S.Geetha, K.Sathiya Kumari / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 5, September- October 2012, pp.1748-1755

Survey on Link Prediction and Page Ranking In Blogs

S.Geetha
M.Phil Scholar
Department Of Computer Science
PSGR Krishnammal College For Women
Coimbatore

K.Sathiya Kumari
Assistant Professor
Department Of Computer Science
PSGR Krishnammal College For Women
Coimbatore

ABSTRACT

This paper presents a study of the various aspects of link prediction and page ranking in blogs. Social networks have taken on a new eminence from the prospect of the analysis of social networks, which is a recent area of research which grew out of the social sciences as well as the exact sciences, especially with the computing capacity for mathematical calculations and even modelling which was previously impossible. An essential element of social media, particularly blogs, is the hyperlink graph that connects various pieces of content. Link prediction has many applications, including recommending new items in online networks (e.g., products in eBay and Amazon, and friends in Face book), monitoring and preventing criminal activities in a criminal network, predicting the next web page users will visit, and complementing missing links in automatic web data crawlers. Page Rank is the technique used by Google to determine importance of page on the web. It considers all incoming links to a page as votes for Page Rank. Our findings provide an overview of social relations and we address the problem of page ranking and link prediction in networked data, which appears in many applications such as network analysis or recommended systems.

Keywords- web log, social networks analysis, readership, link prediction, Page ranking.

I. INTRODUCTION

Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining is also known as Knowledge Discovery in Data (KDD). Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Social networking is the practice of expanding the number of one's business and/or social contacts by making connections through individuals. While social networking gone on almost as long as societies themselves have existed, the unparalleled potential of the Internet to promote such connections is only now being fully recognized and exploited, through web-based groups established for that purpose.

1.1 Social Media

Social media analysis is the practice of gathering data from blogs and social media websites, such as Twitter, Face book, Digg and Delicious, and analyzing that data to inform business decisions. The most common use of social media analytics is gauging customer opinion to support marketing and customer service activities. There are number of types of tools foe various functions in the social media analytics process. These tools include application to identify the best social media sites to serve the purposes, applications to harvest data, a storage product or service, and data analytics software. However, text analysis and sentiment analysis technologies are the foundational components of social media analytics.

1.2 Classification of social media

Social media technologies take on many different forms including magazines, Internet forums, web logs, social blogs, microblogging, wikis, social networks, pod casts, photographs or pictures, video, rating and social book marking. By applying set of theories in the field of media research and social processes Kaplan and Heanlein created a classification scheme for different social media types in their business Horizons article published in 2010.

According to Andreas Kalpan and Michael Heinlein there are six different types of social media: collaborative projects (e.g., Wikipedia), blogs and micro blogs (e.g., Twitter), content communities (e.g., You Tube), social networking sites (e.g., Face book), virtual game words (e.g., World of Watercraft), and virtual social words (e.g., Second Life). Technologies include: blogs, picture sharing, blogs, wall-postings, email, instant messaging, music-sharing, crowd sourcing and voice over IP, to name a few. Many of these social media services can be integrated via social network aggregation platforms. Social media network websites include sites like Face book, Twitter, Bebo and My space.

1.3 Realities of social media mining

The data mining of social media activity is now commonplace in business intelligence circles. Anything and everything on the internet is fair game extreme data mining practices. Once something is pushed out to the World Wide Web, it will forever be fodder for a business intelligence or data mining application somewhere in the cyberspace universe. Content created by a social network's users to outside websites, advertisers, and affiliates for data mining an important component of the social networking business model, as data mining methodologies progress far beyond traditional demographic profiling into interpolation and statistical modeling based on swarms and cluster groups.

So powerful and lucrative is social media data mining that governments around the globe have begun to carefully scrutinize the need for regulation in this apace, especially with respect to protecting the privacy of their citizen's that post data to these networks. While some regulation is probably needed, what concerns user is when the legislators of the free world start demanding social networks involuntarily hand over their user-generated content in order to better enable central governments to carry out their own "citizen intelligence" and data mining programs. That may be closer than any of user care to realize.

1.4 Web Mining

Web Mining aims to discover the informative knowledge from massive data sources available on the Web by using data mining or machine learning approaches. Different from conventional data mining techniques, in which data models are usually in homogeneous and structured forms, Web mining approaches, instead, handle semi-structured or heterogeneous data representations, such as textual, hyperlink structure and usage information, to discover "nuggets" to improve the quality of services offered by various Web applications. Such applications cover a wide range of topics, including retrieving the desirable and related Web contents, mining and analyzing Web communities, user profiling, and customizing Web presentation according to users preference and so on.

Web communities could be modeled as Web page groups, Web user clusters and co-clusters of Web pages and users. Web community construction is realized via various approaches on Web textual, linkage, usage, semantic or ontology-based analysis. Recently the research of Social Network Analysis in the Web has become a newly active topic due to the prevalence of Web 2.0 technologies, which results in an inter-disciplinary research area of Social Networking. Social networking refers to the process of capturing the social and societal characteristics of networked structures or communities over the Web. Social networking research involves in the combination of a variety of research paradigms, such as Web mining, Web communities, social network analysis and behavioral and cognitive modeling and so on.

1.5 Blogs

(Thomas) Ablog-a shorthand term that means "Web log"-is an online, chronological collection of personal communitary and links. Easy to create and use from anywhere with an internet connection, blogs are a form of internet publishing that has become an established communications tool. Blogs represent an alternative to mainstream media publications. The personal perspectives presented on blogs often lead to discourse between bloggers, and many blog circles generate a strong sense of community. Blogs are highly volatile. Bloggers can edit or delete posts, and this transient nature can make blogs difficult to archive or index. Blogs are proliferating rate. Estimates suggest as many as 50 million people are now blogging. Because blogs are easy to create and modify, they occupy a unique niche in cyberspace- that of highly personalized discussion forums that foster communities of interest. Blogs are public and long-lived, and they weave themselves into close relationships with other blogs. As such, they may serve as a tool for reflection, knowledge building, and sharing.

In this diagram explained, 56% percent say blogs are about self-expression. And most post their personal life. And almost 2/3 of blogs read are personal diaries.

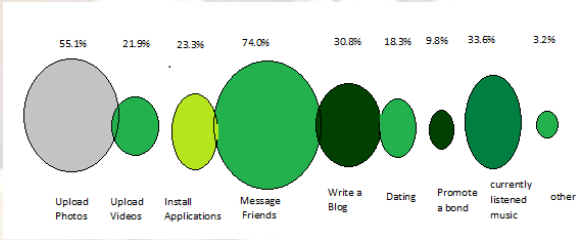

Diagram 1: Blogs vs Social network

Blogs differ from traditional marketing media, however, because they allow you to make high trust impressions with a valuable demographic at a low cost.

Blogs = High Trust, Moderate Reach and High Measurability at a Low Cost

	Trust	Reach	Measurability	Cost
Blog	□ □ □ □	□ □ □	□ □ □	$ $
E-mail newsletter	□ □	□	□ □	$
Press Release	□ □ □ □	□ □	□ □	$ $
Television Ad	□	□ □ □ □ □	□	$ $ $ $ $
Radio Ad	□	□ □ □ □	□	$ $ $ $
Newspaper Ad	□ □	□ □ □ □ □	□	$ $ $ $

Table 1: Comparison of blog advantage

1.6 Link prediction in social network

Link prediction for social network data is a fundamental data-mining task in various application domains, including social network analysis, information retrieval, recommendation systems, record linkage, marketing and bioinformatics. Social networks are interact with people in a group or community and can be visualized as graphs, where a vertex corresponds to a person in some group and an edge represents some form of associations are usually driven by mutual interests that are intrinsic to a group. Link predictions a sub-field of social network analysis. It is concerned with the problem of predicting the (future) existence of links amongst nodes in a social network. The link prediction problem is interesting in that it investigates the relationship between objects, while traditional data mining tasks focus on object themselves.

2. RELEATED WORK

In this section introduced some relative works on social networks. According to that related work in, Rendition behavior in blogosphere (Nardi et al) conducted audio taped ethnographic interviews with 23 bloggers, with analysis of their blog posts. The motivations of blogging are enumerated. Bloggers write blogs to (1) update others on activities and where about, (2) express opinions to influence others, (3) seek others opinions and feedback, (4) think by writing, and (5) release emotional tension. Nardi et al. research leads to speculate that blogging is much about reading as writing, and as much about listening as talking. We specifically examined the production of blogs, but feature research will address blog readers and to assess the relations between blog writers and blog readers precisely.

The spread of media ecstatic over with the blogosphere: In this paper study the trends in the use of blogs as a social medium. Using the HTML links embedded in blog posts from a large data set of blog feeds, we extract the implicit social relationships between blogs and construct the blog graph (Kumar et al. 2003). This allows us to examine dynamic interactions between bloggers. Goal of this paper understand how a specific content propagates in the blog graph and do the spreading characteristics differ when comparing a video of a recent political event, against a music video. This paper focuses the two aspects of blogosphere. (1) Link structure and the content sharing trends across multiple domains and language groups. (2) Posting of you tube links to determine the topics of popular videos and their spreading patterns in the blog graph.

The paper of social media like political blogs can be done by topic modeling for document collections and studies of social media like political blogs. Apply the linkPLSA-LDA model to a blog corpus to analyze its cross citation structure via hyperlinks. The data within blog conversations, focusing on comments left by a comment left by a blog community in response to a blogger's post. In 2008 Nallapati and Cohen introduced the LinkPLSA-LDA model, in which the contents of the citing document and the "influences" on the document, as well as the contents of the cited document, are modeled together.

In 2003 Kumar et al., focused on the evolution of the link structure in blogs over several years and proposed tools and models to study the communities formed by blogs. They called the graph defined by links between the blog graphs. During in 2007 Shi, Tseng, and Adamic 2007 focused in compared the structure of the blog graph, using multiple snapshots in time, against of the web and social networks. Aspect of the spread of media content through the blogosphere focused the link structure and the content sharing trends across multiple domains and language groups and examines the posting of You Tube links to determine the topics of popular videos and their perceptible patterns in the blog graph.

In 2006 Schaller et al., has examined metadata such as gender and age in blogger.com bloggers. They examine bloggers based on their age at the time of experiment, whether in the 10's, 20's or 30's age bracket. They identify interesting changes in content and style features across categories, in which they include blogging words, all defined by the Linguistic inquiry and Word Count (LIWC) by Penne baker et al., 2007. They did not use characteristics of online demeanor (e.g., friends). Their work shows that ease of classification is dependent in part on what division is made between age groups and in turn motivates our decision to study whether the creation of social media technologies can be used to find the dividing line(s).

In predicting the political sentiment of web log posts using supervised machine learning techniques coupled with feature selection, both Hearst and Sack categorized the sentiment of entire documents based on cognitive linguistics models. Other researchers such as Huettner and Subasic, Das and Chen, and Tong manually or semi-manually constructed a discriminate word lexicon help categorize the sentiment of a passage.

Pang, Lee, and Vaithyanathan have successfully applied standard machine learning techniques to a database of movie reviews. They chose to apply Navie Bays, Maximum Entropy and Support Vector Machines to a domain specific corpus of movie formats, the simplest being a unigram representation. Hatzivassiloglou and McKeown, and Turney and Littman chose to classify the orientation of words rather than a total passage. They used the semantic orientation of the containing passage. They pre-selected a set of seed words or applied linguistic heuristics in order to classify the sentiment of a passage.

The collections of data for the sake of performing data/text mining experiments have a long history. In the area of text mining specifically for blog mining task the first attempt to form a collection of blog's dataset was made by, this collection was used for TREC 200 blog-track which was introduced very first time in TREC conference in the same year. Blogs-web log is an online user's diary like application. The political blog-post usually has a clear-cut political bias. The social networks community has also explored issues of interface and diffusion, most frequently using structural and node properties. As this data is mostly static, we are unware of research using timing information for prediction.

Web log data mining referred by MacDougall (2007). That referred to literally millions of blogs on the web, with a new one created every 7-second. Because they store their content electronically in a highly accessible form, blogs are ripe for web mining. Data mining operates on both numerical data and symbolic data, such as text. Text mining provides a means to wade through the massive glut of prose that is generated by our computer-driven society. Chau and Xu (2007) reported an application identifying hate groups and racists through network analysis and visualization. Data sources include HTML documents and link information. Web log mining falls within this category, along with website visitor profile analysis. Yen (2003) presented an approach of pattern analysis of blog linkages.

Predicting response to political; blog posts with topic with topic models (Nallpati and Cohen, 1998) introduced the Link-PLSA-LDA model, in which the contents of the citing document and the "influences" on the documents are modeled together. This study aimed to model the data within blog conversations, focusing on comments left by a blog community in response to a blogger's post. Marlow (2008) using epidemiological terminology, the tools and goals of describing true disease are frequently different than our own. The goal of epidemics research is frequently different to determine how far and how quickly an infection will spread rather than to track the disease through the population.

3. METHODOLOGY

Number of methodology can be used to link prediction, ranking and network analysis in blogs. Blog post hold individual's perception over particular issues. Libe-Nowell and Klienberg propose a link prediction using machine-learning approach. Here the network, the task is to predict whether a link exists or not. If we can predict RR relations hold between two web logs from their social relations using publicly available data.

3.1 Adoption bearing in the blogosphere

Four social networks among web logs (Citation, Blog roll, Comment, and Track back). These relations called social relations because the relations are publicly observable and therefore involve some degree of social consciousness and manifestation. The readership relations and analyze the user data.

3.1.1 Behavioral Relation

First define the bearing relations as relations that are observable only from the user log. Bearing relation includes the readership relations between two weblogs, direct messaging, invitation, and so on. In order to provide the reading behavior, first one is analyze the how often a readership relation exists when other social relations exist. Then consider the probability of reading relations against the number of common neighbors between two bloggers on a social network. Finally view the effect of distance on a social network to the readership relation. For example, if there are 10 bloggers who receive/make comments from blogger A and B, the probability of blogger A reading blog B is about 50%. Blog roll and citation relations include users to read because they create a hyperlink that easily guides a user to the other blog.

3.1.2 Regularly Reading Weblogs

This way used for, how often a blogger reads other blogs when logged in to the system. The interval of reading behavior follows the Poisson distribution. Presume that a user reads another weblog λ times on average in the unit time length. Then the distribution of the interval follows an exponential distribution f (t) $=\lambda e^{-\lambda t.}$ λ is dependent on the weblog; a user might often read interesting weblogs while others do not do so very often.

User 9535 and User 12804	Both have cats as pets. User 9535 often sends a comment to the entries of User 12084. They upload photographs of their cats and comment that "This pose is very pretty", "Cats are fun to watch. They are always mysterious". and so on.
User 10365 and User 5461	They are friend's offline, and often exchange comments. "Followed your advice, and painted my nails pink", "I'm jealous. I want a denim skirt, tool!" and so on.
User 25027 and User 27145	They often write about music. User 27145 put an entry "Please tell me your favorite songs related to the moon or stars". , and User 25027 post an entry in reply "My memorable songs related to the moon or stars" with sending" with sending a track back.

Table 2: Typical examples of RR relations

Machine learning approach is used to define information diffusion on regular reading channels between two blogs. Using these relations for detecting diffusion: readership relations. Two cases include here: (1) How likely is information to diffuse between two blogs with and without RR relations? (2) What kind of information is likely to diffuse through RR relations? In both cases, includes a url in an entry is analyzed. Here, information diffusion on RR relations have following possibilities: Easily identifiable, but information is rarely represented in the form of urls; users do not mention the url in their entries even if the information of the url is propagated: the information might propagate from other over a long distance, or media other than weblogs, which causes coincidental detection of information propagation.

3.2 Web Mining Techniques For Online Social Networks

There is three web-mining techniques are using in online social network analysis: (1) Web content mining, (2) Web usage mining (3) Web structure mining. Web content mining can also be used in on line social networks analysis ro analyze users reading interests, and determine their favorite content. Web usage mining also plays an important role in on line social networks analysis. Wed usage mining is also a tool for measuring centrality degree. The closeness of blog users can be measured by:

Closeness = (f(w*b))+(f*(w*r))+(f*w*I))*

In the equation above, *f* denotes the frequency of a blog behavior, and *w* is the weight of closeness for each blog behavior. The three blog behaviors are *b=browsing, r=reading* and *I=interaction.* This is just a simple example of web usage mining, but the techniques allow many possible means of on line social networks analysis. Web structure mining is the third kind of web mining and it is also useful for extracting and constructing on line social networks to extract the links from WWW,, e-mail or other sources. It can also be used to analyze path length, reachability or to find structural holes, which are very basic and traditional social networks analysis.

3.2.1 Web mining process in online social network analysis

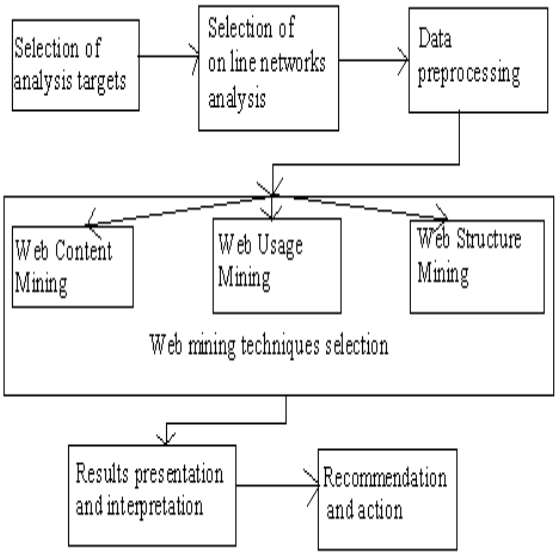

Diagram 2: Process of web mining in online social network analysis

First select the analysis targets, such as web, email, telephone communications, etc. Sometimes, more than one target will be selected. After this step, select what kind on line social networks analysis for proceed with. The analysis targets and on line social networks have been selected, then next one is data preparation. Here data will be collected for analysis, then cleaned and formed as the final format to store in database. Selecting the web mining techniques to be used and then proceeding with them. More than one technique may be selected and sometimes a combination of techniques is necessary. The

selected techniques and then used to analyze the data collected and prepared in the third step of the process. Visualization techniques are sometimes used to assist the presentation of the results of the analysis, such as the extracted social networks. The last one is general process to use web mining for on line social networks analysis is recommendation and action.

3.3 Tacit edifice with Page Rank of Blogspace

In produce the tacit link structure is became obvious that these new graphs differed from the explicit structure of blogspace. These differences may alter the results of ranking algorithms that have largely depended upon the explicit link structure (e.g., www.technorati.com) or have treated blogs in much the same way as regular web pages (e.g., Google's Page Rank). Well read blog may acquire information from a less popular source may miss out on blogs that initially spread infection. To conquer this problem utilizes the timing techniques described above to infer implicit links between blogs and rank the blogs according to those implicit links. For the purpose of ranking blogs do not attempt to infer the most likely infection link but rather all possible infection routes. Page Rank trends to assign high ratings to the sites that contain original materials. iRank tends to assign high ratings to the sites that serve as community portals, such as lists of popular URLs, lists of important blogs, and discussion boards. Another difference is in the way Page rank and iRank treat duplicate blogs.

3.3.1 Classification

Two SVM models using the free LIBSVM toolkit, with a standard radial basis function. The first classifier predicted three classes (reciprocated links, one way links, and unlinked pairs). A second classifier, which used in the final application, distinguish simply between linked (undirected) and unlinked pairs. All classifiers were trained with 10-fold cross validation. Graph inference is achieved through the use of the classifier. For all blogs that are not connected to another blog at an earlier date the classifier proposes links. Graphs are generated using the Graphviz tool, which allows for easy creation of timeline style figures. The coordinates determined by Graphviz are used to render the graph in Zoomgraph, a Java based tool developed to visualize and explore graph structures. Users can use the Zzoomgraph applet to control the threshold for the display of links as well as the types of links that should be displayed.

3.3.2 Clustering graph into communities

Variety of techniques on graph clustering has been used. HITS to discover communities. It is very difficult to set the eigen value threshold to select the communities. If the granularity is course, all communities will merge into one. If the granularity is too fine, the algorithm is very perceptive to noise and outliers. Graph clustering algorithm based on the "edge betweens" of an edge in a graph. The "betweens" of an edge is defined as the number of shortest paths between all pairs of vertices that run along the edge. By removing the edge with the highest "betweens", the graph will be partitioned. Generating a named entity graph, need a collection of documents regarding a certain named entity and then extract all the sentences containing more than one entity's name, which is a person name in this case. To behold named entity tenures from the documents, here using MINIPAR as the named entity parser. All the sentences parsed by MINIPAR, only those sentences containing more than one name are collected. Then map the selected sentences into a weighted undirected graph.

Google Wed search and Google Blog search used by "Tom Cruise" as the blog query because of this is popularity among bloggers. The blog documents were collected manually because most of blog sites have finally provide API for users to access their collections. Set of documents retrieved over with Google Web search engine. This selection is based on except Web documents have a broad coverage on a given person entity and assume that the retrieved top-ranking documents should be of high quality. Because there are millions of pages relevant to our query entities.

3.3.3 Link Induction and visualization

The first classifier (svm) predicted three different classes (restitute links, one way links, and un between linked (undirected) and unlinked pairs. linked pairs). A second classifier (both implemented in SVM and as a logistic regression) imposing simply between linked (undirected) and unlinked pairs. These classifiers using a set of heuristics now able to construct an infection tree, with each node representing a blog. Links in the tree between the nodes show how infection may have spread for a specific URL. The system automatically produces figures representing the specific link structure. The current system only holds edges deduce through the two-class link deduce SVM.

3.3.4 Content sharing

Here examine three aspects of content sharing patterns in the blogosphere. (1) What are the topics and categories of videos that are popular, (2) what is the age of the shared videos, (3) how quickly do links to the same video spread in the blogosphere? The number of HTML links to You Tube videos in the blogosphere follows Zipf's law. This hints us at the existence of a large-scale diffusion of You Tube videos. The median values of the half times and full times of videos for the 4 video categories. Recall that the Spinner3r imprint spans only two months, Spinner3r's web crawler

can discover posts that much older than two months for blogs that published posts rarely.

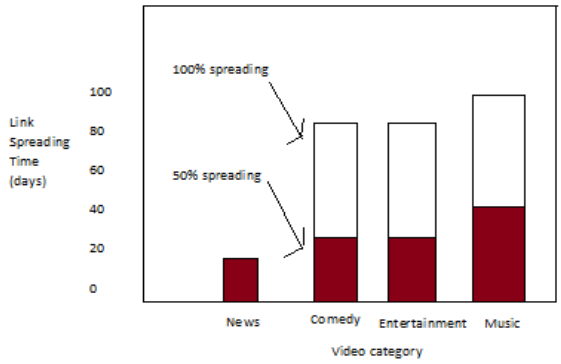

Figure 1: Time lag in the spread of videos in the blogosphere

3.4 Link structure of the web pages

Spreading patterns can greatly us about general information flow in networks. However, to understand how a specific URL has spread requires a closer look at the network structure. Blog creators frequently provide a list of other blogs frequently read by the author or automated track backs. URLs that are reputably cited within the community of bloggers are tracked by several websites. The popularity measurements of these URLs act as a simple filtering and ranking mechanism, allowing users to quickly find potentially interesting URLs that many bloggers are talking about in near real-time.

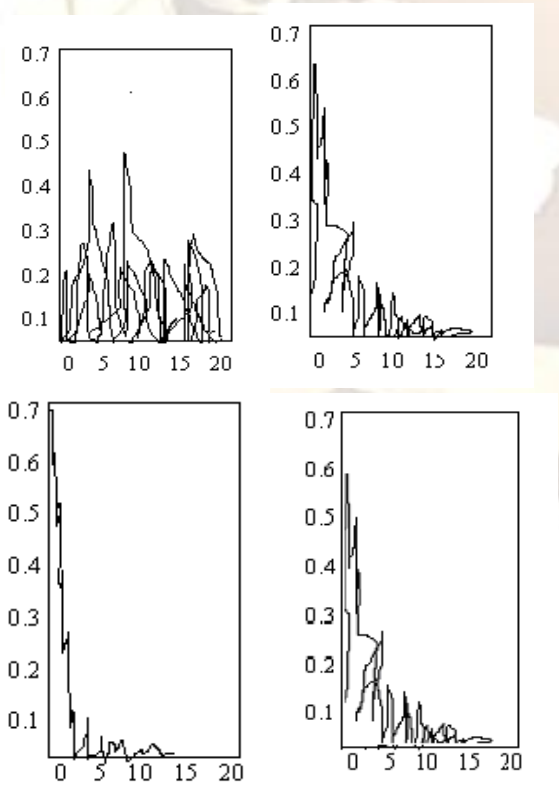

Figure 2: The four epidemic expression profiles resulting from the k-means clustering procedure: (a) 124 of 259 URL time course profiles are of the sustained interest type, (b) 51 URLs peak on day two followed by a slow decay, (c) 38 URLs peak on day one with a very fast decay, and (c) 46 URLs have a peak on day one with a slower decay.

For the analysis 259 URLs selected from all the URLs mentioned by blogs between the period of May 2, 2003 and May 21, 2003 according to the following criteria: (i) the URL was cited at least 40 times, (ii) the URL was not on a stop list os 198 URLs (e.g. the homepages of common news sources such as moveabletype.org), and (iii) the URL was not mentioned by any blogs on May 2, 2003. This tended to eliminate profiles that might have their peak before May 2, but do not sustain interest past that point.

4 CONCLUSION

This study has examined the interaction between social relations and behavioral relations. Several issues require further analysis, but we believe that we have shown a comprehensive overview of social relations that can be associated with readership relations and addressed the problem of link prediction in networked data. The novel model can collectively capture globally predictive intrinsic properties of objects and discover the latent block structure, which shows the success of the coupled benefits feature based approaches and latent block models.

The propagation information through blogspace, both to uncover general trends and to explain specific instances of URL transmission. The success of the simple inference, we would like to augment the simple weighting scheme for the iRank algorithm to include the SVM to calculate from another. Both the availability and quantity of time resolved information is unique to blog data. Using it we were able to not only infer link structure, but also to create a novel-ranking algorithm. iRank for ranking blogs.

Several probabilistic topic models are applied to discourse within political blogs. A novel comment prediction task to assess these models in an objective evaluation with possible practical applications. The results show that predicting political discourse behavior is challenging, in part because of considerable variation in user behavior across different blog sites.

The problem of mining communities from web pages and blogs by exploiting the named entity co-occurrence. Mapped web documents into a named entity graph. An effective hierarchical clustering algorithm, which utilities both the triangle geometry inside a graph and the mutual information

between vertices. The community quality is evaluated by the summery terms. Based on the evaluation, we believe that the technique can enhance our ability to acquire knowledge from web pages and blogs.

Web log mining usually involves mining of text. Text mining is much more challenging than quantitative data mining. Numbers are much easier to process by computer than text. However, many tools have been developed to find patterns in text that lead to human knowledge. This can include learning about the intentions of competitors, as well as study of general customer behavior. Web mining is a natural extension of the use of technology. The vast majority of website content is predominately not of interest to any one individual. Web mining provides more efficient ability to identify relevant content. Along with this technological ability, comes the risk of abuse.

Page Rank algorithm remains complex and quite badly known, maybe because its authors keep it secret for industrial security obvious reasons. Thus, even though our explanations can not be 100% reliable, they reflect the experiences of hundreds of users. Another important point to notice in this conclusion is that Page Rank is only one of the criteria involved in the Google search algorithm: having a good Page Rank isn't up to have very good rankings! Our recommendation is to pass time creating rich content for your visitors, because it's the actual value-added of your site. Read our advices for optimizing your links (anchor texts are very important), choosing your titles and your meta tags. Related inference techniques to produce novel-ranking algorithms for blog search tools and the study of meme “mutations”.

REFERENCES

[1] L. Adamic and N. Glance. The political blogosphere and the 2004 U.S. election: Divided they blog. In linkKDD-2005, 2005.

[2] Adamic, L.A., O. Buyukkokten and E. Adar, A social network caught in the web, First Monday, 8(6).

[3] M. Airoldi, D. Blei, S. Fienberg, and E.Xing, Mixed membership stochastic block models. JMLR, pages 1981-2014, 2008.

[4] Dearstyne, B.W. (2005) ‘Blogs: the new information revolution?’ The Information Management Journal, Vol. 39, No.5, pp.38-44.

[5] Domingos, P., and M. Richardson, Mining the Network Value of Customers, KDD’01: Knowledge and Data Discovery, San Francisco, CA, 2001.

[6] Dredze, M. Crammer, K, and Pereira, F. (2008) Confidence weighted linear classification. In proceedings of the 25th international conference on machine learning, July 5-9, Helsinki, Finland.

[7] Gordon, A., Cao, Q., & Swanson, R. (2007) Automated Story Capture From Internet Weblogs. Proceedings of the Fourth International Conference on Knowledge Capture, October 28-31, 2007, Whistler, Bc.

[8] Kleinberg, J. Authoritative sources in a hyper linked environment, Journal of ACM, 46(5). 604 632, 1999.

[9] P. Kolari, A. Java, and T. Finin. Characterizing the splogosphere. In Proc 3rd Annual Workshop on Weblogging Ecosystem, 2006.

[10] Leskovec, J.; Krause, A.; Guestrin, C., Faloutsos, C.; VanBriesen, J., and Glance, N. 2007. Cost-effective Outbreak Detection in Networks. In ACMSIGKDD.

[11] Y. Lin, H.Sundaram, Y. Chi, J. Tatemura, and B. Tseng Discovery of blog communities based on mutual awareness. In WWW2006 Workshop on Weblogging Ecosystem, 2006.

[12] A. Menon and C. Elkan. Link prediction via matrix factorization ECML-PKDD, 2011.

[13] NILTE Blog Census, http://www.blogcensus.net/

[14] Noble, C. and Cook, D. 2003. Graph-based anomaly detection. In KDD-03, 631-636. Washington, Dc, USA.

[15] Owslev, S., Hammond, K., Shamma, D., Sood S.(2006) Buzz: Telling Compelling Stories. ACM Multimedia, Interactive Arts program, Santa Barbara, CA.

[16] Pastor-Satorras, R. and A. Vesignani, Epidemic spreading in scale-free networks, Physical Review Letters, 86(2001), pp. 3200-3203.

[17] Robertson, S. Spark-Jones. K, Relevance Weighting of Search Terms, Journal of American Society for Information Science, 27, 1988.

[18] Shi X.; Tseng. B.; and Adamic, L. A. 2007. Looking at the Blogosphere Topology through Different Lenses. In ICWSM.

[19] I. Titory and R. McDonald, 2008. A joint model of text and aspect ratings for sentiment summarization. In proceedings of ACL-08: HLT.

[20] Zhou, D., Ji, X., Zha H., Giles, L., Topic Evolution and Social Interactions: How Authors Effect Research, in the proceedings of the 15th ACM CKIM, 2006.