

Analyzing Trends in the Blogosphere Using Human-Centered Analysis and Visualization Tools

Xavier Llorà
NCSA
University of Illinois at
Urbana-Champaign
Urbana, IL, 61801
xllora@uiuc.edu

Noriko Imafuji Yasui
Dept. of IESE
University of Illinois at
Urbana-Champaign
Urbana, IL, 61801
niyasui@uiuc.edu

David E. Goldberg
Dept. of IESE
University of Illinois at
Urbana-Champaign
Urbana, IL, 61801
deg@uiuc.edu

Abstract

In this paper, we present human-centered visualization and analysis tools that can help users to compare and reason synergies and misalignments revolving around a particular topic. We also present and study that shows the discourse misalignment between Google's and the blogosphere discourses.

Keywords

Human-centered visualizations and analysis, market analysis, chance discovery, discussion analysis, blog mining.

1. Introduction

In this paper we present a human-centered approach for analyzing trends in the blogosphere. We use computer-based tools to screen and filter relevant pieces of information that are later presented to the user for evaluation, reasoning and reflection. The information is mostly presented as visual representations of the key extracted elements. We also present some preliminary results after using the proposed techniques to compare different post sources—Google blogs and Technorati's searches on Google.

2. Human-centered Blogosphere Analysis

A key element to any human-centered analysis is the need to deal with uncertainty [2, 4]. Computer-based methods can help filtering, sorting, and annotating key elements in the current environments under analysis, preparing the ground for humans to reason and extract valuable insights. Our work focuses on analyzing discourses in the blogosphere. Posts provide a rich amount of text that, if properly mined, will allow us to present to the user those key elements that may be relevant for the creation of visual maps. Such maps describe the current analysis endeavor and findings. Our approach consists of three layers: data gathering, data processing, and data visualization. This section briefly describes these components.

2.1 Data collection and storage

Tracking the blogosphere requires to have easy access to its contents. Blogs heavily rely on syndication feeds, usually incarnating in the form of RSS or Atom feed—both based on the XML markup language. The first step is to properly

process blog feeds by retrieving, processing, annotating, and storing the posts in the feeds for later analysis. Our approach stores the processed and annotated posts in a RDF metadata store—Mulgara [5]—waiting to be analyzed. Then, we use the extracted text from the post as the input of three different analysis and visualization techniques. The results of these analysis are also stored in the metadata store to facilitate later retrieval.

2.2 BITS: Blog Induced Topic Synthesis

BITS (blog induced topic synthesis) is a ranking algorithm of sentences and terms used in a blog. Higher ranked terms may be regarded as main topics used in a blog. Similarly, higher ranked sentences express how key concepts are used in the posts. BITS is inspired by HITS (hypertext induced topic search) algorithm proposed by Kleinberg [3]. The idea for the ranking is based on mutually reinforcing relationship between sentences and terms: important sentences include many important terms, and important terms are included by many important sentences. Rankings scores are obtained by an iterative calculation—further details can be found elsewhere [3]. Each iteration updates the score of a sentence by the sum of scores of all the terms in the sentence. Likewise, the score of a term is updated by a sum of scores of all the sentences containing the term.

This simple mutually recursive calculation provides two important outputs: (1) the ranking of relevant terms for a post, and (2) the ranking of relevant sentences. The ranking of terms can be regarded as a summarization of the topics of a given post. On the other hand, we regard the ranking of sentences as an excerpt extraction technique of relevant portions of text and, hence, a summarization tool for posts.

2.3 ISNP: Identifying Self/Non-self Post

Each text of a post can be transformed into a n-dimensional vector of features using text mining techniques. Each feature is a word in the text—once stop words are removed—and the vector represent some sort of frequency measure for each of the feature—TFIDF in our case [7]. This transformation enables the usage of machine learning techniques as tools for exploring and understanding the processed posts. ISNP (Identifying Self/Non-self Post) is an algorithm and visualization technique to create predictive models of the posts. ISNP uses the post-based vectors to learn models that describe and predicts pertinence to a feed. In particular ISNP induce linear models based on support-vector machines [1]. Once the models are learned, we can use them to predict pertinence to a

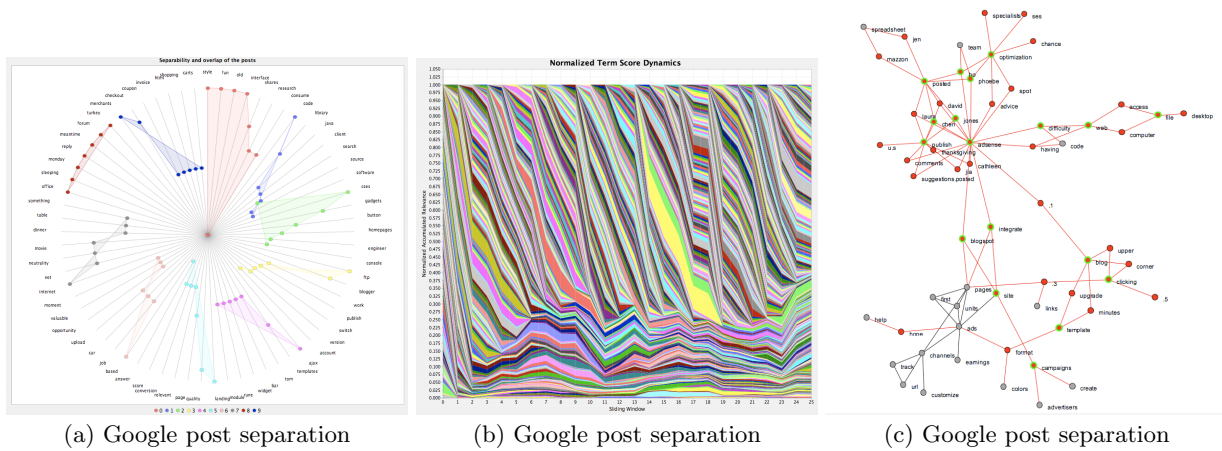


Fig. 1: Results obtained comparing Google Blog discourse (<http://googleblog.blogspot.com/>) from November 10th to November 14th 2006, against Technorati searches for Google during the same time periode. (a) Separation of post, given a post its discriminant terms are plotted in a polar map where magnitude represents the term's discriminant nature. (b) Dynamics of the terms on a given post using a stacked representation of the normalized terms TFIDF values across an sliding window allow an easy visualization of underlying topics and discourse change. (c) Topic map proposed by KeyGraphs presented in three colors: grey to identify high frequency terms and links, red to display key terms and links, and green borders to identify keywords. Results showed a clear misalignment of topics between Google's and bloggers discourses.

feed given a blog, compare multiple feeds to measure degrees of topic overlapping, or simply visualize the key elements that identify self in a post.

ISPN models can be visualized using two different techniques: (1) polar arrangement of the terms that distinguish self and non-self posts and the strength of each of them—see figure 1(a)—, and (2) stacked chart that depicts how topics change across a feed by displaying sliding windows of the TFIDF values across the sequences of blog post—see figure 1(b).

2.4 KeyGraph: Revealing Relevant Connections

When applied to blogs, KeyGraph [6] is a chance discovery technique which provide a visual map of the contents of a blog feed. A KeyGraph is a graph where nodes are terms in the posts and links indicate co-occurrence of terms in a sentence. KeyGraph has been used as tools to support human innovation and creativity in on-line scenarios for market trend detection [4]. KeyGraph computes high-frequency terms and the high-frequency links among them—links are computed inside sentences. Then, relevant low-frequency terms (key terms) and links (key links) are identified. Key terms and key links bridge high frequency clusters together, pointing out interesting transitions between the concepts described by those clusters. Finally, ranking high-frequency and key terms proportionally to the connectivity degree identifies keywords.

KeyGraph visualization depicts concepts and their relations favoring human reflection. Moreover, it provides a simple exploratory method to evaluate bridges between concepts, fundamental building blocks of innovation and creativity. KeyGraphs use three colors: grey to identify high-frequency terms and links, red to display key terms and links, and green borders to identify keywords. An example of the KeyGraph visualization can be found in figure 1(c).

3. Conclusions

In this paper we have presented human-centered visualization and analysis tools that can help users to compare and reason

synergies and misalignments revolving around a particular topic. We have used the proposed techniques to analyze and compare Google blogs posts and the results of Technorati's search on Google. Results showed a clear misalignment of topics between Google's and bloggers discourses.

Acknowledgments

This work was sponsored by the Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant F49620-03-1-0129, and the National Science Foundation under grant IIS-02-09199. The US Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon.

References

- [1] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge Press, 2000.
- [2] D. Goldberg, M. Welge, and X. Llorà. Distributed Innovation and Scalable Collaboration In Uncertain Settings. Technical report, IlliGAL TR No. 2003017. University of Illinois at Urbana-Champaign, 2003.
- [3] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
- [4] X. Llorà, D. Goldberg, Y. Ohsawa, N. Matsumura, Y. Washida, H. Tamura, Y. Masataka, M. Welge, L. Auvil, D. Searsmith, K. Ohnishi, and C.-J. Chao. Innovation and creativity support via chance discovery, genetic algorithms, and data mining. *New Mathematics and Natural Computation*, 2(1):85–100, 2006.
- [5] Mulgara. Mulgara metadata store, 2006. <http://www.mulgara.org/>.
- [6] Y. Ohsawa, N. E. Benson, and M. Yachida. KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor. In *Proceedings of Advances in Digital Libraries*, pages 12–18, 1998.
- [7] S. Weiss, N. Indurkha, T. Zhang, and F. Damerou. *Text Mining: Predictive methods for analyzing unstructured information*. Springer, 2006.