ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 4, April 2014

Predictive Hot set Identification in Social Networks: Approach

Gayatri Kabra[1] ,Mangesh Wanjari[2]

M.Tech. Student, Department of Computer Science & Engineering, Shri Ramdeobaba College Of Engineering & Management, Nagpur, India [1]

Assistant Professor, Department Of Computer Science & Engineering, Shri Ramdeobaba College Of Engineering & Management, Nagpur, India [2]

ABSTRACT: Social networks are well known class of Web-based services that are characterized by novel patterns of access where user operations are not limited to navigation but interaction, knowledge and resource sharing among communities of online users. Here online users upload resources, insert short comments, create links with other users. As the social networks is growing hugely on day-today basis, operating on entire working set is expensive in terms of network, storage and computational power. So it will be beneficial to work just or mainly on the hot set. Hot set is the set of resources that are expected to receive the majority of requests in the near future. These Hot sets can be used by user to get answer for their query. Social networking sites like job sites, technical forums, mobile sites, different product sites, matrimonial sites etc. are used by many users for asking queries in terms of comments. In our approach comments and answers for these comments are been collected and processed with Predictive analytics. As we have large data on social networks Apache-Hadoop is for managing vast amount of data. The generated hot sets can be used to answer query precisely

KEYWORDS: Hot sets, Social network's services, Predictive analytics, Social connection.

I. INTRODUCTION

Social networks has access pattern which requires require a re-design of the traditional strategies for an efficient content management. Some of the strategies for these are catching, replication, pre-fetching etc. of contents. Basis for all these strategies is to determine subset of resources that are expected to receive more requests in near future (the so called *hot set)*.Most algorithms for the hot set identification were dependent on historical data i.e. information about the past resources accesses. They can achieve good results in older web based services where resource popularity doesn't changes frequently. It changes very slowly according to known patterns. One is unable to use these algorithms in social networks where social connection of users are also matters. In social networks presence of user generated resources such as links, comments and uploads along with social connection of users lead to quite new access patterns thus causing frequent changes in popularity of data.

A previous study shows that by adopting predictive models and by taking into account the characteristics of user social connections the accuracy of the hot set identification in social network services can be improved with respect to traditional solutions. However, as we have heterogeneous information scattered over social networks it's very difficult to work with this data. How to merge it in an efficient way on the basis of s prediction on future accesses is a challenge. This is the reason here we have considered review data from different mobile products sites. These reviews consist of erroneous entries in terms of spellings, blank spaces, multiple dots. To solve this problem by building our own dictionary of common as well as uncommon words is still an open issue.

The main contribution of this paper is the proposal of approach for algorithm for deriving hot sets by considering predictive analytics as well as social connection among users. Approach is based on techniques to automatically tune the process of merging predictive- and social-based information according to dynamically variable workload. Accuracy for Hot set identification close to theoretical algorithm can be achieved by our proposed algorithm without getting affected by wide range of parameters. On the other hand, we can have identified hot sets to be correct even after data get added dynamically.

The remainder of the paper is organized as follows. Section 2 describes the related work in identifying the resource hot set for social network services. Section 3 presents the research methodology. Section 4 describes the architectural flow of complete system. Section 5 concludes the paper.

II. RELATED WORK

The combination of user access patterns social connection opens novel issues related to efficient content management in social network services. Our research focuses on identifying hot sets from most popular resources for content management in terms of cost. For traditional Web-based applications, where the resource popularity is been determined on the basis of event sequences and cannot be changed dynamically prediction is based on episode formation of event sequences according to certain order [3]. In this case analysis is based on past resource accesses only. Another traditional approaches include working on log databases which are already available in the form of click through history [8]. These were insufficient to work with dynamically and fast changing workload characteristics. However, in social networks the resource popularities are changing rapidly due to frequent resource uploads and to the social connections among users, thus challenging the effectiveness of already existing static algorithms. These results motivate the need of integrating predictive- and social-based techniques in the algorithms for the hot set identification.

The impact of social connections has been demonstrated by several studies in the last few years [2,5]. A first attempt to combine predictive and social connections for the identification of hot sets has been investigated by the authors in [2]. This study demonstrates the potential benefits of considering both of these metrics to accurately identifying the hot sets as compared to performance of algorithms considering just one metric. The most straightforward approach for this is by a linear combination of metrics. An alternative approach is to use rank merging algorithms to combine heterogeneous metrics, as in the context of search engines. Static coefficients are not suitable for a highly variable Workloads of social network services, because they cannot ensure stable performance. Hence, adaptive mechanisms comes into focus. In [10] the adaptive technique is used to improve the performance of a ranking algorithm for commercial search engines. On the basis of these results, we introduce adaptive techniques to combine predictive social connections to improve the robustness Hot sets in presence of highly variable workloads..

III. RESEARCH METHODOLOGY

Social networks are emerging trends in IT industries which have volume and variety of information exploding over it. Turning this content into useful actionable information is core issue for social networks services. To manage this data cost effectively we need to analyze it. Here predictive Big data analytics comes into focus.

It is evident from the progress of our survey that none of the previous strategies which are based on historical data are not so effective in predicting proper Hot sets as per user's need. So here comes into account a new style of analyzing data which is based on Predictive analytics of dynamically changing worksets and reverse contacts i.e. social connections are also been focused.

Predictive analysis as the name indicates is process of determining events or outcomes before it happen. It is forward looking using past events to anticipate the future which is used to pull out meaningful relationships. Identifying set of resources which are expected to receive majority of requests is basis for this strategy. It basically rely on information about the past accesses of resource. By taking into account user social connection one can improve accuracy of the hot set identification in social network services.

According to design, rank merging technique is used to merge both predictive & social ranks or metrics obtained from data. Some probabilistic formula can be used to merge rank based on some weight factor. Weight factor indicates the importance is to be given the two metrics based on dynamically changing parameter loads.

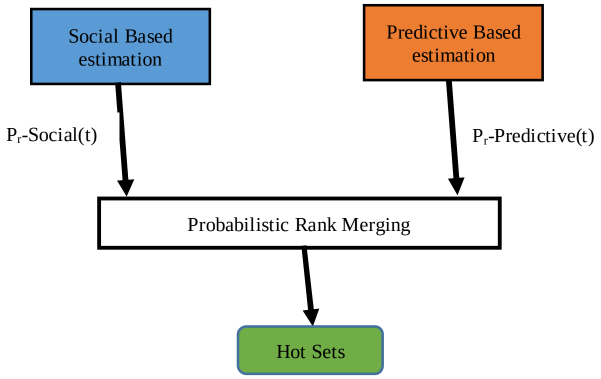

Fig. 1.Design of Predictive-Social method

IV. PROPOSED WORK

Proposed system aim is to design and develop a predictive hot set identification system from user interaction:

- Data collection from social networking sitesis the first step.
- Apache-Hadoop implementation comes into consideration for managing vast amount of data. As we know petabytes of heterogeneous kind of data is been coming over social networks, it becomes quite difficult to manage this large amount.
- Preprocessing operations must be performed on data to keep required data and discard the rest. Preprocessing is also helpful in erroneous data corrections.
- After preprocessing, one is left with exact data which is been required for further processing. Still the data range is wide so one need to identify hotsets.
- Predictive-social connection based technique can be used to identify hotsets(as shown in fig.2.0).
- Proposed system will try to solve the user query on the basis of previous transaction matching to currently asked query.
- This will help user to get quick response on the basis of user experience about the query. Analysis of Hotsets based on user's query is done and these Hotsets can be tested dynamically as the data grows over the system.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 4, April 2014

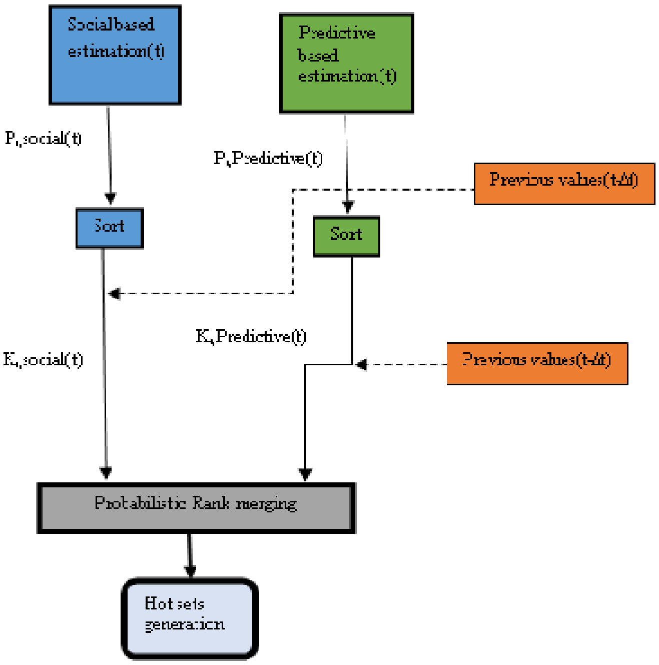

Fig. 2Proposed system of Predictive-Social method

V. ARCHITECTURE

The architectural flow shown in fig. 3 is mainly comprised of following phases:

- Data Collection and Preprocessing
- Hot Set Generation
- Identification of Hot sets as per user's query

Data Collection and Preprocessing:

In this phase social networks data is been collected by any extraction method. Social networks data constitute different classes such as job sites, matrimonial sites, product reviews, tweets, uploads and comments on social networks and so on. The collected data may include structured, unstructured, semi-structured data which is not suitable for further processing. Because of this reason it is mandatory to preprocess data. Preprocessing consist of any technique that lead to keep required data and discard the rest.

Ex. Some review data can have comments like "thank you!!", blank spaces, multiple dots etc. these are needed to be removed. Preprocessing can be used for this purpose.

In other cases user may write wrong spellings of certain common words. These spellings must be corrected in order to get desired output.

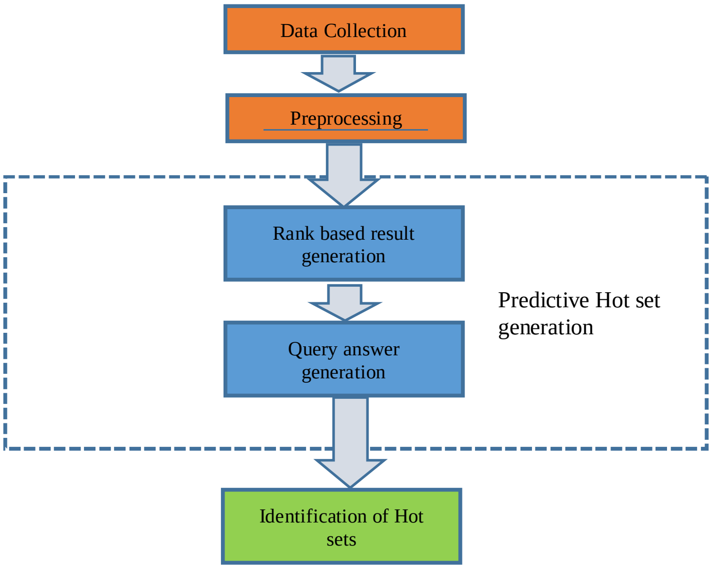
Data Collection
Preprocessing
Rank based result generation
Query answer generation
Predictive Hot set generation
Identification of Hot sets

Fig. 3. Architectural Flow of Overall System

Hot Set Generation:

The proposed algorithms for the hot set identification operates on the whole working set R(t) of the social network and tries to compute the expected resource popularity pr(t) for every resource r.

Predictive based algorithm uses social networks data and gives resource popularity pr,predictive(t). The Social-based algorithm uses information on the user social connections i.e. strong correlation between the amount of accesses to a resource and the number of social links by the user. Here, the number of reverse contactsof the users can be considered as a measure of their social network size, where a reverse contact for A is a user that has designated A as a contact. The algorithm uses the number of reverse contacts as a measure of the resource popularity pr,social(t), because resources uploaded by users with many social connections are likely to receive more accesses in the future.

The probability pr(t) by considering both predictive and social ranks obtained from collected social networks data can be computed as:

$$\mathbf{pr(t) = \gamma(t)kr,predictive(t) + (1 - \gamma(t))kr,social(t),}$$

where,

kr,predictive(t) is predictive rank & kr,social(t) is social rank.

The weight γ(t) is computed by taking into account the workload characteristics.

Identification of Hot sets as per user’s query:

On social networks heterogeneous kind of data is been scattered in huge amount. So, it is quite difficult for user to search for required data. Here in final phase Hot sets will be identified as per user’s query so that one can get the data desired.

VI. CONCLUSION AND FUTURE WORK

In this paper, we propose an approach for hot set identification through an algorithms that exploit an adaptive combination of predictive models and social information to estimate the hot set. Our study demonstrates that any static approach achieves poor performance or too variable results for different workload and scenarios. Indeed, the proposed adaptive algorithms achieve performance close to the ideal algorithm and guarantee robust performance with respect to any considered workload parameter.

REFERENCES.

1. Claudia Canali, Michele Colajanni, Riccardo Lancellotti, University of Modena and Reggio Emilia. "Adaptive Algorithms For efficien Content Management in Social Network Services"
2. Claudia Canali, Michele Colajanni, Riccardo Lancellotti, Department of Information Engineering, University of Modena and Reggio Emilia"Hot Set Identification For Social Network Applications"
3. Yonatan Aumann, Oren Etzioni, Ronen Feldman, Mike Perkowitz, Tomer Shmiel. "Predicting Event Sequences: Data Mining for Prefetching Web-pages"
4. www.cloudera.com/hadoop-training.
5. K. Lerman and L. Jones. "Social Browsing on Flickr", In *Proc. of ICWSM Conference*, March 2007.
6. www.bigdatauniversity.com.
7. Aron Culotta, Ron Bekkerman, Andrew McCallum, *University of Massachusetts –Amherst* "Extracting social networks and contact information from email and the Web"
8. "Predicting Category Accesses for a User in a Structured Information Space" By Mao Chen, Andrea S, LaPaugh, Jaswinder Pal Singh, Department of Computer Science Princeton University Princeton.
9. Benjamin Piwowarski, Hugo Zaragoza, Yahoo! Research, Barcelona, Spain. "Predictive User Click Models Based on Clickthrough History"
10. R. Zhang, Y. Chang, Z. Zheng, D. Metzler, and J.-y. Nie. "Search result re-ranking by feedback control adjustment for time-sensitive query". In *Proc. of Human Language Technologies Conference (HLT'09)*, June. 2009.

BIOGRAPHY

Gayatri Kabra has received his B.E. degree in Information Technology from DKTES Textile and Engineering Institute Ichalkaranji Shivaji University in 2011. She is pursuing M.Tech in Computer Science and Engineering from Shri Ramdeobaba College of Engineering and Management (Autonomous), Nagpur. Her research interests include big data predictive analytics on social networks.

Mangesh Wanjari has received his B.E. in Computer Technology from Nagpur Univeristy in 2002. He has received his Master of Technology (MTech) in Computer Science and Engineering from VNIT, Nagpur in 2009. After having some industrial experience he has joined the teaching field. He is an associate professor in Ramdeobaba College of engineering and Management, Nagpur. His research interests include Database Technologies, Query Optimization and Semantic Analysis.