A survey on various text mining techniques and their Issues

Mr. Jagdish Raikwal
Asst. Professor, Dept. Of Computer Engineering
Institute Of Engineering & Technology,
Devi Ahilya University, Indore, India.

Vikas Ransore
Dept. Of Computer Engineering
Institute Of Engineering & Technology,
Devi Ahilya University, Indore, India.

***Summary*-Text processing and text search is a domain of data mining and knowledge processing. Text is used to represent knowledge and also used for communication. Text mining is process of discovering useful information or knowledge from the web data. There are various techniques available for the text mining. They have their own advantages and disadvantages as well. This survey paper describes the different types of text mining techniques and their issues. In this proposed work an efficient way for text based searching is suggested, along with its proposed mechanism.**

***Keywords: text mining, support vector machine (svm), principal component analysis (pca), biomedical research, nca*.**

1. INTRODUCTION

With the immense measure of data accessible on the web, the World Wide Web is a prolific range for information mining examination. Www is a prominent and intuitive medium to circle data today. The web is immense, assorted, and dynamic. So because of expansive measure of the information on the web some issue's emerges for data looking. Like:

1) Discovering important data (low accuracy and unindexed data)

2) Creating new information out of accessible data on the web (information –triggered procedure).

3) Personalizing the data (individual inclination in substance and presentation of the data)

4) Learning about the buyers (what would the client like to do?)

A. Web Mining

Web mining alludes to the general procedure of finding conceivably helpful and already obscure data or information from the web information. As indicated by examination targets, web mining can be partitioned into three distinct sorts [1], which are depicted underneath.

1)*Web Usage Mining:* Web utilization mining is the procedure of extricating valuable data from server logs e.g. use information from logs, client profiles, client session, treats, client inquiries, bookmarks, and mouse snaps and parchments, and so on.

2***)*** *Web structure mining:* Web structure mining is the procedure of utilizing diagram hypothesis to break down the hub and association structure of a site. E.g. finding powers and centers.

3) *Web substance mining:* Web Content Mining is the procedure of separating helpful data from the substance of Web reports. Content information compares to the gathering of realities a Web page was intended to pass on to the clients. It may comprise of content, pictures, sound, feature, or organized records, for example, records and tables.

Research exercises in this field additionally include utilizing procedures from different teaches, for example, Information Retrieval (IR) and common dialect handling (NLP). Content mining is an application territory of web substance mining.

B. Text mining

As indicated by [2] "the target of content mining is to endeavor data contained in literary records in different ways, including … revelation of examples and patterns in information, relationship among elements, prescient tenets, and so on."

As indicated by [3], "another approach to view content information mining is as a procedure of exploratory information investigation that prompts up to this time obscure data, or to responds in due order regarding inquiries for which the answer is not presently known."

1) How does it identify with information mining by and large?

ISSN: 2278 – 1323

2) How does it identify with computational phonetics?

3) How does it identify with data recovered?

	Finding patterns	Finding "Nuggets"	
Non textual data	General data mining	Novel	Non – novel
Textual data	Computational linguistics	Exploratory data analysis	Data base queries ,information retrieval

2. LITERATURE SURVEY

For text mining techniques we have study of some papers. In this section we describe the different techniques with different authors which are related to the text mining and also machine learning techniques.

According to the paper [4] there are many techniques for the text mining and the idea of the proposed work is takes also from this paper. In this paper author proposed some selected text mining methods i.e. Latent Semantic Analysis, Probabilistic Latent Semantic Analysis, Latent Dirichlet Allocation, Hierarchical Latent Dirichlet Allocation, Principal Component Analysis and Support Vector Machines. Along with some examples from the biomedical domain.

In the paper [5] author proposed the methodology for extracting the useful information from the Medline papers. Because it is a serious problem to extract and find the useful information from the currently available search engines and other tools and according to this paper identify the relationship of an active disease and extract relevant medicine for the patient automatically is tried by system. By applying some special technique and specific keywords it rank those medicines which are frequently used.

In the paper[6] some another methods for text mining are proposed by the author but they have some disadvantage also i.e. The ABC principle, and NLP based method it is for text mining but NLP based systems are more computationally intensive than co-occurrence based methods. And another disadvantage of some NLP based IE methods is that these are trained for detection of specific relationships on a training set and are thus limited by the availability and quality of the training data and do not scale well when a large number of relations needs to be detected

In this paper [7] the objective is to develop and validate automated electronic note search strategies or we can say automated digital algorithm to identify Charlson co morbidities.
When a series of programmatic queries applied to an institutional electronic medical record database then this algorithm built.

In the paper [8] they proposed an operational vision for a digitally based Care system that incorporates data-based clinical decision making. The system would aggregate every patient electronic medical data in the course of care one by one.

3. PROPOSED WORK

The above given figure demonstrate the proposed system architecture of the query improvement technique. The proposed technique involves search interface which accepts the user query for search. There are two aspects of the user query input first during input and after input of user query. During input of user query system tokenize the input query words and using the available tokens the neighbor component analysis is performed and the most relevant previous user queries are extracted from data base. This extracted user queries are sorted and listed using the auto complete text box for guiding the user for writing the correct query. Then after the similar query words are collected as queries are provided as input to the PCA algorithm where the essential features are extracted from data and using selected features and the medicinal data base a SVM classifier is trained. The SVM classifier can accept the user current input values and predict the actual the content which is actually user want to find in medicinal data base. On the other hand the data final user query which is used for finding the outcomes from the database is then preserved into the query database.
In this section some essential functions are listed which are used in the proposed medical record search technique.

1. ***Input query***: this function is help to accept the user query by the system

2. ***Query parser***: the input query is produces in this phase for creating n-grams from the database.

3. ***Feature Extraction***: using the likely component analysis the generated n-grams are evaluated for finding the data features for training and similarity computations

4. ***Trained SVM***: the neighborhood features are produces with the classes for train the SVM.

5. ***Predict***: the trained network is accept the current sequence of characters and generate the next token or complete word from the data base

6. ***Search***: after query support a search methodology is implemented for finding the most relevant data in database.

7. ***Rank***: the results are ranked according to the query relevancy

8. ***Result list***: this function accepts the ranking function outcomes are generate the list of records

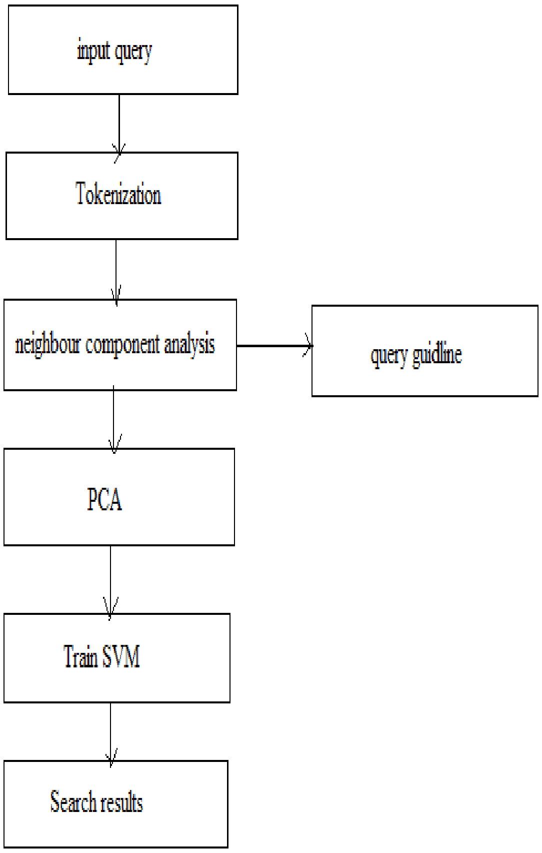

Figure 2 subsystem components

The above given diagram shows the basic and essential feature sets that are help to improve the medical record search technique. Therefore first the user input query is used to find the relevant previous queries which are extracted using the neighbor component analysis. Than after the similar data is consumed with the PCA algorithm for evaluating the data and extraction of features from the raw data. Finally the extracted features are used to train the SVM classifier which accept the current user input and extract the output of the query input.

S.NO.	Method	Responsibilities
1)	*Parse Query*	The current input character is parsed using the provided query interface for example "cancer" is parsed in 'c' 'a' 'n' 'c' 'e' 'r'
2)	*Find likely hood*	Using the neighbor component analysis technique the most likely sequence of characters are search from the data base
3)	*Compute similarity*	After listing of similar token of the words and domain of data is distinguished using the similarity functions
4)	*Train SVM*	A SVM algorithm is used to classify different data for prediction therefore this phase is to provide the training of algorithm
5)	*Predict*	The function accepts the current character of the user query and predicts next character
6)	*Rank List*	The given function is used to rank the predicted outcomes of the neural network
7)	*Display results*	This function is used to display results according to the query relevancy relevancy

4. CONCLUSION

Text mining is an essential contribution of data mining large amount of data is available on the web for text mining. Therefore the proposed system provides the efficient way for text mining on any big database.

References

[1] S. Gowri Shanthi, Dr. Antony Selvadoss Thanamani, Enhanced Approach on Web Page Classification Using Machine Learning Technique, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 1, Issue 7, September 2012

[2] M. Grobelnik, D. Mladenic, and N. Milic-Frayling, "Text Mining as Integration of Several Related Research Areas: Report on KDD'2000 Workshop on Text Mining," 2000.s

[3] M. Hearst, "Untangling Text Data Mining," in the Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, 1999[4] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Second Edition, University of lllinois at urbana-champaing 2006

[5] Ajay S. Patil, B.V. Pawar, Automated Classification of Web Sites using Naive Bayesian Algorithm, IMECS vol-1, 2012

[6] Ms. Darshna Navadiay, Mr. Mehul Parikh, Ms. Roshni Patel, Constructure Based Web Page Classification, International Journal of Computer Science and Management Research, Vol 2 Issue 6 June 2013 ISSN 2278-733X

[7] Victor Fresno, Raquel Martinez, Soto Montalvo, Arantza Casillas, Naive Bayes Web Page Classification with HTML Mark-Up Enrichment

[8] Shalini Puri and Sona Kaushik, A Technical Study And Analysis On Fuzzy Similarity Based Models For Text Classification,, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.2, March 2012

[9] Raj Kumar, Dr. Rajesh Verma, Classification Algorithms for Data Mining: A Survey, International Journal of Innovations in Engineering and Technology (IJIET), Vol. 1 Issue 2 August 2012

[10] Dou Shena, Qiang Yang a, Zheng Chen, Noise reduction through summarization for Web-page classification, Information Processing and management 43 (2007) 1735–1747