Social Media Analytics: Channeling the Power of the Blogosphere for Marketing Insight

Prem Melville
IBM T.J. Watson Research Ctr.
P.O. Box 218
Yorktown Heights, NY 10598
pmelvil@us.ibm.com

Vikas Sindhwani
IBM T.J. Watson Research Ctr.
P.O. Box 218
Yorktown Heights, NY 10598
vsindhw@us.ibm.com

Richard D. Lawrence
IBM T.J. Watson Research Ctr.
P.O. Box 218
Yorktown Heights, NY 10598
ricklawr@us.ibm.com

1. INTRODUCTION

In July 2009, a survey conducted by Universal McCann[1] concluded that 32% of the 200 million bloggers world wide blog about opinions on products and brands. In addition, it was found that 71% of the 625 million active internet users actually read blogs. So not only is there an enormous amount of content on the blogosphere that present opinions on products, this content also has a massive readership rivaling any traditional media. A separate survey on *Trust in Advertising* conducted by Nielson[2] found that 78% of respondents put their trust in the opinion of other consumers. In contrast only 57% of consumers trust advertising in traditional media – newspapers, magazines, TV, and radio. So not only are a large number of consumers reading opinions on products on blogs, they are also more likely to adopt these opinions as opposed to being swayed by traditional advertising. Interestingly, despite the vast amount of resources spent on search engine advertising, only 34% of consumers put their trust in such advertising.

This rise of the blogosphere has empowered the average consumer with the ability to influence the public perception and profitability of brands. As such, marketing organizations need to be mindful of what people in general (and potential customers in particular) are saying in blogs, how the expressed opinions could impact their business, and how to extract (and drive) business insight and value from these blogs. This has given rise to the emerging discipline of *Social Media Analytics*, which draws from Social Network Analysis, Machine Learning, Data Mining, Information Retrieval (IR), and Natural Language Processing (NLP).

The automated analysis of blogs and other social media raises several interesting questions from a marketing perspective:

1. Given the enormous size of the blogosphere, how can we identify the subset of blogs and forums that are discussing not only a specific product, but higher level concepts that are in some way relevant to this product?

2. Having identified this subset of relevant blogs, how do we identify the most authoritative or influential bloggers in this space?

3. How do we detect and characterize specific sentiment expressed about an entity (e.g. product) mentioned in a blog or forum?

4. How do we tease apart novel emerging topics of discussion from the constant chatter in the blogosphere?

In this paper, we discuss techniques from the related sub-disciplines of Social Media Analytics that can be adapted to address the above problems. Within IBM Research, we have several projects underway which focus on Social Media Analytics, and we summarize our experience and other current directions in addressing the issues raised above. We also point out some of the limitations in these approaches, and discuss unexplored directions that we believe are required to advance the state-of-the-art in this rapidly evolving area.

2. SEEKING RELEVANCE

The immediate objective is to effectively filter down the vast blogosphere from millions to the thousands of blogs most relevant to the brand/product being monitored. We want not only the blogs that directly talk about the specific product, but also competing products and/or whitespace (potential new customers).

Current directions: We have explored text-based and network-based techniques, and their combinations, for relevance filtering. Text-based techniques make a judgment about relevance purely guided by the textual content of a blog. There are two natural families of such techniques, one coming from traditional IR where keyword searches are used to retrieve relevant blogs, and the other coming from traditional Machine learning, where relevance filtering is posed as a standard classification problem. These techniques differ in the type of marketer input they require. In the former case, a marketer generates a list of keywords related to the brand/product being monitored, while in the latter case, they point to blogs or blog posts that can serve as positive examples for training a classifier. Note, also, that the filtering problem is intrinsically one-class, i.e., the notion of negative examples is less well-defined. Thus, one needs to consider exclusion patterns in keyword searches, one-class classification methodologies, or the use of a random sample of unlabeled blogs as surrogate negatives.

An alternative to text-based techniques is relevance filtering via graph based methods. Seed relevant blogs may be viewed as positive labels on a large partially labeled blog graph. Under the assumption that links encode similar degrees of relevance, it is natural to apply a classical label diffusion procedure. Starting

[1] http://www.universalmccann.com

[2] http://www.nielsen.com

iteratively from the seed blogs, each node in the graph diffuses positive relevance to its neighbors [23]. A crude instantiation of this idea is *snowball sampling*, in which we start with a small set of well-known seed blogs identified by marketing experts, add all the blogs they link to, and then repeat this process for several iterations (degrees of separation). In the presence of negative examples, one may also use graph transduction [22,24] or graph kernel based classification techniques. An advantage with such approaches is that the intrinsic cluster or community structure in the blog graph can then potentially be used to more accurately access relevance. A large literature exists in Social Network Analysis on community finding in large graphs. A simple snowball sampling has the advantage that it attempts to directly detect a relevant subgraph.

An important consideration is to avoid crawling, parsing and storing parts of the blog sub-universe that are irrelevant from a marketing perspective. Our practical solution is a *focused snowball sampling*, which iteratively expands web crawling from links in blogs deemed to be relevant by a text classifier. We have also explored large-scale versions of classification models that combine graph structure and text content [25, 26, 35].

Opportunities: The goal is to efficiently discover blog sub-universes given minimal human supervision. Ideally, what is needed is a highly scalable methodology that holistically combines the strengths of various techniques to integrate all sources of information in a principled manner. This involves exploiting both the content of blogs and their network structure, while incorporating multiple forms of human supervision, e.g., in the form of a set of keywords as well as few positive examples of relevant blogs. Such a model may then be used to iteratively expand the sub-universe being discovered, and to query marketers to more judiciously provide supervision so as to actively construct an optimal relevance filter. Thus, this problem brings together opportunities in semi-supervised and active learning, use of multiple supervision [27], and the combination of text and network structure in classification models.

3. INFLUENCE AND AUTHORITY

Having identified a subset of relevant blogs, it is then useful to determine the most authoritative bloggers in this space. These are the experts or mavens whose opinions catch on most rapidly. It is important to identify this set of bloggers, since any negative sentiment they express could spread far and wide. These are the bloggers who marketers should keep most accurately informed and actively engaged. In addition to authorities, there are bloggers who are very well connected, who are most responsible for the spread of information in the blogosphere. These *influencers* are often (but not always) blogs with high authority. When presented with a large number of posts relevant to a topic, ordering them by the blogger's influence assists in information triage, given that it is not feasible to read all posts.

Current directions: Since blogs are often linked to by other blogs, this linkage information is commonly used to determine a blog's authority. For instance, Technorati[3] assigns an authority score to a blog based on the number of blogs linking to the website in the last six months. Similarly, Blogpulse[4] ranks blogs based on the number of times it's cited by other bloggers over the last 30 days.

Given that we have a network of directed edges indicating the links between posts/blogs, we can apply more complex measures of prestige from Social Network Analysis. For instance, the authority of a blog can be characterized based on the number and authority of other blogs that link to it, using PageRank. The influence of a blog can be captured by the degree to which the blog contributes to the flow of information between other bloggers, determined by Flow Betweenness [1]. By selecting the subset of nodes (blogs) that are relevant to a topic when computing these measures, one can address the issue of topical influence - i.e., a blogger maybe very influential amongst bloggers in one community, but may have little influence outside this community. In practice we found that these methods not only identified bloggers that marketers had independently determined were key players, but also found other influential blogs that should have been on their radar.

An interesting point concerning influence is that people do not necessarily link to blogs that influence them the most. For instance, many may read and be influenced by CNN, but may still not consider it necessary to have a link to CNN.com. Readership may be better correlated with authority and influence, but it is very difficult to get reliable readership information.

Opportunities: There has been a lot of recent work studying influence and the diffusion of information in social networks. Notably, Bakshy et al. [2] model social influence in the virtual world of Second Life – based on the adoption rate following the actions of a friend. Goetz et al. [3] propose a generative model to simulate the topology and temporal dynamics of the blogosphere. Leskovec et al. [4] provide a graph generator for large scale social and information networks, based on a "forest fire" model. Kossinets et al. [5] propose a framework for analyzing communication in networks, based on inferring the *potential* for information to flow between nodes. These works are very helpful to understand the dynamics of communication in networks, and allows us ways of comparing different kinds of communication within different network. However, they do not directly give us measures of influence and authority in graphs. It would be useful to follow the excellent analyses done in these papers to derive something actionable, such as a better measure of blog influence, or insight into how to increase reach in the blogosphere.

As mentioned earlier, one can use PageRank and Flow Betweenness as measures of authority and influence. However, there are several other measures that could be viable alternatives, such as *Degree Centrality, Closeness Centrality, and Betweenness Centrality* [1]. Using these network measures seem to produce a *qualitatively* appealing ranking of nodes in social networks graphs. However, in order to objectively select the best approach, it is important to *quantitatively* compare approaches. This is problematic, since the notions of authority and influence are ill-defined, and vary based on the application, and how the network was generated. A social network, such as Facebook[5], in which users explicitly declare social connections, may need to be treated

[3] http://www.technorati.com

[4] http://www.blogpulse.com

[5] http://www.facebook.com

differently from a network of possible strangers implied only by links between blogs. In the space of political blogs, bloggers often link to posts they are vehemently denouncing, and such a link should not be construed as conferring authority. Depending on the application, one may just be interested in blogs that receive a lot of links, in which case Degree Centrality may an appropriate measure. However, for the purposes of marketing we are more interested in bloggers who influence the thinking, and subsequently the content blogged by others. If a blogger is indeed influential, we would expect his ideas to propagate to other blogs. Based on this, we propose objectively comparing influence measures on the task of predicting user content generation. If a measure of influence helps you select a blog that more accurately predicts future discussion in the blogosphere, than a randomly selected blog then there is some value to such an influence measure. Furthermore, we can now precisely compare alternative approaches to identifying influencers.

Measures such as PageRank rely solely on the link between posts in the current graph. However, we also have the content of all posts and the evolution of links over time. These additional sources of information should provide much more insight into the influence of a blogger. For instance, if the content of a post "diffuses" to posts that link to it, then it's probably more indicative of the bloggers influence, than if it just receives a link. Also, if a post receives a lot of links in a short period of time, it is indicative of greater potential influence, than a post that has gathered the same number of links over a long period of time.

An associated growing strand of research that attempts to explain content and link structures together with their temporal evolution, is based on tensor factorizations and higher order extensions of classical techniques such as SVD [29,30,31]. Kolda and Bader [29] introduce a topical extension of Klienberg's Hubs and Authorities algorithm (HITS) which is based on the SVD of the link graph. In their approach, content (anchor text) and links are compactly represented in a tensor which is then analogously factorized. Chi et al. [30] apply similar ideas, but to linkage-time tensors for a given keyword. Chi et al. [31] extract blog communities and their temporal evolution using related techniques.

Another interesting approach to quantitatively evaluating the ranking of blogs is through the task of *cascade detection* – selecting a set of blogs to read which link to most of the stories that propagate over the blogosphere. Following the viral marketing work of Richardson and Domingos [6], Leskovec et al. [7] and Kempe et al. [8] provide efficient approximate solutions to the underlying optimization problem. These solutions, do not attempt to address the task of assigning an authority/influence score to individual nodes (bloggers) – since they are focused on optimal set selection. Furthermore they rely only on the links in the network and not the content associated with each node. However, there is a lot of potential for using such approaches to identify influencers.

4. SENTIMENT DETECTION

"The recommendation of someone else remains the most trusted sources of information when consumers decide which products and services to buy… Furthermore, given that nothing travels faster than bad news - with estimates that reports of bad experiences outnumber good service reports by as many as 5:1 - the importance of responsive, high quality customer service is yet again highlighted." – David McCallum, Global managing director for Nielsen's Customized Research Services.

Considering that it's virtually impossible to read all user-generated content, it has become crucial to automatically identify negative sentiment in blogs to enable rapid response. Additionally, significant insight can be gleaned from tracking the trend of sentiment expressed around products, and also observing how consumers respond to marketing actions, such as events and press releases.

Current directions: There has been significant work in NLP on detecting subjective versus objective statements, and the polarity of sentiment expressed in subjective statements. Much of this prior work is on knowledge-based approaches that primarily use linguistic models or other forms of background knowledge to classify the sentiment of passages. A large focus of this area is the use and generation of dictionaries capturing the sentiment of words. These methods range from manual approaches of developing domain-dependent lexicons [9] to semi-automated approaches [10][11], and even an almost fully automated approach [12]. More recently, Pang et al. [13] successfully applied a machine learning approach to classifying sentiment for movie reviews, where they cast the problem as a text classification task using a bag-of-words representation of each movie review. They demonstrated that a learning approach performs better than simply counting the positive and negative sentiment terms using a hand-crafted dictionary.

Opportunities: The expression of sentiment tends to be domain specific; hence it is preferable to use supervised approaches that can adapt to new domains. However, the set of domains to monitor may change often, requiring classifiers to adapt rapidly with the minimum of supervision. Fortunately, the task of sentiment detection lends itself well to emerging techniques for reducing supervision in classifier induction, which we discuss below.

Supervision for a sentiment classifier can be provided not only by labeling documents, but also by labeling words. For instance, labeling a word such as "tragedy" as negative is one way to express our prior belief of the sentiment associated with it (though in the context of movie reviews it is often used in positive reviews). It is possible to learn from such labeled words in conjunction with labeled documents in our proposed framework of Dual Superversion. We have demonstrated that by labeling a few words and few documents we can learn an accurate model that requires less supervision than labeling only documents or only words [14]. In particular, we construct a generative model based on a lexicon of sentiment-laden words, and a second model trained on labeled documents. The distributions from these two models are then adaptively pooled to create a composite multinomial Naïve Bayes classifier that captures both sources of information. By exploiting prior lexical knowledge we dramatically reduce the amount of training data required. In addition, by using some labeled documents we are able to refine the background knowledge, which is based on a generic lexicon, thus effectively adapting to new domains. We have demonstrated the generality of our approach on three, very different domains — blogs discussing enterprise-software products, political blogs discussing US Presidential candidates, and online movie reviews.

Given the vast size of the blogosphere, there is no dearth of unlabeled posts that can be exploited through semi-supervised learning to further reduce the amount of supervision required even in the dual supervision setting. In [15] we proposed such an approach to incorporating labeled words and unlabeled documents within standard regularized least squares. In settings where labeled data is very limited and unlabeled data is abundant, our approach performs better than purely supervised and competing semi-supervised techniques.

Even though there are expressions of sentiment that are domain-specific, there is still a large amount of overlap in how positive and negative emotion is conveyed across domains. This enables the use of *transfer learning* to adapt a classifier trained in one domain to a new domain with little to no labeled data in the target domain [16].

5. EMERGING TOPICS

Using measures of relevance, authority and sentiment help us direct our attention to the posts that are most important for us to read. In addition to focusing on the micro-level of posts, in order to really track the pulse of the blogosphere, it is important to detect phenomena that are only visible at the macro-level of the universe of relevant blogs. Given the plethora of posts being produced continually it is challenging but very valuable to determine large-scale patterns of what people are blogging about, and find emerging areas of discussion.

Current directions: One commonly used approach to capture the notion of *hot* topics is to identify frequently occurring key-phrases; and there has been considerable work in Natural Language Processing in identifying such commonly occurring collocation of consecutive words [17]. For instance, such an approach may identify that "Barack" and "Obama" are words that frequently appear together, and as a phrase they are mentioned many times in political blogs. However, this fact may not be interesting to report, if the phrase is always frequently mentioned. Instead, if we compare the relative frequency of occurrence of phrases in the current week to the occurrences over the past month, we are likely to identify more topical (currently relevant) phrases, such as "public healthcare."

Following on these lines, Tomokiyo and Hurst [18] propose an approach for extracting such key-phrases based on statistical language models which assign a probability of every sequence of words being generated. Given a language model of a background corpus (past blog posts) and a model of a foreground corpus (current blog posts) one can determine if a sequence of words has (1) high *phraseness* – do these words appear more frequently together than one would expect if they were independent, and (2) high *infomativeness* – is a phrase more likely to be generated by the foreground model compared to the background model. In practice, the *phraseness* and *informativeness* of key-phrases can be computed using the Kullback-Leibler divergence between different language models [18]. A similar approach is used by Amazon[6] to identify Statistically Improbable Phrases (SIPs), by comparing all of the books they index and finding phrases in each that are the most unlikely to be found in any other book.

Opportunities: Methods for key-phrase detection are computationally quite expensive and are rarely applied to identifying more than bigrams on large corpora. These method can bring to our attention that certain named entities, like specific products and people are being extensively discussed, but from a marketing perspective we also want to know what is being said about them. Knowing that "Barack Obama" and "public healthcare" are being frequently discussed is less informative than knowing that "Barack Obama downplays public healthcare."

An alternative way to approach this problem is to identify sets of posts that are discussing the same topic. A lot of research in the area of document clustering and topic modeling with Latent Dirichlet Allocation [19], Probabilistic Latent Semantic Analysis [28] and Non-negative Matrix Factorizations [32] can be brought to bear here. We can report each tight cluster of posts detected by presenting the most representative post in the cluster, or by exploiting some of the approaches in automatic multi-document text summarization [20] in order to generate a summary of the collection of posts. Traditional clustering techniques which rely only on the text in documents can be further improved by exploiting the assumption that posts are likely to link to other posts on the same topic; and there is some evidence that this is a viable approach [21].

Although, clustering and topic modeling techniques can find sets of posts expressing cohesive patterns of discussion, for generating marketing insight we need to identify clusters that are also novel or informative compared to previous streams of discussion. This objective can be achieved by clustering current posts while forcing them to be different from clusters of past posts. One approach to doing this is to transform the space in which points (posts) are being clustered, so that distances (dissimilarities) between points that are closer to the background (past) distribution are considered smaller than distance between points that are further away. This has the effect of putting documents that are similar to previous discussions in the same cluster, while encouraging the detection of novel clusters that are discussing new threads.

Several recent papers have developed models of temporal evolution of topics in document streams [see, e.g., 33,34 and references therein]. The main idea is to construct topic models in different time-windows tying them together to maintain some form of temporal continuity. In this direction, we are currently exploring regularized non-negative matrix factorizations. An associated problem is that of *Guided Topic Evolution*, that is, to incorporate feedback from a user in an online fashion as to which topics should be tracked, or discarded, from the analysis.

In addition to detecting the current buzz, we are also developing sophisticated methods designed to predict future buzz today. The idea is to build time-series models to forecast future topics and word-usage patterns given historical data. The goal is two-fold: one, to better prepare marketers for ongoing and emerging discussions on the blogosphere, and two, to use the model to understand which predictive variables are driving tomorrow's buzz, and to study/characterize authority and influence based on notions of causality.

[6] http://www.amazon.com

6. ACKNOWLEDGMENTS

Our thanks to Wojciech Gryc, Estepan Meliksetian, Yan Liu and Claudia Perlich.

7. REFERENCES

[1] S. Wasserman and K. Faust, *Social Network Analysis: Methods & Applications*, Cambridge, UK: Cambridge University Press, 1994.

[2] E. Bakshy, B. Karrer, and L.A. Adamic, "Social influence and the diffusion of user-created content," *ACM Conference on Electronic Commerce*, 2009.

[3] M. Goetz, J. Leskovec, M. Mcglohon, and C. Faloutsos, "Modeling Blog Dynamics," *AAAI Conference on Weblogs and Social Media*, 2009.

[4] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," *ACM Trans. Knowl. Discov. Data*, vol. 1, 2007, p. 2.

[5] G. Kossinets, J. Kleinberg, and D. Watts, "The structure of information pathways in a social communication network," *KDD*, 2008.

[6] P. Domingos and M. Richardson, "Mining the network value of customers," *KDD*, 2001.

[7] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. Vanbriesen, and N. Glance, "a. Cost-effective outbreak detection in networks," *In KDD*, 2007.

[8] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," *KDD*, 2003.

[9] S. Das and M. Chen, "Yahoo! for Amazon: Extracting market sentiment from stock message boards," *Asia Pacific Finance Association*, 2001.

[10] S. Kim and E. Hovy, "Determining the sentiment of opinions," *Intl. Conf. on Computational Linguistics*, 2004.

[11] H. Yu and V. Hatzivassiloglou, "Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences," *EMNLP*, 2003.

[12] P. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," *ACL*, 2002.

[13] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," *EMNLP*, 2002.

[14] P. Melville, W. Gryc, and R. Lawrence, "Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification," *KDD*, 2009.

[15] V. Sindhwani and P. Melville, "Document-Word Co-Regularization for Semi-supervised Sentiment Analysis," *ICDM*, 2008.

[16] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification," *ACL*, 2007.

[17] C.D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, Cambridge, MA: MIT Press, 1999.

[18] T. Tomokiyo and M. Hurst, "A language model approach to keyphrase extraction," *ACL 2003 Workshop on Multiword Expressions*, 2003.

[19] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, 2003.

[20] I. Mani and M. Maybury, *Advances in Automatic Text Summarization*, MIT Press, 1999.

[21] M. Shiga, I. Takigawa, and H. Mamitsuka, "A spectral clustering approach to optimally combining numerical vectors with a modular network," *KDD*, 2007.

[22] M. Belkin, I. Matveeva, P. Niyogi, “Regularization and Semi-supervised Learning on Large Graphs”, *COLT*, 2004.

[23] D.Zhou, O. Bousquet, T.N. Lal, J. Weston, B. Scholkopf, “Ranking on Data Manifolds”, *NIPS*, 2004.

[24] X. Zhu, “Semi-supervised Learning literature survey”, Tech Report 1530, Dept. of Comp. Sci., Univ. of Wisconsin, Madison.

[25] V.Sindhwani, “On Semi-supervised Kernel Methods”, Doctoral Thesis, University of Chicago, 2007.

[26] T. Zhang, A. Popescul, B. Dom, “Linear Prediction Models with Graph Regularization for web-page categorization”, *KDD*, 2006.

[27] V.Sindhwani, P. Melville, R. Lawrence, ”Uncertainty Sampling and Transductive Experimental design for Active Dual Supervision”, *ICML*, 2009.

[28] T. Hoffman, “Probabilistic Latent Semantic Analysis”, *UAI*, 1999.

[29] T. Kolda and B. Bader, “The TOPHITS model for higher order web link analysis”, Workshop on Link Analysis, Counter-terrorism and Security, *SDM*, 2006.

[30] Y. Chi, B.L. Tseng, J. Tatemura, ”Eigentrend: trend analysis in the blogosphere based on singular value decompositions”, *CIKM,* 2006.

[31] Y.Chi, S. Zhu, X. Song, J. Tatemura, B.L. Tseng, “Structural and Temporal Analysis of the Blogosphere Through Community Factorization”, *KDD,* 2007.

[32] D. D Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization”, *Nature*, 401, 788-791, 1999.

[33] D. Blei and J. Lafferty, “Dynamic Topic Models”, *ICML*, 2006.

[34] A. Gohr, A. Hinneburg, R. Schult and M. Spiliopoulou, “Topic Evolution in a stream of Documents”, *SDM*, 2009.

[35] S.A. Macskassy and F. Provost, "Classification in Networked Data: A Toolkit and a Univariate Case Study," *J. Mach. Learn. Res.*, vol. 8, 2007, pp. 935-983.