RESEARCH ARTICLE OPEN ACCESS

Webpage proposal by integrating web usage and content mining

Mahalakshmi.v[1], A.Petrisia[2]
Dhanalakshmi srinivasan college of engineering and technology
Department of Master of Computer Applications

Abstract:

Web-page recommendation plays an important role in intelligent Web systems. Useful knowledge discovery from Web usage data and satisfactory knowledge representation for effective Web-page recommendations are crucial and challenging. This paper proposes a novel method to efficiently provide better Web-page recommendation through semantic-enhancement by integrating the domain and Web usage knowledge of a website. Two new models are proposed to represent the domain knowledge. The first model uses an ontology to represent the domain knowledge. The second model uses one automatically generated semantic network to represent domain terms, Web-pages, and the relations between them. Another new model, the conceptual prediction model, is proposed to automatically generate a semantic network of the semantic Web usage knowledge, which is the integration of domain knowledge and Web usage knowledge. A number of effective queries have been developed to query about these knowledge bases. Based on these queries, a set of recommendation strategies have been proposed to generate Web-page candidates. The recommendation results have been compared with the results obtained from an advanced existing Web Usage Mining (WUM) method. The experimental results demonstrate that the proposed method produces significantly higher performance than the WU method.

I.INTRODUCTION

Million users and have become fertile ground for a variety of research efforts, since they offer an opportunity to study patterns of social interaction among far larger populations than ever before. In particular, Twitter has recently generated much attention in their search community due to its peculiar features, enormous popularity, and open policy on data sharing. Along with the growth in reach of micro blogs, we are also observing the emergence of useful information that can be mined from their data streams. However, as micro blogs become valuable media to spread information, e.g., for marketers and politicians, it is natural that people find ways to abuse them. As a result, we observe various types of illegitimate use, such as spam. we focus on one particular type of abuse, namely political astroturf — campaigns disguised as spontaneous, popular “grassroots” behavior that are in reality carried out by a single person or organization. This is related to spam but with a more specific domain context, and with potentially larger consequences. The importance of political astroturf stems from the unprecedented opportunities created by social media for increased participation and information awareness among the Internet-connected public Online social media tools have played a crucial role in the successes and failures of numerous political campaigns and causes, from the grassroots organizing power of Barack Obama’s presidential campaign, to Howard Dean’s failed 2004 presidential bid and the first-ever Tea Party rally . Moreover, traditional media pay close attention to the ebb and flow of communication on social media platforms, and with this scrutiny comes the potential for these discussions to reach a far larger audience than simply the social media users.

While some news coverage of social media may seem banaland superficial, their focus is not without merit. Social media, such as Twitter, often enjoy substantial user bases with participants drawn from diverse geographic, social and political backgrounds Moreover, the user-as-information-producer model provides researchers and news organizations alike a means of instrumenting and observing, in real-time, a large sample of the nation’s political participants. So relevant

is this discursive space, in fact, that the Library of Congress has recently undertaken the project of archiving a complete record of the discourse produced by Twitter users. Despite the benefits associated with increased information availability and grassroots political organization, the same structural and systematic properties that enable

2. SYSTEM STUDY

2.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential. Three key considerations involved in the feasibility analysis are

- ECONOMICAL FEASIBILITY
- TECHNICAL FEASIBILITY
- SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

3.Literature Survey

Predicting the future with social media.
Technical Report:

social media has become ubiquitous and important for social networking and content sharing. And yet, the content that is generated from these websites remains largely untapped. In this paper, we demonstrate how social media content can be used to predict real-world outcomes. In particular, we use the chatter from Twitter.com to forecast box-office revenues for movies. We show that a simple model built from the rate at which tweets are created about particular topics can outperform market-based predictors. We further demonstrate how sentiments extracted from Twitter can be utilized to improve the forecasting power of social media.

Dynamical Processes on Complex Networks. Cambridge:

The availability of large data sets have allowed researchers to uncover complex properties such as large scale fluctuations and heterogeneities in many networks which have lead to the breakdown of standard theoretical frameworks and models. Until recently these systems were considered as haphazard sets of points and connections. Recent advances have generated a vigorous research effort in understanding the effect of complex connectivity patterns on dynamical phenomena. For example, a vast number of everyday systems, from the brain to ecosystems, power grids and the Internet, can be represented as large complex networks. This new

and recent account presents a comprehensive explanation of these effects

Determining the public mood state by analysis of micro blogging posts:

Extended Abstract Micro blogging is a form of online communication by which users broadcast brief text updates, also known as tweets, to the public or a selected circle of contacts. A variegated mosaic of micro blogging uses has emerged since the launch of Twitter in 2006: daily chatter, conversation, information sharing, and news commentary, among others (Java et al, 2007). Regard-less of their content and intended use, tweets often convey pertinent information about their authors mood status. As such, tweets can be regarded as temporally-authentic microscopic instantiations of public mood state (O'Connor et al, 2010). Here we perform a sentiment analysis of all public tweets broadcasted by Twitter users between August 1 and December 20, 2008. For every day in the timeline, we extract six dimensions of mood (tension, depression, anger, vigor, fatigue, confusion) using an extended version (Pepe and Bollen, 2008) of the Profile of Mood States (POMS), a well-established psychometric instrument (Norcross et al, 2006; McNair et al, 2003). We compare our results to fluctuations recorded by stock market and crude oil price indices and major events in media and popular culture, such as the U.S. Presidential Election of November 4, 2008 and Thanksgiving Day (see Fig. 1). We find that events in the social, political, cultural and economic sphere do have a significant, immediate and highly specific effect on the various dimensions of public mood. In addition, we found long-term changes in public mood that may reflect the cumulative effect of various underlying socio-economic indicators. With the present investigation (Bollen et al, 2010), we bring about the following methodological contributions: we argue that sentiment analysis of minute text corpora (such as tweets) is efficiently obained via a syntac-tic, term-based approach that requires no training or machine learning. Moreover, we stress the importance of measuring mood and emotion using well-established instruments rooted in decades of empirical psychometric research. Finally, we speculate that collective emotive trends can be modeled and predicted using large-scale analyses of user-generated content but results should be discussed in terms of the social, economic, and cultural spheres in which the users are embedded.

Statistical physics of social dynamics.

Statistical physics has proven to be a fruitful framework to describe phenomena outside the realm of traditional physics. Recent years have witnessed an attempt by physicists to study collective phenomena emerging from the interactions of individuals as elementary units in social structures. A wide list of topics are reviewed ranging from opinion and cultural and language dynamics to crowd behavior, hierarchy formation, human dynamics, and social spreading. The connections between these problems and other, more traditional, topics of statistical physics are highlighted. Comparison of model results with empirical data from social systems are also emphasized.

4.Existing system

The way in which information or rumors diffuse in a network has several important differences with respect to infections diseases. Rumors gradually acquire more credibility and appeal asmore and more network neighbors acquire them. After some time, a threshold is crossed and the rumor becomes so widespread that it is considered as ‘common knowledge’ within a community and hence, true. In the case of information propagation in the real world as well as in the blogosphere, the problem is significantly complicated by the fact that the social network structure is unknown. Without explicit linkage data investigators must rely on heuristics.

Disadvantage

Unlike traditional news sources, social media provide little in the way of individual accountability or fact-checking mechanisms, meaning that catchiness and repeatability, rather than truthfulness, can function as the primary drivers of information

diffusion in these information networks. While flame wars and hyperbole are hardly new phenomena online, Twitter's 140-character sound-bytes are ready-made headline fodder for the 24-hour news cycle. More than just the calculated emissions of high-profile users like Sarah Palin and Barack Obama, consider the fact that several major news organizations picked up on the messaging frame of a viral tweet relating to the allocation of stimulus funds, succinctly describing a study of decision making in drug-addicted macaques as for coke monkeys.

5.Proposed System

we intend to add more views to the website, including views on the users, such as the ages of the accounts, and tag clouds to interpret the sentiment analysis scores. We need to collect more labeled data about truthy memes in order to achieve more meaningful classification results, and will also explore the use of additional features in the classifiers, such as account ages for the most active users in a meme, and reputation features for users based on the memes to which they contribute. Another important area to address is that of sampling bias, since the properties of the sample made available in the garden hose are currently unknown. To explore this, we intend to track injected memes of various sizes and with different topological properties of their diffusion graphs.

The fact that many of the memes discussed in this paper are characterized by small diffusion networks, it is important to note that this is the stage at which such attempts at deception must be identified. Once one of these attempts is successful at gaining the attention of the community, it will quickly become indistinguishable from an organic meme. Therefore, the early identification and termination of accounts associated with astroturf memes is critical.

Advantage

social networking sites are sources of a time stamped series of events. Each event involves some number of actors (entities that represent individual users), some number of memes (entities that represent units of information at the desired level of detail), and interactions among those actors and memes. For example, a single Twitter post might constitute an event involving three or more actors: the poster, the user she is retweeting, and the people she is addressing. The post might also involve a set of memes consisting of 'hashtags' and URLs referenced in the tweet. Each event can be thought of as contributing a unit of weight to edges in a network structure, where nodes are associated with either actors or memes. This is not a strictly bipartite network: actors can be linked through replying or mentioning, and memes by concurrent discussion or semantic similarity. The timestamps associated with the events allow us to observe the changing structure of this network over time.

6.Module Description

Feature extraction:

An effective way to detect events is using bursty features in data streams and rich research has been conducted on this area. for example, Kleinberg proposeda formal approach to extract meaningful documents based on modeling the stream using an infinite-state automaton in which bursts appear naturally as state transitions.

Multimedia communication:

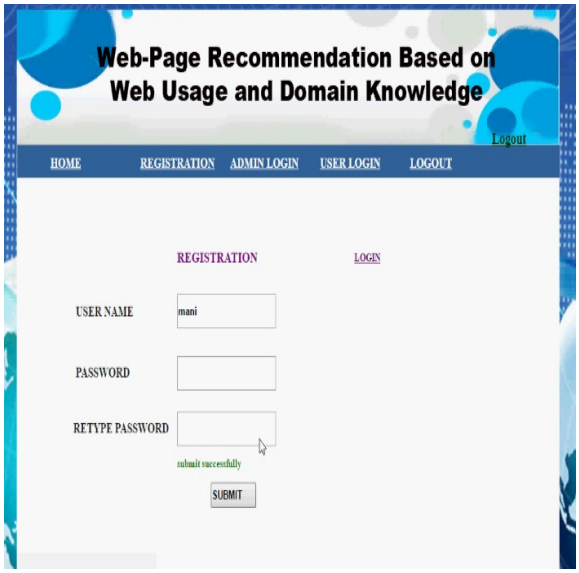

Fig 6.1

The sentiment vector model proposed in our early work in to perform sentiment abstraction. The model contains Chinese words including new Internet words and common emoticons, and is automatically classified into categories. For the messages containing no sentiments, like objective micro blogs, we put them into a candidate For those subjective messages, we perform Principal Component Analysis to detect the main sentiments in time window.

Noise measurement:

A new temporal representation for text streams based on burst features combining with TFIDF was proposed. All those methods, which are very useful on long documents like news or blogs, may encounter disadvantages in micro blog which only contains words. First, it will take a longtime to detect burst features in massive messages. Second, noisy messages contain many burst variation of Chinese words which those approaches may not recognize effectively

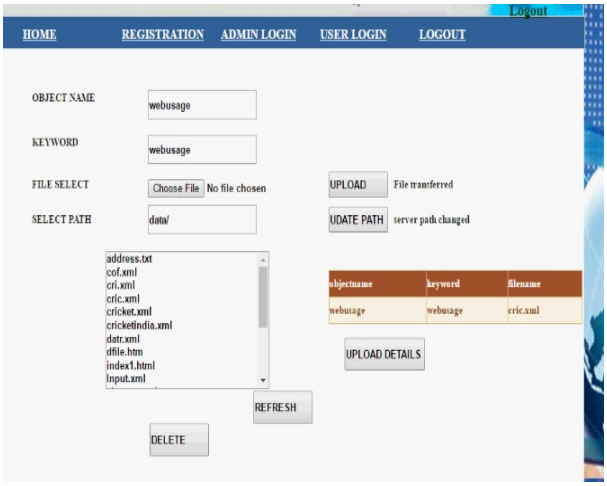

Fig 6.2

Semantics:

A certain sentiments in all subjective messages rather than the absolute number. We can see that the sentiment *happy* almost burst each day, yet *angry* burst only in certain times. In fact, each time China got a medal, there was a growth in sentiment *happy* and we really detected all the medal events.

Graph 6.3

Messages are objective messages, so most messages will be put into candidate set *setC*1 until recycling in the last module. Besides, although there are categories of sentiments in our sentiment vector model, only around types are main sentiment after the process of principal component analysis on average. So it will not take a long time to detect bursts.

7. Conclusion

A new method to offer better Web-page recommendations through semantic enhancement by three new knowledge representation models. Two new models have been proposed for representation of domain knowledge of a website. One is an ontology-based model which can be semi-automatically constructed, namely Domain Onto WP, and the other is a semantic network of Web-pages, which can be automatically constructed, namely Term Net WP. A conceptual prediction model is also proposed to integrate the Web usage and domain knowledge to form a weighted semantic network Of frequently viewed terms, namely Term Nav Net. A number of Web-page recommendation strategies have been proposed to predict next Web-page requests of users through querying the knowledge bases. The experimental results are promising and are

indicative of the usefulness of the proposed models.

8. References

[1] L. Adamic and N. Glance. The political blogosphere and the 2004 U.S. election: Divided they blog. In LinkKDD '05: Proc. of the 3rd International Workshop on Link Discovery, pages 36–43, 2005.

[2] S. Aday, H. Farrel, M. Lynch, J. Sides, J. Kelly, and E. Zuckerman. Blogs and bullets: New media in contentious politics. Technical report, United States Institute of Peace, 2010.

[3] Arbitron/Edison Internet and Multimedia Research Series. Twitter usage in America: 2010. Technical report, Edison Research, 2010.

[4] S. Asur and B. A. Huberman. Predicting the future with social media. Technical Report arXiv:1003.5699, CoRR, 2010.

[5] R. Axelrod. The dissemination of culture a model with local convergence and global polarization the dissemination of culture a model with local convergence and global polarization the dissemination of culture - a model with local convergence and global polarization. J. Conflict Resolution, 41:203, 1997.

[6] A. Barrat, M. Barthelemy, and A. Vespignani. Dynamical Processes on Complex Networks. Cambridge University Press, 2008.

[7] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In Proc. of the 7th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conf. (CEAS), 2010.

[8] Y. Benkler. The Wealth of Networks: How Social Production Transforms Markets and Freedom. Yale University Press, 2006.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. J. Mach. Learn. Res., 3:993–1022, 2003.

[10] J. Bollen, H. Mao, and A. Pepe. Determining the public mood state by analysis of microblogging posts. In Proc. Of the Alife XII Conf. MIT Press, 2010.