Declaration Form

University of Dhaka

We, Nitish Ranjan Bhowmik and Kazi Mazhah Uddin, declare that this work presented in this thesis is the outcome of the investigation performed by us under the supervision of Anna Fariha, Lecturer, Department of Computer Science and Engineering, University of Dhaka. We also declare that no part of this thesis has been or is being submitted elsewhere for the award of any degree or diploma.

Signature:

(Ms. Anna Fariha)
Lecturer
Department of Computer Science and Engineering
University of Dhaka
Thesis Supervisor

Signature:

(Nitish Ranjan Bhowmik)
Candidate

Signature:

(Kazi Mazbah Uddin)
Candidate

Estimating Geo-Location of Social Media Users based on Public Content

by

Nitish Ranjan Bhowmik

Exam Roll No: Curzon-149

Registration No: 2011-012-074, Session: 2011-12

Kazi Mazbah Uddin

Exam Roll No: Curzon-131

Registration No: 2011-012-076, Session: 2011-12

Supervised by

Ms. Anna Fariha

Lecturer

Department of Computer Science and Engineering

University of Dhaka

A thesis submitted in partial fullment of the requirements for the degree of

Bachelor of Science in Computer Science and Engineering
at the University of Dhaka

March 2016

Abstract

With the phenomenal growth of Internet social media services, such as microblogging and social networking, which are offered by platforms such as Twitter, Facebook etc., interactions among people are increasing rapidly. People share their different aspects of daily life in these microblogging sites. They also seek needful information in these sites which varies from location to location. Therefore, finding locations of these users is an interesting idea which also helps to detect fake accounts also. Users of a particular area have some location specific topics in their dialogue and a user has links with other users of that city due to her professional involvement. Terms used in a user's microblog and location of her followers, we can employ a probabilistic framework to estimate city level location of that user. To estimate location of a user, in our approach, we use the contents of a social user's microblogs and locations of her followers. We do not using any external information of a user to estimate her location, such as gazetteer or IP information. Our approach achieves 8% higher accuracy than the sate of art approaches. Experimental results are showed to signify the effectiveness of our proposed approach for estimating a social network user's location.

Acknowledgements

We would like to thank our thesis supervisor Ms. Anna Fariha, Lecturer, Department of Computer Science and Engineering, University of Dhaka, for her proper guidance in selecting the research area and supporting by all means in the process of the research work. Her proper guidance and efforts made this work possible.

We are highly grateful to our parents, family members, friends and senior fellows who helped us and gave us moral support and inspiration.

We would like to thank Amit Mandal and Quazi Marufur Rahman for helping in tools required for thesis writing.

Last, but not the least, we would like to thank our department, Department of Computer Science and Engineering, University of Dhaka, for giving us the opportunity for this research work and facilitate us throughout the whole Bachelor of Science program.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Internet social media services, such as microblogging and social networking, which are offered by platforms such as Twitter and Facebook are getting highly popular in their user bases. These users share their daily activities, feelings, opinions through microblogs, which produce huge amount of data every day. This huge amount of data can be used to mine various interesting information, such as geographic location of users. Geographic location information can be used in many important applications, such as advertisement, recommender system etc. It also can be used to detect fake accounts in these social media. Microblogging sites allow its user to specify their location as user information. This location information is taken from the user while creating account or with the help of Global Position System (GPS). But there are several drawbacks of relying on user's manual information. A user can enter wrong information or a user can live in multiple locations. So, relying on user's manual information is not always effective. To overcome this problem, we provide a new approach to estimate a user's city level location based on the contents of her microblogs and location of followers.

1.1 Data Mining

Data mining is the process of discovering interesting patterns and knowledge from large amount of data. The data sources can include databases, data warehouse, the web, or stream of data that are captured into the system dynamically. Microblogging social media like Twitter and Facebook, are producing vast amount of data every day. Analyzing such data is necessary to mine important knowledge. As a new research field, data mining has made significant progress and covered a broad spectrum of applications since the 1980s. Today, data mining is used in a vast array of areas. Numerous commercial data mining systems and services are available such as market basket analysis, stock market analysis etc.

1.1.1 Mining Web Data

The Word Wide Web is a vast source of global information for advertisements, consumer information, news, government and e-commerce. It contains a rich and dynamic collection of information about web page contents with hypertext structures and multimedia. Web mining can be classified into three areas: web content mining, web structure mining, and web usage mining. Among them, web content mining analyzes web content such as text, multimedia data and structured data.

1.1.2 Text Mining

Text mining is an interdisciplinary field that draws on information retrieval, data mining, statistics, and machine learning. A big portion of information is stored as text such as news, blogs, articles, technical papers, books, and web pages. Mining text from these sources is an important concept. Text mining usually requires structuring the input text (e.g., parsing, along with the addition of some derived linguistic features and removal of others, and subsequent insertion into a database).

Text mining includes text clustering, text categorization, concept extraction, document summarization, and entity-relation modeling. Various types of text mining tools and software are available in academic institutions, open source forums and industry. Text mining often uses WordNet, Semantic web, Wikipedia, and other information sources to enhance the understanding and mining of text data.

1.2 Social Media

Social media are computer-mediated tools that allow people or companies to create, share, or exchange information, career interests, ideas, and pictures/videos in virtual communities and networks. Social media is defined as ”a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content”. Furthermore, social media depend on mobile and web-based technologies to create highly interactive platforms through which individuals and communities share, co-create, discuss, and modify user-generated content. They introduce substantial and pervasive changes to communication between businesses, organizations, communities, and individuals.

There are different forms of social media, such as blogs, microblogs, photo sharing, video sharing, social network, social gaming, virtual worlds etc. Microblogging is a broadcast medium that exists in the form of blogging. A microblog differs from a traditional blog in that its content is typically smaller in both actual and aggregated file size. Microblogs "allow users to exchange small elements of content such as short sentences, individual images, or video links", which may be the major reason for their popularity [7].

1.2.1 Social Media Mining

Social media mining is the way of representing, analyzing, and extracting actionable patterns from social media data. The growth of social network over the last decade has revolutionized the way individuals interact and industries conduct business. Individuals produce data at an unprecedented rate by interacting, sharing, and consuming content through social network. Understanding and processing this new type of data to glean actionable patterns presents challenges and opportunities for interdisciplinary research, novel algorithms, and tool development. Social media Mining integrates social media, social network analysis, and data mining to provide a convenient and coherent platform for students, practitioners, researchers, and project managers to understand the basics and potentials of social network mining. It introduces the unique problems arising from social network data and presents fundamental concepts, emerging issues, and effective algorithms for network analysis and data mining. Suitable for use in advanced undergraduate and beginning graduate courses as well as professional short courses, the text contains exercises of different degrees of difficulty that improve understanding and help apply concepts, principles, and methods in various scenarios of social network mining.

1.2.2 Location Mining of Social Media Users

Social media mining is the technique of analyzing, representing and extracting interesting patterns from social media data. With the immense spread of social media over the last decade, individuals produce data at a revolutionary rate by sharing and interacting through social media. Finding interesting knowledge by processing this huge amount of data is a demanding research work. User's location mining is one of those interesting research works. There are many approaches to find geo-location of users based on different contents such as using text of microblogs, events, friends etc. But there are still a lot researches are needed to improve the existing approaches of location estimation.

1.3 Motivation

Finding locations of users of microblogging sites is an interesting research topic. It helps to provide better recommendation of advertisements, news, necessary sites for the users. It can help to fake accounts also. So, estimate location of users is worthy information to collect. Location can be obtained from the user's manual given information or using GPS technology. But they have several drawbacks and all devices are not GPS enable. In addition, users can provide wrong location information intentionally or unintentionally. Many researchers provide several approaches to estimate a user's geographic location. Most of the approaches estimation have high rate of error and some of the approaches uses external information, such as gazetteer, IP information etc. which motivates us to construct a more efficient approach to estimate the geographic location of the users of microblogging social sites.

1.4 Objective

Objectives of our research are as follows:

1. To propose an approach that estimate geo-location of microblogging site users without using any external information, such as (i.e.) gazetteer, IP information.
2. Our proposed approach should estimate with higher accuracy than the state of art approach.

The potential of utilizing location of followers of a user to estimate her own location was our motivation of this research.

1.5 Thesis Organization

The research work is carried out in order to achieve the objectives of the research. A brief overview of the contents of the chapters is given below:

Chapter 2 describes some background studies and related works of this research work.

Chapter 3 provides our proposed approach to estimate a user's geographic location and the dataset, we used in experimental process.

Chapter 4 discusses the dataset and the results of our experiment.

Chapter 5 concludes the research and provides some future enhancements for more improvement of our work.

Chapter 2

Literature Review

2.1 Preliminaries

Internet social media are getting popular through all over the world. Finding a user's location is an interesting and popular research topic. Researchers proposed many approaches to estimate geographic location of users using the content of social media, blogs [8], web pages [9] etc. Most of the approaches depend on external information to estimate location. But our work does not rely on external information.

2.2 Data Mining

The procedure of discovering pattern or knowledge from data streams, e.g., databases, texts, web etc., is called the Data Mining. It combines our ideas and resources from multiple sectors of computer science and statistics. Data mining techniques are used to find patterns, structure or format and regularities from growing data sets. Many efficient algorithms which are capable of handling large and growing data sets. These algorithms have their scalability or linear complexity with respect to the data size, generally called big data.

Knowledge discovery from data (KDD) implies the process of extracting information from raw data. Figure 2.1 shows the complete version of the KDD process. The point is that Data mining (DM) is the root of KDD process, concerning the deriving of algorithms that test the data, organize the model and discover the already unknown patterns. The model is used for understanding phenomena from the data, analysis and prediction.

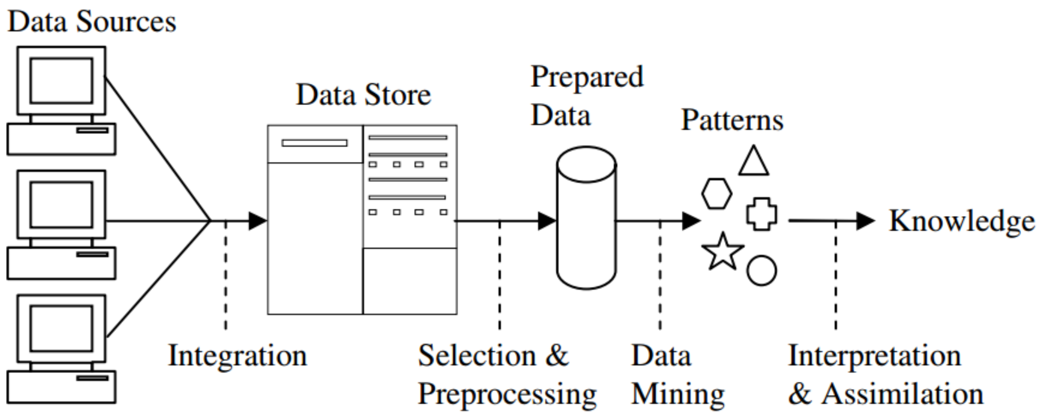


Figure 2.1: Knowledge discovery process [1]

2.2.1 Applications of Data Mining

Data Mining Techniques [10] have been adapted successfully in recent areas such as market basket analysis, analysis of organic compounds, business application, software engineering, automatic abstracting, fraud detection etc.

2.3 Text Mining

Text mining is generally used to denote any system, tools or technique that analyzes large quantities of natural language text [11] and detects lexical or linguistic usage patterns in an attempt to extract probably useful information. It is an intensive process of knowledge mining in which a user interacts with a document (contains of text or words). This document is collected over time by using a suite of analysis tools. It is a strategy of assigning documents with predefined categories that are associated with their contents. According to data mining literature, text mining is used to extract useful information from data sources. This data sources can be collected through identification and exploration of interesting patterns from the text. In text mining, generally, the data sources are document collections, resourceful patterns. Those are not also found among formalized database records but also in the unstructured textual data in the documents.

Text mining is conceptualized as a subsection of text analytics which is focused on applying data mining techniques. The term text analytics is similar with text mining. Text mining considers only syntax that has a relationship between words. It does not deal with phonetics, pragmatics and discourse. In recent years, text mining is believed to have a commercial potential which is higher than that of data mining. In fact, a recent study in Bangladesh indicated that 76% [12] of a company's/business information are contained in text documents. However it is complex task than data mining as it involves dealing with text data that are inherently unstructured and fuzzy. Before starting text mining, we have to consider some tools or techniques that is discussed in 2.3.1.

2.3.1 Framework of Text Mining

Visualization of text mining can be considered into two steps: *Text Refining* and *Knowledge Distillation*. *Text Refining* transforms the unstructured data pattern *Intermediate Form* (IF). IF can be a document or concept based but *Knowledge Distillation* is document base objects. This document can be deduced by the IF according to patterns or knowledge.

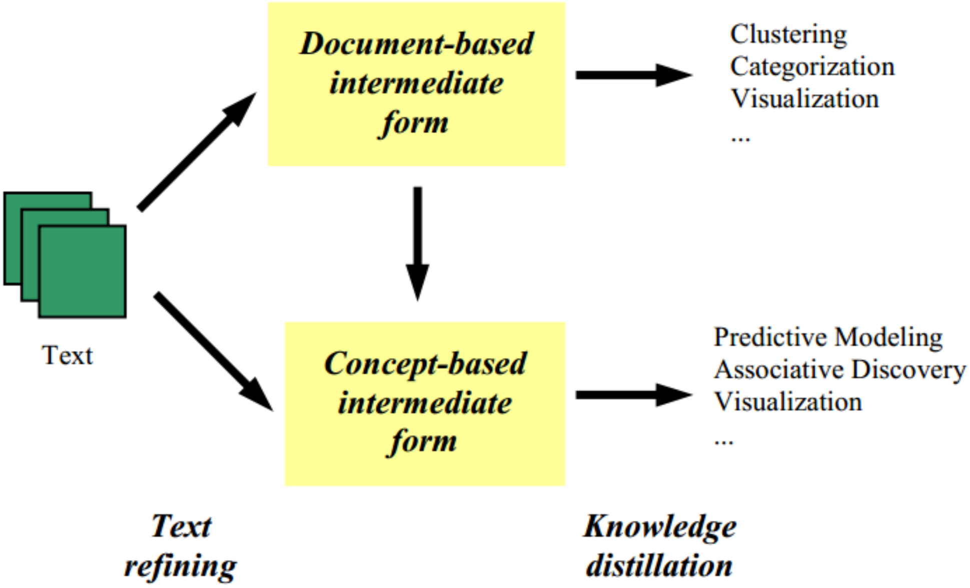
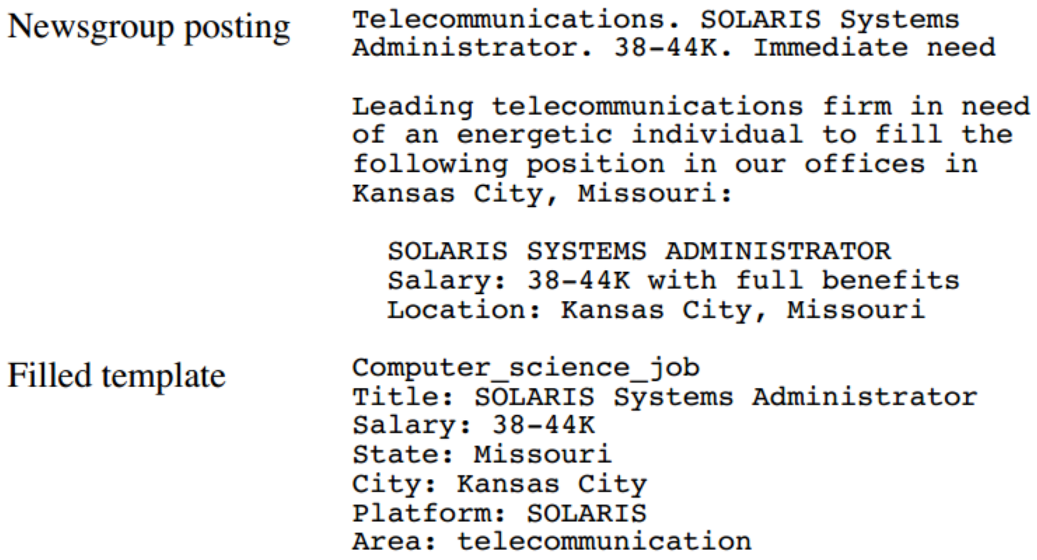


Figure 2.2: A text mining framework [2]

By extracting object information or relevant data domain, a document based IF can be transformed into concept based IF. *Knowledge Distillation* from a concept-based IF deduces the patterns or knowledge relationship across objects or concepts. Figure 2.2 describes the framework of text mining.

2.3.2 Document Collection

Before starting the text mining, it is important to collect the data sources that is called document collection. Document collection may be dynamic or static. Dynamic document can be modified by the insertion of data streams over time but static text document remains unchanged. In text mining process large amount of document can improve the performance of optimization challenges. It does not run its knowledge discovery process algorithm on unprepared document collections. A document is a combination of discrete textual data that has a relationship within text documents eg., business applications, web logs, data press etc. Figure 2.3 implies that Newsgroup Posting and Filled Template are the classification of data and across the text. From this text, we generate the word document with term frequency.



Newsgroup posting

```
Telecommunications. SOLARIS Systems
Administrator. 38-44K. Immediate need

Leading telecommunications firm in need
of an energetic individual to fill the
following position in our offices in
Kansas City, Missouri:

  SOLARIS SYSTEMS ADMINISTRATOR
  Salary: 38-44K with full benefits
  Location: Kansas City, Missouri
```

Filled template

```
Computer_science_job
Title: SOLARIS Systems Administrator
Salary: 38-44K
State: Missouri
City: Kansas City
Platform: SOLARIS
Area: telecommunication
```

Figure 2.3: Combination of discrete textual data [3]

2.3.3 Text Categorization

Text categorization (or text classification) is the assignment of natural language documents to predefined categories according to their content. The set of categorization is referred to as "controlled vocabulary" for analyzing the complex data in text mining. So, *Text categorization*(TC) is given a set of categories (subjects, topics) and

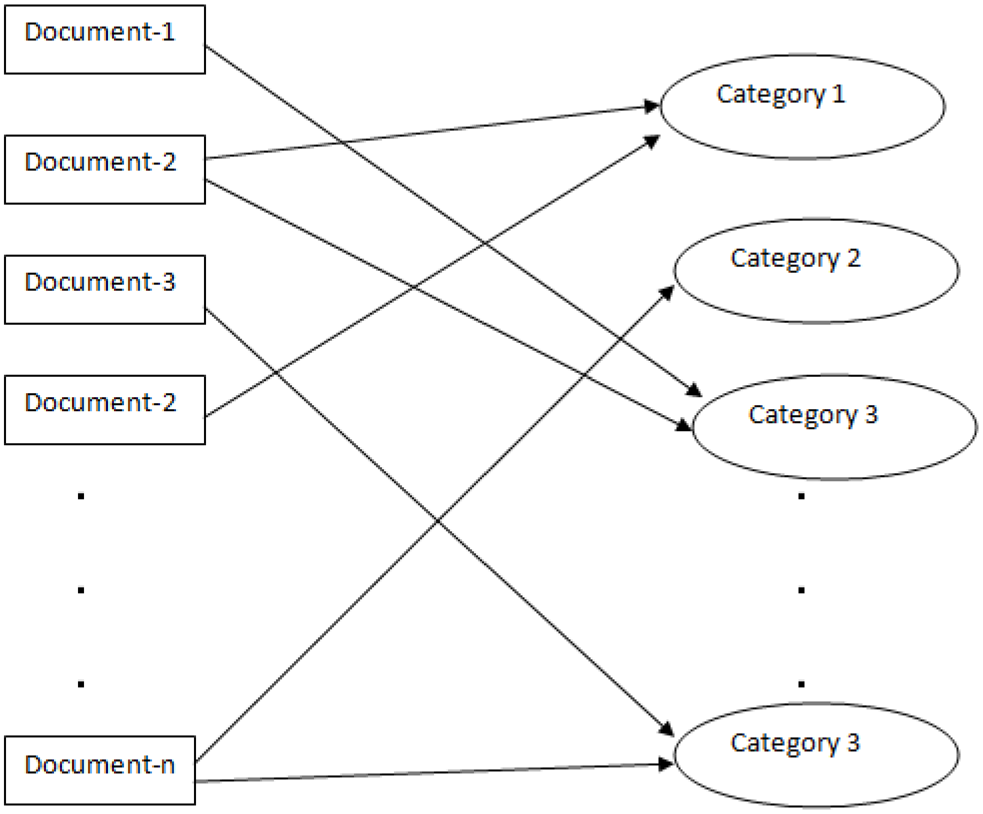


Figure 2.4: Text categorization involving multiple per document

a collection of text documents, the process of finding the correct category for each document. There are three common TC applications of text mining. These are text indexing, document sorting and text filtering and, web page categorization. These are only a small set of possible applications, but they demonstrate the diversity of the domain and the variety of the TC sub-cases.

2.3.4 Text Parsing & Transformation

Text need to extract, clean and create dictionary of words from the documents using parsing algorithms. This includes processing of words without grammatical mistake, determine the parts of speech, identifying the text document. The next step after text parsing is text transformation that means creation of a term by document matrix according to some models like vector, latent semantic indexing(LSI), singular value decomposition(SVD) etc.

Example : Consider a collection of text in three documents as provided below:

Document 1 : Everything happens for a reason so I mean I just got to let you go then I got to.
Document 2 : The place where you can study for a test for six hours and reason still fail miserably happens.
Document 3 : Genuinely I am happy without a few people in my life.
Parsing this document that generates term by document matrix in table 1.

Term	**Document-1**	**Document-2**	**Document-3**
Everything	1	1	0
Happen	1	1	0
Reason	1	1	0
Mean	1	0	0
I	3	0	1
Let	1	0	1
Hour	0	1	0
Happy	0	0	1
Study	0	1	0
You	0	1	0
...	...	...	...

Table 2.1: Term by document matrix

2.3.5 Text Document Reprocessing

Before collection of text documents for analysis, one have to consider some relevant tasks such as stemming, stop ward removal, handling of digits of character or words, phrases, hyphens, cases of the letters and punctuations etc.

2.3.6 Tokenization

Text tokenization means breaking up text into tokens. These tokens can be be phrase, words, symbols etc. It helps to distinguish between words i.e., western languages with no explicit word boundaries. To determine words, tokenization needs. Example: Doc 1:

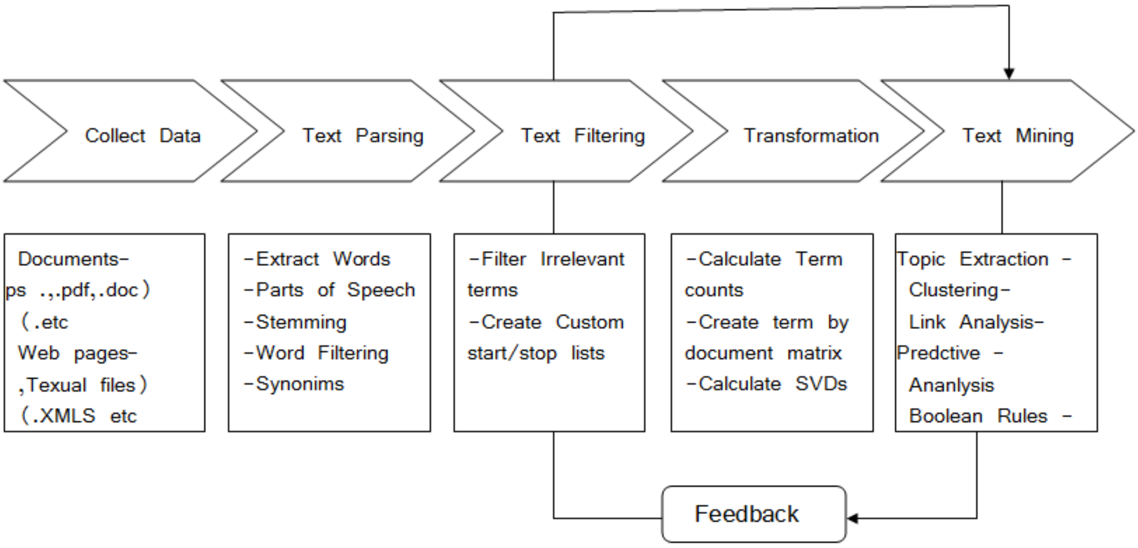


Figure 2.5: Text mining process flow

This is a text.	[This is a text.]
This is a text.	[This, is, a, text]
This is a text.	[T, h, i, s, i, s, a, t, e, x, t, .]

2.3.7 Stemming

A sentence or a word has a collection of grammatical specifications in a language. Like wise in English, it has many grammatical sections like present, past and future tense of verb in singular or plural forms or nouns, pronouns or adverb etc. As token appears in different forms, so stemming is used to normalize these words in original form. Stemming can remove the prefix or suffix from words.

Example: “Drinking” reduces to drink and “simplest” reduces to simple

2.3.8 Stop-Word Removal

Noun, pronoun, adverb, conjunction etc. are the stop words. Those do not contain any significance of the document. Those are less important for document parsing or tokenization. Language identification can be done by using of stop ward removal technique.

Example: at, who, when, which, but, and etc. these are the unnecessary words for the documents. It should be removed from the document during text parsing.

Hyphens

Hyphens usually distract the words. So removing the hyphens from words may generate consistency of the databases.

Example: "faces-of-the-region" may be replaced by the "face of the region" or "faceoftheregion".

Case of Letters

All words of the document can be formalized as either in upper case or lower case letter

Digits

Digits in word may create inconsistency in the database. Removal of these characters from the document, makes it consistent for processing.

Punctuation Marks

'!', '#', '*', '%', '$', '@' these characters should be removed from documents because those characters should be removed from documents.
Text mining is a multidimensional field [13] involving — information retrieval, text analysis, information extraction, clustering, categorization, machine learning, data mining etc.

2.4 Social Media

Social Media is a composition of sociology and technology. It sets up an environment or platform on the internet where people share their views, ideas and experiences. *"Social Media are web-based tools for interaction that, in addition to conversation, allows users to share content such as photos, videos and links to resources"* [14]. Social media are digital tools for sharing conversation and content. It is map of relationship between individuals, ranging from acquaintances to close to familial bonds.

2.4.1 Technical Dimensions of Social Media

The technical dimension refers to the hardware, software, connectivity and devices that enable social media practices. From the current statistics, almost all the USA people, they have regular access to the Internet by the age of eleven and 50% online using a smartphone. Children are using social media significantly increasing as they get older. The majority of Bangladeshi children and young people regularly engage in a variety of activities in social media components like instant messaging, web blogs, online gaming etc. The growth of Smartphones and other Internet enabled devices shows that people use social media, are increasing likely day by day. Now-a-days, with the comparatively changes of social media, people are getting addicted many trends of social media. The technical dimension [5] of social media are collaborative projects e.g., (Wikipedia), Blogs and micro-blogs (Twitter), Content communities etc.

2.4.2 Social Media Levels

From a media perspective, researchers use social presence and media richness principles to classify the social media. According to social presence theory, different types of media have dissimilar degree of social presence. The social presence is defined by the intimacy of the medium, as for example interpersonal e.g., telephones conversation or a face to face conversation. These are two types of immediacy of medium. One is asynchronous and other is synchronous medium. The other dimensions of social media is social presence. Social media can be partitioned by social presence in two sections: High and Low levels of social media which is in Table 2.2.

Levels	Low	Medium	High
High	Blogs	Social networking sites (e.g., Facebook)	Virtual social worlds (e.g., Second Life)
Low	Collaborative projects (e.g., Wikipedia)	Content communities (e.g., YouTube)	Virtual game worlds (e.g., World of Warcraft)

Table 2.2: Social media levels defined by social presence [5]

2.4.3 Functional Blocks of Social Media

There exists seven functional blocks in social media. The seven functional blocks [15] of social media are discussed below:

Identity: It describes how the consumers reveal themselves on a social media platform.

Conversation: Conversation are the way of consumers communication, including motivations, frequency and content.

Sharing: Sharing reveals the rate of content exchange process between the different sectors.

Presence: It delineates the reachability of the users on the social media platforms.

Relationships: It ties between the participant's relationships

Reputation: Reputation is the measure of consumers' identifying themselves, mainly relating to others in the community.

Groups: Groups are the communities or sub-communities, which are the building elements of social media.

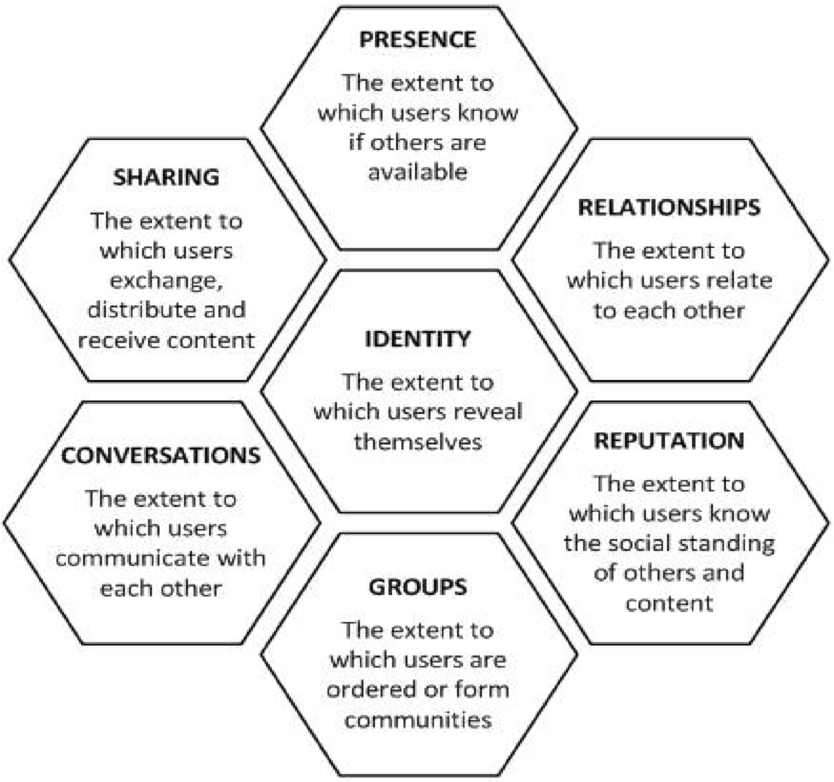

PRESENCE
The extent to which users know if others are available
SHARING
The extent to which users exchange, distribute and receive content
RELATIONSHIPS
The extent to which users relate to each other
IDENTITY
The extent to which users reveal themselves
CONVERSATIONS
The extent to which users communicate with each other
REPUTATION
The extent to which users know the social standing of others and content
GROUPS
The extent to which users are ordered or form communities

Figure 2.6: Seven functional of social media model [4]

2.4.4 Elements of Social Media

Social media has expanded over the recent years to include eight primary categories, each with its own unique set of characteristics.These are given in Table 2.3. [6].

Social Media Elements	**Example**
Collaborative Projects	Wikipedia
Blogs	Company Sponsored Blogs
	User Sponsored Blogs
Micro-Blogs	Twitter

Virtual Worlds	Virtual World Games e.g., COD
	Virtual Social Worlds e.g., Second life
Social News Website	Digg
	Reddit
Network Sites	Business Network Sites e.g., likedin
	Social Network Sites e.g., Facebook
Commerce Communities	Amazon
	ebay
Content Communities	G.i.p.s.s
	Music Sharing e.g., jaemendo
	Photo Sharing e.g., flickr
	Video Sharing e.g., youtube
	C.s.cw.a

Table 2.3: Example of social media by social presence [6]

2.5 Social Media Mining

Social Media Mining (SMM) is the process of representing, analyzing, and extracting interesting patterns from social media data. In simple terms, *SMM* is a systematic way of generating information from social media. SMM also denotes how we analyze the structure and content of the data to understand the behavior of network, online communication and concurrence of our social application. It introduces basic perception and primary algorithms applicable for inspecting huge social media data. It discusses theories and methodologies from different disciplines such as computer science, data mining, machine learning, social network analysis etc.

It encloses the tools to formally represent, measure, model, and mine resourceful patterns from large-scale huge social media data. The data Scientist who are well proficient in social and divination theories, trained to analyze calculation social media data, and skilled to help connecting the bridge into the gap from what we know about the vast social media world with computational tools. Since social media can also be parsed as a form of collective wisdom, we determined to interrogate its

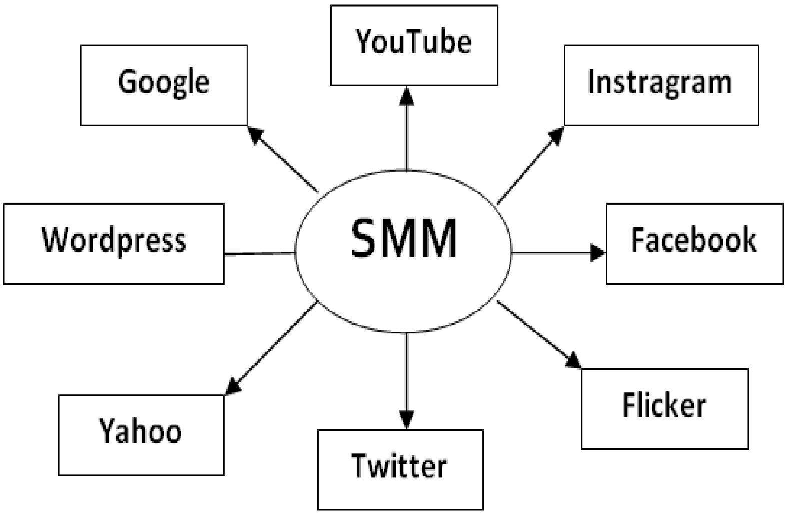


Figure 2.7: Dimensions of social media mining

capacity at predicting real-world outcomes. We discovered that the collection of our real world community can be efficient to mine our media.

2.5.1 Challenges in Social Media Mining

Mining social media data is the task of mining user-generated content with social relations. Data Scientists presents novel challenges detection in social media mining:

Big data Paradox: When all data is is packed up and organized in same place then it is far to easier to analyze the data and take better explanation. But Social media data is surprisingly bigger. We often take some little data to experiment our data. In some cases we have to exploit our data and use its sort of multidimensional data for prediction of mining media with our sufficient statistics tools. Let's make a few assumptions: (1) Data/Knowledge/Information is changing over the time and huge size of data takes more computation that the physical data. (2) A single data point is less effective than the combination of related data sets. Can't be drawn the data as a two dimensional graph with vertices and edges

Obtaining Sufficient Samples: Sufficient samples means collection of sample data to represent our full data. By using of sufficient samples, we can indicate a true or resourceful pattern among the sample data so that it can be understandable the current statistics of data. Collection of data can sampled via some methods or application programming interfaces (API) from social media sites.

Noise Removal Fallacy: Noise removal fallacy means that data often removes its valuable information during the extensive processing of data. That means ‘garbage in garage out’. By nature social media contains a large portion of disturbance data. That steps not only make the actual data application run faster on the cleaned data but also improve overall performance.

Evaluation Dilemma: For testing our data set it is not a standard procedure to get a properly elucidate data sets because of the huge size of data sets. Since there is no truth ground data set in our training set, we can not justify our machine organized algorithm to define our accuracy.

2.5.2 The Generic Process of Social Media Mining

To overlook on data, any data mining procedure follows some activities. On based on blogs there are some theme to process [16] the SMM:

1. Getting authentication from the website
2. Data visualization
3. Cleaning and processing
4. Data modeling using the suitable algorithms such as opinion mining, culturing, spam detection, correlations and segmentation, recommendations.
5. Result visualizations.

2.6 Probabilistic Distribution Model

A probabilistic model describes a set of possible probability distributions for a set of observed data. Probabilistic distribution model (PDM) are extensively used in *Text Mining* and application range from topic modeling, language modeling, document classification and clustering to information extraction. PDM is a system that produces different outcomes with different probabilities. It can simulate a class of objects (events), assign each an associated probability. For predicting problem simulation on some methods, the text and documents can be transformed into measured values. The foundation of most inferential statistical analysis is the concept of a probability distribution. An understanding of probability distributions is critical to use such quantitative methods as hypothesis testing, regression, and time-series analysis. All probability distributions can be classified as **discrete probability distributions** or as **continuous probability distributions**, depending on whether they define probabilities associated with discrete variables or continuous variables. The major probabilistic distribution model [17] are mentioned in below:

Bayesian Nonparametric Models: For computing infinite-dimensional parameters this model can generate the probabilistic models which usually have a stochastic process. It helps to detect number of topics in topic modeling like trends, Hashtags etc.

Hidden Markov Model: A simple case of dynamic Bayesian network in which the invincible states creates a chain and only some possible values for each state can be observed.This models have been applied to a wide variety of problems in information extraction and natural language processing.

Markov Random Fields: The joint density of all random variable network can be defined on cliques by this model.Recently, this model has been widely used in many text mining tasks, such as text categorization and information retrieval.

2.6.1 Related Work & Definition

Definition 1 (Probability). The ratio of the number of favorable outcomes is divided by the number of possible outcomes. Suppose a set of word list is encountered in a data set D, x is a term, $x \in T$, the probability of x occurring in data set,

$$P(x) = term_x / \mid TF \mid \tag{2.1}$$

where $term_x = term_x$ frequency in data set D and $TF =$ total term frequency in data set D Example 1: Consider the given data table 2.4, the probability of "Goofing" word is $P(Goofing) = \dfrac{53}{1000} = 0.053$

Term	**Frequency**
Goofing	53
WillFerrell	12
Last	94
Night's	107
around	117
GoldenGlobes	263

Table 2.4: Dataset of term verses frequency

Definition 2 (Probabilistic Item). A probabilistic item x is an item that appears in an item, the probability of x is the probability of its event.

	Loc-1	**Loc-2**	**Total**
Term-1	255	20	275
Term-2	80	145	225
Total	335	165	500

Table 2.5: A sample space consists of 500 words are distributed according to their cities location and terms.

Definition 3 (Sample Space). A sample space, denoted by S, of an experiment is a set or collection of of all possible outcomes of the experiment such that any outcome of the experiment correspond to exactly one element in the set.

Example 3: Consider the data given in table 2.5, if an experiment consists of two terms T along with two city location L, then the sample space S holds.

$$S = \{T_1L_1, T_1L_2, T_2L_1, T_2L_2\}$$

Definition 4 (Joint Probability Distribution). Let A, B, C, D . . . are events, if they occur simultaneously and probability P, of the events are called joint probability. Thus all events of the form $A \cap B$, $A \cap B \cap C$, $A \cap B \cap C \cap D$ . . .

are joint events.

Example 4: Consider the data given in table 2.5, if an experiment consists of two terms T, along with two city location L, then $T_1 \cap L_1$ is a joint probability that chosen user word is T_1 whose location is L_1.

Definition 5 (Conditional Probability). Two events are A and B, where A event is known that some other event B has occurred is called a conditional probability and is denoted by $P(A \mid B)$. The symbol $P(A \mid B)$ is usually read as *'The probability of A given B'*, where the slash '|' stands for *'given that'.* With two events A and B, the most fundamental formula to compute conditional probability [18]

$$P(A \mid B) = P(A \cap B)/P(B) \qquad if \quad P(B) \neq 0 \tag{2.2}$$

Example 5: Consider the data given in table 2.5, if an experiment consists of two terms, T along with two city location, L then the probability of term-1 which is listed as loc-2 from equation 2.1

$$\begin{aligned} P(term_1 \mid loc_2) &= P(term_1 \cap loc_2) \; / \; P(loc_2) \\ &= \frac{\frac{20}{500}}{\frac{165}{500}} = 0.123 \end{aligned}$$

Definition 6 (Independence of Events). If A and B are two events and if the occurrence of A does not affect, and is not affected by the occurrence of B, then A and B are said to be Independent. In other words two events are said to be independent if and only if equation 2.3 holds.

$$P(A \cap B) = P(A) * P(B) \tag{2.3}$$

Property 1 (Axioms of Probability). Let S is a sample space associated with an experiment. To every event A, in S, where $A \subseteq S$, $P(A)$ called the probability of A, so that the axioms [19] are,

Axiom 1: $P(A) \geq 0$

Axiom 2: $P(S) = 1$

Axiom 3: If A_1, A_2, . . . form a sequence of pair-wise mutually exclusive events in S that is $(A_i \cap A_j) = \emptyset$, if $i \neq j$ then

$$P(A_1 \cup A_2 \cup A_3 \cup ...A_n) = \sum_{i=0}^{n} P(A_i) \tag{2.4}$$

Property 2 (Mutually Exclusive). If A, B, C is mutually exclusive, and then relation 2.5 holds.

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) \tag{2.5}$$

Property 3 (Boole's Inequality). For any event A, the inequality holds.

$$0 \leq p(A) \leq 1$$

$$\text{and if } A \subseteq B \text{, then } P(A) \leq P(B)$$

Property 4 (Complement Event). For any event A, the complement event of A holds in equation 2.6,

$$P(\overline{A}) = 1–P(A) \tag{2.6}$$

Property 5 (Independent Event). If A and B are two independent events, then

$$P(A \mid B) = P(A) \qquad if \quad P(B) > 0 \tag{2.7}$$

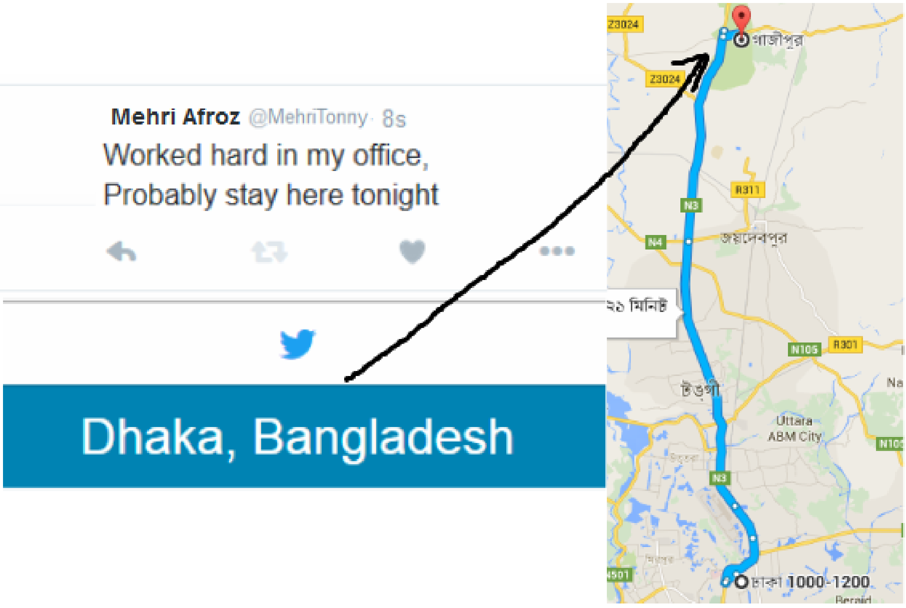
$$P(B \mid A) = P(B) \qquad if \quad P(A) > 0 \tag{2.8}$$

2.7 Location Mining

Mining Location is one of the most essential issues in text mining that need to be explored for social media mining. The increasing availability of large amounts text of data regarding to an individual's direction has given rise to a variety of geographic information of users. Location represents particular challenges within named entity identification. Extraction text from many social media has been a great important process to predict the user location. In web search area, location predicting from text is a value able contribution and useful resource for a multitude applications. Recently several studies have shown that about 20% [20] all of user queries express geographic location information from social media. There are easy ways to obtain user locations, for example, social media service providers allow users to provide their locations, mostly through GPS locating or by manual specification. However, only a small proportion of users actually provide location information. But in our paper we try to give an evolutionary process to give probabilistic perception of a user location from social media.

2.7.1 A Real Life Scenario of Location Mining

Location mining is needed to estimate of location of users. For example, Meheri Afroz is a user who gives other people fake location by tweeting a text that she is working office all day long. But actually she stays at office around 8-9 hours. Then he used to go her home at Gazipur. This tweet proves that she was working at office rather than hanging out all night at home. That arises the inconsistency of location. Figure 2.8 describes the exact scenario of our example.



Mehri Afroz @MehriTonny · 8s

Worked hard in my office,
Probably stay here tonight

Dhaka, Bangladesh

Figure 2.8: Fake location at dhaka

2.7.2 Techniques of Location Mining

Several algorithms of location mining [21] from text have been proposed. Some of those techniques are listed in Table 2.6.

Methods	**Techniques**
Organizing the structure content from text	Categorization
	Classification
	Taxonomy
Statistical Analysis	Term Frequency
	Keyword Frequency
	Distribution Term Matrix
	Term Frequency-Inverse Document Frequency(TF-IDF)
	Document Indexing
Machine Learning	Clustering
	Association Rules
	Classification
	Predictive Modeling
Classification Methods	Naive Bayes
	Support Vector Machine
	K-nearest Neighbor
Model Evaluation	Precision
	Recall
	Accuracy
	Relevance

Table 2.6: Techniques of location mining from text

2.8 Related Work

The geographic location estimation problem has been studied extensively by researchers who propose various ways to extract user location from internet social media platforms [22]. These social media platforms include web pages [8] and blogs [8] etc. Most of works rely on external information as IP, gazetteers, to identify related geographical location [23]. In our approach we do not use external information to a user's geographical location. The work by B. Huberman et al. [24] studies the variation of language usage in Twitter.

There have been works on: relations between geotags [25], geo-location estimations in search engine query logs [26], user privacy of geotags [27], predicting geographical location on proximity [28], and a study of private information trial [24] using correlations between different publicly available pieces of information to extract private information about a person. Z. Cheng et al. [29] studied the use of location sharing services by users of platforms like Foursquare. They use 'Check-In' information to study mobility characteristics of users. Another work by S. Abrol et al. [30] involves location prediction of Twitter users based on her social network. The authors mine implicit attributes associated with the user in her social network, and predict the user location based on these attributes. These works can be used to augment our work in estimating geographic location of a user. Estimating Twitter user location using social interactions – A content based approach by S. Chandra et al. [23] is the most relevant to our approach. They used only tweet and reply tweet to estimate geographic location of a user. They calculated the probability of estimated location based on the terms used in the tweet and reply tweet of a user. The main difference of our approach with their approach is that additionally we are using an important parameter to calculate the probability of a user's location. Usually a large portion of followers of a user live in the same city where she lives. We are using this follower information of a user where they live to estimate her geographic location and we are also using terms used in tweets and reply tweets, combining both probability we are proposing a new probability of a user's geographic location.

2.9 Summary

In this chapter, we discussed data mining, text mining, social media mining for location mining. Then we got familiar with some concepts of probability. We mentioned some techniques of location mining and significance of location mining. We discussed related works and the differences between our approach and other related approaches at the end of related work section.

Chapter 3

Our Proposed Approach

Although RPDM [23] estimates location of social media users more efficiently than PDM [29], still it has some drawbacks. The drawbacks are listed below:

1. The main drawback of RPDM is that it does not consider any other necessary information of user to estimate her location except contents of tweets, which leads to inefficiency.

2. The accuracy of RPDM in location estimation is really low.

These drawbacks of RPDM motivated us to overcome the drawbacks. So we come up with an improved version of the RPDM approach.

3.1 Problem Formulation

We propose a new approach to estimate geographic location of microblogging site users. We are adding a significant parameter, location of friends of a user, with the content based approach, RPDM. In our approach, we estimate city level location of a microblogging site user with higher accuracy than RPDM. To estimate location of a user, we use contents of her microblogs along with her location of followers. Combining both information, we provide an efficient approach.

3.2 Steps of our Approach

The approach for estimating geo-location can be divided into three phases. In each phase we calculate the probability of location of a user belonging to a location and evaluate the final probability in last phase. Three phases are briefly discussed in section 3.2.1, 3.2.2 and 3.2.3.

3.2.1 Finding Probability of Location Based on Contents of Microblog

Estimating geo-location of social medium user, the content based approach by Chandra et al. [23], works on Twitter, a popular one. We followed their approach for calculating probability of users location based on contents of microblog across a city. First, we construct a probability distribution of terms which contains probability of a term being used within city C. This is conducted by the following algorithm 1(Reply Based Probability Distribution Model) from [23].

The input for Algorithm 1 [23] is a list of cities, a list of users, and a list of tweet messages considered in the training data set. Line 1 in the algorithm considers each tweet in the training set. The tweet contains a set of words. These words are normalized in line 2. In lines 3 to line 7, we form a posting list (containing the term frequency and total frequency) to determine the statistics required to calculate the probability in line 10. Here, we check for reply messages, which are tweet messages that begin with a symbol @, followed by a user's screen name. In line 2 of the Algorithm, the tweet considered is that of userA, for instance. We check whether the tweet is a reply-tweet in line 3. If it is, we update the posting list structure instance for userB, who is tagged in the reply-tweet; else, we update the posting list structure instance for userA, who posted the tweet message. This is explained in detail in the next sub-section. Line 8 builds a dictionary that is later used to evaluate the baseline probability estimate, which is explained in the next paragraph. After evaluating all the tweet messages considered in the training set, in line 10 we obtain the distribution matrix (Replydistribution), which is of size c * w, were c is the size of the CityList and w is the size of the Dictionary.

Algorithm 1 Reply Based Probability Distribution Model [23]

Input

Citylist: List of Cities occurring in Data Set.

Tweets: List of Tweet Status considered in the Training Set.

Users: List of users considered in the Training Set.

Output

Probability Distribution of Terms across *cities* ε *citylist*

```
for each $tweet \in$ Tweets do
    $Terms$ = Normalize Words in tweet form
    $userA \in$ Users
    if $tweet$ begin with @UserB then
        PostingList(Terms,userB)
    else
        PostingList(Terms,userA)
    Update Dictionary with $Terms \in Dictionary$
ReplyDistribution = EvalDistribution(PostingList, CityList, Dictionary, Users)
Return $ReplyDistribution$
```

The Evaldistribution function used to evaluate the distribution matrix uses the equation shown in the PDM sub-section to evaluate the probability of a term t given a city *c*. This subroutine is shown in algorithm 2. The input for Algorithm 2 [23] is: the PostingList, which is formed in Algorithm 1 and contains (1) the list of users for each term Dictionary, where each term is used by the user in her tweet, and (2) the corresponding term frequency (Number of occurrences of the term for each user) and (3) the total frequency for the term (Number of occurrences of the term in the total data set); the CityList, which is the list of all cities occurring the training data set; the Dictionary, which is a list of all distinct terms appearing in the tweet messages in the training set; and Users, which is the list of users in the training set. Lines 1 and 2 in the algorithm show the distribution evaluation for each term in the dictionary, for each city. Line 3 initializes the distribution value. In line 4, we consider all the users whose location is the city under consideration. The sum, for all users in the city, of the relevant term frequencies is obtained in line 5. In line 7, the probability for a term given a city, is calculated by dividing the city-wide

Algorithm 2 EvalDistribution Function [23]

Input

PostingList: List of users with term frequency, for each term in the dictionary, and total frequency.

CityList: List of cities occurring in the Data Set.

Dictionary: List of distinct terms obtained from the Data Set.

Users: List of users considered in the Training Set.

Output

Probability Distribution of Terms across *cities CityList*

```
for each $term \in$ Dictionary do
    for each $term \in$ Dictionary do
        $Distribution$[term][city] $= 0$;
        for each $user \in$ Users located in city do
            $Distribution$[term][city] $\mathrel{+}= PostingList.term.user.termFrequency$;
        $Distribution$[term][city] $/= PostingList.term.totalFrequency$;
Return $Distribution$
```

sum of term frequencies obtained in line 5, by the total frequency of the term in the entire collection. Finally, the whole probability distribution is returned in line 10.

The input of algorithm 3 is follower list of a user with their location and city list. Line 1 to 3 calculates frequency of every city of a user's follower. Line 4 evaluates probability of every city where the user may live. This is calculated dividing frequency of a city by the number of total city in city list. The algorithm returns probability of a user location across all cities. This probability distribution contains the probability of use of different terms across different cities. Now, to estimate location of a microblogging site user (e.g. Twitter) across a city, probability of all the terms used in that user's microblogs (e.g. tweet and reply tweet) is calculated using the probability distribution constructed above as follows: The probability P_1 of a user u located in a city c, we calculate the total probability of the terms used in his microblog is:

Algorithm 3 Lcoation Estimation Function

Input

FollowerList: List of followers with their location.

CityList: List of cities.

Output

Probability of a user location across different $cities \in CityList$

```
1: for each $follower \in$ FollowerList do
2:     if $followercity \in cityList$ then
3:         $cityFrequecny_i{+}{+}$
4: for each $city_i \in$ cityList do
5:     $EstimatedLocation$[city][userA] = $cityFrequecny_i$ / | $cityList$ |
6: Return $EstimatedLocation$
```

$$P_1(c \mid u) = \sum (W_{terms}) p_i(c \mid w) * p_i(w) \tag{3.1}$$

where p_i(w)= | term w occurs in a microblog | / | number of terms in microblog |

3.2.2 Finding Probability of User's Location Using Information of Followers

Now we calculate the probability P_2 of a user located in a city using locations of friends of that user. The probability of a user u located in a city c is number of friends in city C divided by total number of distinct cities where her friends live.

$$P_2(u_{city_i}) = \mid c \mid / \mid FC \mid \tag{3.2}$$

where $P_2(u_{city_i})$ = probability that user u belongs to $city_i$, $\mid c \mid$ = number of follower in city C and $\mid FC \mid$ = number of follower in distinct city.

3.2.3 Evaluating Final Probability of Location

We find the final probability, combining this both probabilities, where she lives. In the combination of two probabilities, the ration of P_1 and P_2 is α : (100 - α).

$$P(city_c \mid user) = (\alpha) * P_1(city_c) + (1 - \alpha) * P_2(city_c) \tag{3.3}$$

We use different value of α in our experiment to see which produces the best result.

3.3 Example & Implementation

In our experiment, we consider the text from Twitter, one kind of social media.

3.3.1 Scenario

Figure 3.1 describes that getting text of tweet from twitter(a microblog) alongside with their reply comment status.

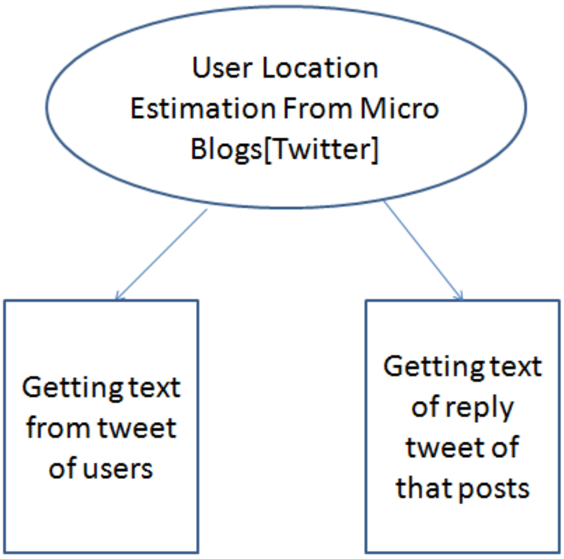
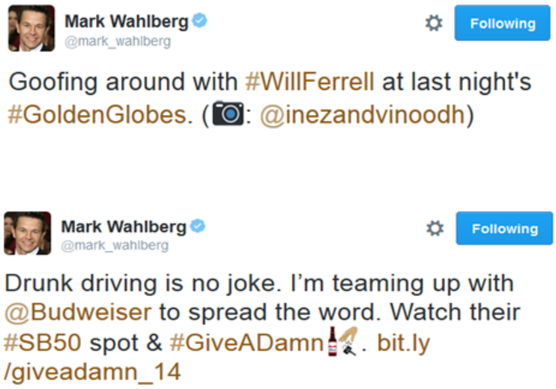

User Location Estimation From Micro Blogs[Twitter]
Getting text from tweet of users
Getting text of reply tweet of that posts

Figure 3.1: Problem specification with Twitter

3.3.2 Tweet Status

For simulating our process we consider the tweet status of ***@mark_whalberg*** in Figure 3.2



Mark Wahlberg
@mark_wahlberg

Goofing around with #WillFerrell at last night's #GoldenGlobes. (📷: @inezandvinoodh)

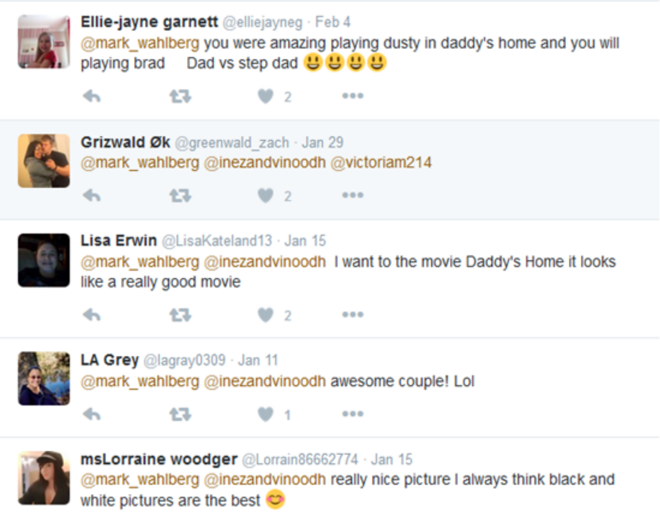
Mark Wahlberg
@mark_wahlberg

Drunk driving is no joke. I'm teaming up with @Budweiser to spread the word. Watch their #SB50 spot & #GiveADamn🍺🔑. bit.ly/giveadamn_14

Figure 3.2: Example of Tweet Status of ***@mark_whalberg*** after crawling the Twitter API

3.3.3 Reply Tweet Comment

Then getting reply status comment of ***@mark_whalberg*** in figure 3.3. We are getting those reply comment terms from the tweet of ***@mark_whalberg*** so that we could make e reply probability distribution model.



Ellie-jayne garnett @elliejayneg · Feb 4
@mark_wahlberg you were amazing playing dusty in daddy's home and you will playing brad   Dad vs step dad 😃😃😃😃

Grizwald Øk @greenwald_zach · Jan 29
@mark_wahlberg @inezandvinoodh @victoriam214

Lisa Erwin @LisaKateland13 · Jan 15
@mark_wahlberg @inezandvinoodh I want to the movie Daddy's Home it looks like a really good movie

LA Grey @lagray0309 · Jan 11
@mark_wahlberg @inezandvinoodh awesome couple! Lol

msLorraine woodger @Lorrain86662774 · Jan 15
@mark_wahlberg @inezandvinoodh really nice picture I always think black and white pictures are the best 😊

Figure 3.3: Example of ***@mark_whalberg*** reply status comment crawling with the API

3.3.4 Calculation of Terms Frequency

Before generating the Probability Distribution Model(PDM) & Reply Probability Distribution Matrix(RPDM) we need to calculate the tweet word term frequencies. Calculating the total status of each word frequencies table from all of the cities which is taken from tweet with the help of global dataset frequencies.
Consider the first status tweet of @mark_wahlberg, we create the term frequency Table in 3.1

Term	Frequency
Goofing	53
WillFerrell	12
Last	94
Night's	107
around	117
GoldenGlobes	263

Table 3.1: Calculating frequencies of terms of tweet of ***@mark_whalberg***

3.3.5 PDM & RPDM

We calculate a probability distribution model. Assuming that each user lives in a particular city, we assign the term of his tweets to that city. The probability distribution of term t over the full dataset, foe each city c, is calculated as

$$P(city_i \mid term) =\mid c_{term} \mid / \mid t_{term} \mid \tag{3.4}$$

where, c_{term} = terms occurred in city and $\mid t_{term} \mid$ = total term count in whole data set

The model will be formulated in Table 3.2

	City-1	**City-2**	**City-3**	. . . **City-n**
Term-1	P($c_1 \mid t_1$)	P($c_2 \mid t_1$)	P($c_3 \mid t_1$)	. . . P($c_n \mid t_1$)
Term-2	P($c_1 \mid t_2$)	P($c_2 \mid t_2$)	P($c_3 \mid t_2$)	. . . P($c_n \mid t_2$)
Term-3	P($c_1 \mid t_3$)	P($c_2 \mid t_3$)	P($c_3 \mid t_3$)	. . . P($c_n \mid t_3$)
Term-4	P($c_1 \mid t_4$)	P($c_2 \mid t_4$)	P($c_3 \mid t_4$)	. . . P($c_n \mid t_4$)
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Term-n	($c_1 \mid t_n$)	P($c_2 \mid t_n$)	P($c_3 \mid t_n$)	. . . P($c_n \mid t_n$)

Table 3.2: Calculation framewrok of PDM & RPDM city Location matrixes

3.3.6 Example of PDM

Calculating the status word frequency table across the cities which is taken from tweet with the help of global dataset frequencies.

3.4 Example of RPDM

Calculating the tweet of reply comment term frequency Table in 3.4, 3.5, 3.6 for each of users from all of the cities which is taken from tweet reply post comment

	Seattle	**Brooklyn**	**LA**	**LakeWood**	**LasVegas**	**Miami**
Goofing	12	19	11	5	4	2
with	11	12	13	14	15	12
last	10	14	12	18	21	19
Night's	13	16	11	19	23	25
around	23	15	14	22	24	19
GoldenGlobes	45	14	71	21	49	63

Table 3.3: Frequencies of terms alongside with cities of tweet

(RT) with the help of global dataset frequencies.

Term	**Frequency**
amazing	61
playing	117
dusty	43
daddy's	107
home	117
Dad	263
Step	263

Table 3.4: Calculating frequencies of reply comment terms of ***@elliejayneg*** of tweet of ***@mark_whalberg***

Summing these reply comment terms from Table 3.4, 3.5 & 3.6 across cities are stored in Table 3.7.

3.4.1 Calculation of Probability of a User's Location Using Tweet & Reply Tweet

So from the Table 3.7 calculate the probabilities of cities by given by term from equation 3.4 is

P(Seattle | Goofing) = 12/53 = 0.226415

Term	**Frequency**
movie	69
Daddy's	85
home	74
looks	94
really	107
like	54

Table 3.5: Calculating Frequencies of reply comment terms of ***@LisaKateLand13*** of tweet of ***@mark_whalberg***

Term	**Frequency**
really	107
nice	45
picture	142
always	71
black	25
white	59
best	16

Table 3.6: Calculating frequencies of reply comment terms of ***@Lorrain8662774*** of tweet of ***@mark_whalberg***

P(Seattle | GoldenGlobes) = 45/263 = 0.171102

P(Brooklyn | Goofing) = 19/53 = 0.358490

P(Brooklyn | GoldenGlobes) = 14/263 = 0.05323

P(LA | Goofing) = 11/53 = 0.207547

P(LA | GoldenGlobes) = 71/263 = 0.2699619

P(Miami | Picture) = 15/142 = 0.105633

	Seattle	**Brooklyn**	**LA**	**LakeWood**	**LasVegas**	**Miami**
amazing	12	31	21	26	10	30
playing	16	19	13	26	18	15
dusty	5	9	11	10	5	3
Daddy's	15	6	18	21	12	13
Dad	18	15	26	26	13	10
picture	27	33	21	19	27	15
. . .	. . .	. . .	. . .	. . .	. . .	. . .

Table 3.7: Add all term frequencies alongside with cities

3.4.2 Probability Calculation of City given by a Specific User

The probability of city given by a specific user formula is stated in equation 3.5.

$$P(city_a \mid user) = \sum \{p(city_a \mid term) * p(term)\} \tag{3.5}$$

Where P(term) = | term t occurs in tweet | / | total no. of terms in that tweet |, Similarly we calculate the every city as follows from equation 3.5 P($city_b$ | user), P($city_c$ | user), P($city_d$ | user) up to P($city_n$ | user). We will also calculate this for reply tweets using this same approach.

3.4.2.1 Calculation of Probability of Terms in Tweet

The probability of terms in tweet formula is stated in equation 3.6.

$$P(term) = \mid t \mid / \mid T \mid \tag{3.6}$$

Where | t | = term t occurs in tweet and | T | = total no. of terms in that tweet From Figure 3.2 the tweet status of ***@mark_whalberg***, total no. of terms in tweet = 7, we crate a Table 3.8 by using the equation 3.6 that is the probability of terms in a tweet.

3.4.2.2 Example of Calculation of City given by a Specific User

From formula 3.1 we calculate the every city by given a user ***@mark_whalberg*** is

Term	Trems in Tweet
Googfing	1/7 = 0.14285
with	1/7 = 0.14285
WillFerrel	1/7 = 0.14285
last	1/7 = 0.14285
Night's	1/7 = 0.14285
around	1/7 = 0.14285

Table 3.8: Calculate frequency of the terms in tweet and probability of the terms in tweet of ***@mark_whalberg***

P(Seatle | ***@mark_whalberg***) = P(Seatle | Goofing) *p[Goofing] + P(Seatle | with) * p[with] + P(Seatle | WillFerrell) * p[WillFerrell] +. . . + P(Seatle | picture) * p[picture] = 0.10842

Similarly for every cities in our data set we calculate this probability by equation 3.1 is:

P(Miami | ***@mark_whalberg***) = P(Miami | Goofing) *p[Goofing] + P(Miami | with) * p[with] + P(Miami | WillFerrell) * p[WillFerrell] + . . . + P(Miami | picture) * p[picture] = 0.51753

3.5 Our Basic Idea for Improving Estimation Location

Calculating the probability of location of a user, we use the contents of tweet, reply tweets and follower's location information, here figure 3.4 describe our approach scenario.

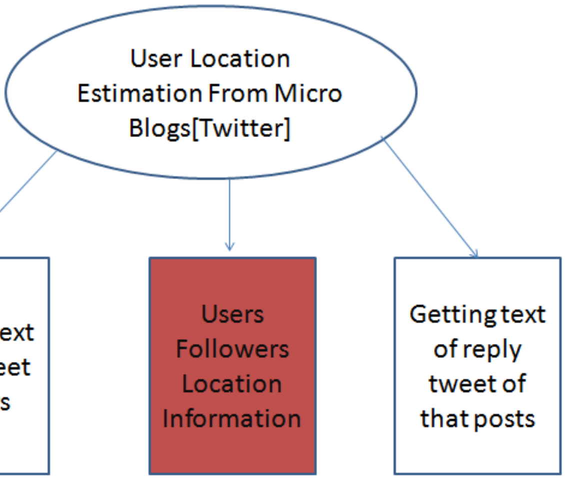

User Location Estimation From Micro Blogs[Twitter]
Getting text from tweet of users
Users Followers Location Information
Getting text of reply tweet of that posts

Figure 3.4: Add Follower Location Information

3.5.1 Calculation of Probability of Location Using Information of Followers

Th formula of probability of location of a specific user, using follower information of followers is calculated from equation 3.2. Table 3.9 indicates the every follower of ***@mark_whalberg*** has with their city location. Then from table 3.9 counting the frequencies of every cities which is stored in table 3.10.

Follower Name	**Location/City**
@Jason	NewYork
@Mathews	LasVegas
@Jhon	LosAngles
@Robert	Brooklyn
@Thomas	NewYork
. . .	. . .

Table 3.9: Example of user follower location information of ***@mark_whalberg***

Location Name	**Frequcncy of Follower in City**
NewYork	5
Seatle	5
LosAngles	16
Brooklyn	4
LakeWood	11
LasVegas	6
Miami	12
. . .	. . .

Table 3.10: Counting the frequency of City of Followers of Location

3.5.2 Example Calculation of City given by a Specific User with Follower Location Information

As we assume that number of total city in citylist is 30 so, from the equation 3,7 we calculate the probability of every city by given user ***@mark_whalberg*** is: P(Seatle | ***@mark_whalberg***) = 5/30 = 0.167, P(LosAngels | ***@mark_whalberg***) = 16/30 = 0.53, P(LakeWood | ***@mark_whalberg***) = 11/30 = 0.37, P(Texas | ***@mark_whalberg***) = 3/30 = 0.1

3.5.3 Calculation of Combined Probability of a User's Location

We took alpha weight from P_1(PDM + RPDM) and (1 - α) weight from P_2(Follower location List) in the combination from equation 3.3 when α = 60%, which is shown in Table 3.11. in decreasing order. We use top k cities to check accuracy of our estmation of location. For example, Here k=3 means top three cities will be considered define accuracy in Table 3.11.

3.6 Data Set

We collected publicly available tweets for constructing our probabilistic model. As a privacy/protocol issue of twitter, Twitter users can make their profile either public or

Location Name	Probability
Miami	0.5175
Seatle	0.4384
LosAngles	0.4123
Brooklyn	0.3987
LakeWood	0.3245
LasVegas	0.2765
Miami	0.2345
Texas	0.1141
. . .	. . .

Table 3.11: Calculate K top estimated city probability

private. All Tweets which is sent by a public profile users that are publicly available for anyone to view, even without having an account. So these public tweets are also combined into a collection of a tweet stream called the public timeline. We could not sample tweet from a private user account because only those tweets can be viewed by other users that they have given permission to follow them.

3.6.1 Data Collection

For collecting the Twitter stream we used code provided by the SNOW challenge organizers [31] based on the Application Programming Interface (API) Twitter 4j. We monitored over 2,000 followers for collecting our training data sets of 20,000 tweets. We considered the test set which contains 50 users of tweet along with their replies/comments and users follower city locations.We describe our data collection methodology by introducing social network enabled API named as Twitter 4j which is an unofficial Java library for the Twitter API. This API has a framework architecture in figure 3.5 [32], which can provide the description of our data sets, exploring strategies for data cleaning, applying filtering techniques from user profile in order to perform for retrieving data sets [33]. The rest stream was collected starting from tweet user name, and had each tweet in the form of a text line, containing of that tweet replies and user follower city info. This API can receive 30 requests of tweet per minute.

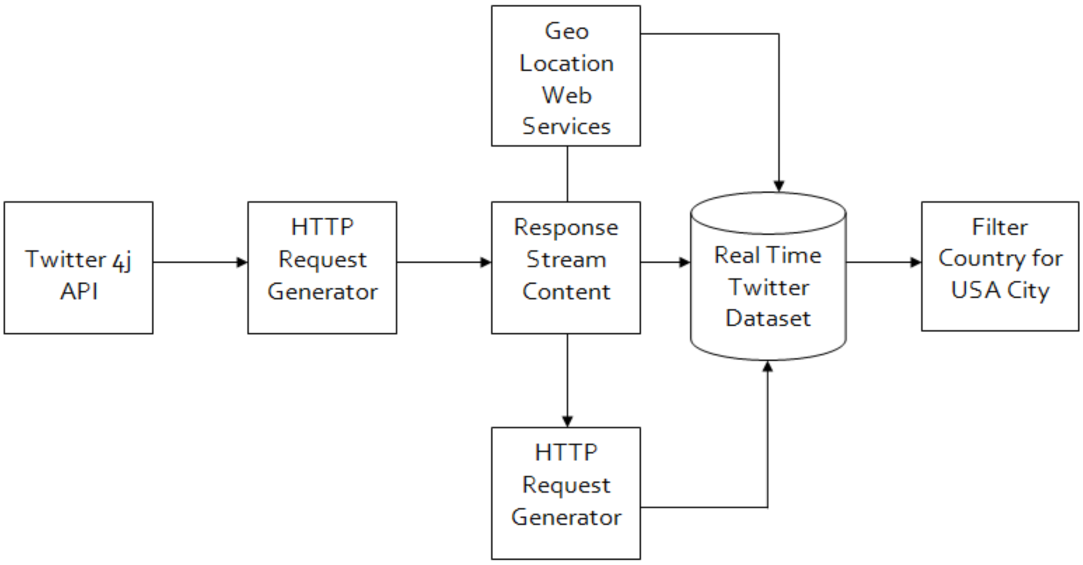


Figure 3.5: Twitter Crawler Framework of the Twitter 4j API

3.7 Summary

In this chapter, we discussed about our proposed approach elaborately. We explained what we are adding with the existing RPDM approach and why it will perform better. We explained the algorithms of our approach. Then we showed a scenario and explained how the processing will be conducted. In the next chapter, we will show our experimental results.

Chapter 4

Experimental Result

4.1 Evaluation of Results

In this section, we will present the experimental results of our proposed approach. Our approach was tested for effectiveness and efficiency. Also we have compared the results with the existing RPDM approach [23].

We will show the experimental results on various scales and a performance comparison between the RPDM algorithm and our proposed algorithm.

We calculate the probability of location of a user combining two probabilities, α % from RPDM and (100 - α) % from the calculated probability based on locations of followers. We have conducted our experiment using different values of alpha. With $\alpha = 60\%$ we get the best result.

We found that about 28% of the 50 users in the test set were assigned an estimated location within 100 miles of their actual location, when using the RPDM method [23]. with the term distribution estimator. However, by using our method with the term distribution estimator and follower's locations, we found that about 36% of the same 50 users were assigned a location within 100 miles to their actual location with $\alpha = 60\%$

4.2 Performance for Different Values of Constant α

The results given above are with the restriction that the accuracy was calculated for the top location (K=1) that was predicted by the estimator. We conduct our experiment changing the value of K. We worked on chemical data set collected from our training data set. The data set was fed to the algorithms sequentially to make them behave as evolving data set. We also applied a brute force method for finding the actual result. By plotting our experimental value we give a overview of our estimating location percentage.

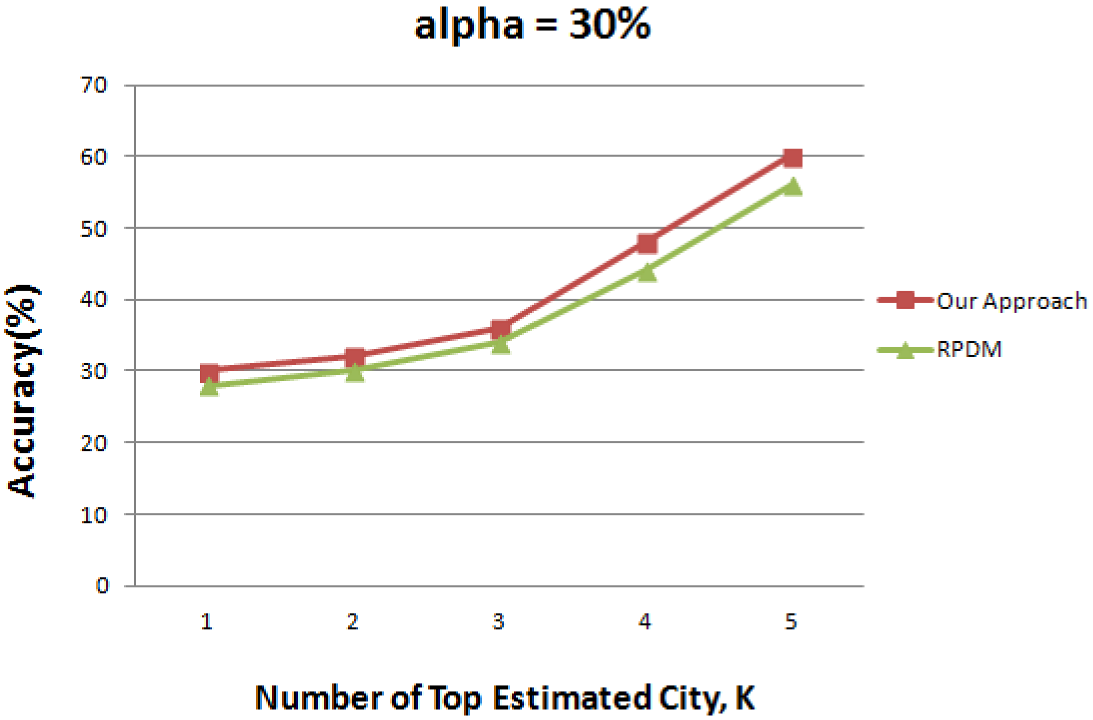


Figure 4.1: Comparison of our approach and RPDM when $\alpha = 30\%$

Figure 4.1 shows an experimental result. Here, we can observe that our approach estimates location more accurately than the state of art approach, RPDM. If we increase the value of K, number of top estimated city for calculating accuracy, overall accuracy increases.

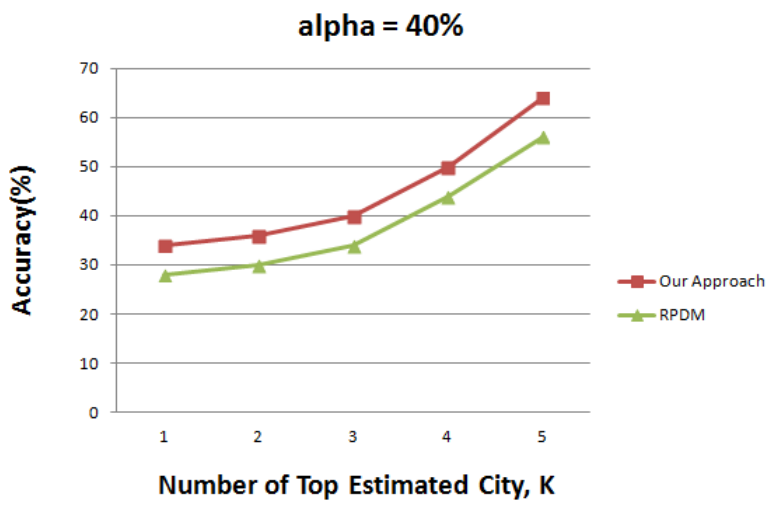


Figure 4.2: Comparison of our approach and RPDM when α = 40%

Figure 4.2 shows an experimental result. The difference between this figure and Figure 4.1 is that here we have increased value of alpha from 30% to 40%. In this experiment, accuracy of our approach is higher than RPDM and it is improved comparing to the experiment in Fig 4.1

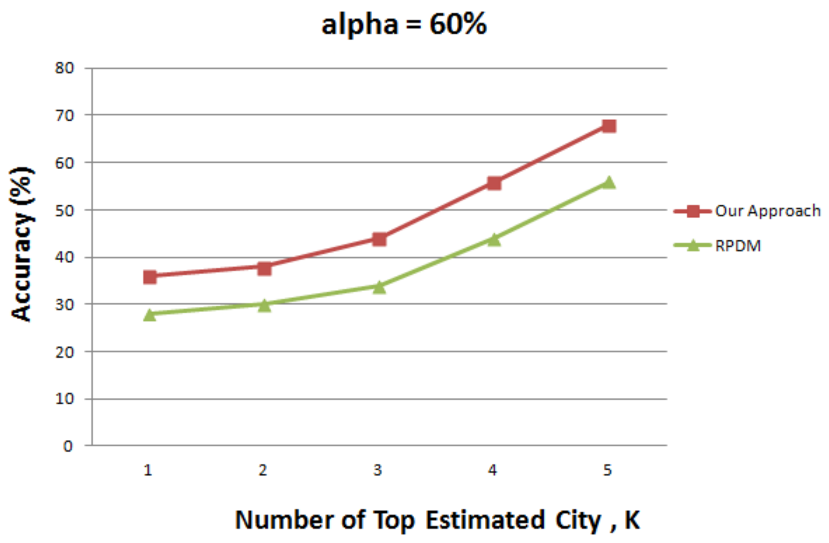


Figure 4.3: Comparison of our approach and RPDM when α = 60%

In Figure 4.3, the value of alpha is increased to 60% and this time, accuracy is still improving with increasing value of alpha.

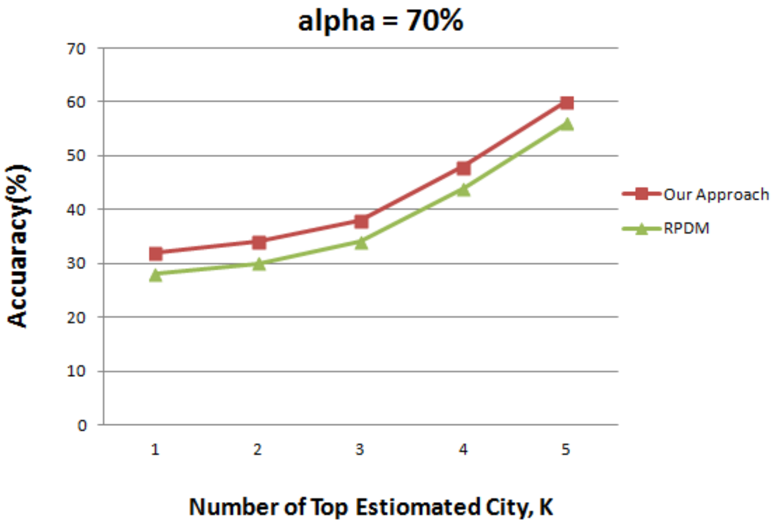


Figure 4.4: Comparison of our approach and RPDM when α = 70%

In Figure 4.4, we have increased the value of alpha for the experiment to 70%, higher than the previous experiment. But this time, accuracy is decreased than previous one.

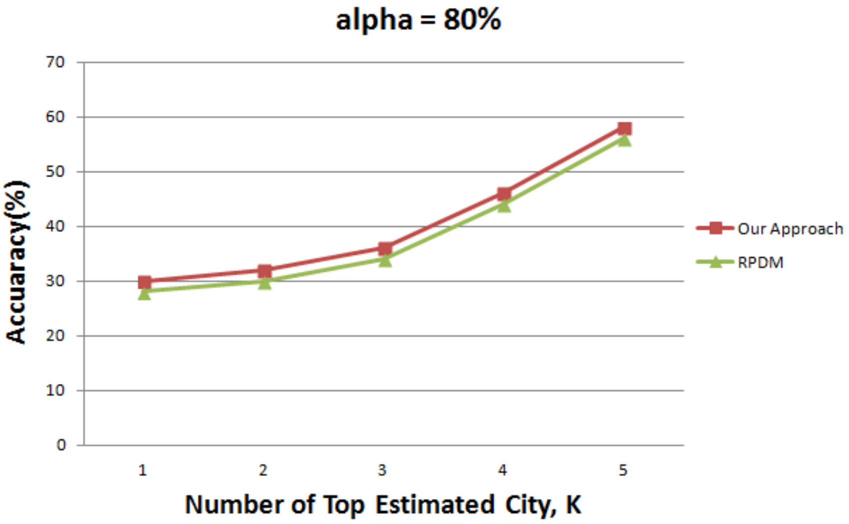


Figure 4.5: Comparison of our approach and RPDM when α = 80%

In Figure 4.5, for the experiment we have increased the value of alpha to 80%. Here, accuracy has decreased more than previous one.

If analyze the results, we can observe that with the increment of value of alpha up to a certain point, performance is increasing. After that value, performance is decreasing again. Here, we have best accuracy for alpha=60% for our dataset.

4.3 Summary

To summarize, our approach estimates location efficiently and effectively. Our approach performs better than the state of art RPDM approach. Taking different value of alpha, we showed how the performance reacts with respect to different values of alpha.

Though the accuracy changes for different values of alpha, but most of the times, accuracy of our approach is higher than RPDM.

Chapter 5

Conclusion

In this research work, we have developed an idea to improve the existing RPDM algorithm. It didn't consider locations of followers to estimate a user's location. As a result, RPDM estimate location with low accuracy. To improve this approach, we proposed a new idea over RPDM.

5.1 Research Summary

Our proposed approach is a modification of the content base approach, RPDM. In our approach, we use a vital factor of user, location of followes, for estimating her geo-location. Our approach estimate location with higher accuracy than RPDM. Database size of our approach is linearly correlated with RPDM as we do not add any extra information except location of followers.

In this paper, we discussed our idea elaborately. Our idea incorporated information about location of followers along with the contents of microblogs of users in the processing of estimation.

We also presented some experimental results to prove our method's efficiency. A comparison between our algorithm and RPDM was illustrated in this work. In most of the cases, our approach works with higher accuracy than RPDM approach.

5.2 Future Work

Researches on estimating location of microblogging site users, including our proposed approach, still needed to be improved a lot as degree of error is comparatively high. Though there are so many complexities to find a user's location with higher accuracy without any external information but still there are some opportunities to improve accuracy. We have detected some cases for which the overall accuracy is reducing and we are working on some propositions to solve those drawbacks.

We have not proposed any model to handle multiple languages which has a bad impact on accuracy. Some existing approaches can be incorporated with our approach to improve accuracy.

We have a plan to find a user's nearby followers. With this information, we can estimate her location more accurately.

Some common words, which are related to most of the cities and are not related to any specific city (e.g., 'is', 'how'), increase database size vastly. We have a plan to identify those words and exclude them from processing, which will reduce processing time and database size.

Appendix A

Tables for Experimental Results in Chapter 4

K	**Our Approach**	**RPDM**
1	30	28
2	32	30
3	36	34
4	48	44
5	60	56

Table A.1: Table for figure 4.1

K	**Our Approach**	**RPDM**
1	34	28
2	36	30
3	40	34
4	50	44
5	64	56

Table A.2: Table for figure 4.2

K	Our Approach	RPDM
1	36	28
2	38	30
3	46	34
4	54	44
5	68	56

Table A.3: Table for figure 4.3

K	Our Approach	RPDM
1	32	28
2	34	30
3	38	34
4	48	44
5	60	56

Table A.4: Table for figure 4.4

K	Our Approach	RPDM
1	30	28
2	32	30
3	36	34
4	44	44
5	58	56

Table A.5: Table for figure 4.5

Bibliography

[1] F. Stahl and I. Jordanov, "An overview of the use of neural networks for data mining tasks," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 193–208, 2012.

[2] A.-H. Tan *et al.*, "Text mining: The state of the art and the challenges," in *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, vol. 8, pp. 65–70, 1999.

[3] R. Menon, L. H. Tong, and S. Sathiyakeerthi, "Analyzing textual databases using data mining to enable fast product development processes," *Reliability Engineering & System Safety*, vol. 88, no. 2, pp. 171–180, 2005.

[4] "Seven fuctional blocks of smm," `https://jonetw.wordpress.com/2013/03/17/how-social-media-has-enhance-communications/`.

[5] A. M. Kaplan and M. Haenlein, "Users of the world, unite! the challenges and opportunities of social media," *Business horizons*, vol. 53, no. 1, pp. 59–68, 2010.

[6] R. Hanna, A. Rohm, and V. L. Crittenden, "We're all connected: The power of the social media ecosystem," *Business horizons*, vol. 54, no. 3, pp. 265–273, 2011.

[7] "Wikipedia. social media mining," `https://en.wikipedia.org/wiki/Social_media_mining`.

[8] C. Fink, C. D. Piatko, J. Mayfield, T. Finin, and J. Martineau, "Geolocating blogs from their textual content.," in *AAAI Spring Symposium: Social Semantic Web: Where Web 2.0 Meets Web 3.0*, pp. 25–26, 2009.

[9] E. Amitay, N. Har'El, R. Sivan, and A. Soffer, "Web-a-where: geotagging web content," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 273–280, ACM, 2004.

[10] M. Bramer, M. Bramer, and M. Bramer, *Principles of data mining*, vol. 131. Springer, 2007.

[11] F. Sebastiani, "Machine learning in automated text categorization," *ACM computing surveys (CSUR)*, vol. 34, no. 1, pp. 1–47, 2002.

[12] M. Rahman, "Information retrieval with text mining for decision support system," *Bangladesh Journal of Scientific Research*, vol. 24, no. 2, pp. 117–126, 2012.

[13] R. Feldman and J. Sanger, *The text mining handbook: advanced approaches in analyzing unstructured data.* Cambridge University Press, 2007.

[14] M. Gould, *The social media gospel: Sharing the good news in new ways.* Liturgical press, 2015.

[15] J. H. Kietzmann, K. Hermkens, I. P. McCarthy, and B. S. Silvestre, "Social media? get serious! understanding the functional building blocks of social media," *Business horizons*, vol. 54, no. 3, pp. 241–251, 2011.

[16] V. G. Sharan Kumar Ravindran, "Mastering social media mining with r," *Business horizons*, vol. 54, no. 3, pp. 241–251, 2011.

[17] C. Z. Charu C. Aggarwal, "Mining text data," *Business horizons*, vol. 54, no. 3, pp. 260–263, 2011.

[18] "Joint probability distribution," `http://goo.gl/PNxEaJ`.

[19] "Axioms of proabability," `http://goo.gl/PNxEaJ`.

[20] H. L. Hartman *et al.*, "Sme mining engineering handbook, vol. 1, society for mining, metallurgy, and exploration," *Inc. Littleton, Colorado*, 1992.

[21] J. S.-F. Tan, E. H.-C. Lu, and V. S. Tseng, "Preference-oriented mining techniques for location-based store search," *Knowledge and information systems*, vol. 34, no. 1, pp. 147–169, 2013.

[22] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 759–768, ACM, 2010.

[23] S. Chandra, L. Khan, and F. B. Muhaya, "Estimating twitter user location using social interactions–a content based approach," in *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pp. 838–843, IEEE, 2011.

[24] R. Heatherly, M. Kantarcioglu, and B. Thuraisingham, "Social network classification incorporating link type values," in *Intelligence and Security Informatics, 2009. ISI'09. IEEE International Conference on*, pp. 19–24, IEEE, 2009.

[25] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing, "A latent variable model for geographic lexical variation," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 1277–1287, Association for Computational Linguistics, 2010.

[26] L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak, "Spatial variation in search engine queries," in *Proceedings of the 17th international conference on World Wide Web*, pp. 357–366, ACM, 2008.

[27] S. S. Lee, D. Won, and D. McLeod, "Tag-geotag correlation in social networks," in *Proceedings of the 2008 ACM workshop on Search in social media*, pp. 59–66, ACM, 2008.

[28] L. Backstrom, E. Sun, and C. Marlow, "Find me if you can: improving geographical prediction with social and spatial proximity," in *Proceedings of the 19th international conference on World wide web*, pp. 61–70, ACM, 2010.

[29] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui, "Exploring millions of footprints in location sharing services.," *ICWSM*, vol. 2011, pp. 81–88, 2011.

[30] S. Abrol and L. Khan, "Tweethood: Agglomerative clustering on fuzzy k-closest friends with variable depth for location mining," in *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pp. 153–160, IEEE, 2010.

[31] S. Papadopoulos, D. Corney, and L. M. Aiello, "Snow 2014 data challenge: Assessing the performance of news topic detection methods in social media.," in *SNOW-DC@ WWW*, pp. 1–8, 2014.

[32] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu, "Predicting flu trends using twitter data," in *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, pp. 702–707, IEEE, 2011.

[33] "Twitter4j java library," `http://twitter4j.org/en/index.html`.