# Social Users Interactions Detection Based on Conversational Aspects

Rami Belkaroui, Rim Faiz, Aymen Elkhlifi

# Social Users Interactions Detection Based on Conversational Aspects

Rami Belkaroui[1], Rim Faiz[2], and Aymen Elkhlifi[3]

[1] LARODEC, ISG Tunis, University of Tunis
Bardo,Tunisia
rami.belkaroui@gmail.com
[2] LARODEC, IHEC Carthage, University of Carthage
Carthage Presidency, Tunisia
rim.faiz@ihec.rnu.tn
[3] LALIC, Paris Sorbonne University,
28 rue Serpente Paris 75006, France
aymen.elkhlifi@paris4.sorbonne.fr

**Abstract.** Last years, people are becoming more communicative through expansion of services and multi-platform applications such as blogs, forums and social networks which establishes social and collaborative backgrounds. These services like Twitter, which is the main domain used in our work can be seen as very large information repository containing millions of text messages usually organized into complex networks involving users interacting with each other at specific times. Several works have proposed tools for tweets search focused only to retrieve the most recent but relevant tweets that address the information need. Therefore, users are unable to explore the results or retrieve more relevant tweets based on the content and may get lost or become frustrated by the information overload. In addition, finding good results concerning the given subjects needs to consider the entire context. However, context can be derived from user interactions.
In this work, we propose a new method to retrieval conversation on microblogging sites. It's based on content analysis and content enrichment. The goal of our method is to present a more informative result compared to conventional search engine. The proposed method has been implemented and evaluated by comparing it to Google and Twitter Search engines and we obtained very promising results.

**Keywords:** Conversation retrieval, Social user interactions, Conversational context, Social Network, Twitter

## 1 Introduction

In the current era, microblogging services gives people the ability to communicate, interact and collaborate with each other, reply to messages from others and create conversations. This behavior leads to an accumulation of an enormous amount of information. Furthermore, microblogs tend to become a solid media for simplified collaborative communication.

Twitter, the microblogging service addressed in our work, is a communication mean and a collaboration system that allows users to share short text messages, which doesn't exceed 140 characters with a defined group of users called followers. Users can reply to each other simply by adding @sign in front name user they are replying to. This set of socio-technical features has made possible for Twitter to host a wide range of social interactions from the broadcasting of personal thoughts to more structured conversations among groups of friends [1]. While communicating people share different kind of information like common knowledge, opinions, emotions, information resources and their likes or dislikes. The analysis of those communications can be useful for commercial applications such as trends monitoring, reputation management and news broadcasting. In addition, one of main characteristic of Twitter is that users are not limited to produce contents, they can get involved indirectly in conversations with other users by liking and sharing user's posts. Furthermore, several works [2] have proposed tools for tweets search focused only to retrieve recent tweets. But, these tools are not powerful enough if we want to get tweets about an event or a product. Users are unable to explore results or retrieve more relevant tweets based on content.

This paper proposed a conversation retrieval method which can be used to extract conversation from twitter in order to provide an informative result for users' information needs based on user's content interactions analysis. Comparing with current methods, the new proposed not only extract directly reply tweets, but also relevant tweets which might be retweets or comments and other possible interactions. The method extracts extensive posts beyond conventional conversation.

The rest of the document is organized as follows: we begin by presenting related work in related domains such as forums discussion, Email threads. Then, we focus on more recent works addressing conversation retrieval on microblogging sites. In section 3, we propose our method allows extracting user's content interactions on microblogging sites. The experimentation and evaluation results are detailed in section 4. The final section presents a summary of our work and future directions.

## 2 Related Work

There are a number of related areas of work, including discussion forums search, email thread detection and conversation search from microblogging sites like Twitter, which is the main domain used in our work. In this section, we will discuss each area in turn.

### 2.1 Email Threads reply Detection

Previous research has been focused on using email structure especially emails threads [3,4,5]. Thread detection is an important task which has attracted significant attention [6]. Email is one of the most important tools for treating conversations between people. Generally, a typical user mailbox encloses hundreds of

conversations. Few works indirectly address to the problem of thread reply reconstruction. Accorded to [3], the detection of these conversations has been identified as an important task. Clustering the messages into coherent conversations useful to applications, among them, it gives users the opportunity to see a messages greater context they are reading and collating related messages automatically. In[6] the authors suggested a method that allows to assemble messages having the same subject attributes and send them among the same group of people. However, conversations may span several threads with similar (but not exact) subject lines. Furthermore, conversations not include all the participants in all the messages. In the same way,[7] developed an email client extension that makes it possible to clusters messages by topic. However, their clustering approach is focused on topic detection, hence messages belonging to different conversations on the same topic will be clustered together. In addition,[4] recreated reply emails chains, called email threads. The authors suggested two approaches, one based on using header meta-information, the other based on timing, subject and emails content. But, this method is specific for emails and the features cannot be easily extended for microblogs conversation construction.

### 2.2 Forums Threads Structure Identification

An online forum is a Web application for holding discussions and posting User Generated Content (UGC) in a particular domain, such as sports, recreation, techniques, travel, etc. In forums, conversations are represented as sequences of posts, or threads, where the posts reply to one or more earlier posts. Several studies have looked at identifying the structure of a thread, question-answer pairs or responses that relate to a previous question in the thread. There are many works on searching forum threads that dealt with the reply-chains structure or reply-trees. [8] has concentrated on identifying the thread structure when explicit connections between messages are missing. Despite the fact that replies to posts in microblogging sites, are commonly explicit, it is proved that different autonomous conversations may be developed inside the same replies thread. Furthermore, distinct threads may belong related to macro-conversations. For example, being Twitter hashtags that connect separate threads by common topic. In [9] authors represent the principal differences between traditional IR tasks and searching in newsgroups. They use a measures combination such as author metrics (posts number, number of replies, etc.) and features threads.

### 2.3 Conversation Retrieval form Microblogging Sites

Conversation retrieval is a new search paradigm for microblogging sites. It results from the intersection of Information Retrieval and Social Network Analysis (SNA). Most of microblogging services provide a way to retrieve relevant information [10], but lack the ability to provide all tweets discussion. There have been few previous researches dealing specifically with conversation detection. In addition, existing conversation retrieval approaches for microblogging sites [11,12,13,14] have so far focused on the particular case of a conversation formed

by directly replying tweets. [11] proposed a user-based tree model for retrieving conversations from microblogs. They considered only tweets that directly respond to other tweets by the use of @sign as a marker of addressivity. The advantage of this approach is to have a coherent conversation based on direct links between users. Furthermore, the downside is that this method does not consider tweets that do not contain the @sign. Similarly [12] proposed a method to build conversation graphs, formed by users replying to tweets. In this case, a tweet can only directly reply to other tweet. However, users can get involved indirectly in conversations communities by commenting, liking, sharing user's posts and other possible interactions. In [15] the authors concentrated on different microblogging conversations aspects. They proposed a simple model that produces basic conversation structures taking into account the identities of each conversation member. Other related works [16,17] focusing on different aspects of microblogging conversation, that deal respectively with conversation tagging and topics identification.

## 3 TCOND: Twitter Conversation Detection Method

We propose an information retrieval method for microblogging sites particularly Twitter based on conversation retrieval concept. Our method combines direct and indirect conversation aspects in order to extract extensive posts beyond conventional conversation. In addition, we defined a conversation as a set of short text messages posted by a user at specific timestamp on the same topic. These messages can be directly replied to other users by using "@username" or indirectly by liking, retweeting and commenting.

### 3.1 Proposed Method

Our method consists in 2 phases:

- Phase 1: Constructing the direct reply tree using all tweets in reply directly to other tweets.
- Phase 2: Detecting the relevant tweets related indirectly to a same reply tree which might be retweets, comments or other possible interactions in order to extract extensive posts beyond conventional conversation.

**Phase1: Twitter Direct Conversations Construction** In this phase, we aim to collect all tweets in reply directly to other tweets. Obviously, a reply to a user will always begin with "@username". Our goal in this step is to create reply tree. The reply tree construction process consists of two algorithms run in parallel recursive root finder algorithm and iterative search algorithm.

Let T0 is the root (first tweet published) of the conversation C and T is a single tweet of the conversation retrieved. Let consider $T_i$ the type of tweet T. A tweet can have three types: root, reply or retweet. The goal of the Recursive Root Finder Algorithm is to identify the conversation root T0 given T. Note that

**Algorithm 1** Recursive Root Finder (A:twitter)

Let T be a tweet collected from Twitter (ID tweet)
**while** ($T_i$ !=root) **do**
  Extract $T_i$- 1 by matching field "in reply to status id"
**end while**
A : twitter = A : twitter – 1

when the algorithm starts,|T| is not known. Once, the conversation root T0 has been established, the Iterative Search Algorithm is used to seek the remainder of conversation C by searching all tweets addressed to $T_i$ using matching field "in reply to status id". It is run repeatedly until some conditions, indicating that the conversation has ended, are met.

**Phase2: Indirect Reply Structure Using Conversational Features** To the best of our knowledge, there has not been previous work on the structure of reply-based on indirectly conversation. Therefore, we define new features that may help to detect tweets related indirectly to a same conversation. The goal of our method is to extract tweets that may be relevant to the conversation without the use of the @symbol. We use the following notations in the sequel:

- $t_i$ is a set of tweets present in direct conversation (tweets in reply to other tweets directly).
- $t_j$ is a tweet that can be linked indirectly to conversation.

The features we used are:

- ***Using the same URL***:

Twitter allows users to include URL as supplement information to their tweets. By sharing an URL, an author would enrichment the information published in his tweet. This feature is applied to collect tweets that share the same URL.

$$P1(t_i, t_j) = \begin{cases} 1 \text{ if t contains the same URL.} \\ 0 \text{ otherwise.} \end{cases} \quad (1)$$

- ***Hashtags Similarity***:

The # symbol, called hashtag, is used to mark a topic in a tweet or to follow conversation. Any user can categorize or follow topics with hashtags. We used this feature to collect tweets that share the same hashtags.

$$P2(t_i, t_j) = \begin{cases} 1 \text{ if t contains the same hashtag.} \\ 0 \text{ otherwise.} \end{cases} \quad (2)$$

– ***Tweets Time Difference***:

The time difference is highly important feature for detecting tweets linked indirectly to the conversation. We use the time attribute to efficiently remove tweets having a large distance in terms of time compared to conversation root.

– ***Tweets Publication dates***:

Date attribute are highly important for detecting conversations. Users tend to post tweets about conversational topic within a short time period. The euclidean distance has been used to calculate how similar two posts publication dates are.

– ***Content***:

The criterion content refers to the thematic relevance traditionally calculated by IR systems standards. We compute the textual similarity between each element in $t_i$, $t_j$ taking the maximum value as the similarity measure between two messages. The similarity between two elements is calculated using the well-known tf-idf cosine similarity, $\mathrm{sim}(t_i, t_j)$.

– ***Similarity Function***:

Finally, the similarity between tweets indirectly linked to conversation and tweets which are present in the reply tree is calculated by a linear combination between their attributes.

## 4 Experiments and Results

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

- Using a set of conversational features improves the conversation search task 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 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 method. The second involves several different combination functions to evaluate the criteria applicability for conversation detection.

### 4.1 Quantitative Analysis

The following experiment has been designed to gather some knowledge on the impact of our results on end-users. For this experiment we have selected

three events and queried our dataset using Google[4], Twitter search engine[5] and our method (TCOND). Then, we have asked a set of assessors to rate the top-10 results of every search task, to compare these approaches. In order to measure the results quality, we use the Normalized Discounted Cumulative Gain (NDCG) at 10 for all the judged events. In addition, we used a second metric which is the Precision at top 10. In the following, we first describe the experimental setting, then we present the results and finally we provide an interpretation of the data.

**Experimental Settings** The dataset has been obtained by monitoring microblogging system Twitter posts over the period of July-August 2013. In particular, we used a sample of about 113 000 posts containing trending topic keywords using Twitter's streaming API. Trending topics (the most talked topic on twitter) have been determined directly by Twitter, and we have selected the most frequent ones during the monitoring period.

To evaluate our results search tasks we have used a set of 100 assessors with three relevance levels, namely highly relevant (value equal to 2), relevant (value equal to 1) or irrelevant (value equal to 0). The assessors selected among students and colleagues of the authors with backgrounds in computing and social sciences, on a voluntary base, and no user was aware of the details of the underlying systems. Every user was informed of three events happened during the sampling period. For each event we performed three searches:

1. One using Google.
2. One using Twitter Search.
3. One using our method (TCOND).

The evaluators were not aware of which systems had been used. Every user for each search task was presented with three conversations selections, one for each of the previous options with the corresponding top-10 results.

**Experimental Outcomes and Interpretation Results** We compare our conversation retrieval method with the results returned by Google and by Twitter search engine using two metrics namely the P@10 and the NDCG@10. From this comparison, we obtained the values summarized in Table 1 where we notice that our method overcomes the results given by both of Google and Twitter. The reason of these promising values is the fact that we combine a set of conversational features and direct replies method to retrieve conversation may have a significant impact on the users' evaluation.

Focusing on the three messages selections, we observe that all conversations obtained with our method receive higher scores with compared to Google and Twitter's selection. According to the free comments of some users and following the qualitative analysis of the posts in the three selections we can see that Google and twitter received lower scores not because they contained posts judged as less

[4] www.google.com
[5] Search.twitter.com.

interesting, but because some posts were considered not relevant with regard to the searched topic.

Concentration on the three messages selections we observe that all conversations selections obtained with twitter search has higher scores with respect to Google's selection. These results lead us toward a more general interpretation of the collected data. It appears that the social metrics usage have a significant impact on the users' degree interest in the retrieved posts. In addition, the retrieving conversations process from Social Network differs from traditional Web information retrieval; it involves human communication aspects, like the degree interest in the conversation explicitly or implicitly expressed by the interacting people.

**Table 1.** Table of Values for Computing our Worked Example

| | P@10 (Average%) | NDCG (Average%) |
|---|---|---|
| **Task1** | | |
| Google | 59.62 | 56.86 |
| Twitter | 65.73 | 59.71 |
| TCOND | **73.28** | **64.52** |
| | | |
| **Task2** | | |
| Google | 57.31 | 56.02 |
| Twitter | 62.78 | 58.45 |
| TCOND | **67.27** | **62.73** |
| | | |
| **Task3** | | |
| Google | 63.21 | 66.52 |
| Twitter | 65.88 | 68.46 |
| TCOND | **77.27** | **69.33** |

### 4.2 Qualitative Analysis

For this experiment, we worked on our same social dataset mentioned in the previous section. Our goal is to evaluate the usefulness of all proposed criteria on tweets that have not been appeared in direct conversation (created on the basis of our two algorithms detailed in the section 3) to detect tweets which are indirectly related to conversation such as retweets or comments and other possible interactions.

**Features Evaluation** In this part, we evaluate the usefulness of single features and combinations of features for indirect conversation detection. To evaluate the applicability of criteria, we experiment with different coefficient weights of the similarity function. The results are given in Table 2 and demonstrate the significance of considering all message features by the similarity function. In addition,

the results show that to get higher-quality conversations, the content attribute should be weighted higher than other attributes. However, Table 2 shows also that the dates attribute as almost insignificant for the conversation detection task, and that the content attribute is more important than the time attribute. The previous experiment supports the hypothesis that a combination of a set conversational feature can improve the results of search tasks in microblogging sites.

**Table 2.** Conversation Detection Results, Using Different Coefficient Weights

| | Precision | Recall |
|---|---|---|
| Presence of same URL | 0.68 | 0.61 |
| Hashtags Similarity | 0.74 | 0.86 |
| Time Difference | 0.63 | 0.71 |
| Tweets Publication Dates | 0.58 | 0.64 |
| Content | **0.82** | **0.86** |
| | | |
| SameURL+Hashtags Simi | 0.69 | 0.71 |
| SameURL+Content | 0.81 | 0.68 |
| Hashtags+Content | **0.87** | **0.85** |
| | | |
| Hashtags Similarity+Time | 0.81 | 0.70 |
| SameURL+Hashtags+Content | **0.88** | **0.86** |
| Time+Hashtags+Content | 0.82 | 0.76 |
| | | |
| All | **0.96** | **0.92** |

## 5 Conclusion

This work explored a new method for detecting conversation on microblogging sites: an information retrieval activity exploiting a set of conversational features in addition to the directly exchanged text messages to retrieve conversation. Our experimental results have highlighted many interesting points. First, including social features and the concept of direct conversation in the search function improves the relevance of tweets informativeness and also provides results that are considered more satisfaction with respect to a traditional tweet search task. Future work will further research the conversational aspects by including human communication aspects, like the degree of interest in the conversation and their influence/popularity by gathering data from multiple sources from Social Networks in real time.

## References

1. Boyd, D., Golder, S., Lotan, G.: Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In: Proceedings of the 2010 43rd Hawaii International

Conference on System Sciences, HICSS '10, IEEE Computer Society, 1–10, (2010)
2. Jabeur, L., Tamine, L., Boughanem, M.: Uprising Microblogs: A Bayesian Network Retrieval Model for Tweet Search. In: Proceedings of the 27th Annual ACM Symposium on Applied Computing, 943–948, (2012)
3. Kerr, B. : Thread arcs: an email thread visualization. In: Proceedings of the Ninth annual IEEE conference on Information visualization, 8, 211–218, (2003)
4. Yeh, J.: Email Thread Reassembly Using Similarity Matching. The Third Conference on Email and Anti-Spam (CEAS), Mountain View, California, USA, (2006)
5. Erera, S., Carmel, D.: Conversation detection in email systems. In: Proceedings of the IR research, 30th European conference on Advances in information retrieval, 498–505. Springer, Heidelberg (2008)
6. Klimt, B., Yang, Y.: Introducing the Enron Corpus. The Third Conference on Email and Anti-Spam (CEAS), Mountain View, California, USA, (2006)
7. Cselle, G., Albrecht K., Wattenhofer, R.: BuzzTrack: topic detection and tracking in email. In: Proceedings of the 12th international conference on Intelligent user interfaces, 190–197, (2007)
8. Wang, Y., Joshi, M., Cohen, W., Rosé, C.: Recovering Implicit Thread Structure in Newsgroup Style Conversations. the Second International Conference on Weblogs and Social Media, ICWSM 2008, Seattle, Washington, USA, The AAAI Press, (2008)
9. Xi, W. Lind, J., Brill, E.: 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, Sheffield, United Kingdom, 394–401, (2004)
10. Cherichi, S., Faiz,R.: New metric measure for the improvement of search results in microblogs. In: Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics, WIMS '13 24:1–24:7, (2013)
11. Magnani, M., Montesi, D., Nunziante, Gabriele, Rossi, L.: Conversation retrieval from twitter. In: Proceedings of the 33rd European conference on Advances in information retrieval, ECIR'11. 780–783, (2011)
12. Cogan, P., Andrews, M., Bradonjic, M., Kennedy, W. S., Sala, A., Tucci, G.: Reconstruction and analysis of Twitter conversation graphs. In: Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research 7, 25–31, (2012)
13. Matteo, M., Danilo, M. and Luca, R.: Information Propagation Analysis in a Social Network Site ASONAM, IEEE Computer Society 296–300, (2010)
14. Magnani, M., Montesi, D. and Luca, R.: Conversation retrieval for microblogging sites. In: Information Retrieval Journal, 15,354-372. Springer, Heidelberg (2012)
15. Kumar, R., Mahdian, M., McGlohon, M.: Dynamics of conversations. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining Washington, DC, USA, 553–562, (2010)
16. Huang, J., Thornton, K.M., Efthimiadis, E. N.: Conversational tagging in twitter. In: Proceedings of the 21st ACM conference on Hypertext and hypermedia 173–178, (2010)
17. Song, S., Li, Q., Zheng, N.: A spatio-temporal framework for related topic search in micro-blogging. In: Proceedings of the 6th international conference on Active media technology, 63–73. Springer, Berlin/Heidelberg (2010)