Collaborative Topic Tracking in an Enterprise Environment[⋆]

Conny Franke and Omar Alonso

Department of Computer Science
University of California at Davis
One Shields Ave, Davis, CA 95616
franke@cs.ucdavis.edu, oralonso@ucdavis.edu

Abstract. Business users in an enterprise need to keep track of relevant information available on the Web for strategic decisions like mergers and acquisitions. Traditionally this is done by the user performing standing queries or alert mechanisms based on topics. A much richer tracking can be done by providing a way for users to initiate and share topics in a single place. In this paper we present an alternative model and prototype for tracking topics of interest based on a continuous user collaboration.

1 Introduction

In today's enterprises users need to keep track of relevant content in internal and external sources for strategic reasons. These reasons are usually business decisions about acquiring companies, making sense of new technologies, and exploration of partnerships just to name a few. This means that users must take advantage of the wealth of available information on the Web to track content that is business relevant on a daily basis.

Foraging, scanning, and keeping track of vast amounts of information from a variety of sources is a very time consuming task, so there is clearly a need for automation. There are some tools that can help a single user with some tasks, like Google alerts [1], which are email updates of relevant Google content based on queries or topics. The user has to define certain keywords like in standing queries and gets a summary of the results via email. A more collaborative approach is to use a typical mailing list where people who have the same interest can post and respond accordingly. This is powerful as long as there is enough activity on the list. Wikis have also appeared as an alternative platform for collaboration. So far, all these solutions require active participation from users.

We propose an alternative approach where the user seeds the topics of interests and the system, in a very proactive manner, finds relevant content that is shared within the user's enterprise community. Furthermore, this tracking is an ongoing activity until the user decides that the topic is no longer of interest. Detecting new topics of interest among the vast amount of new information from

[⋆] Part of this work was performed while the authors were affiliated with SAP Research, 3410 Hillview Avenue, Palo Alto, CA 94304.

different sources like news, blogs, etc. that is published daily is a non-trivial task. This is also true for finding interesting themes that are related to topics already popular with an individual or a group of people. Leveraging the collaboration among peers in an enterprise and the increased agility of the group's knowledge that results from the shared interest, is at the center of our approach.

Related Work
The informal network of collaborators and colleagues is one of the most effective channels for dissemination of information and expertise within an organization [4]. In order to extract the most relevant information, concepts of collaborative filtering can be applied to the communication among the members of these networks. Methods like Amazon.com's recommendation engine [5] produce high quality results. However, they only consider content within the same domain as the seeds, i.e., if we applied collaborative filtering on blog posts, the output would only contain related blog posts, not any other kind of articles like news articles in addition to that. The input for traditional topic detection and tracking approaches [2] is a set of keywords. Usually, these keywords are not generated dynamically based on the automated analysis of seed emails, as we do in our method. Recently, the social aspect is having a huge impact in the way people perform information discovery [6] on the Web. Finally, part of this work is based on a sensemaking-based application for technology trends [3].

2 Collaborative Topic Tracking

Our approach aims at getting additional input for accurate topic tracking by utilizing the ongoing discussion about topics of interest within a company. We assume that a group of users has similar interests and goals and thus all members are eventually interested in the same trends. We consider a trend as a general direction expressed in news and blogs that are available on the Web.

In the following sections, we describe how we tap into a discussion to seed the topic tracker, and how we augment the ongoing discussion with results found by the topic tracking system.

2.1 Blok

A pivotal element in our topic tracking system is the "blok", a cross between a blog and a talk/chat client. A blok is similar to a long log session that contains all conversations. The seeds of the blok are individual messages, each consisting of a title or subject, the actual message, a user name to identify the author of the seed, and the time when the seed was created. The idea behind the blok is to enable discussion about enterprise related topics between users with similar interests and augment their discussion with automatic tropic tracking. The blok allows users to see all entries and their associated topics over time. This concept in combination with the variety of filtering options we offer makes it easy to keep track of discussions on specific topics as well as the development of "hot topics" over time.

Users post findings and possibly URLs containing news articles they think are interesting to the blok. Others that read about it can comment on these initial posts and contribute their own facts, links, and findings. To further guide and help the user to explore existing blok posts, we automatically detect the sentiment of each blok post to give the user an indication about whether the contribution is positive, neutral, or negative.

Users can interact with the blok by sending an email. This interface also makes it easy to post to the blok as a byproduct of a discussion on a mailing list or newsgroup. Additionally, we replicate the content that users generate in their enterprise internal blogs in our blok. Users can also seed the system directly by entering a message in the user interface.

2.2 Seeding the Topic Tracker

Given a seed message, we automatically find articles from news and blogs that are related to the message. As new seeds arrive, they undergo a processing pipeline as follows. We use named entity extraction on the blok posts to find names of people and organizations, as these entities represent the seed's content best. Discovered named entities are added to the set of tracked keywords for use in finding related articles. Additionally, we use heuristics on the seed's URLs to determine if they contain worthwhile topic to track, e.g., `http://www.powerset.com`. For the example of `http://www.powerset.com`, "powerset" would be extracted as an additional keyword for seeding the topic tracker.

We augment the content of a blok entry by performing sentiment detection. We do this by first applying a subjectivity analysis to detect neutral blok posts. Then, for all posts that are found to be subjective, we apply a sentiment detector to determine if it is positive or negative. As a last step, we remove all stop words from the seed message and analyze the term frequencies within the message.

2.3 Feedback Loop

After a user posted to the blok, he can explore the additional articles our system found based on his contribution. Ideally, the automatically discovered content contains new and valuable information for the user and he is likely to report his new insights back to the blok or write about them on his private blog. In either case our system picks up his response and includes it in its knowledge base where all other users can pick it up.

3 Prototype Implementation

The main parts of the prototype are the email parser with named entity extractor, URL parser, and sentiment detector, the topic tracker, and the user interface for interaction. Figure 1 shows the processing pipeline. A user's seed message that is input via email or blog post triggers the email parser, which generates a set of topics as input for the automated tracker. Periodically, the topic tracking system is invoked. From a user given set of sources it retrieves articles that contain at least one of the tracked topics and stores it into a database for later exploration by the user.

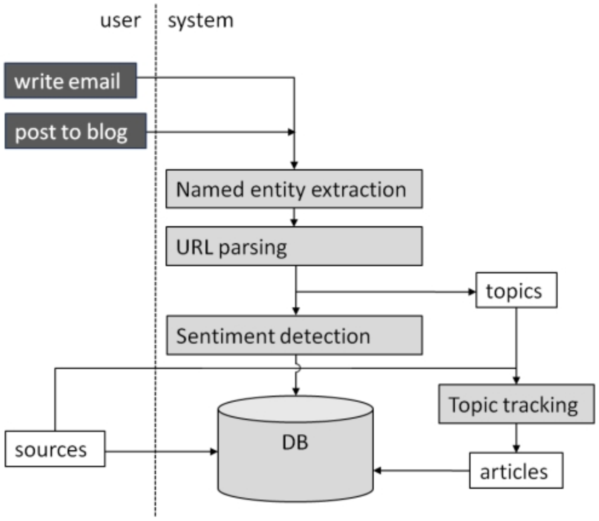
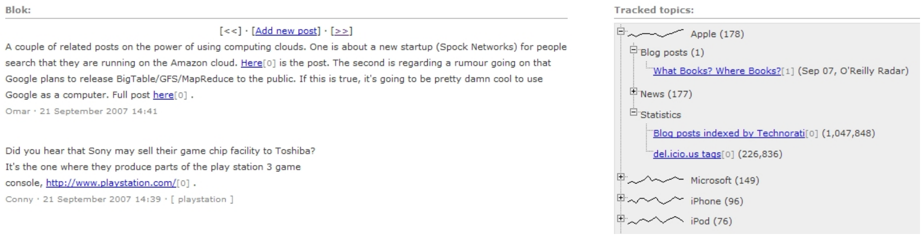


Fig. 1. Text processing pipeline

The actual email parser is a Perl script that extracts the content of the email and stores it in a database. The script also recognizes and parses all URLs contained in the email. After storing the relevant content, the Perl script calls a Java program that conducts a more elaborate analysis. First, we apply heuristics to all URLs in the email to extract new topics from the domain name. Then, we use an open source named entity extractor to get seeds for the topic tracker. The last step in processing each incoming email is sentiment detection. We use a combination of subjectivity analysis and sentiment detection to distinguish between positive, neutral, and negative sentiment in the email.

There is a task scheduler that starts the topic tracking to find new articles. In this step, we extract all articles from a set of news sources and blogs and match them with the set of tracked topics. We then store all relevant articles in the database. In addition, we grab all new posts in enterprise internal blogs and include them in our blok. We then apply named entity extraction and sentiment detection on these blok posts just like we do for incoming emails. Once a week, the topic tracking includes an additional step to collect some statistics about the tracked topics. For example, we look up the number of tags listed on the Web site `del.icio.us` as well as the number of blog posts about this topic indexed on `technorati.com`. These numbers give some additional indication about the popularity of a specific topic over time.



Blok:

[<<] · [Add new post] · [>>]

A couple of related posts on the power of using computing clouds. One is about a new startup (Spock Networks) for people search that they are running on the Amazon cloud. Here[0] is the post. The second is regarding a rumour going on that Google plans to release BigTable/GFS/MapReduce to the public. If this is true, it's going to be pretty damn cool to use Google as a computer. Full post here[0] .

Omar · 21 September 2007 14:41

Did you hear that Sony may sell their game chip facility to Toshiba?
It's the one where they produce parts of the play station 3 game
console, http://www.playstation.com/[0] .

Conny · 21 September 2007 14:39 · [playstation]

Tracked topics:

⊟ Apple (178)
  ⊟ Blog posts (1)
    What Books? Where Books?[1] (Sep 07, O'Reilly Radar)
  ⊞ News (177)
  ⊟ Statistics
    Blog posts indexed by Technorati[0] (1,047,848)
    del.icio.us tags[0] (226,836)
⊞ Microsoft (149)
⊞ iPhone (96)
⊞ iPod (76)

Fig. 2. Collaborative tracking user interface: blok entries and tracked topics.

Figure 2 shows part of the topic tracking user interface. The current view shows a couple of recent blok posts and an overview over the tracked topics. The automatically generated sparkline next to every topic shows its popularity within the last two weeks, as it summarizes the overall number of news and blog articles found each day.

4 Conclusions and Future Work

We have presented an approach for tracking topics of interest in a collaborative fashion. The proposed technique differs from current standing queries solutions by adding the community factor and by providing a feedback loop to the tracking status. An initial prototype that includes some of the ideas presented was developed in an enterprise environment. At time of writing, the system has been in use with a handful of users, who are tracking business information. The initial feedback on using the blok as input has been very positive. This encourages us to continue working on other aspects that were left out due to time constraints. Future work includes evaluation of the accuracy of the information over significant periods of time.

References

1. http://www.google.com/alerts.
2. J. Allan. Introduction to topic detection and tracking. pages 1–16, 2002.
3. O. Alonso, F. James, C. Franke, J. Talbot, S. Xie, and K. Klemba. Sensemaking in the enterprise: A framework for retrieving, presenting, and providing feedback on information sources. *Submitted for SIGMOD Industrial Session*, 2008.
4. H. Kautz, B. Selman, and M. Shah. Referral web: Combining social networks and collaborative filtering. *Commun. ACM*, 40(3):63–65, 1997.
5. G. Linden, B. Smith, and J. York. Amazon.com recommendations: Item-to-item collaborative filtering. *Internet Computing, IEEE*, 7(1):76–80, 2003.
6. R. Ramakrishnan and A. Tomkins. Toward a peopleweb. *IEEE Computer*, 40(8):63–72, Aug. 2007.