Conversation Retrieval From Social Networks

Rami Belkaroui

Abstract

Retrieving information from Twitter is always a hard task given its volume, inconsistent writing and noise. Existing systems focused only to retrieve separate tweets. However, the entire conversation is often neglected, leading to less satisfactory results while searching information from tweets.

In this work, we propose an information retrieval approach for microblogging sites based on conversation retrieval concept. Our approach detects automatically conversations generated on the popular twitter social network site, particularly conversations about trending topics. Thus, we focus our work not only on the particular case of a conversation formed by users replying to tweets, but also tweets which are linked indirectly to conversation.

In order to evaluate our framework performance, we start by modeling the conversation retrieval approach according to a set of data collected from Twitter, i.e a social networking and microblogging service. Then, we compare our results with two baseline systems on conversation models. Finally, Suggestions on how the system performance can be improved are given.

Dedications

Believing that there is no mountain higher as long as **Allah** is on our side,

I dedicate this thesis to My beloved parents **Habib** and **Hamida** ,

and My brother **Ramzi** for believing in me and encouraging me during all of my years of study.

I also dedicated it to all of my dear friends for their confidence and support.

Ramy

Acknowledgements

First, my sincere gratitude goes to **Pr. Rim FAIZ**, my master thesis supervisor for her encouragement and guidance that truly help the progression of my research internship. I am also thankful for her availability, supervision, constructive criticism and valuable advices, allowing me achieving and improving this work.

It is with immense honor that, I thank all of the **LARODEC** members for their help during my internship.

The special thank also goes to **Pr. Mohand BOUGHANEM**, the supervisor of this master thesis at University of Toulouse III, for accepting me among the team **SIG-IRIT**, introducing me the field of information retrieval and inspiring me throughout this research. I also want to thank him for his uninterrupted encouragement, time, and efforts.

Furthermore, I acknowledge the guidance of **Mr. Firas Damak** who assisted me to be better integrated into the workplace within the Research Institute on Computer Science of Toulouse (IRIT).

Finally, it is with immense gratitude that I thank all the jury members for agreeing to evaluate my work and all the professors for the knowledge and skills they gave me during all my years of study.

Contents

Introduction **1**

1 A Review of Information Retrieval **4**

1.1 Introduction 4

1.2 Information Retrieval 5

1.2.1 Information Retrieval Definition 5

1.2.2 Basic Concepts 6

1.3 The Standard Process of Information Retrieval 8

1.3.1 Indexing 9

1.3.1.1 Tokenization 10

1.3.1.2 Stop Words Removal 11

1.3.1.3 Stemming and Lemmatization 11

1.3.1.4 Weighting 12

1.3.2 Matching 14

1.3.3 Relevance Feedback 14

1.3.3.1 Explicit Relevance Feedback 15

1.3.3.2 Pseudo Relevance Feedback 15

1.3.3.3 Implicit Relevance Feedback 15

1.4 Information Retrieval Modelling 15

1.4.1 The Boolean Model 16

1.4.2 The Vector Space Model 18

1.4.3 The Probabilistic Model 20

1.5 Utility, Relevance, and IR System Performance 21

1.6 Problem Description 23

1.7 Conclusion 25

2 Information Retrieval In Social Media 26

2.1 Introduction 26

2.2 Information Retrieval and Social Analysis Area 27

2.2.1 Web 2.0: Social Field Evolution 28

2.2.2 User-generated Content (UGC) 29

2.3 Information Retrieval on the Web 30

2.4 Information Retrieval in Microblogs 31

2.5 Social Networks 31

2.5.1 Microblogging 32

2.5.2 Microblog Conversation Definition 34

2.6 Twitter 34

2.6.1 Twitter Background 34

2.6.2 Twitter Conventions 37

2.6.3 Replies and Retweets on Twitter 39

2.7 Conclusion 41

3 Conversational Approach 42

3.1 Introduction 42

3.2 Conversation Retrieval From Social Media 43

3.2.1 Email Threads Detection 43

3.2.2 Summarizing Threaded Conversations 45

3.2.3 Forum Thread Search 46

3.2.4 Retrieval of Twitter Microblogs 49

3.3 Microblog Conversation Approach 50

3.4 Microblog Conversation Architecture 52

3.4.1 Replies Conversation Method 54

3.4.2 Conversational Features Selected 55

3.5 Conclusion 58

4 Experiments and Results **59**

4.1 Introduction 59

4.2 Implementation Modules 59

4.2.1 Tweet Retrieval Module 61

4.2.2 Conversation Server Module 61

4.2.3 Display Module 62

4.3 TweetsConvs System Description 62

4.4 Experiments 66

4.4.1 Data description 66

4.4.2 Evaluation Metrics 68

4.5 Our Approach Results 69

4.5.1 Properties of Conversations 70

4.5.1.1 Depth 70

4.5.1.2 Size 71

4.5.2 Experimental Outcomes 72

4.5.2.1 Conversation' Relevance 72

4.5.2.2 Conversation'Consistency 73

4.6 The Kappa Statistic 73

4.7 Comparison with Baseline Approachs 75

4.8 Discussion 76

4.9 Conclusion 77

Conclusion **78**

Appendix **80**

Bibliography **82**

List of Figures

1.1 General Model of the IR Process 9
1.2 A stop list of 25 semantically non-selective words which are common 11
1.3 Diagram of a Boolean query 17
1.4 A query and document representation in the vector space model . . 19
1.5 Retrieved versus relevant documents 22

2.1 Example of Hashtags 36
2.2 Function of @Reply 39
2.3 The first Retweet's Type 40
2.4 The second Retweet's Type 41

3.1 Microbog Conversation Architecture 53
3.2 Tweets' Time Difference 57

4.1 Our Framework Components 60
4.2 User's Identification 63
4.3 User's Failed Identification 63
4.4 Track Conversation 64
4.5 TweetsConvs Search Results 65
4.6 Tweets Linked to Conversation 65
4.7 Based on "in reply" and "not in reply" tweets representation 67
4.8 Conversation Levels Deep 70
4.9 Conversation Length 71
4.10 Conversation'Relevance 72
4.11 Conversation'Consistency 73

List of Tables

2.1 Functions of the @sign in English tweets (Honeycutt and Herring, 2009) 38

4.1 Data Set Statistics 67

4.2 Evaluation of Our System "TweetsConvs" 69

4.3 Table of Values for Computing our Worked Example 74

4.4 Interpretation of Kappa 75

4.5 The Comparison Result 76

List of Algorithms

1 Iterative Root Finder 54
2 Iterative Search 55

List of Equations

1.1 Term Frequency 13
1.2 Inverse Document Frequency 13
1.3 Term Frequency-Inverse Document Frequency 13
1.4 The Cosine Formula 19
1.5 The Similarity Score 20
1.6 The Recall Formula 22
1.7 The Precision Formula 22
1.8 The F-measure Formula 22
3.1 Presence of the same URL Formula 56
3.2 Presence of the same Hashtag Formula 56
3.3 Similarity Function 58
4.4 The Kappa Formula 69
4.5 The P Formula 74
4.6 The P_eFormula 74
4.7 The Kappa Result 75

Introduction

Communication over social networks has been an emergent phenomenon over the last several years. A growing amount of online information is produced by users (User Generated Content) and published on social network sites. In addition, Social Networks Sites (SNS) are among the most relevant source of knowledge where information is created, exchanged and transformed, as witnessed by the important number of their users and their activities during events or campaigns like the terror attack in Mumbai[1] in 2008 or the terror attack in Boston[2] in 2013.

Recently, microblogging services draw the attention of users due to the ease and the speed of information sharing especially in real time. Those services 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. For instance, at the beginning of the XXX Olympic Games in London in 2012 spectators are believed to disturb television coverage of a cycling road race1. People tweeted so much about their experience that mobile network could not handle the load. This is just one of many examples on how social media connect the people. Clearly, in the vast amount of produced tweets, there is a certain number of discussions that mirror the event and reveal spectators attitude to the event which is very fruitful to analyze.

In addition, Twitter[3] which is one of the most famous social media platforms

[1] http://edition.cnn.com/2008/WORLD/asiapcf/11/27/mumbai.twitter/

[2] http://fullcomment.nationalpost.com/2013/04/16/matt-gurney-boston-attacks-show-traditional-journalists-that-twitter-has-come-of-age/

[3] www.twitter.com

in the world, with over 340 million tweets[4] and over 1.6 million search queries[5] each day. In addition, this service gives people the ability to communicate, interact and collaborate with each other. However, users can get involved indirectly in conversations communities by commenting, liking and sharing other user's posts. Furthermore, conversations still emerge between dyads and groups, even though, users can find who share the same passions, beliefs, hobbies, or lifestyles. Therefore, the ability of users to interact with each other, reply to messages from others and create conversations is a new phenomenon.

The use of the Social Networks gives a new global importance as an additional incentive to find a mechanism for retrieving information that is not limited to pure research on text messages. Existing systems focused only to retrieve the separate tweets. However, the entire conversation with direct and indirect relationships is often neglected, leading to less satisfactory results while searching information from Twitter. Therefore, search functionality provided by Twitter is limited to keyword-based retrieval in order to return the most recent posts, but, it is not powerful enough if the goal of the search is to get tweets about an event or a product. 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 this work, we propose an information retrieval approach for microblogging sites based on conversation retrieval concept. Our approach detects automatically conversations generated on the popular Twitter social network site, particularly conversations about trending topics. The 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 reach our target, we describe two steps for constructing twitter conversation. The first step is the replies conversation method which produces the reply tree. A reply to a user will always begin with "@username", thus a search call will return all the replies to him. In order to

[4]https://blog.twitter.com/2012/twitter-turns-six
[5]http://engineering.twitter.com/2011/05/engineering-behind-twitters-new-search.html

determine whether the returned tweet is an element of the conversation, we can check the "in-reply-to-status-id" field. The second one is the conversation method features selected, which produces tweets that may be related to the conversation without the use of the @symbol that can be linked indirectly to the conversation (commenting, liking and sharing other users' posts). The objective is to distinguish relevant tweets from irrelevant ones regarding a discussion. In order to achieve this task, we define selection features namely the presence of the same hashtags and the same URLs, the time difference between posts and the content similarity.

Document Organization

This document is composed by four chapters:

- Chapter 1: We begin by describing the background of information retrieval as the framework for this work. We start with its basic concepts, main issues. Then, its existing potential applications in the literature. Finally, our thesis problem is stated by means of important questions related to conversation retrieval.

- Chapter 2: We we begin with introducing the social information retrieval. Then, we studie social information retrieval include social realm evolution and the different types of information retrieval.

- Chapter 3: We begin by presenting related works focusing on the problem of conversation retrieval and we identifies their limits. Then, we present our contribution. Finally, we get into more details about the proposed conversation retrieval approach by describing the system architecture and its components.

- Chapter 4: We describe the experimental methodology used to evaluate our proposed approach followed by the corresponding results and their discussion. The last part of this thesis, is devoted to summarize the conclusions and the guidelines for future work.

Chapter 1

A Review of Information Retrieval

1.1 Introduction

The Information Retrieval (IR) can be defined as an activity whose purpose is to locate and deliver a set of documents to a user based on its need for information. The major limitation of traditional search systems is returning the result's same list to a same query submitted by different users. However, users may have different interests and therefore have different information needs. The challenge is to be able, from the large volume of documents available, to find the ones that best fit user's expectations. This problem has become more developed with the rise of the Web 2.0 that has handed over IR new challenges of access to information; it is the time to find relevant information in a diverse area and of considerable size. These difficulties have led to a new discipline called Information Retrieval on the Web.

Previously, we highlight the details of our problem and contribution, introduce in this chapter a general overview of the IR field. This chapter is organized as follows: section 1.1, presents the IR field. Section 1.2, presents the basic concepts of the IR field. Section 1.3, describes the process of IR with its steps. Section 1.4, outlines the main classical models. Section 1.5, we describe the IR system

performance metrics and finally, we highlight the details of our problem.

1.2 Information Retrieval

1.2.1 Information Retrieval Definition

The discipline of information retrieval is almost as old as the computer itself. The definition of information retrieval is: "Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text)that satisfies an information need from within large collections (usually stored on computers)".

The Information Retrieval (IR) can be also defined as the science of extracting useful knowledge from data stored in documents, relational databases, or on the World Wide Web is known as information retrieval. Due to their data-oriented nature, computers are the tools best suited for such a task.

An information retrieval system is a software programme that stores and manages information on documents. It is designed specifically as purpose of locating information stored on computer systems are called search engines. The system assists users in finding the information they need. The goal of search engines is to reduce the amount of time necessary to extract useful information from a data source, and to provide an interface which hides the complex details of these engines from their users. In order, to interact with a search engine, a user must specify certain criteria about their interest's subject. This is known as a query. The engine therefore makes its best attempt to locate the most relevant search results relating to the search query. Some of the most prominent examples of search engines are the variety of online Web search engines which allow WWW's users (World Wide Web) to locate Web pages, images, videos, and other information over the Internet.

A perfect retrieval system would retrieve only all relevant documents and no irrelevant document. However, perfect retrieval systems do not exist and will not exist because search statements are necessarily incomplete and relevance depends

on the subjective opinion of the user. Two users may pose the same query to an information retrieval system and give different relevance judgments on the retrieved documents.

1.2.2 Basic Concepts

We quote below the definition of IR given by Sparck Jones and Willett (1997) in its original form: "The user expresses his information need in the form of a request for information. Information retrieval is concerned with retrieving those documents that are likely to be relevant to his information need as expressed by his request".

More specifically, an Information Retrieval System (IRS) is a software tool for data representation, storage and research. The selected documents are relevant when they meet the user's need expressed in the submitted query. From this definition, we define the following key concepts:

- **Document**

  A Document is known as any unit of information that can be an answer to a need expressed by the user. A document can be a text, a piece of text, image, video, URL, etc.

- **Need Information**

  The notion of information need in information retrieval is often likened to user's requirements. Three types of user's need were defined by Ingwersen (1992):

  - Verificative needs, or locational information problems, i.e; the user wants to verify or locate items, e.g., some specific articles (or a list of client addresses). Characteristic bibliographic data (formal data), e.g., sources, pages, author name, titles words, are in this case known to the user.

  -Conscious topical needs, i.e; the user want to clarify, review or pursue aspects of known subject matter.

- Muddled topical needs, or ill-defined information problems i.e; the user wants to explore some news concepts or concepts relations outside known subject matter.

- **Query**

  A query is a representation of the user's information need. For a given search session, the user must submit a query to the search engine in which he/she specifies the keywords describing his/her information need.

- **Relevance**

  According to preliminary definitions(Saracevic, 1970), relevance is the matching between a document and a query, or a measure of how a document corresponds to a query. The concept of relevance is the primary criteria for evaluating IRS. In addition, many studies have been conducted around the concept of relevance and show that the relevance judgment is based on the degree of similarity between the query's representation and the content of the documents retrieved by the search system. We cite the following types of relevance (Tamine-Lechani and calabretto, 2008).

  - The algorithmic relevance (or system): It is often presented by a score of similarity between document's content and the query's terms.

  - The topical relevance: It reflects the quality of matching between the information of retrieved documents and the topic expressed in the query's keywords.

  - The cognitive relevance: It is defined as the relationship between the knowledge state of the user and the information carried by the documents.

  - The situational relevance (utility): It is the relevance of the resources according to the situation, task, or problem of the user.

1.3 The Standard Process of Information Retrieval

According to (Smeaton, 1990) "The purpose of an information retrieval system is to retrieve documents in response to a request from the user, so that the contents of documents are relevant to the initial information need of the user". Information Retrieval (IR) is the activity of finding information by entering a text or word using the search engine. Information Retrieval is used to search for text or graphics on the World Wide Web. The main purpose of Information Retrieval system (IRS) is to select all information resources relevant to an information need from a collection of information resources. But, the result of the output may be relevant or irrelevant. The origin module within the searcher's system is responsible for establishing the connection to the Target system, formulating the query according to the database, interpreting the results in a format familiar to the searcher, keeping track of the results, terminating a search session, etc.

A user neither have time, not wish to spend the time reading the entire document collection, apart from the fact that it may be physically impossible for the user to do so. The purpose of an automatic retrieval strategy is to retrieve all the relevant documents at the same time retrieving as few of the not-relevant as possible. Information retrieval system it merely informs on the existence (or non-existence) and whereabouts of documents relating to his request information storage and retrieval is simple. Suppose there is a store of documents and a person (user of the store) formulates a question (request or query) to which the answer is a set of documents satisfying the information need expressed by his question. He can obtain the set by reading all the documents in the store, retaining the relevant documents and discarding all the others.

Although information retrieval can be subdivided in many ways, it seems that there are three main areas of research which between them make up a considerable portion of the subject. They are: content analysis, information structures, and evaluation. Briefly, the first is concerned with describing the contents of documents in a form suitable for computer processing; the second with exploit-

ing relationships between documents to improve the efficiency and effectiveness of retrieval strategies; the third with the measurement of retrieval's effectiveness.

Bringing out such a system consists mainly in implementing a process, shown in Figure 1.1, which lies in two main phases: indexing and matching.

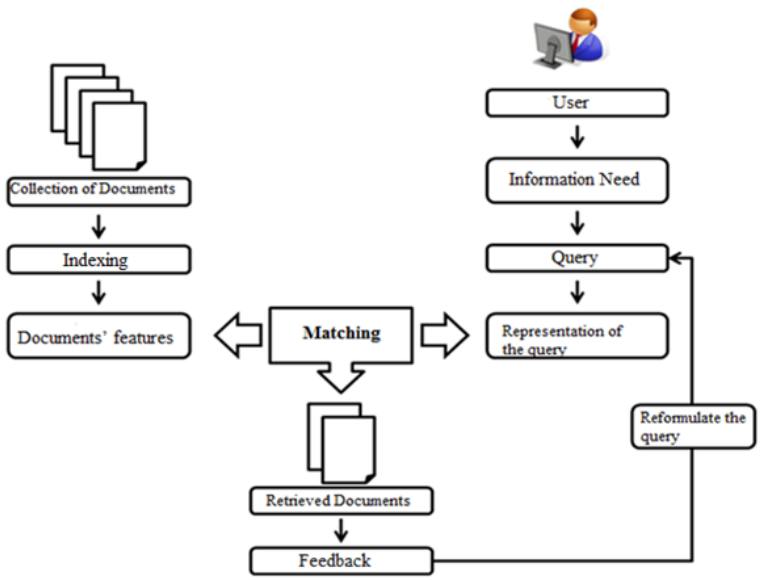

Figure 1.1: General Model of the IR Process

1.3.1 Indexing

Indexing is a set of techniques employed to transform the documents (or queries) into substitutes or descriptors which can represent their contents (Salton and McGill, 1986). These descriptors characterize the indexing language which is often represented through a structure. This structure is based on a set of keywords or phrases representing the document's textual content. Therefore, the indexing consists in detecting the most representative terms of the document's content. There are different methods existing in IR indexing: manual, automatic or semiautomatic indexing.

- **Manual Indexing:** An expert in the field chooses the terms that he considers relevant in describing the semantic content of the document by using

a controlled vocabulary (hierarchical list, thesaurus, dictionary...).

- **Automatic Indexing:** This type of indexing does not involve expert. The indexing process is completely automated based on algorithms associating descriptors to documents.

- **Semiautomatic Indexing:** It is a combination of the two previous methods: the automatic process allows the extraction of the document's terms. However, the final choice of descriptors is left to the expert, which uses a controlled vocabulary or thesaurus. The task of indexing involves not only deriving which parts and keywords are extracted from the documents in the collection, but also a complex computational process: Tokenization, Stopwords removal, Stemming and Weighting of words.

1.3.1.1 Tokenization

Given a character sequence and a defined document unit. Tokenization is the task of chopping it up into pieces, called tokens, using a simple lexical analysis to identify words by recognizing spaces separating words, special characters, numbers, punctuation, etc. Here is an example of tokenization:

Input: Friends, Romans, Countrymen, lend me your ears;
Output: Friends/ Romans/ Countrymen/lend/ me/your/ ears

These tokens are often loosely referred to as terms or words, but it is sometimes important to make a type token distinction. A token is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing. A type is the class of all tokens containing the same character sequence. A term is a (perhaps normalized) type that is included in the IR system's dictionary.

1.3.1.2 Stop Words Removal

Sometimes, some extremely common words which would appear to be of little value in helping select documents matching a user need are excluded from the vocabulary entirely. These words are called stop words. The stop words removal refers to eliminating very frequent terms, which are mostly prepositions and articles as "in", "a", "the", "to" or pronouns like "I", "he", "she", "it" and so on. In fact, these words are likely a good indicator of poor content and therefore can be filtered out. The general strategy for determining a stop list is to sort the terms by collection frequency (the total number of times each term appears in the document collection), and then to take the most frequent terms, often hand-filtered for their semantic content relative to the domain of the documents being indexed, as a stop list, the members of which are then discarded during indexing.

Traditionally, stop words are chosen from a fixed list, which is constructed from lexical information. This has led to the appearance of different stop words lists which vary in size, like Smart's list, (Buckley et al., 1995), comprised 571 terms, or Okapi's one (Robertson et al., 2000), made up of 220 terms. An example of a stop list is shown in Figure 1.2.

a	an	and	are	as	at	be	by	for	from
has	he	in	is	it	its	of	on	that	the
to	was	were	will	with					

Figure 1.2: A stop list of 25 semantically non-selective words which are common

1.3.1.3 Stemming and Lemmatization

For grammatical reasons, documents are going to use different forms of a word, such as organize, organizes, and organizing. Additionally, there are families of derivationally related words with similar meanings, such as democracy, democratic, and democratization. In many situations, it seems as if it would be useful for a

search for one of these words to return documents that contain another word in the set.

The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. For instance

$$\text{am, are, is} \longrightarrow \text{be}$$
$$\text{car, cars, car's, cars'} \longrightarrow \text{car}$$

The result of this mapping of text will be something like:

$$\text{the boy's cars are different colors} \longrightarrow \text{the boy car be differ color}$$

However, the two words differ in their flavor. Stemming refers to the conflation of words to their lemmatized base form, where the morphological variants of words are stripped to one single suffix entry. Thus, "computer", "computing", "compute" is reduced to "compute". This process employs shallow linguistic information (e.g., removing endings of nouns), as well as language pattern matching rules (Porter, 1997).

Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma. If confronted with the token saw, stemming might return just an s, whereas lemmatization would attempt to return either see or saw depending on whether the use of the token was as a verb or a noun. The two may also differ in that stemming most commonly collapses derivationally related words, whereas lemmatization commonly only collapses the different inflectional forms of a lemma.

1.3.1.4 Weighting

The weighting is the final step in most of the information Retrieval indexing applications. The weighting models are used to assign weights to terms. They include local weighting, global weighting or both.

- **Term Frequency (TF):** If local weights are used, then term weights are generally expressed as term frequencies TF (Manning et al., 2008). The underlying idea is that the more frequent a term is, the more important becomes in the document's representation. Let assume that dj is the document and ti is the term, then the frequency TFij of the term in the document is often used directly or expressed as one of the following sorts. Where tdij is the number of occurrences of term t in the document dj. The denominator is the number of occurrences of all terms in the document dj. The last variation normalizes the term frequency to avoid bias related to the length of the document.

$$\mathrm{TF_{ij}} = 1 + \log \mathrm{td_{ij}} \tag{1.1}$$

- **Inverse Document Frequency (IDF):** Using the global weights, the weight of a term is specified by IDF values (Manning et al., 2008). The main idea is to measure the overall representativeness of the term with respect to the collection of documents. However, rare terms have high IDF, contrary to frequent terms. Where N is the size of the collection (number of documents) and ni is the number of documents containing the term i.

$$\mathrm{IDF_i} = \log \frac{\mathrm{N}}{\mathrm{n\ i}} \qquad ; \qquad \mathrm{IFD_i} = \log \frac{\mathrm{N} - \mathrm{n_i}}{\mathrm{n_i}} \tag{1.2}$$

- **Term Frequency-Inverse Document Frequency (TF-IDF):** In this type of weighting, both the local and global weights are used. This is commonly referred to as TF-IDF weighting (Robertson and Sparck Jones, 1988):

$$\mathrm{TF{*}IDF{=}\log\,(1{+}TF)\,{*}IDF} \tag{1.3}$$

1.3.2 Matching

The IRS includes a research process for selecting the information deemed relevant to the user. It involves a process of user's interaction with the IRS which includes the following steps: First, the user expresses his information need under the shape of a query. Second, the system creates the query's index which will be compatible with the model of documents' index. Finally, the system evaluates the match between the query and the documents.

To achieve the previous step, the IRS calculates a measure of similarity between the query and the indexed descriptors of the document's collection. Only the documents whose similarity exceeds a predefined threshold are selected by the IRS. This function of similarity uses the index generated in the indexing phase in order to calculate a similarity score denoted RSV (Q, D) (Retrieval Status Value) where Q is a query and D is a document in the collection. The matching process is closely linked to the process of indexing and term weighting. In fact, there are two matching models:

- **Exact-match retrieval model:** The result is a list of documents respecting exactly the submitted query according to specific criteria. However, the documents returned are not sorted (Salton, 1971).
- **Best-match retrieval model:** The result is a list of documents supposed to be relevant to the query. The returned documents are sorted according to the relevance score(Robertson and Sparck Jones, 1988).

1.3.3 Relevance Feedback

An essential component of any retrieval system is the feedback mechanism. The relevance feedback is the process of judging whether a search result is relevant to the user's query or not. Once these judgments are available, the retrieval system is able to learn from the judged samples to improve the search accuracy.

The basic idea is that relevant and non-relevant search results present strong

evidence about the user's intention. We can distinguish between three types of feedback, namely explicit feedback, pseudo feedback, and implicit feedback.

1.3.3.1 Explicit Relevance Feedback

In explicit relevance feedback, the user explicitly marks (some) search results as relevant or irrelevant. User studies have shown, however, that users are often unwilling to make such judgments, raising concerns about the practical value of relevance feedback.

1.3.3.2 Pseudo Relevance Feedback

The pseudo relevance feedback simply assumes some of the top ranked documents to be relevant and thus does not require the user to explicitly mark the search results.

1.3.3.3 Implicit Relevance Feedback

In implicit relevance feedback, the system tries to automatically infer what is relevant and what it is not from the user's interactions with the search results. This mainly relies on heuristics, such as assuming that the search results clicked by the users are relevant and those non-clicked are irrelevant, even if the latter are ranked higher than the clicked but irrelevant ones. To overcome noise behavior and to make the inference more robust, other measures can be included to infer relevance of a search result, for example how long the user spends viewing a search result, or how much scrolling takes place on the result page.

1.4 Information Retrieval Modelling

Information Retrieval (IR) is concerned with identifying documents in a collection that best match a description of a user's information need. Research in this field spans many subdisciplines of computer and information science. However, central to any effective retrieval system is the identification and representation of

document content. A retrieval model specifies the details of the document representation, the query representation, and the matching function.

The models are important in the IR process. There are many good reasons for having models of information retrieval. In fact, they allow guiding search and providing a formal description of the IR process and a theoretical framework for modeling the relevance measure. In addition, models can serve as a blueprint to implement an actual IR system. According to (Baeza-Yates and Ribeiro-Neto, 1999), an IR model is defined by a tuple (D, Q, F, R (q, d)) where:

- D is a representation for documents.
- Q is a representation for queries.
- F is a modeling framework for D, Q, and the relationships among them.
- R (q, d) is a relevance or similarity function which ranks the documents with respect to a query.

Many retrieval models have been proposed as the basis of text retrieval systems. Three main classes of retrieval models are in current use: Boolean models which form the basis of most commercial retrieval systems, vector space models which view documents and queries as vectors in a high dimension vector space and use distance as a measure of similarity, and probabilistic models which view retrieval as a problem of estimating the probability that a document representation matches or satisfies a query. All these methods use term distributions within the documents to decide on the relevance to a query.

1.4.1 The Boolean Model

The Boolean model is the first model of information retrieval which is based on set theory and Boolean algebra. We can distinguish the pure Boolean model (Boolean model), the extended Boolean model (extended Boolean model) and the model based on fuzzy sets (fuzzy set model). It is probably the most criticized model. In

this model, each document is represented by a logical conjunction of unweighted terms which represent the index of the document. Indeed, the Boolean model considers that the index terms are present or absent in a document. Hence, the weights of terms in the index are binary, i.e either 0 or 1. Using the operators of George Boole's mathematical logic, query terms and their corresponding sets of documents can be combined to form new sets of documents (Frakes and Baeza-Yates, 1992). Boole defines three basic operators, the logical product called AND, the logical sum called OR and the logical difference called NOT. The result of this function is a binary score written as follows: RSV (qi, dj) = {1, 0 }.

An example of a Boolean query may be seen in Figure 1.3. Here the information need is for documents which are relevant to the three terms; 'information', 'social' and 'Networks'. It is important to note that in the example query given, the documents which will be returned (those corresponding to the shaded area in Figure 1.3) are those which contain all the desired query terms.

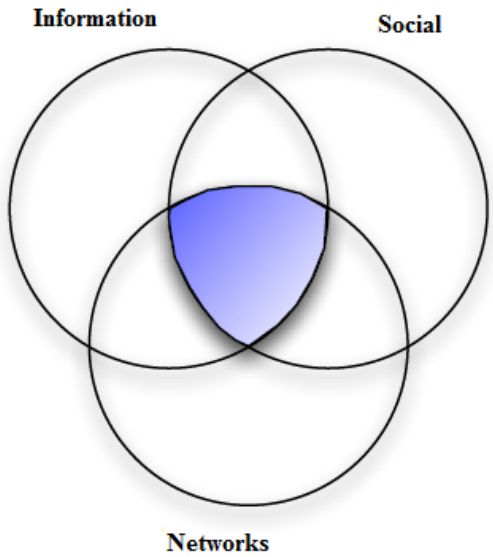

Figure 1.3: Diagram of a Boolean query

The Boolean model has much strength such as being a very simple model based on sets which makes it easy to understand. An advantage of the Boolean model is that it gives (expert) users a sense of control over the system. It is simple to understand why a document has been retrieved in response to a given query. However, for untrained users, the model has a number of disadvantages. Its central disadvantage is that it does not provide a ranking of retrieved documents. The system has to make decisions because the model either selects a document or not. Furthermore, users find hard the construction of a successful Boolean queries with AND, OR or NOT that have a different meaning when used in a query based on natural language.

To remediate these disadvantages, the extended Boolean model(Salton and McGill, 1986) has been proposed. It takes into account the importance of terms in the representation of documents and in the query, by assigning weights to each term in the document and the query. (Losee, 1998) has shown, using sorting that the extended Boolean model is a special case of the probabilistic search.

1.4.2 The Vector Space Model

Gerard Salton and his colleagues suggested a model based on Luhn's similarity criterion that has a stronger theoretical motivation (Salton and McGill, 1986). The basic idea of the Vector Space Model (VSM) is to represent documents and queries as vectors in a multidimensional space, which dimensions are the terms used to build an index representing the documents. This model is based on the mathematical foundations of vector spaces. The query and the document substitute are compared by emphasizing the similarity between their vectors, such as using the cosine similarity measure. This measure computes the cosine of the angle formed by the two vectors $\vec{d}$ and $\vec{q}$, respectively the document vector and the query vector. The cosine of an angle is 0 if the vectors are orthogonal in the multidimensional space and 1 if the angle is 0 degrees. The cosine formula is given by:

$$\text{score}\left(\underset{\text{d}}{\longrightarrow};\ \underset{\text{q}}{\longrightarrow}\right) = \frac{\sum_{\text{k}=1}^{\text{m}} \text{dk . qk}}{\sqrt{\sum_{\text{k}=1}^{\text{m}} (\text{dk})^2}\ \sqrt{\sum_{\text{k}=1}^{\text{m}} (\text{qk})^2}} \tag{1.4}$$

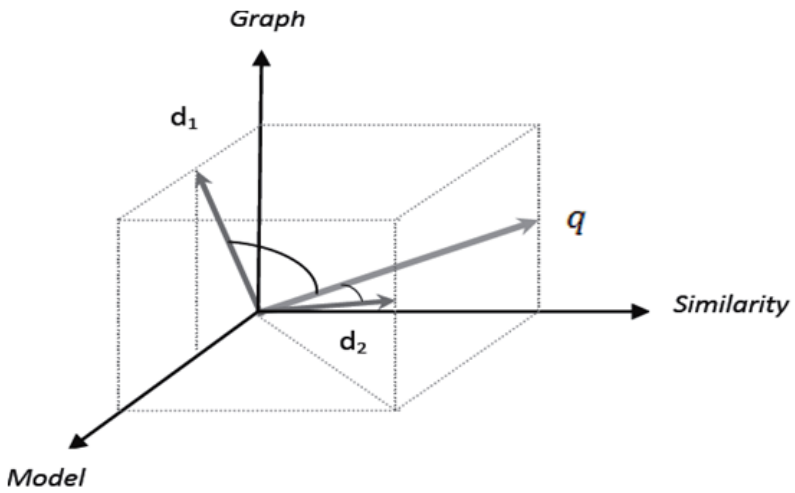

Figure 1.4: A query and document representation in the vector space model

In fact, the more similar two vectors are, smaller the angle is and larger the cosine of this angle is. The VSM assigns a high ranking score to a document that contains only a few terms belonging query, if those terms occur occasionally in the collection but frequently in the document. This model makes the following assumption: The more similar a document-vector to a query-vector, is the more relevant document to that query.

Unlike the Boolean model, the matching function evaluates a best-match between a document and a query, allowing users to find documents that satisfy approximately the query. Furthermore, the results can be ranked in decreasing order of similarity. However, this model assumes that the terms used to define the dimensions of the space, are orthogonal or independent, which is not realistic because relations among terms may exist.

The vector model is similar to the probabilistic model in different aspects, excepting that it lacks a sound theoretical basis (Turtle and Croft, 1991). Moreover,

it sorts the documents using to a measure of similarity to the query instead of a probability of relevance of documents as is the case in the probabilistic model. Although, the vector model is criticized as an "ad hoc" model, it is one of the classic models RI, the most studied and well accepted.

1.4.3 The Probabilistic Model

The probabilistic model aims to capture the IR problem in a probabilistic framework. The Probability Ranking Principle, defined by Robertson and Sparck Jones (1988), is used in this model as it states that an information retrieval system is supposed to rank the documents based on their probability of relevance to the query. The fundamental idea is as follows: Given a query q and a collection of documents D, a subset R of D is assumed to exist because it contains exactly the relevant documents to q (the ideal answer set).

This model uses the statistical distribution of the terms in both the relevant and non-relevant documents. For a given query Q, a document D is selected, if the probability that D is relevant, denoted by p $(\bar{R}|D)$, is greater than the probability that D is not relevant, denoted by p $(R|D)$ and $\bar{R}$ is the irrelevance event. The similarity score RSV (Q, D) is given by:

$$\text{RSV}\ (Q,\ D) = \frac{p(R|D)}{p\ (\overline{R}|D)} \tag{1.5}$$

This model has the followings advantages: First, the retrieved documents are ranked based on their relevance. Thus, it allows users controlling the results according to two options: fixing a relevance threshold, or limiting the number of documents to display. Second, the users can use the natural language when formulating a query rather than using a specific one. However, this approach has some shortcomings: the need to guess the initial separation of documents into relevant and non-relevant sets. In addition, the queries have to contain a large number of words to improve the retrieval performance. Moreover, the users need to choose

the most expressive words that also exist in the relevant documents. Furthermore, the fact that the classic model does not take into account the frequency of index terms inside documents (i.e., all weights are binary).

Despite these shortcomings, variations of the probabilistic model have led to the development of one of the most successful ranking models, BM25 (Robertson and Jones, 1976; Jones et al., 2000). The first system that was implemented to incorporate this function, was the Okapi information retrieval system, implemented at London's City University in the 1980s and 1990s. This ranking methodology takes into account the present/absence of relevant information and incorporates a document-specific component, which measures term frequencies and documents lengths.

1.5 Utility, Relevance, and IR System Performance

Utility and relevance underlie all IR operations. In order to evaluate the performance of an IR system we need to measure three things, topical relevance, pertinence, and novelty. A document is topically relevant for a query, a question, or a task if it contains information that either directly answers the question or can be used, possibly in combination with other information, to derive an answer or perform the task. In addition, it is pertinent with respect to a user with a given purpose if it gives just the information needed; is compatible with the user's background and cognitive style so he can apply the information gained; and is authoritative. It is novel if it adds to the user's knowledge. In the literature, the term "relevance" is used imprecisely; it can mean utility or topical relevance or pertinence. Many IR systems focus on finding topically relevant documents, leaving further selection to the user. From relevance assessments, we can compute measures of retrieval performance such as:

$$\text{Recall(R)} = \frac{\text{relevant items correctly retrieved}}{\text{all relevant items in the collection}} \tag{1.6}$$

$$\text{Precision(P)} = \frac{\text{relevant items retrieved}}{\text{all items retrieved}} \tag{1.7}$$

$$\text{F} - \text{measure} = \frac{2\text{PR}}{\text{P} + \text{R}} \tag{1.8}$$

Commonly, evaluation studies use recall and precision or their combination, whether these are the best measures is debatable. With low precision, the user must look at several irrelevant documents for every relevant document found. More sophisticated measures consider the gain from a relevant document and the expense incurred by having to examine an irrelevant document. For ranked retrieval document, performance measures are more complex. All of these measures are based on assessing each document on its own, rather than considering the usefulness of the retrieved set as a whole. For example, many relevant documents that merely duplicate the same information just waste the user's time, so, retrieving fewer relevant documents would be better.

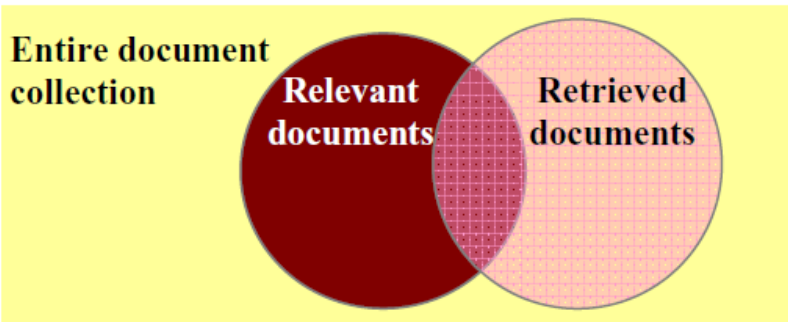

Figure 1.5: Retrieved versus relevant documents

Figure 1.5 shows the distribution of the retrieved documents versus the relevant documents. In the upper part of the figure, the intersection of the two circles is the part that needs to be maximized by an IR system. In the lower part of the

Figure, the number of documents that need to be maximized is in the lower left corner and the upper right corner. The other two corners would contain zeros for an ideal IR system.

1.6 Problem Description

Social Information Retrieval consists on incorporating information from social network in the Information Retrieval process. The nature of this technology makes an easy and popular way to communicate information to either a select group of people or to a much wider audience. A change in peoples' interaction with information came in the early 2000's. This decade saw the development and maturation of social media and online environments that allow people to interact in novel ways.

In recent years, with the arrival of Web 2.0, user-generated content have realized a significant increase. The popularity of user-generated content, particularly in the context of social media that has spawned a host of novel IR problems (King et al., 2009). One issue that makes microblog retrieval a compelling research area is its youth. Even the most basic problems in microblog IR remain to be identified. Among these basic problems there are defining useful units of retrieval. Given a corpus of tweets and a user with a particular information need, what should an IR system present to the searcher?

Facebook, Twitter, blogs, wikis and other social media tools give people the ability to communicate, interact and collaborate with each other, on a scale and in ways that they can't interact with traditional media. 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. These online groups are made up of those who share passions, beliefs, hobbies, or lifestyles. The ability of users to interact with each other, reply to messages from others and create conversations, it forms a complex reality that is interesting. Nevertheless, using social information to retrieve conversation is a new phenomenon and constitutes a hard task. The use of the

Social Networks give a new global importance as an additional incentive to find a mechanism for retrieving information that is not limited to pure research on text messages.

Companies are getting interested in joining social media as well. Obviously, they obtain yet another way to promote their product. Track a brand's tweets to see the type of conversations, the "matter" of conversations, how frequently they interact, and make assessments. This is certainly something that upcoming start-up businesses might love to tap into especially because it gives the kind of insight about how a particular company's social media marketing allows it to be successful. Furthermore, social media reveals customer preferences giving an opportunity to reduce advertisement costs, to receive immediate product feedback, and most importantly to alter business strategy. However, raw social data hardly meet expectations and fulfill needs of both humans and companies. This happens simply because of the amount of the information, which cannot be handled manually in any reasonable amount of time, and data's noise level. This is why automatic analysis of massive unstructured text data receives a lot of attention from the scientific community and support from the industry.

Discussion retrieval is a challenging task. In addition, conversations still emerge between dyads and between groups. The medium of online conversation allows for sharing ideas, asking questions or discussing issues and solutions interactively along with others. It is an age-old communications practice that helps cultivate creativity, exploratory ideas, perspectives and experiences to take better decisions individually or collectively in the process. The problem of extracting information from Social Networks Sites (SNSs) has already been addressed in the past because of its theoretical and practical relevance. However, Social Network Sites are more than just collections of text messages, making the retrieval process not trivial and requiring specific concepts and techniques. In particular, there is a plethora of tools that can be used to monitor the usage of keywords or tags. However, these tools work on simple text collections, and do not consider their structure. Some services, like Twitter, provide search functionality based on keywords, but it is not

powerful enough if the goal of the retrieval is to get tweets about an event or a product. In general, there is little work on high recall retrieval of messages relevant to a discussion. In the case of information retrieval in microblogs, we found that most of the existing works only separate tweets or tweets with same hastags are returned. Hashtag based search is widely used for data retrieval from social media. This work shows limitations of this approach, because the majority of the relevant messages do not even contain any hashtag, and unpredictable hashtags are used as the conversation evolves in time. In addition, when looking for information, one may obtain the entire context of given microblog. In other words, it will be more helpful when retrieving to present the entire conversation, that has been published and not isolated tweets.

The conversation retrieval problem that information retrieval (IR) tools are confronted with, more specific questions are stated in our work:

Which approach can we use for online conversations building? In particular, how can we build a model which allows the retrieval of millions of conversations generated on the popular Twitter social network site? How can we understand the conversation graphs that are formed when users in the network discuss a particular topic?

1.7 Conclusion

In this chapter we give a briefly review of IR technologies, theories and systems. The architecture of a general IR system introduce, with special emphasis on the different steps involved in the IR task: indexing, query processing, searching and ranking. Finally, we discuss our thesis problem.

Before presenting the conversation retrieval concept from social media, we will study in the next chapter the social information retrieval background.

Chapter 2

Information Retrieval In Social Media

2.1 Introduction

With the current development of the World Wide Web, termed as "Web 2.0" by Oreilly (2007), more people are increasingly well connected on the internet through some social network sites. It is an ongoing trend that people increasingly reveal very personal information on social network sites in particular and in the World Wide Web in general (Boyd and Ellison, 2007; Donath and Boyd, 2004; Joinson, 2001; Gross and Acquisti, 2005). Some popular Web sites such as Facebook[1], Flicker[2], YouTube[3], MySpace[4] and Windows Live provide virtual platforms for all users to share their ideas, photos, videos and music, and all users are free to post comments and communicate with each other. This environment presents clear challenges to conventional information retrieval research, which is based mainly on individual activities. Methods for retrieving information effectively and efficiently for social media have become an important research direction for modern

[1] www.facebook.com
[2] www.flickr.com
[3] www.youtube.com
[4] www.myspace.com

information retrieval.

In this chapter, we begin with introducing the social information retrieval. The following section expounds the evolution of Web 2.0 and identify its different services. Then, we present the information retrieval in microblog. After that, we are interested in the available data and new ways of interactions of users in social web. Finally, we describe the Twitter social network site background.

2.2 Information Retrieval and Social Analysis Area

(Wenger, 1998) introduced the idea of the 'community of practice': the notion that a person can satisfy his information needs more efficiently if he is embedded in a community of practitioners with similar interests and problems. Indeed, before the advent of modern information retrieval systems, most information needs were satisfied by social means: by asking friends and acquaintances, by going to the library and asking the librarian for help, or by enquiring at specialized agencies.

It is an ongoing trend that people increasingly reveal very personal information on social network sites and in the World Wide Web in general (Boyd and Ellison, 2007; Donath and Boyd, 2004; Joinson, 2001; Gross and Acquisti, 2005). As this information becomes more and more publicly available in these various social network sites and the Web in general, the social relationships between people can be identified. This enables the automatic extraction of social networks.

Information retrieval meets the social realm at another, more subtle point: Information is also produced in social situations. Few authors work in a social vacuum. Participation in the community and active exchange with like-minded persons fosters information production and improves the quality of the work.

(Granovetter, 1973) notes that 'weak ties-ties' between acquaintances rather than between close friends or family are particularly important for information dissemination and diffusion: Weak ties allow information to spread from one closely-knit community to another. Individuals with many weak ties 'hubs' in the social network are important for the adoption of new ideas, since their authority is ac-

cepted by a large number of immediate acquaintances.

We conclude that social networks are an important factor for finding and spreading information, and that an individual's position in the social network of his peers is indicative of his authority and influence. Accordingly, we define social information retrieval as the incorporation of information about social networks and relationships into the information retrieval process.

2.2.1 Web 2.0: Social Field Evolution

Web 2.0 is a term introduced by Oreilly (2007) as an attempt to describe new practices in comparison to the conventional Web practices. Web 2.0 is the term used to describe a variety of Web sites and applications that allow anyone to create and share online information or material they have created. A key element of the technology is that it allows people to create, share, collaborate and communicate. In addition, Web 2.0 which is also mentioned as Social Web due to its plethora of social dimensions, has become an issue of high interest among Web experts and Web users, and it is considered to be the next generation of World Wide Web. The basic asset of Web 2.0 is its strong focus on the user satisfaction.

Web 2.0 has an intense social dimension, which allows users to create networks of trusted users within the same areas of interest. In addition, to that, Web 2.0 differs from other types of Websites as it does not require any Web design or publishing skills to participate, making it easy for people to create and publish or communicate their work to the world.

The nature of this technology makes it an easy and popular way to communicate information to either a select group of people or to a much wider audience. There are number of different types of Web 2.0 applications such as Wikipedia[5], YouTube, Facebook, MySpace, Flickr. Thus, the result is a data intensive and annotated Web. Despite of its preliminary and unstructured state, it is already an advance compared to the conventional Web. This is due to the current Web 2.0

[5] http://en.wikipedia.org

sites, which are annotated with a primitive "semantic layer", even if this is weak and uncontrollable.

Web 2.0 has come to dominate almost every aspect of debate and discussion of the contemporary practices, technologies, and meanings of the Internet. Conversation is the informal expression of views and exchange of ideas through people talking among themselves. It is a social skill and the defining characteristic of social media. Conversations allow people with different information to learn from each other. Facebook, Twitter, blogs, wikis and other social media tools give people the ability to communicate easily with each other, on a scale and in ways that they can't with traditional media.

Collaboration is the underpinning of social media online and, as can be observed in news of political movements, frequently links offline to the world's hotspots. Social networking sites on the Web let people communicate, interact and collaborate with friends, family, neighbors, coworkers, associates, colleagues, customers, teachers, students and other acquaintances.

Creation of content is relatively easy for people using digital media tools. The communication philosopher Marshall McLuhan told us: "The content of a new medium usually is built on and expands from the content of an older medium". Social media incorporate traditional media technologies and the digital culture surrounding social media freely borrows from traditional media to create new content, which in turn frequently muddles the ownership of intellectual property.

The choice of media content abounds on social network sites where people use a variety of ways to interact. Through search engines, recommendations from friends, RSS feeds, and both traditional and social media, people today are proactive and successful in locating what they seek.

2.2.2 User-generated Content (UGC)

A change in peoples' interaction with information came in the early 2000's. Web 2.0 is currently one of the most widely used buzz word in the Internet industry. This

term describes a new generation of Web interfaces enabling users to read, share and write content over the Web. User-generated content (UGC) like Facebook or YouTube are challenging the traditional media. The nature of information sharing has begun to change also. Instead of sharing information on a one-to-one basis such as e-mail, there is an increasing trend towards the one-to-many style of publication. This decade saw the development and maturation of social media, online environments that allow people to interact in novel ways.

The social media sites offer users the ability to share information with close friends, family, or the Web at large. This information may be in the form of video, audio, images or text, and the amount of content being shared continues to grow. With respect to information science, a key development brought by these media was the proliferation of user-generated content. In contrast to early Web resources, newer services (characterized by the ubiquitous label Web 2.0) allow and often rely upon their users to create, edit, and mediate access to information.

2.3 Information Retrieval on the Web

Searching information is commonly an individual task which aims to solve any information need. To achieve this task, one may go to a library, or go surfing the Web in order to find relevant information. Indeed, due to the large amount of available documents, the Web has become a favorite information source for solving daily information needs. An issue remains: the Web is in perpetual evolution; so the problem is less the existence of relevant information rather than the way users find it. One may compare searching for information on the Web with "looking for a needle in a haystack".

For the information retrieval (IR) community, the Web now presents a new paradigm, while also generating new challenges and attracting growing interest from around the world. An important example of these challenges is managing huge text collections and evaluating the usefulness of hyperlinks contained within them. Furthermore, information seeking on the Web is much more diverse and

unpredictable than in traditional IR contexts. Consequently, determination of the applicability of traditional IR methods to Web IR should begin with a consideration of the contextual differences between traditional IR and Web IR research. The findings of traditional IR are mostly based on experiments conducted with small, stable document collections and sets of relatively descriptive and specific queries on the other hand. Web information retrieval must deal with mostly short and unfocused queries posed against a massive collection of heterogeneous, hyperlinked documents that change dynamically. Thus, one focus of Web IR research has been the characteristics of the Web search environment (notably, document characteristics and searcher behavior).

2.4 Information Retrieval in Microblogs

The Microblogs are services that reflect social relations among people, such as user communities and common interest groups, and allow users to share content, through typically short, but informative, text messages (Rocha et al., 2011), which may include links to images, videos or Web pages.

Information retrieval in microblogs was introduced for the first time in the work of (Java et al., 2007). This work focuses on the analysis of data and the structure of social networks. More recently, some works such as (Teevan et al., 2011) have addressed the search tweets as a problem independent research differs from traditional Web.

Compared to Web search (Java et al., 2007) notes that research in microblogging provides information brief concise and accurate on a current topic. It can also receive real-time information about an event that has just occurred.

2.5 Social Networks

Social networks, a very old and pervasive mechanism for mediating distal interactions among people, have become prevalent in the age of the Web. With interfaces

that allow people to follow the lives of friends, acquaintances and families, the number of people on social networks has grown exponentially since the turn of this century.

For instance, MySpace, Facebook, and Twitter have attracted millions of users, many of whom have integrated these sites into their daily practices. Such networks provide an extremely suitable space to instantly share multimedia information between individuals and their neighbors in the social graph. Social networks provide a powerful reflection of the structure and dynamics of the society of the 21st century and the interaction of the Internet generation with both technology and other people. As of this writing, there are hundreds of Social Networks Sites (SNSs), with various technological affordances, supporting a wide range of interests and practices. While, their key technological features are fairly consistent, the cultures that emerge around Social Network Sites(SNSs) are varied. Indeed, the dramatic growth of social multimedia and user generated content is revolutionizing all phases of the content value chain including production, processing, distribution and consumption. It also originated and brought to the multimedia sector a new underestimated and now critical aspect of science and technology: social interaction and networking. The importance of this new rapidly evolving research field is clearly evidenced by the many associated emerging technologies and applications including online content sharing services and communities, multimedia communication over the Internet, social multimedia search, interactive services and entertainment, health care and security applications. A new research area has been generated called, social multimedia computing, in which well established computing and multimedia networking technologies are brought together with emerging social media research.

2.5.1 Microblogging

Microblogging is a new phenomenon defined as "a form of blogging that lets you write brief text updates (usually less than 200 characters) about your life on the

go and send them to friends and interested observers via text messaging, instant messaging (IM), email or the Web." [6]

According to (Java et al., 2007), a microblogging service is a new communication medium for information broadcasting and network collaboration. Unlike other social networking services, microblogging posts are particularly short and submitted in real-time to report an actual topic.

Several microblog services exist. For example, Facebook, Tumblr[7], Orkut[8], and FourSquare[9] all offer microblog-type services. However, at this time the most visible microblogging platform is Twitter. Microblogs have many specificity. On the one hand, people's motivations for microblog search are diverse (Teevan et al., 2011): some are similar to those of Web search (e.g., "astronomy or science stuff"), while many users reported an interest in searching Twitter to find timely information. Many users having looked for information related to news (e.g., "technology news, trends"), topics gaining popularity (e.g., "information on a currently trending topic"), or summaries of events colleagues were attending (e.g., "event hashtags"). Another type of timely information sought, mentioned by Twitter users, was real-time information. For example, looking for regional/local information ("police incident, weather, etc"). Others want reports of traffic (e.g., "traffic jam") or the status of online services (e.g., "down services") or social information (e.g., people with similar interests), are relatively specific to microblog search.

On the other hand, microblogs are different from traditional Web. Tweets (i.e., microblogs in Twitter), for example, cannot exceed 140 characters. They are usually composed of a single sentence. Users publish microblogs on a wide variety of topics in real time. Thus, the messages contain mistakes and some are written in netspeak. Besides, some microblogging platforms like Twitter use specific syntax (e.g., retweets, hashtags, mentions), that is likely to challenge conventional natural language processing techniques. Finally, microblogs often point to external sources

[6] http://en.wikipedia.org/wiki/Micro-blogging
[7] http://tumblr.com
[8] http://orkut.com
[9] http://foursquare.com

by including hyperlinks to documents or other contents on the Web.

2.5.2 Microblog Conversation Definition

(Magnani et al., 2011) defined conversation as a "tree where nodes represent short text messages posted by a user at specific timestamps in reply to their parent nodes". Based on this definition, (Magnani et al., 2011) detected a conversation using the tweets that directly respond to other tweets. The advantage of this method is to have a coherent conversation based on the direct links between users. However, the downside is that this method eliminates the tweets that does not contain the "@sign" but you can have tweets that are in the conversation even though they are not directly linked with other tweets.

(Cogan et al., 2012) defined a conversation as a reply tree which is a graph where vertices are tweets and a directed edge represents one tweet that is a reply to another. In this case a tweet can only reply to one other tweet and a retweet are ignored, but users can get involved in conversations with other sharing other user's posts (retweet). However, a retweet can be found in a conversation.

We define a conversation as a set of short text messages posted by a user at specific timestamps. This messages can be directly replied to other users by using "@username" (a message is a reply to another) or indirectly by liking, commenting and sharing other users' posts (retweets, comment, anser).

2.6 Twitter

2.6.1 Twitter Background

Twitter, the most popular microblogging service to date, has experienced exponential growth in recent years. Launched in October, this service rapidly gained worldwide popularity with over 500 million registered users[10] as of 2012, generat-

[10] Twitter Passed 500M Users In June 2012, 140M Of Them In US; Jakarta 'Biggest Tweeting' City

ing over 340 million tweets[11] daily and handling over 1.6 billion search queries[12] per day. Twitter is distinguished from similar Websites by some key features. The main one consists on the following social relationship. Twitter a Web-based microblogging service that allows registered users to connect profiles to others, but these connections are directed rather than undirected.

Users write short textual messages "tweets" on different topics that include their current activities and share links to interesting media. A tweet is a short message from a user to her followers. A key property of tweets is that they are restricted to 140 characters and hence are much more concise than blog posts, which is probably the reason why users of Twitter send them more often, and usually on diverse topics. The short allowance on tweet size means that the users must be concise, and sometimes have to resort to phrase abbreviations.

Twitter allows a user to "follows" others and see their updates (tweets). The central feature of twitter, which users see when they log in, is a stream of tweets posted by those that they follow, listed in reverse chronological order. Participants have different strategies for deciding who they follow some follow thousands, while others follow few; some follow only those that they know personally, while others follow celebrities and strangers that they find interesting.

Whenever a user receives a tweet, she/he may retweet the tweet (the ability to forward an interesting tweet to their followers) or alternatively may to reply to the tweet(respond to it by using her own original tweet). Tweets are usually directed towards a general audience, but there are special provisions by which a user can direct a tweet at another user, or send it as a response to an earlier tweet. This is usually done by appending a "@username" to the intended user name and including this keyword towards the beginning of the tweet. Direct messages, are useful as they provide a way of establishing connections between two users, or between two tweets.

Another peculiarity of tweets is that sometimes they contain words

[11] https://blog.twitter.com/2012/twitter-turns-six
[12] https://blog.twitter.com/2011/engineering-behind-twitter E28099s-new-search-experience

prepended by "#" (hash symbols), which are referred to as hashtags that allows do a search on a theme to follow a conversation, or to mark a topic. By including a hashtag in a tweet, the user suggests that the word denoted by the hashtag makes for a good candidate as a search key for the tweet. For example, tweets about the Tunisie revolution in 2011 typically contain the hashtag "#TunisieRevolution", "#Tunisie" or some number of other variations as described in Figure 2.1.

Figure 2.1: Example of Hashtags

Finally, it is important to note that 140 characters does not allow for tweets that are verbose, or those that contain a well developed idea. This is probably the reason that tweets of tentimescontain URL link directs the reader to breaking news, a blog post, or in many cases interesting artifacts such as related Websites, videos, photographs, etc. This fairly simple set of socio-technical rules has made possible for Twitter to host a wide range of social interactions (Boyd et al., 2010) from the broadcasting of personal thoughts to a large public to more structured conversations among groups of friends.

2.6.2 Twitter Conventions

Twitter participants are strained to expressing themselves in 140 characters. Moreover, an increasing number of people are using Twitter to communicate with others, sometimes in extended exchanges. To facilitate their interactions a series of conventions become visible that allowed users to add structure to tweets. Twitter users begin using the novel symbol "@user" syntax to refer to specific users.

This convention stems from an older IRC (Internet Relay Chat) practice and serves multiple purposes in Twitter, including directing messages to specific people (also known as @replies), and to obliquely reference another user. Discussing the appearance of the @sign as an indicator, (Java et al., 2007) indicated that 21% of the users in their study used Twitter for this purpose, and that one-eighth of the messages (12.5%) were part of conversations. Furthermore, (Mischaud, 2007) found that in his sample, that when the @sign appeared in a tweet, it was clear that a fellow Twitter user was being interacted with.

Besides in (Honeycutt and Herring, 2009), the authors point out that the use of "@user" is a form of addressivity, or indicating intended recipients of messages that are posted in an otherwise public forum in order to gain the target person's attention, which is essential for conversation to occur. (Honeycutt and Herring, 2009) also indicate the use of "@user" messages to reference other users. Moreover, it is not obvious that all uses of the "@sign" in tweets are destined to create conversation, or that all Twitter conversations use the "@sign". Additional factors may need to be considered for to determine, to what extent, twittering is conversational, including the users intention behind displaying a "@sign". The breakdown of the functions of "@sign" is summarized in table 2.1.

Retweeting, is another emergent practice. Unlike @replies and hashtags, the conventions for retweeting are hugely inconsistent. Retweets are a way of rebroadcasting one of your followers messages to everyone that follows you. This is a simple way to share content that you find valuable or interesting with your followers. The prototypical way of retweeting is to copy the message, precede it with

RT and address the original author with "@sing".

While these practices were not built into Twitter. This service responds to some user convention by building the collectively established syntax into the system. The "@user" now results in a hyperlink to that user's Twitter page. A special page exists so users can see all the @messages mentioning them in a single page. Users are also provided with a button to "@reply" to any given message which helps track the referent.

Category	**Number**	**%**
Addressivity	1339	90.96%
Reference	80	5.43%
Locational 'at'	11	0.75%
Non-locational'at'	5	0.43%
Email	1	0.07%
Emoticon	1	0.07%
Locational 'at'	11	0.75%
Non-locational'at'	5	0.43%
Addressivity	1339	90.96%
Reference	80	5.43%
Other	7	0.48%
Address/Refer	20	1.36%
Address/Location	3	0.20%
Address/Email	2	0.14%
Address/Other	1	0.07%
Refer/Other	2	0.14%
Total	1472	100.00%

Table 2.1: Functions of the @sign in English tweets (Honeycutt and Herring, 2009)

In this sample, 1472 English tweets employing an @sign were found. Nearly 91% of the "@signs" appearing in the sample were used to direct a tweet to a specific addressee. The next most common use of the "@sign" was to refer to another user. In this sample, 28 users included multiple "@signs" in their tweets for different purposes.

2.6.3 Replies and Retweets on Twitter

- **Replies**

Replying to tweets is the way to remain friendly and have conversations on Twitter. There are two ways to reply to tweets. There is an @reply where you use "@username" in your message. Such replies are public i.e., visible on your Twitter page. If you want to reply privately, you can send a DM (direct message) which is sent only to the recipient like a private e-mail.

A reply is defined as including the @symbol before the username of any Twitter user. Using "@username" in your tweet is a way to mention or reference someone on Twitter.

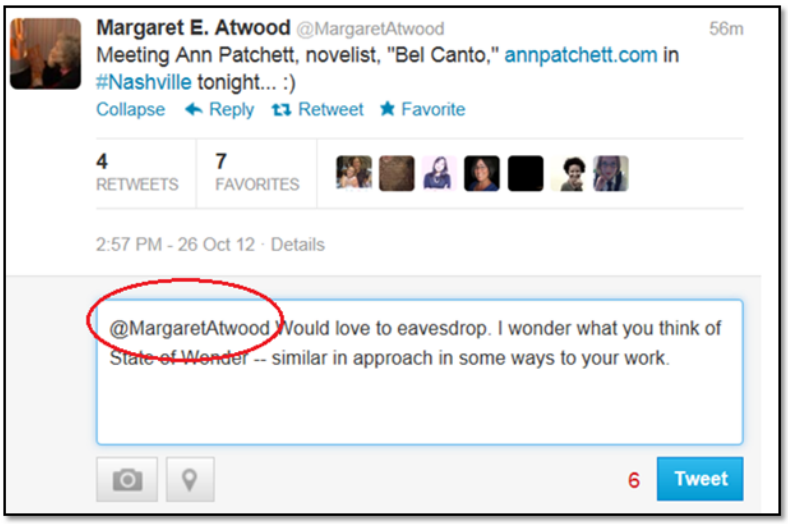

Figure 2.2: Function of @Reply

For example, if you tweet "@ABC Hello World" then you are addressing your messages to the user @ABC. This is the most common way to talk with some back and forth on Twitter, primarily because it is the simplest. Figure 2.2 show the method of @reply.

- **Retweets**

A retweet is a post originally made by one user that is forwarded by another user. This is a simple way to share content that you find valuable or interesting with your followers. These retweets are a popular means of propagating interesting posts and links through the Twitter community.

The first type of Retweetas presented in Figure 2.3 is the official version as released by Twitter. This will post the Tweet to all of your follows exactly as it appears in your news feed. To do this, you use the little Retweet icon at the bottom of the twitter message.

Figure 2.3: The first Retweet's Type

The second type of Retweet presented in Figure 2.4 is the manual version where you actually quote and credit the original author all in the same post. This was the original method for retweeting on Twitter, and still preferred by most users. The benefit of this method is that you are still "credited" in the tweet, which means that the content is still associated with you. The most common method for this is the "RT @username" method. When a tweet generates a reply (@ABC) or a retweet (ABC RT), it suggests the tweet has resonated enough with someone that it sparks a conversation or encourages someone to share it with their followers.

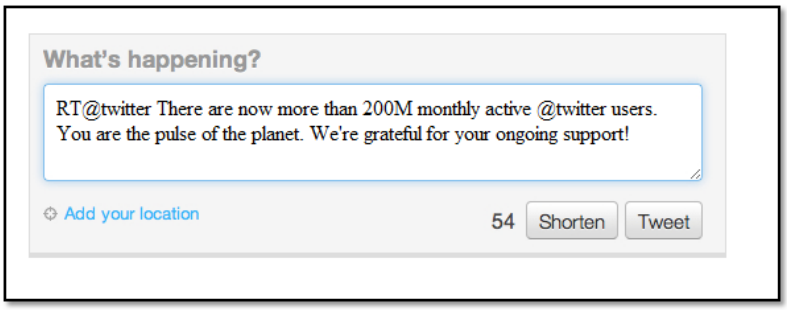

Figure 2.4: The second Retweet's Type

According to the sysoms research 29% of all tweets produced a reaction - a reply or a retweet. From this group of tweets, 19.3% were retweets and the rest replies. This means that of the 1.2 billion tweets they examined, 6%, (or 72 million) were retweets.

2.7 Conclusion

In this chapter, we studied social information retrieval include social realm evolution and the different types of information retrieval, such as on the Web, in microblogs.

Before presenting our contribution that focus on the problem of social information retrieval, we will present in the next chapter the conversation retrieval concept into social media data to better situate our approach contributions.

Chapter 3

Conversational Approach

3.1 Introduction

The concept of conversation retrieval from social media is a novel search paradigm for microblogging sites resulting from the intersection of Information Retrieval and Social Network Analysis(SNA). Furthermore, conversations still emerge between dyads and groups, even though, users can find who share the same passions, beliefs, hobbies, or lifestyles. Therefore, the ability of users to interact with each other, reply to messages from others and create conversations is a new phenomenon but a complex reality. However,before analyzing the data it has to be collected.

This chapter is organized as follows: In the first section, we take into account the fields that are specifically related to the topic of conversation retrieval in microblogging sites. In particular, the aspect of conversation has been specifically treated in works on forum/thread search, where people reply to previous messages in an explicit way. Second, we present more details about the proposed conversation retrieval approach which integrates the social media information. Then, we describe our framework architecture adopted including its different components. Finally, we define the features applied in our conversation retrieval approach.

3.2 Conversation Retrieval From Social Media

The application of the notion of conversation retrieval from social media data is based on very strong established research works. It can be recognized as an intersection of two major fields which are: **Information Retrieval** and **Social Network Analysis**. Two main corpora of related work will be reviewed in this section as following: the relevant information retrieval (IR) literature, the main concepts of Social Network Analysis (SNA) used in our work specifically addressing Twitter, which is the principal field used in our work.

3.2.1 Email Threads Detection

Previous research has been focused on using email structure, and especially emails threads, in several email-based applications (Kerr, 2003; Lam et al., 2002). Thread detection is an important task which has attracted significant attention as well (Klimt and Yang, 2004; Lewis and Knowles, 1997; Rudy, 1996).

Electronic mail (email) is considered one of the most important tools for treating conversations between people. Generally, a typical user mailbox encloses hundreds of conversations. There are few works in the literature that directly address to the problem of thread reply reconstruction. Over the past years, accorded to (Kerr, 2003), the detection of these conversations has been identified as an important task. Clustering the messages into coherent conversations has many useful applications, among them, it gives users the opportunity to see a greater context of the messages they are reading and collating related messages automatically. (Lewis and Knowles, 1997) and recently, (Yeh, 2006) point out that by using text matching techniques to the textual portions of messages, it will be possible to detect threads effectively.

(Cohen et al., 2004) described the concept of 'email speech act' for classifying the intention of an email sender by applying supervised learning to recognize them. However, (Carvalho and Cohen, 2005) proposed in their work a solution consisting at classifying email speech acts by using a dependency-network based on collective

classification method. In addition, (Feng et al., 2006) presented a novel approach to discover the conversation by applying similar collective classification approaches.

(Klimt and Yang, 2004) suggested a method that allows to assemble messages having the same subject attributes and are sent among the same group of people. On the other hand, Conversations may span several threads with similar (but not exact) subject lines. Furthermore, a conversation may not include all the participants in all the messages. (Yeh, 2006) recreated reply chains of emails, called email threads. They suggested two approaches, one based on using header meta-information, the other based on timing, subject and content of the emails themselves. The reconstruction heuristic consists of a number of threshold tests on the features. The main difference is that this approach is specific for emails and the features cannot be easily extended for forum thread reconstruction.

(Shen et al., 2006) introduced in their work a partial solution to the thread reconstruction task for synchronous chat data. In their approach, they modified a single-pass clustering algorithm. This algorithm was amended in order to exploit the temporal information in the streams by computing centroids that update dynamically as the analysis proceeds in a sequential way through the data stream. Identifying subtopic clusters of contributions is the main purpose of this analysis.

In addition, (Cselle et al., 2007) developed an email client extension that makes it possible to clusters messages by topic. The similarity between two messages, as applied by the clustering algorithm, is based on the email message subject, date, participants, and content attributes. However, as they noted, their clustering approach is focused on topic detection, hence messages belonging to different conversations on the same topic will be clustered together.

Currently, (Erera and Carmel, 2008) submitted an approach for conversation detection based on email attributes. They started by sharpening the distinction between email threads and conversations. The task was to assemble messages into consistent conversations by utilizing a function of similarity that takes into consideration all relevant email attributes, for example message subject, participants, date of submission, and message content. Their approach is the most similar to

ours seen that in our approach we will use criteria for detecting tweets that will be in the same conversation.

3.2.2 Summarizing Threaded Conversations

Some research in relation with thread structure has been accomplished in the context of email summarization process, where recovering thread structure is essential for constructing coherent summaries of email conversations. All of these researches resume the conversations through defining a similarity on posts, typically content, and then clustering similar posts into segments of conversations.

The proposition of (Muresan et al., 2001) was to work on summarizing individual email messages using machine learning approaches in order to learn rules for salient noun phrase extraction. In addition, (Lam et al., 2002) addressed also to the problem of email summarization using the thread structure of email conversation. They exploited common features such as named entities and dates. They summarize the message only, though the content of the message to be summarized is "expanded" using the content from its ancestor messages.

The problem of summarizing archived discussion lists was the topic that (Newman and Blitzer, 2003) focused on. The authors start by clustering messages into topic groups, and thereafter they extract summaries for each cluster. The summary of a cluster is extracted using many features, such as lexical similarity of a sentence to cluster centroid, a scoring metric based on sentence position, and a feature based on quotation.

According to (Dalli et al., 2004), FASIL is an email summarization system for use in a voice-based Virtual Personal Assistant. The ranking function that the system utilizes is aimed to rank original sentences in individual email messages, and select the top required number of sentences to generate an extractive summary of the email. To accomplish these tasks, the function uses the occurrence of named entities and other empirically determined parameters.

(Rambow et al., 2004) carry out a machine learning approach to email summa-

rization. In their approach, RIPPER is used as a classifier wich helps to determine which sentences should be included in a summary. Many features are used for learning including linguistic features, features describing the email as well as the threading structure. Such an approach requires a large number of positive examples and cannot produce summaries with varying length based on the user's request. In fact, how this approach can handle hidden emails is still unclear.

The proposition of Zhou and Hovy (2005) consists in summarizing threaded discussions in an equivalent fashion to multidocument summarization. However, their work does not take into consideration afterwards the relative importance of different messages in a thread.

Yeh (2006) represent in their study how to use quoted messages, string matching metrics, and email headers in order to reassemble the tree like thread structure of email conversations. As an inspiration from Yeh et al.'s work, (Carenini et al., 2007) suggested a new email summarization framework, ClueWordSummarizer. Their approach is based on using what they refer to as a fragment quotation graph for organizing conversations and subsequently constructing lexical chains.

(Elsner and Charniak, 2010) used logistic regression to gather utterances in a chat room conversation to separate topics, while (Wang and Oard, 2009) probabilistically expanded the messages by information about their context.

In summary, all these works in topic detection and tracking and discourse disentanglement are successful both in summarizing and detecting topic segments, but this is a different task from reply structure reconstruction, as posts can be replies to each other but not related to each other.

3.2.3 Forum Thread Search

Online forums can be recognized as one of the popular ways which helps in finding useful information on the Web. An online forum is a Web application for holding discussions and posting user generated content in a particular sector, such as sports, recreation, techniques, travel, etc. In forums, conversations are repre-

sented as sequences of posts, or threads, where the posts reply to one or more earlier posts. There have been few previous researches dealing specifically with search in online forums. 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 previous works on searching forum threads that dealt with the structure of reply-chains or reply-trees. (Wang et al., 2008) has concentrated recently on identifying the structure of a thread when explicit connections between messages are missing. Despite the fact that replies to posts in microblogging sites, are commonly explicit. This work is pertinent because since analyzing real microblog conversations, it is proved that different autonomous conversations may be developed inside the same thread of replies. Furthermore, distinct threads may belong to related macro-conversations one example being Twitter hashtags that connect separate threads by common topic. In (Xi et al., 2004) the authors represent a resume of the principal differences between traditional IR tasks and searching in newsgroups. Besides, they use a combination of measures such as author metrics (number of posts, number of replies, etc.) and features of the thread.

In question and answering, there are two streams of similar work: the first one is to find the best set of answers for a query, and the second is to identify question and answer pairs to build a querying system knowledge base. (Xi et al., 2004) associated posts that are most likely to answer a user's query. They used content, thread based, and author based features to train a linear regression and SVM classifier.

In (Feng et al., 2006), the authors aimed to discover the most relevant answer in question threads. They implemented a discussion-bot to automatically answer student queries in a threaded discussion by combining lexical similarity, speech acts and reputation of the author of posts into a similarity measure. To discover best potential answers, they used the HITS algorithm to find posts that are most likely to be answers to the initial question post. However, extract possible answers

(the most informative message) using analgorithm with a rule-based traverse that is not optimal for selecting a best answer; consequently, the result may comprise redundant or incorrect information.

Similar to Feng et al. (2006), (Hong and Davison, 2009) detected threads in which first posts are questions and its corresponding answers belong to the thread. To detect answers, they used features including the positions of the candidate answer posts, authorship and likelihood models based on content. (Seo et al., 2009) tested different combinations of these features with an SVM classifier and found post position and authorship result in the highest accuracies.

In (Cong et al., 2008), the authors choose a more common approach, in a way they permit questions to exist anywhere in a thread. Questions were found thanks to a rule classifier which is used to discover which frequent sequential Part-of speech patterns in sentences lead to questions. For answers, (Ding et al., 2008) used language models to construct weighted similarity graphs between each question and the set of candidate answers. For each detected question, a page-rank like propagation algorithm was utilized to determine and rank the set of candidate answers.

In (Ding et al., 2008) and (Yang et al., 2009), the identical set of authors extended their previous work (Cong et al., 2008) to incorporate the idea of context sentences and posts that provide indirect links between a question and answer pair. The authors used conditional random fields and structural support vectors to learn the contexts.

In summary, much novel works were conducted in question and answer detection, but the focus of this type of work is to detect a question and rank the candidate answers. While in the reply reconstruction task, we are trying to link posts together, regardless whether they are question-answer pairs or not.

In fact, (Seo et al., 2009) has been demonstrated that high quality thread structures learned which are based on manually created training data are able to ameliorate retrieval performance. However, it is still doubtful how we can enhance retrieval accuracy by using only the raw thread structures naturally available in

all the forums.

3.2.4 Retrieval of Twitter Microblogs

Extracting information from Social Networks Sites is one of several problems which have already been addressed in the past because of its theoretical and practical importance. Twitter has become a digital place where users can discuss public issues, share critical information during natural disasters and it will be possible to comment TV shows they are following. Users examine twitter particularly for answers, timely information (e.g., news, events), people information and topical information (Efron and Winget, 2010; Teevan et al., 2011) (Cherichi and Faiz, 2013). However, search functionality provided by Twitter is limited to keyword-based retrieval to return the most recent posts. 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 (Bernstein et al., 2010).

Recently, various researches focused on these phenomena with a closer perspective (Bruns and Burgess, 2011; Magnani et al., 2010, 2011, 2012). According to (Magnani et al., 2011) proposed a user-based tree model for retrieving conversations from microblogs. They considered only tweets that directly respond to other tweets by using the "@username" as a marker of addressivity. The advantage of this method is to have a coherent conversation based on the direct links between users. The downside is that this method neglected the tweets that does not contain the "@sign". However, we can have tweets that are in the conversation even though they are not directly linked with other tweets.

(Cogan et al., 2012) proposed a method to build conversation graphs. The authors focused on the particular case of a conversation formed by users replying to tweets. In this case, a tweet can only reply to one other tweet and a retweet are ignored, but users can get involved in conversations with other by commenting, liking and sharing other user's posts. However, a retweet can be found in a conversation. Other related works concentrated on different aspects of microblogging

conversations are (Huang et al., 2010; Song et al., 2010), they deal respectively with the tagging of conversations and the identification of topics. (Kumar et al., 2010) in their work have proposed a simple mathematical model that produces basic conversation structures taking into account the identities of each member of the conversation.

Besides, some recent studies aimed to identify how information flow depends on the structural properties of user interactions (Cha et al., 2009; Kwak et al., 2010). Moreover, modeling the interaction among users plays a crucial role in order to understand how the information is disseminated in the network (Wang et al., 2011; Gómez et al., 2011) and how to maximize its spread (Kempe et al., 2003; Chen et al., 2009). (Kwak et al., 2010) proposition involves the first large scale analysis of both topological and temporal structures of retweets in order to catch users' popularity of, trending topics and temporal patterns describing the propagation of information in the network. The temporal dimension of the information spread plays an important role in characterizing users' retweets and identifying influential's as well (Lee et al., 2010; Yang and Leskovec, 2011). Eventually, we have to mention that in dynamic interactions social cascades have also been used to reveal that, differently from static properties, the user's geographic locality is a key factor to characterize the way how the information circulates in the network (Scellato et al., 2011).

In summary, retrieving information from Twitter represented always a challenge given its volume, inconsistent writing and noise. Existing systems focus on term-based approach, but retrieving the context of conversation is often neglected, leading to less satisfactory results while searching information from tweets.

3.3 Microblog Conversation Approach

In recent years, microblogging services, like Twitter, draws the attention of users due to the ease and the speed of information sharing especially in real time. This service can be seen as a very large information repositorie containing millions of

text messages usually organized into complex networks involving users interacting with each other at specific times.

In our work, we propose an information retrieval approach for microblogging sites based on conversation retrieval concept. One of the main characteristic of Social Network Sites is that users do not just produce contents but they can get involved indirectly in conversations related to their interests with other users by commenting, liking and sharing other users'posts. Thus, our conversation retrieval approach consider that within a social Web of interactions, not only the informative content has to be observed as relevant but also its context.

Little recent studies focus their works on the conversation retrieval task. To the best of our knowldge, existing conversation retrieval approaches for microblogging sites have so far focused only on the particular case of a conversation formed by users replying to tweets. According to (Magnani et al., 2011) and more recently (Cogan et al., 2012) retrieved a conversation using only the tweets that are directly respond to other tweets by the use of "@user". Compared to these approaches, the basic idea of our conversation approach is to consider not only tweets in reply to other tweets especially by the use of "@sing" as a marker of addressivity, but also, tweets that are in reply indirectly.

To achieve this task, our work is divided in two steps for constructing twitter conversations. The first step is the replies conversation method, which produces the reply tree. A reply to a user will always begin with "@username". Thus, a search call will return all the replies to that user. In order, to determine whether the returned tweet is an element of the conversation, we can check the "in reply to status id" field. The second step is the conversation features selected which indicates tweets that are related to the conversation without using the @symbol. The task is to distinguish relevant tweets from irrelevant ones regarding a discussion. To do this, we define the features of selection namely the presence of the same hashtag and the same URL, the time difference between posts and the content similarity.

Summarizing, our contribution has the following outcomes:

- Representing the entire context of conversation not only the isolated content.
- During the retrieving of the conversation considering not only the tweet in reply to other tweet directly but also, we consider the tweet that in reply to another tweet indirectly (retweets, comment, anser).
- Finally, the advantage of representing a coherent conversation based on the specific relation between users talking about specific topic.

3.4 Microblog Conversation Architecture

In this section, we describe in details the process of detecting conversations from a collection of tweets. Figure 3.1 shows an overview of our conversation approach.

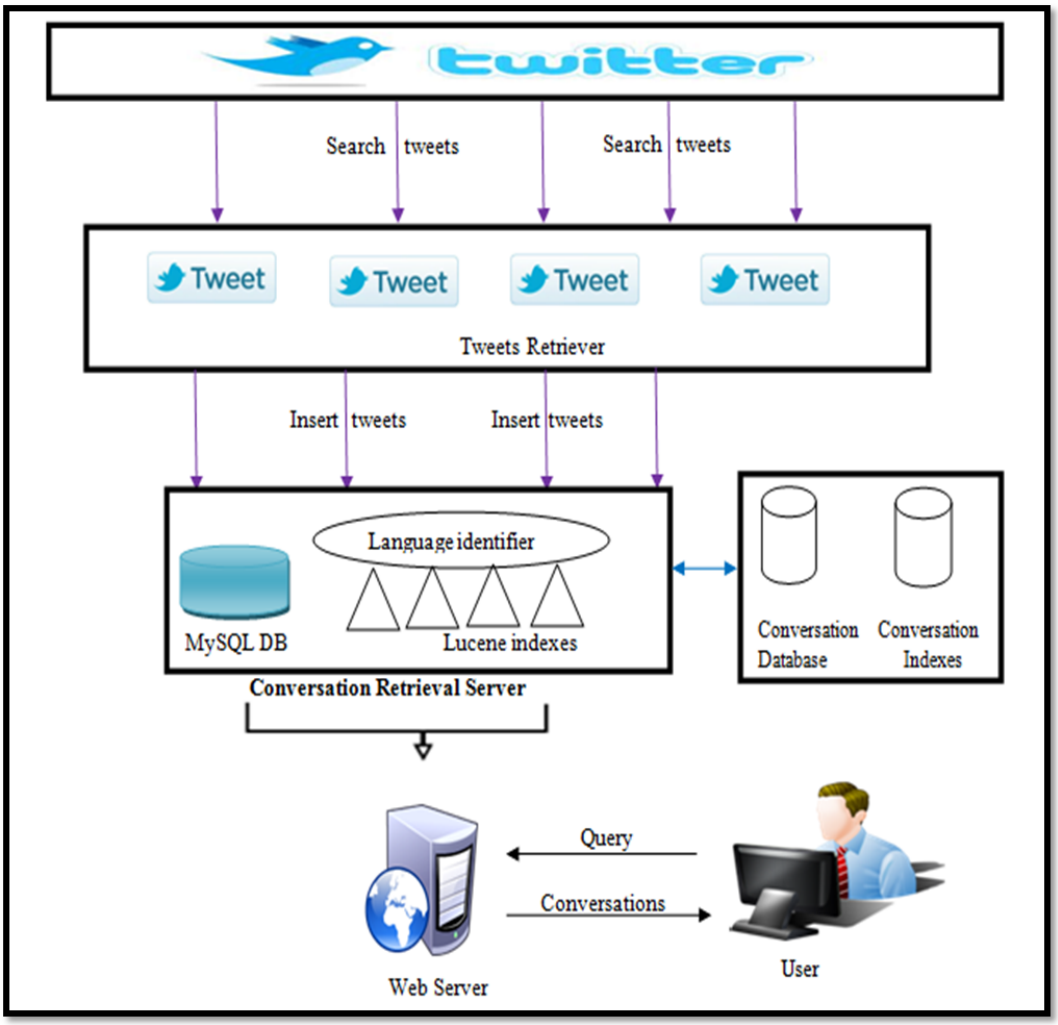

Figure 3.1: Microbog Conversation Architecture

Our approach operates on two steps. The first step is the replies conversation method. It is focusing on the functions and uses the "@sign". The second one is the conversation method without the use of the "@username" which produces tweets that are related to the conversation. We will present more details about this two steps in the next part.

3.4.1 Replies Conversation Method

In this subsection, we formally describe the first step of our architecture. We aim to collect all tweets in reply to other tweets directly. Obviously, a reply to a user will always begin with " @username", thus a search call will return all the replies to this user. In order, to determine whether the returned tweet is an element of the conversation, we can check the "in reply to status id" field.

The goal of this method is to create reply trees. The construction process consists of two algorithms run in parallel. First, twitter API is used to filter tweets with subjects of interests called specific keywords. The set of tweets retrieved from the filter operation is denoted S. Each tweet Si in S is a part of a conversation C. However, it may not be the root of the conversation C. A tweet can have three types: root, reply or retweet. Let consider (tSi) the type of tweet t. The goal of the construction algorithm, given Si, is to determine the root of the conversation C, and then to determine the set of all tweets which compose C.

Algorithm 1 Iterative Root Finder

```
T is the tweet passed as an input ( ID tweet)
L = 0
repeat
  if (type ((T -L)=Reply) or (type(T -L)=retweet)) then
    Extract Tn-L- 1 by matching field "in reply to status id"
  end if
  L = L + 1
until (type (T-L-1)=root)
```

Let T be a subset of conversation C where S0 is the root of the conversation C and Tn is a single tweet of the conversation retrieved using the filter (i.e., Si in the above context). T is defined as Tn-L is a reply to Tn-L-1 for k=(0,|T|-2), with this pattern repeating to S0. The goal of the Iterative Root Finder algorithm is to identify T0 given Tn. Note that when the algorithm starts, |T| is not known. The complexity of this algorithm depends on the number of tweets N in the conversation. The complexity is $\theta(g(n) * f(n))$.

Once, the conversation root R0 has been established, the iterative search al-

gorithm is used to seek the remainder of conversation C. The iterative search algorithm is run repeatedly until some conditions, indicating that the conversation has ended are met.

Algorithm 2 Iterative Search

```
N = the number of all tweets in C
for j = 1 to j = < |N| do
   Search for tweets addressed to author.
   Extract replies to Ti by matching field in reply to status id
end for
```

The complexity of this algorithm depends on the number of tweets N in the conversation. T(n)= n-1+1=n is in θ(n).

3.4.2 Conversational Features Selected

In this subsection, we describe features used to retrieve tweets without the use of the @symbol that are linked indirectly to the conversation. We evaluate our features in a context of information retrieval, i.e., the evaluation of a tweet relevance given to a tree reply structures.

We collect a set of tweets in which the features of our approach will be applied. We begin by grouping tweets with an identical "core" the same URL as the reply tree of conversation. Thereafter, the feature of hashtags is applied to collect tweets that share the same hashtags present in the tree of the conversation. In addition, sorting tweets by their time attribute allows us to efficiently remove tweets having a large distance in terms of time compared to the conversation. Finally, we calculate the similarity between candidates' tweets and tweets present in the reply tree to find those that are indirectly related to the conversation. If a good candidate is found (if the similarity of the tweets to a conversation is higher than a certain threshold), we can add this tweets to that conversation. The goal of our methods is to extract tweets that may be relevant to the conversation without the use of the @symbol. The system needs to decide whether a message is part of the discussion.

When choosing the features for our approach, we carefully focus on features

that are simple, easy to obtain and fast to compute. Features are normalized to lie between 0 and 1.

We use the following notations in the sequel:

- **T** is a set of tweets present in the reply tree of conversation.
- **t** is tweets that can be linked indirectly to the conversation.

- **Presence of the same URL as the reply tree:** Twitter allows users to include URL as a supplement information in their tweets. By sharing an URL, an author would confirm the information published in his tweet. This feature is applied to collect tweets that share the same URL present in the reply tree. This is a binary feature.

$$\mathrm{F1(T,t)} = \begin{cases} \text{1 if t contains the same URL of reply tree.} \\ \text{0 otherwise.} \end{cases} \tag{3.1}$$

- **Presence of same Hashtags as the reply tree:**

  The # symbol, called a hashtag, is used to mark a topic in a tweet or to follow conversation. Any user can categorize or follow topics with hashtags. In this step, we integrate the use of the same hashtags in order to find tweets that contain the same hashtags present in conversation's reply tree. This is a binary feature.

$$\mathrm{F2(T,t)} = \begin{cases} \text{1 if t contains the same hashtag of reply tree.} \\ \text{0 otherwise.} \end{cases} \tag{3.2}$$

- **Time Difference :**

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

conversation. Therefore, the temporal feature is considered as a useful one. To better explain the time difference we presented in Figure1, a set of tweets published at specific timestamps. For example, the time difference between post #1 and post #2 is 20 minutes whereas the time difference between post #1 and post #7 is 1 hour and 14 minutes. Typically, the duration of a conversation is short. Our analysis has shown that the vast majority of conversations are not continued if the oldest tweet in the conversation is over 4 hours old. After that, reply activity dramatically declines. Therefore, we use a max time difference threshold, above which the time similarity is zero.

ID	Time	@UEFAcom tweet
1	00:32	GER 0-0 ITA Kick-Off #GERITA #EURO2012 http://uefa.to/NSfroA
2	00:52	GER 0-1 ITA Goal: #Balotelli (20) #GERITA #EURO2012 http://uefa.to/NSfroA
3	01:08	GER 0-2 ITA Goal: #Balotelli (36) #GERITA #EURO2012 http://uefa.to/NSfroA
4	01:09	GER 0-2 ITA Yellow card: #Balotelli (37) #GERITA #EURO2012 http://uefa.to/NSfroA
5	01:18	GER 0-2 ITA End 1st half #GERITA #EURO2012 http://uefa.to/NSfroA
6	01:33	GER 0-2 ITA Substitution: #Klose In #Gomez Out (46) #GERITA #EURO2012 http://uefa.to/NSfroA
7	01:34	GER 0-2 ITA Substitution: #Reus In #Podolski Out (46) #GERITA #EURO2012 http://uefa.to/NSfroA
8	01:34	GER 0-2 ITA Start 2nd half #GERITA #EURO2012 http://uefa.to/NSfroA
9	01:46	GER 0-2 ITA Substitution: #Diamanti In #Cassano Out (58) #GERITA #EURO2012 http://uefa.to/NSfroA

Figure 3.2: Tweets' Time Difference

- **Content :**

  It measures how many tweets of the conversation tree are similar in content with the tweet candidate, and then we classify the results according to their relevance. We calculate a cosine similarity score for every pair of tweets. Cosin calculates the angle between two vectors. In this case, these two vectors are the contents of two messages because message content is typically a plain text tf-idf (term frequency-inverse document frequency) is used to convert text into vectors with numerical elements.

- **Similarity Function :**

  Finally, the similarity between tweets linked indirectly to the conversation t (anser, commenting or sharing other users' posts), and tweets which are

present in the reply tree T, is computed by a linear combination of the similarities between the criteria that we used from twitter:

$$\mathrm{Sim}(\mathrm{ti}, \mathrm{T}) = \mathrm{PURL}(\mathrm{ti}, \mathrm{T}) + \mathrm{PHashtags}(\mathrm{ti}, \mathrm{T}) + \mathrm{Time}(\mathrm{ti}, \mathrm{T}) + \mathrm{Content}(\mathrm{ti}, \mathrm{T}) \tag{3.3}$$

3.5 Conclusion

In this chapter, we described a synthesis of a variety of work in the research of social information specifically looking for conversation retrieval. we also criticized the antecedent research works. Then, we studied the application of the conversation retrieval concept into social media. Thus, after analyzing their specifications, we presented our contribution.

In order to evaluate the performance of our method, the next chapter will describe the implementation of our system and the the results generated by the framework and evaluate its effectiveness.

Chapter 4

Experiments and Results

4.1 Introduction

In this chapter, we present the implementation of our solution and we describe the different modules that compose our framework. For each module, we describe the process adopted and the tools used for the implementation. Finaly, we present the experiments setup and evaluation.

4.2 Implementation Modules

Based on our system architecture described in section 3.3 of chapter 3, we illustrate in this section the modules of our system including mechanisms and tools as demonstrated in Figure 4.1.

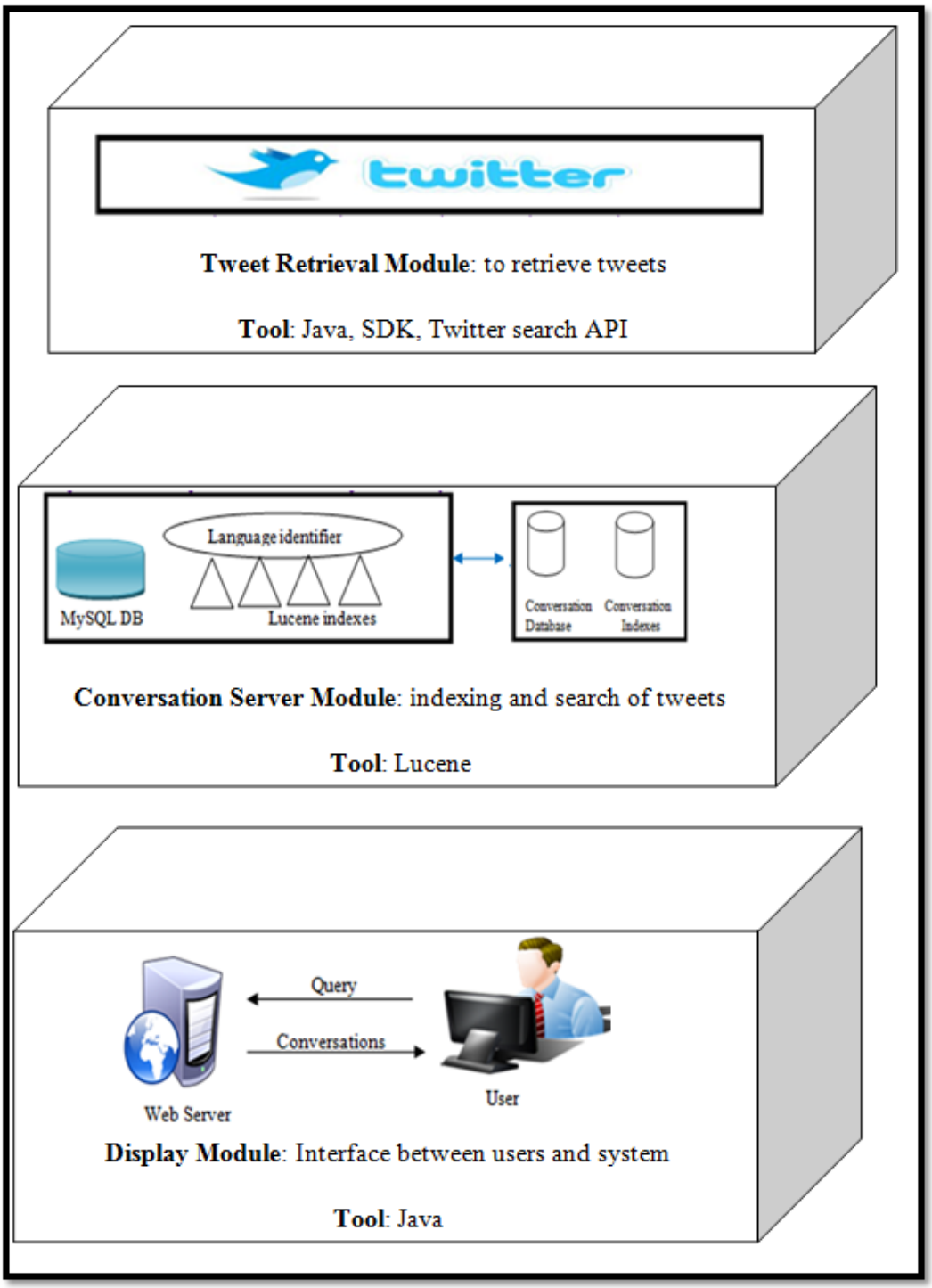

Figure 4.1: Our Framework Components

A Conversation Retrieval system is constituted at least of three modules. First,

we will explain tweets retrieve module. Then, we will define the retrieval tool. Second, we describe Conversation Server module. After that, we will explain the tool that allowed us to index tweets. Finally, we will see an overview of the display tool.

4.2.1 Tweet Retrieval Module

The Tweet Retrieval Clients that use the **Twitter Search API**[1] to get the corresponding tweets and also to retrieve the conversation chains when tweets respond to others. The REST API allows access to specific tweets, follower lists, retweet lists, etc. Each tweet is assigned a unique unsigned 64 bit integer identifier -this is the ID. Usage of many of the APIs are limited to 350 calls per hour for authenticated applications, thus any conversation collection algorithm which relies on the REST API will be severely limited.

In order to mitigate against this issue, we used Twitter API search. One of the key components of this algorithm is the ability to find replies to tweets. The Twitter REST API does not provide a mechanism to directly find replies, thus we use the search API to find replies to users. A reply to a user will always begin with "@username", thus a search call will return all the replies to that user. In order to determine whether the returned tweet is an element of the conversation, we can check the "in reply to status id" field.

4.2.2 Conversation Server Module

In this module, **Lucene**[2] is used to index basic feature from tweets and to index the text of the conversations. Apache Lucene which is an open source information retrieval software library and originally created in Java by Doug Cutting. Lucene contains principal classes including:

- Analyzer: it includes a number of "stemming" algorithms for various lan-

[1] http://dev.twitter.com/doc/get/search
[2] http://lucene.apache.org/

guages, which can improve document retrieval accuracy. Besides, there are

- IndexWriter: it creates and maintains an index.
- TokenStream: it enumerates the sequence of tokens created by the analyzer. Indexing is an important element to analyze texts in order to extract terms and their frequency. Terms frequency are used to obtain both the users and documents vectors.

4.2.3 Display Module

To create our application we used java swing which establishes the standard for building server-side user interfaces. In fact, Java Swing provides the multiple platform independent APIs interfaces for interactions between the users and GUIs components. Java provides an interactive feature for design the GUIs toolkit or components like: labels, buttons, text boxes, checkboxes, combo boxes, panels and sliders etc. All AWT flexible components can be handled by Java Swing. The Java Swing supports the plugging between the look and feel features. The look and feel that means the dramatically changing in the component like JFrame, JWindow, JDialog etc. for viewing it into several types of window.

4.3 TweetsConvs System Description

We describe in the following steps the process adopted to run our system:

- **User identification:** Figure 4.2 shows how each user accesses to our system according to the identification via a username and a password.

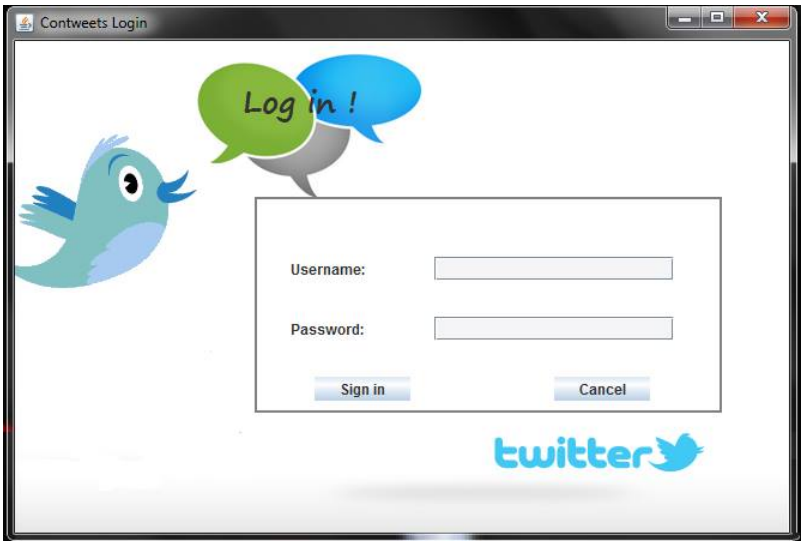

Figure 4.2: User's Identification

Besides, as presented in Figure 4.3, our system contains a user identification control in order to secure the access.

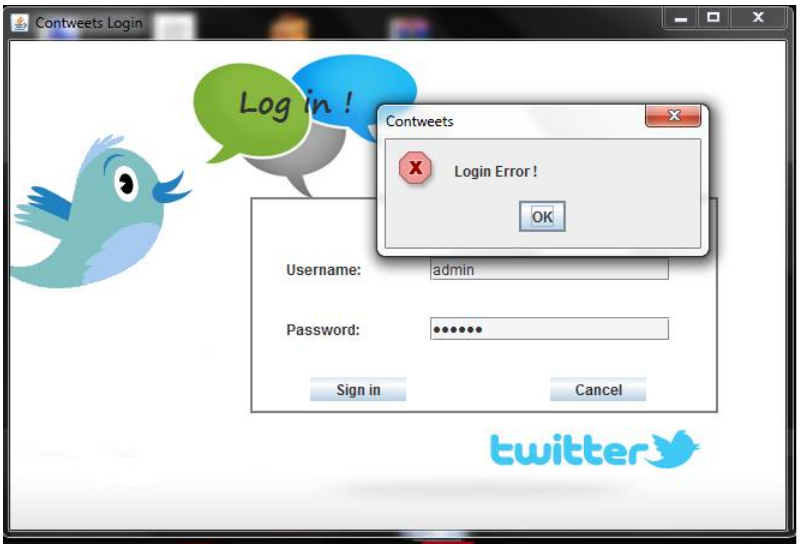

Figure 4.3: User's Failed Identification

- **Search Conversation:** Figure 4.4 present the conversation track task. It's possible for the user, simply to insert the topic that interests him. Our ap-

plication will be scanning quickly through a stream of tweets and we present the conversations between people related to that topic.

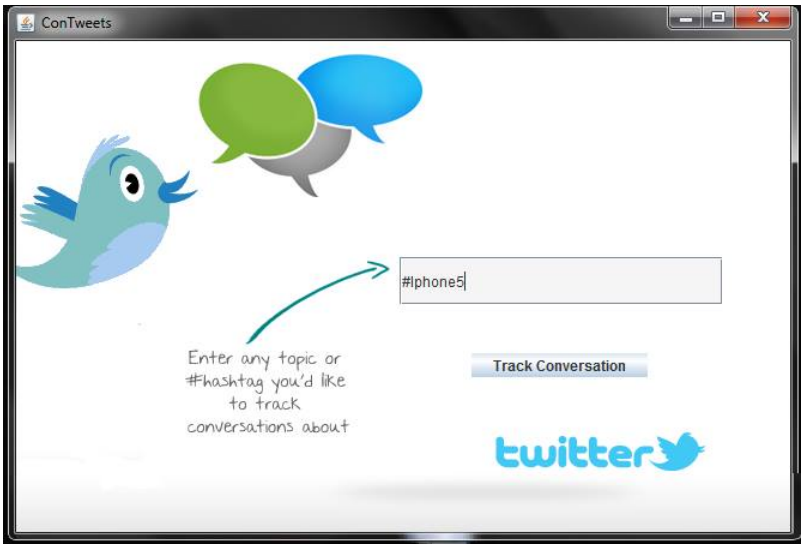

Figure 4.4: Track Conversation

- **Conversations Results:** This conversation is constituted 5 messages long and lasted 11 minutes. It was initiated by "Looping", who publicly announces (in one undirected tweets) the problem of the music player in his Iphone. Others respond quickly (using "@sign") with reactions, further observations about this problem based on their own examination, suggestions regarding how to respond to it, and suggestions in response to "Looping". The topical evolution of this group conversation is quickly, as shown in Figure 4.5. Because, the response to "Looping" initial tweets came quickly, many messages overlapped each other temporally.

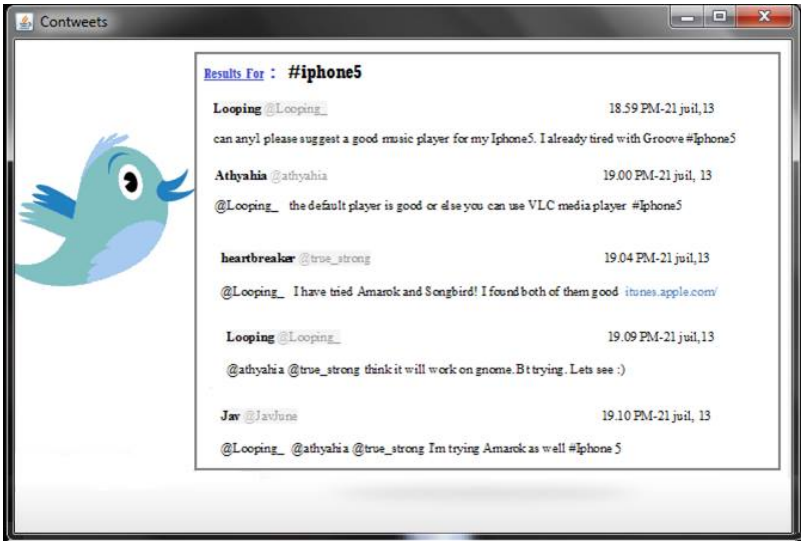

Figure 4.5: TweetsConvs Search Results

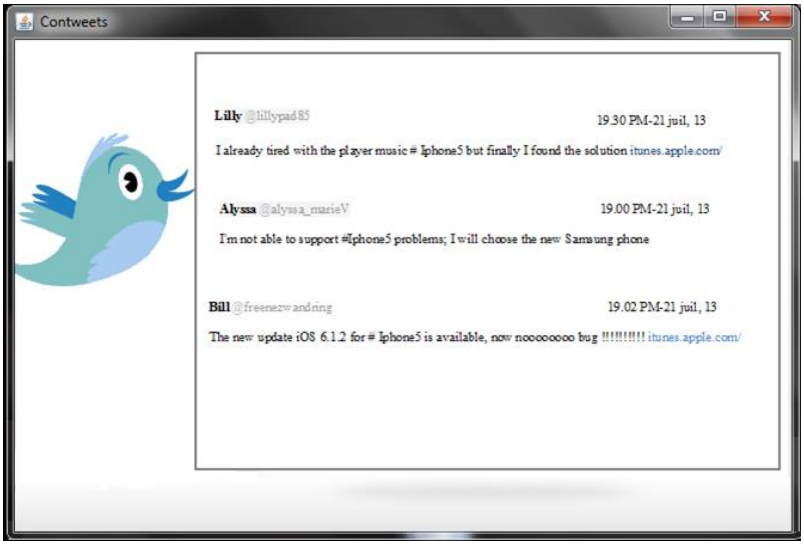

Figure 4.6: Tweets Linked to Conversation

In Figure 4.6, we present the tweets that are indirectly related to the conversation and that satisfy our conversational features selected mentionned in Chapter 3. Tweets posted are not direct reply about the conversation but they are status posted by users that address the same subject "Iphone5". We note the presence

of hasthags # Iphone5 confirming that the tweets have the same subject as the conversation (reply tree) and the presence of a URL that was already present in the reply tree confirming the information published in the tweet. In addition, tweets are in the same time interval as the conversation. Finally, by calculating the similarity score, the tweets that had the best scores will be displayed.

4.4 Experiments

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

- Using a conversational model 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.

  To evaluate these two hypotheses. We used a set of people who have been asked to score the relevance and consistency of retrieved conversation. These people have been selected among students and colleagues (with backgrounds in computing).

4.4.1 Data description

Our implementation exploits a part of the corpus Tweets 2013. Table 4.1, represents the general statistics of this corpus. In fact, the collection is composed of 13 477 exchanges (tweets) expressed in various languages, where an exchange was defined as at least one initiation and one response involving at least two people, with or without use of the "@sign" and posted on Twitter between June 23, 2013 and July 27, 2013. Each tweet is characterized by its identifier (ID), author, and date of publication.

The dataset has been obtained by retrieving the tweets and conversations containing trending topic. Trending topics have been determined directly by Twitter, and we select the ten most frequent ones during the monitoring period. The ten

topics were: [S1: Tour de France], [S2: La France], [S3: Justin Bieber], [S4: 3year-sofonedirection], [S5: FIFA], [S6: Iphone5], [S7: derek], [S8: Germany-Italia], [S9: cassano], [S10: bari].

tweets / Topic	Tweets in reply	Tweets not in reply	Total
Topic1	542	758	1300
Topic2	751	749	1500
Topic3	930	420	1350
Topic4	730	490	1220
Topic5	781	649	1430
Topic6	351	649	1000
Topic7	900	421	1321
Topic8	899	700	1599
Topic9	850	650	1500
Topic10	807	450	1257
Total	7541	5936	13477

Table 4.1: Data Set Statistics

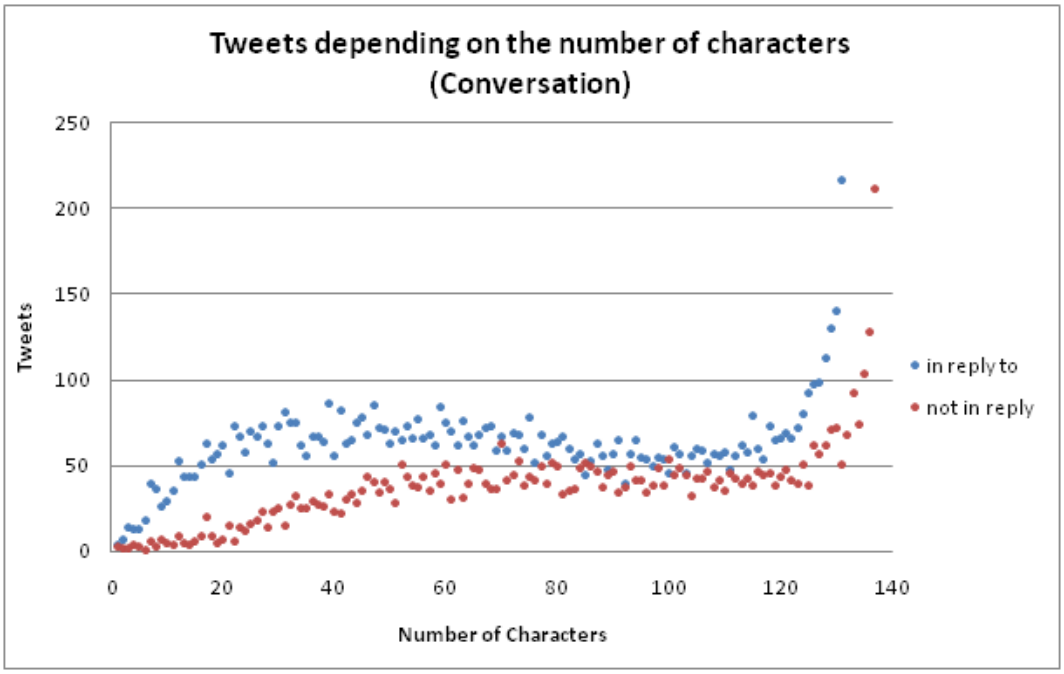

Figure 4.7: Based on "in reply" and "not in reply" tweets representation

In general, the majority of the tweets were relevant, we manually evaluate the

relevance of the tweets marking in some cases some tweets as half-relevant when they were consistent with the query but without a clear informative content.

After retrieving the data from Twitter, we use it to find out the tweets that are in reply by using the "@sign" and there are not in reply. After the corpus analysis, we found that 7541 tweets are in reply and 5936 tweets are not in reply as presented in Figure 4.7.

4.4.2 Evaluation Metrics

This section describes the experiment run in the study to evaluate the performance of our approach using the standard Recall, Precision and F-score.

The **Precision** (also known as specificity) measures how often the system is correct when it detect a conversation. It is calculated by dividing the number of correct outputs (true positive, TP) by the total number of the outputs. The total number of the outputs is the number of correct outputs plus the number of incorrect ones (false positive, FP).

$$\text{Precision} = \frac{|\text{TP}|}{(|\text{TP}| + |\text{FP}|)} \tag{4.1}$$

The **Recall** (also known as sensitivity) measures how often the system correctly finds the right classes to output. It is defined as proportion of true positives against potential correct outputs. The total number of potential correct outputs is the number of correct output (true positive, TP) plus the count of objects that should have been output but where not (true negative, TN).

$$\text{Recall} = \frac{|\text{TP}|}{(|\text{TP}| + |\text{TN}|)} \tag{4.2}$$

The **F-score** is the harmonic mean between Precision and Recall:

$$\text{F} - \text{Score} = \frac{(2 * \text{Precision} * \text{Recall})}{(\text{Presicion} + \text{Recall})} \tag{4.3}$$

Fleiss' kappa coefficient [3] is a statistical measure for assessing the reliability of agreement between a fixed number of raters when assigning categorical ratings to a number of items or classifying items. This contrasts with other kappas such as Cohen's kappa, which only work when assessing the agreement between two raters. The measure calculates the degree of agreement in classification over that which would be expected by chance. There is no generally agreed-up on measure of significance, although guidelines have been given.
Agreement can be thought of as follows, if a fixed number of people assign numerical ratings to a number of items then the kappa will give a measure for how consistent the ratings are. The kappa,κ can be defined as,

$$\text{kappa} = \frac{\bar{P} - \bar{P_e}}{1 - \bar{P_e}} \tag{4.4}$$

4.5 Our Approach Results

Topic	Precision(%)	Recall(%)	F-score(%)
Topic1	60.37	52.97	56.42
Topic2	74.49	59.84	66.36
Topic3	57.96	37.50	45.53
Topic4	69.75	50.43	58.53
Topic5	78.63	57.69	66.55
Topic6	68.28	32.15	43.71
Topic7	65.62	60.98	63.20
Topic8	48.52	50.12	49.30
Topic9	**79.21**	**65.53**	**71.72**
Topic10	52.36	41.23	46.13

Table 4.2: Evaluation of Our System "TweetsConvs"

The results are given in Table 4.2 demonstrate the significance of considering the convresation features such as the similarity function. The overall performance scores of our framework are 65.53% recall, 79.21% precision and 71.2% F-Score.

[3] http://en.wikipedia.org/wiki/Fleiss'$_k$appa

Precisions are higher than recall in all topics. The best F-Score around 71.22% in the ninth topic. Our results presented a best retrieval accuracy. This is due to the fact that we combined a set of features and the direct replies method to retrieve conversation.

4.5.1 Properties of Conversations

In this section we state the main observations about the conversations from our datasets. We study the distribution of conversation sizes and depth (which is the length of the maximum path to a leaf from the root).

4.5.1.1 Depth

We examined the distance between an original tweet and the replies it attracts. Most conversations that occur in Twitter appear to be dyadic exchanges of three to five messages sent over a period of 15 to 30 minutes. Of all tweets that generated a reply, 85% have only one reply. Another 10.7% attracted a reply to the original reply - the conversation was two levels deep. Only 1.53% of Twitter conversations are three levels deep - after the original tweet, there is a reply, reply to the reply, and reply to the reply of reply. The Figure 4.8 shows conversation levels deep.

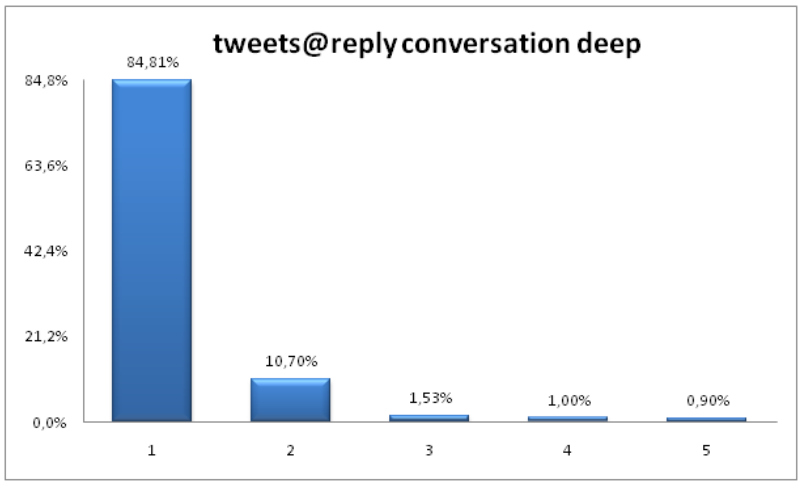

Figure 4.8: Conversation Levels Deep

4.5.1.2 Size

The analysis we made has demonstrated that the majority of conversations are not continued if the oldest tweet in the conversation is more than 4 hours old. We found that 97.87% of @replies take place within the first hour of the original tweet being published, while an additional 0.98% of replies happen in the second hour. Subsequently, reply activity dramatically declines as it is shown in the following figure. Thus, in our algorithm, if the oldest tweet in a conversation is more than 4 hours old we stop tracking it. There is no way to know for certain that a conversation will not be replied to at some indeterminate time in the future, however the algorithm must eventually stop following conversations since there are practical limits to how many search calls can be made while still aggregating active conversations.

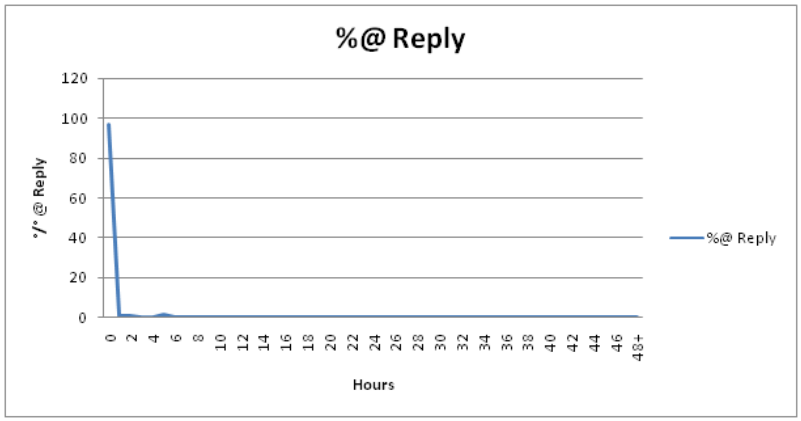

Figure 4.9: Conversation Length

Some circumstances prevent us to correctly recreate a reply tree as it originally occurred. For example, the user has the option of deleting a tweet, or a user with a public profile could write a tweet in reply to a user with a private profile. This way, less than 7% of conversations are affected. Figure 4.9 below shows the fraction of @replies as number of hours since the original tweet.

4.5.2 Experimental Outcomes

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 an event and queried a Twitter Social Network Site using our approach. Then we have asked a set of 100 people to rate the conversation. These people have been selected among students and colleagues (with backgrounds in computing and social sciences). Every user was informed of conversation topic and he been asked to judge the relevance and the consistency of conversation.

4.5.2.1 Conversation' Relevance

It can be observed in the Figure 4.10 the inclusion of conversation metrics may have a significant impact on the users' evaluation of the results of a search task. Looking at the histogram in Figure 4.10, we can see that the evaluators have expressed their appreciation of the conversation relevance. The results are very satisfactory people have seen that most conversations are relevant which the percentage is 79%.

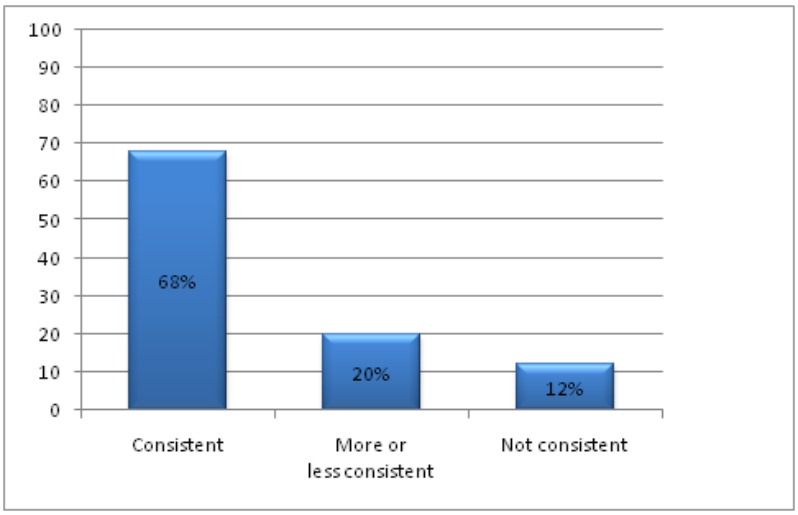

Figure 4.10: Conversation'Relevance

4.5.2.2 Conversation'Consistency

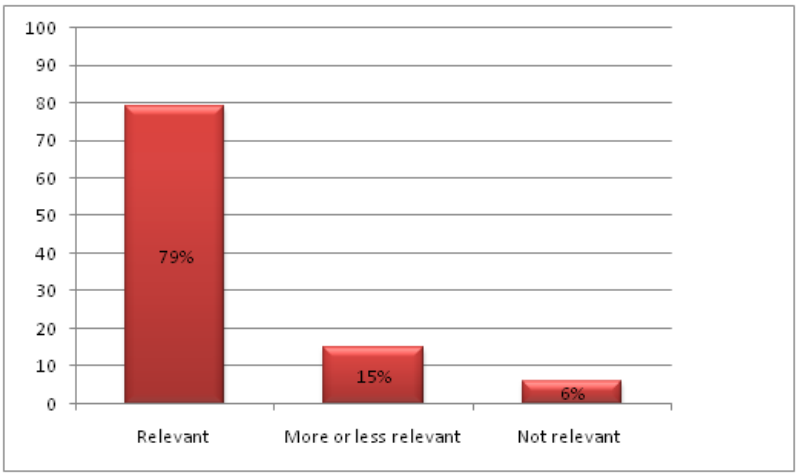

Figure 4.11: Conversation'Consistency

We also tested the consistency of tweets related indirectly to the conversation. The test of the consistency of conversations also presented a good result. The people have found that 68% of conversation are consistency. This high score show that considering the date attribute and the content attribute as significant for the conversation detection task.

4.6 The Kappa Statistic

In this example, 100 raters (n) assign ten "subjects" (N) to a total of three categories (k). The categories are presented in the columns, while subjects are presented in the rows. Each cell is filled with the number of raters who agreed that a certain subject belongs to a certain category.

topic \ Relevance	Relevant	Not Relevant	More or less relevant
Topic1	76	14	10
Topic2	62	10	28
Topic3	72	15	13
Topic4	75	12	13
Topic5	80	10	10
Topic6	52	20	28
Topic7	74	12	14
Topic8	62	28	10
Topic9	85	5	10
Topic10	80	10	10
Total	708	141	151

Table 4.3: Table of Values for Computing our Worked Example

- **Data**

See table to the right.

N = 10, n= 100, k = 3

Sum of all cells = 100 Sum of P_i, = 7.780

- **Calculations**

For example, taking the first column,

$p_1 = \frac{76+62+72+75+80+52+74+62+85+80}{1000} = 0.708$

And taking the second row,

$P_2 = \frac{1}{100(100-1)}\left(62^2 + 10^2 + 28^2 - 10\right) = 0.467$

In order to calculate , we need to know the sum of P_i,

$\sum_{i=1}^{N} P_i = 0.603 + 0.467 + \cdots + 0.535 + 0.656 = 8.580$

Over the whole sheet,

$$\bar{P} = \frac{1}{(10)}(7.780) = 0.778 \tag{4.5}$$

$$\bar{P}_e = 0.708^2 + 0.141^2 + 0.151^2 = 0.543 \tag{4.6}$$

$$\kappa = \frac{0.858 - 0.543}{1 - 0.543} = 0.689 \tag{4.7}$$

- **Interpretation of Kappa Statistic**

What does a specific kappa value mean? We can use the value of 0.68 from the example above. Not everyone would agree about whether 0.68 constitutes "good" agreement. However, Landis and Koch (1977) gave the following table for interpreting κvalues.

It turns out that, using this scale, a kappa of 0.68 is in the "Substantial" agreement range between our observers. Remember that perfect agreement would equate to a kappa of 1, and chance agreement would equate to 0. Table 4.4 allows to "visualize" the interpretation of kappa.

κ	Interpretation
<0	Poor agreement
0.01-0.20	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-1.00	Almost perfect agreement

Table 4.4: Interpretation of Kappa

4.7 Comparison with Baseline Approachs

After evaluating our framework's effectiveness, we compare our approach with two baseline systems (Cogan et al., 2012) and (Aumayr et al., 2011) for each subject in our topic sets. (Cogan et al., 2012) proposed an approach to detect a conversation by considering only the tweets in reply to other tweets especially by the use of "@sing" as a marker of addressivity. The authors focused on the particular case of a conversation formed by users replying to tweets and they did not consider

the retweet. (Aumayr et al., 2011) proposed an approach to reconstruct the reply structure in online discussion forums by using a set of features. The following table shows the comparison result.

Baseline Approachs	Precison(%)	Recall(%)	F-score(%)
Cogan et al.,2012	74.52	60.23	66.61
Erik Aumayr et al.,2011	78.12	62.56	69.47
TweetsConvs	**79.21**	**65.53**	**71.72**

Table 4.5: The Comparison Result

For the 10 topic, our system outperforms the first baseline approach(Cogan et al., 2012) with 4,69% using the Precision and with a significant improvement of 5.30% using the Recall. By comparing our approach and the (Aumayr et al., 2011) approach, we found an improvement of 2.9% using the Recall metric and an slight improvement of 1.09% using the Precision (cf. Table 4.5).

Our results achieved the best retrieval accuracy. This is due to the fact that we combine a set of features and the direct replies method to retrieve conversation. However, it worth noticing that the method which uses a conversational model without consedering conversation features does not reach the same level of precision and recall.

4.8 Discussion

Our experiments demonstrate that there is a clear advantage to use social information in the aim to conversation detection. There is, also, an opportunity for using Twitter to show the significance of extracting the conversation from microblogs which constitutes essential sources of hot topics and recent events.

The results reveal our approach's effectiveness with a very high efficiency. Fortunately, by comparing our system with two baseline systems, we deduce that our system performs better results. We notice that the results illustrate that it is important to combine tweets which are in reply to other tweets and tweets are in reply indirectly. The purpose of the combination is to present the entire conversation,

that has been published and also its context.

Finally, we have to note that conversation retrieval is a new concept that emerged and was highlighted with the explosion of the social Web and the proliferation of microblogsand social networks. Recently, the recent projects are emphasizing on conversation retrieval. Track how a particular brand is communicating with its customers. What kind of PR it has, the kind of customer service it provides its customers via Twitter, how companies receive feedback and suggestions etc, is certainly something that businesses might love to tap into especially because it gives the kind of insight about how a particular company's social media marketing allows it to be successful.

4.9 Conclusion

We expose in this chapter in details our system implementations and its evaluation. The conversation retrieval is constructed according to our method described in the previous chapter via the microblogging service (Twitter). Users' tweets are extracted, analyzed and integrated into the conversation. Our experimental results have highlighted many interesting points. The fact that we combine a set of features and the direct replies method to retrieve conversation improves the best retrieval accuracy. Furthermore, the comparison of our system with two baseline systems using three metrics proved the higher performance of our search results confirming our model's effectiveness.

Conclusion

Recently, we moved from using the Web as a mean of information access to generating data, a so called user-generated data, especially with the increase of social services that merged with the Web 2.0. In fact, those services, like tagging, microblogging, sharing media, require users to be active and allow expressing their interests, opinions and preferences.

This is the backdrop to the work presented in this thesis. In fact, we have introduced a microblogging search task called conversation retrieval: an information retrieval activity exploiting a set of conversational features in addition to the directly exchanged text messages to retrieve conversation. Our approach allows the retrieval of millions of conversations generated on the popular Twitter social network site, and in particular conversations about trending topics.

To achieve our framework's goals, we proceeded into two steps:

The first step is the replies conversation method. Obviously, a reply to a user always begin with "@username", thus a search call will return all the replies to this user. In order, to determine whether the returned tweet is an element of the conversation, we can check the "in reply to status id" field. The goal of this method is to create reply trees. The construction process consists of two algorithms run in parallel.

The second step is conversation selected features. We describe a set of conversational features used to retrieve tweets without "@sing" that can be linked indirectly to the conversation (commenting or sharing other users' posts).

In order to evaluate our approach, we implement the generic framework. Our

implementation exploits a part of the corpus Tweets 2013. Given this collection, the framework retrieve all conversation related to topic need. We obtain a recall around 65.53%, a precision around 79.21% and F-score around 71.72%. The experimental evaluation of our contribution demonstrated its effectiveness in several aspects. In addition, our experimental results have highlighted many interesting points. First, the fact that we combine a set of features and the direct replies method to retrieve conversation improves the best retrieval accuracy and also provides results that are considered more satisfactory with respect to a traditional Web search task.

Our future work is to investigate the real time search. The evaluation of the conversational approach can be improved by evaluating our approach with a stream of tweets in real time (Twitter API streaming). On the other hand,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. We will focus to include these aspects into the search process. In addition, we will focus on ranking conversation by using ranking measures including Social Network Analysis (SNA) centrality metrics, time-related conversational metrics and other specific features of current microblogging sites.

Appendix: Evaluation of Conversation Retrieval from Twitter

We aim to propose an information retrieval approach for microblogging sites based on the conversation retrieval concept. We will examine Twitter network. the basic idea of our conversation approach is to consider not only the tweets in reply to other tweets especially by the use of "@sing" as a marker of addressivity, but we consider also the tweets are in reply indirectly. we define a conversation as a set of short text messages posted by a user in reply to other users directly by using "@username" (a message is a reply to another) or indirectly by commenting and sharing other users' posts. Thus, our work is divided in two steps for constructing twitter conversation. The first step is the replies conversation method with "@username", which produces the reply tree. The second one is the conversation method which produces tweets that may be related to the conversation without the use of the @symbol. -we did this survey to evaluate this definition of a conversation[4].

[4]https://docs.google.com/forms/d/1tcSOUMpYT0T5FMbug9$_4$sO3yoLkliiamH9b6LDlpKIU/

Questions	**Réponses**
1. Sex?	☐ Woman ☐ Man
2. How do you assess the relevance of this conversation regarding the subject "iphone5"?	☐ Relevant ☐ Not relevant ☐ More or less relevant
3. How do you rate the consistency between the previous conversation and the following articles?	☐ Consistent ☐ Not consistent ☐ More or less consistent
4. Do you rate the usefulness of this application?	☐ Useful ☐ Unnecessary ☐ More or less useful

Bibliography

Aumayr, E., Chan, J., and Hayes, C. (2011). Reconstruction of threaded conversations in online discussion forums. In Adamic, L. A., Baeza-Yates, R. A., and Counts, S., editors, *ICWSM*. The AAAI Press.

Baeza-Yates, R. A. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Bernstein, M., Hong, L., Kairam, S., Chi, H., and Suh, B. (2010). A torrent of tweets: Managing information overload in online social streams. In *Workshop on Microblogging: What and How Can We Learn From It?*

Boyd, D. and Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. volume 13, page article 11, Washington, DC, USA. IEEE Computer Society.

Boyd, D., Golder, S., and Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences*, HICSS '10, pages 1–10, Washington, DC, USA. IEEE Computer Society.

Bruns, A. and Burgess, J. E. (2011). #Ausvotes: how twitter covered the 2010 australian federal election. *Communication, Politics and Culture*, 44:37–56.

Buckley, C., Singhal, A., and Mitra, M. (1995). New Retrieval Approaches Using SMART: TREC 4. In *Proceedings of the Fourth Text Retrieval Conference (TREC-4)*, pages 25–48.

Carenini, G., Ng, R. T., and Zhou, X. (2007). Summarizing email conversations with clue words. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 91–100, New York, NY, USA. ACM.

Carvalho, V. R. D. and Cohen, W. W. (2005). On the collective classification of email "speech acts". In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 345–352, New York, NY, USA. ACM.

Cha, M., Mislove, A., and Gummadi, K. P. (2009). A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 721–730, New York, NY, USA. ACM.

Chen, W., Wang, Y., and Yang, S. (2009). Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 199–208, New York, NY, USA. ACM.

Cherichi, S. and Faiz, R. (2013). 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, pages 24:1–24:7, New York, NY, USA. ACM.

Cogan, P., Andrews, M., Bradonjic, M., Kennedy, W. S., Sala, A., and Tucci, G. (2012). Reconstruction and analysis of twitter conversation graphs. In *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*, HotSocial '12, pages 25–31, New York, NY, USA. ACM.

Cohen, W. W., Carvalho, V. R., and Mitchell, T. M. (2004). Learning to classify email into"speech acts". In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 309–316.

Cong, G., Wang, L., Lin, C.-Y., Song, Y.-I., and Sun, Y. (2008). Finding question-answer pairs from online forums. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 467–474, New York, NY, USA. ACM.

Cselle, G., Albrecht, K., and Wattenhofer, R. (2007). Buzztrack: topic detection and tracking in email. In *Proceedings of the 12th international conference on Intelligent user interfaces*, IUI '07, pages 190–197, New York, NY, USA. ACM.

Dalli, A., Xia, Y., and Wilks, Y. (2004). Fasil email summarisation system. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ding, S., Cong, G., yew Lin, C., and Zhu, X. (2008). Using conditional random fields to extract contexts and answers of questions from online forums. In *ACL*. The Association for Computer Linguistics.

Donath, J. and Boyd, D. (2004). Public displays of connection. *BT Technology Journal*, 22(4):71–82.

Efron, M. and Winget, M. (2010). Questions are content: a taxonomy of questions in a microblogging environment. In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47*, ASIS&T '10, pages 27:1–27:10, Silver Springs, MD, USA. American Society for Information Science.

Elsner, M. and Charniak, E. (2010). Disentangling chat. volume 36, pages 389–409, Cambridge, MA, USA. MIT Press.

Erera, S. and Carmel, D. (2008). Conversation detection in email systems. In *Proceedings of the IR research, 30th European conference on Advances in information retrieval*, ECIR'08, pages 498–505, Berlin, Heidelberg. Springer-Verlag.

Feng, D., Shaw, E., Kim, J., and Hovy, E. (2006). Learning to detect conversation focus of threaded discussions. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 208–215, Stroudsburg, PA, USA. Association for Computational Linguistics.

Frakes, W. B. and Baeza-Yates, R. A., editors (1992). *Information Retrieval: Data Structures and Algorithms*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Gómez, V., Kappen, H. J., and Kaltenbrunner, A. (2011). Modeling the structure and evolution of discussion cascades. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*, HT '11, pages 181–190, New York, NY, USA. ACM.

Granovetter, M. (1973). The strength of weak ties. *The American Journal of Sociology*, 78(6):1360–1380.

Gross, R. and Acquisti, A. (2005). Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, WPES '05, pages 71–80, New York, NY, USA. ACM.

Honeycutt, C. and Herring, S. C. (2009). Beyond microblogging: Conversation and collaboration via twitter. In *Proceedings of the Forty-Second Hawaiâi International Conference on System Sciences (HICSS-42). Los Alamitos, CA.*, pages 1–10, Los Alamitos, CA, USA. IEEE Computer Society.

Hong, L. and Davison, B. D. (2009). A classification-based approach to question answering in discussion boards. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 171–178, New York, NY, USA. ACM.

Huang, J., Thornton, K. M., and Efthimiadis, E. N. (2010). Conversational tagging in twitter. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, HT '10, pages 173–178, New York, NY, USA. ACM.

Ingwersen, P. (1992). *Information Retrieval Interaction.* Taylor Graham, London.

Java, A., Song, X., Finin, T., and Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA. ACM.

Joinson, A. N. (2001). Self-disclosure in computer-mediated communication: the role of self-awareness and visual anonymity. *European Journal of Social Psychology*, 31(2):177–192.

Jones, K. S., Walker, S., and Robertson, S. E. (2000). A probabilistic model of information retrieval: development and comparative experiments. *Inf. Process. Manage.*, 36(6):779–808.

Kempe, D., Kleinberg, J., and Tardos, E. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, pages 137–146, New York, NY, USA. ACM.

Kerr, B. (2003). Thread arcs: an email thread visualization. In *Proceedings of the Ninth annual IEEE conference on Information visualization*, INFOVIS'03, pages 211–218, Washington, DC, USA. IEEE Computer Society.

King, I., Li, J.-Z., Xue, G.-R., and Tang, J., editors (2009). *SWSM '09: Proceedings of the 2nd ACM workshop on Social web search and mining*, New York, NY, USA. ACM. 605099.

Klimt, B. and Yang, Y. (2004). Introducing the enron corpus. In *First Conference on Email and Anti-Spam (CEAS).*

Kumar, R., Mahdian, M., and McGlohon, M. (2010). Dynamics of conversations. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 553–562, New York, NY, USA. ACM.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 591–600, New York, NY, USA. ACM.

Lam, D., Rohall, S. L., Schmandt, C., and Stern, M. K. (2002). Exploiting e-mail structure to improve summarization. Massachusetts Institute of Technology.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. volume 33, pages 159–174. blackwell publishing ltd.

Lee, C., Kwak, H., Park, H., and Moon, S. (2010). Finding influentials based on the temporal order of information adoption in twitter. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 1137–1138, New York, NY, USA. ACM.

Lewis, D. D. and Knowles, K. A. (1997). Threading electronic mail: a preliminary study. volume 33, pages 209–217, Tarrytown, NY, USA. Pergamon Press, Inc.

Losee, R. M. (1998). *Text retrieval and filtering: analytic models of performance.* Kluwer Academic Publishers, Norwell, MA, USA.

Magnani, M., Montesi, D., Nunziante, G., and Rossi, L. (2011). Conversation retrieval from twitter. In *Proceedings of the 33rd European conference on Advances in information retrieval*, ECIR'11, pages 780–783, Berlin, Heidelberg. Springer-Verlag.

Magnani, M., Montesi, D., and Rossi, L. (2010). Information propagation analysis in a social network site. In *Advances in Social Networks Analysis and Mining(ASONAM)*, pages 296–300.

Magnani, M., Montesi, D., and Rossi, L. (2012). Conversation retrieval for microblogging sites. *Information Retrieval*, 15(3-4):354–372.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval.* Cambridge University Press, New York, NY, USA.

Mischaud, E. (2007). *Twitter: Expressions of the Whole Self.* MSc dissertation, London School of Economics and Political Science.

Muresan, S., Tzoukermann, E., and Klavans, J. L. (2001). Combining linguistic and machine learning techniques for email summarization. In *Proceedings of the 2001 workshop on Computational Natural Language Learning - Volume 7*, ConLL '01, pages 19:1–19:8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Newman, P. S. and Blitzer, J. C. (2003). Summarizing archived discussions: a beginning. In *Proceedings of the 8th international conference on Intelligent user interfaces*, IUI '03, pages 273–276, New York, NY, USA. ACM.

Oreilly, T. (2007). What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *Communications & Strategies*, (65):17+.

Porter, M. F. (1997). Readings in information retrieval. chapter An algorithm for suffix stripping, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Rambow, O., Shrestha, L., Chen, J., and Lauridsen, C. (2004). Summarizing email threads. In *Proceedings of HLT-NAACL 2004: Short Papers*, HLT-NAACL-Short '04, pages 105–108, Stroudsburg, PA, USA. Association for Computational Linguistics.

Robertson, S. E. and Jones, K. S. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–146.

Robertson, S. E. and Sparck Jones, K. (1988). Document retrieval systems. chapter Relevance weighting of search terms, pages 143–160. Taylor Graham Publishing, London, UK, UK.

Robertson, S. E., Walker, S., and Beaulieu, M. (2000). Experimentation as a way of life: Okapi at trec. volume 36, pages 95–108, Tarrytown, NY, USA. Pergamon Press, Inc.

Rocha, E., Francisco, A. P., Calado, P., and Pinto, H. S. (2011). On tweet/retweets time lags. In *Making Sense of Microposts (MSM2011) Workshop at ESWC 2011*. CEUR.

Rudy, I. (1996). *A critical review of research on electronic mail*. European Journal of Information Systems, University of Cambridge, The Judge Institute of Management Studies, Trumpington Street, Cambridge CB2 1AG, UK.

Salton, G. (1971). *The SMART Retrieval System Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Salton, G. and McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA.

Saracevic, T. (1970). The concept of relevance in information: A historical review. pages 111–151, New York: R.R.Bowker. Introduction to Information Science.

Scellato, S., Mascolo, C., Musolesi, M., and Crowcroft, J. (2011). Track globally, deliver locally: improving content delivery networks by tracking geographic social cascades. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 457–466, New York, NY, USA. ACM.

Seo, J., Croft, W. B., and Smith, D. A. (2009). Online community search using thread structure. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 1907–1910, New York, NY, USA. ACM.

Shen, D., Yang, Q., Sun, J.-T., and Chen, Z. (2006). Thread detection in dynamic text message streams. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 35–42, New York, NY, USA. ACM.

Smeaton, A. F. (1990). Natural language processing and information retrieval. *Inf. Process. Manage.*, 26(1):19–20.

Song, S., Li, Q., and Zheng, N. (2010). A spatio-temporal framework for related topic search in micro-blogging. In *Proceedings of the 6th international conference on Active media technology*, AMT'10, pages 63–73, Berlin, Heidelberg. Springer-Verlag.

Sparck Jones, K. and Willett, P. (1997). *Readings in information retrieval*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Tamine-Lechani, L. and calabretto, S. (2008). Recherche d'information contextuelle et web. In *Recherche d'information: état des lieux et perspectives*, pages 201–224. Hermes.

Teevan, J., Ramage, D., and 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, pages 35–44, New York, NY, USA. ACM.

Turtle, H. R. and Croft, W. B. (1991). Evaluation of an inference network-based retrieval model. *ACM. In Transactions on Information Systems*, 9:187–222.

Wang, D., Wen, Z., Tong, H., Lin, C.-Y., Song, C., and Barabási, A.-L. (2011). Information spreading in context. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 735–744, New York, NY, USA. ACM.

Wang, L. and Oard, D. W. (2009). Context-based message expansion for disentanglement of interleaved text conversations. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 200–208, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wang, Y.-C., Joshi, M., Cohen, W., and Rose, C. P. (2008). Recovering implicit thread structure in newsgroup style conversations. In *International Conference on Weblogs and Social Media*. The AAAI Press.

Wenger, E. (1998). *Communities of Practice: Learning, Meaning, and Identity (Learning in Doing: Social, Cognitive and Computational Perspectives)*. Cambridge University Press, 1 edition.

Xi, W., Lind, J., and 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, pages 394–401, New York, NY, USA. ACM.

Yang, J. and Leskovec, J. (2011). Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 177–186, New York, NY, USA. ACM.

Yang, W.-Y., Cao, Y., and Lin, C.-Y. (2009). A structural support vector method for extracting contexts and answers of questions from online forums. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 514–523, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yeh, J.-Y. (2006). Email thread reassembly using similarity matching. In *Conference on Email and Anti-Spam*.

Zhou, L. and Hovy, E. (2005). Digesting virtual "geek" culture: the summarization of technical internet relay chats. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 298–305, Stroudsburg, PA, USA. Association for Computational Linguistics.