REVEALING MICROBLOGGER INTERESTS BY ANALYZING CONTRIBUTIONS

by

Duygu Saide Akman

BS, Computer Engineering, Bogazici University, 2006

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Master of Science in Computer Engineering

Boğaziçi University

2010

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to my supervisor Dr. Suzan Üsküdarlı for her time, guidance, and understanding. You provided me with many helpful suggestions, important advice and constant encouragement. It has been an honor and a pleasure to work with you.

I am indebted to my colleagues Melkon and Arda for helping me in performance issues. You always had time to help no matter how busy you were.

I am grateful to Tolga, for his support and patience.

Special thanks and appreciation to my dearest friends, Elçin, and Gökhan for their continual encouragement.

I would like to thank Dr. T. B. Dinesh for his contributions during the final preparation of this menuscript, and all members of SosLab for their contributions, and sharing their knowledge.

For financial support, I thank TÜBİTAK, The Scientific and Technological Research Council of Turkey. This work also is partially funded by B.U. Research Funds (BAP 08A103 and BAP 09HA102P).

The most special thanks goes to my parents, my grandfather, and my grandmother. You gave me your unconditional support and love through all this long process.

ABSTRACT

REVEALING MICROBLOGGER INTERESTS BY ANALYZING CONTRIBUTIONS

Personal blogs are online diaries. Bloggers share their comments, opinions, feelings and experiences on their blogs. People, who share similar interests but typically do not know each other in person, follow each other's updates through their blogs.

Microblogging is a kind of blogging in which users' contributions consist of shorter messages. Microblogs may express what the microblogger is doing or thinking, or inform about something like news, entertainment, good deals, etc. Since microblogging is suitable for mobile use, and short microblog contributions do not require much attention as long, well-structured blog posts, microbloggers tend to post their updates more frequently than regular bloggers, which results in a larger number of microblog posts. As a result, the microblogosphere presents a vast amount of short messages that arrive at high speed.

In microblogging systems, there is the problem of finding users of interest – as people are multifaceted and often escape notice. When deciding whether to follow a user who may be following us or followed by a friend, it would be useful to know something about them. Usually, a person who wants to get an opinion about a microblogger can look at the metadata supplied by the system, examine other microbloggers in communication with that particular microblogger, or read her contributions. A user's dynamic and continuously updated contributions reveal her interests in that particular system. However analyzing microblog posts is more difficult than analyzing blog posts. Compared to well written, structured blogger posts, microblogger posts are restricted

in size and plenty. The sheer volume and fragmented nature of microblogs make it difficult to assess the characteristics and interests of a user.

Manually analyzing microblog contributions would be overwhelming due to their quantity and fragmented nature. In this study, a model for automatically revealing microbloggers' characteristics and interests is proposed. Proposed approach supports the following:

- analyze all significant words uttered in posts,
- analyze external references existing in posts,
- analyze internal references existing in posts, and
- examine user meta information in the microblogging system

An implementation of this model, which uses the API of the highly successful and widely used microblogging service Twitter is presented.

The results of this work are discussed in terms of determining the specific characteristics of particular groups of users as well as the comparison of individual microblogger contributions. Such information could be utilized in deciding who to follow for what purpose.

ÖZET

KATKILARIN İNCELENEREK, MICROBLOG KULLANICILARININ İLGİ ALANLARININ ANLAŞILMASI

Kişisel bloglar çevrimiçi günlüklerdir. Blog sahipleri yorumlarını, fikirlerini, duygularını ve tecrübelerini bloglarında paylaşırlar. Birbirini kişisel olarak tanımayan, ancak benzer ilgi alanlarına sahip kişiler birbirlerinin güncellemelerini bloglar aracılığı ile takip ederler.

Microblog, güncellemelerin daha kısa mesajlardan oluştuğu bir blog türüdür. Kullanıcılar microbloglarında ne yaptıklarını, ne düşündüklerini belirtebilirler, ya da bir haber, etkinlik hakkında bilgi verebilirler. Microblog kullanımı, mobil kullanıma uygun olduğundan ve kısa microblog güncellemeleri, uzun, iyi yapılandırılmış blog güncellemeleri kadar özen gerektirmediğinden, microblog kullanıcıları blog kullanıcılarına göre daha sık güncelleme yapmaya eğilimlidirler. Bu da çok büyük sayılarda microblog güncellemelerine neden olur. Sonuç olarak microblog sistemlerinde, çok büyük miktarda, çok yüksek hızla biriken küçük güncellemeler oluşur.

Microblog sistemlerinde aynı ilgi alanlarına sahip kullanıcıları bulmak bir sorundur. Çünkü kullanıcılar genelde birden çok alan hakkında yazarlar, ve bir çok kullanıcı gözden kaçabilir. Bir arkadaşımızın takip ettiği, ya da bizi takip eden bir kullanıcıyı takip edip etmemeye karar verirken, onun hakkında bir şeyler bilmek faydalı olur. Bir microblog kullanıcısı hakkında fikir edinmek isteyen bir kişi, microblog sisteminin o kullanıcı hakkında sağladığı bilgiyi, ya da o kullanıcının iletişim halinde olduğu diğer kullanıcıları inceleyebilir, ya da o kullanıcının güncellemelerini okuyabilir. Bir kullanıcının devingen ve sürekli yenilenen güncellemeleri, o kullanıcının ilgi alanlarını ortaya çıkarır. Ancak microblog güncellemelerini incelemek, blog güncellemelerini incelemekten daha

zordur. İyi yazılmış, iyi yapılandırılmış blog güncellemeleri ile karşılaştırıldığında microblog güncellemeleri boyut olarak daha kısıtlı ve sayı olarak daha çoktur. Microblog güncellemelerinin sayıca çokluğu, ve dağınık yapısı, microblog kullanıcılarının ilgi alanlarını tayin etmeyi zorlaştırır.

Sayıları ve dağınık yapıları dolayısıyla microblog güncellemelerinin bir kişi tarafından incelenmesi yorucudur. Bu çalışmada, microblog kullanıcılarının karakteristiklerinin ve ilgi alanlarının otomatik olarak anlaşılabilmesi için bir model önerilmektedir. Önerilen yöntem aşağıdaki adımları içerir:

- güncellemelerde kullanılan dikkate değer kelimelerin incelenmesi,
- güncellemelerde geçen harici referansların incelenmesi,
- güncellemelerde geçen dahili referansların incelenmesi, ve
- microblog sistemi tarafından kullanıcı hakkında verilen bilginin incelenmesi

Bu modelin, çok başarılı ve yaygın olarak kullanılan bir microblog servisi olan Twitter'ın uygulama programlama arayüzünü kullanan bir uygulaması sunulmaktadır.

Bu çalışmanın sonucunda, belirli grup kullanıcıların karakteristikleri belirlenmiş, ve bireysel kullanıcıların microblog güncellemeleri karşılaştırılmıştır. Bu şekilde bir bilgi hangi microblog kullanıcısının, hangi amaçla takip edileceği kararını verirken kullanılabilir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

C	Total number of contributions of 30 users.
C_c	Collected contributions of 30 users. (Twitter API limit is 3200 for each user)
C_{daily}	Average daily contributions of 30 users.
H	Hashtags in collected contributions of 30 users.
L	Titles' texts of links in collected contributions of 30 users.
L_{wos}	Titles' texts without stopwords of links in collected contributions of 30 users.
M	Mentions in collected contributions of 30 users.
P	Plain texts in collected contributions of 30 users.
P_{wos}	Plain texts without stopwords in collected contributions of 30 users.
Rc_H	Ratio of hashtag usage in contributions (H/T).
Rc_M	Ratio of mention usage in contributions (M/T).
Rc_P	Ratio of plain text usage in contributions (P/T).
$Rpos_n$	Ratio of nouns in significant tokens (S_n/S).
$Rpos_{n\_u}$	Ratio of noun unidentified tokens in significant tokens ($(S_n + S_u)/S$).
$Rpos_u$	Ratio of unidentified tokens in significant tokens (S_u/S).
Rs_H	Ratio of hashtags in significant tokens ($H/H + P_{wos} + L_{wos}$).
Rs_L	Ratio of link title text in significant tokens ($L_{wos}/H + P_{wos} + L_{wos}$).
Rs_P	Ratio of plain text in significant tokens ($P_{wos}/H+P_{wos}+L_{wos}$).
$Rstop$	Ratio of stopwords in plain text and link titles ($((P - P_{wos}) + (L - L_{wos}))/(P + L)$).
S	Significant tokens ($H + P_{wos} + L_{wos}$).
S_n	Significant noun tokens.
S_{pos}	Significant noun and unidentified tokens. These tokens are used in generating users' word clouds. ($S_n + S_u$)
S_u	Significant unidentified tokens.

T	Single word tokens in collected contributions of 30 users (Links and retweet tokens are excluded).

1. INTRODUCTION

A *public space* is a kind of space where people from different backgrounds (age, gender, religion, education, economic etc.) get together for various reasons such as exchanging ideas, socializing, learning, or fun [1]. Some examples of *physical public space*s are streets, parks, museums, libraries, city centers, town squares, public beaches and playgrounds. According to their functionality, different physical public spaces serve different purposes for people. People go to Disneyland, or playgrounds for fun, where as parks or promenades offer them relaxation and recreation.

The most significant properties of public spaces are that they are easily accessible and they promote diversity. Internet itself provides these to people by its very nature. Every person who has access to Internet can benefit from almost all of the services it provides regardless of their age, gender, or background. Accessibility to the Internet can be compared to the accessibility in physical world in terms of transportation. A person who cannot go to a public space since she cannot afford a bus or plane ticket is similar to someone who cannot afford the cost of Internet. Free and paid services exist on the Internet. This is similar to physical places that do and do not charge money for entrance (e.g. city centers or streets versus some museums or art galleries).

When we look at social networking services (such as Twitter [2], and SecondLife [3]), collaborative knowledge bases (such as Wikipedia [4]), or collaborative art services (such as SwarmSketch [5]) on the Internet, we can see that they share most of the common properties of conventional (physical) public spaces, such as being flexible to change, accommodate temporal use, and provide an environment for exchanging ideas, learning and socializing. We refer to these kinds of spaces as *digital public space*s.

Another significant common characteristic of public spaces is that people contribute to them. There are various ways to contribute to physical public spaces. In the Dreaming Wall Project [6] in Milan, people send short messages [7], which are randomly displayed on a wall with a chemical reaction between an UV laser projection and

phosphorescent panels. These messages fade away in time before the eyes of gathered people. The contributions in this public space are in the form of text. In Burble London [8], a giant structure of balloons floats in the sky and moves in response to controls manipulated by public below. The control directives are the form of contribution in this public space. Street graffiti is an example of image contribution. However, most of the contributions in physical public spaces are simply speech - people contribute to the public space by speaking with each other. In social networking services, people contribute with text (Wikipedia [4]), pictures (Flickr [9]), videos (YouTube [10]) and other types of media.

Microblogging is a kind of social networking. As in physical public spaces, microblogging services have their own rules, and their own visitors - the microbloggers. The people in a physical public space communicate by talking and listening to each other, whereas the microbloggers communicate by sending posts to general space and reading others'. Subscription is the key mechanism to interact in microblogging systems. For example, in the popular microblogging service, Twitter, user A's contributions are displayed on user B's home page, if user B has subscribed to posts of user A. Subscriptions between users may be asynchronous in microblogging systems. In the previous example, user A may or may not have subscribed to posts of user B. This can be compared to public speaking in a physical spaces. A person may be interested in what the speaker is talking about, and listen, but the speaker may not be interested in the ideas or comments of the listener. Also, the influence range of a speaker in a physical public space is restricted by the number of people who can hear her voice. This range is equal to the number of subscribers of a Twitter user.

In physical public spaces, we meet many others who we do not know personally. We try to get to know them by watching their gestures, and listen to their conversations. These are their contributions in physical spaces. Given sufficient information, we get an idea about their characteristics and their interests. Then we decide whether to communicate with those strangers. In digital public spaces, we also evaluate users by their contributions.

Microblogs are very small posts. The limited length of microblogging contributions allow users to post their messages as short messages via mobile phones, or via web through various applications and web pages. Since microblogging is suitable for mobile use, and it is easier to post short microblog contributions than long, well-structured blog posts, microbloggers tend to post their updates more frequently than regular bloggers, which results in a larger number of microblog posts. The sheer volume and fragmented nature of microblogs makes it difficult to assess the interests of a user.

When deciding to reciprocate a following microblogger or upon encountering a reference to a microblogger, it would be useful to know something about them. Furthermore, it would be nice to know what all a followed user contributes about the most– as people are multifaceted. Another criteria of interest is whether the candidate to follow is a human or autonomous agent?

This study proposes an approach for examining the nature of contributions and the characteristics of a microblogger. In describing a microblogger only their contributions are utilized and any interests they don't contribute in their microblogs are not of interest. The goal of this work is to identify the interests that a microblogger contributes about and, therefore, can be followed in the microblogs.

A model for describing users in terms of their contributions is proposed (see Chapter 5). An implementation of this model is developed using the popular microblogging system Twitter (see Chapter 6).

2. BACKGROUND

2.1. Blogging and Microblogging

Blogging is one of the activities that have become popular with Web 2.0. According to Wordnet [11] a blog *is a shared on-line journal where people can post diary entries about their personal experiences and hobbies.* The most significant characteristics of blogs are that they are usually maintained by a single author, and the blogs are time stamped entries that are presented in reverse-chronological order.

Microblogging is a kind of blogging in which users' contributions consist of very short messages. It has become very popular with contributors ranging from average persons to celebrities to commercial organizations. Individuals users such as politicians, actors, musicians, academicians, students use it regularly. Organizations such as businesses, institutions, and activists use it as well.

Microblogs may express what the microblogger is doing or thinking. Microblogs may also inform about something like news, entertainment, good deals, etc. Microblogs that inform typically provide reference to an external resource, since their limited size is insufficient to convey the news. Broadcasting is spreading information over a large range of audience. Microblogs can be used to broadcast just about anything its contributor desires.

Microblogs are especially suitable for mobile users since it is very easy to make small contributions in mobile circumstances. The increasing support for the *always-on* internet access and rich media content generation intensifies the allure for participating in such platforms. References to pictures, audio, and videos are shared in microblogs with web links. As a result the microblogosphere presents a vast amount of short messages that arrive at high speed, which are user-filtered by follow relationships. The consequence of this quick, easy, anytime, and anywhere publishing is the huge volume of fragmented contributions [12]. Microblogging application users choose the microblogs

they wish to view through (typically a large number of) subscriptions.

Microblogs, like weblogs, tend to be publicly accessible. Their update rates tend to be much higher – typically several times a day. They have strict size limits on the number of characters that can be contributed. Thus, microbloggers develop conventions for creating short posts, such as using abbreviations, short urls (services that dramatically shorten regular URLs), omitting words, etc. A typical contribution is: *FeedDemon no longer owned by NewsGator. http://r2.ly/kyau* with a timestamp.

The massive amount of fresh and diverse posts from a large user base has inspired many studies – such as what kind of people microblog; why and how they contribute; and identifying trends based on what is being contributed.

There are many different microblogging services such as Jaiku [13], Tumblr [14], and the most popular of all - Twitter [2].

2.2. Twitter

Twitter [2](launched in August 2006) is a highly successful and widely used microblogging application. It became very popular after it won the Web Awards of South by Southwest [15] conference in March 2007.

In April 2010, at the official Twitter Developer Conference - Chirp [16], the following statistics regarding the popularity of Twitter were revealed [17]:

- 105,779,710 registered users.
- New users sign up at the rate of 300,000 per day. (Of the new accounts, over 60 percent come from outside the US [18])
- 180 million unique visits every month.
- 75 percent of Twitter traffic comes from outside Twitter.com (i.e. via third party applications.)
- 3 billion requests per day via Twitter API.

- An average of 55 million tweets a day.
- Approximately 600 million search queries per day.
- Of active users, 37 percent use their phone to tweet.
- In the past year, the Twitter company has grown from 25 to 175 employees.

In Twitter, posts - *tweet*s, are limited to 140 characters. Posts are composed of plain text, links, and keywords that have a special meaning in Twitter (hashtags, mentions, and retweets). Hashtags are single word tokens that are preceeded by a hash symbol ('#'). Hashtags can occur anywhere in a tweet. Hashtags are used to tag a tweet, and a tweet can only be hashtagged by its creator. Mentions are Twitter usernames that are preceded by an at symbol ('@'). A twitterer, who wants to reference a user, does so with the pattern @¡username¿. Retweets are used whenever a twitterer wants to spread a tweet. To denote that a tweet is a repeat(retweet) of another tweet, twitterers use RT or RETWEET in their tweets (In 2009, Twitter developed a new feature for retweets. Instead of adding keywords, a twitterer who wants to retweet a tweet can just click the retweet link next to the original tweet. An icon is automatically placed next to her tweet, denoting that it is a retweet. RT and RETWEET keywords can be seen used in older tweets).

Unlike earlier social applications, where users are privy to each other's contributions through friendship networks, Twitter supports unidirectional (asymmetric) follow relationships. Microblogger M_1 is a *follower* of Microblogger M_2 if she follows the updates of Person B, who does not necessarily follow Microblogger M_1. Twitter users-*twitterers* - choose whether to publish their tweets publicly or privately. In the latter case, only the followers of a user are allowed to see the tweets, whereas in the former, updates are published in the *public timeline*, making them visible to anyone. Majority of Twitter feeds are public.

Among its millions of users, celebrities and organizations use Twitter such as the U.S. president Barack Obama, actor Ashton Kutcher, and Google. Automated agents such as CNN Breaking News, and CFA (Country Fire Authority) [19] use Twitter to share the recent news with their followers. The microblogger sfearthquakes posts

earthquake news in SF Bay area [20]. The asymmetric relationship in Twitter enables the number of followers of celebrities and organizations to reach millions (In April 2010, The Twitter accounts BarackObama, aplusk, google, and cnnbrk have well over 3.7, 4.8, 2.2, and 3.0 million followers respectively).

One needs a name, a username, a password, and a valid mail address for creating a Twitter account. After getting an account, a user can optionally specify a time zone, a personal URL, a one line bio, location, picture, and language. Also there are settings for protecting users' updates, and settings for email notifications. Notification settings include being notified of new followers and new direct messages via email.

As well as the Twitter web page itself, tweets can be sent via mobile texting, instant messaging services, and third party applications and other web applications that use Twitter API [21]. Numerous web and mobile applications streamline content creation and microblog posting, such as Twitpic [22] for sharing photos on Twitter and TweetDeck [23] that integrates Twitter with the popular social networking services Facebook [24] and Myspace [25]. Accordingly, users can choose to read the tweets of the people that they follow via their Twitter home pages, IM, or applications on their personal computers or their mobile phones, where the latest is the one that makes Twitter that much mobile and popular.

3. RELATED WORK

Microblogging, as a social media tool, has gained an enormous interest among people and commercial organisations. With the rise of its popularity, microblogging has been studied by researchers in various areas. For example in a study by Sandler et al.[26], current limitations of microblogging services are investigated and a more efficient protocol is proposed.

Further studies investigate microblogging usage in different contexts:

- for mobile learning and educational purposes [27, 28],
- for scientific writing [29],
- for collaborative work [30],
- for informal communicating at work [31],
- as a communication tool for health librarians [32]

Over recent years, Twitter, the most popular microblogging service, has been investigated in many studies.

In [33], Huberman et al. discuss whether online social networks really represent actual social interactions. They study social interactions within Twitter. They found the existence of two different types of networks: a dense one made up of followers and followees, and a smaller network of actually interacting friends.

Microblogging helps retrieving, producing, and spreading information. In a study of Vieweg et al.[34], Twitter posts generated during the Oklahoma Grassfires and the Red River Floods were analyzed. They identified the features of information generated during emergencies for the development of software systems that employ information extraction strategies.

In their study[35], Jansen et al. investigate microblogging as a form of online word

of mouth branding. About 150,000 tweets were analyzed and it was concluded that microblog posts provide valuable competitive intelligent information to brand owners, like changes in sentiments for brands.

In [36], Honeycutt et al. analyzed a corpus of about 35,000 tweets, focusing on the mentions(uses of the '@' sign). The results of this study are:

- Different language groups make use of the @ sign with almost equal frequency.
- More than 90% of the @ signs in English tweets were used to direct a tweet to a specific addressee.
- Tweets with @ signs are more interactive. On the other hand, tweets without @ signs are more self-focused, and they make more general announcements.
- 31.2% of the tweets that include the @ sign -to direct the message to a particular individual- received a public response in half an hour.

In [37], Shamma et al. investigated Twitter posts during the 2008 Presidential Debates. They discovered that the structure of Twitter traffic can provide insights into changes in topics in the media event.

A. Java et al., [38], categorize Twitter users based on link structures:

- Information Source: Automated or human agents, who are categorized as information sources post valuable content and/or they post frequently, and they tend to have a large number of followers.
- Friends: Most relationships fall into this broad category. Such users follow friends, family and co-workers.
- Information Seeker: These type of users post rarely, but follow others regularly.

In the same study[38], tweets are manually categorized as follows:

- Daily Chatter: About daily routines or present activity.
- Conversations: Tweets that have *mention*s in them.

- Sharing Information/URLs: Tweets that have URLs in them.
- Reporting News: Users who report latest news or comment about current events. Automated agents that post updates like weather reports and news stories from RSS feeds fall into this category.

In a study by B. Krishnamurthy et al.[39], users of Twitter are categorized as *broadcaster*s, *acquaintance*s, or *miscreant/evangelist*s according to two criteria: The number of microblogs users follow and how many users follow their microblogs, and second, the number of tweets of users.

- Broadcasters have a much larger number of followers than the ones they follow. Plus, they tend to tweet a lot. This category includes online radio stations' automated users, and news sources such as New York Times, BBC, etc.
- Acquaintances tend to exhibit reciprocity in their relationships, meaning that the number of users they are following and the number of users that are following them are close to each other.
- Miscreants / Evangelists follow a much larger number of people than they have followers. They tend to tweet less than other types of users.

There is Research on Twitter users' influence. In [40], Weng et al. propose a ranking algorithm to measure the topic-sensitive influence of twitterers. In [41], Lee et al. propose considering both the link structure and the temporal order of information adoption in Twitter for finding influentials.

4. MOTIVATION AND PROBLEM STATEMENT

Personal blogs are online diaries. Bloggers share their comments, opinions, feelings and experiences on their blogs. People, who share similar interests but typically do not know each other in person, follow each other's updates through their blogs.

Microblogging is a kind of blogging in which users' contributions consist of shorter messages. This size limitation allows users to update their microblogs and read others' via their mobile devices. Being able to use microblogging systems anytime and anywhere, microbloggers tend to post their updates more frequently than regular bloggers.

In microblogging systems, it is difficult to

- locate microbloggers who contribute regarding a specific topic,
- locate microbloggers who have a specific characteristic,
- discover what a microblogger – whose references is encountered – contributes about, and
- discover what a microblogger's characteristics are.

In the former two cases, there is an intent to locate microbloggers of desired kind (such as automated agents or celebrities) who are actively contributing in a given area, i.e. birdwatching, Android OS, etc. In the latter two cases, a reference to a microblogger is somehow encountered and the context of the encounter has motivated the user to find out more about that microblogger. It is common to come across references in other posts, emails, and web pages.

A microblogger's characteristics and contribution topics may be revealed by looking at

- user's self provided information,
- people in communication with that user,

- information supplied by the system about that user, or
- user's own contributions

Homepage and biography sections are examples of self provided information. These sections are typically optional, so, microbloggers may prefer not to share such information on their profiles. Besides, the context of homepages and biographies may be different from their contributions on microblogging systems. For describing microbloggers, self provided information may be missing or misleading. Any avocations shared on other platforms, but not on that particular microblogging system cannot be used to correctly describe that microblogger's subjects. For example, a professor who likes bowling may list his publications and the courses he teaches in his homepage, while sharing his ideas and experiences in bowling in his microblog posts. By looking at his homepage only, Person A, searching for a microblogger writing about bowling, can skip this professor.

One may get an idea of a microblogger's interests by looking at the people she is in communication with. A common friend may be a sign of common interests, or the common interests of people following her may reveal her interests. In microblogging systems, following other microbloggers is costless. Thus, a microblogger may follow many others, even the irrelevant ones, without deliberation. Also, generally, microbloggers allow other microbloggers to follow their updates without authorization. This results in uncontrolled followers, i.e. spammers or automated agents. For describing microbloggers, checking every follower /followee is exhausting since their numbers may reach to millions and is not very informative.

Microblogging systems may also supply information about its users. This may include name, location, user creation date, number of posts, number of followers, and such. This kind of data is usually only numerical and /or generic, so that is insufficient for describing microbloggers' subjects. However, these statistical data can be used to understand microbloggers' general nature.

A microblogger's subjects can best be revealed by analyzing user's dynamic and

continuously updated contributions. However analyzing microblog contributions is more difficult than analyzing blog posts. Compared to well written, structured blogger posts, microblogger posts are restricted in size and plenty. A person, who wants to describe a microblogger needs to combine all those fragmented microblog posts. This makes it difficult to identify a microblogger's characteristics, and subjects of contributions.

This study proposes a method for automatically revealing microbloggers' characteristics and subjects by analyzing information supplied by the system and users' own contributions.

5. MODEL

Manual inspection of microbloggers to identify what they contribute about is not feasible due to the quantity and nature of their contributions.

This work proposes an automated approach for identifying microbloggers based on their contributions.

The basic idea is to collect and process the contributions and meta-data to yield two types of identifications:

- Content-specific identification, which reveals the subjects the microblogger contributes about
- General identification, which provides insight about the type of the contributor.

Content-specific identification concerns *what* a user contributes about. Microbloggers may focus on one or a few topics or they may contribute about numerous topics. This work proposes an approach for processing contributions in order to yield a weighted list of words representing relevant topics.

Determining a set of relevant words is not as easy as one may think. An inspection of the nature of microblog contributions shows that they contain:

- approximately seven words
- abbreviations and special syntax specific to microblogger communities
- many references to external links, internal links, other users
- special names
- very up to date concepts and instances.

Furthermore they are grammatically incorrect partial sentences. Thus, processing microblog contributions presents some interesting challenges. An approach to these

challenges is detailed in this chapter.

Microblogging systems supply quantitative meta information about its users, i.e. total number of contributions, date of account creation, etc. Other quantitative information can also be calculated from user contributions, i.e. the number of other microbloggers in communication. In this model, users are characterized as *automated*, *spam*, *bot*, *celebrity*, or *social*.

- *automated* users are software agents acting as microbloggers. These users can update their microblogs more frequently than regular human users. Also, software agents can be used for advertisement or information purposes. These agents include a single or a few external domains in most of their contributions. A user is categorized as *automated* if her ac or fdd is above a threshold value. *bot*s are automated users which usually provide beneficial information such as breaking news, traffic conditions, or weather. These users do not subscribe to many users, but lots of other users subscribe to *bot*s. A user is categorized as *bot* if that user has been categorized as *automated* and her rs is above a threshold value. *spam*s are automated users which rarely provide beneficial or interesting information. These users subscribe to many users for advertising purposes, but a few other users subscribe to *spam*s. A user is categorized as *spam* if that user has been categorized as *automated* and her rs is below a threshold value.
- *celebrity* users are well-known non-automated users. Due to their popularity, many people subscribe to these kind of users, while they subscribe to a small number of other users. A user is categorized as *celebrity* if that user has not been categorized as *automated* and her rs is above a threshold value.
- *social* users are in communication with many other users. They are chitchatters, and they mention many other users in their contributions. A user is categorized as *social* if the number of other users she mentions is above a threshold value. Automated and non-automated users can be social.

Due to the limited size of microblogging, external and internal references are highly used in posts. These references contain important information for revealing

microbloggers' subjects. In this model, for revealing user's subjects of contributions, Due to the limited size of microblogging, external and internal references are highly used in posts. As well as the words uttered by users themselves, these references contain important information for revealing what microbloggers talk about. In order to reveal user's subjects of contributions, in this model,

- words uttered in posts are collected,
- information in external and internal references are analyzed,
- a set of candidate tokens is constructed,
- words that are irrelevant to user's subjects are discarded from candidate token set, and
- human-readable, visual weighted token sets revealing what microbloggers contribute about are generated.
- Resulting weighted token sets of different users are compared using cosine similarity measures.

Figure 5.1 shows the overview of the processing for generating user tagging.

The basic idea is to analyze the contributions of a microblogger in order to reveal a set of words (keywords) that identify their subjects. A microblogger may contribute regarding a few or several subjects. The distribution of the weights of the words used by a microblogger who predominantly contributes regarding a specific area will be different than someone who contributes about many different topics.

5.1. Microblogging System Specification

In order to describe microblogger analysis, some types corresponding to their fundamental aspects are introduced. A microblogging system essentially consists of a set of users, a set of microblogs, and a set of subscriptions between users and microblogs. *IRef* is an internal reference whose syntax whose denotation is microblogging system specific (In Twitter, hashtags are used for this purpose. They are represented with a hash sign (#) followed by a sequence of characters – i.e. #myTag). Microbloggers

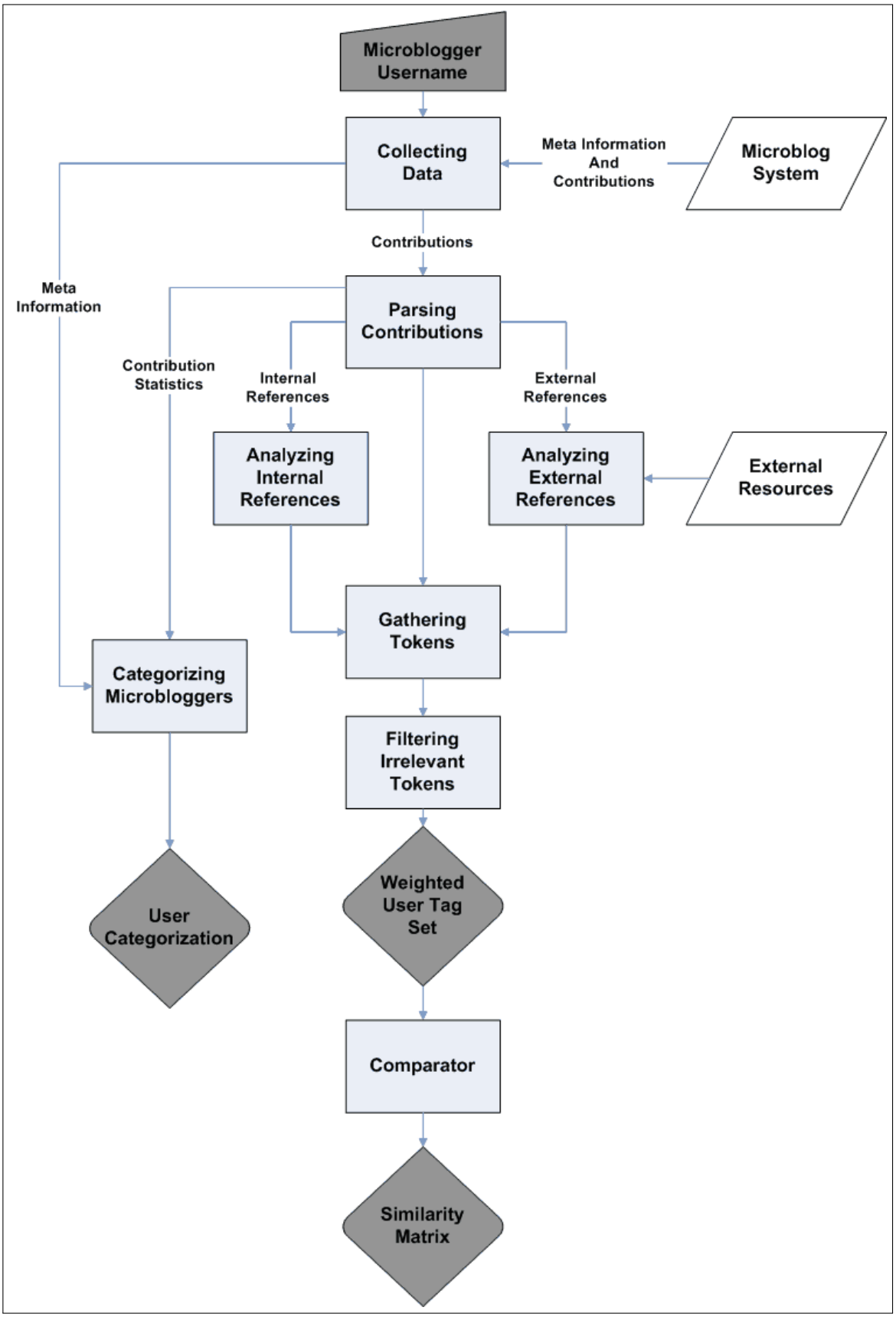

Figure 5.1. System architecture for microblogger tagging.

explicitly or implicitly choose to use them in contributions of a given characteristic.

In order to auto-tag users, their contributions are processed. In order to discuss the fundamentals of a microblogging system various types (Figure 5.2) and functions (Figure 5.3) are introduced. Essentially a microblogging system consists of a set of users, a set of microblogs and a set of subscriptions. Each user has a microblog, which consists of a sequence of posts. Posts are time stamped small textual contributions.

The proposed auto-tagging approach consists of gathering the posts of a user and process them so as to discover which words (more accurately tokens) they use and how frequently they use them. The processing of the posts is dependent on what and how microbloggers contribute. Microblogging is social activity and is subject to a strict size limitation, which strongly influences the nature of contributions. An analysis of microblog content revealed that indeed many socialization related words were used, the average number of words in a post is approximately 7, and many stop words are used.

The idea is to create a tag cloud out of what users with what they contribute. When contributions are parsed, many kinds of tokens are attained – such as words in natural language, tokens that have special meaning within the microblogging system (The tokens RT, RETWEET are examples of tokens with special meaning in Twitter), references to other users, and external links. Since the tag cloud is meant to describe a user, irrelevant tokens must be removed from the set. The following types of tokens are deemed insignificant for the purpose of tagging users.

- verbs, adverbs, adjectives
- internal references
- stop words – the words used so commonly that they have no distinguishing property

The function *pos* returns the *part of speech* (pos) related to a given token. In this model *Noun* and *Unidentified* are considered significant types of words and *Verb*, *Adverb*, and *Adjective* are ignored.

Type name	Specification
MicroblogSystem	$\langle Users, Microblogs, Subscriptions \rangle$
User	$\langle Name, SelfDescription, CreationDate \rangle$
Users	$\{u_1, u_2, ...\} \mid u_i : User$
Contribution	text of limited length
Post	$\langle User, Contribution, TimeStamp \rangle$
Microblog	$\langle u, \{p_1, p_2, ...\} \rangle where \;\; u \in Users, p_i : Post.$
Microblogs	$\{m_1, m_2, ...\} \mid m_i : Microblog$
Subscription	$\langle u, \{m_1, m_2, ...\} \rangle where \; u \in Users \wedge m_i \in Microblogs.$
Subscriptions	$\{s_1, s_2, ..., s_n\} \mid s_i : Subscription$
Token	is a sequence of non space characters
Label	"ExternalRef" | "UserRef"|"InternalRef"|"Plain"
LabeledTokens	$\{t_1 : l_1, t_2 : l_2, \cdots\} where \;\; t_i : Token \wedge l_i : Label$
Stopwords	$\{w_1, w_2, ...\} \mid w_i : Token$
ExternalRef	$\{r_1, r_2, ...\} \mid r_i : URL$
InternalRef	internal reference as represented by the microblogging system
wTag	$\{(w_1, wt_1), (w_2, wt_2), ...\} \mid w_i : Token \wedge wt_i : Integer$

Figure 5.2. Main data types of microblog systems.

Internal references are references users select to associate with posts relevant to a given criteria (In the microblogging system Twitter internal references are hastags denoted by bf # followed by a sequence of characters, i.e. #www2010). They are used as a collective filtering mechanism for posts.

Microbloggers are described with a set of weighted tags. These tags are derived from the microblogger's contributions. This model focuses on describing microbloggers in terms of what they say (more accurately, what they contribute). This approach is sensible only if microbloggers do contribute and contribute consistent with their interests, rather than lurk or simply repeat what others say. Accordingly, one of the inquires of this research is to inspect microblogger contribution behavior.

$posts(u : User) = \{p_1, p_2, ..., p_n\} where\ \langle u, \{p_1, p_2, ..., p_n\}\rangle \in Microblogs \wedge u \in Users$

$parse(c : Contribution) = \{t_1, t_2, ...\} \mid t_i : Token$

$filter(P : Posts) = parse(P) - Stopwords \mid P = \{p_1, p_2, ...\}$

$tokenFreq(t : Token, u : User) = |select(t, filter(posts(u)))|$

$Iref(m_i : Microblog) = \{\langle t_1, val_1\rangle, \langle t_2, val_2\rangle \cdots\} \mid t_i \in tokens(user(m_i)) \wedge$

$isIref(t_i) = true \wedge val_i = |t_1|$

$isExternalRef(t : Token) =$ **true** if t is an url **false** otherwise.

$externalRefs(p : Posts) = \{e_1, e_2, \cdots\} \mid e_1 :$ ExternalRef

$partOfSpeech(token) =$ "Noun"|"Verb"|"Adverb"|"Adjective"|"Unidentified"

$subscriptions(u : User) = \{m_1, m_2, ..., m_n\} \mid \langle u, \{m_1, m_j, ...\}\rangle \in Microblogs$

$activity(u : User) = \frac{|posts(u)|}{date(lastPost(u)) - registrationDate(u)}$

Figure 5.3. Functions for processing microblogs.

5.2. Processing Microblogger Contributions

In order to gain insight about the nature and subject(s) of microbloggers, their contributions must be gathered, filtered, and analyzed. Basically the contributions are processed in order to identify a list of weighted words, which are considered significant. The computation of a set of weighted words involves the following steps: collecting and parsing user contributions, extracting information from the content in external references, analyzing internal references, tokenizing contributions, removing insignificant tokens, categorizing users. Given such microblogger descriptions, the similarity between two microbloggers are computed with a comparison function (Section 5.4). These tasks are further described in the following sections.

5.2.1. Collecting and Parsing User Contributions

For revealing what a user is contributing about, user's microblog posts are gathered. Also, a user's meta information shared by the system is retrieved for understanding the characteristics of the user.

Microbloggers are described based on their posts. This is achieved by examining *what* and *how* they contribute. Their posts are analyzed for this purpose. The aim is to reduce the posts of a microblogger to a set of weighted tokens. These tokens are considered to be tags. The weighted set of tags can be considered tag clouds, which are commonly used to summarize keywords. Algorithm 5.4 describes how a set of posts is reduced to a weighted set of tokens.

5.2.2. External Reference Analysis

In microblogging systems, length of a post is very limited. Consequently, the real content typically is at the external resource. External references may refer to news, conferences, videos, pictures, and much more. These references are important in that they contain what the microbloggers are not able to express in the limited space.

In this module, for each token labeled as "ExternalRef" are taken as metadata, and data in those references are fetched and analyzed. The result is a multiset, where the elements are single word tokens, and multiplicities are the number of occurences of the respective tokens in the resulting data.

$$\text{ExternalReferences}(u : User) = \{\langle t_1, val_1\rangle, \langle t_2, val_2\rangle, ...\} \mid t_i : Token \wedge val_i = |select(t_i, filter(posts(m)))|$$

5.2.3. Internal Reference Analysis

Different microbloggers' contributions surrounding the same interest (event, topic, person, etc.) can be organized under related topics. Internal reference points in microblogging systems are used for this reason. These internal reference points are community generated. A microblogger initiates a category and lists her related contributions under that category. Since single users' contributions are visible to a wide audience, that category gets adopted by other users. Other microbloggers interested in a particular topic can go to that internal link and see all related contributions. Internal reference points also help users sharing similar interests to find and follow each other.

```
for each post in posts(user) do
   contributionWithoutER, tokenList:  Tokens[]
   contributionWithoutER ⇐ []
   tokenList ⇐ parse(post)
   for each token in tokenList do
      if isExternalReference(token) then
         label(token, ‘‘ExternalRef’’)
      else
         contributionWithoutER.append(token)
      end if
   end for
   contributionWithoutER ⇐ removePunctuation(contributionWithoutER)
   for each token in contributionWithoutER do
      if isUserReference(token) then
         label(token, ‘‘UserRef’’)
      end if
      if isInternalReference(token) then
         label(token, ‘‘InternalRef’’)
      end if
      if isSystemSpecificToken((token)) then
         label(token, ‘‘Special’’)
      else
         label(token, ‘‘PlainText’’)
      end if
   end for
end for
```

Figure 5.4. Parsing Contributions Algorithm

In this module, each token labeled as “InternalRef” are analyzed to have a better opinion about user’s contribution subjects. The result is a multiset, where the elements are single word tokens, and multiplicities are the number of occurences of the respective tokens in the resulting data.

Iref$(m_i) = \{\langle t_1, val_1\rangle, \langle t_2, val_2\rangle \cdots\} \mid t_i \in tokens(user(m_i)) \wedge isIref(t_i) = true \wedge val_i = |t_i|$

5.2.4. Gathering Tokens

WeightedTokens $= \{\langle t_1, val_1\rangle, \langle t_2, val_2\rangle, ...\} \mid t_i : Token \wedge label(Token) =$ “Plain” $\wedge$ $val_i : |t_i|$

PlaintText$m_i = \{\langle t_1, val_1\rangle, \langle t_2, val_2\rangle \cdots\} \mid t_i \in tokens(user(m_i)) \wedge label(t_i) =$ “Plain” $\wedge$ $val_i = |t_i|$

CandidateTokens $= \{\langle t_1, val_1\rangle, \langle t_2, val_2\rangle \cdots\} \mid (label(t_i) =$ ExternalRef $\vee$ InternalRef $\vee$ Plain) $\wedge$ $val_i = |t_i|\}$

5.2.5. Filtering Irrelevant Tokens

Tokens that are considered irrelevant to the subjects of user contribution are deleted from CandidateTokens. The resulting set is the desired weighted user tag set.

UserTags $\subset$ CandidateTokens:

$UserTags = CandiateTokens - StopWords - verbs(CandidateTokens) - adverbs(CandidateTokens) - adjectives(CandidateTokens)$

where the functions verbs, adverbs, and adjectives are functions that take a set of tokens and return a set of tokens that are verbs, adverbs, or adjectives respectively.

Category	Formula	Threshold
bot	$\tau_s(u : User) = \frac{subcriptionNo(u)}{follower(u)}$	θ_{auto}
automated	$\mu_{c/day}(u : User) = \frac{contributionNo(u)}{lastContribution(u) - creationDate(u)}$	$\theta_{automated}$
social	distDomains$(u : User) = max($externalReferences$(u))$	θ_{social}
celebrity	userRefs$(u : User) =$ |userReferences(u)|	$\theta_{celebrity}$

Figure 5.5. Metrics used in categorization of microbloggers.

As a result UserTags is a set of nouns, unidentified words, and internal references.

5.3. Microblogging Categorization

Microbloggers are categorized in terms of their general contribution characteristics. Contribution characteristics are used to compute a set of metrics used for this purpose. The fundamental parameters that describe microbloggers are their contributions, whose microblogs they subscribe to, and who subscribes to their microblog. While the content of the contributions and the identities of who subscribes to whom can be used for detailed analysis, the mere quantities of these parameters can also be quite informative. Furthermore, information about account creation and most recent contribution indicative levels of activity. Figure 5.5 provides a set of metrics used to broadly categorize microbloggers.

The following information is retrieved for each user:

- $subcriptionNo(u : User) = |subscriptions(u)|$
- $follower(u : User) = |\{u_1, u_2, \cdots\}| where \;\; u! = u_i \wedge u \in subscriptions(u_i)$
- $contributionNo(u : User) = |posts(u)|$
- $creationDate(u : User) = dwhere \;\; u = \langle n, p, d\rangle, n : String, p : Post, d : Date$
- $lastContribution(u : User) = date(p_n) where \;\; u = \langle n, \langle p_1, \cdots, p_n\rangle, d\rangle$

Based on these calculations, users are characterized as *automated*, *spam*, *bot*, *celebrity*, or *social*.

5.4. Comparison of Two Microbloggers

If the proposed microblogger identification is appropriate, a similarity measurement which compares two such identifications must be yield a higher score for similarly contributing microbloggers. In other words, this approach assumes that microbloggers contributing regarding similar topic use the same vocabulary.

Microblogger comparison is done with cosine similarity functions:

- Unweighted Comparison: In unweighted comparison, only the tokens in users' word clouds are taken into consideration. Weights of tokens are not used. Two set of words are constructed using top n tokens in each user's $UITS$. A similarity value is generated depending on the number of the common words in both sets.
- Weighted Comparison: In weighted comparison, weights of tokens in words clouds are taken into consideration. Two set of words, and their weights, using top n tokens in each user's $UITS$ are constructed. A similarity value is generated depending on the number of the common words and their related weights.

Comparator module uses the results of previous modules. Two users are taken as input, and their word clouds are compared. A similarity value for two users is given as output.

Results of comparator are used to evaluate our model. Weighted word tokens are generated for users who have declared similar interest in external systems. Similarities between these users are calculated and compared with manually declared similarities (see Section 7.4).

Results of comparator may further be used for automated microblogger suggestion.

6. IMPLEMENTATION

A code to analyze microbloggers in order to evaluate our approach is implemented for the popular microblogging service Twitter. Application is composed of 9 modules: Data Collector, Post Parser, External Reference Analyzer, Internal Reference Analyzer, Token Gatherer, Filterer, Categorizer, Word Cloud Visualizer, and Comparator. All modules take a Twitter username as input.

The following modules are used for revealing the characteristics of a user:

- Data Collector,
- Post Parser,
- External Reference Analyzer, and
- Categorizer

For revealing user descriptions, following modules are used:

- Data Collector,
- Post Parser,
- External Reference Analyzer,
- Internal Reference Analyzer,
- Token Gatherer,
- Filterer, and
- Word Cloud Visualizer

Finally, the Comparator module, compares microbloggers in terms of their descriptions.

The implementation uses Java programming language, MySQL Server 5.1 for database, Twitter4J library, Twitter API, WordNet 2.1, MIT Java Wordnet Interface, and IBM Word-Cloud Generator.

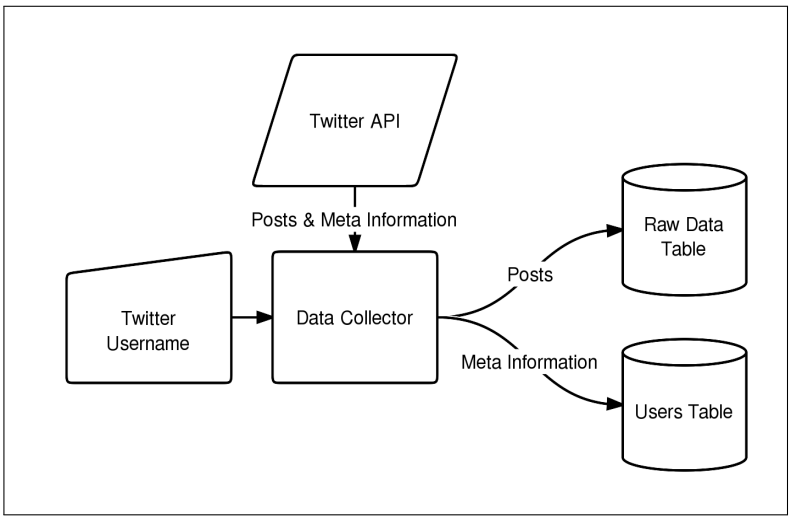

Figure 6.1. System architecture for Data Collector.

6.1. Data Collector

There are two kinds of user data gathered in the data collection module. First, the maximum number of microblog posts are gathered. Second, user's meta information shared by the system is retrieved. (see Figure 6.1)

6.1.1. Microblog Posts

The Twitter API currently allows a maximum of 3200 tweets (Twitter posts) to be fetched. Following data for each tweet is collected and stored in the raw data database table.

- tweetId (a unique id given by Twitter service for each tweet)
- userId (a unique id given by Twitter service for each user)
- userName (screen name for that user)
- text (tweet itself)
- timeStamp (day and time that the tweet was created)

6.1.2. Meta Information

The Twitter API provides access to many user details. The following information is fetched for a microblogger and stored in the users database table.

- $follower(u)$: Number of users following the user u,
- $friends(u)$: Number of users the user u is following (subscribed to),
- $postCount(u)$: The total number of tweets the user u has posted,
- $creation(u)$: The date at which the account for user u was created,
- $latestPost(u)$: The date of the most recent post (tweet) of user u.

6.2. Post Parser

Tweets can be composed of plain text, punctuation marks, URLs, hashtags, mentions, and retweet keywords. An example to a typical tweet is:

```
"RT @lynda_hardman: \#ssms10 The website for this year's Summer School on
Multimedia Semantics : http://www.smart-society.net/ssms10"
```

In this tweet, "#ssms10" denotes a hashtag. **Hashtags** are single word tokens that are preceeded by a hash symbol ('#'). Hashtags can occur anywhere in the tweet. Hashtags are used to tag (give a category) to a tweet, and a tweet can only be hashtagged by its own creator. **Mentions** are Twitter usernames that are preceded by a at symbol ('@'). A twitterer, who wants to reply to or mention another user, puts @other_user, anywhere in her tweet. **Retweets** are used whenever a twitterer wants to spread another twitterer's tweet to her followers. To denote a tweet is a retweet of another tweet, twitterers use RT or RETWEET keywords in their tweets. The example tweet is a retweet of an original tweet posted by lynda_hardman.

In Post Parser module, first, to keep the original tweets of the user, user's tweets are copied from raw data database table to analysis database table. All the parsing is

executed on analysis database table. Second, tweets of user are parsed in the following order:

Tweets are tokenized by white space characters. Tokens starting with strings *http* or *www* (case insensitive) are categorized as (LINK)s. From the rest of the tweet, punctuation marks that have no special meaning in Twitter are removed. These punctuation marks are: `! " % ( ) * + , - . / : ; < = > ? [ \ ] ^ ' { | } ~`

The tweets are re-tokenized by white space characters. The resulting tokens are categorized according to the following criteria (See Section 5.1):

- Tokens starting with @ are categorized as MENTION corresponding to "UserRef".
- Tokens starting with # are categorized as HASHTAG corresonding to "InternalRef".
- Tokens *RT* or *RETWEET* (case insensitive) are categorized as RETWEET.
- The rest of the tokens are categorized as PLAINTEXT corresonding to "Plain".

Finally, the stem of each PLAINTEXT is found using WordNet 2.1 and MIT Java Wordnet Interface (JWI). In order to stem a word, WordNet requires the part of speech of the word as input. Finding correct part of speech of tokens in tweets is out of the scope of this study. Thus, findStems function of JWI is called with every possible part of speech value(Noun, Verb, Adjective, Adverb). For each part of speech, JWI function returns a list of possible stems. Among these lists for different part of speech, Noun results are preferred the most, and Adverb the least. If the preferred list contains more than one stem, the first stem in the list is assumed the correct one and assigned to PLAINTEXT_STEM. If JWI does not return any candidate stems, PLAINTEXT itself is assigned to PLAINTEXT_STEM (see Algorithm 6.3).

All tokens labeled as either PLAINTEXT_STEM, LINK, MENTION, or HASHTAG are stored in a database table (see Algorithm 6.2 and Figure 6.4).

```
for each post in posts(user) do
   contributionWithoutLINK : String
   tokenArray : String[]
   tokenArray = []
   tokenArray ⇐ parse(post)
   for each token in tokenArray do
      if startsWith(upperCase(token),‘‘WWW’’) ∨
startsWith(upperCase(token),‘‘HTTP’’) then
         storeinDB(user, token, ‘‘Link’’, ‘‘Analysis Table’’)
      else
         contributionWithoutLINK.append(token)
      end if
   end for
   contributionWithoutLINK ⇐
removePunctuation(contributionWithoutLINK)
   tokenArray ⇐ parse(contributionWithoutLINK)
   for each token in tokenArray do
      if startsWith(token,‘‘@’’) then
         storeinDB(user, token, ‘‘Mention’’, ‘‘Analysis Table’’)
      end if
      if startsWith(token,‘‘#’’) then
         storeinDB(user, token, ‘‘Hashtag’’, ‘‘Analysis Table’’)
      else
         storeinDB(user, getStem(token), ‘‘PlainTextStem’’,
‘‘Analysis Table’’)
      end if
   end for
end for
```

Figure 6.2. Post Parser Algorithm

```
stemsNoun, stemsVerb, stemsAdjective, stemsAdverb : List
stemsNoun ⇐ jwi.findStems(token, ‘‘NOUN’’)
stemsVerb ⇐ jwi.findStems(token, ‘‘VERB’’)
stemsAdjective ⇐ jwi.findStems(token, ‘‘ADJECTIVE’’)
stemsAdverb ⇐ jwi.findStems(token, ‘‘ADVERB’’)
if ¬(stemsNoun.isEmpty()) then
   return(getFirstItem(stemsNoun))
end if
if ¬(stemsVerb.isEmpty()) then
   return(getFirstItem(stemsVerb))
end if
if ¬(stemsAdjective.isEmpty()) then
   return(getFirstItem(stemsAdjective))
end if
if ¬(stemsAdverb.isEmpty()) then
   return(getFirstItem(stemsAdverb))
else
   return(token)
end if
```

Figure 6.3. Get Stem Algorithm

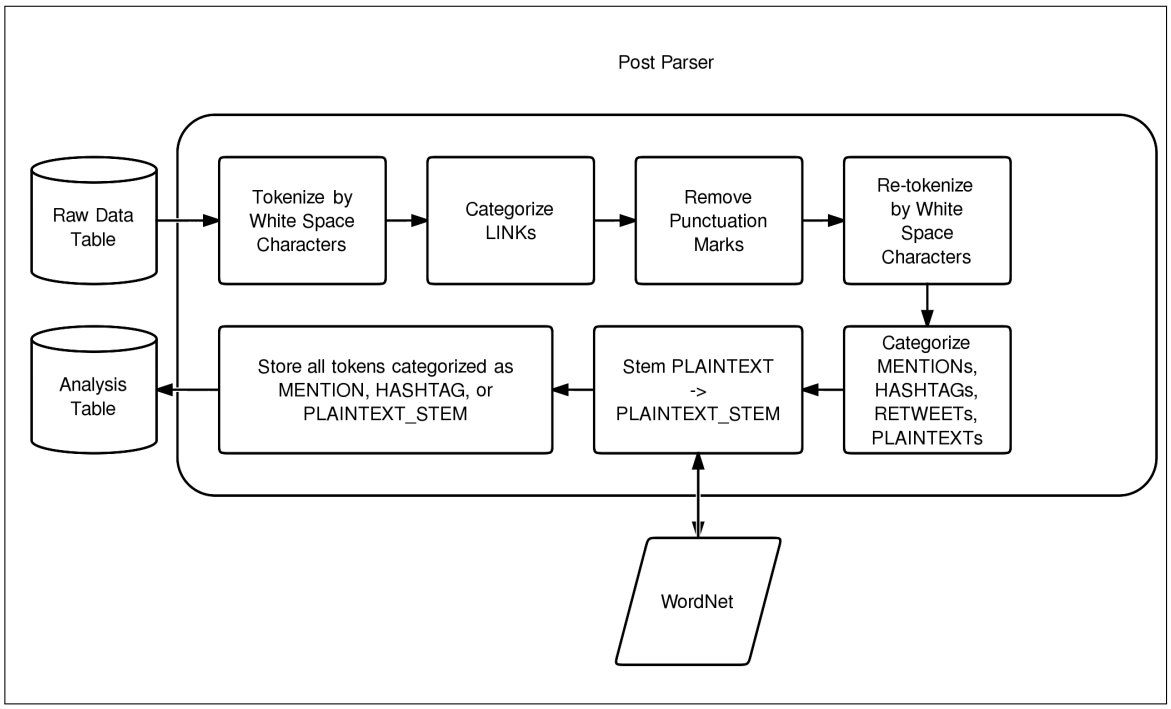

Figure 6.4. System architecture for Post Parser.

6.3. External Reference Analyzer

Links are frequently used in tweets. Since posts in Twitter are limited to 140 characters, most of these external references are shortened. Url shortening services such as TinyURL and bit.ly are heavily used in tweets.

For each token labeled as LINK in the Post Parser module, an HTTP connection is opened to the external resource. Although it is possible to gather all the page content from this connection, due to processing time limitations, only the original URL (LONG_URL), and the HTML title of the page (TITLE) are collected. LONG_URLs are stored in links database table. The same set of punctuation marks in Post Parser module are removed from the TITLEs. Result is tokenized by white space characters. The stems of resulting tokens are found using WordNet 2.1 and MIT Java Wordnet Interface (TITLE_STEMs). All tokens labeled asTITLE_STEM, are stored in link titles table for future use. (see Algorithm 6.5 and see Figure 6.6).

```
LINKArray : String[]
LINKArray = []
LINKArray ⇐ getFromDB(user, ‘‘Link’’, ‘‘Analysis Table’’)
for each LINK in LINKArray do
   TITLE, TITLE_STEM, LONG_URL : String
   titleArray : String[]
   titleArray = []
   conn : HTTPConnection
   conn ⇐ openConnection(LINK)
   TITLE ⇐ conn.getTitle()
   LONG_URL ⇐ conn.getLongUrl()
   storeinDB(user, LONG_URL, ‘‘Links Table’’)
   TITLE ⇐ removePunctuation(TITLE)
   titleArray ⇐ parse(TITLE)
   for each token in titleArray do
      TITLE_STEM ⇐ getStem(token)
      storeinDB(user, TITLE_STEM, ‘‘TitleStem’’,
‘‘Link Titles Table’’)
   end for
end for
```

Figure 6.5. External Reference Analyzer Algorithm

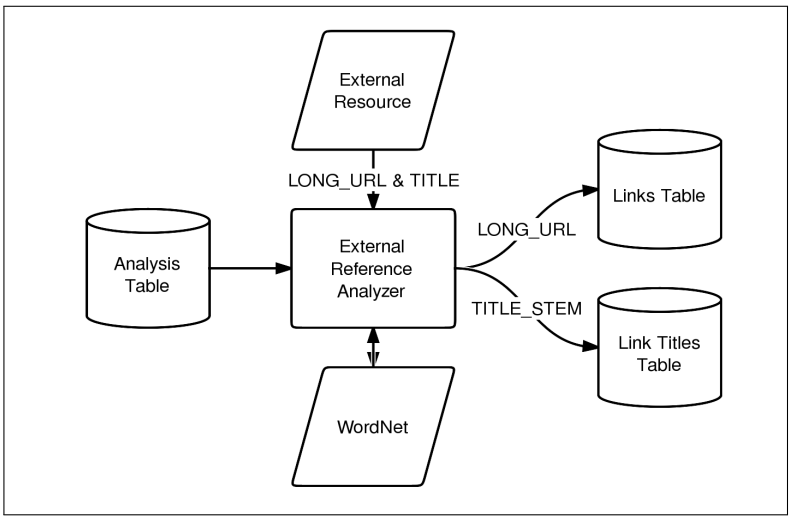

Figure 6.6. System architecture for External Reference Analyzer.

6.4. Internal Reference Analyzer

Some twitterers use hashtags as plain text also. An example is:

```
One good thing about #music, when it hits you, you feel no pain ~Bob Marley
```

In this tweet, #music is a hashtag, and also a part of the sentence. To detect such usage, each token labeled as HASHTAG in the Post Parser module is checked in the Wordnet dictionary. WordNet 2.1 and MIT Java Wordnet Interface are used. In the case of existence in the dictionary, besides being labeled and stored as a HASHTAG, the stem of that token is also labeled as PLAINTEXT_STEM, and is stored in analysis database table. (see Algorithm 6.7)(see Figure 6.8)

6.5. Token Gatherer

In previous modules, Post Parser, External Reference Analyzer, and Internal Reference Analyzer, tokens were labeled as either LINK, MENTION, HASHTAG, PLAINTEXT_STEM orTITLE_STEM. These tokens and labels were stored in two different data tables: Analysis Table and Link Titles Table. This module gathers all tokens labeled as either PLAINTEXT_STEM, HASHTAG, orTITLE_STEM. The

```
HASHTAGArray : String[]
HASHTAGArray = []
HASHTAGArray ⇐ getFromDB(user, ‘‘Hashtag’’, ‘‘Analysis Table’’)
for each HASHTAG in HASHTAGArray do
   if inDictionary(HASHTAG) then
      storeinDB(user, getStem(HASHTAG), ‘‘PlainTextStem’’,
‘‘Analysis Table’’)
   end if
end for
```

Figure 6.7. Internal Reference Analyzer Algorithm

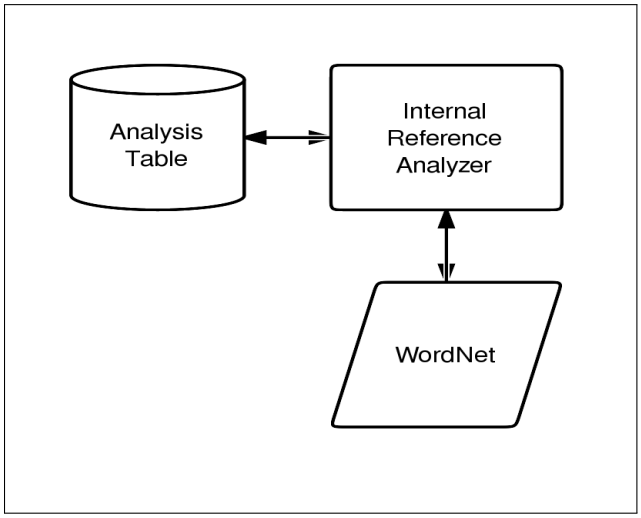

Figure 6.8. System architecture for Internal Reference Analyzer.

```
PLAINTEXT_STEMArray, HASHTAGArray, TITLE_STEMArray : String[]
PLAINTEXT_STEMArray, HASHTAGArray, TITLE_STEMArray = []
GATHERED_WORDSArray : String[]
GATHERED_WORDSArray = []
PLAINTEXT_STEMArray ⇐ getFromDB(user, ‘‘PlainTextStem’’,
‘‘Analysis Table’’)
HASHTAGArray ⇐ getFromDB(user, ‘‘Hashtag’’, ‘‘Analysis Table’’)
TITLE_STEMArray ⇐ getFromDB(user, ‘‘TitleStem’’,
‘‘Link Titles Table’’)
GATHERED_WORDSArray.add(PLAINTEXT_STEMArray)
GATHERED_WORDSArray.add(HASHTAGArray)
GATHERED_WORDSArray.add(TITLE_STEMArray)
for each token in GATHERED_WORDSArray do
   if ¬(inDB(token, ‘‘Gathered Words Table’’)) then
      weight : Integer
      weight ⇐ getTotalWeight(token, GATHERED_WORDSArray)
      storeinDB(user, token, weight, ‘‘Gathered Words Table’’)
   end if
end for
```

Figure 6.9. Token Gatherer Algorithm

number of occurrences of each token is summed and stored as the token's weight in gathered words table. MENTIONs are not used further in the model.(see Algorithm 6.9 and Figure 6.10).

6.6. Filterer

A list of stop words for English (570 words) is constructed using SMART system's list[42]. The list is found at ftp://ftp.cs.cornell.edu/pub/smart/english.stop and we also include it in the appendix A. Tokens in this list are discarded from the set gathered in the Token Gatherer module with their related weights.

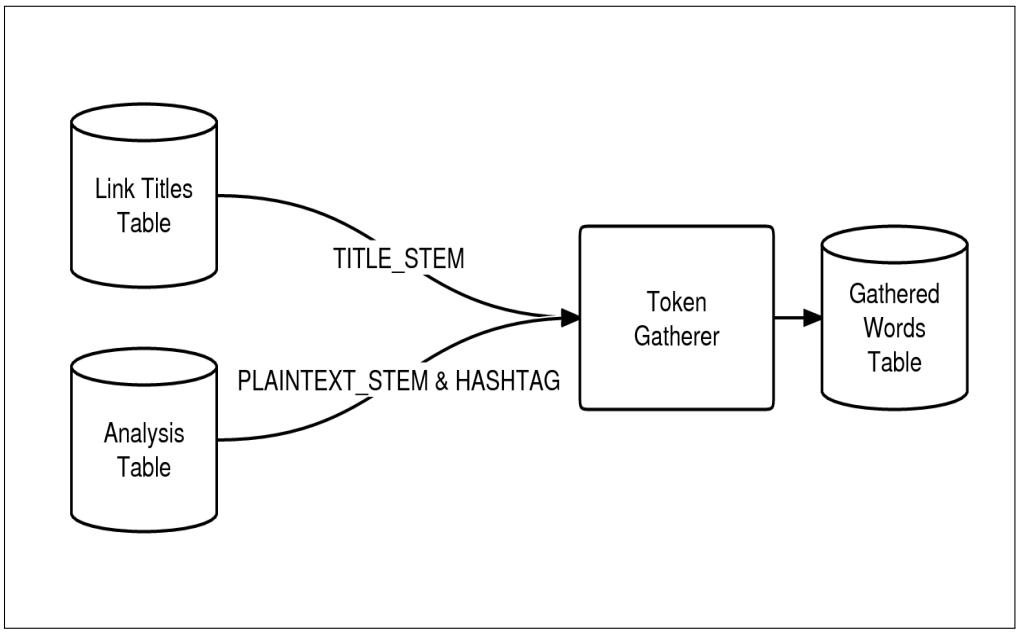

Figure 6.10. System architecture for Token Gatherer.

Also, part of speech of each token is found using WordNet 2.1 and MIT Java Wordnet Interface. A token's part of speech can be one of the following:

- Adjective,
- Adverb,
- Verb,
- Noun, or
- Unidentified if none above.

Tokens which are found to be adjectives, adverbs, or verbs are also discarded from the set gathered in the Token Gatherer module with their related weights. Result is stored in filtered words table. (see Algorithm 6.12) (see Figure 6.11)

6.7. Categorizer

For each LONG_URL gathered by the External Reference Analyzer module, the domain of the link is extracted and stored.

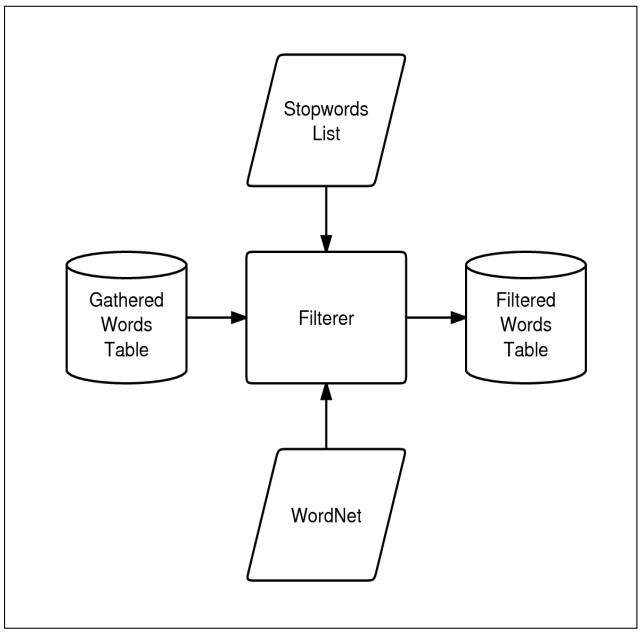

Figure 6.11. System architecture for Filterer.

The following values are computed using the meta information gathered and information processed for a user:

- $\tau_{fol,fr}(u) = \frac{f|ollowers(u)|}{|friends(u)|}$
- $\rho_{fol,fri}$(u)= $\frac{|followers(u)|}{|friends(u)|}$
- $\mu_{tweets} = \frac{status(u)}{latestPost(u) - creation(u)}$
- $domainsPercent(u)$ = Percentage of all domains posted by the user
- $mentionCount(u)$ = The number of distinct mentions that occurs in all of u's posts.

Based on these calculations, users are characterized as *automated*, *spam*, *bot*, *celebrity*, or *social* (see Figure 6.13).

Two thresholds automatically categorize users as *automated*:

- Update Frequency Threshold: Based on known human twitterers with high tweet frequencies (aplusk: 14.7, guykawasaki: 39.6, mrskutcher: 12.9), an upper threshold of 80 tweets/24 hours was chosen to indicate the maximum number of tweets

```
GATHERED_WORDSArray, FILTERED_WORDSArray : String[]
GATHERED_WORDSArray, FILTERED_WORDSArray = []
GATHERED_WORDSArray ⇐ getFromDB(user, ‘‘Gathered Words Table’’)
for each token in GATHERED_WORDSArray do
   if inList(token, ‘‘Stop Words List’’) then
      GATHERED_WORDSArray.delete(token)
   else
      if hasAdverbSense(token) then
         GATHERED_WORDSArray.delete(token)
      end if
      if hasAdjectiveSense(token) then
         GATHERED_WORDSArray.delete(token)
      end if
      if hasVerbSense(token) then
         GATHERED_WORDSArray.delete(token)
      end if
   end if
end for
for each token in GATHERED_WORDSArray do
   storeinDB(user, token, weight, ‘‘Filtered Words Table’’)
end for
```

Figure 6.12. Filterer Algoritm

that a human user is likely to post. Users whose μ_{tweets} are greater than 80 are categorized as **automated**.

- Domain Frequency Threshold: When the same domain occurs in a significant portion of a user's tweets, that user is also categorized as **automated**. Currently this threshold is 50%. Users who have a domain with $domainsPercent(u)$ greater than 50% are categorized as **automated**.

Users who are categorized as **automated**, are attempted to be further categorized as **spam** or **bot** by examining $RATIO_{follower,friend}$.

Spammers are users who post very frequently, generally referencing a single external domain. They have few followers (since most users find them irritating), but follow many others.

Bots also post very frequently and generally reference a single external domain. However, the number of their followers is usually far greater than the number of users they follow. This is because **bot**s usually provide beneficial information such as breaking news, traffic conditions, or weather.

Users with

$$RATIO_{follower,friend} > \theta_{bot}$$

are categorized as **bot**s. Currently $\theta_{bot} = 100$.

Users with

$$(RATIO_{follower,friend}) < 1/\theta_{bot}$$

are categorized as **spammer**s.

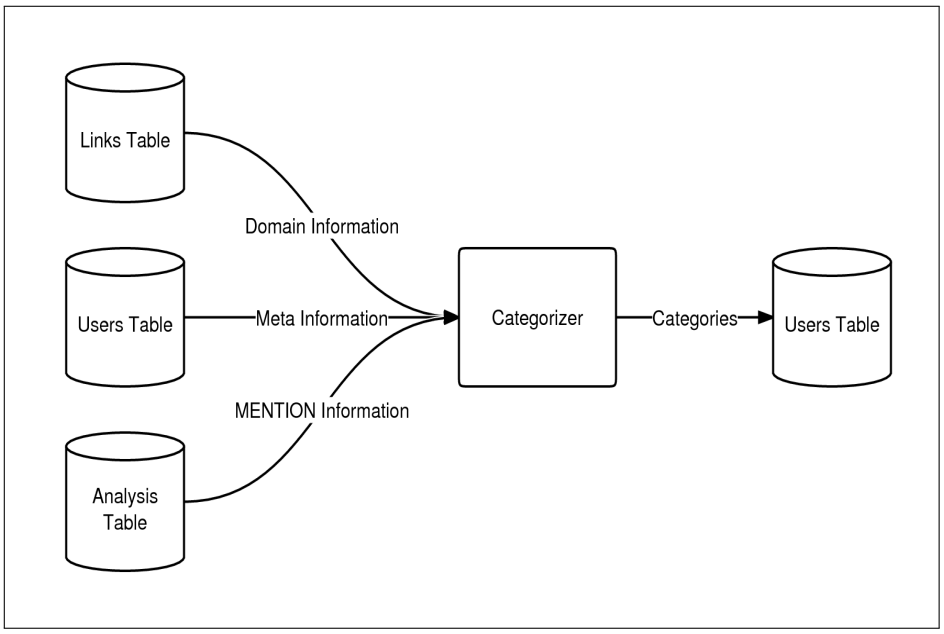

Figure 6.13. System architecture for Categorizer.

Celebritys are human users whose number of followers is usually far greater than the number of users they follow. Users who are not categorized as **automated** but with

$$RATIO_{follower,friend} > \theta_{bot}$$

are categorized as **celebrity**s.

Social users are in communication with many other users. Users with

$$COUNT_{mention} > \theta_{social}$$

are categorized as **social**s. Currently $\theta_{social} = 500$.

6.8. Word Cloud Visualizer

This module generates a visual (weighted) word cloud. A text file is created using PLAINTEXT_STEM, HASHTAG, andTITLE_STEM tokens that has not

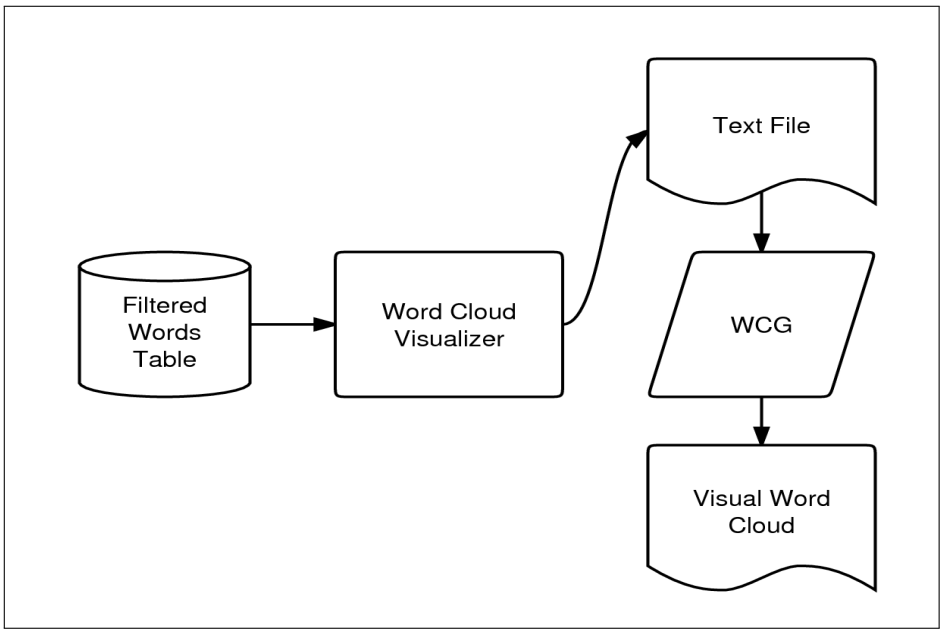

Figure 6.14. System architecture for Word Cloud Visualizer.

been filtered by the Filterer and their number of occurences as weights. This file serves as input to IBM®(International Business Machines Corporation is abbreviated as IBM) Word-Cloud Generator (WCG). IBM WCG generates a tag cloud as a PNG image (Portable Network Graphics (PNG) is a bitmapped image, which uses lossless data compression). Font sizes are proportional to the frequencies of the tokens (see Figure 6.14).

6.9. Comparator

This module compares the user's generated word cloud with other user's in the system. Top n tokens from each users' words clouds are taken and compared using Cosine Similarity measures [43]. For each pair of users, comparator executes two different algorithms: Unweighted Cosine Similarity and Weighted Cosine Similarity. Both algorithms give a number between 0 and 1 as output. This number denotes the similarity between two users. A similarity of 1 indicates equality (see Figure 6.17).

$Set\ sTokenList_i \Leftarrow$ getSet(top n tokens from $user_i$'s $UITS$)

$Set\ sTokenList_j \Leftarrow$ getSet(top n tokens from $user_j$'s $UITS$)

$int\ commonWords \Leftarrow$ getCount(getCommonWords($sTokenList_i$, $sTokenList_j$))

$int\ magnitude_i \Leftarrow$ squareRoot(getCount($sTokenList_i$))

$int\ magnitude_j \Leftarrow$ squareRoot(getCount($sTokenList_j$))

store($user_i$,$user_j$,$\frac{commonWords}{magnitude_i \times magnitude_j}$)

Figure 6.15. Unweighted Cosine Similarity Algorithm.

6.9.1. Unweighted Cosine Similarity

For two users $user_i$ and $user_j$, two sets are constructed using top n tokens from each user's word clouds. The number of tokens that exist in each set and both of the sets are calculated.

$unweightedCosineSimilarity(u_i, u_j) = \frac{tokens(u_i) \cap tokens(u_j)}{\sqrt{tokens(u_i)} \times \sqrt{tokens(u_j)}}$ is stored as the Unweighted Cosine Similarity value between u_i and u_j (see Algorithm 6.15).

6.9.2. Weighted Cosine Similarity

For two users $user_i$ and $user_j$, top n tokens from each user's word clouds are collected with their related weights. A vector of common words is constructed from the union of these token sets. Two weight vectors for $user_i$ and $user_j$ are initialized. For each token in the common words, related weights are added to users' weight vectors. Value of 0 is inserted when a token does not exist in top n tokens in a user's word cloud. Dot product and magnitudes of two vectors are calculated.

$dotProduct =$ dot product of weight vectors of u_i and u_j, where $u_i, u_j \in Users \wedge u_i ! = u_j$

$\frac{dotProduct}{|weightVector(u_i)| \times |weightVector(u_j)|}$ is stored as the Weighted Cosine Similarity of u_i and u_j (see Algorithm 6.16).

$HashMap\ hTokenList_i \Leftarrow$ getHashMap(top n tokens with their associated weights from $user_i$'s $UITS$)

$HashMap\ hTokenList_j \Leftarrow$ getHashMap(top n tokens with their associated weights from $user_j$'s $UITS$)

$Vector\ vCommonWords \Leftarrow$ getUnion(getKeys($hTokenList_i$), getKeys($hTokenList_j$))

initialize($Vector\ vTokenWeightList_i$)

initialize($Vector\ vTokenWeightList_j$)

for each $token_k$ in $vCommonWords$ **do**

  if exists($token_k$,getKeys($hTokenList_i$)) **then**

    add(getValue($hTokenList_i$,$token_k$),$vTokenWeightList_i$)

  else

    add(0,$vTokenWeightList_i$)

  end if

  if exists($token_k$,getKeys($hTokenList_j$)) **then**

    add(getValue($hTokenList_j$,$token_k$),$vTokenWeightList_j$)

  else

    add(0,$vTokenWeightList_j$)

  end if

end for

$int\ magnitude_i \Leftarrow$ getMagnitude($vTokenWeightList_i$)

$int\ magnitude_j \Leftarrow$ getMagnitude($vTokenWeightList_j$)

store($user_i$,$user_j$,getDotProduct($vTokenWeightList_i$, $\frac{vTokenWeightList_j)}{magnitude_i \times magnitude_j}$))

Figure 6.16. Weighted Cosine Similarity Algorithm

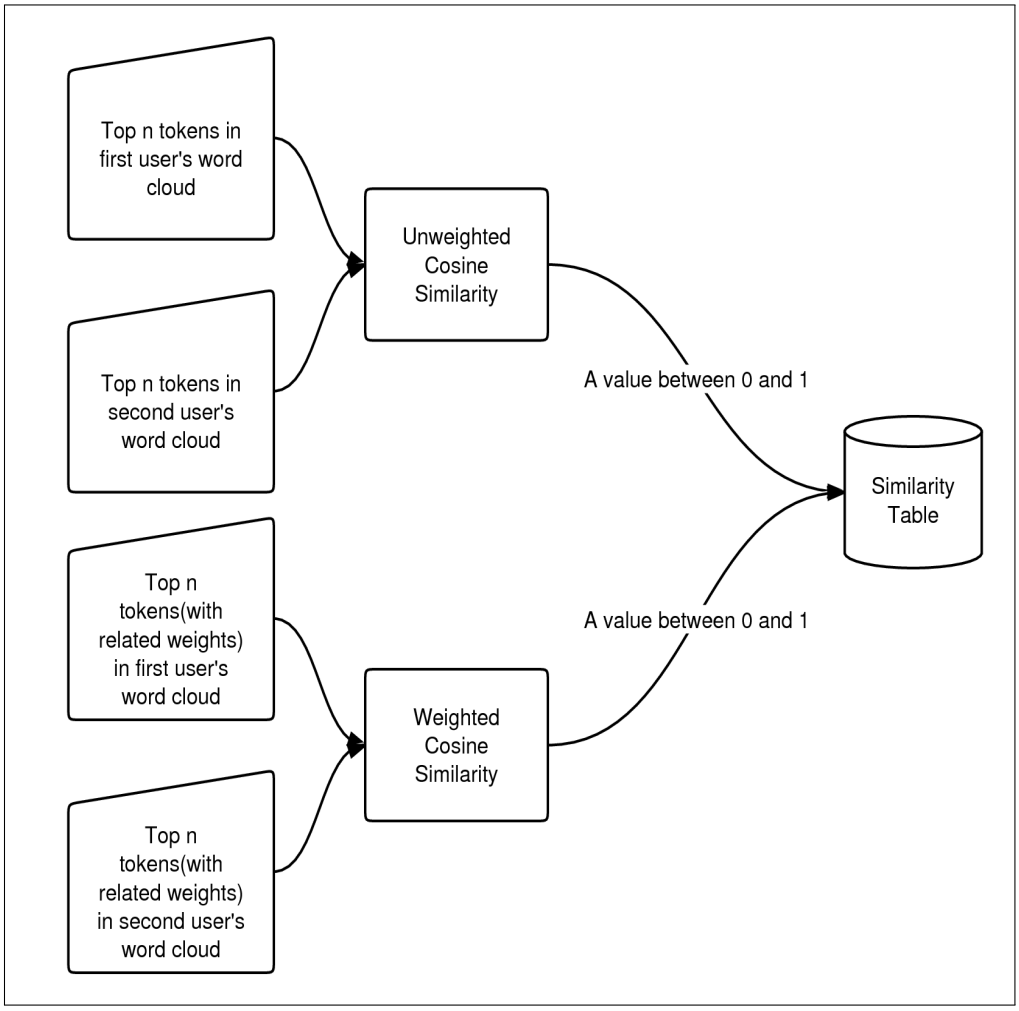

Figure 6.17. System architecture for Comparator.

7. RESULTS

Various Twitter user lists group users according to some criteria. WeFollow [44] allows users to declare five self selected interests. These interests – such as music, socialmedia, tech, and tv – are interpreted as tags. The frequencies of the tags most chosen are in the tens of millions, whereas the number of users following those with such interests are in millions (These figures taken as of 08.05.2010).

We chose a few such tags from WeFollow to examine the contributions of similarly interested Twitter users. Tables 7.1, 7.2, 7.3, 7.4, 7.5 show five sets of users selected using the tags *socialmedia*, *microblogging*, *music*, *indie*, and *birdwatching*. For each set we chose the top 30 users associated with that tag (as of 11.04.2010), whose contributions are public. Table 7.6 and Table 7.7 describes the properties examined in these data sets.

In Section 7.1, some statistics about data gathered are shown to reveal general Twitter usage among users. In Section 7.2, results of user categorization for 150 users are given and discussed. Section 7.3 lists commonly used words among 5 sets of users.

Table 7.1. Selected WeFollow users who declared interest in *socialmedia*.

socialmedia
ijustine problogger rww Jasoncalacanis Veronica WholeFoods steverubel feliciaday MCHammer ChrisPirillo sacca Ustream QueenRania someecards youtube ScottMonty threadless eMarketer armano Mediabistro prsarahevans briansolis SocialMedia411 mashable TechCrunch aplusk kevinrose GuyKawasaki zappos rainnwilson

Table 7.2. Selected WeFollow users who declared interest in *microblogging*.

microblogging
m140z microversos DirkRoehrborn boehr svb elcario rev2tweet101 filos kleverson tuitwit thejournaldotme guillembaches kevinblakeley Scabr radarnet Tommaso microblogging zoomer49 maccimum MicroPoesia kuckuvn tweetconvo tweetcrunch hdzimmermann starpath SportSpotter justincron TheRyanColby jeos AyaMai

Table 7.3. Selected WeFollow users who declared interest in *music*.

music
johncmayer coldplay petewentz snoopdogg pitchforkmedia ashleytisdale MariahCarey ashsimpsonwentz questlove souljaboytellem 50cent markhoppus alyankovic johnlegend LennyKravitz MCHammer twtfm SaraBareilles TheRealJordin jimmyeatworld samantharonson PaulaAbdul DaveJMatthews AFineFrenzy QtipTheAbstract amazonmp3 ryanleslie EmilyOsment mitchelmusso iamdiddy

Table 7.4. Selected WeFollow users who declared interest in *indie*.

indie
pitchforkmedia indieBandFollow indiefeed indiemusicfinds Under_Radar_Mag MeLikeGoodMusic atpfestival IndiescreetBlog victoryrecords TeamClermont indiespotting welistenforyou Think_Indie eardrums indiemusicfiltr inertiamusic DeadOceans shelflife carybrothers TheMusicMan81 MadeLoud copelandband oswaldband eeniemeenierecs FameGames indiefeeds portobrien indierockgirl theflyingchange deathrockstar

Table 7.5. Selected WeFollow users who declared interest in *birdwatching*.

birdwatching
burdr birdingbev smido BirdWatchingMag MaineBirder RGVBirdingFest _BTO birdpost LadyWoodpecker gonolek Sly102 birdfeeders birdinggirl picusblog jpperret gwendolen adaptive ontdeksafaris OP_Birding BirdingBliss chuq AustinBirder WatchBirds kennysalazar FatFinch Birdsafariswede AWFN simbirds penelopedia agru

Table 7.6. Descriptions of properties examined in contributions.

C:	Total number of contributions of 30 users.
C_{daily}:	Average daily contributions of 30 users.
C_c:	Collected contributions of 30 users. (Twitter API limit is 3200 for each user)
T:	Single word tokens in collected contributions of 30 users (Links and retweet tokens are excluded).
H:	Hashtags in collected contributions of 30 users.
M:	Mentions in collected contributions of 30 users.
P:	Plain texts in collected contributions of 30 users.
P_{wos}:	Plain texts without stopwords in collected contributions of 30 users.
L:	Titles' texts of links in collected contributions of 30 users.
L_{wos}:	Titles' texts without stopwords of links in collected contributions of 30 users.
S:	Significant tokens ($H + P_{wos} + L_{wos}$).
S_n:	Significant noun tokens.
S_u:	Significant unidentified tokens.
S_{pos}:	Significant noun and unidentified tokens. These tokens are used in generating users' word clouds. ($S_n + S_u$)

Table 7.7. Descriptions of properties examined in contributions.

Rc_H:	Ratio of hashtag usage in contributions (H/T).
Rc_M:	Ratio of mention usage in contributions (M/T).
Rc_P:	Ratio of plain text usage in contributions (P/T).
Rs_H:	Ratio of hashtags in significant tokens ($H/H + P_{wos} + L_{wos}$).
Rs_P:	Ratio of plain text in significant tokens ($P_{wos}/H + P_{wos} + L_{wos}$).
Rs_L:	Ratio of link title text in significant tokens ($L_{wos}/H + P_{wos} + L_{wos}$).
$Rstop$:	Ratio of stopwords in plain text and link titles ($((P - P_{wos}) + (L - L_{wos}))/(P + L)$).
$Rpos_n$:	Ratio of nouns in significant tokens (S_n/S).
$Rpos_u$:	Ratio of unidentified tokens in significant tokens (S_u/S).
$Rpos_{n\_u}$:	Ratio of noun unidentified tokens in significant tokens ($(S_n + S_u)/S$).

Finally, in Section 7.4, word clouds generated for all users in the dataset are compared.

7.1. Twitter usage

Based on 150 users belonging to 5 different sets, numerical analysis revealing general Twitter usage are shown in Table 7.8 and Table 7.9.

Observing the total number and daily average of users' contributions, it can be observed that the users in socialmedia group post more frequently than others. Twitter itself is a social media tool. Thus, users who have declared interest in socialmedia being more active than others is not surprising.

Among all words uttered by users in tweets,

- hashtag usage is 2% in average, and is more common in more specialized groups (i.e indie and bird watching).
- mention usage is 4% in average, and is more common in more specialized groups (i.e bird watching).

Set of users who are a part of a community with more specific interests tend to be closely connected with each other. They refer to each other more, and they use a common vocabulary. So, above two results were expected. This information is not used further in the model.

Of all significant tokens gathered for 150 users,

- 2% are hashtags,
- 72% are tokens uttered by users in tweets, and
- 26% are tokens collected from titles of external links.

41% of significant tokens are stopwords, and are filtered. Of the remaining,

Table 7.8. Comparison of Twitter usage in numbers

	social media	micro blogging	music	indie	bird watching	TOTAL
C	**325,534**	**127,050**	**99,118**	**68,098**	**71,211**	**691,011**
C_{daily}	**11**	**8**	**6**	**5**	**5**	**7**
C_c	83,745	46,548	56,293	51,411	41,806	279,803
T	1,087,917	589,458	764,595	703,724	554,911	3,700,605
H	8,806	13,434	4,096	18,491	19,067	63,894
M	47,453	22,629	31,887	29,779	32,187	163,935
P	1,031,658	553,395	728,612	655,454	503,657	3,472,776
P_{wos}	515,054	382,058	348,042	374,977	281,055	1,901,186
L	312,938	206,201	102,199	194,595	140,780	956,713
L_{wos}	214,046	158,918	74,141	144,314	103,888	695,307
S	737,906	554,410	426,279	537,782	404,010	2,660,387
S_n	398,129	229,023	221,480	268,075	223,195	1,339,902
S_u	150,149	248,934	95,598	145,077	90,620	730,378
S_{pos}	548,278	477,957	317,078	413,152	313,815	2,070,280

- 50% are nouns, and
- part of speech of 28% cannot be determined by our part of speech function.

This 78% of significant tokens are used in users' word clouds to describe users. 22%(adjectives, adverbs, and verbs) are filtered.

7.2. User categorization

Users are categorized as *automated*, *spam*, *bot*, *celebrity*, or *social* depending on the below rules (see Section 6.6):

- Using two values, Update Frequency Threshold(80) and Domain Frequency Threshold(50%), users are categorized as *automated*.

Table 7.9. Comparison of Twitter usage in percentage

	social media	micro blogging	music	indie	bird watching	TOTAL
Rc_H	1%	2%	1%	3%	3%	**2%**
Rc_M	4%	4%	4%	4%	6%	**4%**
Rc_P	95%	94%	95%	93%	91%	94%
Rs_H	1%	2%	1%	3%	5%	**2%**
Rs_P	70%	69%	82%	70%	69%	**72%**
Rs_L	29%	29%	17%	27%	26%	**26%**
$Rstop$	46%	29%	49%	39%	40%	**41%**
$Rpos_n$	54%	41%	52%	50%	55%	**50%**
$Rpos_u$	20%	45%	22%	27%	23%	**28%**
$Rpos_{n\_u}$	74%	86%	74%	77%	78%	**78%**

- Users who are categorized as *automated* are further categorized as *bot* or *spam* if their Follower Friend Ratio is above 100 or below 1/100.
- Users who are not categorized as *automated*, but have a Follower Friend Ratio above 100, are categorized as *celebrity*.
- Independent of being *automated*, *spam*, *bot*, or *celebrity*, users who use more than Distinct Mention Threshold(500) different mentions in their tweets are categorized as *social*.

Categorization results can be seen in Table 7.10.

- 18 users are automated. Among these, 9 are further categorized as bots. No automated users are further categorized as spam. This was expected since users examined were the top 30 users of 5 categories in WeFollow. The rest of the automated users is not categorized as bot or spam.
- 132 users are not categorized as automated. Among these, 38 users are celebrities. In music group, 25 of 30 users are categorized as celebrity. 10 other celebrities are from socialmedia group. These results were expected since these two groups

Table 7.10. Comparison of Twitter users' categorization by groups

	socialmedia	microblogging	music	indie	birdwatching
automated	7	3	3	3	2
bot	5(1)	0	3	1	0
spam	0	0	0	0	0
not bot nor spam	2	3	0	2	2
non-automated	23	27	27	27	28
celebrity	10(6)	1	25(12)	2	0
non-celebrity	13(10)	26(4)	2(1)	25(5)	28(3)
social	17	4	13	5	3

are popular groups among twitterers, and celebrities add these groups to their interests. The rest of the non-automated users is not categorized as celebrity.

- 42 users are categorized as social. The distribution of these users to other categories is shown in parantheses.

7.3. Commonly used words in Twitter

For each set of users, tables 7.12, 7.13, 7.14, 7.15, 7.16 show words that at least $UserPercentage\%$ of users contributed above threshold $CommonWordThreshold$.

This analysis is made to understand words that are frequently used by many users. To eliminate words that are rarely used by many people, $CommonWordThreshold$ is used.

$CommonWordThreshold$ is calculated with respect to S_{pos} for each set. Average number of words used in word clouds is calculated by:

$$S_{pos_{average}} = \frac{S_{pos}}{30}$$

$$CommonWordThreshold = \frac{S_{pos_{average}}}{1000}$$

To obtain words that are commonly used by absolute majority of the group,

$$UserPercentage = 50$$

Words that exist in 3 or more tables (7.12, 7.13, 7.14, 7.15, 7.16) are considered as commonly used words among microbloggers in general, and these words are shown in Table 7.11. In Table 7.11, numbers in parantheses declare the number of groups the word is commonly used in.

In Tables 7.12, 7.13, 7.14, 7.15, 7.16, the first numbers in parantheses represent the number of users contributed above the threshold $CommonWordThreshold$. The numbers after colon represent how many times that word is used by these users in total. Words that do not exist in Table 7.11 are considered as *Group Specific Words* and are shown in bold.

One can observe that the common words of the set of users who declared more specific interests, such as *birdwatching* and *indie*, are more descriptive. General interests, such as *music* and *socialmedia* are less descriptive. By examining common words, one can observe community-specific vocabulary – such as band, mp3, track for music,

Table 7.11. Common words of microbloggers in general.

Common Words
news(5) post(5) time(5) twitter(5) video(5) 2009(4) blog(4) check(4) day(4) love(4) make(4) photo(4) tweet(4) watch(4) week(4) work(4) year(4) fan(3) friend(3) life(3) man(3) morning(3) music(3) night(3) people(3) show(3) thing(3) twitpic(3)

and bird, wildlife for birdwatching. All common words are tokens either uttered by users in tweets, or in titles of external links within their tweets. Since hashtags are rarely used by twitterers, there are no hashtags among common words. In future work, it would make sense to have lower thresholds for hashtags. Analysis on larger groups and statistical analysis should be applied to improve results.

In the initial phases of this study, contributions of many users were examined. As a result, it was observed that some words are generally used by twitterers. Table 7.11 shows the common words for the examined groups. Consistent with earlier observations, in these cases the following common word use was observed:

- Twitter related: Twitter itself, Twitter applications, and other Twitter related words (e.g. tweet, twitpic, twitter, etc.).
- Time related: Names of days, months, years, dates, times. (e.g. morning, night, week, year, 2009, 2010, etc.). It is common to observe digits that are related to the day of post. Time is an important aspect of microblogs, where freshness can be measured by seconds. So this is not surprising.
- Emotional: Words like love, hope, miss, etc.
- Instructional: Words like check, watch, make, etc.

7.4. User Tagging and User Comparison

This study is based on the opinion that contributions of users can be used to describe users, and users who share similar interests contribute similarly. To under-

Table 7.12. Commonly used words among Twitter users who declared interest in *socialmedia*.

social media	
$CommonWordThreshold = 18$	$UserPercentage = 50$
twitter(29:8421) **facebook(25:4323)** video(28:4257) blog(26:3777) **google(19:3585)** time(30:3536) day(30:3495) make(28:3117) **app(22:2656)** tweet(29:2552) **iphone(21:2346)** check(26:2186) love(28:2142) news(25:2138) **web(19:2099) $(22:2073) online(16:1990)** people(25:1695) **job(19:1626)** show(26:1613) post(23:1597) year(28:1566) **business(17:1506)** work(27:1437) photo(22:1399) **user(15:1354)** thing(25:1351) twitpic(15:1315) **2010(17:1239)** week(25:1235) watch(24:1230) **search(18:1225)** 2009(15:1128) **site(23:1093)** **book(16:1038)** **list(20:1023)** **internet(16:993)** **start(26:979)** **share(19:963)** **email(20:932)** man(22:930) **interview(23:922)** life(23:900) **miss(23:887)** night(19:820) friend(23:819) **company(15:815)** **hour(20:803)** **story(18:802)** **call(22:795)** **guy(17:786)** **tv(18:783)** **find(23:763)** **buy(21:762)** music(15:741) **update(16:721)** **service(15:709)** **add(16:703)** **question(20:701)** **page(18:698)** fan(17:680) **feature(16:679)** **lot(16:657)** **give(21:642)** **party(17:628)** **link(16:617)** **hope(17:615)** **talk(18:592)** **idea(19:589)** **win(17:582)** **change(18:574)** **read(15:573)** **thought(19:568)** **team(15:563)** **interest(15:536)** **bit(18:530)** morning(16:527) **kid(17:514)** **feel(15:426)**	
Number of words commonly used among microbloggers: 28	
Number of words specific to group: 51	

Table 7.13. Commonly used words among Twitter users who declared interest in *microblogging*.

micro blogging	
$CommonWordThreshold = 16$	$UserPercentage = 50$
de(15:9954) twitter(28:9920) blog(23:3591) **google(21:3542)** **web(21:2701)** news(18:2529) **facebook(19:1805)** tweet(21:1659) **online(17:1592)** video(18:1383) time(19:1328) **iphone(17:999)** **app(17:919)** post(15:872) **search(15:799)** 2009(18:780)	
Number of words commonly used among microbloggers: 8	
Number of words specific to group: 8	

Table 7.14. Commonly used words among Twitter users who declared interest in *music*.

music	
$CommonWordThreshold = 11$	$UserPercentage = 50$
love(27:2916) **album(23:2606)** day(29:2568) twitpic(23:2308) show(26:2275) **song(28:2215)** time(27:2136) video(26:2135) twitter(26:2064) music(27:2047) make(28:1781) photo(23:1479) **guy(18:1373)** night(26:1337) check(25:1297) **share(19:1172)** people(22:1042) watch(22:972) **tour(17:971)** tweet(18:922) life(19:883) **gonna(15:876)** friend(21:869) news(16:860) man(20:844) **hey(18:830)** morning(18:826) week(21:781) **hope(19:779)** year(19:739) **boy(15:734)** work(20:727) thing(20:718) **play(20:713)** **rock(17:698)** **call(20:661)** **la(19:657)** fan(17:642) **miss(21:636)** **studio(18:632)** **feel(16:606)** **wait(19:605)** **head(18:587)** **band(15:546)** **record(15:546)** **party(18:538)** **movie(20:534)** **give(17:520)** **girl(16:517)** **birthday(15:499)** **hit(15:490)** **itune(16:450)** **start(16:440)** **city(16:401)** **stop(17:387)** **kid(17:384)** post(15:381) **win(16:370)** **tune(15:348)**	
Number of words commonly used among microbloggers: 26	
Number of words specific to group: 33	

Table 7.15. Commonly used words among Twitter users who declared interest in *indie*.

indie	
$CommonWordThreshold = 14$	$UserPercentage = 50$
music(28:11501) **album(26:4514)** **song(25:3993)** day(28:3386) video(26:3302) **2010(15:3157)** 2009(20:2767) **rock(19:2704)** **band(27:2224)** love(26:2146) **mp3(20:1899)** blog(19:1699) check(25:1692) show(26:1648) **record(20:1524)** **release(20:1522)** **tour(19:1428)** **track(20:1402)** time(26:1352) twitter(22:1328) **artist(17:1295)** **review(16:1284)** post(20:1247) **ep(15:1214)** make(24:1125) week(22:910) year(22:882) news(18:857) **play(18:785)** **remix(15:750)** friend(15:704) night(21:684) photo(15:648) people(19:640) fan(17:625) thing(16:620) man(16:613) watch(15:600) **interview(17:578)** work(17:458) **feature(16:445)** **friday(15:377)**	
Number of words commonly used among microbloggers: 24	
Number of words specific to group: 18	

Table 7.16. Commonly used words among Twitter users who declared interest in *birdwatching*.

bird watching	
$CommonWordThreshold = 10$	$UserPercentage = 50$
bird(29:15324) photo(23:4919) twitpic(15:3274) news(19:3062) blog(25:2889) twitter(21:2551) day(25:1850) **nature(16:1664)** post(20:1647) time(20:1274) **wildlife(16:1228)** love(18:957) year(20:812) make(20:793) morning(18:787) watch(16:757) 2009(16:710) video(15:639) week(16:596) work(17:550) check(17:549) life(16:516) tweet(15:431)	
Number of words commonly used among microbloggers: 20	
Number of words specific to group: 3	

stand whether this approach can be pursued to locate and describe users based on their contributions, descriptions of users who declared similar interests in other systems are compared.

Twitter users can declare five self selected interests in WeFollow. We chose five groups (socialmedia, microblogging, music, indie, birdwatching) from WeFollow [44] to examine the contributions of similarly interested Twitter users. For each group we chose the most popular 30 users. Among popular users, there are:

- technology news agents, celebrities, bloggers, entrepreneurs, technology geeks in the socialmedia group.
- many different types of users of various interests in the microblogging group. They are all interested in microblogging in common.
- artists, bands, dancers, music guides, radio stations, music critics, web pages focused on music, electronic commerce compaines in the music group.
- music guides, music magazines, festivals, artists, bands, radio stations, music critics interested in independent music in the indie group.
- photographers, individuals, organisations, naturalists, magazines focused on birds and bird watching in the birdwatching group.

Music is a popular interest of microbloggers, people in general. Microblogging is a social media tool, hence *Social media* is another popularly declared interest among microbloggers. *Microblogging* is a more specific area of *social media*, and *indie* is a genre of *music*. *Bird watching* is a very narrow area, since it is a very specific interest. Before the comparisons were computed, the expectations were:

- users of the same group to be more similar to each other than to users of another group
- users of group *indie* to be more similar to the users of group *music* than any other group and users of group *music* to be more similar to the users of group *indie* than any other group
- users of group *microblogging* to be more similar to the users of group *social media*

than any other group and users of group *social media* to be more similar to the users of group *microblogging* than any other group

- users of group *bird watching* to be very similar to each other, but not so similar to other users in other groups.

During the experiments, top 10, 100 and 1000 tokens of each of 150 users are compared with each other using cosine similarity.

Figures 7.1, 7.2, and 7.3 show the results of unweighted cosine similarity (see Section 5.4), meaning that the frequencies of tokens are not taken into consideration. Figures 7.4, 7.5 and 7.6 show the results of weighted cosine similarity. In weighted cosine similarity, vectors for each user are constructed using the frequencies of tokens in their word clouds (see Section 5.4).

In Figures 7.1, 7.2, 7.3, 7.4, 7.5, and 7.6, users are figured as follows:

- 1 - 30 are users of group socialmedia
- 31 - 60 are users of group microblogging
- 61 - 90 are users of group music
- 91 - 120 are users of group indie
- 121 - 150 are users of group birdwatching

Inside a group, users are sorted according to their influence in WeFollow. The most influential user in WeFollow is the closest user to origin on the figure.

Compared to Figures 7.2, 7.3, 7.5, and 7.6, Figures 7.1 and 7.4 are not very clear and descriptive. This suggests that using the top 10 tokens is not enough for describing users. Observing how Figure 7.2 is similar to Figure 7.3 and Figure 7.5 is similar to Figure 7.6 suggests that using the top 1000 tokens instead of 100 does not change the results much. More investigation is required to determine where the breaking point is.

In this study, weighted token sets are used for describing users. When comparing

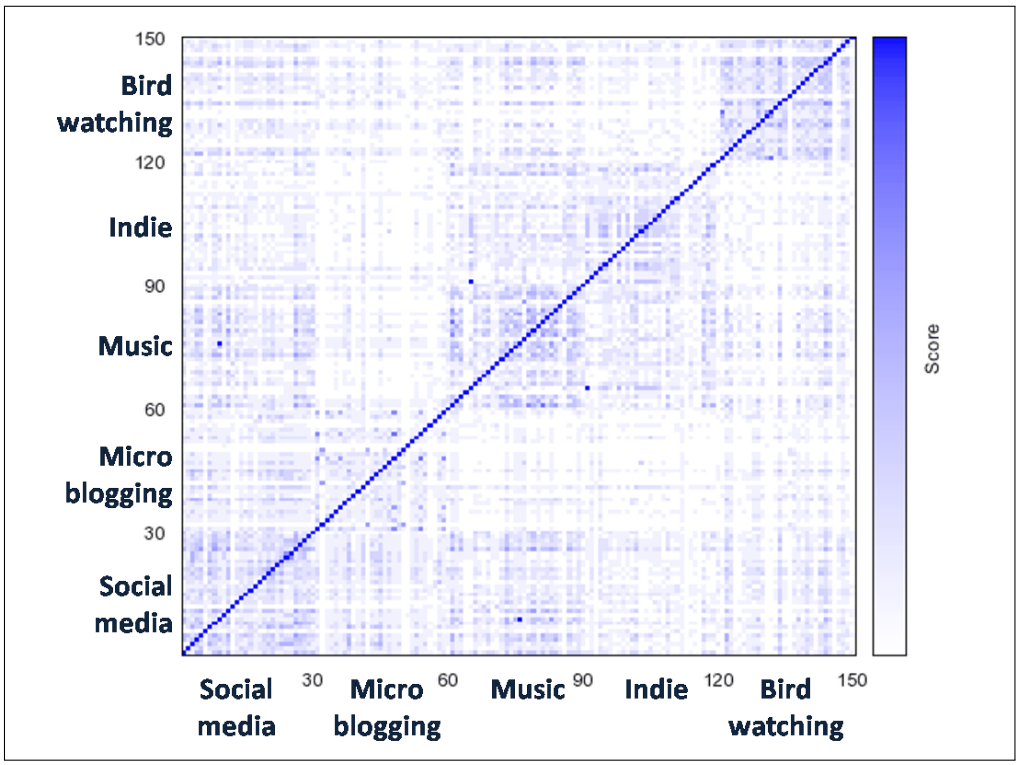

Figure 7.1. Comparison of microblogger groups using unweighted cosine similarity over the top 10 tokens.

two sets, ignoring weights may produce such a result: Two users, who use same tokens, but with very different frequencies may seem very similar to each other. As a result, really similar users, who use similar tokens with similar frequencies, may be overlooked. So, weights of tokens are important metrics, and are very useful in comparing users. One can observe that weighted cosine similarity figures are quite different from unweighted cosine similarity figures and weighted cosine similarity figures(Figures 7.5 and 7.6) are more effective in comparing similarities.

Results observed from Figures 7.5 and 7.6 are as follows:

- As expected, users in the same group are more similar to each other than any other groups in general.
- Social media is a popular interest, and popular people (e.g. celebrities) choose socialmedia as one of their interests. This results in a group of users, who utter about a variety of topics instead of areas specific to social media. These topics

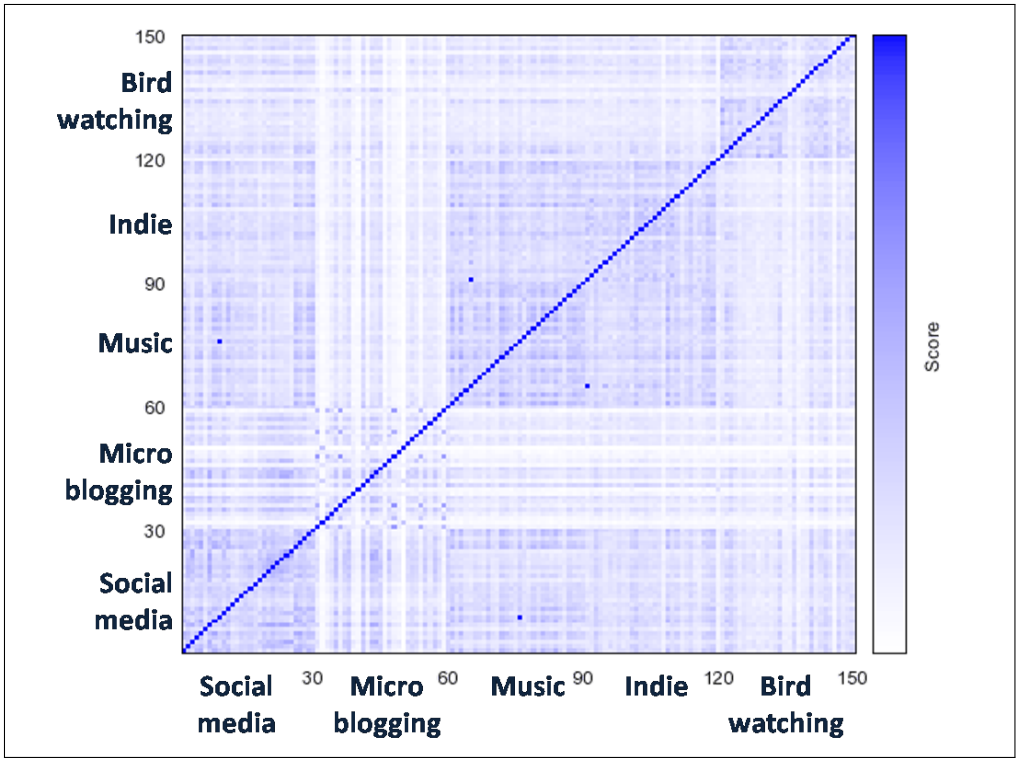

Figure 7.2. Comparison of microblogger groups using unweighted cosine similarity over the top 100 tokens.

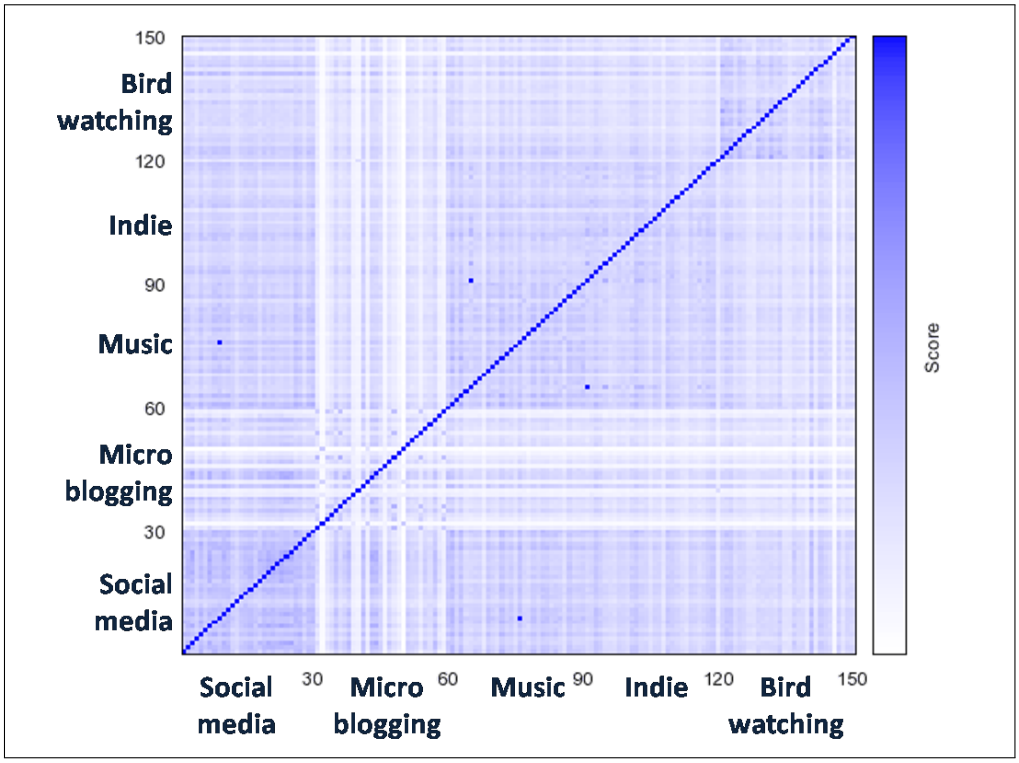

Figure 7.3. Comparison of microblogger groups using unweighted cosine similarity over the top 1000 tokens.

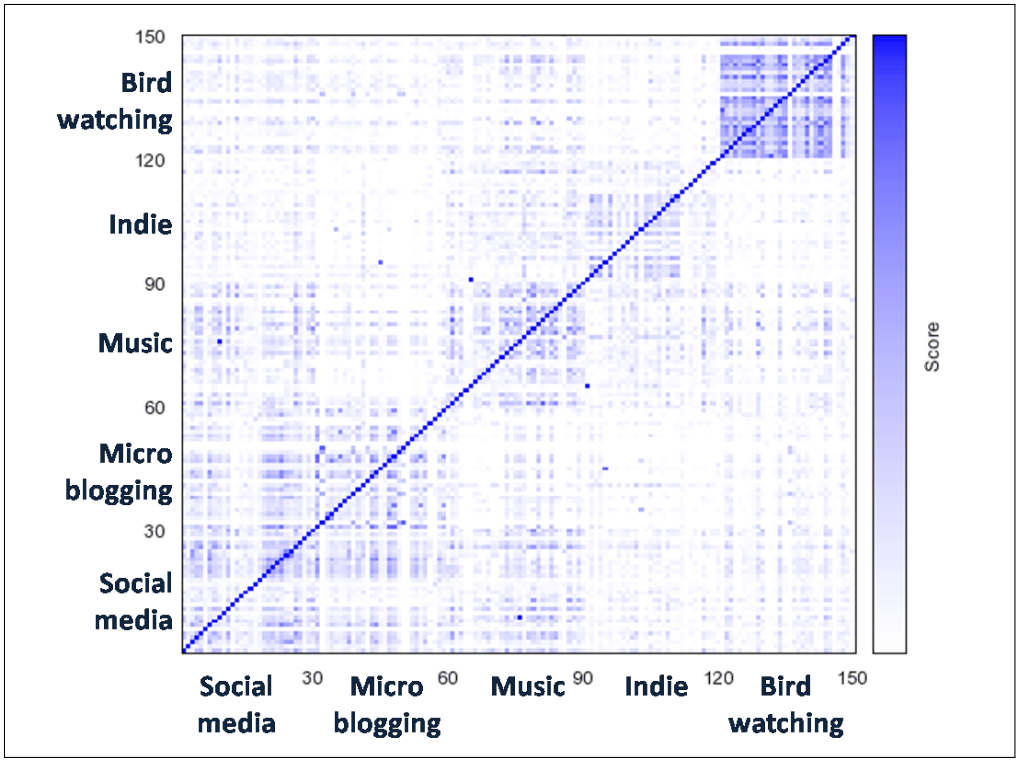

Figure 7.4. Comparison of microblogger groups using weighted cosine similarity over the top 10 tokens.

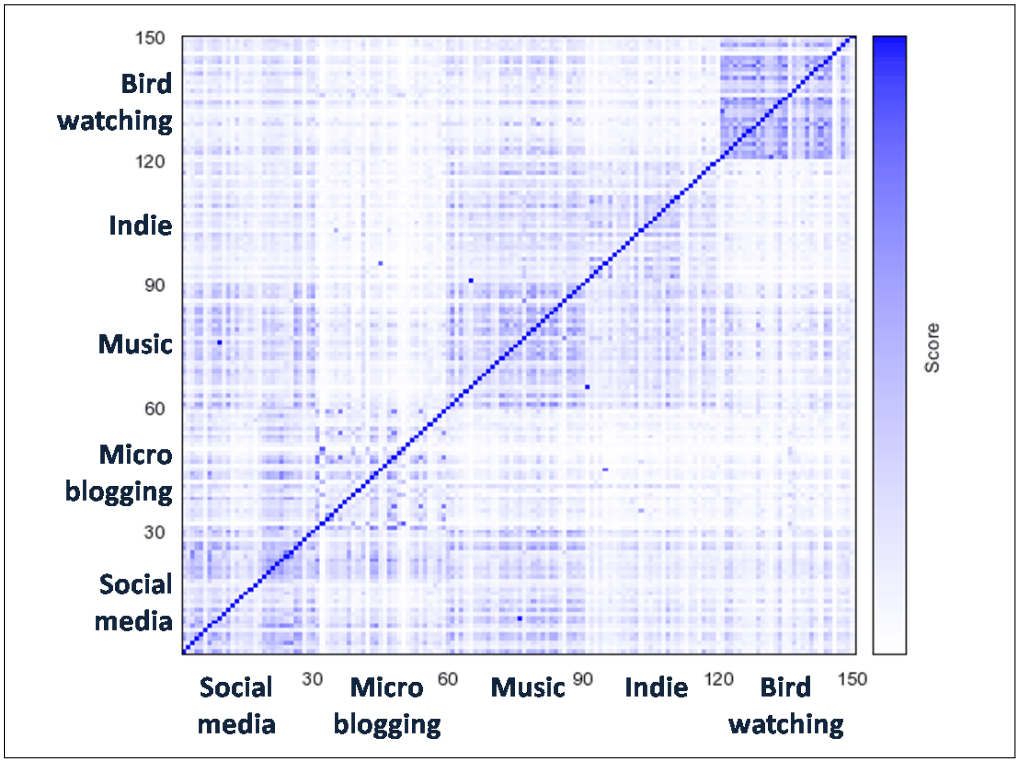

Figure 7.5. Comparison of microblogger groups using weighted cosine similarity over the top 100 tokens.

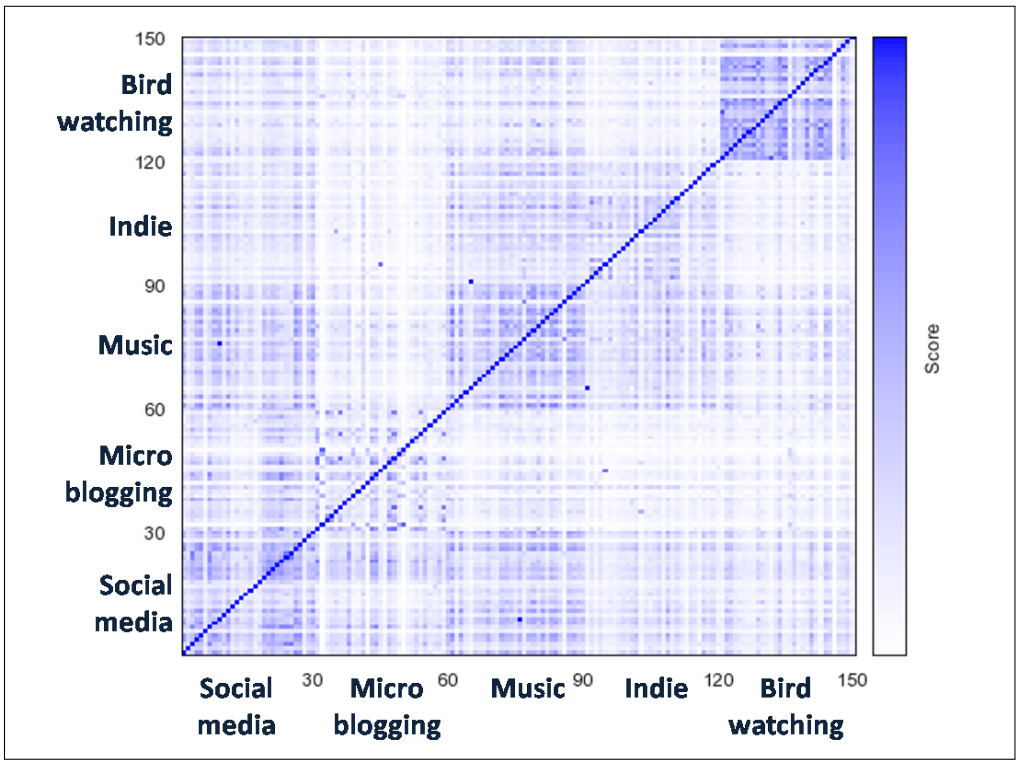

Figure 7.6. Comparison of microblogger groups using weighted cosine similarity over the top 1000 tokens.

include tokens which are also used by other users in other groups. So, socialmedia users being similar to users in other groups is not surprising.

- When it was observed that microblogging group did not resemble any other group much, the contributions of users in this group were examined. It was discovered that microblogging set had many non-English contributions. So, this result is plausible. But among all other four groups, microblogging users are most similar to socialmedia users. As microblogging is a social media service, this was expected.
- Music is another generally popular interest, and like social media, popular users who utter about general topics add music to their interests. Since users in both socialmedia and music groups typically contribute about general topics, socialmedia users and music groups being similar to each other is not surprising.
- In indie group, users are most similar to music users aside of themselves. As indie is a music genre, this was expected.
- Birdwatching users do not resemble any other group users. Only users who are

really interested in and contribute about this specific area add birdwatching to their interests. So, this result was expected. It is more probable that birdwatching users would also be interested in general topics, rather than other specific interest. In the figures, it can be observed that birdwatching users are more similar to popular groups (social media and music) than specific groups (microblogging and indie).

- Besides the diagonal line, there are four points of 1.0 similarity in the figures. When users in 5 groups were examined, it was discovered that two users belonged to two groups (One user belonged to music and indie, and one user belonged to music and social media). These points represent these users.

8. DISCUSSION AND FUTURE WORK

Due to computational constraints and time limitations, our model is tested on 150 users. Although the results are consistent and promising, large number of users over more interests should further be used.

Presently, tweets are tokenized by white space characters. Hence, every word in microblogger descriptions is a single token. Phrases and multi-word expressions are not taken into consideration. Also, an abbreviation like 'U.S.' is considered as two one-letter tags, 'U' and 'S', since punctuation symbols are removed from the contributions. Natural language processing will greatly improve these results.

In processing contributions, the contents of tweets and the titles of web links were used. This choice was due to performance reasons. In Figure 7.9, it can be observed that 26% of significant tokens were collected from the titles of web links. Given the encouraging results, using more content from the web pages accessed from the external links should be investigated. It may be useful to compare user descriptions with using only the titles and using more content from the web pages.

Hashtags are user specified references. Gathering more information from these references, and associating this information with users will be interesting. Hastags are used much more deliberately. They are also conventionally adopted by users to relate relevant content. These properties should render hashtag processing highly relevant. Token clouds of hashtags can be generated in a similar manner to describe hashtags. Later, these descriptions can be used to improve users' descriptions. Although this process may be useful, due to the number of distinct hashtags used by a microblogger, analyzing every hashtag in a user's description may be exhausting. Also, typically, hashtag use is temporal (e.g. hashtags for conferences). Thus, gathering data for a formerly popular hashtag is difficult.

A case study of information diffusion in social networks may be pursued for

hashtags. How hashtags emerge and propagate among Twitter users may be analyzed.

A microblogger is followed by others who share similar interests. Similarly, microbloggers follow others who contribute contents of their interests. For describing users, their followers, or the users followed by them may be analyzed. Token clouds of followers or followees may be generated and merged with microblogger's own token cloud to describe a microblogger.

Threshold values used in constructing common words lists and user categorization were set heuristically. A more formal approach, such as statistical analysis methods must be applied for determining threshold values.

Descriptions of microbloggers generated as a result of this study may be used for prediction and recommendation. Observing changes in other microbloggers' descriptions in time, by analyzing past tokens in a microblogger's contributions, microblogger's future subjects may be predicted. Comparisons of microblogger descriptions may be utilized for recommendation. Observing the present microbloggers followed by a user, other microbloggers who contribute similarly may be recommended. For prediction and recommendation, a wide range of microbloggers must be analyzed and described.

The most significant direction being pursued is to determine the context of tokens using semantic web techniques. Semantic web processing can be applied to cluster tokens to reveal different interests, such as tagging a user with 'music' by grouping the words 'Hendrix', 'guitar', and 'tune' in a word cloud. Semantic tagging is expected to give much better results for comparing, describing and searching microbloggers. Ontologies and semantic data about many *things* like people, places, concepts, etc. is present [45]. This information can be used greatly enrich our results.

9. CONCLUSION

This work proposed an approach for examining microblogger content to reveal the subjects of their contributions and their characteristics.

A system was designed and implemented to analyze microbloggers (Twitter users). Meta information about users were used to reveal their characteristics and their contributions were processed resulting in a weighted set of tokens. These tokens are visualized as a tag cloud. In order to assess the the process, sets of users who declared their interests on the user list creator service *WeFollow* were selected. Word clouds for the individuals in these lists were constructed, and inspected for similar word usage. As a byproduct, common words specific to groups and common words of microbloggers in general were listed.

The results are encouraging. Common words found to be specific to groups are related to the group interests. Comparison of users show that users belonging to a same group are more similar to each other than any other users, as expected.

Results indicate that the proposed approach can be used to describe microbloggers. Future work indicated in Chapter 8 can improve the results.

APPENDIX A: STOP WORDS LIST

Stop words are the words used so commonly that they have no distinguishing property. There are many stop words lists for various languages. In this study, a list of stop words for English (570 words) is constructed using SMART system's list [46]. These words are listed below.

a a's able about above according accordingly across actually after afterwards again against ain't all allow allows almost alone along already also although always am among amongst an and another any anybody anyhow anyone anything anyway anyways anywhere apart appear appreciate appropriate are aren't around as aside ask asking associated at available away awfully b be became because become becomes becoming been before beforehand behind being believe below beside besides best better between beyond both brief but by c c'mon c's came can can't cannot cant cause causes certain certainly changes clearly co com come comes concerning consequently consider considering contain containing contains corresponding could couldn't course currently d definitely described despite did didn't different do does doesn't doing don't done down downwards during e each edu eg eight either else elsewhere enough entirely especially et etc even ever every everybody everyone everything everywhere ex exactly example except f far few fifth first five followed following follows for former formerly forth four from further furthermore g get gets getting given gives go goes going gone got gotten greetings h had hadn't happens hardly has hasn't have haven't having he he's hello help hence her here here's hereafter hereby herein hereupon hers herself hi him himself his hither hopefully how howbeit however i i'd i'll i'm i've ie if ignored immediate in inasmuch inc indeed indicate indicated indicates inner insofar instead into inward is isn't it it'd it'll it's its itself j just k keep keeps kept know knows known l last lately later latter latterly least less lest let let's like liked likely little look looking looks ltd m mainly many may maybe me mean meanwhile merely might more moreover most mostly much must my myself n name namely nd near nearly necessary need needs neither never nevertheless new next nine no nobody non none noone nor normally not nothing novel now nowhere o obviously of off often oh ok okay old on once one ones only

onto or other others otherwise ought our ours ourselves out outside over overall own p particular particularly per perhaps placed please plus possible presumably probably provides q que quite qv r rather rd re really reasonably regarding regardless regards relatively respectively right s said same saw say saying says second secondly see seeing seem seemed seeming seems seen self selves sensible sent serious seriously seven several shall she should shouldn't since six so some somebody somehow someone something sometime sometimes somewhat somewhere soon sorry specified specify specifying still sub such sup sure t t's take taken tell tends th than thank thanks thanx that that's thats the their theirs them themselves then thence there there's thereafter thereby therefore therein theres thereupon these they they'd they'll they're they've think third this thorough thoroughly those though three through throughout thru thus to together too took toward towards tried tries truly try trying twice two u un under unfortunately unless unlikely until unto up upon us use used useful uses using usually uucp v value various very via viz vs w want wants was wasn't way we we'd we'll we're we've welcome well went were weren't what what's whatever when whence whenever where where's whereafter whereas whereby wherein whereupon wherever whether which while whither who who's whoever whole whom whose why will willing wish with within without won't wonder would would wouldn't x y yes yet you you'd you'll you're you've your yours yourself yourselves z zero

REFERENCES

1. Ryan, Z., *the good life—new public spaces for recreation*, Von Alen Institute, 30 W 22 Street, 6th Floor, New York, NY 10010, USA, 2006.

2. Twitter, "A very popular microblogging service", `http://www.twitter.com`, April 2010.

3. Second Life, "A virtual world where users can socialize, participate in individual and group activities, and create and trade virtual property and services with one another, or travel throughout the world", `http://secondlife.com/`, June 2010.

4. Wikipedia, "A free, web-based, collaborative, multilingual encyclopedia project", `http://www.wikipedia.org`, December 2009.

5. SwarmSketch, "An online collective sketching tool", `http://swarmsketch.com`, June 2010.

6. dreaming wall, "A wall where messages sent by public are displayed randomly and continuously", `http://www.dreamingwall.net`, June 2010.

7. SMS, "Short Message Service", `http://en.wikipedia.org/wiki/SMS`, June 2010.

8. Burble, "A floating interactive architecture made with balloons", `http://www.haque.co.uk/burble.php`, October 2009.

9. Flickr, "An online photo management and sharing application", `http://www.flickr.com/`, May 2010.

10. YouTube, "A web site to discover, watch, upload and share videos", `http://www.youtube.com/`, April 2010.

11. WordNet, "A lexical database for the English language", `http://wordnetweb.princeton.edu`, March 2010.

12. McFedries, P., "Technically Speaking: All A-Twitter", *Spectrum, IEEE*, Vol. 44, pp. 84–84, 2007, `http://ieeexplore.ieee.org/xpls/abs\_all.jsp?arnumber=4337670`.

13. Jaiku, "A microblogging system", http://www.jaiku.com, June 2010.

14. Tumblr, "A short form blog system", http://www.tumblr.com, June 2010.

15. South by Southwest, "A private company based in Texas that is known for Music and Media Conference and Festival", http://sxsw.com, May 2010.

16. "The official Twitter Developer Conference", http://chirp.twitter.com/index.html, April 2010.

17. Biz Stone at Chirp, "Statistics regarding the popularity of Twitter", http://www.justin.tv/twitterchirp/b/262219316, June 2010.

18. Twitter Blog, "New Twitter account rate outside the US", http://blog.twitter.com/2010/04/growing-around-world.html, April 2010.

19. cfa_updates, "RSS feeds from the CFA website to Twitter, posting fire incidents with their locations", http://twitter.com/cfa_updates, November 2009.

20. sfearthquakes, "A Twitter user that posts earthquake news in SF Bay area", http://twitter.com/sfearthquakes, November 2009.

21. Twitter API, "Twitter Application Programming Interface wiki", http://apiwiki.twitter.com, September 2009.

22. Twitpic, "A website that allows users to easily post pictures to the Twitter microblogging service", http://twitpic.com, December 2009.

23. Tweetdeck, "A desktop application for Twitter, Facebook, LinkedIn, Google Buzz, Foursquare, and MySpace", http://www.tweetdeck.com, June 2010.

24. Facebook, "A social networking website", http://www.facebook.com, May 2010.

25. Myspace, "A social networking website", http://www.myspace.com, January 2010.

26. Sandler, D. R. and D. S. Wallach, "Birds of a FETHR: Open, Decentralized Micropublishing", *8th International Workshop on Peer-to-Peer Systems (IPTPS '09) April 21, 2009, Boston, MA*, 2009, http://www.usenix.org/events/iptps09/tech/.

27. Ebner, M. and M. Schiefner, "Microblogging - more than fun?", *Proceedings of IADIS Mobile Learning Conference 2008,*, pp. 155–159, 2008, http://lamp.tu-graz.ac.at/~i203/ebner/publication/08_mobillearn.pdf.

28. Grosseck, G. and C. Holotesku, "Can we use Twitter for educational activities?", *4th Scientific Conference eLSE "elearning and Software for Education"*, 2008, http://www.scribd.com/doc/2286799/Can-we-use-Twitter-for-educational-activities.

29. Ebner, M. and H. Maurer, "Can Microblogs and Weblogs change traditional scientific writing?", *E-Learn 2008*, pp. 768–776, 2008, http://lamp.tu-graz.ac.at/\~{}i203/ebner/publication/08_elearn01.pdf.

30. Zhao, D. and M. B. Rosson, "How Might Microblogs Support Collaborative Work?", *Workshop on Social Networking in Organizations, November 9, 2008, San Diego*, 2008, http://research.ihost.com/cscw08-socialnetworkinginorgs/.

31. Zhao, D. and M. B. Rosson, "How and why people Twitter: the role that microblogging plays in informal communication at work", *GROUP '09: Proceedings of the ACM 2009 international conference on Supporting group work*, pp. 243–252, ACM, New York, NY, USA, 2009.

32. Dean, "Twitter: an introduction to microblogging for health librarians", *JCHLA*, Vol. 30, No. 1, 2009, http://pubs.nrc-cnrc.gc.ca/jchla/jchla30/c09-009.pdf.

33. Huberman, B. A., D. M. Romero, and F. Wu, "Social networks that matter: Twitter under the microscope", *CoRR*, Vol. abs/0812.1045, 2008, http://uk.arxiv.org/abs/0812.1045v1.

34. Vieweg, S., A. L. Hughes, K. Starbird, and L. Palen, "Microblogging during two natural hazards events: what twitter may contribute to situational awareness", *CHI '10: Proceedings of the 28th international conference on Human factors in computing systems*, pp. 1079–1088, ACM, New York, NY, USA, 2010.

35. Jansen, B. J., M. Zhang, K. Sobel, and A. Chowdury, "Micro-blogging as online word of mouth branding", *CHI '09: Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*, pp. 3859–3864, ACM, New York, NY, USA, 2009.

36. Honeycutt, C. and S. C. Herring, "Beyond Microblogging: Conversation and Collaboration via Twitter", *Hawaii International Conference on System Sciences*, Vol. 0, pp. 1–10, 2009.

37. Shamma, D. A., L. Kennedy, and E. F. Churchill, "Tweet the debates: understanding community annotation of uncollected sources", *WSM '09: Proceedings of the first SIGMM workshop on Social media*, pp. 3–10, ACM, New York, NY, USA, 2009.

38. Java, A., X. Song, T. Finin, and B. Tseng, "Why we twitter: understanding microblogging usage and communities", *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pp. 56–65, ACM, New York, NY, USA, 2007, `http://dx.doi.org/10.1145/1348549.1348556`.

39. Krishnamurthy, B., P. Gill, and M. Arlitt, "A few chirps about twitter", *WOSP '08: Proceedings of the first workshop on Online social networks*, pp. 19–24, ACM, New York, NY, USA, 2008, `http://dx.doi.org/10.1145/1397735.1397741`.

40. Weng, J., E.-P. Lim, J. Jiang, and Q. He, "TwitterRank: finding topic-sensitive influential twitterers", *WSDM '10: Proceedings of the third ACM international conference on Web search and data mining*, pp. 261–270, ACM, New York, NY, USA, 2010.

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

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

43. Cosine Similarity, "A measure of similarity between two vectors of n dimensions by finding the cosine of the angle between them", http://en.wikipedia.org/wiki/Cosine_similarity, May 2010.

44. WeFollow, "A directory of Twitter users organized by interests", http://wefollow.com, April 2010.

45. Linked Data, "A semantic web project to describe a method of exposing, sharing, and connecting data via dereferenceable URIs on the Web", http://linkeddata.org, June 2010.

46. Stop Words List, "A stop words list for English", ftp://ftp.cs.cornell.edu/pub/smart/english.stop, May 2010.