Current Approaches to Data Mining Blogs

Introduction

Clai Rice (Information Architecture Summit 2007) provides a useful summary of the current direction of blog research using data mining techniques. He notes that one set of researchers has utilized blog mining to explore the networking element of blogs, such as the use of hyperlinks between blogs and to outside sources of information. Networks of links are thereby analyzed to facilitate an understanding of social networks, blogging communities, and their contextual foundations. Another area of research deconstructs the literary aspects of bloggers' writing style and attempts to identify personal characteristics like gender, nationality, geographical location or political affiliation based on this style. Thirdly, blogs are a prime source for opinion mining and sentiment extraction due to the personal and candid nature of posts and comments. Opinion mining particularly appeals to commercial interests because of the rich potential to better understand target demographics (cf. RelevantNoise.com). These three fields accurately represent the majority of readily available and ongoing research projects (described below). It appears that a majority of the attempts at utilizing data mining for any of these means maintains a very commercial and/or computational approach due to the nature of their development by computer scientists, programmers and corporate representatives. Still, the interests, methodologies and applications of current research into blog data mining overlap considerably with those of social researchers such as sociologists and anthropologists and could effectively be pursued to these ends.

A summary of recent research dealing with data mining blogs for various purposes is outlined below. The papers have been organized into four categories based on the approach taken to data mining, the purpose of the research, and the type of analysis provided. The four categories therefore include articles (1) relating to tagging, classification and folksonomy in the blogosphere; (2) mining comments and links to determine blogging community networks; (3) focusing on spatio-temporal data; (4) extracting information regarding bloggers' identity, behavior and/or mood. A brief conclusion along with the possibilities of future research is also presented.

Tags, Classification and Folksonomy

Improved Annotation of the Blogosphere via Autotagging and Hierarchical Clustering (from WWE 2006[1])
Christopher H. Brooks and Nancy Montanez

This research analyzed the use of tagging in blog posts by mining Technorati, a site which indexes and collects information from the blogosphere. It examined whether or not blog tagging is an effective means of labeling the content of blogs or, likewise, of conveying similarity of content between equally tagged posts. Their findings indicate that bloggers utilize tags for three main purposes: personal notes like “to do” or “to read”; simple categorization or organization; and to annotate and describe the content of articles. Further experimentation into the method of categorization mined from some hundreds of blogs shows that automated tagging produces more focused, topical clusters, whereas human-assigned tags produce broad categories. One possible further

application of this research for social scientists interested in understanding bloggers relates to the development of a working "folksonomy" (community-defined tagging practices). As such, this research may provide a beginning for using data mining techniques (including popular tools like Technorati) to understand how people organize their blog entries and what different methods of tagging can tell us about personal organization and classification. The researchers' future interests in examining tagging as an evolving social phenomenon may also prove promising.

Browsing System for Weblog Articles based on Automated Folksonomy (from WWE 2006)
Tsutomu Ohkura, Yoji Kiyota, and Hiroshi Nakagawa

Also on the topic of understanding the "folksonomy" of the blogosphere, the authors attempt to design a program which crawls the web for new blog posts and then automatically tags them. The "tagger" is meant to incorporate the classification scheme of ordinary users and determine when it should be applied to a particular item. The benefit of this program is to provide a more universally accurate system which facilitates searching and organization. Tagging folksonomy, as a classification scheme, is both intriguing as a social construct and can be easily extracted as raw data. Replacing human tagging with an automated system is an interesting exercise in artificial intelligence but, from the perspective of social research, this method may prove lacking in terms of understanding the context and subjectivity of human classification.

BlogHarvest: Blog Mining and Search Framework (from International Conference on Management of Data 2006)
Mukul Josh and Nikhil Belsare

The researchers developed a blog mining program called BlogHarvest which searches for, and extracts, a blogger's interests in order to recommend blogs with similar topics. The program uses classification, links, topic similarity clustering and tagging based on opinion mining to provide these features. The program design is based on the knowledge that blogging communities are not formed randomly, but as a result of shared interests. It is also designed to provide a useful search facility to bloggers while generating large amounts of revenue for advertising services and providers. A program like this (and other services like Blogpulse, Technorati and Google Blog Search, etc) can be utilized as ready-made data mining aids for social research.

Mining Comments, Links and Tracing Networks

Experiments on Persian Weblogs (from WWE 2006)
Kyumars Sheykh Esmaili, Mohsen Jamali, Mahmood Neshati, Hassan Abolhassani and Yasaman Soltan-Zadeh

This research used crawlers to locate and mine Persian blogs. The resulting information was utilized to rank the popularity and significance of the blogs based on the numbers of in- and out-links from each page. The primary goal was to enable researchers interested in understanding bloggers' social networks to apply social network analysis to the mined data, such as through hyperlink and recommendation

analysis. The potential of this research appears to be its attention to the easily overlooked aspect of inter-linking and trackbacks between blogs and bloggers, which, along with comments, are the main ways in which blogging communities tend to be initiated and sustained. As such, in order to understand the meaning and context of blog topics and posts, mining techniques should be able to take into account themes spread across blogging communities as well as individual textual content. Ranking systems using backlinks within a finite subject or location range may thereby shed light on the maintenance of blogging communities.

Leave a Reply: An Analysis of Weblog Comments (from WWE 2006)
Gilad Mishne and Natalie Glance

The authors argue that the area of weblogs devoted to comments from other users tend to be ignored in typical data mining exercises because they are more difficult to extract and process than main blog posts. Hence, they seek to provide the first large-scale study of blog comments, which, significantly, represent up to 30% of the blogosphere. They note that some weblog domains (such as LiveJournal) have an extremely high proportion of internal comments that must be examined in order to understand social networks and identify blogging communities. On the one hand, attention to comments results in weaker topical precision from mined content; on the other, they provide access to the perspectives of blog readers. Lastly, by mining both the blog post and its comments, the researchers argue that they can determine the degree and level of 'controversy' or conflict prompted by particularly themes or blog authors. The implications of this paper for social researchers include the attention to the dynamics of inter-blogger communication, the impact of the topical 'subject matter' of posts on blogger community interaction, and, finally, the suggestion that data mining techniques can be designed to extract subjective qualitative data (or at least to analyze it).

Discovery of Blog Communities based on Mutual Awareness (from WWE 2006)
Yu-Ru Lin, Hari Sundaram, Yun Chi, Jun Tatemura and Belle Tseng

The researchers used data mining to divide blogs into communities by topic and to examine the means by which they link to each other. They propose a process which analyzes 'mutual awareness' between bloggers based on link semantics and/or keyword extraction. This method of mining content and blogger interactions therefore enables the discovery of blogging communities and an analysis of their evolution and sustainability; the possibility of topical extraction for further content analysis; and the identification of specific "roles" applied to individual bloggers within an interactive blogging community. The focus here is using data mining techniques to trace links 'between' bloggers as well as to better understand individual blogging patterns.

Enhancing Clustering Blog Documents by Utilizing Author/Reader Comments (from ACM Southeast Regional Conference 2007)
Beibei Li, Shuting Xu, and Jun Zhang

The authors also advocate mining blogs for the personal opinions and feelings of their authors and other users who leave comments. They argue that blog comments left by both the author and other readers can contain data which is more effective for classifying blog documents than the body of blog entries themselves.

Mining Spatiotemporal Data

A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs (from WWE 2006)
Qiaozhu Mei, Chao Liu, Hang Su, and ChengXiang Zhai

The authors address the problem of mining spatiotemporal theme patterns from blogs and propose a means by which to extract common themes from multiple weblogs; analyze them by time period (duration) and location; and record a 'snapshop' of a particular thematic element for a specific time period. They advocate an approach to utilizing data mining for research on multi-topic themes which cannot afford to be glossed as a singular entity. Equally, the time and location of blog posts can make a qualitative difference to the sort of information sought around the world blogging population and should therefore be correlated with the multiple themes being addressed. Citing spatiotemporal instances of the term "Hurricane Katrina" as an example, they suggest that this type of analysis can enable the prediction of blogger behaviour and a greater understanding of the evolution of the blogosphere. Thus, rather than focusing on individual bloggers, this method mines a great deal of data from text, timestamps and location labels on documents to extract wider thematic trends. See also: **Extracting Topics From Weblogs Through Frequency Segments** (Mizuki Oka, Hirotake Abe and Kazuhiko Kato), where extracted weblog topics (such as "London Bombings") and human perceptions of events were compared to rank how closely mining techniques could reflect actual human response.

Mining Blog Stories Using Community Based And Temporal Clustering (from ACM Conference on Information and Knowledge Management)
Arun Qamra, Belle Tseng, and Edward Y. Chang

The researchers draw attention to the existence of popular web services for mining and analyzing blog content such as Blogpulse, Technorati and Google Blog Search, which allow keyword searchers, automated popularity ranking and identifying keyword trends in the blogosphere. They argue, however, that such programs do not address the necessity to mine and extract "cohesive discussions" from blog communities over time. Thus, they suggest a time- and community-sensitive model to cluster blog entries into "stories". Once the most recent "hot stories" or discussion topics are discovered, issues of interest to various domains and communities can be determined and analyzed. This method was also designed with marketing research in mind, such as the ability to mine product opinions across many blogs over time, thereby enhancing market intelligence. While this is the case, it equally holds opportunities for social research in tracking and grouping topics of interest across multiple blog communities.

Blog Map of Experiences: Extracting and Geographically Mapping Visitor Experiences from Urban Blogs (from WISE 2005)
Takeshi Kurashima, Taro Tezuka, and Katsumi Tanaka

These researchers are also attentive to the spatio-temporal nature of blog articles and examine this in the context of personal experiences such as those of tourists who blog

about their trips. They attempt to geographically map the activities and impressions of tourist blogs which, they argue, can be mined more specifically and accurately than a simple web search and yield more personalized data and experiences than a generic information search. The final product appears to be across between Google Maps (geographic search) and Blogger (blog posts), with tags. It may be useful for examining a geographically-bounded unit of bloggers because it allows online data to be traced and mapped onto offline locations.

Mining Identity, Mood, and Behavior

A Framework for Locating and Analyzing Hate Groups in Blogs (from the Pacific-Asia Conference on Information Systems 2006)
Michael Chau and Jennifer Xu

The researchers provide a more specific application of blog mining techniques for examining a particular social phenomenon; namely, how racist hate groups are formed and maintained through blogging communities (namely anti-Black hate groups on Xanga.com). They provide a framework for analysis consisting of a blog 'spider', information extraction, network analysis and related visualization programs. While constructed for the use of law enforcement and social work professionals, the researchers believe their methodology can be applicable to data mining projects across other fields, including security informatics, marketing analysis, and business intelligence. The benefits of this type of study is the ability to extract information regarding a relatively closed, coherent group of bloggers, as might be the case with anthropological or social research into blogging communities.

Gender Classification of Weblog Authors (from AAAI 2006 Symposia on Computational Approaches to Analyzing Weblogs).
Xiang Yan and Ling Yan

The authors are interested in identifying bloggers by gender by extracting stylistic, textual information which may be used to label its writer. They use a model which searches out key words programmed to be associated with different 'genders' as well as background colors, particular fonts and cases, punctuation marks, and emoticons. The applicability and efficacy of this method initially appears lacking and superficial; however, if other researchers are able to program discriminating features from more clearly defined categories to be mined and automatically classified or sorted into any number of groups, this application may be worthwhile.

Capturing Global Mood Levels using Blog Posts (from AAAI 2006 Symposia on Computational Approaches to Analyzing Weblogs)
Gilad Mishne and Maarten de Rijke

The researchers describe how many bloggers indicate what their mood is at the time of posting a blog entry, and suggest that this information can be mined in order to ascertain a "blogosphere state-of-mind" tracking the intensity of different moods among bloggers at that time. Their intention is to estimate aggregate mood levels across a large domain to determine the global intensity of mood within the entire "blogosphere", with the resulting goal of predicting mood trends *without* the self-

ascribed blogger mood labels. They mined specific mood labels produced by bloggers and attempted to identify words and phrases which indicate mood. The commercial applications, such as tracking public opinion regarding certain products or brands, are clear, as are those of media analysts tracking public sentiment in light of social policies. The association between certain moods and other key words (for example, "London Bombings" with "shock" or "sadness") can also be utilized for other types of research. Limitations include the subjective nature of mood and mood labeling, and the difficulty in accurately conveying mood through blog text (with all its stylistic elements) and to diverse audiences. The researchers report a strong correlation between their mood-estimating models and those reported by bloggers and conclude that predicting the intensity of moods over a time span can be done with a high degree of accuracy. (For more on this see http://moodviews.com/Moodteller.)

Conclusion and Future Research

Naturally, work on blogs and blog mining techniques is relatively new. The findings summarized above represent work which remains highly statistical and computational but with tangible linkages to the social sciences. Because of the mass marketing appeal of powerful data extraction programs which can summarize market trends[2], new programs with more sophisticated algorithms and graphic visualization tools will predictably continue to represent a majority of new developments. It is also expected that work on data mining blogs will begin to cross over with other sites which combine the features of blogging with other social networking facilities (such as MySpace, FaceBook, Bebo, etc). Blogging will continue to overlap with social bookmarking, tag sharing, photo-sharing, social networking sites and wikis as Web 2.0 environments become more elaborate. This expectation is supported by the contents of two upcoming conferences in 2007 which will hopefully expand upon the data mining tools and methods described above:

International Workshop on Data Mining in Web 2.0 Environments, held in conjunction with the IEEE International Conference on Data Mining (ICDM 2007) on October 28, 2007 in Omaha, United States.

Description:
"Users feel very attracted by currently emerging Web 2.0 environments, that [provide] content in a simple, unrestricted, and ad hoc way. Providing annotations (such as tags) in a Web 2.0-like way is applicable to a wide range of resources and data types, such as web pages, images, multimedia, etc. There is, however, a disadvantage: the freedom to provide arbitrary (personal) content and tags in ubiquitous, uncoordinated ways results in very large amounts of poorly structured information. Behind the current hype around Web 2.0 applications, this raises several important challenges for future data and web mining methods. The workshop aims to bring together researchers and professionals in the areas of data and web mining, information systems and collaborative systems to discuss challenges and solutions of applying data mining to highly unstructured, user-created data." Some of the topics of interest include analysis of blogs, visual and textual information extraction, application of web and text mining to wiki content, discovering social structures and communities and predicting user behaviour.

Workshop on Web Mining and Social Network Analysis, held in conjunction with The 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007) from August 12-15, 2007 in San Jose, California, United States.

Description:
"The Joint 9th WEBKDD and 1st SNA-KDD Workshop 2007 aims to bring together practitioners and researchers with a specific focus on the emerging trends and industry needs associated with the traditional Web, the social Web, and other forms of social networking systems. The workshop solicits experimental and theoretical work on Web mining and social network analysis, including (1) data mining advances on the discovery and analysis of communities, on personalization for solitary activities (like search) and social activities (like discovery of potential friends), on the analysis of user behavior in open fora (like conventional sites, blogs and fora) and in commercial platforms (like e-auctions) and on the associated security and privacy-preservation challenges; (2) social network modeling, scalable, customizable social network infrastructure construction, dynamic growth and evolution patterns identification and discovery using machine learning approaches or multi-agent based simulation."

Footnotes

1. The 3rd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics was held in conjunction with WWW 2006, the 15th Annual International World Wide Web conference. It built upon conferences and workshops held in previous years (WWE 2004 and WWE 2005). WWW-2006 was hosted by a mix of academic and commercial researchers, including representatives from Nielsen BuzzMetrics (Creators of Blogpulse.com), Yahoo! and Google. The conference covered all "dynamics of the blogosphere, found in trackbacks, citation links, blog-rolls, comments, tags, shared topics and interests" and researchers applied methods including "text mining, social network analysis, computational linguistics, business and marketing intelligence, library sciences, taxonometrics, graph theory and data visualization" (see WWE 2006). This varied list of interests commercial and academic interests appears to have enabled this series of conferences to beginning bridging the gap between highly commercial, computational methods of data mining and the needs of social researchers interested in understanding blogging.

2. The popularity of blogging in the teenager to under-30 population has led to new opportunities for marketing and advertising that have not gone unnoticed. See: **Giving it up for Free: Teens, Blogs, and Marketers' Lucky Break**.

All links last accessed 26 July 2007.

Francine Barone
University of Kent