Clinical Informatics by Using Cross Reference Method

Komal S Munde[1], Prof. Shiv Sutar[2]
[1, 2] Department of Computer Engineering
[1, 2] MITCOE, Savitribai Phule Pune University, Pune, India

Abstract*-Topic detection in a text corpus is the detection of semantic units from the underlying texts that can function as building blocks of knowledge discovery. Topic detection provides a powerful tool for text summarization and information navigation across a corpus of documents. Topic detection from text documents using statistical models and natural language processing techniques has been extensively studied by the researchers. In the recent years, it has drawn considerable research interest in the field of data mining too. The problem is also interesting and challenging in the domain of academia, considering the ever growing size of academic literature and the fast changing fields of research, which requires the algorithms to be fast and highly scalable in order to efficiently mine research topics and their features over the time-line. In this survey we are going to see some of the work done by several researchers on the same topic.*

Keywords-Classification, clustering, document networks, post marketing product surveillance, similarity.

I. INTRODUCTION

With the rapid progress of computer technology in recent years, there has been an explosion of electronic information published on the Internet. The world wide web represents vast stores of information. However, the sheer amount of such information makes it practically impossible for any human user to be aware of much of it. Neither the depth nor the extent of possibly useful information is known to the human user of the web. This problem will only get worse, without efforts to deal with such information overload.

In particular, such a phenomenon is more serious for the decision maker, who has to determine what to do at the right time. Timely decision-making can be accomplished by awareness of the ever-changing surroundings. In order to make a decision on a timely basis, the decision maker or decision supporter should monitor information broadly and track, in depth, specific items relevant to the decision. However, considering the tremendous volume of surrounding information, it is practically impossible for decision makers or decision supporters to discover and monitor all of the pertinent information or knowledge. Therefore, it would be very helpful to have a system that automatically discovers relevant, yet previously unknown information, and reports it to users in human-readable form.

Topic Detection and Tracking (TDT) was proposed as the endeavor to find the solution for problem of "well-awareness" on a user's surroundings. The topic detection is the problem of identifying stories in several continuous news streams that pertain to new or previously unidentified events. It consists of discovering previously unidentified events in an accumulated collection ("retrospective detection"), or flagging the onset of new events from live news feeds ("on-line detection"). Both forms of detection by design lack advance knowledge of the new events, but have access to (unlabeled) data in chronological order. Given the facts that it should deal with unlabeled data and a subset of news reports with similar contents is grouped together, the clustering algorithms are a good choice to discover unknown events.

In order to manage the large amount of information published, there is a clear need for systems that provide automatic organization of its content, in order to exploit the information more efficiently and retrieve only the information required for a particular user. Document clustering –the assignment of documents to previously unknown categories—has been used for this purpose. Topic detection aims to extract meaningful patterns from text datasets. These topics or semantic units, once detected can be used in multiple domains like trend analysis, document summarization, recommender systems, information navigation etc. With the growing volume of text documents generated on web and specialized digital archives, topic detection has become an extremely important tool for browsing, summarizing and clustering the documents.

Basically, topic detection software applications input a collection of temporal textual data and recognize the topic trend in time series. It can be viewed as a branch of research in Text or Web Mining involving the time features analysis. Many researchers have proposed different measurement as the significant of a topic at a particular time point. These measurements can be information gain, topic term weight, topic cluster size, number of documents containing the topic keyword, and etc. Higher the value of this measurement at a certain time, more important the topic is likely to be. These topics significant are usually measured towards a bigger time unit, for example day, week, month or year. This is different to the time-sensitive stock prices that may move drastically in minutes or even seconds. We don't report a topic of “stock price” movement in minutes range, but we may track the topic of the stock's drastic move in minutes range if it became popular and become news for a certain period.

II. LITERATURE REVIEW

This paper goal was to investigate the creation of document networks based on various thresholds of shared data also, distinctive clustering algorithms on those networks to recognize document clusters describing similar clinical cases. We made networks from vaccine adverse event report sets utilizing seven methodologies for linking reports. We then applied three clustering algorithms [visualization of similarities (VOS), Louvain, k-means] to these networks and evaluated their ability to identify known clusters. The report sets included one simulated set and three sets from the Vaccine Adverse Event Reporting System; each was split into training and testing subsets. Training subsets were utilized to estimate parameter values for the clustering algorithms and testing subsets to assess clusters.

As a cure, this paper [2] proposes a Subject–Action–Object (SAO)- based patent intelligence framework. SAO structures that can be extracted from textual patent data are known as the skill and innovative discoveries of the pertinent patent. On the premise of semantic investigation of patent SAO structures, this proposed knowledge framework develops patent maps and patent systems. Expanding on the maps and systems, the framework gives particular functionalities including identification of technology trends and significant patents, recognition of novel technologies, and recognizable proof of potential infringement. This paper depicts the design of this proposed patent intelligence framework in detail, and represents the framework's functionalities utilizing case studies.

This paper [3] present a model that combines aspects of mixed membership stochastic block models and topic models to enhance entity-entity link modeling by jointly modeling links and text about the entities that are linked. We apply the model to two datasets: a protein-protein interaction (PPI) dataset supplemented with a corpus of abstracts of scientific publications annotated with the proteins in the PPI dataset and an Enron email corpus.

The aim of this study [4] was to utilize the data encoded in the Medical Dictionary for Regulatory Activities (MedDRA®) in the US Vaccine Adverse Event Reporting System (VAERS) to support and assess two classification approaches: a various data retrieval procedure and a rule-based methodology. To assess the performance of these methodologies, they chose the conditions of anaphylaxis and Guillain–Barré syndrome (GBS). Techniques they utilized MedDRA® Preferred Terms stored in the VAERS, and two standardized medical terminologies: the Brighton Collaboration (BC) case definitions and Standardized MedDRA® Queries (SMQ) to classify two sets of reports for GBS and anaphylaxis. Two methodologies were utilized the rule-based instruments that are accessible by the two terminologies (the Automatic Brighton Classification [ABC] tool and the SMQ algorithms); and the vector space model.

In [5], proposes a SAO-based patent knowledge framework. SAO structures are the syntactically sequenced sentence of a subject, a verb, and an article that can be separated by exploiting natural language processing (NLP) of textual patent data. In particular, SAO structures that are possible from a patent are considered as having the capacity to give the skill, know-how, and significant discoveries of the patent. Expanding on the semantic similitudes of patents' SAO structures, the framework develops patent maps and patent systems that give a few functionalities for patent knowledge. Utilizing the patent maps and patent systems, the framework makes significant data to support decision-making for technology planning. In this paper, authot depict the framework of the proposed knowledge framework and show the framework's functionalities utilizing a few reasonable contextual analyses. We foresee that the proposed patent insight framework will be consolidated into the innovation arranging procedure to help specialists in the definition of technology techniques.

In paper [6], present a model that uites features of mixed membership stochastic piece models and topic models to enhance substance element link demonstrating by together displaying connections and content about the entities that are connected. Author apply the model to two datasets: a protein-protein interaction (PPI) dataset supplemented with a corpus of edited compositions of logical distributions annoted with the proteins in the PPI dataset and an Enron email corpus. The model is assessed by examining induced points to comprehend the nature of the data and by quantitative strategies, for example, functional classification forecast of proteins and perplexity which display changes when joint demonstrating is utilized over baselines that utilization just connection or text information.

In paper [7] authors implemented novel sequential memristor crossbar that is regarded as the first memristor circuit which copes with the sequential data. The designed crossbar is composed of two layers which are the base layer and the sequential one, respectively. The base layer can recognize only static items one by one. The sequential layer can detect the serial order of items and find the best match with the detected sequence among many reference sequences stored in the memristor array.

In paper [8] authors generated a networks from vaccine adverse event report sets using seven approaches for linking reports. After that they applied three clustering algorithms[visualization of similarities (VOS), Louvain, k-means] to genrated networks and evaluated their ability to identify known clusters. The report sets included one simulated set and three sets from the Vaccine Adverse Event Reporting System; each was split into training and testing subsets. Training subsets were used to estimate parameter values for the clustering algorithms and testing subsets to evaluate clusters.

Table 1: Survey Paper

Sr. No.	Title	Paper Details	Method Used	Advantages	Disadvantages
1.	Identifying Similar Cases in Document Networks Using Cross-Reference Structures	explore the creation of document networks based on different thresholds of shared information and different clustering algorithms	visualization of similarities (VOS), Louvain, *k*-means	flexible method for identifying clinically similar cases	---
2.	A patent intelligence system for strategic technology planning	On the basis of semantic analysis of patent SAO structures	Subject–Action–Object (SAO)- based patent intelligence framework	constructs patent maps and patent networks	system is not incorporated into the technology planning process
3.	Block-LDA: Jointly modeling entity-annotated text and entity-entity links	jointly models links between entities and text annotated with entities	functional category prediction of proteins and perplexity	useful in understanding the structure of the data both in terms of the topics discussed	---
4.	Application of information retrieval approaches to case classification in the vaccine adverse event reporting system	information encoded to support and evaluate two classification approaches	MedDRA Preferred Terms stored in the VAERS	more efficient than current rule-based approaches	Low performance
5.	A patent intelligence system for strategic technology planning	depict the framework of the proposed knowledge framework	SAO based patent intelligence framework	intelligence system	---
6.	Block-LDA: Jointly modeling entity-annotated text and entity-entity links	jointly models links between entities and text annotated with entities	entity-entity link modeling by jointly modeling links	outperforms approaches that use only a single source of information	---
7.	Sequential Memristor Crossbar for	new sequential memristor crossbar which is regarded as the first memristor circuit	sequential memristor crossbar	The new crossbar was verified to recognize speech	0-% variation to 20-% variation in the proposed sequential crossbar which

	Neuromorphic Pattern Recognition			patterns which are changing dynamically over time.	can be reduced.
8.	A Novel Approach for Decision Support in Uncertain Environments: The Case of Identifying Similar News Tickers in Natural Gas Trading	features of a novel method that encourages price prognosis in gas trading	task-technology-fit theory and technology-acceptance-mode	lower scattering within the clusters	---

III.CONCLUSION AND FUTURE SCOPE

This paper analyses various techniques used for topic detection and pattern reorganization detection task. In this paper an overview of existing topic discovery works in document clustering algorithms is presented, which gives the summarization of recent research work on various methods in document clustering for topic discovery.

REFERENCES

[1] Taxiarchis Botsis, John Scott, Emily Jane Woo, and Robert Ball, "Identifying Similar Cases in Document Networks Using Cross-Reference Structures", IEEE Journal of Biomedical and Health Informatics, Vol. 19, No. 6, November 2015.

[2] H. Park, K. Kim, S. Choi, and J. Yoon, "A patent intelligence system for strategic technology planning," Expert Syst. Appl., vol. 40, no. 7, pp. 2373–2390, 2012.

[3] R. Balasubramanyan and W. W. Cohen, "Block-LDA: Jointly modeling entity-annotated text and entity-entity links," Proc. Workshop Topic Models: Structure, Appl., Evaluation, Extensions, 2011,pp. 450–461.

[4] T. Botsis, E. J. Woo, and R. Ball, "Application of information retrieval approaches to case classification in the vaccine adverse event reporting system," Drug Safety, vol. 36, pp. 1–10, 2013.

[5] Jasmine Irani, Nitin, Madhura Phatak, "Clustering Techniques and the Similarity Measures used in Clustering", nternational Journal of Computer Applications (0975 –8887) Volume 134 –No.7, January 2016

[6] Rekha Baghel, Dr. Renu Dhir, "A Frequent Concepts Based Document Clustering Algorithm", International Journal of Computer Applications (0975 –8887) Volume 4 –No.5, July 2010

[7] S. N. Truong; K. V. Pham; W. Yang; K. S. Min, "Sequential Memristor Crossbar for Neuromorphic Pattern Recognition," in IEEE Transactions on Nanotechnology , vol.PP, no.99, pp.1-1

[8] S. Dreikorn, C. Felden, M. Pospiech and C. Koschtial, "A Novel Approach for Decision Support in Uncertain Environments: The Case of Identifying Similar News Tickers in Natural Gas Trading,"2015 IEEE 12th Intl Conf on Ubiquitous Intelligence, pp. 1676-1681.

[9] Hao Lin, Bo Sun, J. Wu and H. Xiong, "Topic Detection from Short Text: A Term-based Consensus Clustering method," 2016 13th International Conference on Service Systems and Service Management (ICSSSM), Kunming, 2016, pp. 1-6.

[10] K. Nur'aini, I. Najahaty, L. Hidayati, H. Murfi and S. Nurrohmah, "Combination of singular value decomposition and K-means clustering methods for topic detection on Twitter," 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Depok, 2015, pp. 123-128.