Framework for Social Network Data Mining

Gayana Fernando
Management and Science University
Malaysiya

Md Gapar MdJohar
Management and Science University
Malaysiya

ABSTRACT
Social networks have become a vital component in personal life. People are addicted to social network features, updating their profile page and collaborating virtually with other members have become daily routines. Social networks contain massive collection of data. Web data mining is a new trend in the current research body. This conceptual paper introduces a framework that can be used to mine social network data. The proposed framework tries to handle the major limitations in current web mining frameworks by handling the unstructured and dynamic behavior of web data. Framework adopts the Hidden Markov Model to the data mining algorithm to predict the next status of web data.

Keywords
Social Networks, Web Data Mining, Framework, Social Network Analysis, Hidden Markov Model

1. INTRODUCTION
Social networks have opened a new path for people to communicate with each other. That has influence the traditional way of collaborating with others, marketing, running virtual businesses, teaching and learning, conducting researches and many more. Social networks have made to carry out all those activities in efficient and effective manner.

Large collection of data is available on Social networks. There are confidential information plus general details. General details comprise of members' interests, fan pages, birthdays, relationship status, networks etc. Most of the members of these networks display this information on their profiles. According to IEEE Spectrum research firm IDC the amount of data created globally in 2010 surpassed 1 zettabyte and that is enough to fill a billion 1-terabyte hard drives. This has made a new research area called big data analysis [1].

Social network analysis is the study of social networks to understand their structure and behavior.Social network analysis can be carried out using data collected from online interactions and from explicit relationship links in online social network platforms (e.g., Facebook, LinkedIn, and Flickr , etc.). On one hand, it has brought a huge increase in the availability and in the size of social network data and it has changed the types of data that can be collected and analyzed [2].Many researcheshave been carried out in social network analysis along with web mining techniques. This paper introduced a framework that can be used in social network data mining. In the first phase the research team carried out an empirical study on all web mining techniques [3]. During the study it was found out that many algorithms are developed in the web structure mining area. Web content mining, which is similar to traditional data mining, can be used to identify interesting patterns. So it was beneficial to use both mining algorithms together, as a hybrid approach. The framework is designed to overcome the disadvantages of the existing algorithms.

First the paper gives an overview of social networks and existing frameworks used to mine web data. Then it introduces a new framework. Finally it mentioned the usages, recommendations and future work.

2. SOCIAL NETWORKS
Social networks are defined as virtual spaces where people of all ages can make contacts, share information and ideas, and build a sense of community [4]. They allow the users to join as members, add friends, join to groups, and post their ideas and to virtually collaborate with others. Facebook, Twitter and LinkedIn are giants' among the social network sites.

H.Ma et.al states in "Mining Social Networks Using Heat Diffusion Processes for Marketing Candidates Selection" that millions of users participate in these social networks, and act as different roles. All of these social networks provide valuable information for decision making in marketing campaigns, especially in marketing of new products from start-up businesses [5]. Based on the Empirical analysis done in research Tucker et.al stated that when the private information of the profile holders is given to public is beneficial in advertising supported media and advertisers on social network sites [6].

3. WEB DATA MINING FRAMEWORKS
Web content mining is as mentioned above mining the content of the web pages. There are two approaches to do web content mining, as mentioned in paper [7]. Namely agent-based web mining systems having three variations like intelligent search agents, information filtering/categorization and personalized web agents. The second one is the databases approach with multilevel databases or web query systems [7].

Markov and Larose states that clustering and classification can be used as a solution to handle the unstructured nature of the web data. The clustering can be done by categorizing the web content in to groups based on their similarities. The classification can be done with the title, usage or the type if the web content. They suggest using concept learning methods to generate unique descriptions of sets of web content. This can be used to find the similar qualities of a new web page or a document [8]. In order to group or cluster the web content several techniques like Hierarchical Agglomerative Clustering, k-Means Clustering and Probability-Based Clustering can be used. To evaluate the data models generated by the clusters Similarity based criterion functions, Probabilistic criterion function, MDL based model and Feature evaluations can be used [8].

Yanagimoto in 2010 have proposed web page clustering method using social bookmarking data with dimension reduction regarding web pages' similarity by using k-means to partition web pages [9]. Rangaranjan et.al in 2014 have developed a clustering algorithm that groups users according to their Web access patterns. The algorithm is based on the ART1 version2 of adaptive resonance theory. ART1 offers an

unsupervised clustering approach that adapts to changes in users' access patterns over time without losing earlier information. It applies specifically to binary vectors [10].

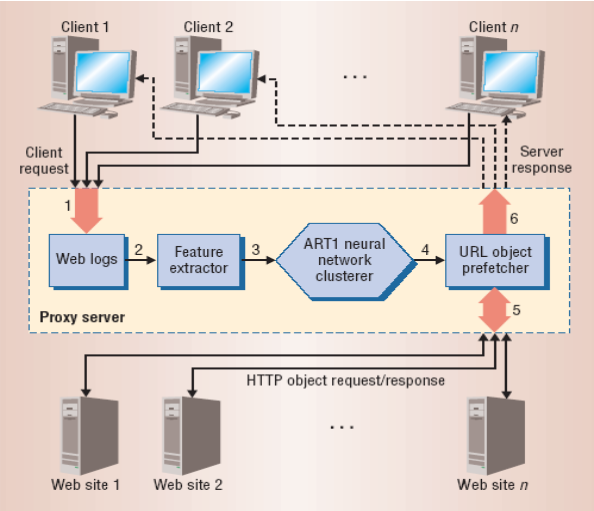

Fig 1- ART1-based clustering and prefetching architecture [10]

Aghabozorgi et.al in "Using Incremental Fuzzy Clustering to Web Usage Mining" presents a model that is made in "off-line" mode, and then they have change it to a dynamic model to predict user's interests in "online" mode as given in figure 2 [11].

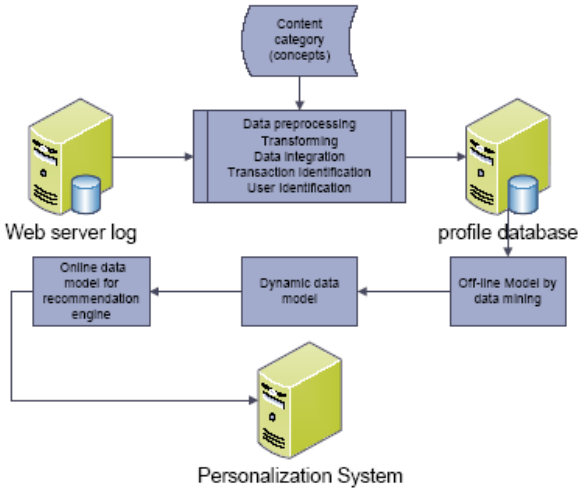

Fig 2 Summary of the architecture steps in usage mining [11]

4. PROPOSED FRAMEWORK

After performing an empirical study [3] the research team found that statistical methods like Markov Models can be used to tackle the dynamic behavior of web data. It can be used to cluster profile holders based on their qualities plus it can be used in customization. Customization would be beneficial in social network marketing [3].

4.1 Hidden Markov Model

A hidden Markov model (HMM) is a statisticalMarkov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states. An HMM can be considered as the simplest dynamic Bayesian network [12]. Figure 3 explains the architecture on the Model. HMM is a porpular techinique used in bioinformatics and character recogniztion. Da Silva and Ferreira in 2009 have conducted a research on appliyinh HMM for process mining. They have succefully used HMM with sequence clustering for log tracing [13].x — states y — possible observations a — state transition probabilities b — output probabilities

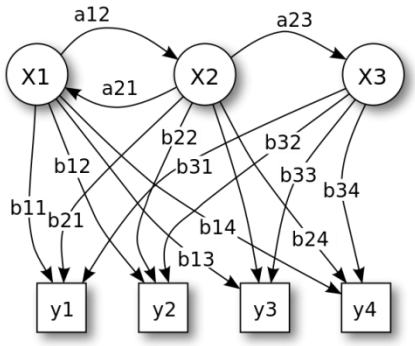

Fig 3 : Probabilistic parameters of a hidden Markov model (example) [12]

It is a probabilistic model, it would predict the behavior of a variable. Once sequences of observations are entered, the model would learn from them. Then the model can be treated as a black box, when new inputs are entered it would predict the relevant observation [14].

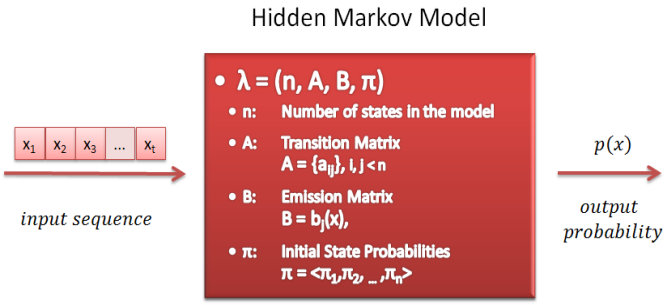

Fig 4 : Hidden Markov Model (de Souza, 2014).

Many new algorithms have been created based on HMM. Zaki et.al introduced a variable order hidden Markov model with state durations combining pattern mining and data modeling named VOGUE. This algorithm has been tested on web usage mining, intrusion detection and as a spell checker [15]. HMM is widely used in data mining. Andrea in 2006 mentioned following as the main characteristics of 1 using HMM in data mining.

1. Sequence matching for sets where elements differ for insertions, deletions and they can evaluate the associated penalty.
2. Model training can be unsupervised.
3. Work with variable length sequences.
4. Used for alignment purposes, data mining and classification, structural analysis and pattern discovery.
5. Used for recognition purposes in hierarchical processes.

However the major limitations would be when number of states increases the number of variable increase as well. This will make algorithm performance to decrease. The other limitation would be emission function and the transition

function are independent and both depend only on the hidden state [16].

4.2 Framework

The proposed framework consists of three major modules. The web crawler to extract the data from social network sites, the data repository and the data mining component based on HMM model. The web crawler is the software component developed to extract the data. The extracted details can be public information available in social network profile pages, for an example it can include birthdays, email, likes, interests, professional qualification etc. The crawler is capable of starting from a given profile page, and access the friends network. Different social networks provide different platforms for the developers to access the data available. For an example Facebook API provides social graph tool (Graph API) , robust well-structured platform to access user details. The developer needs to create an application in the social network in order to use relevant tools. Based on the access permission set by the profile holder, the data can be extracted. It will not violate privacy policies. These details would be stored in a database or excel files. This process would be executed periodically.

The data warehouse is created by importing the data. A data mining structure can be created with a mining model, where clustering based on HMM can be used as the algorithm. The output consists of the possible states each attribute can hold in future. The results can be saved in data warehouse. The new values can be mined again by using classification rules or clustering algorithms. The second mining algorithm can be changed based on the business interest and relent attributes. In order to enhance the system, graphical user interfaces can be designed to get business interest. The results of the miming model can be interpreted via a reporting service.

Figure 5 illustrates the structure of the proposed framework.

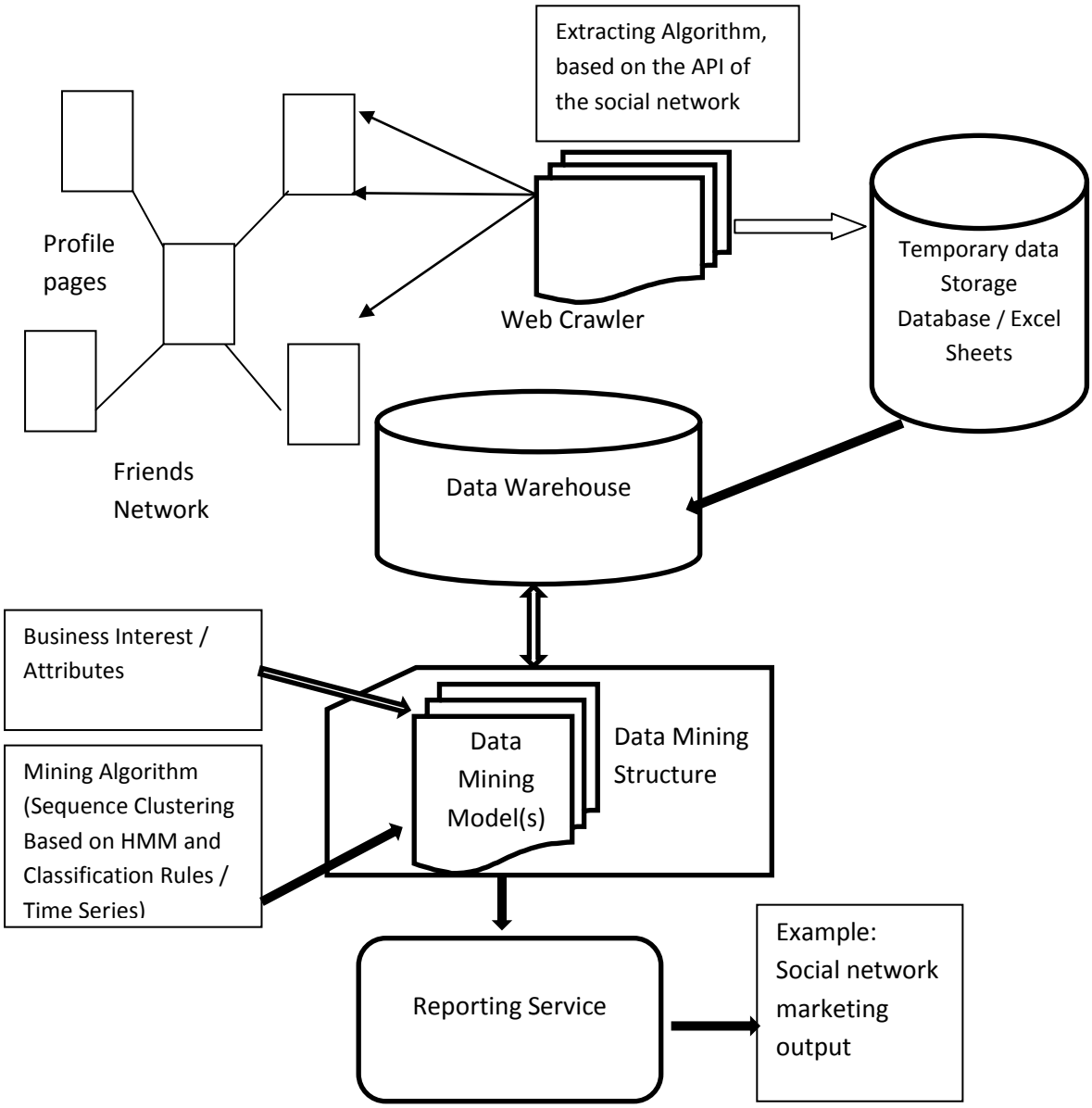

Fig 5. Proposed Framework

5. CONCLUSION

The proposed framework tries to handle the dynamic behavior by predicting the possible sates via a probability algorithm and mine the predicted results. It is still in conceptual level. The next step would be to implement a software system based on the framework proposed and test for the reliability and the performance. A comprehensive study need to be carried out to compare and contrast the system developed based on the framework with existing systems.

Since the proposed framework depends on the HMM, the mentioned limitations of the HMM model is applicable to the framework as well. Another limitation of the framework would be the scalability. The social network data should be accessed via their APIs. Accessing the social network data without the APIs provided would not be ethical. When a new social network to be added, then a new module should be created specifically to the social network API. The user of the system should have a basic knowledge on the social network platforms in order to configure the system. The result accuracy depends on the truthfulness of the data published by the profile holders of socialnetwork sites. If they have provided incorrect data, the predictions also would be inaccurate. If many profile holders are reluctant to provide information publically, the system might not be that useful.

This framework can be used by companies to predict target audience in social network marketing. Most of the companies maintain their company pages in common social network sites. The details of the users who follow the page can be taken as test data, and the system can be trained and used to predict new candidates to market and promote their service. The same logic can be applied to any field where you need to identify the patterns of social network data. For an example social science researches, this can be used in user profiling. It can be also used to analyses life patterns and behaviors of users.

6. ACKNOWLEDGMENTS

We sincerely thank DrNishanthaPerera for the help and guidance given. Further we would like to extend the gratitude to Professor Dr. Ali Khatibiall the encouragement given.

7. REFERENCES

[1] E. Ackerman and E. Guizzo, '5 technologies that will shape the web', IEEE Spectr., vol. 48, no. 6, pp. 40-45, 2011.

[2] F. Bonchi, C. Castillo, A. Gionis and A. Jaimes, 'Social Network Analysis and Mining for Business Applications', ACM Transactions on Intelligent Systems and Technology, vol. 2, no. 3, pp. 1-37, 2011.

[3] S. G. S. FernandoG. Johar, and S. N. Perera. "Empirical Analysis of Data Mining Techniques for Social Network Websites.", An Internatational Journal of Advance computer Technology, Vol 3, February 2014.

[4] Clemons, K. Eric, S. Barnett, and A. Appadurai. "The future of advertising and the value of social network websites: some preliminary examinations." Proceedings of the ninth international conference on Electronic commerce.ACM, 2007.

[5] M. Hao, H. Yang, M. R. Lyu, and I. King. "Mining social networks using heat diffusion processes for marketing candidates selection." Proceedings of the 17th ACM conference on Information and knowledge management, pp. 233-242.ACM, 2008.

[6] Tucker, E Catherine. "Social networks, personalized advertising, and privacy controls." Journal of Marketing Research 51, no. 5 (2014): 546-562.

[7] J. Srivastava, R. Cooley, M. Deshpande and P. Tan, 'Web usage mining', ACM SIGKDD Explorations Newsletter, vol. 1, no. 2, p. 12, 2000.

[8] Z. Markov and D. Larose, Data mining the Web. Hoboken, N.J.: Wiley-Interscience, 2007.

[9] H. Yanagimoto, M. Yoshioka, and S. Omatu."Web clustering using social bookmarking data with dimension reduction regarding similarity." Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on, pp. 386-390. IEEE, 2010.

[10] S. Rangarajan, V. Phoha, K. Balagani, R. Selmic and S. Iyengar, 'Adaptive neural network clustering of Web users', Computer, vol. 37, no. 4, pp. 34-40, 2004.

[11] S.R. Aghabozorgi , and T.Y. Wah. "Using incremental fuzzy clustering to web usage mining."Soft Computing and Pattern Recognition, 2009.SOCPAR'09. International Conference of, pp. 653-658. IEEE, 2009.

[12] Wikipedia, 'Hidden Markov model', 2015.[Online]. Available: http://en.wikipedia.org/wiki/Hidden_Markov_model. [Accessed: 14- Apr- 2015].

[13] Da Silva, G. Aires, and D. R. Ferreira. "Applying hidden Markov models to process mining." Sistemas e Tecnologias de Informação: Actas da 4ª ConferênciaIbérica de Sistemas e Tecnologias de Informação, AISTI/FEUP/UPF. 2009.

[14] C. Souza, 'Sequence Classifiers in C# - Part I: Hidden Markov Models - CodeProject', Codeproject.com, 2014. [Online]. Available: http://www.codeproject.com/Articles/541428/Sequence-Classifiers-in-Csharp-Part-I-Hidden-Marko. [Accessed: 14- Apr- 2015].

[15] M. Zaki, C. Carothers and B. Szymanski, 'VOGUE', ACM Transactions on Knowledge Discovery from Data, vol. 4, no. 1, pp. 1-31, 2010.

[16] M. Andrea. "Hidden Markov Models applied to Data Mining." (2006).