Topic Trend Identification in Blogs

Il-Chul Moon, Young-Min Kim, Hyun-Jong Lee, Alice H. Oh
Department of Computer Science
KAIST

Presenter : Hyun-Jong Lee

Problem Statement

- People read/write hundreds of thousands of blog posts everyday.
- What are they writing and reading about?
- Previous research has not considered the enormous size of the blogosphere.

Challenges

1. 1 day of Blogs : 250,000 posts
   - includes spam blogs
   - includes personal stories that only a few users read
   - Objective : Identify representative blogs that we should look at to find the meaningful topics

2. Topics are not pre-specified, and change over time
   - need a fast way to track topic changes over time
   - need to be able to identify previously unseen topics

Our Approach

- We built a system to look at the real-life scale of the blogosphere to identify meaningful topics
- We organized blogs by their indegree scores. We found that this method identifies the representative blogs.
- We made an issue identification metric that runs fast for a large corpus.
- We tested the new metric on the identified representative blogs in a massive blog dataset from Spinn3r.

Previous Research : Topic Identification in blogs

- Oka et al(2006) extracted the topics from weblogs by selecting terms of interests based on the over periods and the number of occurrences.
  - drawback
    - 42,000 posting

- Wang et al(2006) extracted the topics over time by an LDA(Latent Dirichlet Allocation) model.
  - drawbacks
    - 6,427 (3-paragraph) documents
    - LDA is inherently slow and thus cannot process a large dataset.

Dataset Description (1)

- Blog dataset used at data track, the 3rd ICWSM 2009 (International AAAI Conference on Weblogs and Social Media) provided by Spinn3r
- An attempt to provide a standard (free) blog corpus for the research community
- There is an XML data structure for a single blog post but not a specific data structure for a group of posts with in a single blog site.

Dataset Description (2)

Time span	Aug. 1. 2008 – Sep. 31. 2008 (62 days)
Number of sites	1,071,156
Number of posts	14,970,428
Average Number of words/post	69.2
Number of unique words/post	51.2
Average indegree	175.62
Size	60 GB

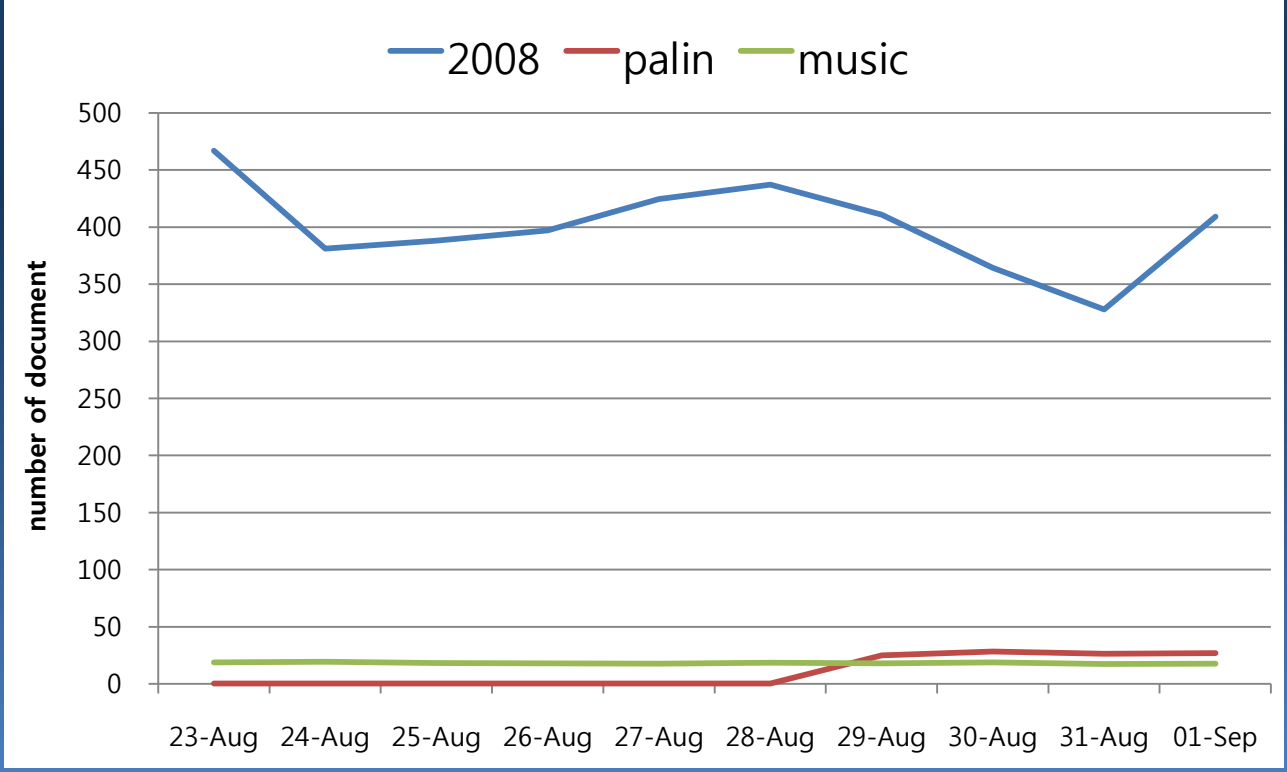
Dataset Description (3)

Tier	1	2	3	4	5	6
# of sites	174,191	26,022	132,218	107,754	40,559	26,141
# of posts	7,642,461	91,523	512,219	275,389	102,700	70,194
Avg indegr	234.97	3.56	0.18	0.02	0.32	1.27

Tier	7	8	9	10	11	12	13
# of sites	24,583	22,058	11,912	11,286	10,452	32,376	102,354
# of posts	70,194	59,636	57,531	57,723	171,570	17,587	2,487,654
Avg indegr	2.09	2.29	100.25	11.48	0.76	1.28	96.59

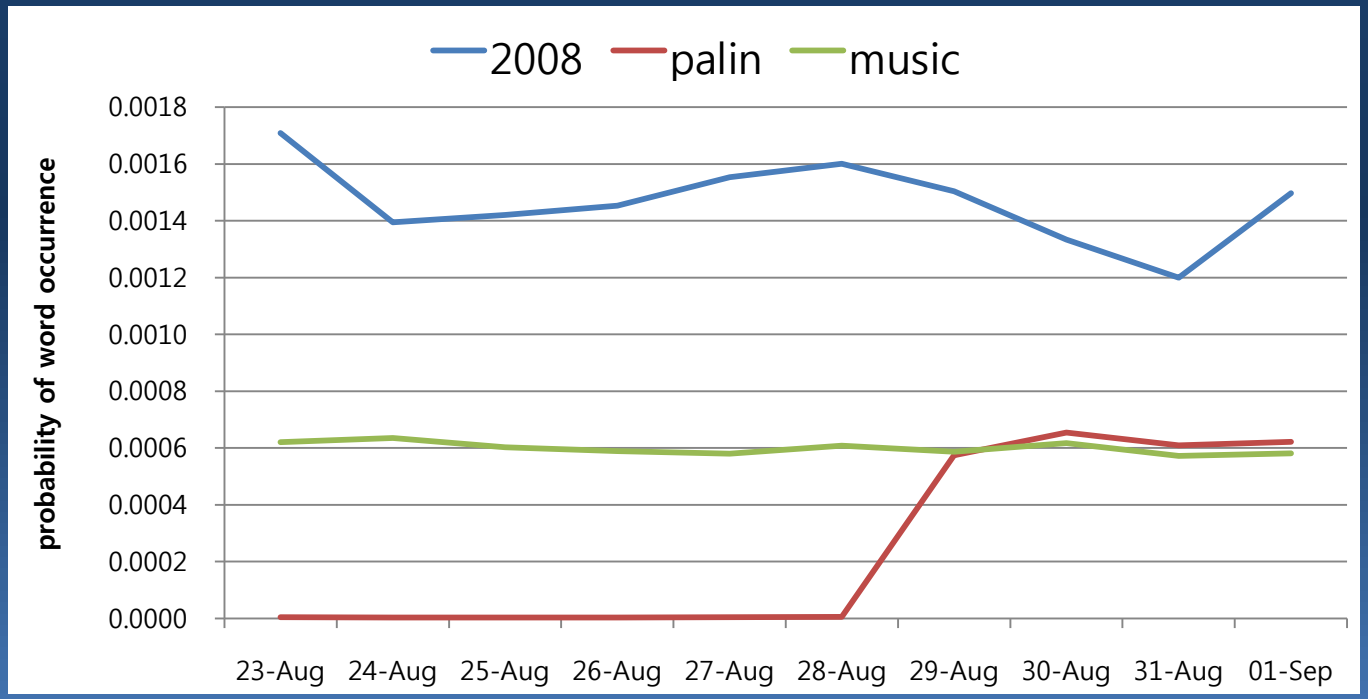
Method: Issue Identification Metric (1)

- Extract English blog posts and calculate the frequency of the unigram by Bernoulli presence-based model
- Y-axis is the number of documents that contain the word for each day



Method: Issue Identification Metric (2)

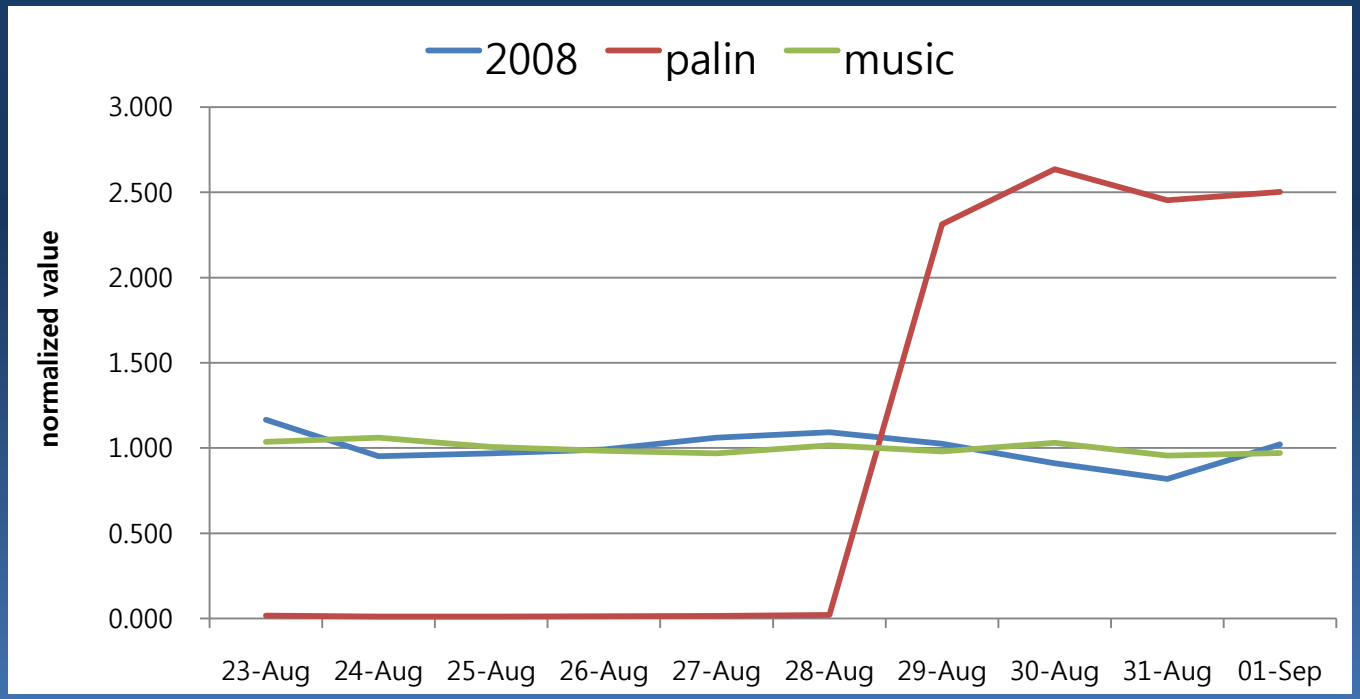
- Y-axis is the probability that the word occurs in a document



$$P_i^j = \frac{Occurrence_i^j}{\sum_{i \in Vocabulary^j} Occurrence_i^j}$$

Method: Issue Identification Metric (3)

- Y-axis shows a normalized value with the average of a word`s probability to be 1



$$NormP_i^j = \frac{P_i^j \times |Period|}{\sum_{j \in Period} P_i^j}$$

Method: Issue Identification Metric (4)

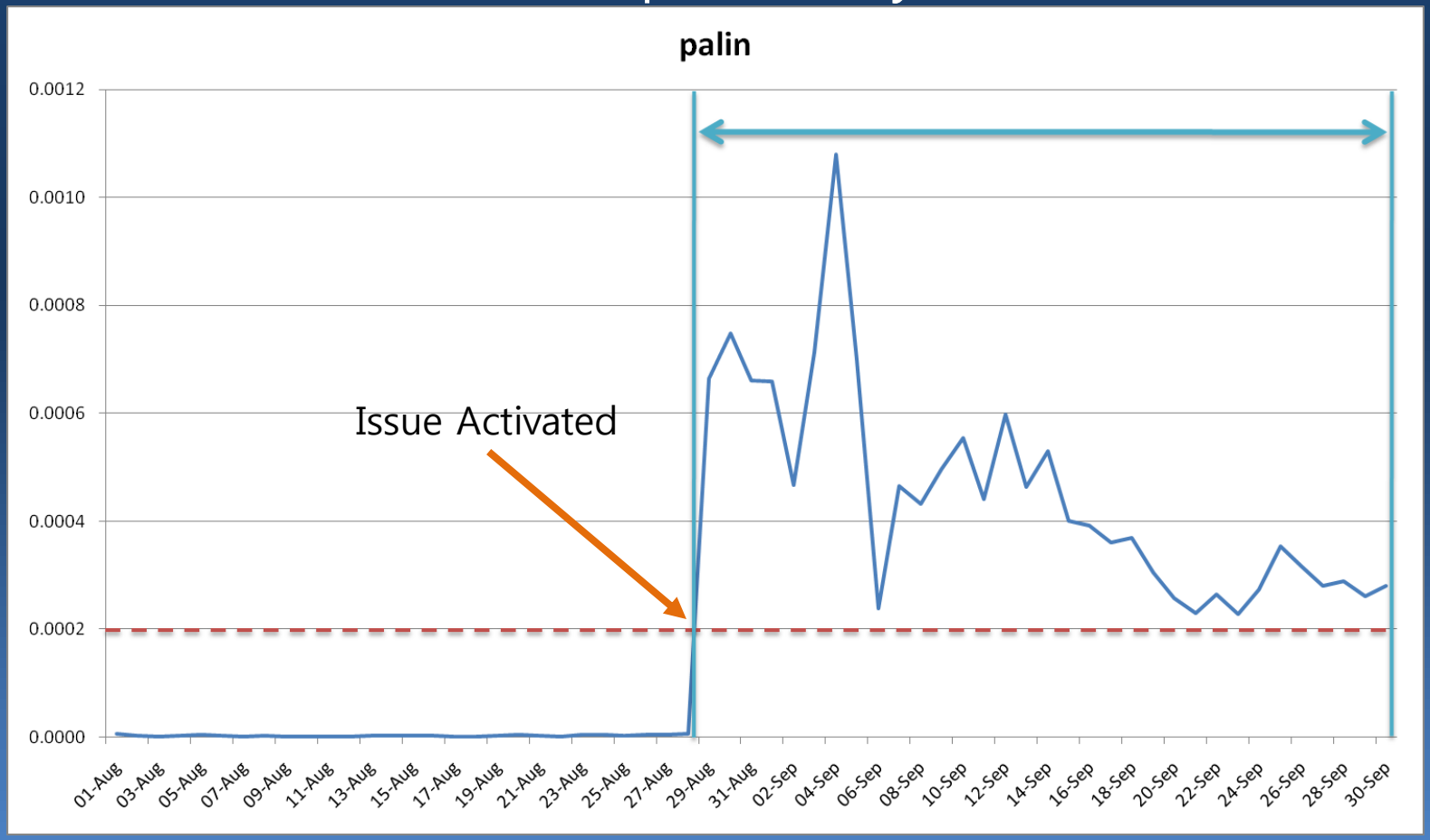
- We then rank the words by standard deviation. This process discovers words that show high fluctuation over time. These are meaningful topic words.

	2008	**palin**	**music**
Average Frequency	390	16	18
STDEV	0.156154	1.088005	0.057355
IIM Ranking	2	1	3
Freq/Fluctuation	High frequency Low fluctuation	**Topic word** High fluctuation Low-medium frequency	Low-medium frequency Low fluctuation

"palin" is the 1st ordered with IIM. Because of it`s fluctuation.

Method: Issue Identification Metric (5)

- Threshold value = max probability x 0.15



Initial Result from Issue Identification Metric

- Top 10 words by IIM

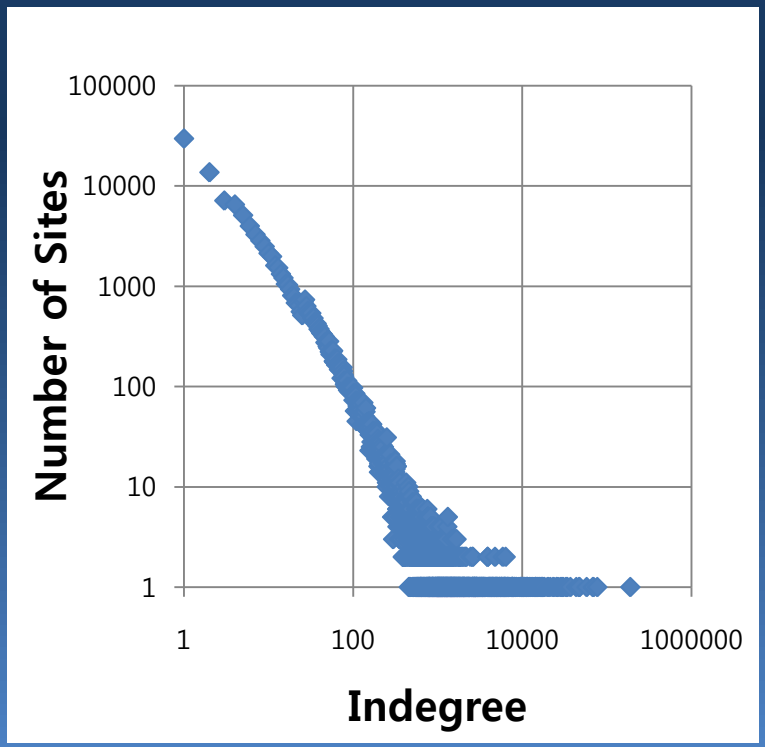
IIM Rank	ALL	TIER1
1	newman	palin
2	gustav	sarah
3	bailout	convention
4	jul	speech
5	aig	financial
6	merrill	vmware
7	$700	billion
8	worm	preferences
9	lehman	republican
10	ike	reserved

Problem with this

- using IIM on 62 days of data for all posts identified good topics
  - But it took 3 days
- Spinn3r has divided the dataset into “Tiers”, with Tier1 being (supposedly) most relevant
  - But Tier1 has non-topic words
- Objective : To re-define tiers such that the newly organized groups enable faster identification of “meaningful” topics.

Method: Grouping by indegree

- Hypothesize : The sites over 10,000 indegree scores might have more representative words to identify temporal issues.



New Group	Indegree Threshold	Number of Sites	Average Indegree
1	> 10000	65	24536.28
2	10000 > Indegree ≥ 1000	773	2499.28
3	1000 > Indegree ≥ 100	8299	260.00
4	100 > Indegree ≥ 10	35194	32.74
5	10 > Indegree ≥ 0	662742	0.33

Result from Issue Identification Metric on Newly Organized Corpus(1)

- New Group 1 has more representative words than Tier 1.

IIM Rank	ALL	TIER1	New Group 1
1	newman	palin	gustav
2	gustav	sarah	lehman
3	bailout	convention	eagles
4	jul	speech	$700
5	aig	financial	brothers
6	merrill	vmware	bailout
7	$700	billion	ike
8	worm	preferences	gulf
9	lehman	republican	medal
10	ike	reserved	denver

Result from Issue Identification Metric on Newly Organized Corpus(2)

- Compare the appearance date at NYT with the activated date from the new group 1
- Newly organized group found topic words that match well with the NYT topics

New Group1 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)

Conclusion

- Our objective is to find topic trend identifications in representative blogs.
- We tested new metric in a massive blog dataset to track topics over time.
- We organized new groups by indegree score to find representative blog sites.
- We found that many topic words from the entire corpus matched with New Group 1.
- We can find topic words from the representative blog sites by our new metric.

Acknowledgements

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

References

- 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.

- X. Wang, A, McCallum. Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends, Proc. 12th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 2006.

Thank you!