Conversation Retrieval from Twitter

Conference Paper · April 2011

DOI: 10.1007/978-3-642-20161-5_93

Matteo Magnani
Uppsala University
96 PUBLICATIONS **970** CITATIONS
SEE PROFILE

Danilo Montesi
University of Bologna
180 PUBLICATIONS **1,146** CITATIONS
SEE PROFILE

Conversation Retrieval from Twitter

Matteo Magnani[1], Danilo Montesi[1], Gabriele Nunziante[1], and Luca Rossi[2,⋆]

[1] Dept. of Computer Science, University of Bologna
{magnanim,montesi,nunziant}@cs.unibo.it
[2] Dept. of Communication Studies, University of Urbino Carlo Bo
luca.rossi@uniurb.it

Abstract. The process of retrieving conversations from social network sites differs from traditional Web information retrieval because it involves human communication aspects, like the degree of interest in the conversation explicitly or implicitly expressed by the interacting people and their influence/popularity. Our demo allows users to include these aspects into the search process. The system allows the retrieval of millions of conversations generated on the popular Twitter social network site, and in particular conversations about trending topics.

Keywords: Conversation retrieval, Twitter.

1 Introduction

Today a growing amount of online information is produced by users (User Generated Content) and published on social network sites. Social motivations laying below this process are various and go from the users' need for information to their desire for social interaction such as *online conversations* or *online dating.*

Even if it is possible to gather many different services under the Web 2.0 label it is important to highlight that each service shows its own peculiarities and deserves to be understood and studied according to those. Twitter, the service we are dealing with in our demo, belongs to the sub-category of microblogging sites: services that allow users to share short text messages (tweets) with a defined group of users called *followers.* Users can reply to each other simply by adding a @ sign in front of the name of the user they are replying to. This fairly simple set of socio-technical rules has made possible for Twitter to host a wide range of social interactions [6] from the *broadcasting* of personal thoughts to a large public to more structured *conversations* among groups of friends. The analysis of those communications can be useful from many points of view (from marketing research to political consensus analysis) but in order to be fruitful it requires to take into careful consideration not only the textual relevance of the searched keywords but also the social relationships existing among the users involved in the conversation.

In our demo we provide a full conversation retrieval system extracting conversations about trending topics from Twitter and providing a query system where

⋆ This work has been partly funded by Telecom Italia.

P. Clough et al. (Eds.): ECIR 2011, LNCS 6611, pp. 780–783, 2011.

users can specify the impact of social and communication aspects on the ranking of conversations.

2 Conversation Retrieval Model

The concept of *conversation retrieval for social network sites.* has been introduced in [10] and builds over previous research on structured [8,9,7,3,2,4], hypertext and Web information retrieval [1,5].

In our demo system a conversation is modeled as a tree where nodes represent short *text messages* posted by a *user* at specific *timestamps* in *reply* to their parent nodes, as exemplified in Figure 1. The ranking of a conversation depends on all these aspects, as follows.

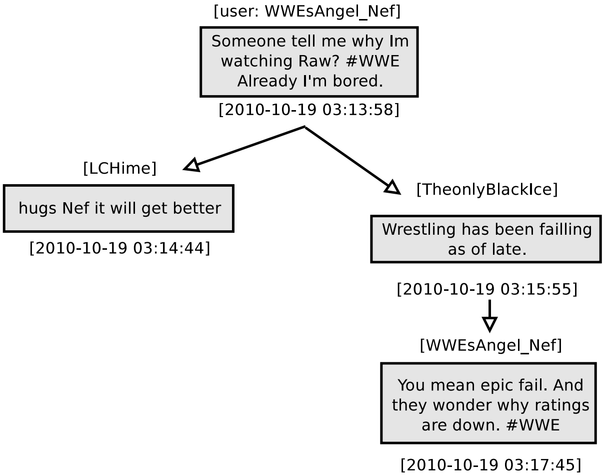
[user: WWEsAngel_Nef]
Someone tell me why Im watching Raw? #WWE Already I'm bored.
[2010-10-19 03:13:58]
[LCHime]
hugs Nef it will get better
[2010-10-19 03:14:44]
[TheonlyBlackIce]
Wrestling has been failling as of late.
[2010-10-19 03:15:55]
[WWEsAngel_Nef]
You mean epic fail. And they wonder why ratings are down. #WWE
[2010-10-19 03:17:45]

Fig. 1. Part of a Twitter conversation composed of 4 tweets

The **text relevance** of a conversation is obviously one of the ranking criteria. However, the same tweets posted by different users may have different degrees of importance — for instance a tweet about a product or brand assumes special importance in a brand monitoring system if it has been posted by a well known blogger. We will thus use a concept of **popularity** of users and conversations. In addition, the same tweet posted at different times may be more or less important — for example, a five-year-old tweet can often be regarded as less important than very recent news. We call this the **timeliness** of a conversation. Moreover, the rate at which tweets are exchanged can be indicative of the level of interest/emotion attached to the conversation — in the following this aspect is indicated as the **density** of a conversation. Finally, the number of tweets exchanged during a conversation (**size**) and the number of participating users (**audience**) can also be regarded as ranking criteria.

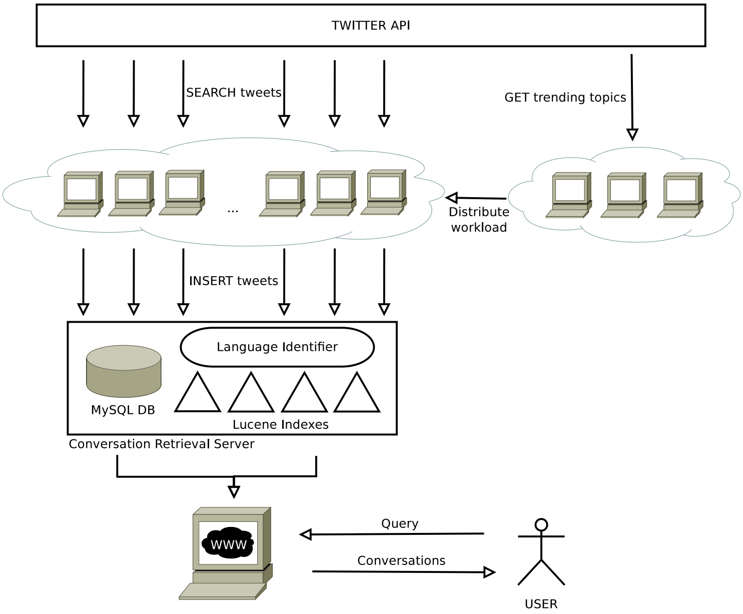

Fig. 2. Architecture of the Conversation Retrieval Demo

In our demo the text relevance is computed using an existing IR engine, and size and audience are computed by analyzing the tweets composing the conversation. As measures of user popularity we consider the number of followers and the ratio between posted tweets and replies received, while we use the number of retweets as a measure of popularity of a conversation, i.e., how many times its tweets have been shared by other users. The density of a conversation is computed as the average inverse time interval between tweets. Finally, the timeliness is computed by comparing the timestamps of the tweets with the date specified in the query.

3 Conversation Retrieval System

In Figure 2 we have illustrated the architecture of our system. A small number of programs periodically get a set of keywords from the Twitter API[1] indicating the trending topics, i.e., the most discussed topics at that instant. These topics are then distributed to several programs retrieving related tweets using the Twitter Search API[2]. Whenever a tweet indicates that it is part of a reply chain, all the chain of tweets is collected.

The extracted tweets and conversations are then uploaded to a conversation retrieval server. This server hosts a relational database management system (MySQL) to store tweets and other information like the number of followers of

[1] `http://dev.twitter.com/doc`

[2] `http://dev.twitter.com/doc/get/search`

every user participating to the conversations. The IR engine Lucene[3] is used to index the text of the conversations and to associate it to their identifiers in the database, from which they can be later efficiently retrieved.

At this point users can ask queries through a web interface where they can specify how much each of the aforementioned social aspects should be weighted in computing the ranking of the result. While we are currently evaluating some predefined query profiles to ease the specification of the search parameters, one important feature of the system is that users can change these values to improve the result of previous research tasks by analyzing the retrieved conversations and updating the weights of the ranking criteria.

4 Demo Session and Concluding Remarks

During the demo session users will be able to query directly the system and tune the ranking parameters to evaluate the impact of different social aspects on the results of their search tasks. One of the innovative features of the system and the underlying theoretical model that will be appreciated during the session is the focus on user social relations in addition to the traditional emphasis on words and document relationships.

References

1. Agosti, M., Smeaton, A.F.: Information retrieval and hypertext. Kluwer Academic, Boston (1996)
2. Amer-Yahia, S., Botev, C., Shanmugasundaram, J.: Texquery: a full-text search extension to XQuery. In: WWW (2004)
3. Amer-Yahia, S., Fernandez, M.F., Srivastava, D., Xu, Y.: Phrase matching in XML. In: Proceedings of the International Conference on Very Large Data Bases (2003)
4. Amer-Yahia, S., Lakshmanan, L.V.S., Pandit, S.: Flexpath: Flexible structure and full-text querying for xml. In: SIGMOD Conference (2004)
5. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: Computer Networks and ISDN Systems, pp. 107–117 (1998)
6. Danah, B., Scott, G., Gilad, L.: Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In: HICSS-43, January 6. IEEE, Kauai (2010)
7. Fuhr, N., Großjohann, K.: XIRQL: A query language for information retrieval in XML documents. In: SIGIR Conference (2001)
8. Fuhr, N., Rölleke, T.: A probabilistic relational algebra for the integration of information retrieval and database systems. ACM Transactions on Information Systems 15(1), 32–66 (1997)
9. Lalmas, M.: Dempster-Shafer's theory of evidence applied to structured documents: modelling uncertainty. In: Proceedings of the 20th Annual International ACM SIGIR Conference, pp. 110–118. ACM Press, New York (1997)
10. Magnani, M., Montesi, D.: Toward conversation retrieval. In: Agosti, M., Esposito, F., Thanos, C. (eds.) IRCDL 2010. Communications in Computer and Information Science, vol. 91, pp. 173–182. Springer, Heidelberg (2010) isbn: 978-3-642-15849-0

[3] `http://lucene.apache.org`