ISSN:0975-9646

A Survey on Mining and Statistical Analysis of Social Website Data using Syntatic Analysis and Hybrid Classification

Niranjan C Kundur[1], Rohitaksha K[2] , Sreenatha M[3], Abhilash C B[4]

Department of Computer Science & Engineering,
JSS Academy of Technical Education, Bangalore
Visveswaraya Technological University, Belgaum- 60.

***Abstract* - Abstract - Social networking websites create lots of data by the interaction among people leading to mutual learning and sharing of valuable knowledge comments, discussions, and wishes. Data in social networking websites are very much semi-structured, unstructured and confusing in nature. In every day conversations, people do not care about the spellings and accurate grammar in the construction of a sentence that may leads to different types of confusions, such as syntactic, and understanding the meaning leads to errors. Therefore, analyzing and extracting information patterns from such data sets are not so easy. Here it is proposed that the unstructured data has to be extracted first and re analysed by the nearest grammar checker and the final analyzing tool should be composed of many primitive analysis tools.**

Keywords: Extraction, Classification, Text -Mining.

I. Introduction

Social network sites have large beneficial effects on people's social communications through internet communication. Many social networking sites, which provide various services to meet different needs of the people, are emerging and growing rapidly with easier and better interactive experience. A large number of people register their personal accounts at social websites with social communities in order to extend their social relationships in local, national or at international ranges by connecting with people on the other part of the world. Such connections between individuals are usually based on their manifest or latent relationships fashioned by their self-disclosed information such as personal profiles, locations. Up to now, more than a hundred OSNs have been launched and many of them focus on providing local services to users in their own countries with mother-tongue-only interfaces by means of their inherent advantages in the aspects of local culture, daily lifestyle, online or offline behaviour.

A social network is usually formed and constructed by daily and continuous communication between many people and therefore which includes different relationships, such as the positions, closeness and betweenness among individuals or groups. In order to understand the social structure, social relationships and social behaviours, social network analysis is an essential and important part in analysing and getting the statistics. Research on social networks could be traced back to sociology, anthropology and epidemiology. Social network analysis developed with the kinship studies of Elizabeth Bott in England in the 1950s and the 1950s-1960s urbanization studies of the University of Manchester group of anthropologists. Initially, the studies of social networks analysis focused on small groups and small social networks. However, it became difficult and harder to manually look in to and analyse the very broad social networks. Therefore, strong computation power and information technology has became a very important tool for social networks analysis and the direction of the research is therefore now moving from sociology to computer science. Social Networks has become omnipresent in today's life. It gives way to share information between people anywhere and at anytime. Many Social Network sites Friendster, and MySpace. Face book is a social networking service and website which was launched in February 2004, operated behind the closed doors of Face book, Inc. As of February 2012, Face book has more than 845 million active users.

Social networking can solve coordination problems among people that may arise due to geographical distance and can increase the effectiveness of social campaigns by sharing the required information anywhere and anytime across the globe. However, in social networking websites, people generally use unstructured or semi-structured language for communication. In everyday life conversation, people do not care about the spellings and accurate grammatical construction of a sentence that may leads to different types of ambiguities, such as lexical, syntactic, and semantic. Therefore, extracting logical patterns with accurate information from such unstructured form is a critical task to perform.

Text mining can be a solution of above mentioned problems. Due to the increasing number of readily available electronic information , text mining is gaining more importance. Text mining is a knowledge discovery technique that provides computational intelligence. The technique comprises of multidisciplinary fields, such as information retrieval, text analysis, natural language processing, and information classification based on logical and non-trivial patterns from large data sets. In, the authors defined text mining as an extension of data mining technique. The data mining techniques are mainly used for the extraction of logical patterns from structured database. Text mining techniques become more complex as compared to data mining due to unstructured and fuzzy nature of natural language text. Social networking websites, such as Facebook are rich in texts that enable user to create various text contents in the form of comments, wall posts, social media, and blogs. Due to ubiquitous use of social networks in recent years, an enormous amount of data is available via the Web. Application of text mining techniques on social networking websites can reveal significant results related to person-to-person interaction behaviours. Moreover, text mining techniques in conjunction with social networks can be used for finding general opinion about any specific subject, human thinking patterns, and group identification in large-scale systems. Recently, researchers used decision trees and hierarchical clustering for group recommendation in Facebook where user can join the group based on similar patterns in user profiles .

II. LITERATURE SURVEY

Several works has been done in the area of text mining classification. The rapidly growing phenomenon of creating, editing, processing and transferring and storing data in digital form has made the automated machine learning classification systems become important due to the advent of large amount of electronic data. These classification systems are discovering, organising, analyzing and mining data in order to transform data into information since some decades now ,an increasing number of machine learning approaches have been developed to perform the tasks of classifying data into groups for some specified purposes , including decision tree induction, Rocchio Algorithm (RA) [10], K-Nearest Neighbour (K-NN) [3] Case Based Reasoning (CBR) [8], Decision Trees and Support Vector Machine [4] Support Vector Machine (SVM) [7], Artificial Neural Networks (ANN) [6], Genetic Algorithm (GA) [5]; [6]), Ontology based Text Classification[3];[2], Hybrid Approach [8]; [1]; [9];[10].

Electronic textual documents are extensively available due to the emergence of the Web. Many technologies are developed for the extraction of information from huge collections of textual data using different text mining techniques. However, information extraction becomes more challenging when the textual information is not structured according to the grammatical convention. People do not care about the spellings and accurate grammatical construction of a sentence while communicating with each other using different social networking websites (Facebook, LinkedIn, MySpace). Extracting logical patterns with accurate information from such unstructured form is a critical task to perform.

Improving prediction accuracy of text classifiers has been an important issue and many studies have been conducted in this area. Rocchio is a linear classifier. When the decision boundary is non-linear, the classification accuracy of Rocchio classifier is low. proposed a generalized pattern set algorithm to overcome the weakness of Rocchio algorithm. The main idea for this method is to construct more than one prototype vector for a category, in contrast to only one prototype vector for a category in the Rocchio algorithm. The drawback of this method is the difficulty to choose an appropriate k and the order in which positive patterns are chosen to construct each local prototype vector as the performance of the method depends on both of them. [13] combined KNN and SVM to construct a classifier for[14] Introduced fuzzy-rough uncertainty to enhance classification performance of the KNN algorithm. Some drawbacks still exist for this method. Firstly, it need to store all training data and hence for a large training set it may take large space. Secondly, for every new pattern, the distance should be computed between the new pattern and all training data. Thus, the efficiency of this method may be low.

Machine Learning techniques are most widely used in the field of clustering of data. The K-means algorithm is one which is widely used algorithm for clustering of data sets and is easy to understand and simulate on different datasets. In our paper work we have used K-means algorithm for clustering of yeast dataset and iris datasets, in which clustering resulted in less accuracy with more number of iterations. We are simulating an improved version in K-means algorithm for clustering of these datasets, the Improved K-means [5].

The main objective of the system is to make the stemmed document without stop words and tokens of that final document. With the help of this system we can search the document easily. By calculating the index term frequency can get the relevance of the document with the logic that a document in which more query items are present that document shall get a top ranking. can also retrieve and estimate relevant (according to query) information from the document.

Some of the techniques require a complete recalculation of the data set in case of adding just one new document or a term to the term-document matrix. This is especially a significant disadvantage for multi-user, real-time systems. It requires a flexible architecture to handle constant data input in an acceptable time frame. Another problem is the processing of very large data sets which are normal for text mining systems. An essential requirement for the data model is the reduction of very high dimensional data into low dimensional data, without the loss of important data. Furthermore, it should reduce the noise of the data. There has still research to be done to satisfy those requirements.

Each of the schemes previously mentioned has its own unique properties and associated strengths and problems. The simple decision tree induction and rule induction approaches are easy to be understood and interpreted , as the trade-off, the low classification performance of these approaches has restricted them to be widely implemented in real world application, especially when the number of distinguishing features within the data is large , k-nearest neighbour (KNN) is another approach which is easy to be implemented with high degree of effectiveness in many classifications tasks of varying problem domains . However when the amount of training data is large , k-nearest neighbour approach suffers from the problem of computational intensive and its classification accuracy could be drastically degraded when the number of attributes grows, artificial neural networks outperform the other classification approaches with its ability in handling data with high-dimensional features and also noisy and contradictory data. However, the high computing cost which consumes high CPU and physical memory usage has become the main disadvantage of the artificial neural networks. Bayesian approach is outstanding with its simplicity and low computational cost in both the training and classifying stage and it has been widely implemented in various types of domains and applications .however , this generative method has been reported to be less accurate than the discriminative methods such has support vector machines.

III. Methodologies

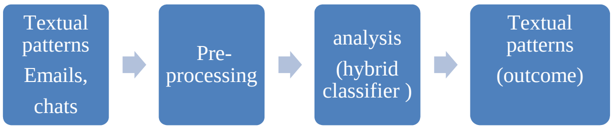

Fig 1. Block diagram of text mining in social network.

The textual data, the chart history and the log files of emails are been processed using text pre-processor. The tokenised data have to be analysed by hybrid classifier to give the enhanced statical information. Finally analysed and extracted textual patterns given as an outcome.

IV. Conclusion

In social network textual documents are extensively available due to the emails, comments, chats and blogs, etc. Many technologies are developed for the extraction of information from the huge collection of textual data using different text mining technique user data extraction becomes challenging when the textual information is not structured according to the grammatical construction. In our project introduced tokenization technique that can be used for extraction of logical patterns from the unstructured and grammatically incorrect the textual data and also introduced hybrid classification technique that are useful for the analysis of text data for best performance .

References

[1] Aci, M., Inan, C. & Avci, M. 2010. A hybrid classification method of k-nearest neighbour, Bayesian method and genetic algorithm. Expert Systems with Applications.

[2] Baumer, E. P. S., Sinclair, J. & Tomlinson, B. 2010. America is like Metamucil: Fostering critical and creative thinking about metaphor in political blogs. In Proceedings of 28th International Conference on Human Factor in Computing Systems (CHI 2010) ACM, Atlanta, GA, USA.

[3] Chang, M. & Poon, C. K. 2009. Using phrases as features in e-mail classification. Journal of System and Software, 82(6).

[4] Forman, G. & Kirshenbaum, E. 2008. Extremely fast text feature extraction for classification and indexing. In Proceedings of 17th ACM Conference on Information and Knowledge Management, California, USA.

[5] CB Abhilash, K Rohitaksha, Shankar Biradar, 2014. “A comparative analysis of data sets using Machine Learning techniques”. Advance Computing Conference (IACC), 2014 IEEE International. IEEE.

[6] Jo, T. 2010. NTC (Neural Text Categorizer): Neural network for text categorization. International Journal of Information Science, 2(2), 83-96.

[7] Luger, G. F. 2008. Artificial Intelligence: Structure and Strategies for Complex Problem Solving. 6th edn. Addison Wesley Luger, G. F. 2008. Artificial Intelligence: Structure and Strategies for Complex Problem Solving. 6th edn. Addison

[8] Li, J., Wang, H. & Khan, S. U. 2012. A semantics-based approach to large-scale mobile social networking, Mobile Networks and Applications, 17(2), 192-205

[9] Miao, D., Duan, Q., Zhang, H. & Jiao, N. 2009. Rough set based hybrid algorithm for text classification. Journal of Expert Systems with Applications, 36(5), 9168-9174.

[10] Meesad, P., Boonrawd, P. & Nuipian, V. 2011. A Chi-square-test for word importance differentiation in text classification. In Proceedings of International Conference on Information and Electronics Engineering, Singapore, 110-114.

[11] Zhao, Y. & Dong, J. 2009. Ontology classification for semantic-web-based software engineering. IEEE Transactions on Service Computing, 2(4), 303-317

[12] Jimmy Lin and Chris Dyer University of Maryland, College Park Data-Intensive Text Processing with MapReduce Manuscript prepared April 11, 2010