International Journal of Enhanced Research in Management & Computer Applications, ISSN: 2319-7471
Vol. 3 Issue 3, March-2014, pp: (48-52), Impact Factor: 1.147, Available online at: www.erpublications.com

An Analytical research on social media system by web mining technique: blogs & blogosphere Case

Dr. Manoj Kumar Singh[1], Dileep Kumar G[2,] Mohammad Kemal Ahmad[3]
[1,2]Asst Professor, Department of Computing, Adama Science & Technology University, Adama, Ethiopia
[3]Head of Department, Department of Computing, Adama Science & Technology University, Adama, Ethiopia

Abstract: Now a days blogosphere are very popular platform for users to post and share articles with each other. Blogs have become increasingly popular and have been widely used for such purposes as online diaries, commentaries, and socialization and such social media system has become a very popular application of Web 2.0 ages. In this paper, we work on building systems that analyse these emerging social media systems to recognize spam blogs, find opinions on topics, identify communities of interest, derive trust relationships, and detect influential bloggers.

Keywords: blogs, blogosphere, social media, Web mining.

Introduction

In recent years, social media is a very popular application in the age of web 2.0, which allows user to communicate, interact and share in the World Wide Web. Such Web 2.0 systems have a significant amount of user-generated content, and have become an important new way to publish information, engage in discussions, and form communities on the Internet. Their reach and impact is significant with tens of millions of people providing content on a regular basis around the world. Governments, corporations, traditional media companies, and NGOs are working to understand how to adapt them and use them effectively. Citizens, both young and old, are also discovering how social-media technology can improve their lives and give them more voice in the world. We use the term “social media” for new genre is the weblog or blog. Following the current fashion, we will refer to the set of weblogs as blogosphere. This is more than just a name coined to reference a set of Web pages—the rich interlinking of blogs does make it seem like a distinct network of resources embedded into the larger Web. While every social-media genre has its unique and interesting features, the blogosphere does embody many of the key features that typify social media in general. In this paper, we focus our discussions on blogs and the blogosphere, to a wider range of social-media systems. We begin by describing an overarching task of discovering which blogs and bloggers are most influential within a community or about a topic. Pursuing this task uncovers a number of problems that must be addressed, several of which we describe here. These include recognizing spam in the form of blogs and comments, and developing more effective techniques to recognize the social structure of blog communities.

Blogs have received a high profile in the last couple of years. They are well known with anybody who is comfortable with the web. Also beyond the web, they are even referenced and issued in traditional media as well. Blogs have grown to a permanent phenomenon that has become a significant component of the global information and communication infrastructure. With the massive growth of the global blogging infrastructure several meta-services emerged, which offer services such as searching for blog postings. Well, searching actually is one of the more simple services. The more exciting services are described in detail below. Apart from this, blogging also penetrates corporate information and communication infrastructures.

As a social-media genre, blogs have many ancestors, most of which still exist today. These include message forums, topical mailing lists, newsgroups, and bulletin board systems. As the Internet and Web became widely available and used in the mid-1990s, a number of people began to maintain sites as personal diaries in which they made chronologically organized records of interesting Web resources discovered, thoughts on current events, and activities. In 1997, one of them, Jorn Barger, coined the term “weblog” to describe his diary site, which was subsequently shortened to blog. During the late 1990s, the appearance of hosted and open-source blogging platforms like Open Diary, LiveJournal, and Blogger made it easy for people to adopt the form. At the same time, the now common blog structure, protocols, and supporting infrastructure comments, feeds, permalinks, blogrolls, trackbacks, pings, ping servers, etc.became standardized. Since 2000, the blogosphere has seen a continuous, exponential grown, doubling every 6 months for most of this period and only within the last year beginning to tapper off. Technorati claimed [1] in the Spring of 2007 to be following over 70Mblogs that are generating about 1.5M posts per day. This report also shows evidence that blogging has become a global activity by analyzing the most common languages used in blogs: Japanese 38%; English 35%; Chinese 9%; Spanish 4%; Russian, French, and Portuguese 2%; German and Farsi 1%; and all others 3%.

One reason that the blogosphere continues to grow is that popular blogging platforms like Blogspot, WordPress, and TypePad can serve as simple content management systems and for many provide a very easy way to add content to the Web. Moreover, the ping-based infrastructure exploited by blog search engines means that it provides the fastest way to get that content indexed by major search engines.* The blogosphere is part of the Web and therefore shares most of its general characteristics. It differs, however, in ways that impact how we can model it and use the model to help extract information. The common model for theWeb in general is as a graph of Web pages with undifferentiated links between pages. The blogosphere has a much richer network structure in that there are more types of nodes that have more types of relations between them. For example, the people who contribute to blogs and author blog posts form a social network with their peers, which can be induced by the links between blogs. The blogs themselves form a graph, with direct links to other blogs through blogrolls and indirect links through their posts. Blog posts are linked to their host blogs and typically to other blog posts and Web resources as part of their content.

A typical blog post has a set of comments that link back to people and blogs associated with them. The blogosphere trackback protocol alerts a blog site that it is being linked to from another blog and enables it to create a back link. Still more detail can be added by taking into account post tags and categories, syndication feeds, and semi structured metadata in the form of XML and RDF content. Finally, a link’s anchor text as well as the text around the link provides significant information. Believe that adapting and extending the work done by many sub communities in the data-mining arena can help develop new techniques to analyze social media. The blogosphere and social media in general, continue to evolve.. Finally, users are increasingly interested in being able to extract and use the data that social-media sites have about them.

Blogosphere

Viewing merely the technical intricacies of blogs is insufficient to catch the overall phenomenon. Technology and culture always develop in tandem, and this report hence takes a broad perspective that includes both. This chapter aims to draw a broad picture of the global blogging environment, the public blogosphere. It discusses what blogs are, makes a historic synopsis and illustrates the blogosphere’s path from an isolated phenomenon to a massive and vibrant ecosystem. Estimations about the size of the blogosphere vary greatly. By August 2005, large blog search engines including BlogPulse and Technorati were watching more than 15 million blogs each. While estimations vary greatly, some put the global blogosphere to a stunning size of more than 70 million blogs[9]. At the same time, the blogosphere is extremely dynamic with many tens of thousands of blogs created each day. While a great amount of discontinued blogs and novel occurrences like ‘spam blogs’ somewhat distort numbers, Technorati estimates the blogosphere to double in size about once every five months. With regard to the connectedness of this huge and quickly growing space, there is some disagreement in the community. Despite the common assumption and some scientific indications that the blogosphere is densely connected, other initiatives contradict this notion. More research based on common concepts and accepted reference points is hence needed to establish a clear characterisation of the blogosphere.

Modeling Influence in the Blogosphere

The blogosphere provides an interesting opportunity to study social interactions including spread of information, opinion formation, and influence. Through original content and commentary on topics of current interest, bloggers influence each other and their audience. We aim to study and characterize this social interaction by modeling the blogosphere and providing novel algorithms for analyzing social-media content. Content and commentary on topics of current interest, bloggers influence each other and their audience. We aim to study and characterize this social interaction by modeling the blogosphere and providing novel algorithms for analyzing social-media content. Figure shows a hypothetical blog graph and its corresponding flow of information in the influence graph. Studies on influence in social networks and collaboration graphs have typically focused on the task of identifying key individuals (influential) who play an important role in propagating information. This is similar to finding authoritative pages on the Web. Epidemic-based models like linear threshold and cascade models [2–4] have been used to find a small set of individuals who are most influential in social network. However, influence on the Web is often a function of topic. A post comparing the Apple iPhone with the Nokia N95 will likely be influential. A blog that is relatively low ranked on conventional measures can be highly influential in this small community of interest. In addition, influence can be subjective and based on the interest of the users. Thus, by analyzing the readership of a blog, we can gain some insights into the community likely to be influenced by it. We have implemented a system called Feeds That Matter [5] that aggregates subscription information across thousands of users to automatically categorize blogs into different topics.

- These were the top 10 blogs in 2013 based on readership metrics from the feeds that matter system.

(1) http://www.huffingtonpost.com/
(2) http://www.tmz.com/
(3) http://www.businessinsider.com/
(4) http://www.huffingtonpost.com/
(5) http://www.tmz.com/
(6) http://www.businessinsider.com/
(7) http://gawker.com/
(8) http://lifehacker.com/
(9) http://mashable.com/
(10) http://gizmodo.com/

.

The BlogVox system [6] retrieves opinionated blog posts specified by ad hoc queries. After retrieving posts relevant to a topic query, the system processes them to produce a set of independent features estimating the likelihood that a post expresses an opinion about the topic. These are combined using an support vector machine (SVM)-based system and integrated with the relevancy score to rank the results. Since blog posts are often informally written, poorly structured, rife with spelling and grammatical errors, and feature nontraditional content, they are difficult to process with standard language analysis tools. Performing linguistic analysis on blogs is plagued by two additional problems:

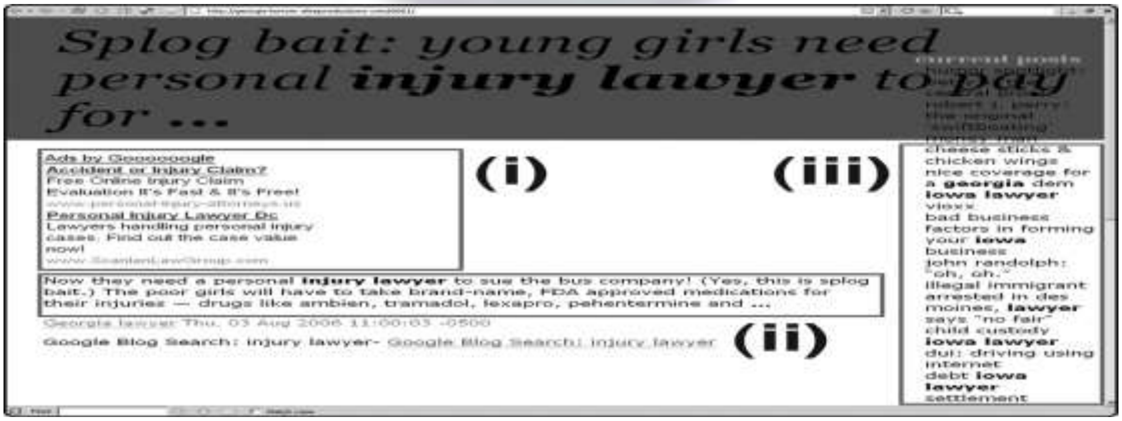
(1) The presence of spam blogs and spam comments
(2) Extraneous noncontent including blogrolls, linkrolls, advertisements, and sidebars.

Detecting Blog Spam

As with other forms of communication, spam has become a serious problem in blogs and social media, effecting both users and systems that harvest, index, and analyze blog content. Two forms of spam are common in blogs: spam blogs, or splogs, where the entire blog and hosted posts consist of machine-generated spam, and spam comments attached by programs to authentic posts on normal blogs Figure. Though splogs continue to be a problem for Web search engines and are considered a special case of Web spam, they present a new set of challenges for blog analytics. We limit our discussion to that of splogs. Blog search engines index new blog posts by processing pings from update ping servers, intermediary systems that aggregate notifications from updated blogs. Pings from spam pages increase computational requirements, corrupt results, and eventually reduce user satisfaction.

Detecting Splogs

Over the past two years, we have developed techniques to detect spam blogs. We discuss highlights of our effort based on splog detection using blog home pages with local and relational features. Results reported are based on a seed dataset of 700 positive (splogs) and 700 negative (authentic blog) labeled examples containing the entire HTML content of each blog home page. All of the models are based on SVMs[8] .We used a linear kernel with top features chosen using mutual information and models evaluated using one fold cross-validation. We view detection techniques as local and relational, based on feature types used.



Splog bait: young girls need personal **injury lawyer** to pay for ...

current posts
humor spotlight:
...
robert l. perry:
the original
'swiftboating'
money from
cheese sticks &
chicken wings
nice coverage for
a **georgia** dem
iowa lawyer
views
bad business
factors in forming
your **iowa**
business
john randolph:
"oh, oh."
illegal immigrant
arrested in des
moines, **lawyer**
says "no fair"
child custody
iowa lawyer
dui: driving using
internet
debt **iowa**
lawyer
settlement

Ads by Goooooogle
Accident or Injury Claim?
Free Online Injury Claim
Evaluation It's Fast & It's Free!
www.personal-injury-attorneys.us
Personal Injury Lawyer Dc
Lawyers handling personal injury
cases. Find out the case value
now!
www.SeanlanLawGroup.com

Now they need a personal **injury lawyer** to sue the bus company! (Yes, this is splog bait.) The poor girls will have to take brand-name, FDA approved medications for their injuries — drugs like ambien, tramadol, lexapro, pehentermine and ...

Georgia lawyer Thu, 03 Aug 2006 11:00:03 -0500

Google Blog Search: injury lawyer- Google Blog Search: injury lawyer

Figure: The performance of local models, as measured by the standard, area under the curve metric, varies for different feature types and sizes

Local Features

A blog's local features can be quite effective for splog detection. A local feature is one that is completely determined by the contents of a single Web page. A local model built using only these features can provide a quick assessment of the degree to which a blog appears to be spam. These features include bag-of-words (BOW), word n-grams, anchor text, and URLs.

(1). **Words.** To verify their utility, we created bag-of-words for the samples based on their textual content. We also analyzed discriminating features by ordering features based on weights assigned to them by the linear kernel. It turns out that the model was built around features that the human eye would have typically overlooked. Blogs often contain content that expresses personal opinions, so words like I, we, my, and what appear commonly on authentic blog posts. To this effect, the bag-of-words model is built on an interesting blog-content genre. In general, such a content genre is not seen on the Web, which partly explains why spam detection using local textual content is less effective there.

(2). **Word n-grams.** An alternative methodology to using textual content for classification is the bag-of-word-n-grams, where n adjacent words are used as a feature. We evaluated both bag-of-word-2-grams and bag-of-word-3-grams, which turned out to be almost as effective as bag-of-words. Interesting discriminative features were observed in this experiment. For instance, text like comments-off (comments are usually turned-off in splogs), new-york (a high paying advertising term), and in-uncategorized (spammers do not bother to specify categories for blog posts) are features common to splogs, whereas text like 2-comments, 1-comment, i-have, and to-my were some features common to authentic blogs. Similar features ranked highly in the 3-word gram model.

(3). **Tokenized anchors.** Anchor text is the text that appears in an HTML link (i.e., between the <a...> and </a> tags) and is a common link-spamming technique around profitable contexts. We used a bag-of-anchors feature, where anchor text on a page, with multiple word anchors split into individual words, is used. Note that anchor text is frequently used for Web page classification, but typically to classifying the target page rather than the one hosting the link. We observed that comment and flickr were among the highly ranked features for authentic blogs.

(4). **Tokenized URLs.** Intuitively, both local and outgoing URLs can be used as effective attributes for splog detection. This is motivated by the fact that many URL tokens in splogs are in profitable contexts. We term these features as bag-of-urls arrived at by tokenizing URLs using "/" and ".". Results indicate this can be a useful approach complementing other techniques.

Relational Features

A global model is one that uses some nonlocal-features, that is, features requiring data beyond the content of Web page under test. We have investigated the use of link distributions to see if splogs can be identified once they place themselves on the blog (Web) hyperlink graph. The intuition is that authentic blogs are very unlikely to link to splogs and that splogs frequently do link to other splogs. We have evaluated this approach by extending our seed dataset with labeled in-links and out-links, to achieve AUC values of close to 0.849. Though current techniques work well, the problem of spam detection is an adversarial challenge. In our continuing efforts, we are working toward better addressing concept drift and leveraging community and relational features. The problem of spam in social media is now extending well beyond the blogs and is quite common in popular social tools like MySpace and Facebook. The nature of these social tools demands additional emphasis on relational techniques, a direction we are exploring as well.

Conclusions

This research portrayed the social space blogosphere and, from a technical perspective, it outlined the diverse blog activities in the international community. Social-media systems are increasingly important on the Web today and account for the majority of new content. The various kinds of social media are alike in that they all have rich underlying network structures that provide metadata and context that can help when extracting information from their content. We have described some initial results that is focused on extracting and exploiting this structural information. We note that there is a lack of adequate datasets to fully test the new approaches. To some extent, synthetic blog graphs can be useful to this end. In recent work [7], we have shown that existing graph-generation techniques do not work well for this, and proposed a new model. As the Web continues to evolve, we expect that the ways people interact with it, as content consumers as well as content providers, will also change. The result, however, will continue to represent an interesting and extravagant mixture of underlying networks—networks of individuals, groups, documents, opinions, beliefs, advertisements, and scams. These interwoven networks present new opportunities and challenges for extracting information and knowledge from them.

References

[1]. D. Sifry. Available at http://www.sifry.com/alerts/archives/000493.html, April 2007.

[2]. D. Kempe, J.M. Kleinberg, and ´ E. Tardos.Maximizing the spread of influence through a social network. In Proceeding of the 9th ACMSIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 137–146, New York, 2003. ACM Press.

[3]. D. Kempe, J. M. Kleinberg, and ´ E. Tardos. Influential nodes in a diffusion model for social networks. In Proceedings of the Internatonal Colloquium on Automata, Languages and Programming, pp. 1127–1138, Berlin/Heidelberg, Germany, 2005. Springer.

[4]. J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. In SIAM International Conference on Data Mining (SDM 2007), Philadelphia, PA, 2007. SIAM.

[5]. A. Java, P. Kolari, T. Finin, A. Joshi, and T. Oates. Feeds That Matter: A study of bloglines subscriptions. In Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007), Menlo Park, CA, March 2007.

[6]. A. Java, P. Kolari, T. Finin, A. Joshi, J. Martineau, and J. Mayfield. The BlogVox opinion retrieval system. In Proceedings of the 15th Text Retrieval Conference (TREC 2006), Gaithersburg, MD, February 2007.

[7]. A. Kale, A. Karandikar, P. Kolari, A. Java, A. Joshi, and T. Finin. Modeling trust and influence in the blogosphere using link polarity. In Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007),Menlo Park, CA, March 2007. AAAI Press.

[8]. P. Kolari, A. Java, T. Finin, T. Oates, and A. Joshi. Detecting spam blogs: A machine learning approach. In Proceedings of the 21st National Conference on Artificial Intelligence (AAAI 2006), Menlo Park, CA, 2006. AAAI Press.

[9]. P. Kolari, T. Finin, and A. Joshi. SVMs for the blogosphere: Blog identification and splog detection. In AAAI Spring Symposiumon Computational Approaches to Analyzing Weblogs, Menlo Park, CA, 2006. AAAI Press.