I J C T A, 9(28) 2016, pp. 319-324

TRAFFIC PREDICTION AND FORECASTING USING CLASSIFICATION OF TWITTER STREAM ANALYSIS

Boopalan. K* Rajesh. A and Nalini. P*****

Abstract: Web Usage Mining is that area of Web Mining which deals with the extraction of interesting knowledge from logging information produced by Web servers and apply that knowledge in specific field to enhance the usability of that field. In this paper we present a survey of the recent developments in this area that is receiving increasing attention from the Data Mining community. Also we have deployed the obtained knowledge towards the application area of traffic analysis in order to predict and forecast the traffic updates to the users.

Keywords: Web Mining, Knowledge discovery, Traffic analysis, Prediction, Forecasting

1. INTRODUCTION

With the rapid increasing popularity of the WWW, Websites are playing a crucial role to convey knowledge and information to the end users. Discovering hidden and meaningful information about Web users usage patterns is critical to determine effective marketing strategies to optimize the Web server usage for accommodating future growth. Most of the currently available Web server analysis tools provide only explicitly and statistical information without real useful knowledge for Web managers. The task of mining useful information becomes more challenging when the Web traffic volume is enormous and keeps on growing. Recently, social networks and media platforms have been widely used as a source of information for the detection of events, such as traffic congestion, incidents, natural disasters (earthquakes, storms, fires, etc.), or other events. An *event* can be defined as a real-world occurrence that happens in a specific time and space [12], [13]. In particular, regarding traffic related events, people often share by means of an SUM information about the current traffic situation around them while driving. For this reason, event detection from social networks is also often employed with Intelligent Transportation Systems (ITSs). An ITS is an infrastructure which, by integrating ICTs (Information and Communication Technologies) with transport networks, vehicles and users, allows improving safety and management of transport networks. ITSs provide, e.g., real-time information about weather, traffic congestion or regulation, or plan efficient (e.g., shortest, fast driving, least polluting) routes[14], [15].

For Perfect execution of above described procedure, personalization plays major role as it is the first and foremost step in web mining. Personalization can be done either via information brokers (e.g., Web search engines) or in an *end-to-end* manner by making websites adaptive. The latter solution is further attractive since it also cuts down on network traffic. Initial work in this area has basically focused on creating broker entities, often called recommender systems. An important component of personalization is Web mining.

* Research Scholar, Bharath University,Chennai.
Corresponding author: erbookumar@gmail.com
** Professor/CSE,C.Abdul Hakeem College of Engineering and Technology,Vellore
*** Professor,Bharath University,Chennai

Web mining can be viewed as the extraction of structure from unlabeled semi-structured data containing the characteristics of users or information. The logs kept by Web servers provide a classic example of such data. Web mining can be viewed as a special case of the more general problem of knowledge discovery in databases [16]. It can be said to have three operations of particular interest: clustering (e.g., finding natural groupings of users, pages, etc.), associations (e.g., which universal resource locators (URLs) tend to be requested together), and sequential analysis (the order in which URLs tend to be accessed).

In this paper we proposed a sequence of steps which delivers a quality output in traffic prediction strategy.

The Steps involved are

1. Tweet segmentation
2. Knowledge Discovery
3. Classification
4. Web Personalization
5. Web graphs for recommendations
6. Prediction and Forecasting

2. RELATED WORK

In paper[1], the authors present a novel 2-step unsupervised NER system for targeted Twitter stream, called TwiNER. In the first step, it leverages on the global context obtained from Wikipedia and Web N-Gram corpus to partition tweets into valid segments (phrases) using a dynamic programming algorithm. Each such tweet segment is a candidate named entity. It is observed that the named entities in the targeted stream usually exhibit a gregarious property, due to the way the targeted stream is constructed. In the second step, TwiNER constructs a random walk model to exploit the gregarious property in the local context derived from the Twitter stream.

In paper[2],the authors proposed to combine a K-Nearest Neighbors (KNN) classifier with a linear Conditional Random Fields (CRF) model under a semi-supervised learning framework to tackle the challenges during named entity recognition. The KNN based classifier conducts pre-labeling to collect global coarse evidence across tweets while the CRF model conducts sequential labeling to capture fine-grained information encoded in a tweet. The semi-supervised learning plus the gazetteers alleviate the lack of training data.

In paper[3] the authors proposed a complete preprocessing methodology that allows the analyst to transform any collection of web server log files into structured collection of tables in relational database model. The log files from different Web sites of the same organization are merged to apprehend the behaviors of the users that navigate in a transparent way. Afterwards, this file is cleaned by removing all unnecessary requests, such as implicit requests for the objects embedded in the Web pages and the requests generated by non-human clients of the Web site (i.e. Web robots). Then, the remaining requests are grouped by user, user sessions, page views, and visits. Finally, the cleaned and transformed collections of requests are saved onto a relational database model. They have provided filters to filter the unwanted, irrelevant, and unused data

In paper[4],the authors presented a real-time monitoring system for traffic event detection from Twitter stream analysis. The system fetches tweets from Twitter according to several search criteria; processes

tweets, by applying text mining techniques; and finally performs the classification of tweets. The aim is to assign the appropriate class label to each tweet, as related to a traffic event or not. The traffic detection system was employed for real-time monitoring of several areas of the Italian road network, allowing for detection of traffic events almost in real time, often before online traffic news web sites. We employed the support vector machine as a classification model,

In paper[5], the authors used association-rule mining based on frequent item sets and introduced a data structure to store the item sets. They split Web logs into user sessions and then mined these sessions using their suggested association association rule algorithm. They argue that other techniques based on association rules for usage data do not satisfy the real-time constraints of recommender systems because they consider all association rules prior to making a recommendation. Ming-Syan Chen and colleagues[17] proposed a somewhat similar approach that uses a different frequent itemset counting algorithm.

In paper[6], the authors aiming at providing a general framework on mining Web graphs for recommendations, 1) we first propose a novel diffusion method which propagates similarities between different nodes and generates recommendations; 2) then we illustrate how to generalize different recommendation problems into our graph diffusion framework.

In paper[7] the authors presented the algorithm which uses breadth-first enumeration and is based on the Apriori algorithmThe algorithm consists of three phases. In phase 1, they find all frequent paths (including paths with cycles), starting with frequent nodes and frequent edges. In phase 2, they find all graphs composed of two paths, in other words, they find all possible intersections between pairs of paths from phase 1. In phase 3, they merge pairs of frequent graphs, each consisting of n - 1 paths, such that the graphs have a common core of n - 2 paths in an attempt to produce graphs with n paths. Throughout, they assume that some admissible support measure is used. In phases 1 and 3, they construct frequent graph patterns recursively, using the Apriori approach

In paper[8] the authors proposed a concurrent neuro-fuzzy model to discover and analyze useful knowledge from the available Web log data. The authors made use of the cluster information generated by a self organizing map for pattern analysis and a fuzzy inference system to capture the chaotic trend to provide short-term (hourly) and long-term (daily) Web traffic trend predictions.

In paper[9] we propose a novel traffic surveillance system for detecting, tracking, and recognizing vehicles from different video sequences .In this approach, for robustness consideration, a background update method is first used to keep background static. Then, desired vehicles can be detected through image differencing and then tracked by a Kalman filter. Furthermore, through shadow elimination and feature extraction, several features including vehicle size and linearity features can be extracted.

Paper[10] presents a Cloud-based system computing customized and practically fast driving routes for an end user using (historical and real-time) traffic conditions and driver behavior. In this system, GPS-equipped taxicabs are employed as mobile sensors constantly probing the traffic rhythm of a city and taxi drivers' intelligence in choosing driving directions in the physical world. Meanwhile, a Cloud aggregates and mines the information from these taxis and other sources from the Internet, like Web maps and weather forecast.

In paper[11] the authors examine the effectiveness of a clustering technique, i.e. latent class clustering, for identifying homogenous traffic accident types. Firstly, a heterogeneous traffic accident data set is segmented in seven clusters, which are translated into seven traffic accident types. Secondly, injury analysis is performed for each cluster.

Sl. No	Title	Author	Issues	Method used	Tools	Advantages/disadavantages
1	TwiNER: Named Entity Recognition in Targeted Twitter Stream	Chenliang Li Jianshu Weng, Qi He3, Yuxia Yao	SIGIR'12, August 12–16, 2012,.	TWINER: A Dynamic Programming Algorithm Segment ranking	Segment Stickiness Function Random Walk Model	TwiNER, h, achieves comparable F1 performance with the other supervised systems. TwiNER performs much better than LBJ-NER and T-NER on SGE_g. Disadvantage: This paper does not address the problem of entity type classification.
2	Recognizing Named Entities in Tweets	Xiaohua Liu , Shaodian Zhang Furu Wei, Ming Zhou	Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 359–367,	NER for Tweets KNN Training KNN predication	Evaluation Metrics-Baselines	KNN consistently yields comparable performance, while enjoying a faster retraining speed Disadvantage: The gazetteers used in our work contain noise, which hurts the performance. Moreover, they are static,
3	Knowledge Discovery from Web Usage Data: Complete Preprocessing Methodology	G T Raju and PS Satyanarayana	IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.1, January 2008	Data Preprocessing	KDWUD Toolbox	Reducing the size of the log file along with increasing the quality of the data available through the new data structures Disadvantage: The process itself does not fully guarantee that we identify correctly all the transactions
4	Real-Time Detection of Traffic From Twitter Stream Analysis	Eleonora D'Andrea, Pietro Ducange, Beatrice Lazzerini, Francesco Marcelloni,	IEEE transactions on intelligent transportation systems, vol. 16, no. 4, august 2015	Traffic detection system	Text mining elaboration	Shown the superiority of the SVMs, which have achieved accuracy of 95.75%, for the 2-class problem, and of 88.89%for the 3-class problem, in which we have also considered the traffic due to external event class. Disadvantage: May take some time and effort, or to wait for getting the information from the radio traffic news.
5	Effective Personalization Based on Association Rule Discovery from Web Usage Data	B. Mobasher et al.,	Proc. 3rd ACM Workshop Web Information and Data Management (WIDM 2001), 2001, pp. 9–15.	Collaborative Filtering	Association Rule	Satisfy the real-time constraints of recommender systems because they consider all association rules prior to making a recommendation.
6	Mining Web Graphs for Recommendations	Hao Ma, Irwin King, ichael Rung-Tsong Lyu,	IEEE transactions on knowledge and data engineering, vol. 24, no. 6, june 2012	Graph diffusion Model	Query Suggestion Image Recommendation	This work is a general framework which can be effectively, efficiently, and naturally applied to most of the recommendation tasks on the Web.

Table 1 contd...

7	Discovering Frequent Graph Patterns Using Disjoint Paths	Ehud Gudes, Solomon Eyal Shimony, Natalia Vanetik	IEEE transactions on knowledge and data engineering, vol. 18, no. 11, november 2006	Apriori algorithm	Operations on Graph Composition Bijective sum	Produces fewer candidate patterns and therefore performs fewer support computations than the edge addition algorithm. Disadvantage: Synthetic graphs are not very regular. As the number of distinct labels in synthetic database increases, the chance of finding nontrivial frequent patterns in that database decreases drastically.
8	Intelligent web traffic mining and analysis	Xiaozhe Wanga,*, Ajith Abrahamb, Kate A. Smitha,	Journal of Network and Computer Applications 28 (2005) 147–165	Hybrid neuro-fuzzy approach for web traffic mining and prediction	WUDA	WUDA provided useful information related to the user access patterns, which could not be possible by using conventional statistical approaches. Disadvantage: Due to incomplete details, the authors had to analyze the usage patterns for different aspects of log files separately.
9	An Automatic Traffic Surveillance System for Vehicle Tracking and Classification	JunWei Hsieh, Shih-Hao Yu, Yung-Sheng Chen and Wen-Fong Hu	IEEE Transactions on Intelligent Transportation Systems, Vol. 7, No. 2, 175-187, 2006	Detection of Lane Dividing Lines Feature Extraction and Vehicle Classification	Kalman filter	Method is superior in terms of accuracy, robustness, and stability in vehicle classification.
10	Driving with Knowledge from the Physical World	Jing Yuan, Yu Zheng, Xing Xie, Guangzhong Sun	2011 ACM KDD'11, August	Online Mining Online Inference	h-step-ahead transition matrix	This prediction method considering both historical patterns and real-time traffic, , outperforms the approaches separately using and in predicting the future traffic conditions, especially, in handling road segments with relatively more links Disadvantage: T-Drive only employs the historical traffic patterns in the routing process.
11	Traffic Accident Segmentation by Means of Latent Class Clustering	Benoît DepaireGeert Wets Koen Vanhoof		Cluster Analysis	Finite Mixture Models	Latent class clustering succeeds in finding various clusters in a heterogeneous traffic accident data set. Disadvantage: The restriction on the cluster-specific covariance matrix can be removed from the model with a high cost of parametric complexity and computing time.

3. CONCLUSION

In this paper, literature survey on traffic prediction and forecasting using classification of twitter stream analysis was useful to understand the techniques used and how to discover knowledge, classify, recommend and forecast the web usage data. The fundamental approach is to accurately acquire knowledge about the

classification of tweet segmentation through supervised learning. The acquired knowledge has to be analysed carefully in order to provide the traffic prediction with the help of graph recommendations.

References

1. Chenliang Li , Jianshu Weng, Qi He, Yuxia Yao, Anwitaman Datta, "TwiNER: Named Entity Recognition in Targeted Twitter Stream", *ACM 978-1-4503-1472-5/12/08, SIGIR'12,* August 12–16, 2012.
2. Xiaohua Liu , Shaodian Zhang, Furu Wei , Ming Zhou, "Recognizing Named Entities in Tweets", *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 359–367.
3. G T Raju and P S Satyanarayana ,"Knowledge Discovery from Web Usage Data: Complete Preprocessing Methodology (IJCSNS) *International Journal of Computer Science and Network Security*, VOL.8 No.1, January 2008.
4. Eleonora D'Andrea, Pietro Ducange, Beatrice Lazzerini, and Francesco Marcelloni , "Real-Time Detection of Traffic From Twitter Stream Analysis", *IEEE Transactions On Intelligent Transportation Systems*, Vol. 16, No. 4, August 2015
5. B. Mobasher et al., "Effective Personalization Based on Association Rule Discovery from Web Usage Data" .*Proc. 3rd ACM Workshop Web Information and Data Management (WIDM 2001*), pp. 9–15.
6. Hao Ma, Irwin King, and Michael Rung-Tsong Lyu , "Mining Web Graphs for Recommendations*", IEEE Transactions On Knowledge And Data Engineering*, VOL. 24, NO. 6, JUNE 2012.
7. Ehud Gudes, Solomon Eyal Shimony, Natalia Vanetik, " Discovering Frequent Graph Patterns Using Disjoint Paths",*IEEE Transactions On Knowledge And Data Engineering*, Vol. 18, No. 11, November 2006.
8. Xiaozhe Wanga,, Ajith Abrahamb, Kate A. Smitha, " Intelligent web traffic mining and analysis" , *Journal of Network and Computer Applications* 28 (2005) 147–165.
9. JunWei Hsieh, Shih-Hao Yu, Yung-Sheng Chen and Wen-Fong Hu, "An automatic traffic surveillance system for vehicle tracking and classification, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 7, No. 2, 175-187, 2006.
10. Jing Yuan, Yu Zheng, Xing Xie, Guangzhong Sun, "Driving with Knowledge from the Physical World", *ACM 978-1-4503-0813-7/11/08*, August ,21–24, 2011
11. Benoît Depaire, Geert Wets, Koen Vanhoof, " Traffic Accident Segmentation by Means of Latent Class Clustering".
12. F. Atefeh and W. Khreich, "A survey of techniques for event detection in Twitter,*" Computer Intelligence*, vol. 31, no. 1, pp. 132–164, 2015.
13. J. Allan, "Topic Detection and Tracking: Event-Based Information Organization" Norwell, MA, USA: Kluwer, 2002
14. G. Anastasi *et al.*, "Urban and social sensing for sustainable mobility in smart cities," *in Proc. IFIP/IEEE Int. Conf. Sustainable Internet ICT Sustainability*, Palermo, Italy, 2013, pp. 1–4.
15. T. Sakaki, M. Okazaki, and Y.Matsuo, "*Tweet analysis for real-time event detection and earthquake reporting system development," IEEE Transactions on Knowledge on Data Engineering.*, vol. 25, no. 4, pp. 919–931, Apr. 2013.
16. R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proc. 20th VLDB Conf., Santiago, Chile, 1994, pp. 487-499.
17. M.-S. Chen, J.S. Park, and P.S. Yu., "Efficient Data Mining for Path Traversal Patterns,*" IEEE Transactions on Knowledge and Data Engineering*., vol. 10,no. 2, 1998, pp. 209–221.