A Topic Collection and Context Mining System

Tsun Ku[1], Chen-Ming, Wu[2], Wen-Tai, Hsieh[2], and Gwo-Dong Chen[1+]

1National Central University, Computer Science Information Engineering, Taoyuan, Taiwan

[2]Institute for Information Industry, Taipei, Taiwan

Abstract. With the rapid growth of the social Internet, the task of assisting users with the collection of data on topics of particular interest in an efficient way has become very important. Social networks generate vast streams of text data with very rich content from many different types of source. Efficient organization and summarization of the embedded semantics has also become an important issue. In order to achieve this we have proposed a topic collection and context mining summarization system, we use a topic detection module, in which terms weighting and domain knowledge leverage is employed to isolate the specific topic or event. This paper describes how the major data source is used to efficiently and effectively collect information about specific issues.

Keywords: Topic Detection, Context Mining, Text Data

1. Introduction

The dramatic growth of the social Internet has resulted in the generation of a vast stream of rich text data of many kinds from many different sources. The efficient organization and summarization of data on a specific topic has become an interesting and important issue.

The monitoring functionality of the proposed system provides real-time and on-line access to the service platform. All scalable mass social data coming from a social network, forum, news portal, blogosphere, social account and content are monitored and recorded. A cell phone product is used as an example in the proposed system to demonstrate topic collection. A chart is provided that allows efficient observation of the process.

The rest of this paper is organized as follows. Preliminaries and related works are reviewed in Section 2. The primary functionality and academic theory are covered in Section 3. The monitoring dashboard and operation details are discussed in Section 4. Section 5 is the conclusion.

2. Related Work

Topic modeling has been popularly used for data analysis in several domains that include topic discovery, document classification, citation analysis, and social network analysis. Topic models, such as Probabilistic Latent Semantic Indexing [7] and Latent Dirichlet Allocation [2] have shown impressive empirical success in revealing hidden structure in documents and in related applications like document classification and collaborative filtering. Based on the above models, a set of variants and extensions [1][5][9] have been proposed to further address document modeling problems in different scenarios.

There have been several regularized topic models proposed to incorporate auxiliary knowledge as a constraint in the topic model learning process and to show the resulting benefits. For example, Cai et al. proposed two topic models, Laplacian pLSI [3] and Locally-consistent Topic Modeling [4], which incorporate manifold structure information as a constraint in the PLSI model to smooth the probability density functions. Similarly, Mei et al. [8] regularized the statistical topic model PLSI with an harmonic regularizer based on the structure of a graph of the data. In [6], Guo et al. introduced a weakly supervised topic model, i.e. WS-LDA, by incorporating human labels as a soft constraint into the LDA model to supervise the topic alignment.

Even within communities with very similar interests, there are many different topics of discussion. In order to extract these subjects, cluster-like methods [11][13] are proposed to explore interesting subjects.

+ Corresponding author. Tel.: +0088634227151 ext.35802/35303; fax: +0088634260120.
E-mail address: chen@csie.ncu.edu.tw

Topic-based events may have high impact on articles in the blogosphere. However, it is impossible to view all the topics because of their vast size. By using the technique of topic detection and tracking (TDT) [10] [12], the related stories can be identified within a media stream.

3. Topic Detection Mechanism

In a forum style website, the format used to publish an article is usually fixed. The collected forum data includes unchanged elements such as title, content, reply, author and time. However, here we focus on topic detection and its theoretical practice. The author and time are not relevant in this case. We use the basic element of article as a analysis feature to merge the topic as shown in Figure 1. We calculate the terms intensity and refer to the domain knowledge which is pre-processed by domain experts as shown in formula 1 to find the relevant topics in an accurate and efficiency way.

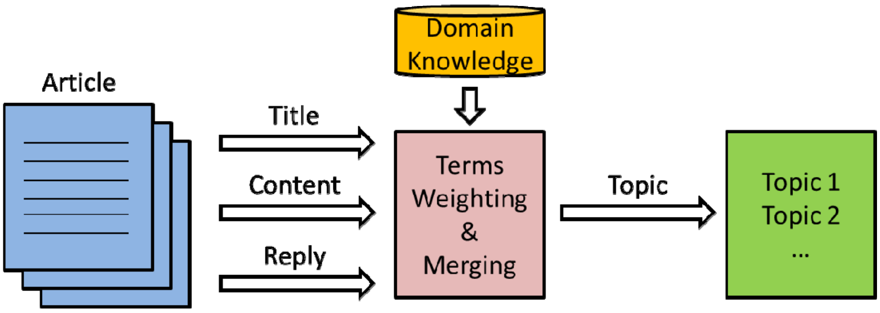


Fig. 1. Bilingual Sentiment Opinion Analysis

The Score (M) decides the topics to be merged and should between 0.7 and 0.8, as shown in Table 1. However, the Score (M) can be adjusted for different cases. In our experiment, the relationships between the terms are highly intensive, so 0.7 to 0.8 shows the best case for aggregation of relevant topics. The Score (Title), Score (Content) and Score (Reply) show the accumulated scores which stand for the terms appearing in the title, content or reply in the article and refer to domain knowledge. The weighting variable α, β and γ are used to reconcile the importance within title, content and reply and the weighting variable sum is 1.

$$Score(M) = \frac{1}{n}\left(\alpha \cdot \sum Score(Title) + \beta \cdot \sum Score(Content) + \gamma \cdot \sum Score(\mathrm{Re}\,ply)\right) \tag{1}$$

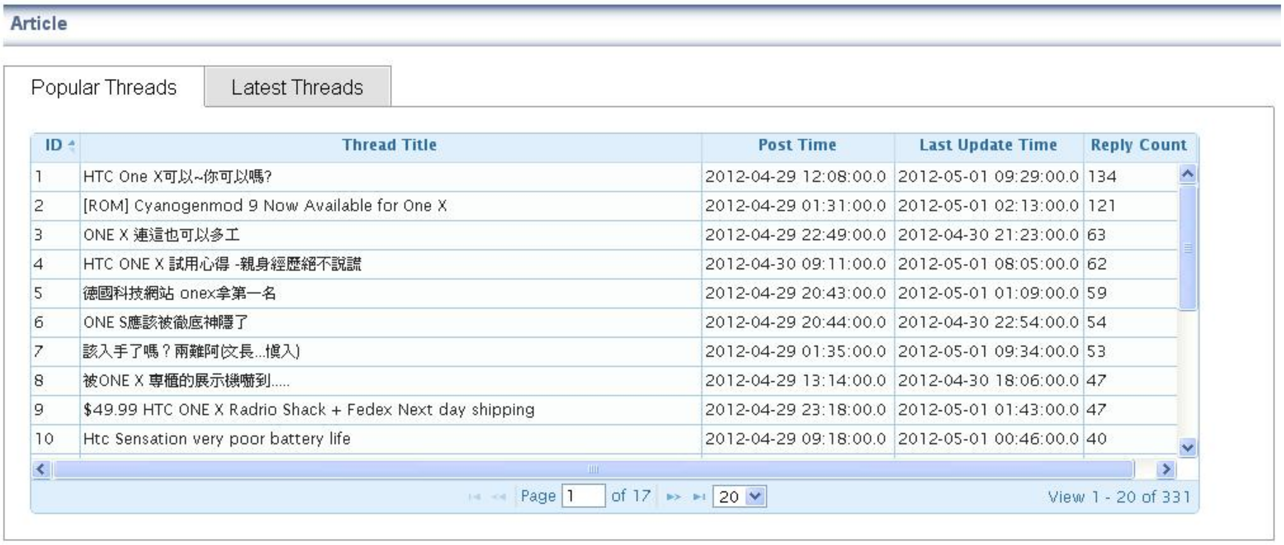
In order to explore the hidden issues in the topics, we use the semiautomatic TF-IDF to generate the importance issue. This semi automation means it is necessary for some keywords to be identified no matter what TF-IDF score they have.

TABLE 1: Experiment of Merged Threshold

Score(M)	Precision	Recall	F-Measure
0.8≦M＜0.9	98.25%	96.79%	97.51%
0.7≦M＜0.8	98.85%	97.52%	98.18%
0.6≦M＜0.7	97.90%	96.55%	97.22%
0.5≦M＜0.6	98.00%	93.60%	95.75%

4. Monitoring Dashboard

Figure 2 is a partial screenshot from the proposed system and shows “Popular Threads” and “Latest Threads“. The popular threads include the high reply frequency threads accumulated within 3 days. Some old threads already have lots of replies and few recent ones. We designed this protocol to avoid this phenomenon. The latest thread uses the same method to present the latest reply within 3 days. As can be seen the topic is cell phones. We collected data from more than 100 popular websites and parsed their common collection into a uniform style and sorted them by reply count. This step makes it easy for the user to see the relevant importance of the entire issue in a single frame.



Article

Popular Threads | Latest Threads

ID	Thread Title	Post Time	Last Update Time	Reply Count
1	HTC One X可以~你可以嗎?	2012-04-29 12:08:00.0	2012-05-01 09:29:00.0	134
2	[ROM] Cyanogenmod 9 Now Available for One X	2012-04-29 01:31:00.0	2012-05-01 02:13:00.0	121
3	ONE X 連這也可以多工	2012-04-29 22:49:00.0	2012-04-30 21:23:00.0	63
4	HTC ONE X 試用心得 -親身經歷絕不說謊	2012-04-30 09:11:00.0	2012-05-01 08:05:00.0	62
5	德國科技網站 onex拿第一名	2012-04-29 20:43:00.0	2012-05-01 01:09:00.0	59
6	ONE S應該被徹底神隱了	2012-04-29 20:44:00.0	2012-04-30 22:54:00.0	54
7	該入手了嗎？兩難阿(文長...慎入)	2012-04-29 01:35:00.0	2012-05-01 09:34:00.0	53
8	被ONE X 專櫃的展示機嚇到.....	2012-04-29 13:14:00.0	2012-04-30 18:06:00.0	47
9	$49.99 HTC ONE X Radrio Shack + Fedex Next day shipping	2012-04-29 23:18:00.0	2012-05-01 01:43:00.0	47
10	Htc Sensation very poor battery life	2012-04-29 09:18:00.0	2012-05-01 00:46:00.0	40

Page 1 of 17 20 View 1 - 20 of 331

Fig. 2. Article collection – Example uses posts about a cell phone product

We list the trends for the specific product which depends on the discussed volume in different professional forums. As can be seen, the rising line shows the product is becoming popular. In Figure 3, the cell phone 'One X' has become more popular than the others. From the news, we know this cell phone is an upcoming market release and has become a topical subject.

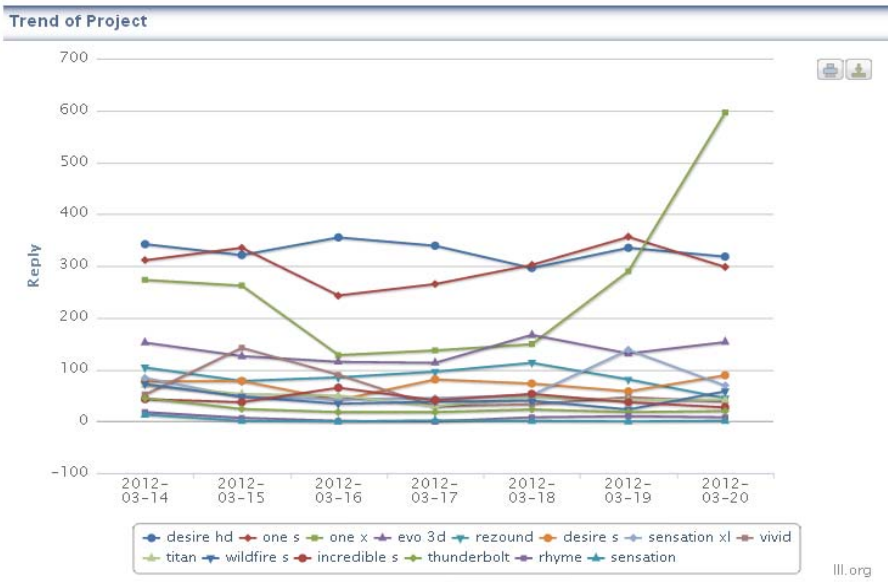


Fig. 3. Trend of Project – Example uses posts about a cell phone product

Beyond the products, some details are discussed with product as a topic and TF-IDF is used with a fixed keyword to extract the important issue. These issues are coordinated with a time slice and generated dynamically. This highlights the most discussed issues about the product. In Figure 4, the X-axis represents the different issue distribution in chronological order and the Y-axis represents the accumulated replies for each day. In this case, ICS (Ice Cream Sandwich, an Android software version) has become the issue of most concern to people.

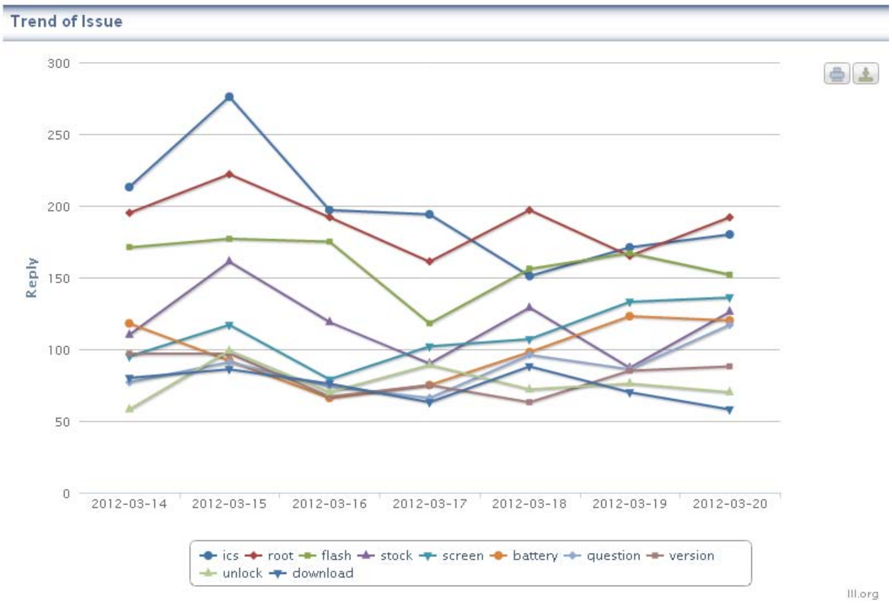


Fig. 4. Trend of Issue – The example uses cell phone issues

The proposed system supports a training mode for different projects to help users with their specific domain knowledge. A user can easily tag an important keyword to a customized category. This is an advantage when using different domain sources and one need not worry about data anomaly. The training facility is automatically enabled if the tagging word arrives within the training standard.

Figure 5 shows the first post of a thread with the analysed information of the entire topic, including the source, category, sentiment analysis and topic thread authors. As can be seen three training modes are provided: Category, Sentiment and Same Problem. In this paper, we only discuss the category training mode. The topic may be classified into different issues, so the system initially provides a classified issue in the category. The system also provides a self-training mechanism to adjust for miss classification.

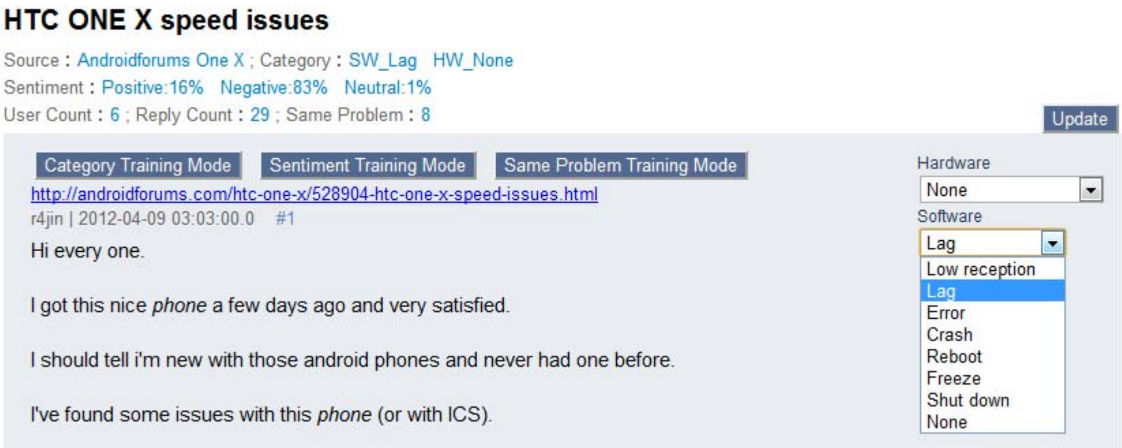


Fig. 5. Self training – Detail issue classification

5. Conclusions

In this high speed, high volume information era with its vast array of different types of information, it is not possible to provide Big Data Analytics for all the small and medium-sized enterprises, government or task-force oriented bodies. Here we include: food-safety related issues, the unexpected appearance of diseases or Government campaign issues, enterprise issues of all kinds and even elections. The proposed system introduces an easy way to access and monitor any topic or issue. Technologically, the system offers the benefits of correct and real-time access and monitoring that can not only be shown at any time but also provides the details of issues within the topics.

6. Acknowledgements

This study is conducted under the "Social Intelligence Analysis Service Platform" project of the Institute for Information Industry which is subsidized by the Ministry of Economy Affairs of the Republic of China.

7. References

[1] D. M. Blei and J. D. Lafferty. Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning, ICML ’06, pages 113–120, New York, NY, USA, 2006. ACM.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. J. Mach. Learn. Res., 3:993–1022, March 2003

[3] D. Cai, Q. Mei, J. Han, and C. Zhai. Modeling hidden topics on document manifold. In Proceeding of the 17th ACM conference on Information and knowledge management, CIKM ’08, pages 911–920, New York, NY, USA, 2008. ACM.

[4] D. Cai, X. Wang, and X. He. Probabilistic dyadic data analysis with local and global consistency. In Proceedings of the 26th Annual International Conference on Machine Learning, ICML ’09, pages 105–112, New York, NY, USA, 2009. ACM.

[5] T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum. Integrating topics and syntax. In In Advances in Neural Information Processing Systems 17, pages 537–544. MIT Press, 2005.

[6] J. Guo, G. Xu, X. Cheng, and H. Li. Named entity recognition in query. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, SIGIR ’09, pages 267–274, New York, NY, USA, 2009. ACM.

[7] T. Hofmann. Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR ’99, pages 50–57, New York, NY, USA, 1999. ACM.

[8] Q. Mei, D. Cai, D. Zhang, and C. Zhai. Topic modeling with network regularization. In Proceeding of the 17th international conference on World Wide Web, WWW ’08, pages 101–110, New York, NY, USA, 2008. ACM.

[9] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004.

[10] I. Varlamis , V. Vassalos, and A. Palaios. Monitoring the Evolution of Interests in the Blogosphere. In Proceedings of the 24th International Conference on Data Engineering Workshops, 2008.

[11] M. Viermetz, and M. Skubacz. Using Topic Discovery to Segment Large Communication Graphs for Social Network Analysis, In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, 2007, pages 95-99.

[12] Wang, C., Zhang, M., Ru, L., and Ma, S. Automatic Online News Topic Ranking Using Media Focus and User Attention based on Aging Theory, In Proceeding of the 17th ACM Conference on Information and Knowledge Management, 2008, pages 1033-1042.

[13] Yoon, S.-H., Shin, J.-H., Kim, S.-W., and Park, S. Extraction of a Latent Blog Community based on Subject, In Proceeding of the 18th ACM Conference on Information and Knowledge Management, 2009, pages 1529-1532.