Temporal Issue Trend Identifications in Blogs

Il-Chul Moon[1] Young-Min Kim[2] Hyun-Jong Lee[2] Alice H. Oh[2]
[1]Department of Electrical Engineering, [2]Department of Computer Science
KAIST
Daejeon, Republic of Korea
icmoon@smslab.kaist.ac.kr, creatcross@hanmail.net, lhjgine@kaist.ac.kr, alice.oh@cs.kaist.ac.kr

Abstract— **Many blog posts deal with current issues, so much attention has been paid to identifying topic trends in blogs. This paper suggests a new metric of selecting topic words. We empirically tested the accuracy and the performance of the metric with a massive blog corpus. First, we created blog site groups to their indegree influence. Second, we ran the metric with blog posts of each group. The test was encouraging because the metric identified key issues matching to the headlines of New York Times when it is applied to the top indegree blog group. We expect that this metric and the source grouping methods will be developed to a new topic analysis framework of a large blog corpus.**

Keywords-component; Blog, Social Media, Issue Identification

I. INTRODUCTION

Weblogs (blogs) are a significant subset of social media that have gained much popularity among academic researchers in recent years. Researchers from diverse fields [22, 23] including social science, political science, journalism, and computer science, are actively studying blogs because they present new phenomena and challenges that had not been seen in traditional media and Web documents. Blogs pose especially interesting problems for social computing researchers because 1) blogs promote social interactions through links, comments, subscriptions, and other tools, 2) blogs contain highly time-sensitive information, and 3) blogs can be widely accessed and searched through the World Wide Web. These three characteristics of blogs translate respectively to a need for 1) analyses of complex social interactions through this novel medium, 2) (near) real-time processing of a large dynamic corpus of blog contents and metadata, and 3) a fresh perspective on Web search where users, contents, and information needs are highly dynamic and diverse.

We present in this paper our recent research that takes an initial look at those three aspects. We analyze a large blog corpus, released to the research community with the intention of bringing a reference corpus of blogs, for temporal trend identification. It is a successful experiment at looking at a large, temporal blog contents and metadata, the second aspect listed above. It looks at the social interactions through an automatic categorization of blogs by the level of social interactions (via degree of in-links), and the effect of that new categorization on issue identification. Lastly, it looks at the diversity of the blog types, in a simple analysis of the level of authorship and readership of a particular blog.

Identifying topic trends require solutions in three critical problems. First, we need to look into a large corpus of media contents collected from different sources. This problem is difficult because of possible privacy concerns over the course of data collection. Even we collected large corpus, the corpus makes the data processing harder due to its large volume. Second, finding hot issues is like finding a needle in a haystack. Since these multiple sources focus on diverse issues, we need a method to identify which issue is more popular or more prolonged than others. Third, we might have to distinguish the importance of sources. The internet media is networked, and each media site has its own influence on this network. Some sources would be more important than others, which will impact the issue identification process.

This paper suggests headways to resolve the three problems by analyzing large blog corpus from multiple blog sites. First, we create an issue identification metric that run reasonably fast in a large corpus. Second, we analyze the blog influence from blogs' indegree scores, and we discriminate the sources to reduce noises. Third, we test the metric and the source discrimination on a massive blog dataset from Spinn3r [8]. Our analysis indicates that a simple data processing can reveal popular issues hidden in flat texts. Also, we observed the temporal trends of the found issues. For instance, an issue about a short political convention ended quickly while a discussion about a prolonged economic depression lasted long period. This issue and temporal trend identification provides insights into how to analyze and build models upon general blog datasets.

II. DATASET DESCRIPTION

We used the blog dataset[1] provided by Spinn3r [8]. This dataset includes useful metadata in addition to textual contents of blog posts, allowing the users of the dataset to analyze the data in several ways. There are a few metadata tags that we found interesting, and there are some questionable ways in which the data are grouped, so we discuss these interesting tags and data groupings in this section.

A. Data Structure

This blog corpus is a collection of multiple posts from multiple sources over a certain time period. The dataset is structured such that there is an XML data structure for a single blog post, but not a specific data structure for a group of posts within a single blog site. However, this XML structure for a blog post contains information to specify when and where the blogs are posted and what the contents were. Additionally, there are three interesting metadata tags for each blog post in the dataset. Other tags, such as permalink, date, and author

[1] This is the same dataset used at Data track, the 3rd International AAAI Conference on Weblogs and Social Media.

TABLE I. BLOG CORPUS META-DATA

Variables	Total	Tier1	Tier 2	Tier 3	Tier 4	Tier 5	Tier 6	Tier 7	Tier 8	Tier 9	Tier 10	Tier 11	Tier 12	Tier 13
Time span	Aug. 1. 2008 – Sep. 31. 2008 (62 days)													
# of sites	707117	174191	26022	132218	107754	40559	26141	24583	22058	11912	11286	10452	32376	102354
# of posts	11636303	7642461	91523	512219	275389	102700	70194	70194	59636	57531	57723	171570	17587	2487654
Avg # of wds/post	69.2	76.6	121.7	20.5	15.6	22.1	50.1	68.6	51.8	91.5	120.9	69.0	34.9	62.7
# unique wds/post	51.2	56.6	81.2	17.6	14.3	18.6	36.1	47.2	37.8	63.5	82.0	49.1	27.5	46.4
Avg indegr	175.62	234.97	3.56	0.18	0.02	0.32	1.27	2.09	2.29	100.25	11.48	0.76	1.28	96.59

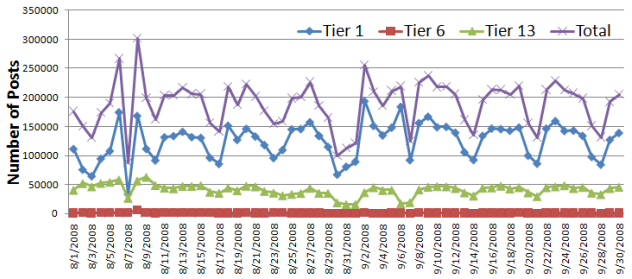

Figure 1. Blog Post Counts by Date

information, are described in the Spinn3r website, and since we did not use any of the other metadata tags other than the post date, we will not include explanations of them in this section.

1) Indegree: According to the data provider, Spinn3R, the indegree score is "the raw number of inbound links to the blog since this blog has been part of our index." We believe that this score is similar to the number of trackbacks receiving from other sites. However, there is no specification on the nature of "inbound", and there are no lists of the sites making those inbound links to a certain site.

2) IRanking: IRanking is the influence ranking of a specific web site. Spin3r [8] says that IRanking is a measure of how successful the site was "at creating and participating in memes on the Internet." However, there is no details about the IRanking calculation except that it is a function combining *memetracker* algorithm and inbound counts.

3) Tier: Tier is a metric that results by taking the IRanking and dividing it by 1000. Therefore, this metric has an approximated distribution to IRanking. Though the data provider gave us IRanking and Tier values, their contribution and interpretation are unclear due to the lack of information on how exactly IRanking is calculated. We decided to use only the Tier value in analyzing our data because the Tier value groups the blog posts nicely into 14 groups, and the IRanking value has a similar distribution to the Tier value.

B. Data Preprocessing

Before our data analysis, the blog corpus went through a couple of preprocessing steps. The purpose of preprocessing was to obtain clear natural language texts from corpus and to discard entries written in different language codes. The first preprocessing step was excluding blog posts whose language code is other than 'EN'. This limits our analysis to the English texts only. The second preprocessing was clearing HTML codes and tags from the blog texts.

C. Metadata Statistics

Table 1 displays the corpus metadata. By observing the metadata distribution, we found an anomaly. The numbers of posts from Tier 1 to Tier 13 are skewed. Tier 1 and Tier 13 have the most of blog posts, and the other Tiers have far fewer posts. Considering that the tier group is determined by the site influence ranking, there must be some mid-level ranked sites, but there are far fewer sites than we expected. Furthermore, the average indegree scores of tiers are anomalous. The highest indegree average of tier 1 is expected, but tiers 9 and 13 have the second and the third highest indegree scores, respectively. Interpreting this result with the rough definition given in the IRanking explanation above, this means that either the sites in tiers 9 and 13 have created and participated in far fewer memes on the Internet; or the influence of indegree score in determining IRranking is minimal. There is no way to interpret this anomaly further without more specifications on the IRranking calculation. In this paper, the anomaly motivated us to create alternative tier groups based on only the indegree scores.

Also, some blog texts, tagged with the xml tag description in the dataset, were not recorded as a whole but were cut off in the middle of the posts. We hypothesize that this was because some blog managing sites (e.g., MySpace, wordpress, etc.) truncate the contents of the blog posts when they are sent out via RSS. The resulting corpus, consisting of partial documents rather than the entire posts, may introduce biases if users of the dataset analyze the data in terms of words, phrases, and other natural language characteristics of the description part (i.e., textual content) of the blog posts. As described in the next section, we use words in the blog description to analyze and identify issue trends in the corpus, and we made the assumption that the truncated descriptions are representative of the entire descriptions. This may or may not be problematic, but we cannot confirm it without re-retrieving the original posts with the permalinks given in the dataset.

After observing the metadata, we examined the number of blog posts by date (See Figure 1). This figure suggests two points. First, the number of posts other than tiers 1 and 13 are too small to be included in the analysis. We show a comparison among the number of posts in tiers 1, 6, and 13 to illustrate this point. All other tiers are similar in terms of the post counts. Second, the post counts show a regular pattern

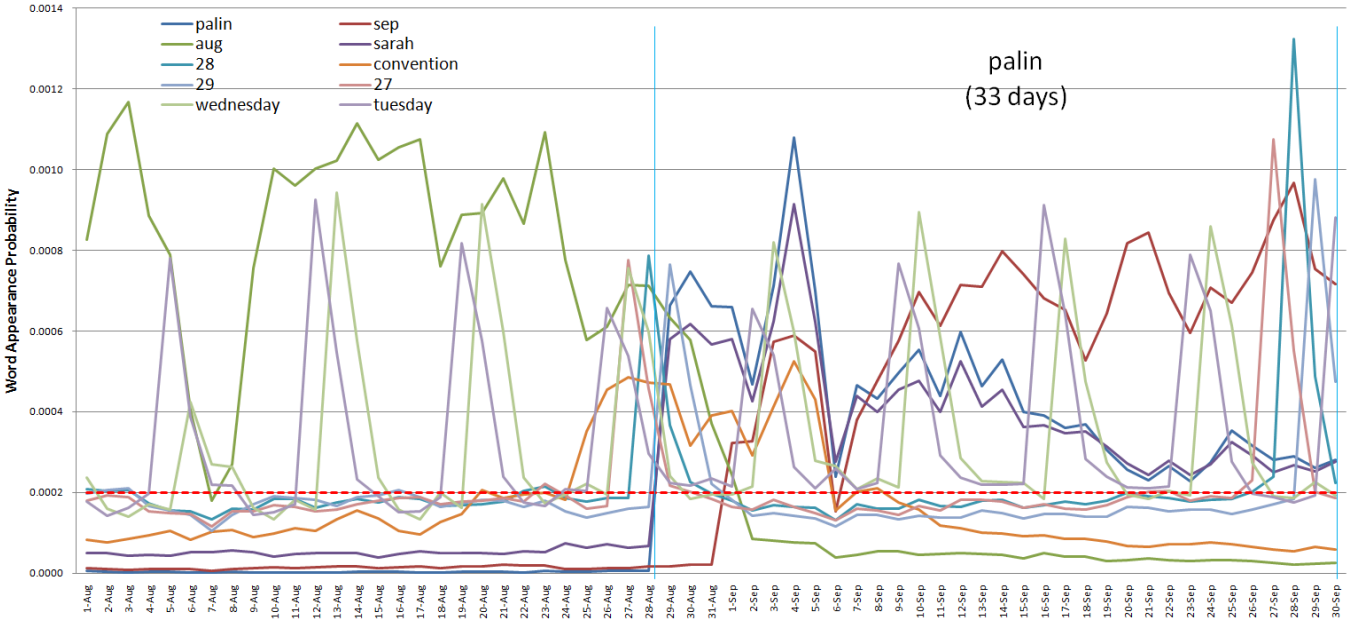

Figure 2. Old Tier-1 word appearance probability fluctuations. The IIM identified top 10 issue words displayed in the legend. The dotted line represents the issue activation threshold which is arbitrarily decided.

TABLE II. ISSUE WORDS IDENTIFIED BY IIM FROM OLD TIER GROUPING

IIM Rank	Total	Tier1	Tier 2	Tier 3	Tier 4	Tier 5	Tier 6	Tier 7	Tier 8	Tier 9	Tier 10	Tier 11	Tier 12	Tier 13
1	palin	palin	links	olympics	olympics	olympics	aug	hurricane	palin	palin	tags	penguin	palin	10
2	sep	sep	content	palin	palin	palin	sarah	beijing	sarah	sarah	posted	sarah	region	msn
3	aug	aug	olympic	speech	republican	hurricane	mccain	olympic	aug	republican	financial	mccain	sarah	tel
4	sarah	sarah	olympics	29	hurricane	29	october	olympics	october	st	politics	aug	georgia	aug
5	wednesday	28	beijing	hurricane	sarah	sarah	august	palin	financial	paul	september	posted	comments	wednesday
6	28	convention	website	republican	july	28	obama	subscription	china	speech	economy	china	troops	hotmail.com
7	29	29	visit	sarah	oct	aug	letter	timjblair	9	financial	august	features	aug	tuesday
8	tuesday	27	posted	financial	sep	27	28	convention	july	democratic	election	alt	afghanistan	blog
9	august	wednesday	september	sep	aug	related	haha	unsubscribe	mccain	gold	26	img	china	thursday
10	palin	tuesday	boys	aug	28	republican	26	subscribed	august	wednesday	october	chinese	forces	sept

within a one-week cycle. For reasons not easily seen from the dataset, the post counts predictably drop on Saturdays and Sundays. This may be due to lifestyle patterns of bloggers, but a separate study is needed to verify that guess.

III. DATA ANALYSIS

We analyzed the dataset with the main purpose of understanding topic and issue trends within the dataset, particularly with respect to the different tier groups that the dataset is composed of. First, we analyzed the natural language texts in the descriptions to identify the temporal issue trends. We found meaningful issue words by date and tier groups. The results of this analysis led us to the second step of forming new tier groups and performing the same analysis again. Because of the anomalies discussed in the previous section with respect to the original tier groups, we regrouped the posts according to the indegree scores and performed the same analysis to see the impact of the new tier groups based purely on indegree scores.

A. *Issue Identification Measure*

We developed a metric to identify popular issue terms for a given day. This metric is based on the unigram language model of the dataset. Before the actual metric description, we define the following basic concepts.

$$\text{Occurrence}_i^j = (\text{number of posts with word}_i \text{ on day}^j) \quad (1)$$

$$\text{Vocabulary}^j = (\text{a set of unique words on day}^j) \quad (2)$$

$$\text{Period} = (\text{number of days during the data collection}) \quad (3)$$

The first step is calculating the probability of word occurrence. We experimented with two probabilistic modeling approaches. First is a word frequency-based multinomial model in which we counted the number of occurrences of a word if that word occurs multiple times in a single blog post. Second is a Bernoulli presence-based model in which we assigned a value of 0 to indicate the absence of a word in a post, or 1 to indicate the presence of a word, regardless of the number of word occurrences. We found that the latter

TABLE III. TIER-1 ISSUE WORD INTERPRETATION FROM OLD TIERS

Words	First appearance date on NYT	First issue activation date on blog corpus	Issue description
palin	25-Aug	29-Aug	Running mate of McCain
sep	-	01-Sep	Name of the next month
aug	-	01-Aug	Name of the current month
sarah	25-Aug	29-Aug	Running mate of McCain
28	-	01-Aug	Name of current date
convention	25-Aug	20-Aug	2008 Democratic National Convention, in denver. Aug. 25 – 28.(found in NYT)
29	-	02-Aug	Name of current date
27	-	23-Aug	Name of current date
wednesday	-	01-Aug	Name of current day
tuesday	-	05-Aug	Name of current day

approach is better because a few sites, such as spam sites luring search engines, repeatedly placed the same words. Also, this can mitigate the bloggers' word usage biases.

$$P_i^j = \frac{\text{Occurrence}_i^j}{\sum_{i \in \text{Vocabulary}} \text{Occurr ence}_i^j} \quad (4)$$

The next step is normalizing the word occurrence probabilities. As will be explained in more detail later, our metric is based on the standard deviation of the word occurrence through the period. However, the amount of standard deviation is influenced by the average level; a small fluctuation of high average is larger than a big fluctuation of low average. Hence, we need to normalize the average of the occurrence probability to exclude the effect from the high average. The variable defined below models the normalized occurrence probability.

$$\text{NormP}_i^j = \frac{P_i^j \times |\text{Period}|}{\sum_{j \in \text{Period}} P_i^j} \quad (5)$$

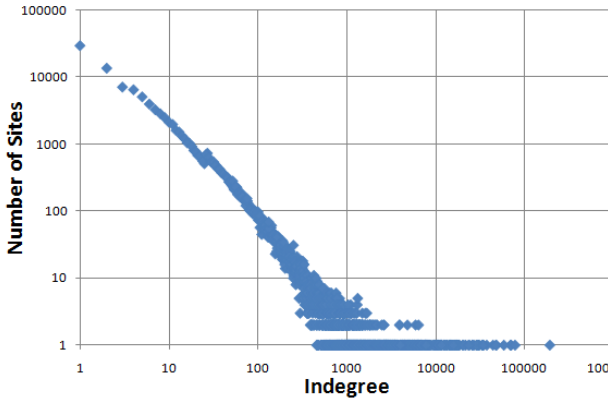

Figure 3. Indegree distribution from collected blog sites

TABLE IV. META INFORMATION OF THE NEW INDEGREE BASED TIERS

Tier	Indeg. Thresh.	# of Sites	Avg. Indeg.	URL with 'news'	URL with 'word-press'	URL with 'myspace'
New Tier 1	Indeg. >10000	65	24536.28	16.92%	0.00%	0.00%
New Tier 2	10000> Indeg. ≥1000	773	2499.28	5.43%	0.26%	0.00%
New Tier 3	1000> Indeg. ≥100	8299	260.00	2.71%	2.95%	0.00%
New Tier 4	100> Indeg. ≥10	35194	32.74	1.05%	10.94%	0.00%
New Tier 5	10> Indeg. ≥0	662742	0.33	0.47%	20.53%	53.76%

The final step is calculating the standard deviation of the normalized occurrence probability. For example, if a word, 'Russia', appears suddenly during the period (a skewed distribution), we weigh that the word might be an issue word. On the other hand, a word, 'Open', appears regularly (a uniform distribution), then the word is not an issue word. Hence, using the standard deviation might yield the issue word identification. We named this variable as Issue Identification Measure, or IIM.

$$\text{IIM}_i = \sqrt{E_j\left(\text{NormP}_i^j - E_j\left(\text{NormP}_i^j\right)\right)^2} \quad (6)$$

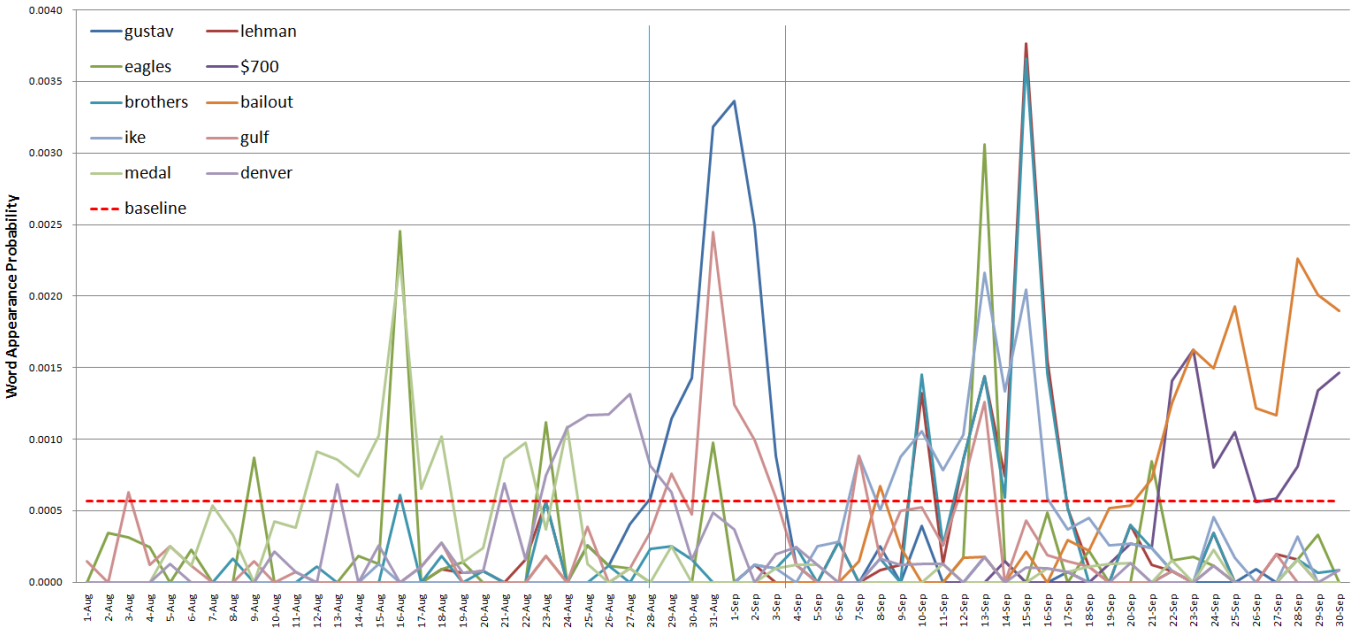

Figure 4. New Tier-1 word appearance probability fluctuations. The IIM identified top 10 issue words displayed in the legend. The dotted line represents the issue activation threshold which is 0.0005 which is arbitrarily chosen.

We applied this metric to words which were in the top 1000 ranks in terms of the word occurrence probability over the course of the collection period. Table 2 is the top ten selected words for each tier, and Figure 2 shows the word occurrence probability fluctuations through the observation period. We looked for the interpretation of the frequently appearing IIM issue words (see Table 3). This result indicates that this IIM measure can find the issue words, but there are some noises as well. For instance, 'palin', 'convention', and 'sarah' are issue words about the Republican Vice Presidential nominee. On the other hand, the algorithm also identified some words that are obviously not popular issue words, such as '29', '27' (presumably dates), 'Wednesday', and 'Tuesday'.

B. Indegree Based Corpus Reorganization

While the above data analysis utilized only text and tier information, we hypothesized that the sites with high indegree scores might have much cleaner texts to identify key issues. As we noted in Table 1, the indegree distribution was not the main factor in deciding tier groups because tier 13 has higher indegree than tier 2. On the other hand, some papers, such as [9] argue that a higher degree centrality is an indicator of issue diffusion. Therefore, this time, we tested our issue analysis on more inlink oriented tier groupings to see whether we find more meaningful and clean results.

In order to test our hypothesis, we reorganized the given tier groups based on the indegree score distribution. Figure 3 is the indegree distribution of this blog corpus. We counted the number of blog sites with a certain indegree score and created a scatter plot for indegree scores and numbers of sites. It should be noted that both axes are in logarithmic scales. This distribution is a typical modified power-law distribution discovered in many Internet researches [10]. This log-log distribution suggests that we should set the indegree threshold exponentially if we regroup the sites by the indegree variable. Also, this distribution predicts that the exponentially increasing threshold setting will result in the exponentially decreasing new tier group sizes. Table 4 illustrates the new tier grouping results. First, the new tier 1 has the smallest number of sites whereas the old tier 1 has the largest one. Second, the average indegree scores are now exponentially decreasing as going to lower tiers while the old tier groupings did not show this trait. Third, the brief site characteristic identification is done by looking into whether a site URL contains a certain word or not. For instance, if we see 'myspace' in a URL, we may conjecture that the site is hosted by Myspace, and the site owner is an individual doing a personal blogging. On the other hand, if a URL contains 'news', we might guess that the blog site is actually a news blog site maintained by a traditional news media. The new tier grouping shows that the new tier 1 is more like a news blogging site cluster, and the new tier 5 looks like a personal blog cluster. Through this new clustering, we might be able to find key issues from influential news outlets by analyzing the top tier; and to identify common topics among personal blogs from the lower tiers.

C. Issue Identification on New Tiers

After the tier reorganization, we applied the same IIM to the new tiers. Table 5 enumerates the top 10 issue words, and Figure 4 displays the top 10 issue word temporal trends from new tier 1. Compared to Table 2, we can see that the issue words from the Table 5 are more reasonable. For instance, the old tier results have common words, like month names 'sep', 'aug', and weekday names, 'wednesday'. These are hardly issue words. On the other hand, the new tier 1 issue words in Table 5 lists 'gustav', the name of hurricane; 'lehman', the company bankrupted; and $700, the amount of public policy fund which was a hot issue. You can compare the implication

TABLE V. ISSUE WORDS IDENTIFIED BY IIM FROM NEW TIER GROUPING

IIM Ranking	Tier 1	Tier 2	Tier 3	Tier 4	Tier 5
1	gustav	tuesday	leecher	palin	sep
2	lehman	wednesday	seeders	sarah	Aug
3	eagles	26	hash	aug	sarah
4	$700	thursday	mb	convention	28
5	brothers	sunday	category	governor	wednesday
6	bailout	29	palin	debate	29
7	ike	debate	privacy	speech	tuesday
8	gulf	monday	status	sept	august
9	medal	palin	sarah	crisis	27
10	denver	saturday	olympics	financial	thursday

TABLE VI. TIER-1 ISSUE WORD INTERPRETATION FROM NEW TIERS

Words	First appearance date on NYT	First issue activation date on blog corpus	Issue description
gustav	26- Aug	28-Aug	Hurricane name landed
lehman	-	10-Sep	A bankrupted company
eagles	-	09-Aug	-
$700	19-Sep	22-Sep	$700 Million dollar bailout plan
brothers	-	16-Aug	A bankrupted company
bailout	05-Aug	08-Sep	A heavily discussed government policy
ike	06-Sep	07-Sep	Hurricane name landed
gulf	27-Aug	03-Aug	Hurricane Gustav as it heads into the Gulf of Mexico
medal	-	12-Aug	Olympic medal
denver	25-Aug	13-Aug	2008 Democratic National Convention, in denver. Aug. 25 – 28.(found in NYT)

of found issues between new tiers and old tiers by going through Table 3 and 6.

Additionally, we can see the impact of the new tier grouping by comparing the first new tier and the new last tier. While the issue words from the new tier 1 are reasonable issue words as explained above, the issue words from the new tier 5 are just like the old tier results that contains only common words. We conjecture that this difference is coming from the nature of sources. As in Table 4, the new tier 1 are likely to be related to news services, and the new tier 5 seems to be personal blogs. Since the new outlets blogged the popular issues consistently, such constantly issued words become outstanding in the metrics. On the other hand, the posts in the lower tiers blogged their own personal issues, not in a consistent manner, so the issue words have not emerged from the metric perspective. We need further investigation to confirm this conjecture.

Finally, the time gap between the first appearance on NYT[2] and the issue activation date in Table 6 is interesting. Most issue words were identified in the NYT earlier than in the blogs. However, some words took longer time to be activated than the other words. For example, 'bailout' almost took a month to be active in the blogs whereas 'gustav' took only two days. This might indicate the different issue focus than the traditional news media and a blog collective. However, it should be noted that the issue activation identification required us to hypothesize the activation threshold that we assumed that an issue word is active if its appearance probability is higher than 0.0005 (this thresold is depicted as a dotted line in Figure 4).

IV. RELATED WORKS

We consider that this data analysis on this blog dataset demonstrates two possibilities. First, this issue identification shows interesting issue dispersion patterns in blogs which can be considered as emerging social media. Already, there are research efforts establishing blogs as social media and characterizing from such perspective, and we reviewed them. Second, we found some works performing similar issue identificiation tasks though they were often using either natual language processing or social network analysis methods.Blogs as Social Media

A. Blogs as Social Media

First, we found a work comparing the traditional and the emerging social media. Lloyd et al [11] tried to compare topics between blog and newspapers over the same six week period. *T*hey *concluded that a* blog provides a forum for a much larger and potentially representative group of correspondents than conventional media. *Also,* blogging analysis can be used to determine the collective public opinion on current events. And one of their contributions is that they contrast the relative blog/new interest among a wide range of topics. Galitsky and Kovalerchuk [12] introduced that weblogs are frequently viewed as an upcoming substitute for mass media. Readership is increased by 58% in 2004; 7% of Internet users are blogging (8 million Americans), and 27% of Internet users read weblogs (32 million Americans)• 38% of Internet users know what a weblog is (46 million Americans), in accordance to BlogPulse.

Another noticed distinction between two media types is the existence of replies, which is an author-reader interaction. Mishne and Glance [13] analyzed the relations between body posts and their comments. They came up with the popular weblogs based on the average page view counts and the average incoming links, which we did in this paper. Then, they used a decision tree to find disputes in comments. Fusing content analysis and link analysis is same as this paper, but our paper is focused on the issue identification, rather than the dispute identification. Furthermore, their paper was limited to find disputed comments, not the actual disputed topics.

[2] We used the New York Times website and its search function. The activation date is the date when the issue words appear first time in the search result.

However, we presented an automated identification on issue words, not the posts with active issues.

Finally, the distrinction between two media types is the existence of spam among blogs. Three papers [14, 15, 16] showed the existence of spam blog; provided their basic statistics, i.e. number of visits and number of terms in a post; and suggested filtering methods. As discussed in the later section of our paper, we suspect that one source of error in identifying issues across the blogs is the existence of spam blogs. We also conjecture that our clustering somewhat filters such blogs in our blog cluster process. However, this front of research still remained as a further research topic in our paper.

B. Blog Issue Identification

According to our reviews, there have been blog issue identification works, but they were limited to using either natual language processing or social network analysis approaches. For example, Wu and Tseng [27] tried to rank weblogs weighted link-based. They summarized hot story where a hot story is the discussion that attracts various weblogs` attention. Thelwall [18] researched the blogger activities after the London terrorist attack. His identification of top information sources is similar to our Tier reorganization method, and the results are in the same format: his identification found BBC, Guardian, NYTimes as the top information sources, and we found such traditional new outlets as the top tier clusters. However, he chose the topic words based on the number count while we used a standard deviation method to pick the issue words. Hurst [19] introduced the growth of online personal media (weblogs, message boards, usenet, etc.) has resulted in the emergence of a new type of business and marketing intelligence solutions space. And online data is mined to determine a number of issues important to many corporate functions, including brand monitoring, alerting, competition tracking, sentiment mining, customer care and so on.

Among recent works, Oka et al [20] tried to extract the topics from weblogs, which is quite similar to our objective. They selected terms of interests based on the erratic occurrences over periods and the number of occurrences, which are similar to our issue identification method. However, their selected topics were far from the news issue, intuitively, and they present no more validation or filtering efforts for their topic selection. Our methods of corpus reorganization can be their filtering method.

Another research area in the trend identification is the topic identification. Rather catching the issue words, the topic models, such as LDA [21], are used to find the hot issues. We consider this result is a demonstration of a possibly useful feature (IIM) that support such machine learning algorithms.

C. Other Research on Blogs

There have been numerous research studies on blogs. Blogs are, by their nature, unedited content freely and irregularly generated by a large number of diverse bloggers on the Web. Hence, traditional models of search and retrieval do not work very well on blogs, leading to research on new retrieval models [1]. Users of blog search have different needs [2], and thus, researchers have started to look at ways to design the user interaction for blog search [3]. Another popular and interesting characteristic of blogs is that they contain much more subjective content, and this has attracted attention from the sentiment and opinion mining community [4]. While the contents of blogs are actively researched, authors and readers of blogs are also the focus of another line of research. People are involved in many aspects of blogs, as they write the posts, read the posts, comment on the posts, and link to posts. These different roles that bloggers play can connect them into a community structure, and there are many recent papers on that topic [5]. Because the blogosphere is much more anonymous and diversified compared to the offline world, issues such as trust among bloggers have become important [6]. Lastly, much of the content in weblogs deals with current issues, and hence, much attention has been paid to identifying and predicting topic trends in blogs [7].

V. CONCLUSION

This paper proposes an issue identification method and provides a data analysis to test our ideas. By combining blog source discrimination and a text analysis metric, we identified issue words from a massive blog corpus in a reasonable time frame. The found issue words were checked by looking into word appearance in NYT and examining the issue contexts qualitatively. These tests agree that our found issues are reasonably selected.

In overall, this data analysis revealed two high level implications. First, the issue word identification can be done in well designed metrics. We do not believe that the current IIM is the best metric for issue identification. However, it is a start of building such a metric. We are also interested in building a machine learning model to find an issue, but it should be examined whether nowadays complex machine learning model can run on a massive blog data in a reasonable time frame. Our vision is that well designed metrics will be an input feature for a fast machine learning model. Having said that, we are plan to experiment three models: 1) LDA with a bag of words, 2) Only using IIM, and 3) LDA with a bag of words and IIM. We conjecture that IIM can be a useful feature in the topic trend identification using a machine learning algorithm.

The second implication is the importance of source discrimination. Blogs are, in their nature, decentralized, individual and not consistent. Finding an issue from such media will require selecting which blogs are more influential, trustworthy, and informative. We selected influential blogs based on their indegree scores, but there must be other sophisticated methods, i.e. a new version of page rank, to determine which blogs to examine.

While this paper provides simple metrics and groupings, these simplistic methods achieve our issue identification purpose reasonably well. We expect to build up these models to improve the identification accuracy and the calculation performance.

ACKNOWLEDGMENT

This work was partially supported by Brain Korea 21 Project, the School of Information Technology, KAIST, in 2008.

REFERENCES

[1] Arguello, J., Elsas, J., Callan, J., & Carbonell, J. (2008). Document representation and query expansion models for blog recommendation. Proc. of the 2nd Intl. Conf. on Weblogs and Social Media (ICWSM) .

[2] Mishne, G., & de Rijke, M. (2006). A Study of Blog Search. Proceedings of ECIR '06 .

[3] Hearst, M., Hurst, M., & Dumais, S. (2008). What should blog search look like? Proceedings of Search and Social Media Workshop, CIKM 2008 .

[4] Chesley, P., Vincent, B., Xu, L., & Srihari, R. (2006). Using Verbs and Adjectives to Automatically Classify Blog Sentiment. AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs .

[5] Java, A., Joshi, A., & Finin, T. (2008). Approximating the Community Structure of the Long Tail. Proceedings of the International Conference on Weblogs and Social Media .

[6] Kale, A. (2007). Modeling Trust and Influence on Blogosphere using Link Polarity. Proceedings of International Conference on Weblogs and Social Media (ICWSM) .

[7] Glance, N., Hurst, M., & Tomokiyo, T. (2004). BlogPulse: Automated trend discovery for weblogs. WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation .

[8] Spin3r (2009) http://www.spin3r.com

[9] Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004) Information diffusion through blogspace, Proceedings of WWW2004.

[10] Adamic, L.A. & Huberman, B.A. (2000) Power-Law Distribution of the World Wide Web, Science, vol. 287, no. 5461, Mar. 2000, p. 2115.

[11] Lloyd, L., Kaulgud, P., and Skiena, S. (2006) Newspapers vs. Blogs: Who Gets the Scoop?, AAAI Symp. Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006), Stanford University, March 27-29, 2006

[12] Galitsky, B. and Kovalerchuk, B. (2006) Mining the blogosphere for contributor's sentiments, AAAI Symp. Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006), Stanford University, March 27-29, 2006

[13] Mishne, G. and Glance, N. (2006) Leave a reply: An analysis of weblog comments, in WWW 2006 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2006, at the 15th World Wide Web Conference, Edinburgh, Scotland, May 2006.

[14] Kolari, P., Java, A. and Finin, T. (2006) Characterizing the Splogosphere, in WWW 2006 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2006, at the 15th World Wide Web Conference, Edinburgh, Scotland, May 2006.

[15] Narisawa, K., Yamada, Y., Ikeda, D., and Takeda, M. (2006) Detecting Blog Spams using the Vocabulary Size of All Substrings in Their Copies, in WWW 2006 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2006, at the 15th World Wide Web Conference, Edinburgh, Scotland, May 2006.

[16] Han, S., Ahn, Y., Moon, S., and Jeong, H. (2006) Collaborative Blog Spam Filtering Using Adaptive Percolation Search, in WWW 2006 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2006, at the 15th World Wide Web Conference, Edinburgh, Scotland, May 2006.

[17] Wu, Y. and Tseng, B. L. (2006) Important Weblog Identification and Hot Story Summarization, AAAI Symp. Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006), Stanford University, March 27-29, 2006

[18] Thelwall, T. (2006) Bloggers during the London attacks: Top information sources and topics, in WWW 2006 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2006, at the 15th World Wide Web Conference, Edinburgh, Scotland, May 2006.

[19] Hurst, M. (2006) Temporal Text Mining, AAAI Symp. Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006), Stanford University, March 27-29, 2006

[20] Oka, M., Abe, H., and Kato, K. (2006) Extracting Topics From Weblogs Through Frequency Segments, in WWW 2006 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2006, at the 15th World Wide Web Conference, Edinburgh, Scotland, May 2006

[21] Blei, D. and Lafferty, J. (2006) Dynamic Topic Models, In the Proceeedings of the 23rd International Conference on Machine Learning

[22] Nardi, B.A. et al.(2004) Why we blog. In Communications of the ACM, Vol. 47, No. 12, 2004, 41–46.

[23] Wang, X. and McCallum, A. (2006) Topics over time: a non-Markov continuous-time model of topical trends, Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining table of contents, Philadelphia, PA, USA, pp. 424 - 433