).

[33] D. Cheng, C.Knox, N.Young P. Standards, S. Damaraju. Nucleic acids Res 36 (2008) "W399-W405" (Webserver Issue).

[34] S.W.Zhang, Y.J.Li.Xia, Q.Pan. BMC Bio-Informatics (2010) 326.

[35] Y.Tsuroka, Y.Tateisht, J.D.Kam. T.Ohta, J.McNaught, S, Ananiadou, J.Tsoji, P.Bozans, E.N. Hoeslis "Advances in Informatics" vol 3748 Springer 2005. PP "382-392"(Lectures notes in Computer Science).

[36] A.Franschini, D.Szklarczyk, S.Franktld, M.kuhn, M.Simonovic, A.Roth, J Lin, P. Minguez, P.Bork C.Mering, L.J Jenson, Nucleic acids Res.41(2013) "D808-D815" (Database Issue).

[37]S.Balaji, McClendon, R.Smaikwski, F.Gobels. B.Wachinger, A.Ruepo, O.fishman, Nucleic Acids Res 42 (2014) "D369-400".

[38] M.Huang, X.Zhu, Y.Hao, Bio-Informatics (Oxford, England) 20 (2004) "3604-3612".

[39] Y.Tsuroka, M.Miwa, K.Harnamoto, J. TSujii, S.Ananiadou, Bio-Informatics (Oxford England) 27 (2011) pp.no:"2759-2765".

[40] L.Licta, L.Briganti, D.Peluso, L.Perfetto, M.Lannucelli, E.Galeota, F.Sacco, A.Paloma, nucleic acids res.40 (2012) "D857-D861" (Database Issue).

[41] J.dsas, H, Yu, BMC Syst.Biol 6(2012) 92.

[42] L.Salwinski, Nucleic Acids Res. 32 (2004) "D449-D451" (Database Issue)

[43] S.Kerrien, B.Aranda, L.Breuza, A.Fridge, F.Broackers, A.Winter, C.Stark, J. Nixon, L.Ramage, N.Kolas, L.O. Donnell, T.Rrguly, M.Tyres, Nucleic acids Res.(40/92012) D841-D8$6 (Database Issue).

[44] Uniport Consortium, Nucleic Acids Res.41 (2013) "D43-D47" (Database Issue).

[45] M.Ashburner, Nat, Genet 25 (2000), pp."25-29".

[46] P.D. Thomas, M.J. Campbell, A. kejariwal, H.Mi, B. KarlaK, R. Daverman. K. Diarner. A. Narechama Genome, Res 13(2003) “2129-2141".