Inf Retrieval (2012) 15:354–372
DOI 10.1007/s10791-012-9189-9

INFORMATION RETRIEVAL FOR SOCIAL MEDIA

Conversation retrieval for microblogging sites

Matteo Magnani · Danilo Montesi · Luca Rossi

Received: 1 April 2011 / Accepted: 23 January 2012 / Published online: 14 February 2012

Abstract In this article we introduce a novel search paradigm for microblogging sites resulting from the intersection of Information Retrieval and Social Network Analysis (SNA). This approach is based on a formal model of on-line conversations and a set of ranking measures including SNA centrality metrics, time-related conversational metrics and other specific features of current microblogging sites. The ranking approach has been compared to other methods and tested on two well known social network sites (Twitter and Friendfeed) showing that the inclusion of SNA metrics in the ranking function and the usage of a model of conversation can improve the results of search tasks.

Keywords Conversation retrieval · Social network sites · Social media · Monitoring · Social network analysis metrics

1 Introduction

Contemporary Web 2.0 is often defined as *social*.[1] The social Web, and by extension the broader category of social media, has changed the relationship between Web authors and

[1] http://en.wikipedia.org/wiki/Web_2.0.

M. Magnani (✉)
Department of Computer Science, Aarhus University, Aabogade 34, 8200 Aarhus-N, Denmark
e-mail: magnanim@cs.au.dk

D. Montesi
Department of Computer Science, University of Bologna, via Mura A. Zamboni 7, 40127 Bologna, Italy
e-mail: montesi@cs.unibo.it

L. Rossi
Department of Social Sciences, University of Urbino Carlo Bo, via Saffi 15, 61029 Urbino, Italy
e-mail: luca.rossi@uniurb.it

Web audiences by giving to the latter the ability to create their own contents, and it is rapidly changing the way in which we perceive the Web and its functionality. By allowing the audiences to became *active*, social Web technologies are changing the Internet from a landscape made of a few *speakers* and many *readers* to a crowded place of *speakers*.

This scenario is offering new opportunities and challenges. A growing number of on-line services is no longer generating any kind of content: these services are just providing their users with the tools to produce their own contents. Blogging platforms like Wordpress or Blogger, microblogging sites like Twitter of FriendFeed and Social Network Sites like Facebook can be basically described as infrastructures to be filled by the users' activities.

While this evolution has been quickly understood by service providers, that are offering more and more services and platforms based on the users' active involvement, the search methods provided by these services are still very similar to traditional Web search engines, accepting keyword queries and sometimes providing in the advanced search options the possibility to express filters on the number of comments/likes and the user's context, e.g., including only contents produced by his/her direct contacts. However it has been found that searching in social media sites and searching in the Web present many differences and thus require different search methods (Teevan et al. 1935).

In this paper we introduce an information retrieval approach for microblogging sites based on the concept of conversation. One of the main features of microblogging and Social Network Sites is that users do not just produce contents but they can get involved in conversations with other users by commenting, liking and sharing other users' posts. Switching the perspective to a conversation-based approach means that within a social web of interactions not only the informative content has to be observed as relevant but also its context. To the best of our knowledge, existing ranking approaches for microblogging sites have so far focused on single microblogs, e.g., single tweets from Twitter. In particular, the contributions of this paper are the following:

- We define a model of on-line conversation.
- We define the properties and the constituents of a ranking function for conversations.
- We evaluate the proposed approach to assess its impact on search results with respect to existing search methods.

The main result of the paper is the evidence that the inclusion of social network analysis (SNA) metrics and the adoption of a conversational model can improve the effectiveness of search tasks in microblogging sites. This confirms previous experimental results regarding the ranking of single tweets (Huang et al. 2010, Song et al. 2010, Teevan et al. 1935) and improves these methods by considering the structure of conversations, as suggested in works about forum thread search (Davis et al. 2007, Liu et al. 2010, Weinberger and Saul 2009, Xing et al. 2002). The objective of this article is to show that the joint usage of multiple conversational and social metrics can improve the search results. Therefore, in our experiments we test different aggregation functions for the SNA parameters defined in the paper and we also use a state of the art classifier to automatically compute an effective combination of these parameters.

The paper is structured as follows: in Sect. 2 we provide a description of related works and of the state of the art in this research field, in Sect. 3 we propose a model for on-line conversations and in Sect. 4 we describe the desired properties and basic components of a ranking function for conversations. Finally, in Sect. 5 we show the result of our experimental evaluation, which consists in two complementary experiments: a quantitative comparison of different ranking approaches on popular (trending) Twitter queries and a user evaluation performed by asking a group of users to compare the outcome of two

search tasks using a standard Web search engine and our socially-extended ranking functions. We conclude the paper with a summary and discussion of our results.

2 Related work

The application of the concept of conversation retrieval to social media data builds over very well established research works and can be seen as an intersection of two main disciplines: Information Retrieval and Social Network Analysis. In this section we review three main corpora of related work: the relevant **IR** literature, the main concepts of **SNA** used in our model and the works specifically addressing **Twitter**, which is the main domain used in our experimental evaluation. In addition to the following references, a preliminary version of the ideas presented in this article has been presented in (Magnani and Montesi 2010) and a Conversation Retrieval system based on our model has been demonstrated at the ECIR conference (Magnani et al. 2011).

2.1 Related work in IR

In the context of IR we consider three main domains that are specifically relevant for the topic of conversation retrieval in microblogging sites. **Structured and Hypertext/Web IR** have considered the problem of evaluating the relevance of textual information in presence respectively of an internal structure, like in XML documents, and of an external structure, e.g., hyperlinks between text documents/web pages. This is relevant because microblogging conversations can be seen as (short) interconnected text documents. Then, the aspect of conversation has been specifically treated in works on **forum/thread search**, where people explicitly reply to previous messages. Finally, ranking a set of conversations requires the usage of several parameters, therefore it is important to mention existing works on the **combination of parameters** in the ranking process.

Researchers in Structured Information Retrieval (Fuhr and Rölleke 1997), with specific reference to structured documents (Amer-Yahia et al. 2003, Amer-Yahia et al. 2004, Amer-Yahia et al. 2004, Fuhr and Großjohann 2001, Lalmas 1997), have considered the problem of retrieving parts of documents, which however have not the same structure of Social Network conversations where there are no overlapping messages *but* connections between them. In addition, conversation ranking is also influenced by the importance/popularity of the authors of the posts, an aspect not considered in this field. Studies in Hypertext Information Retrieval (Agosti and Smeaton 1996) and Web Information Retrieval developed methods to consider connections between text documents, like Google's PageRank (Brin and Page 1998), but these approaches do not include user interaction (e.g., the popularity of the author of a Web page) and do not provide means to compute the aggregate relevance of trees of text messages, like in SNS conversations. It appears however that Google's ranking of Twitter status updates considers the number of followers of the users, even if messages are indexed alone and not inside larger conversations.

The structure of reply-chains or reply-trees has been considered in works on searching forum threads. (Wang et al. 2008) has focused on the problem of identifying the structure of a thread when explicit connections between messages are missing. Although replies to posts in microblogging sites are usually explicit, this work is relevant because by analyzing real microblog conversations it appears that different autonomous conversations may develop inside the same thread of replies. In addition, distinct threads may belong to

related macro-conversations—one example being Twitter hashtags that connect separate threads by common topic. In this paper we focus on the reply structure exposed by the APIs of the microblogging platform. In (Xi et al. 2004) the authors outline the main differences between traditional IR tasks and searching in newsgroups. In addition, although experimentally compared only against their baseline system, they use a combination of measures including author metrics (number of posts, number of replies, etc.) and features of the thread. In our work we apply these ideas to the context of microblogging sites, using additional microblogging-specific features (e.g., conversational density) and SNA metrics (centrality measures). The fact that including thread structure in searching online forums can improve the results of the search tasks is also experimentally verified in (Elsas and Carbonell 2010, Seo et al. 1907–1910). (Smith et al. 2000) extends the discussion on the differences between traditional Web interactions and chats/threads to the design of specialized interfaces, which is not object of our work.

In this paper we show that a simple aggregation of the proposed metrics improves the ranking of microblogs and conversations. However in the literature many approaches have been studied to additionally improve the aggregation of multiple parameters by finding the best weights. This topic is specifically addressed in (Dwork et al. 2001, Taylor et al. 2006)—the first work deals with the aggregation of results from multiple search engines, but also suggests the aggregation of ranking functions as a potential application, while (Davis et al. 2007, Liu et al. 2010, Weinberger and Saul 2009, Xing et al. 2002) deal with the problem of distance metric learning. Computational aspects of the aggregation of multiple rankings are presented in the well known work by Fagin et al. (2003) whose algorithm can be applied to a broad class of aggregation functions including Average, Min and Max, also used in our experimental evaluation where we show that they are not all effective in our context.

2.2 Related work in SNA

Social Network Analysis is a standard area that crosses several disciplines—many textbooks may be found originating from fields like statistics, computing/economics and physics (Easley and Kleinberg 2010, Newman 2010, Wasserman and Faust 1994). The advent of on-line Social Network Sites has introduced new problems and boosted research in this area, in particular on active topics like community detection, network evolution, information propagation in social media and multi-layer networks (Magnani et al. 2011, Magnani and Rossi 2011, Yang et al. 2011).

In this work we include information about the popularity of users into the search paradigm. This information is traditionally known as *centrality* in the Social Network Analysis literature. Centrality is generally recognized as an important attribute to describe both networks and node relevance within a network. Since the pioneering work by Freeman (1979) who defined a set of geodesic centrality measures able to provide a good description of centrality in undirected networks, the concept of centrality proved its empirical validity several times during the years (White and Borgatti 1994). Further works such as those by White and Borgatti (1994) and more recently Opsahl et al. (Opsahl et al. 2010) extended Freeman's work toward more complex kinds of network trying to widen the domains where the intuitively working concept of centrality could have been effectively used. The best known centrality measures for Social Networks are the degree, betweenness, closeness and pagerank centralities, whose definition can be found in any handbook on Social Network Analysis. In the context of on-line Social Network and

microblogging sites, some works have also considered specific algorithms to find leaders and influent users (Agarwal et al. 2008, Goyal et al. 2008, Weng et al. 2010).

2.3 Retrieval of Twitter microblogs

The problem of extracting information from SNSs has already been addressed in the past because of its theoretical and practical relevance. In particular, there is a plethora of tools that can be used to monitor the usage of keywords or tags, e.g., for Twitter. However, these tools work on simple text collections, and do not consider their structure. Other tools like Twitter's built-in trend analysis service or the one presented in (Mathioudakis and Koudas 2010) can be used to identify keywords that are very popular at some specific instant—this approach is complementary to our work, because our aim is not to find keywords based on their frequency but to find conversations based on keywords and social metrics.

In (Teevan et al. 1935) the authors discuss the differences between the Web and microblogging sites during search tasks. This work supports and motivates our model, that introduces a microblog-specific ranking functionality. Other related works focusing on different aspects of microblogging conversations are (Huang et al. 2010, Song et al. 2010), that deal respectively with the tagging of conversations and the identification of topics. (Das et al. 2010, Nagmoti et al. 2010) discuss and experimentally assess alternative methods to evaluate ranking mechanisms, based on the comparison of pairs of tweets (preference judgement). However in our experiments we have not adopted preference judgments because most of the tweets were relevant with respect to the query and it was often difficult to assign meaningful preferences between different tweets.

Another general approach using social information to improve search tasks is commonly known as *social search*. This approach, that has already been studied for some years in the context of microblogging and Social Network Sites, consists in considering the actions of our siblings to find potential resources of interest, following a locality principle assuming that connected people (e.g., friends) tend to have similar preferences (Bao et al. 2007, Evans and Chi 2008). Also this approach, that is strictly connected to Social Recommendation Systems (not reviewed here as not specifically concerning search tasks) and that has been implemented in Google's Social Search, is a complementary method with respect to our work—we do not consider the actions of our friends but the role of the authors of messages in the network and the structure of the discussions themselves. In fact, while we are interested in ranking conversations, social search is more generally used to search for Internet resources like Web pages, goods, etc.

3 Conversation modeling

On a very simple level of abstraction we can assume that an on-line conversation is made of a series of messages exchanged between users using an on-line SNS. Given that level of abstraction conversations happening in SNSs may be described as very similar to many other off-line technological mediated forms of communication. On the contrary, SNS-based interactions have some unique features that must be considered while developing a model for conversations.

According to boyd (danah 2008) content in a SNS space is defined by four properties: *persistence, replicability, searchability* and *addressed to an invisible audience*. These properties generate a set of specific dynamics in SNS interactions and create a specific background for conversational practices.

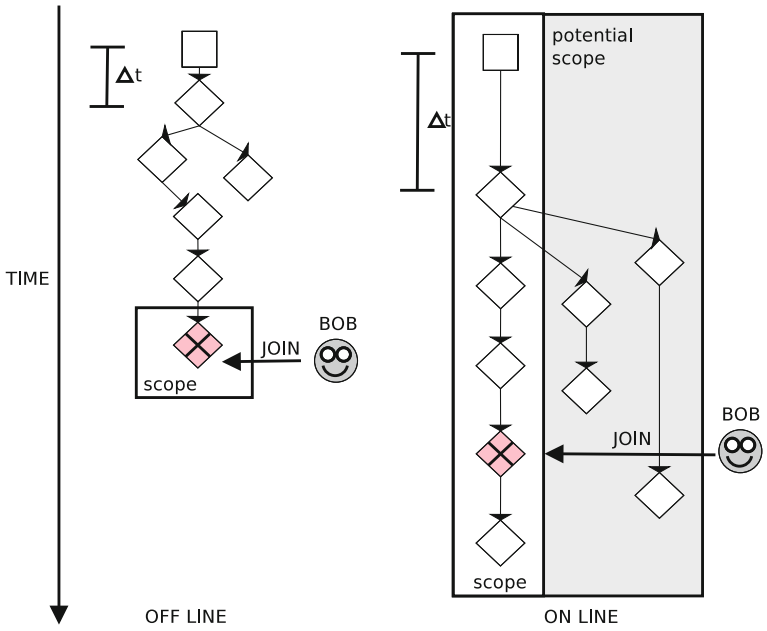

Fig. 1 Main differences between on-line and off-line conversations

With *persistence* we define the characteristics of a message of being available on-line after the first publication for an undefined time. Persistence is strengthen by the *search-ability* of on-line digital contents: every content published on-line is not only permanently accessible but also potentially easier and easier to find thanks to improving searching algorithms and techniques.

In Fig. 1 we point out the main differences between off-line and on-line conversations related to these unique features. Squares represent the first message of a conversation and diamonds follow-up messages.

- Synchrony is not required: the time Δt between the original message and any answer can be **arbitrarily long**. This is increasing the life span of conversations since they will always be available no matter when a user becomes interested in a specific topic: he/she will always be able to search on-line for that topic and even *restart* a conversation which became inactive long time before. In the figure, the time elapsed between the first message and the first reply may be much longer in the on-line case.
- Joining an on-line conversation allows the new user to have a **complete view** of what has been said until that time. In the figure Bob joins the conversation on the message marked with a cross. In the off-line case, he will know only the messages exchanged from that point, while on an on-line conversation his scope will cover all the interactions occurred before his decision to join the conversation.
- On-line persistence of the whole conversation, including the original messages, can bring many users to post messages not directly addressed to the last comment available but referring to any previous message. This may end up in many **concurrent conversations** starting from a single message. This cannot be modeled as a quasi-chain structure with a rigid chronological sequence of interactions where almost every message refers to the previous one, as it usually happens with off-line conversations.

- Finally, the on-line environment enables the collaboration of many more people compared with typical off-line physical conversational environments.

Having these features of on-line conversations in mind we can now provide a formal model describing them. The basic communication step of a conversation involves an actor performing a communicative act in the on-line environment at a precise timestamp. In this work we focus on communicative acts expressed as textual interactions. These actions, that we call *polyadic interactions*, remain as a persistent communicative object that can be later interpreted by a set of other actors to whom the object is available.

Definition 1 *(Polyadic interaction)* Let $\mathcal{U}$ be a set of people, $\mathcal{T}$ a set of timestamps and $\mathcal{M}$ a set of text messages. A polyadic interaction is a tuple (t, u, U, m) where $t \in \mathcal{T}$ is the timestamp of the communicative act, $u \in \mathcal{U}$ is the actor performing it, $U \subseteq \mathcal{U}$ is the set of actors to whom the message is available and $m \in \mathcal{M}$ is the text of the message. If $I = (t, u, U, m)$ is a polyadic interaction, we will notate $\text{ts}(I) = t$ (timestamp of the interaction), $\text{post}(I) = u$ (poster), $\text{read}(I) = U$ (readers), and $\text{msg}(I) = m$ (text message).

It follows that a polyadic *conversation* is a chronological sequence of text messages exchanged between actors where the people involved may change during the conversation. Each message will refer to a previous one, constituting a tree-structure.

Definition 2 *(Polyadic conversation)* A polyadic conversation is a directed graph (V, E) where:

- V is a set of polyadic interactions.
- $E \subset V \times V$.
- (V, E) is a tree.
- $\forall (I, J) \in E \ \text{ts}(I) < \text{ts}(J)$.

In Fig. 2 we have illustrated a set of messages composing a polyadic conversation.

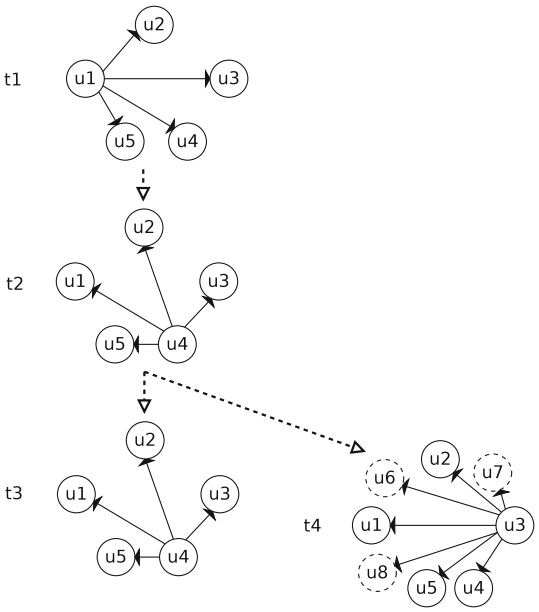

Fig. 2 A graphical representation of a polyadic conversation highlighting the actors involved in each interaction. Notice that during the conversation the set of readers may change

Table 1 Abstract ranking measures for microblogging sites and concrete measures used for our Twitter evaluation

Abstract measure	Concrete measure
Text relevance	Distance (vector space model)
Information quantity	Length of tweet
User popularity	Degree centrality
Message popularity	Number of retweets
Timeliness	Time difference
Density	Sum of inverse time differences

4 Conversation ranking

In the last section we defined a model to represent an on-line conversation. Now we introduce a ranking function that can be used to order the results of a conversation search task. As we have aforementioned, this is an aggregation of other functions representing the relative importance of different aspects of the conversation. It is worth noticing that most of the measures indicated in the following have been defined in other contexts, and their practical usefulness has been proved several times. Here we propose their joint application to the task of ranking microblogs.

The first aspect regards the exchanged text message. To rank text messages we can compute their **relevance** with regard to some information requirements, e.g., using a vector space model and a list of keywords. However, the same sentence pronounced by two different people will have different degrees of importance—a message from a Prime Minister will probably be more important than a message from one of the authors of this paper, at least in some contexts. In addition the identity of the speaker may be much more important than what he is saying, making his social interactions very popular even when he is not saying anything meaningful. We will thus use a concept of **popularity of the posters** depending again on some contextual information, and we can similarly define a concept of **message popularity**. Finally, the same people may exchange the same message, but at different times this may be more or less important—for example, a five-year-old message can be less important than very recent news, and the rate at which messages are exchanged can be indicative of the level of interest/emotion attached to the conversation. Therefore we will also use **time-related** measures.

Given a microblogging site, for each kind of measure we can define a specific way to compute it. In Table 1 we have indicated the concrete measures used for the Twitter platform, that are defined in the following sections and will be used in our experimental evaluation.

4.1 Text-centric measures

Text relevance of single tweets can be evaluated using any IR model, and to evaluate the relevance of an entire conversation we can calculate the average relevance of its interactions. Many standard models such as the boolean, vector-space or more complex models can be used, but this is a traditional topic in IR for which we do not present details here—in our implementation we use the Lucene library with its built-in ranking functions[2]

[2] http://lucene.apache.org.

Definition 3 *(Text conversation relevance)* Let (V, E) be a conversation and rel(m) the relevance of message m. We define the text relevance of conversation (V, E) as $\frac{\sum_{I \in V} \text{rel}(\text{msg}(I))}{|V|}$.

To compare our method to existing proposals to rank single tweets we will also use a measure proposed in (Nagmoti et al. 2010) to capture the amount of information provided by a microblog. In the case of Twitter, where the length of messages cannot exceed 140 characters, the authors use the number of characters to estimate the amount of available information:

Definition 4 *(Information quantity)* Let m be a message. We define the information quantity of m as $\frac{|m|}{140}$ where $|m|$ is the length of the message.

4.2 User-centric measures

The popularity of a user can be defined in several different ways and inside a SNS we can simply compute its degree centrality, i.e., a function of the number of its followers. Other centrality metrics are more related with the *role* of the user than its popularity, e.g., closeness and betweenness, and more complex degree-based metrics like page-rank are more complex to compute and usually provide significantly different user rankings only for users with a small degree centrality or for a few users with a specific position in the network. Also in this case we can define an aggregated popularity measure for whole conversations.

Definition 5 *(User Popularity)* We define the popularity of a user as its in-degree centrality, i.e., the number of its followers

Definition 6 *(User conversation popularity)* Let (V, E) be a conversation and pop(u) the popularity of user u. We define the user popularity of (V, E) as $\frac{\sum_{I \in V} \text{pop}(\text{post}(I))}{|V|}$.

4.3 Message-centric measures

In the same way as we can use the popularity of the authors of a conversation to evaluate its rank, we can also consider the popularity of the posted messages, that can be different from the popularity of their authors. This can be usually computed easily in Social Network Sites, e.g., counting the number of likes, sharings or re-tweets received by the message. Also in this case we can define an aggregated popularity measure for whole conversations.

Definition 7 *(Message popularity)* We define the popularity of a message as its number of re-tweets.

Definition 8 *(Message conversation popularity)* Let (V, E) be a conversation and pop(m) the popularity of message m. We define the popularity of (V, E) as $\frac{\sum_{I \in V} \text{pop}(\text{msg}(I))}{|V|}$.

4.4 Time-centric measures

The **timeliness** of an interaction can be defined in many ways, e.g., returning a result inversionally proportional to the difference between an input timestamp and an internal timestamp of the conversation (starting, medium or ending). In the case of Twitter, timeliness is the main measure used by the standard query interface and the reference time

corresponds to the time of the request. Therefore, messages are returned in inverse chronological order, from the most recent one. We will include also this approach in our experimental evaluation.

More interestingly, time-related measures can be used to associate other attributes to a conversation. If SNSs do not give us a simple way to understand the loudness of a message and other fundamental aspects like non-verbal signals (Watzlawick et al. 1967), conversational density may tell us something more than a single message can. As an example, consider a passionate political discussion: some people will not wait their turn to speak, and the more the conversation will touch sensitive topics the more people will increase the frequency of their interactions starting speaking together.

From this discussion, it seems important to be able to model a concept of **density** of a conversation.

Definition 9 *(Density)* Let (V, E) be a conversation. We define its density as $\sum_{(I,J)\in E} \frac{1}{\mathrm{ts}(J)-\mathrm{ts}(I)}$.

4.5 Aggregation of basic ranking functions

In our experimental analysis we test different aggregation functions for these metrics: Average, Min and Max, all *monotone* functions in the sense of (Fagin et al. 2003) making them easy to be implemented also with very large databases, and we will see that this combination has the effect of improving the relevance of the results of microblogging search tasks. Many works have also dealt with the optimization of parameter aggregation, to find the best weights for each parameter. As the objective of this paper is to show the improvement of ranking tasks using SNA and conversational models, which is already obtained using basic aggregation functions, the application of more complex weight learning methods lies outside the scope of our contribution. However, in the next section we also show the result of the application of a state of the art classifier used to learn the correlation of the aforementioned ranking parameters with the relevance of the conversations.

We conclude this section with an example of the metrics used in this paper. Consider the following sequence of two posts:

1. At **2011-08-09 23:58:20** user **Enrico** with **53 followers** writes: `Still feels real`
2. At **2011-08-09 23:59:24** user **Luca** replies: `@EnriFatigati bro sneijder isnt at manU yet, there hasnt even been an offer yet`

The *information quantity* of the first message is 16 (the number of characters). The popularity of the first user is 53, and the popularity of the message would be determined by the number of times it has been retweeted (which can be extracted using the Twitter API). The second message was posted 64 s after, therefore the density of the conversation would be $\frac{1}{64}$, and its timeliness with respect to an input timestamp **2011-08-09 24:00:00** would be $\frac{1}{36}$, where 36 is the number of seconds between the end of the conversation and the input timestamp.

5 Experimental analysis

The objective of this experimental analysis is to verify the following hypotheses:

- Using a conversational model and SNA measures improves the relevance of search tasks in microblogging sites.

- User satisfaction increases when social aspects are included in the search task, i.e., there is a visible and measurable impact on end users.

These two hypotheses are addressed by two separate evaluations. The first involves several queries and methods to compare different approaches and different combination functions. The second consists in an on line user evaluation, where users have been asked to compare the result of queries performed using a traditional search engine and using our social metrics.

5.1 Quantitative analysis - Twitter

The objective of this experiment is to compare existing methods for microblog ranking with our approach, consisting in using SNA and conversational metrics. The approaches used in the experiment are the following:

Auth. Applied in (Nagmoti et al. 2010), consists in using the degree centrality of the users to evaluate single tweets.

Chrono Default in the Twitter platform, it consists in retrieving the most recent tweets (i.e., the one with higher timeliness).

Length Applied in (Nagmoti et al. 2010), uses the length of the tweets to score them (information quantity).

Avg Used to score conversations, takes the average of text relevance, user popoularity, message popularity and density.

Max Used to score conversations, takes the max of text relevance, user popoularity, message popularity and density.

Min Used to score conversations, takes the min of text relevance, user popoularity, message popularity and density.

Density Used to score conversations, ranks them according to their density.

Then, after having manually annotated the conversations extracted by these approaches, we used them to learn some classifiers and to extract additional conversations, as described in the following.

5.1.1 Experimental setting

The dataset has been obtained by retrieving the tweets and conversations containing trending topic keywords for 24 h (this has been done using the conversation retrieval system presented in (Magnani et al. 2011)). Italian trending topics have been determined directly by Twitter, and we have selected the ten most frequent ones during the monitoring period. The ten topics (and the corresponding queries used in the experiments) were: [Q1: eto], [Q2: vota], [Q3: berlusconi], [Q4: what makes you beautiful, Q5: wmyb, Q6: one direction], [Q7: derek], [Q8: italia-spagna, Q9: cassano, Q10: bari] —we have grouped together trending topics related to the same event.

From these datasets we have then removed the tweets not belonging to conversations, i.e., outside a reply tree, and re-tweets. The first cleaning operation was necessary because we need to compare methods retrieving single tweets against methods retrieving conversations—for single tweets our conversation-based methods reduce to existing approaches, therefore no valuable comparison is necessary or possible. The second cleaning operation is used to remove duplicate information—for example, if a single long tweet (140 characters) is retweeted 9 time the *Length* approach proposed in (Nagmoti et al. 2010) would

Table 2 Top-10 precision of the tested methods for the top-10 trending queries, with average performance and standard deviation

Method	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Avg	StDev
Auth.	.83	**1.0**	.60	.80	.85	.85	.60	**.75**	.75	.95	.80	.13
Chrono	.62	**1.0**	.65	**1.0**	.75	.70	.75	.60	.75	**1.0**	.78	.16
Lenght	**.93**	.80	.65	.55	.60	.75	.72	.70	.75	.85	.73	.11
Avg	.80	**1.0**	**.67**	.90	**1.0**	.90	.80	**.75**	**.85**	.90	.86	.10
Max	.22	.90	.61	.80	.65	.75	.60	.65	.75	.94	.69	.20
Min	.55	.90	.44	.85	.95	**1.0**	**.89**	.70	.80	.95	.80	.18
Dens.	.90	.80	.65	.90	.95	.80	.75	.70	.80	.80	.81	.09
SVM	.63	.50	.50	.80	.75	.70	.80	.60	.70	.80	.68	.12

Bold numbers indicate the best performance for each query

risk to return 10 times the same piece of information in the top-10 results. Notice that this is a standard operation in Twitter search tasks—also the standard Twitter search site has an option to remove re-tweets.

As a result of these cleaning operations we ended up with ten datasets, one for each trending topic, with a total of about 20 000 tweets. At this point we applied each tested method to every dataset to get the top-10 results, and for all the 70 result sets we manually evaluated the relevance of the tweets or conversations. In general most of the tweets were relevant, therefore we did not apply the evaluation methods proposed in (Nagmoti et al. 2010, Das et al. 2010) and based on the comparison of pairs of tweets, but we assessed directly the relevance of every tweet—marking in some cases some tweets as half-relevant when they were consistent with the query but without a clear informative content[3].

In Table 2 we have represented the top-10 precision for each query and the average performance and standard deviation for all methods has been represented in Fig. 3. From these results we can derive many interesting considerations. First, the method that averages the contribution of SNA metrics and conversational metrics has the best average performance (and it also provides the best results in half of the evaluated queries). This is partly due to the fact that structured conversations tend to develop the news and therefore to provide additional information. However, it is worth noticing that the *Dens* method which uses a conversational model without SNA metrics does not reach the same levels of precision. In addition, the next experiment will highlight how long conversations may become counterproductive.

The second consideration is that no method performs constantly better than the others. For example the *Chrono* approach may find very relevant results in some cases, but its performances are not very good in general and they vary significantly. On the contrary the *Average* approach is very stable (stability is indicated by a small standard deviation).

Finally, we can notice how the *Max* and *Min* approaches are not very stable, as they depend too much on the presence of one single high or small value for one of the metrics without considering the indications given by the other parameters.

To conclude the quantitative experimental evaluation we have also tried to further improve the precision of the search results by using the annotated tweets extracted by the seven approaches. For every query (Q1–Q10) we have used the other nine annotated query results to train a support vector machine designed to learn ranking functions, using all the

[3] In a limited number of cases we could not understand the meaning of the tweet, that was not considered in the evaluation

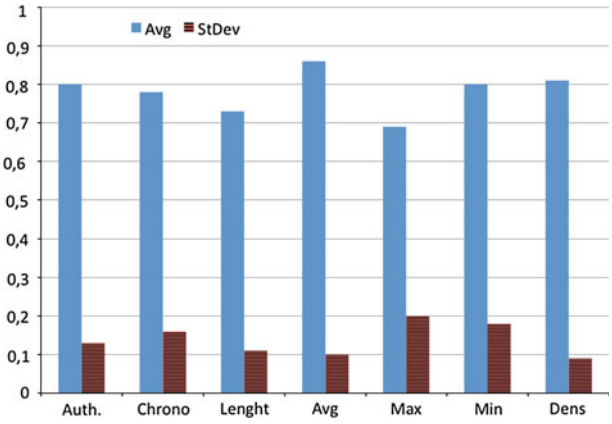

Fig. 3 Mean and standard deviation for the evaluated methods

previous ranking parameters (length, number of retweets, user popularity, conversation popularity, density, text relevance)[4] (Joachims 2002). For each of the ten queries we have manually annotated the relevance of the retrieved tweets. The results of these search tasks are indicated in the last row of Table 2. As it appears from these values, the approach based on the average of the social network analysis metrics still performs better than the others.

The previous experiment supports the hypothesis that a combination of a conversational model and SNA metrics can improve the results of search tasks in microblogging sites. However, another important research questions regards the impact of this functionality on end-users. The following experiment has been designed to gather some knowledge on this aspect, assessing the perceived difference between a traditional search engine used to browse a microblogging site and our approach.

5.2 Qualitative analysis: friendfeed

For this experiment we have selected two events and queried a Social Network Site using Google and our approach (with two different configurations). Then we have asked a set of users to rate the top-10 results of every search task, to compare these approaches. In the following we first describe the experimental setting, then we present the results and finally we provide an interpretation of the data.

5.2.1 Experimental settings

The analysis presented in this section is based on a real social database extracted by monitoring the FriendFeed SNS. In particular, we used a sample of about 3.5 million posts collected by monitoring its public feed on August/September 2010. The complete database of all posts extracted during this monitoring period can be downloaded from the project website[5]. The choice of this SNS, a microblogging service created in 2007 and acquired by Facebook in 2009, is justified by several features making it and ideal case to provide a general analysis: it provides a public API (most of the data are accessible), it is small with respect to other services and aggregates contents from several other SNSs, e.g., Facebook,

[4] http://www.cs.cornell.edu/people/tj/svm_light/, `svm_learn` with option `-z p`

[5] http://larica.uniurb.it/sigsna

Twitter, YouTube, and Flickr. In addition it presents all the messages composing a single conversation into a separate Web page. In this way conversations can be searched using existing Web search engines.

To evaluate the results of our search tasks we have used a set of 30 people who have been asked to score different selections of posts from 1 to 5. These people have been selected among students and colleagues of the authors (with backgrounds in computing and social sciences) and among the users of the FriendFeed service (about half of the evaluators), on a voluntary base, and no user was aware of the details of the underlying systems. Notice that this sample size is comparable with the one used in recent similar experiments (Evans et al. 2010, Morris et al. 2010) and that we do not provide here the details about the background, age and gender of the participants because our aim is not to study the behavior of specific populations from a statistical point of view.

Every user was informed of two events happened during the sampling period: a global event (the mining accident in San José, Chile) and an event of national relevance (the death of a former Italian President). For each event we performed three searches:

1. One using Google.
2. One using our approach set to give priority to user popularities.
3. One using our approach set to give priority to conversation densities.

The evaluators were not aware of which search produced which result, and they were not aware of which systems had been used.

The keywords used for the two tasks were (MINERS OR MINEROS) AND CHILE and COSSIGA (the surname of the President) and the weight assigned to popularity (search 2) and density (search 3) was three times the weight assigned to the other metrics, e.g., text relevance. While this weight may seem arbitrary, it is worth noticing that here our objective was not to determine the best weight for each parameter or to identify the best measure or combination of measures (the latter was object of the previous experimental evaluation): the only objective was to present to the users result sets where different social aspects were emphasized.

At this point every user for each search task was presented with three selections of conversations, one for each of the previous options with the corresponding top-10 results. In Fig. 4 we have represented part of one of these selections. Users where not aware of which system or method had produced the selections, and in the two tasks the order of the selections was changed, e.g., for the first task the first selection corresponded to the results of the Google search while for the second task the results obtained using Google were the second to be presented to the users. In addition to the numerical evaluation users' were free to describe their impressions qualitatively.

5.2.2 Experimental outcomes

In the first row of Table 3 we have represented the average score of the three approaches with regard to the global event. Here we can find that the best performing approach has been indicated as our method with a high weight associated to the density of conversations, the second has been indicated as our method with a high weight associated to the popularity of conversations and the worst scores were given to the result of the Google search. The score distribution of this first task is illustrated in Fig. 5.

The second row of Table 3 indicates the average score of the three approaches with regard to the national (local) event. Here the best scores have been assigned to Google,

"Thirty-three miners trapped 2,300 feet (701 meters) below ground in Chile are depending on food, medicine and supplies being dropped to them through a 4-inch-wide tube. What comes out of that tube will have to sustain the men both physically and mentally for a long time -- perhaps four months, experts say -- while a shaft wide enough to pull a man through is drilled. The miners already have been trapped for 18 days, since a rockslide inside the San Esteban gold and copper mine cut off their exit route. A probe retrieved a note from the miners Sunday saying all were alive and well in a cramped, 530-square-foot (50-square-meter) shelter. They survived by sharing tiny portions of canned fish stored in the shelter room." - SteVe C da Bookmarklet

Incredible. That they survived, and that it might take 4 months to save them. - SteVe C

CNN Reporters
In >2hrs on #WR leading w/ Chile miners trapped under 700metres of unstable rock & not being told their rescue will take 4mths #news #cnn

Fig. 4 Part of a selection of posts and conversations related to the mining accident, extracted using our approach and presented to the users to assess their interest in the selection

Table 3 Average scores for the two tasks and three systems

Bold numbers indicate the best performance for each task

	Google	Popularity	Density
TASK 1	2.57	3.1	**3.43**
TASK 2	**2.86**	2.38	2.03

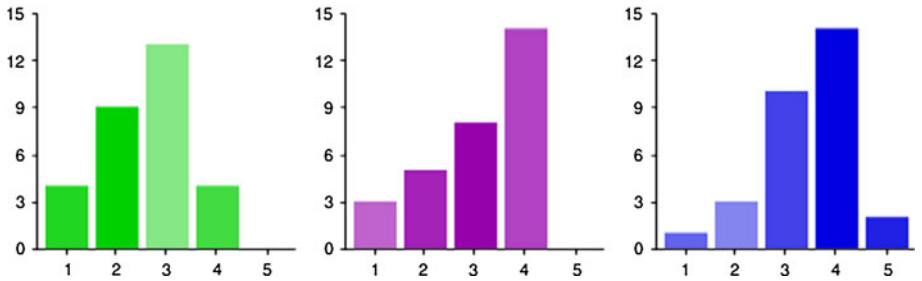

Fig. 5 Chilean mining accident: user evaluation of Google search, Conversation Retrieval with high popularity and Conversation Retrieval with high density (x: score, y: number of votes)

followed by the popularity-based approach followed by the density-based approach. The score distribution of this second task is illustrated in Fig. 6.

Finally, in Table 4 we have indicated the results computed considering only the frequent users of FriendFeed, the platform from which we have extracted the data. These users corresponded to about half of the evaluators. In this case the scores related to the first task (global event) do not change significantly, while in the second task (local event) the approach with the higher scores is the one based on popularity, with the other two scoring the same.

5.2.3 Interpretation of the experimental results

As it can be observed by comparing the histograms in Fig. 5 the inclusion of social metrics may have a significant impact on the users' evaluation of the results of a search task.

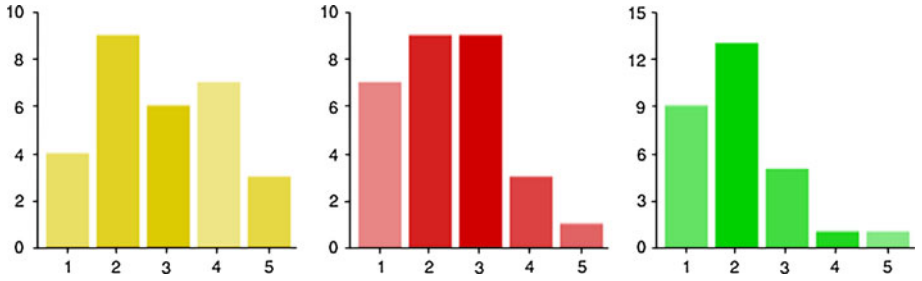

Fig. 6 Italian President death: user evaluation of Google search, Conversation Retrieval with high popularity and Conversation Retrieval with high density (x: score, y: number of votes)

Table 4 Average scores for the two tasks and three systems (only Friendfeed users)

Bold numbers indicate the best performance for each task

	Google	Popularity	Density
TASK 1	2.46	3	**3.31**
TASK 2	2.38	**2.62**	2.38

Focusing on the first selection of messages (related to the San José Mining accident) it appears that both selections of posts/conversations obtained with our approach receive higher scores with respect to Google's selection.

Looking at the three selections we can see that the approach based on popularity presents seven out of ten posts generated by the CNN account on FriendFeed, and it appears that the evaluators have expressed their appreciation of this option because of the authority of the sources of information, presented together with a few longer conversations. The approach based on density extracted these and other long conversations, that appear to have been well regarded by the evaluators because they provided some reasoning on the news and discussed related themes enriching the provided information.

By switching to the second selection of news we can see a slightly different scenario. In the case of the death of former Italian President Francesco Cossiga Google's selection is the one with the higher scores. According to the free comments of some users and following the qualitative analyses of the posts in the three selections we can see that the approaches based on social metrics received lower scores not because they contained posts judged as less interesting, but because some posts were considered not relevant with regard to the searched topic.

This behavior is caused by an interesting feature of the Italian community of this Social Network Site, producing almost all messages on this topic. Italian users tend to discuss much more than in other cultures. If we compare the average number of comments for each posted message in the Italian and English subsets of the messages we can see that Italian users comment ten times more than English users. The result of this behavior is the creation of very long and dense conversations addressing many related topics. These conversations were ranked high using our approach because of their high social activity but were considered out of topic by the evaluators because in practice they addressed other subjects. Also the low scores obtained by the popularity-based approach depends on the inclusion of the same very dense messages in the top-10 results.

These results lead us toward a more general interpretation of the collected data. It appears that the usage of social metrics can have a significant impact on the users' degree of interest in the retrieved posts. Nevertheless the ratio between social aspects and content

aspects has to be taken into careful consideration. Otherwise this could lead toward a loss of relevance as it seems happening in the *social* selection about Cossiga's death when the highly conversational environment of FriendFeed produced several conversations only lightly related with Cossiga's death but highly relevant according to their social parameters.

Finally, focusing only on the frequent users of this social service we can appreciate a very interesting result of our evaluation, indicated in Table 4. The messages selected by our popularity-based approach had been posted by users very well known in the Italian community. This can explain the different results obtained by the two subgroups: people not using this Social Network Site could not be aware of the important role played by these posters in the community, therefore they had no reason to consider these messages better than others—as they could do with respect to the CNN, which is known world-wide. On the contrary, users of the site know very well who is important in their community and the fact that the popularity-based approach is the best one according to their evaluation indicates that they have recognized and appreciated the presence of popular users in the top results.

6 Conclusion

In this paper we have introduced a microblogging search task called *conversation retrieval*: an information retrieval activity exploiting structural aspects in addition to the exchanged text messages. In particular, we have defined a formal conversational model, functions of relevance, popularity, timeliness and density to be used in the computation of the ranking of a conversation.

Our experimental results have highlighted many interesting points. First, including social features and the concept of conversation in the ranking function improves the relevance of the top-ranked tweets (quantitative analysis) and also provides results that are considered more satisfactory with respect to a traditional Web search task not taking these aspects into account (qualitative user evaluation). At the same time, especially in highly conversational environments there is the risk of providing too much heterogenous information to the users and to reduce the relevance of this information. Finally, we have seen that people knowing the context from where conversations are retrieved appreciate more the usage of social aspects in the search tasks, recognizing the alignment of these results with their knowledge of the system and of the different importance of conversations generated inside it.

While the results presented in this article show the importance of including SNA and conversational models into the ranking of microblogs, several aspects of microblogging search remain open. Among these, we consider of particular interest the study of how the size of the social network influences the ranking process, the further optimization of the parameter weights, and the adoption of alternative SNA metrics.

Acknowledgments This work has been partly funded by Telecom Italia.

References

Agarwal, N., Liu, H., Tang, L., & Yu, P. S. (2008). Identifying the influential bloggers in a community. In: *Proceedings of the international conference on web search and web data mining, WSDM '08*, pp. 207–218. ACM, New York, NY. doi:10.1145/1341531.134155.

Agosti, M., & Smeaton, A. F. (1996). *Information retrieval and hypertext*. Boston: Kluwer.

Amer-Yahia, S., Botev, C., & Shanmugasundaram, J. (2004). Texquery: a full-text search extension to XQuery. In: *WWW*.

Amer-Yahia, S., Fernandez, M. F., Srivastava, D., & Xu, Y. (2003). Phrase matching in XML. In: *Proceedings of the international conference on very large data bases*.

Amer-Yahia, S., Lakshmanan, L. V. S., & Pandit, S. (2004). Flexpath: Flexible structure and full-text querying for xml. In: *SIGMOD conference*.

Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., & Su, Z. (2007). Optimizing web search using social annotations. In: Proceedings of the 16th international conference on world wide web, WWW '07, pp. 501–510. ACM, New York, NY. doi:10.1145/1242572.1242640.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In: *Computer networks and ISDN systems*, pp. 107–117

danah, Boyd (2008). *Taken out of context: American Teen sociality in networked publics*. PhD thesis, University of California-Berkeley, School of Information.

Das Sarma, A., Das Sarma, A., Gollapudi, S., & Panigrahy, R. (2010). Ranking mechanisms in twitter-like forums. In: *Proceedings of the third ACM international conference on web search and data mining, WSDM '10*, pp. 21–30. ACM, New York, NY. doi:10.1145/1718487.171849.

Davis, J.V., Kulis, B., Jain, P., Sra, S., & Dhillon, I. S. (2007). Information-theoretic metric learning. In: *Proceedings of the 24th international conference on machine learning, ICML '07*, pp. 209–216. ACM, New York, NY. doi:10.1145/1273496.127352.

Dwork, C., Kumar, R., Naor, M., & Sivakumar, D. (2001). Rank aggregation methods for the web. In: *WWW*, pp. 613–622.

Easley, D. A., & Kleinberg, J. M. (2010). *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge: Cambridge University Press.

Elsas, J. L., Carbonell, J. G. (2010). It pays to be picky: an evaluation of thread retrieval in online forums. In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, SIGIR '09*, pp. 714–715. ACM, New York, NY. doi:10.1145/1571941.157209.

Evans, B. M., & Chi, E. H. (2008). Towards a model of understanding social search. In: *Proceedings of the 2008 ACM conference on computer supported cooperative work, CSCW '08*, pp. 485–494. ACM, New York, NY. doi:10.1145/1460563.146064.

Evans, B. M., Kairam, S., & Pirolli, P. (2010). Do your friends make you smarter? An analysis of social strategies in online information seeking. *Information Processing & Management, 46*, 679–692. doi:10.1016/j.ipm.2009.12.00.

Fagin, R., Kumar, R., & Sivakumar, D. (2003). Efficient similarity search and classification via rank aggregation. In: *Proceedings of the 2003 ACM SIGMOD international conference on management of data, SIGMOD '03*, pp. 301–312. ACM, New York, NY. doi:10.1145/872757.87279.

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks, 1*(1), 215–239.

Fuhr, N., & Großjohann, K. (2001). XIRQL: A query language for information retrieval in XML documents. In: *SIGIR conference*.

Fuhr, N., Rölleke, T. (1997). A probabilistic relational algebra for the integration of information retrieval and database systems. *ACM Transactions on Information Systems, 15*(1), 32–66.

Goyal, A., Bonchi, F., Lakshmanan, L. V. (2008). Discovering leaders from community actions. In: *Proceeding of the 17th ACM conference on information and knowledge management, CIKM '08*, pp. 499–508. ACM, New York, NY. doi:10.1145/1458082.145814.

Huang, J., Thornton, K. M., & Efthimiadis, E. N. (2010) Conversational tagging in twitter. In: *Proceedings of the 21st ACM conference on hypertext and hypermedia, HT '10*, pp. 173–178. ACM, New York, NY. doi:10.1145/1810617.181064.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In: *ACM conference on knowledge discovery and data mining (KDD)*. ACM.

Lalmas, M. (1997). Dempster-Shafer's theory of evidence applied to structured documents: modelling uncertainty. In: *Proceedings of the 20th annual international ACM SIGIR conference on research and development in information retrieval*, pp. 110–118. ACM Press. doi:10.1145/258525.25854.

Liu, W., Ma, S., Tao, D., Liu, J., & Liu, P. (2010). Semi-supervised sparse metric learning using alternating linearization optimization. In: *Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '10*, pp. 1139–1148. ACM, New York, NY. doi:10.1145/1835804.183594.

Magnani, M., & Montesi, D. (2010). Toward conversation retrieval. In: *Italian research conference on digital libraries*.

Magnani, M., Montesi, D., Nunziante, G., & Rossi, L. (2011). Conversation retrieval from twitter. In: *ECIR conference*. Accepted for presentation.

Magnani, M., Montesi, D., & Rossi, L. (2011). *Social Networks Analysis and Mining, chap. A study on factors influencing information diffusion in a social network site*. Lecture Notes in Social Networks. Springer, Berlin (in press).

Magnani, M., & Rossi, L. (2011). The ml-model for multi layer network analysis. In: *IEEE International conference on advances in social network analysis and mining*.

Mathioudakis, M., & Koudas, N. (2010). Twittermonitor: trend detection over the twitter stream. In: *Proceedings of the 2010 international conference on management of data, SIGMOD '10*, pp. 1155–1158. ACM, New York, NY. doi:10.1145/1807167.180730.

Morris, M. R., Jaime, T., & Panovich, K. (2010). A comparison of information seeking using search engines and social networks. In: *Proceedings of 4th international AAAI conference on weblogs and social media*, pp. 291–294. URL http://teevan.org/work/publications/posters/icwsm10.pd.

Nagmoti, R., Teredesai, A., & De Cock, M. (2010). Ranking approaches for microblog search. In: *Proceedings of the 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology–Volume 01, WI-IAT '10*, pp. 153–157. IEEE Computer Society, Washington, DC. doi:10.1109/WI-IAT.2010.17.

Newman, M. (2010). *Networks: An introduction*. New York, NY: Oxford University Press, Inc.

Opsahl, T., Agneessens, F. & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks, 32*, 245–251.

Seo, J., Croft, W. B., & Smith, D. A. (2009). Online community search using thread structure. In: *Proceeding of the 18th ACM conference on information and knowledge management, CIKM '09*, pp. 1907–1910. ACM, New York, NY, USA. doi:10.1145/1645953.164626.

Smith, M., Cadiz, J. J., & Burkhalter, B. (2000). Conversation trees and threaded chats. In: *Proceedings of the 2000 ACM conference on computer supported cooperative work, CSCW '00*, pp. 97–105. ACM, New York, NY. doi:10.1145/358916.35898.

Song, S., Li, Q., & Zheng, N. (2010). A spatio-temporal framework for related topic search in microblogging. In: *Proceedings of the 6th international conference on active media technology, AMT'10*, pp. 63–73. Springer, Berlin, Heidelberg. URL http://portal.acm.org/citation.cfm?id=1886192.188620.

Taylor, M., Zaragoza, H., Craswell, N., Robertson, S., & Burges, C. (2006). Optimisation methods for ranking functions with multiple parameters. In: Proceedings of the 15th ACM international conference on information and knowledge management, CIKM '06, pp. 585–593. ACM, New York, NY. doi:10.1145/1183614.118369.

Teevan, J., Ramage, D., & Morris, M. R. (2011). #twittersearch: a comparison of microblog search and web search. In: Proceedings of the fourth ACM international conference on Web search and data mining, WSDM '11, pp. 35–44. ACM, New York, NY. doi:10.1145/1935826.193584.

Wang, Y. C., Joshi, M., Cohen, W. W., & Rosé, C. P. (2008). Recovering implicit thread structure in newsgroup style conversations. In: Adar, E., Hurst, M., Finin, T., Glance, N. S., Nicolov, N., & Tseng, B. L. (eds). *ICWSM*. Menlo Park, CA: The AAAI Press.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. New York: Cambridge University Press

Watzlawick, P., Bavelas, J. B., & Jackson, D. D. (1967). *Pragmatics of human communication: A study of interactional patterns, pathologies, and paradoxes*. New York: W. W. Norton and Co.

Weinberger, K. Q., & Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res, 10*, 207–244. URL http://dl.acm.org/citation.cfm?id=1577069.157707.

Weng, J., Lim, E. P., Jiang, J., & He, Q. (2010). Twitterrank: finding topic-sensitive influential twitterers. In: *Proceedings of the third ACM international conference on Web search and data mining, WSDM '10*, pp. 261–270. ACM, New York, NY. doi:10.1145/1718487.171852.

White, D., & Borgatti, S. (1994). Betweenness centrality measures for directed graphs. *Social Networks, 16*(4), 335–346. doi:10.1016/0378-8733(94)90015-9.

Xi, W., Lind, J., & Brill, E. (2004). Learning effective ranking functions for newsgroup search. In: *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '04*, pp. 394–401. ACM, New York, NY. doi:10.1145/1008992.1009060.

Xing, E. P., Ng, A. Y., Jordan, M. I., & Russell, S. (2002). Distance metric learning, with application to clustering with side-information. In: *Advances in neural information processing systems 15*, pp. 505–512. MIT Press.

Yang, T., Chi, Y., Zhu, S., Gong, Y., & Jin, R. (2011). Detecting communities and their evolutions in dynamic social networks–a bayesian approach. *Mach. Learn, 82*, 157–189. doi:10.1007/s10994-010-5214-.