© 2016 IJSRSET | Volume 2 | Issue 2 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Ranking and Quick Access of Information Hybrid Content Based and Hmm in Health Care Social Media

S. Sudhakarilango, V. Hemalatha

Department of Computer Science, Anna University, Sri Krishna College of Engineering of Technology, Tamil Nadu, India

ABSTRACT

Internet has become a media of communication. Web 2.0 increases its' impact and make it as social media where people can share all the information. Using this people is forming a group, forum for discussion etc. This discussion group becomes a vital source of information. This forum discussion is available for all fields like medical, computer science, engineering and technology. Since millions and millions of people is using net as forum information evolved over a net is huge and decision making based on these huge information is also complex task. Hadoop framework provides a solution for this big data analysis. As a research work, forum of health care community is selected. In the forum, patients can post quires, share therapy followed, prescription prescribed. To make this forum more beneficial for beneficiary, the comments posted are ranked. For accurate ranking and quick access of information Content Based Mining along with hidden markov model (HMM) is used.

Keywords: Hadoop, (HMM) Hidden markov model, social media, Big Data

I. INTRODUCTION

Hadoop is an unlock-source software infrastructure for storing data and running applications on clusters of commodity hardware. It provides big storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks. The danger, of course, in running on commodity machines is how to handle failure. Hadoop is architected with the expectation that hardware will fail and as such, it can delightfuly handle lot of failures. Furthermore, its architecture allows it to scale nearly linearly, so as processing capacity demands increase, the only constraint is the amount of budget you have to add more machines to a cluster. At a high-level, hadoop operates on the philosophy of pushing analysis code close to the data it is aim to analyze rather than requiring code to read data across a network. A hidden Markov model (HMM) is a statistical Markov model in which the system being modelled is assumed to be a Markov process with undefined (hidden) states. A hidden Markov model can consists a common of a collaborative model where the latent variables, which control the combine component to be selected for each observation, are related through a Markov process rather than independent of each other. Currently, hidden Markov models have been generalized to pair wise Markov models and triplet Markov models which allow consideration of more complex data structures and the modelling of non-stationary data.

In the literature, there are a number of works related to influential user identification in social networks, namely, (i) influence maximization, (ii) social network construction and influence probability computation, (iii) expert findings, and (iv) combining content based and network based approaches.

Attribute to the revolutionary development of web 2.0 technology, individual users have become major contributors of web content in online social media. In light of the growing activities, how to measure a user's influence to other users in online social media becomes increasingly important.

The user influence is computed by combining content-based and network-based approaches. Online medical support forum is analyzed and a social network is constructed according to replying relationships and compute weights of edges by making using of forum

features. To quantify user influence two effective methods, User Rank and Weighted in-degree where used. Alcoholism forum and Smoking Addiction forum are considered for user influence computation. Weighted in-degree measure does not take into consideration the global structure of the network.

1. These algorithms utilize more memory.
2. Computational complexity is high.
3. Execution time of this algorithm is high.
4. Accuracy and efficiency is less.
5. Ranking accuracy is less.

II. METHODS AND MATERIAL

A. Related Work

Neelam Tyagi et al,[1] discussed the World Wide Web consists billions of web pages and hugs amount of information available within web pages. To retrieve required information from World Wide Web, search engines perform number of tasks based on their respective architecture. When a user refers a query to the search engine, it generally returns a large number of pages in response to user's query. To support the users to navigate in the result list, various ranking methods are applied on the search results. Most of the ranking algorithms which are given in the literature are either link or content oriented. Which do not consider user usage trends. In this paper, a page ranking mechanism called Weighted PageRank Algorithm based on Visits of Links (VOL) is being devised for search engines, which works on the basis of weighted pagerank algorithm and takes number of visits of inbound links of web pages into account. The original Weighted PageRank algorithm (WPR) is an extension to the standard PageRank algorithm. WPR takes into account the importance of both the inlinks and outlinks of the pages and distributes rank scores based on the popularity of the pages. The proposed algorithm is used to find more relevant information according to user's query. So, this concept is very useful to display most valuable pages on the top of the result list on the basis of user browsing behavior, which reduce the search space to a large scale. He also presents the comparison between original and VOL method.

Examining other forum features to further improve the performance of our technique. We shall investigate features that are particularly important for healthcare social media. For example, we shall utilize the ontology such as Unified Medical Language System (UMLS) and Consumer Health Vocabulary (CHV) to improve the content analysis component. We shall also utilize the user contributed tags related to symptoms, medications, and health conditions to enhance the link analysis.

Khalid Hussain Zargar et al [3]., discussed that text categorization is the task of assigning predefined category to a set of documents. Several different models like SVM, Naïve Bayes, KNN have been used in the past. In this paper we present another approach to automatically assign a category to a document. The approach is based on the use of Markov Models. It consider text as bag of words and use Hidden Markov Model to assign the most appropriate category to the text. The proposed approach is based on the fact that while creating documents the user uses the specific vocabulary related to the particular category. Hidden Morkov models have been widely used in automatic speech recognition, part of speech tagging, and information extraction but has not been used extensively for text categorization.

Anagnostopoulos et al.,[4] proposes two tests that can identify influence as a source of social correlation when the time series of user actions is available. In online social systems where social influence exists, ideas, modes of behaviour or new technologies can diffuse through the network like an epidemic. Therefore, identifying and understanding social influence is of tremendous interest from both analysis and design points of view. This is a difficult task because the factors such as homophily or unobserved confounding variables that can induce statistical correlation between the actions of friends in a social network. Distinguishing influence from these is essentially the problem of distinguishing correlation from causality, a notoriously hard statistical problem. To solve this defined a model that implements an aforementioned two tests. And give a theoretical justification of one of the tests by proving that with high probability it succeeds in ruling out influence in a rather general model of social correlation. To apply these test to real tagging data on Flickr.

Matthew Richardson et al., [5] optimize the amount of marketing funds spent on each customer, rather than just

making a binary decision on whether to market to him. Taken into account the fact that knowledge of the network is partial, and that gathering that knowledge can itself have a cost. Showed how to find optimal viral marketing plans, use continuously valued marketing actions, and reduces computational costs.

B. Methodology

The discussion group becomes a vital source of information. This forum discussion is available for all fields like medical, computer science, engineering and technology. Since millions and millions of people is using net as forum information evolved over a net is huge and decision making based on these huge information is also complex task. Hadoop framework provides a solution for this big data analysis. As a research work, forum of health care community is selected. In the forum, patients can post quires, share therapy followed, prescription prescribed. To make this forum more beneficial for beneficiary, the comments posted are ranked. For accurate ranking and quick access of information Content Based Mining along with hidden Markova model (HMM) is used. It has more of merits with comparing existing related systems are; Hidden markov model measure considers the global structure of the network and has good ranking accuracy.

C. System Implementation

Thus the overall system implementation classified as following four modules, Collection of data using web crawler, Construction of social network, Ranking user based on hidden markov model, Performance Evaluation.

D. Collection of Data Using Web Crawler

A Web crawler is an Internet bot which step by stem browses the World Wide Web, typically for the purpose of Web indexing. A Web crawler another name is a Web spider, in the forum data collecting phase, a crawler was built to collect all threads and replies on the discussion board of the forum of interest. In addition, a parser was built to parse and filter the collected data. For each thread, we generated a formatted thread record which consisted of TID (unique ID for each thread), Thread Title, Thread URL, Thread Initiator ID, Timestamp, List of Replier ID and Thread Content. For each reply of a thread, we created a formatted reply record which was composed of RID (unique ID for each reply), Thread Title, Thread URL, Replier ID, Timestamp and Reply Content. The formatted data was stored in a database which provided inputs to the social network constructing phase.

1. Choose a framework/library/language
2. To need something to make HTTP GET requests to the pages in question.
3. The GET will return you html data. Use whatever language/framework you chose to parse out the data you are interested in
4. store crawled data (in a database, or xml file, or text file etc)

E. Ranking User Based On Hidden Markov Model

Given the weighted social network G', proposed two different approaches of computing user influence

i. Weighted in-degree and User Rank
ii. Content Based Mining along with hidden markov model

In a directed graph, In-degree of a node is the number of head endpoints adjacent to this node. With edge weights computed, weighted in-degree is a straightforward way of computing user influence. Given a weighted social network, user's influence score is equal to the sum of weights on all in-link edges of the network.

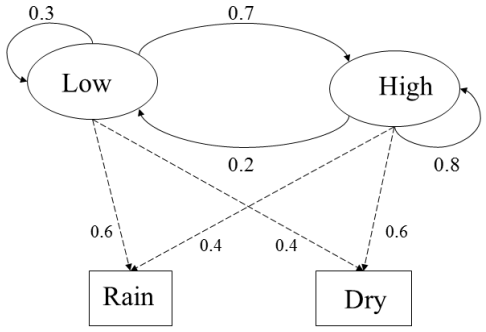

Figure 1. Example of HMM

Since the user influence within a social network is similar to the web page popularity in a hyperlink network, User Rank algorithm is used to quantify user influence in a weighed social network that is constructed.

To incorporate the content similarity and response immediacy, content base mining with hidden markov model is used. In content base mining, based on the content similar users are identified and the identified similar users are grouped by hidden markov model. By

this approach user influence over the medical forum is evaluated and ranked efficiently.

III. RESULTS AND DISCUSSION

A. Construction of Social Network

Social network is a convenient and effective way to represent user interactions. Each vertex of a social network represents a social actor. Two social actors who are interacting with each other are connected by an edge in a social network. Depending on the specific applications and interactions, a social network can be constructed in different way.

A forum consists of a number of threads. A forum thread is composed of a number of messages. A social network is constructed by extracting the users and their interactions in a hierarchical tree of a thread based on three observations:

1. Direct reply from a user to another user in a thread represents an interaction. In other words, when A replies to a message posted by B, there should be an edge connecting vertex A and vertex B in the social network to capture their interaction.
2. Indirect reply is B replies to A and then C replies to B, it is possible that C is not only replying to B but also addressing to the message posted by A. In this case, three edges should be created in a social network corresponding to the interactions between A and B, B and C, and A and C.
3. Most of the threads in medical support forum is initialized by the members who are seeking information help. When B is replying the question posted by A, it is considered that B is offering some kinds of support and influence to A. As a result, the direction of an edge should be made from the one who receives the reply to the one who makes the reply to confer authority.

A weight function is incorporated by both content similarity and response immediacy to compute weights, leading to a weighted social network.

$$G' = (V,E,W)$$

where the node set V is a set of nodes corresponding to the members of a forum and edge set E is a set of edges corresponding to the interactions between members and the weight set W corresponds to a collection of weights $\{w_{i,j}\}$, for each edge in E.

Given a forum, there are a collection of N threads, , , , , which consist of messages posted by n users $v_1, v_2, \ldots v_n$. Let $M_{k,l}$ to be the l^{th} message of , $V(M_{k,l})$ to be the user who posts , and time $(M_{k,l})$ to be the timestamp of message . In addition, let ( ) denotes the content similarity between messages $M_{k,a}$ and $M_{k,b}$, and $(M_{k,a}, M_{k,b})$ represents the response immediacy between two messages. The weight $W_{i,j}$ of edge $e_{i,j}$ between v_i and v_j is computed.

B. Performance Evaluation

Table 1. Statistics of Collected Datasets

	High Blood Pressure Forum1	Heart Disease Forum2
Total Number of Registered Members	500	482
Average Number of Friends for each Members	20	16
Total Number of Threads	800	600
Total Number of Replies	3000	1777
Average Number of Replies per Thread	6.7	8.3

1.URL:
http://www.medhelp.org/forums/Heart-Rhythm/show/92
2.URL:
http://www.medhelp.org/forums/Heart-Disease/show/72

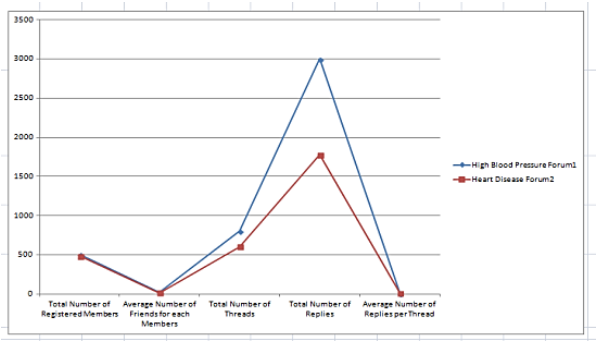

Figure 2. Statistics of Collected Datasets

In above a crawler to fetch all HTML web pages from the discussion board of two medical support forums, including high blood pressure forum and heart disease forum. A parser was also built to extract threads, replies, timestamps, and user information from each HTML page. Table 1 describes the statistics of the two collected datasets.

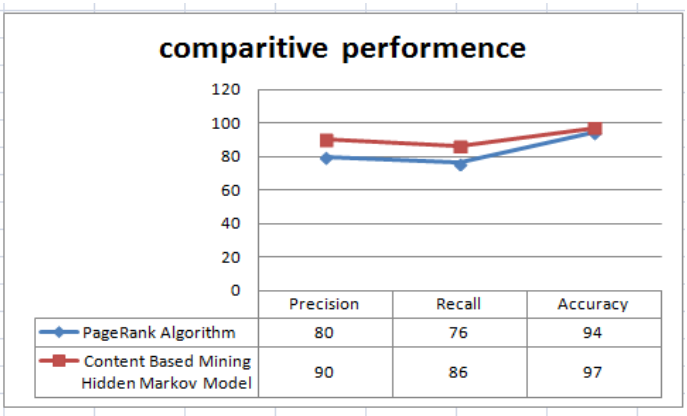

Figure 3. Comparison of Weighted Page rank And Hmm Algorithms

IV. CONCLUSION

User influence on medical forum is analyzed using link and content based approach. A social network is constructed by collecting data using a crawler which collects all threads related to the user conversation. From the collected data set, social network is constructed. From the constructed network, user behaviour is understood by assigning a weight to the links between user and the conversation. User influence is quantified using weighted-in degree and user rank algorithm. Along with this procedure, the content based mining is used with hidden markov model. Content based mining is used to mine a people of similar behaviour and markov model is used to group similar behaviour user.

V. FUTURE WORK

As a future work, ontology can be built for the user and communication relation to make a form. Ontology can be trained to retrieve a relation easier and efficiently than algorithm. Along with this, some other forum features and characteristics can examine. System is developed to accommodate further changes made.

VI. REFERENCES

[1] Neelam Tyagi, Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Page, IJSCE, ISSN: 2231-2307, Volume-2, Issue-3, July 2012.

[2] Wenpu Xing and Ghorbani Ali, "Weighted Page Rank Algorithm", Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR '04), IEEE, 2004.

[3] Khalid Hussain Zargar ,Content Based Text Classification Using Morkov Models, International Journal of Scientific Engineering and Research, ISSN (Online): 2347-3878, 2014.

[4] J.Premkumar, Image Retrieval using Markovian Semantic indexing (MSI), International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 4 Issue 4 April 2015.