On-line Communities Making Scense - A Hybrid Micro-blogging Platform Community Analysis Framework

Cheng-Lin Yang and Yun-Heh Chen-Burger

Artificial Intelligence Applications Institute, Centre for Intelligent Systems and their Applications, School of Informatics, University of Edinburgh
{s0969605,jessicac}@inf.ed.ac.uk

Abstract. The upsurge of Micro-blogging platform attracts enterprises to use it as a public relationship tool. It also act as a new form of news source, journalists can hunt for next upcoming breaking news. It is worth to identify communities from it and reveal social relationships among community members in a timely manner. However, traditional SNA approaches are insufficient to achieve the requirement in a reasonable time. In this paper, we proposed a hybrid framework to tackle the problem. It is designed to identify the community with real social relationships automatically, that withstand dynamically changing content, have the ability to process fast and live-streaming data and provide a self-feedback mechanism to refine the result without human interference. The benefit of this framework is that average users should be able to employ it and to really understand communities in micro-blogging platforms without any or limited domain knowledge.

Keywords: Social Network Analysis, Micro-blogging System, Machine Learning

1 Introduction

What is community? Traditionally speaking, a community is a group of people who are gathered to embrace the same values or share the same responsibility. From a geographical perspective, a community can be defined by a street, a town or a specified area [1]. From the relationship aspect, a community is formed by human interaction [2]. To summarise, a community is a group of people who have the same values and live in a specified area. Members of the community feel dependence and through social interactions that build up the collective spirits.

With the rapid development of communication technology, the Internet has become an indispensable utility in daily life. People exchange information and knowledge on the Internet through various devices, forming a large social network and developing different types of on-line communities. The user with new communication technology uses the forum or blog system to share his knowledge or experience with multimedia resources. By using search engine such as Yahoo and Google, the user who is interested in particular topics can reach the information much quicker and easier. He is able to discuss the topic with users from different countries by Internet. Therefore, a new type of community is formed. In this paper, we call them on-line communities.

Members of an on-line community are not restricted to the same geographical area unlike the community as defined in the traditional sense. An on-line community can be defined as a social phenomenon formed by a group of people who communicate with each other through the Internet. The interaction on the Internet satisfies people's interests and fantasies as well as developing social relationships [7].

1.1 Upsurge of Micro-blogging platform

Micro-blogging platforms such as Twitter [3], Yahoo meme [4], Plurk [5] and Sina Weibo [6] have gained popularity since 2007. They quickly attracted the attention of many users, including many celebrities, politicians, and television and sports stars, who to share their daily life and personal opinions. Since the growth rate of registered users is staggering, enterprises have begun using it as a public relationship tool to announce news or provide the latest promotion information. These platforms also act as a new form of news source, journalists are able to hunt for next upcoming breaking news. For example, the first report of the emergency landing of US Airway airplane on the Hudson River was posted on Twitter.

When using a traditional blog system, e.g., Blogger and Wordpress, the user has to organize several paragraphs to form a blog-post, which becomes a problem if the user wants to share his current feelings in only a few words or sentences or to share interesting photos with a few comments. Posts like this would be considered odd by blog viewers since they expect a full article rather than a few sentences or photos without descriptions. A Micro-blogging system successfully fills the need of this kind of user. It is called "micro" because it normally limits either the word count or number of characters to 200 words. In addition, unlike the traditional blog system, where the user has to compose the article on personal computer, a micro-blogging system makes use of various communication technologies. The user can submit a micro-blog post from his desktop, laptop, tablet, smartphone and even by sending SMS. An individual micro-blog post can also contain video or image links with a few comments. Friends of micro-blog will be notified automatically by the system, so they can respond to the post immediately after it is posted.

2 Motivation and Challenges

Most micro-blogging posts are public, which means that everyone can view it and participate if they are interested in the topic. Any user can share his topics and let others participate in the discussion. These interactions form an on-line community. If we translate this behavior to the real world, most would find it very strange because people do not typically allow a stranger to join their daily life. The interesting way in which an on-line community forms through a micro-blogging system causes our curiosity. We want to understand the micro-blogging community deeply. However, we need to know the answers to several questions as follows: a) How is a micro-blogging community identified? b) How can a micro-blogging community be aligned with a human community, and how can we define those behaviours which cannot be aligned and understood? c) What kind of behaviour can be identified in a micro-blogging system? The challenges inherent in these questions are: a) The types of communities are complex, so how do we identify micro-blogging communities automatically, efficiently and accurately? b) The relationships of members lack definition, which means we need to carefully identify and interpret relationships that are important to the community. c) Unlike human communities, which have boundaries such as geographical boundaries, physical activities and gatherings, etc., on-line micro-blogging communities do not. Therefore, defining its boundaries will be challenging.

3 Related Work

The key to understanding the micro-blogging system is to identify communities behaviours of on-line communities. Social network analysis (SNA) is commonly used. SNA considers the human community and their relationships as a graph. Any node represents an individual, and the edge which connects two nodes indicates the relationship between two individuals [8]. By applying traditional clustering algorithms of graph theory like Hierarchical Clustering [9] and k-means [10], we can find the clusters in the given graph.

3.1 Traditional SNA Aprroach

Hierarchical Clustering treats each node as an unique cluster. It finds the most similar nodes, which is determined by the distance between two nodes calculated by the predefined distance function, and merges them as a new cluster. The procedure continues until all nodes are merged into new clusters. The benefits of Hierarchical Clustering are: 1) The algorithm is simple and easy to implement. 2) It does not require the centre point among all nodes. As long as we have the distance between the nodes, the algorithm is able to find the clusters in the given data. However, the drawbacks of Hierarchical Clustering are: 1) We can not predict how many clusters will be generated. 2) The computational complexity is $O(Ed\log N)$ where N is number of nodes and E is number of edges, which requires large amount of computing resource. When the given data is too large, the algorithm cannot generate the results in a reasonable amount of time.

Comparing the need of understanding the micro-blogging system, Hierarchical Clustering is able to generate a maximum number of clusters since it will merge the nodes close to each other into a cluster, which is beneficial since we cannot predict how many types of communities will emerge in the given data. However, in terms of the data size of a real micro-blogging system. Hierarchical

Clustering algorithm cannot deliver results in an acceptable timeframe, which renders this algorithm infeasible for our scenario since we need to identify the on-line communities in a realtime micro-blogging system.

In contrast, in k-means algorithm, k is the number of clusters we except to observe. Using that information, the algorithm will choose k data points randomly and consider those points as the centre point of clusters. After that, it calculates the distance from each data point to all centre points. Centre points then move to their closest data points. The procedure stops when no centre points can be moved. The benefits of k-means algorithm are: 1) Fast converge 2) The computational complexity is $O(NVk)$ where N is number of nodes and V is the number of vetoers, which means that requires less computing resource. Unfortunately, the drawback is that the result is easily affected by the initial centre point easily and gives a local maximum rather than a global one.

3.2 Local, Global and Score-based Approach

Researchers have discovered that with traditional methods, the definition of community is not well-defined. Hence, the new methods try to analyse the network based on its density. They believe that the relationships within a community will be much denser than those relationships outside the community. Based on this concept, new methods are proposed by other researchers.

Local Methods Since researchers believe that a community in a social network should have denser relationships, we should be able to find a subgraph that is closely related in the given graph. The most well-known method is to find the clique. A clique is a subset of a graph, which consists three or more nodes. Each node in the clique has a edge connected to all other nodes.

It would be meaningless if the discovered *clique* is too small. On the other hand, finding the largest clique is a NP-hard problem [11]. Moreover, the definition of clique is restricted. A clique can collapse if some of the edges are lost, which is easy to observe in practical data. In order to solve this problem, several clique relaxations were proposed, such as k-cliques [12], and k-clubs [13].

The original definition of clique restricts the distance from a node to all other nodes to be one. k-clique eases the distance limitation by requiring that the distances between any two nodes be less than k. Since k-clique does not require that the shortest path between two nodes must pass through k-clique itself, the shortest path may go through a node which does not belong to k-clique. Situations like this cause two problems: 1) The distances between k-clique nodes may be longer than k. 2) The nodes in k-clique are possibly not connected. In our scenario, the community identified by k-cliques may have loose relationships or even no relationship at all.

To overcome k-clique's problem, k-clubs [13] was proposed. The definition of k-clubs is that any distance between node pairs in k-clubs group must be less than k, and the paths of node pairs must go through k-clubs' group.

Global Methods Instead of finding cliques, global methods such as the G-N algorithm focus on calculating relationship density [14]. The procedure of the G-N algorithm is to calculate the betweenness of each edge in the given graph. The betweenness is used to measure the importance of a node between any other two nodes.

G-N algorithm will purge the edge with the largest betweenness each time until the graph is partitioned into several subgraphs. It is obvious that when applying the G-N algorithm, the betweenness calculations are performed repeatedly, which costs extensive computing time. The time complexity of the G-N algorithm is $O(N^E)$ where N is number of nodes and E is number of edges. The G-N algorithm is not suitable to large scale network.

[15] consider the social network as an electrical circuit, with the edge of the graph containing different levels of resistances. The main idea of this approach is that human relationships are fading out from layers of friends, which is exactly like the voltage growing smaller and smaller via flowing through the power circuit. The procedure of this approach is to select a node and assign one volt to it. Then assign 0 volt to randomly selected nodes from the network. The voltages of all nodes are calculated by applying Kirchoff's Laws. The value of each node should between 0 to 1. After that, a threshold is given and nodes have the value higher than the threshold belong to one cluster and the rest nodes belong to another. It can also be extended to find multiple clusters by given range of threshold.The computational complexity of this method is $O(N+E)$ where N is number of nodes and E is number of edges. Although the speed of it is attractive to us, two drawbacks are shown: 1) The quality of generated results is based on iteration time rather than the size of the graph, which will make it difficult for us to find the best value of iterations. 2) The number of generated communities needs to be defined before the procedure, which is not suitable for us since the number of communities in a social network is unknown.

Score-based Methods Algorithms suchs as PageRank [16], Hypertext Induced Topic Selection (HITS) [17] and Betweenness Centrality[18] are also used by researchers. The concept of them is to compute an authority score based on the relationships of nodes in the network. Betweenness Centrality was used by [18] to identify terrorist groups. Also, J. Qin and et. al., [19] applied the PageRank algorithm in their hunt for the global Salafi Jihad network. However, the dataset they used were relatively small, and their target result had clearly predefined attributes, which makes they unsuitable to our scenario.

3.3 Finding by Text

Q.-M Li and et.al, [20] and D. Shen and et al [21] believed that if two individuals have the same interests, they are possible to be *hidden* friends or in the same community. Besides on this assumption, they then collected the blog posts from the Internet and applied Latent Semantic Indexing (LSI) [22] on each post. Using keyword filters, the interests of users will be generated. Finally the users with the same interests will be considered to be in the same community. This approach lacks of consideration for user relationships, as people may have the same interests but never interact with other users, which make it hard for us to define them as being in the same community.

Using techniques such as LSI helps researchers to observe blog's features. Sometimes, the discovered features are dangerous, such as if they reveal that the owner of the tends to harbour racial discrimination or hatred of democracy or a particular country. However, the accurate level of danger of a blog owner cannot determined by his posts alone. M. Chau and J, Xu [23] believe that if a blog owner who has several dangerous features within his posts also recommends blogs or websites with dangerous features or add lots of blog owners who also have dangerous features as friends, he is highly likely to be a member of a criminal group.

4 Proposed Hybrid Framework

A hybrid system with a three-layered framework: collection, classification and reasoning layers. The architecture of proposed system is shown in Figure 1.

4.1 Collection Layer

The collection layer contains components that fetch the data from the micro-blogging system, process the raw data into a pre-defined format and convert the data into numerical parameters. The work will be performed by three components: crawler, database and data condenser. The requirement of them will be presented below.

Crawler The crawler is responsible for retrieving the user data from the concerned micro-blogging system. All fetched data will be stored in the storage for further usage. It is designed to be a lightweight but targeted daemon so it can be deployed on multiple machines easily to increase the throughput.

The Storage In order to support fast lookup and flexible schema, the proposed system will take advantage of a distributed key-value database system such as Cassandra [24] or MongoDB [25], which allows us to change the table schema without altering the entire table. Scalability is another concern for any system that handles tremendous amount of data. A traditional relational database, if to be used, needs to be well-designed before development and it may take a great deal of effort to expand the system scale. In contrast, a distributed database system provides a simple procedure to add new node into the system.

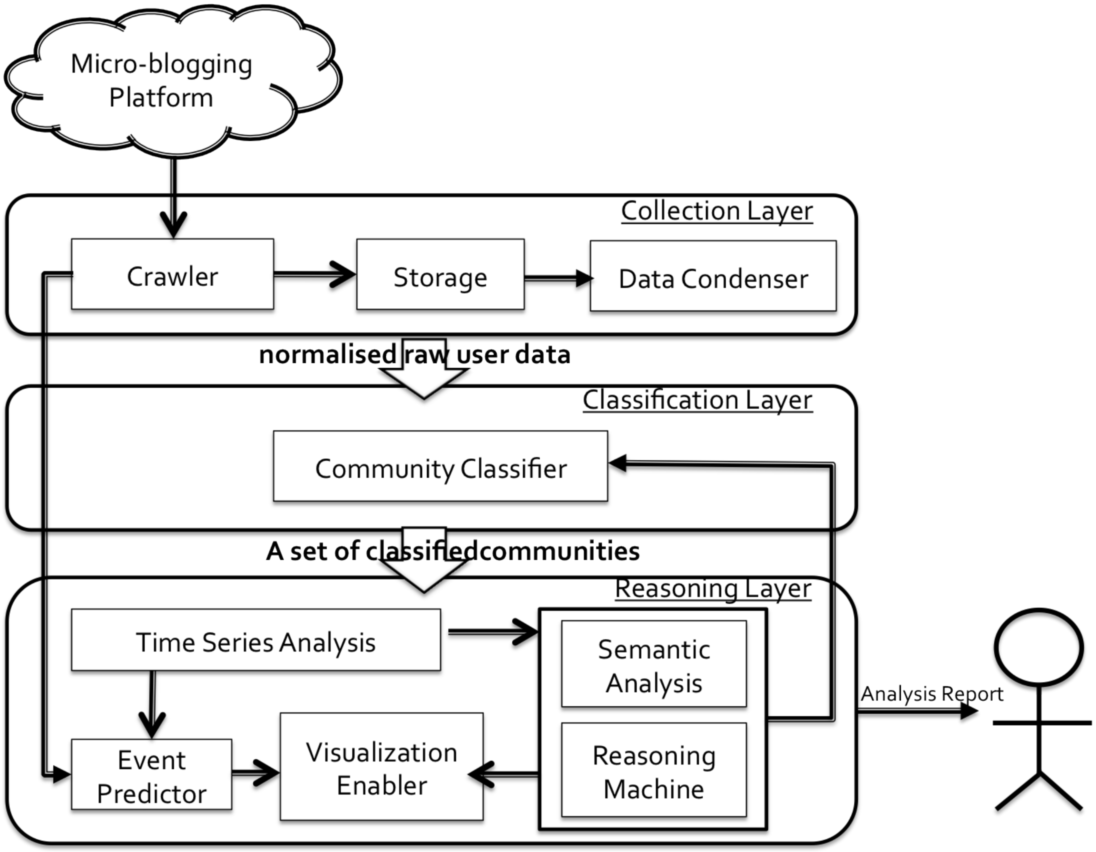

Fig. 1. Overview of proposed hybrid three-layered framework

Data Condenser The data condenser reads the raw data from the database. The raw data contains noises like auxiliary words, emoticons or random characters, so, it is the data condenser's responsibility to remove these noises. It is also responsible for converting and normalising the selected fields such as relationships among users into numerical parameters for the classification layer.

4.2 Classification Layer

The classification layer is the main interface which processes numerical parameters from the collection layer. It generates a set of classified communities to the reasoning layer that allows users to manipulate the final results easily.

Community Classifier The community classifier plays an important role in the proposed framework. The process flow can be summarised as follows. The crawler retrieves the raw data from the micro-blogging system and stores it in a database. After that, the data condenser fetches the raw data from the database. It removes all noises from the raw data and transforms the purified data into numerical parameters. Based on parameters, the community classifier applies the machine learning classification method to generate communities. This is then fed to the reasoning layer for further processing. The reasoning layer will analyse the results and text content with a logic-based reasoning engine, which will feedback to the community classifier to improve the quality of the result. Finally, the community classifier provides a finely tuned result to the user.

4.3 Reasoning Layer

The reasoning layer consists of a set of components which will infer the outcome of the classification layer based on its reasoning machine and feed the suggestion back to the upper layer. A framework is inaccessible if the user cannot visualise and manipulate the result. Hence, the reasoning layer also needs to provide a simple interface to the user.

Time Series Analysis The timestamp of each user's responses to a conversation or statement is a crucial factor to a community. It reveals how the user interacts with the community as well as revealing the characteristics of each member. Therefore, we plan to carry out several time-based anaysis on user interactions, e.g. the result from the classification layer will be analysed on the timestamp field. This then will be fed to other components as an additional parameter.

Event Predictor The value of an on-line community is that it reflects real daily life. The event predictor consists of a topic model based on observations from various micro-blogging conversations since each message within the conversation contains an unique timestamp. The topic model defines types of topics and how they grow through the timeline. By utilising the topic model, the event predictor will notify the user of which topics will be popular or even controversial in a given community.

Semantic Analysis The focus of our research is to understand the on-line community. Therefore, the identified community requires further investigation. Traditional techniques cannot handle large amount of content within a reasonable time. We simplify the content into categories. By utilising the user generated hashtag and keyword spotting along with the pre-defined dictionary, we are able to achieve the simple categorization. It improves the performance and provides us a flexible way to adjust the category if needed.

Reasoning Machine Reasoning machine is a logic rule-based inference engine with a set of rules which are given in advance and able to expand manually. It is used to verify the quality of the community generated by classification layer.

The engine examines the social relationships among community members, how the user interacts with others and so on. The reasoning machine will evaluate all factors to see if it needs to ask the community classifier to change the parameter settings. It is also an interrupter which converts the query from visualize enabler into logic rules and returns the result to the user.

Visualize Enabler A system will be inaccessible if the provided information cannot be easily understood by the user. The user may also want to make his own query to manipulate the result. The Visualize enabler tackles these problems by providing a simple user interface that makes the system as intuitive as possible.

5 Evaluation

Our proposed framework needs to be evaluated to see if it fits the requirements. The following criteria is designed to cover tests cases from different aspects.

1. Ground Truth
   A set of experts are invited to examine the generated result from our framework. The will examine the results and calculate the precision/recall rates to prove the accuracy of automatically classified communities generated by our framework.
2. Performance
   The test of performance is divided into two parts to cover overall execution and internal feedback speed. Candidate test cases to be considered are:
   - Feeding the realtime streaming data from the micro-blogging platform, it helps us to assess the overall performance on speed of our framework from the input to the final result. A threshold time is assigned and our system should generate the result faster than threshold time.
   - The design of our framework allows it adjust its model by reasoning machine. It should be agile enough to tune the model when new rules are applied. In this test, we will change the logic-based rule on reasoning machine. Our system should regenerate a new result in a reasonable time.
3. Robustness
   In a live micro-blogging system, its content usually contains certain noises such as spams or emoticons. Therefore, it is important to evaluate how accurate our framework to be performed under highly noise data. Our system needs to be able to handle the noise and the generated result should be acknowledged by human experts.
4. Usability
   A user interaction system should be manipulated by the user easily. Hence, in usability test, we will invite users who have rich and basic computer science background. Each user will score our framework from 1 to 5, which 1 means that he totally do not know how to use it and 5 represents that he feel trouble at all, after he uses the system. Our system should obtain 3.5 on average score on both groups.

References

1. F. T. Rothaermel and Sugiyama. Virtual internet communities and com- mercial success: individual and community-level theory grounded in the atypical case of timezone.com. Journal of Management, 27(3):297312, 6 2001.
2. J. Koh and Y. G. Kim. Knowledge sharing in virtual communities: an e-business perspective. Expert Systems, 26(2):155166, 2004.
3. Twitter. http://www.twitter.com/
4. Yahoo meme. http://meme.yahoo.com/
5. Plurk. http://www.plurk.com/
6. Sina weibo. http://www.weibo.com/
7. P. N. Romm, C. and C. R. Virtual communities and society: Toward and integrated three phase model. International Journal of Information Management, 17(4):261270, 1997.
8. S. Wasserman and K. Faust. Social network analysis: Methods and applications. Cambridge University Press, 1994.
9. B. S. Everitt. Cluster analysis. 1979.
10. J. A. Hartigan. Clustering algorithms. Probability and Mathematical Statistics, 1975.
11. M. Garey and D. Johnson. Computers and intractability: A guide to the theory of np-completeness. Freeman, 1979.
12. R. Alba. A graph-theoretic definition of a sociometric clique. Journal of Mathematical Sociology, 3:113126, 1973.
13. R. Mokken, cliques, clubs, and clans. Quality and Quantity, 13:161173, 1979.
14. M. Girvan and M. E. J. Newman. Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 99(12):78217826, 2002.
15. F. Wu and B. A. Huberman. Finding communities in linear time: A physics approach. The European Physical Journal B-Condensed Mater,, 38(2):331338, 2004.
16. L. Page, S. Brin, R. Motwani and T. Winograd. The pagerank citation ranking: Bringing order to the web. Stanford Digital Libraries Working Paper, 1998.
17. J. Kleinberg. Authoritative sources in a hyperlinked environment. Pro- ceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algo- rithms, pages 668677, 1998.
18. V. Krebs. Mapping networks of terrorist cells. Connections, 24(3):43 52, 2001.
19. J. Qin, J. Xu, D. Hu, M. Sageman, and H. Chen. Analyzing terrorist networks: A case study of the global salafi jihad network. Intelligence and Security Informatics, pages 287304, 2005.
20. Q.-M. Li, M.-W. Xu, J. Hou, and F.-Y. Liu. Web classification based on latent semantic indexing. Journal of Communication and Computer, 3(1):2427, 2006.
21. D. Shen, J.-T. Sun, Q. Yang, and Z. Chen. Latent friend mining from blog data. Proceedings of the Sixth International Conference on Data Mining, pages 552561, 2006.
22. Deerwester, Dumais, Furnas, Lanouauer, and Harshman. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41:391407, 1990.
23. M. Chau, J. Xu. Mining communities and their relationships in blog: A study of online hate groups. International Journal of Human-Computer Studies, 65(1):5770, 2007.
24. The apache cassandra project. http://cassandra.apache.org/
25. MongoDB. http://www.mongodb.org/