HashTag Prediction

December, 2015

Submitted in partial fulfillment of the requirement for the degree of

Bachelor of Technology

in

Information Technology

By

Shruti Thakur(121408)

Under the supervision of

Mrs. Sanjana Singh

to

Department of Computer Science & Engineering

and Information Technology

Jaypee University of Information Technology

Waknaghat, Solan-173234

Himachal Pradesh

CERTIFICATE

This is to certify that the work presented in this report entitled “**Hashtag Prediction”** in the partial fulfillment for the award of degree of **Bachelor of Technology** in **Information Technology** from **Jaypee University of Information Technology, Waknaghat** is an authentic record of our own work carried out over a period from August 2015 to December 2015 under the supervision of **Mrs. Sanjana Singh,** Assistant Professor(Department of Computer Science and Engineering).

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Student’s Signatures :

Student’s Names : Shruti Thakur

Student’s Roll No.s’ : 121408

This is to certify that the above statement made by the candidates is true to the best of my knowledge.

Signature of Supervisor :
Name of Supervisor : Mrs. Sanjana Singh
Designation : Assistant Professor(Grade I)
Department Name : Computer Science and Engineering
Date :

ACKNOWLEDGEMENT

I would like to express our gratitude to all those who gave us the possibility to complete this project. I want to thank the Department of CSE & IT in JUIT for giving us the permission to commence this project in the first instance, to do the necessary research work.

I am deeply indebted to my project guide Mrs. Sanjana Singh, whose help, stimulating suggestions and encouragement helped us in all the time of research on this project. I felt motivated and encouraged every time we got her encouragement. For her coherent guidance throughout the tenure of the project, I feel fortunate to be taught by her.

I am also grateful to **CSE Project lab staff** for their practical help and guidance.

Shruti Thakur (121408)

Contents

List of Figures

Abstract

Social media has demonstrated quick growth, in both directions of becoming the most popular activities in internet and of attracting scientific researchers to get better insights into the understanding of the underlying sociology. Real time micro-blogging sites such as Twitter use tags as an alternative to traditional forms of navigation and hypertext browsing. The tag system of those micro-blogging sites has unique features in that they change so frequently that it is hard to identify the number of clusters and so effectively carry out classification when new tags can come out at any time. We basically take the example of twitter for the discussion purposes. Twitter is one popular web application nowadays. Twitter allows users to use "Hash tags" to classify their tweets. It is called a micro-blog because people can post short, quasi-public messages up to 140 characters in length. People create lists of others and are shown a list of all of the posts of those people. The substantive nature of the social tie on Twitter is attention-based . In addition to paying attention to one another by "following," Twitter users can address tweets to other users and can mention others obliquely in their tweets. Another common practice is "retweeting," or rebroadcasting someone else's message (with attribution) so as to direct attention toward that person's tweets. Twitter differs from other online social networking services in that ties are asymmetric. Consider friendship ties in LinkedIn, Facebook, or MySpace; in these services, when two people share a friendship tie, the tie is symmetrical; A being friends with B implies B is friends with A. This is not the case in Twitter; A can "follow" B, but B needs not follow A. People who are popular, such as basketball players or actors, can be followed by millions of people, but can barely pay attention to all of those who follow them. Hashtags (single tokens often composed of natural language n-grams or abbreviations, prefixed with the character '#') are ubiquitous on social networking services, particularly in short textual documents (a.k.a. posts). Authors use hashtags to diverse ends, many of which can be seen as labels for classical tasks: disambiguation (chips #futurism vs. chips #junkfood); identification of named entities (#sf49ers); sentiment (#dislike); and topic annotation (#yoga). The hash tag enables Twitter users to create searchable subject groups and so to be able to navigate the hypertext structures of the whole site. The power of the hash tag is that it creates very specific sets of content. If you want to know what other people think of the superbowl that just came on you can find it easier by searching for the hash tag than by searching for something similar in a normal search engine. Every day, many new hash tags are formed and this process can happen right before your eyes-heck. The frequent creation of new tags makes the prediction of tags challenging. *Hashtag prediction* is the task of mapping text to its accompanying hashtags. Hash tag prediction is different from normal texts classification. Here we don’t know how many clusters we need to find. In addition, the tag set changes so frequently that it is almost impossible to effectively carry out classification or clustering, since a new tag would force us to establish a new class and a new classification rule. Our intuition is: if we can measure the correlation between various tweets as the mathematical metric we can treat the collected tweets as points in a high dimensional space, and construct a network by the latent space model. We show that simple techniques are sufficient to extract key semantic content from tags and also filter out extraneous noise. We demonstrate the efficacy of this approach by comparing it with other classification functions and show that our model maintains a false positive rate lower than 15%.

Chapter 1: Introduction

1.1 Introduction

Social media has demonstrated quick growth, in both directions of becoming the most popular activities in internet and of attracting scientific researchers to get better insights into the understanding of the underlying sociology. Real time micro-blogging sites such as Twitter use tags as an alternative to traditional forms of navigation and hypertext browsing. The tag system of those micro-blogging sites has unique features in that they change so frequently that it is hard to identify the number of clusters and so effectively carry out classification when new tags can come out at any time.

Social tagging is a method for Internet users to organize, store, manage and search for resources online. Trant categorizes the existing works on social tagging into three broad topics:

(a) on the *folksonomy* that results from the collective wisdom of users of the social tagging system.

(b) on the *tagging behaviour* of users, such as the incentives and motivation for tagging.

(c) on the software aspects of the *social tagging systems*, for improving system performance and enhancing user satisfaction.

Over the past few years, social media services have become one of the most important communication channels for people. According to the statistic reported by the Pew Research Center’s Internet & American Life Project in Aug 5, 2013, about 72% of adult internet users are also members of at least one social networking site. Hence, microblogs have also been widely used as data sources for public opinion analyses, prediction, reputation management, and many other applications. In addition to the limited number of characters in the content, microblogs also contain a form of metadata tag (hashtag), which is a string of characters preceded by the symbol (#).

An estimate of the dense and wide use of hashtags can be taken from the figure on next page.

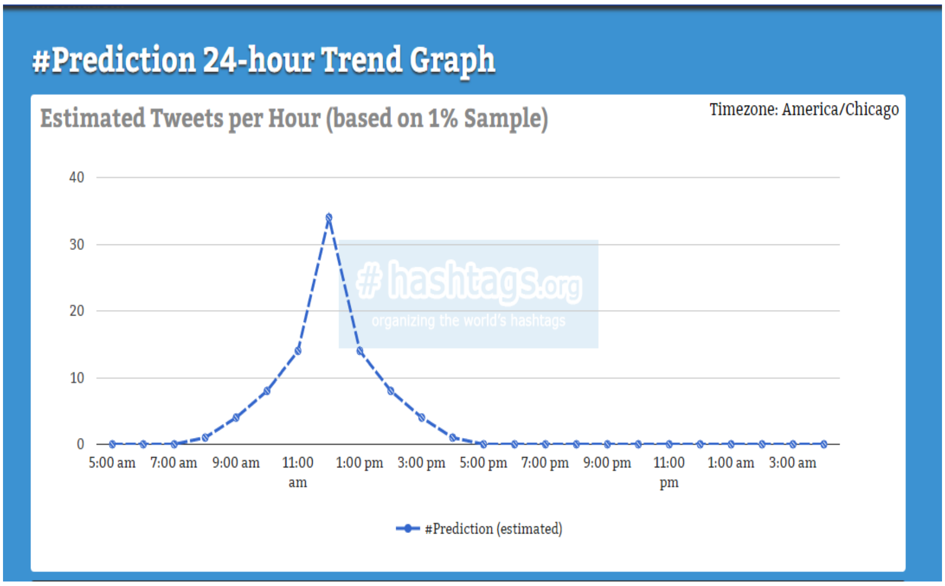


Figure 1.

Hashtags are used to mark the keywords or topics of a microblog. They can occur anywhere in a microblog, at the beginning, middle, or end. Hashtags have been proven to be useful for many applications, including microblog retrieval (Efron, 2010), query expansion (A.Bandyopadhyay et al., 2011), sentiment analysis (Davidov et al., 2010;Wang et al., 2011). However, only a few microblogs contain hashtags provided by their authors. Hence, the task of recommending hashtags for microblogs has become an important research topic and has received considerable attention in recent years. Various approaches have been proposed to study the problem from different aspects.

1.2 Problem Description

1.2.1. Problem Statement

To present an interface that would provide a platform to view the related posts (from twitter) and pictures (from instagram) on a particular topic together at one place.

1.2.2 Description

As mentioned above the task of hashtag prediction is not that simple. Task of predicting hashtags can be categorized in accordance with different aspects. Therefore, due to viral, dense and random nature of hashtags, the task of hashtag prediction can be divided into different sub-categories like **temporal** factor, **similarity** factor, **geographical** factor etc, to make the prediction work easier.

The main category that we have considered for our project purposes is the similarity factor between hashtags. One way to deal with this problem (i.e. predicting tags) is by classifying the tags according to the similarity between them. And one of the most efficient way to measure similarity is by using *Euclidean Distance* method.

Euclidean distance method can be proved highly effective for this purpose. This method by following the following process gives us a highly accurate rate of more than 86% :-

First we constructed and estimated the cosine matrix using the dataset. Then we applied the normalisation on the constructed data and then finally predicted the similar or related hashtags.

Apart from the high accuracy rate this method has some other advantages like its dynamic nature i.e. it is easy to update when the wordlist changes. Moreover, it is quite easy to store and handle.
(this technique is discussed later in detail).

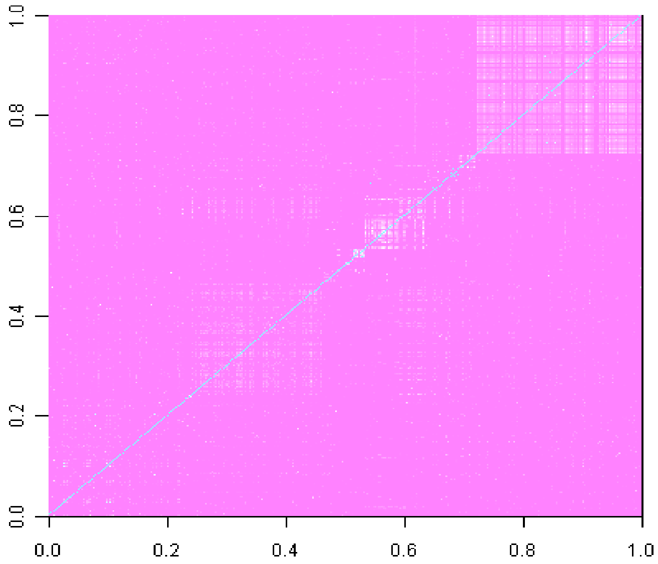


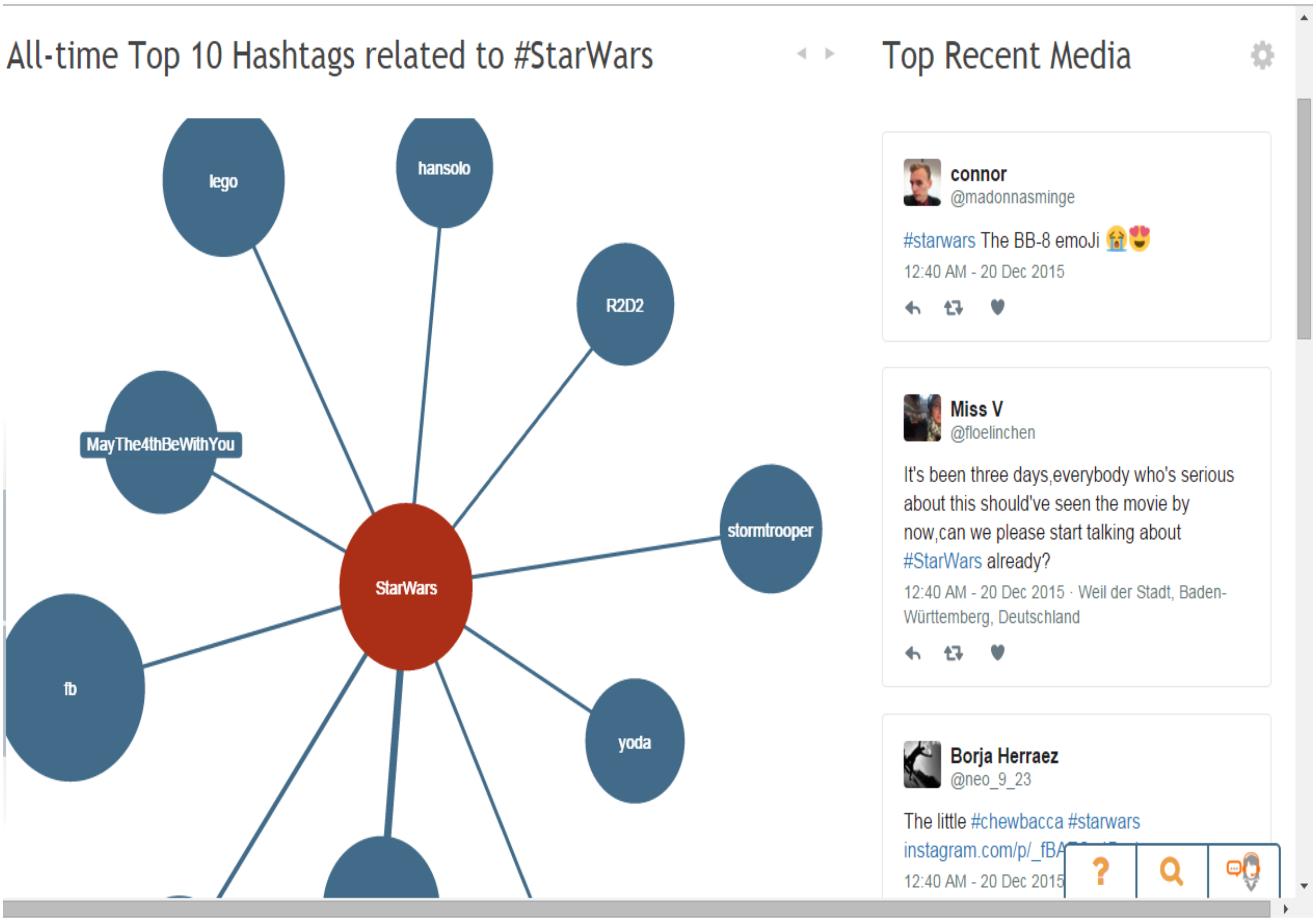
Figure 2. Distance to one point and the distribution of sample distance matrix

1.3 Objectives

The main aim of our project is:-

To present an interface that would provide a platform to view the related posts (from twitter) and pictures (from instagram) on a particular topic together at one place.

This figure shows a sort of proposed output of what we are trying to achieve.



connor
@madonnasminge

#starwars The BB-8 emoJi 😭😍

12:40 AM - 20 Dec 2015

Miss V
@floelinchen

It's been three days,everybody who's serious about this should've seen the movie by now,can we please start talking about #StarWars already?

12:40 AM - 20 Dec 2015 · Weil der Stadt, Baden-Württemberg, Deutschland

Borja Herraez
@neo_9_23

The little #chewbacca #starwars
instagram.com/p/_fBA

12:40 AM - 20 Dec 2015

Figure 3.

1.4 Methodology and Organisation

To achieve the above mentioned objective we tend to follow the following methodology:-

Phase 1: Fetching and Retrieval

In this phase we try to fetch the related posts and images from both the interfaces i.e. Twitter and Instagram and store them on a local database. And to achieve this we make use of the API's(Application Program Interface) of both the interfaces.

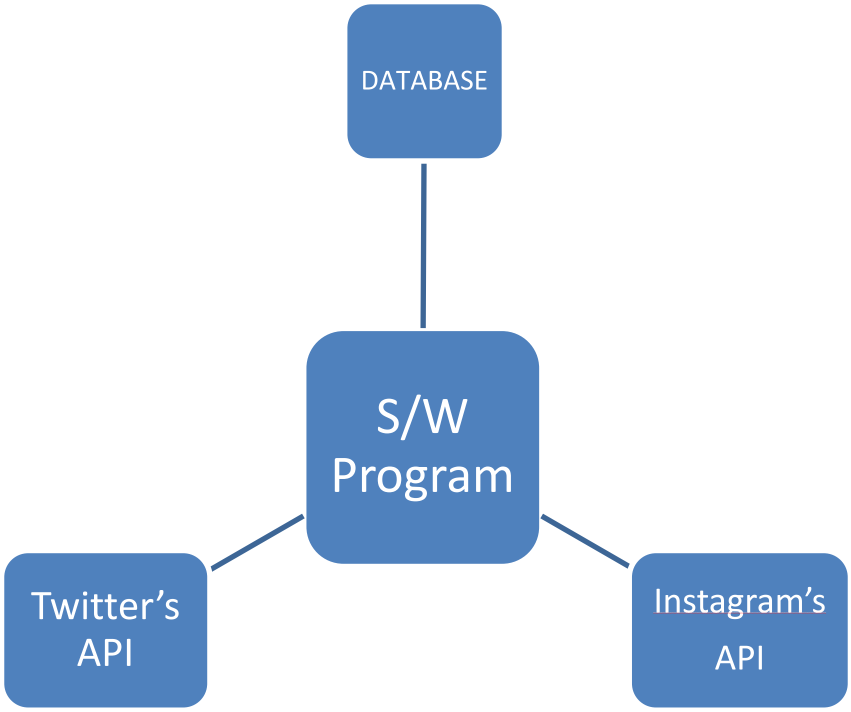


Figure4

Phase 2: Recommendation

Now, in this phase we already have the top related posts and we recommend new hashtags that can be related to the mentioned posts. For this purpose we use some of the recommendation techniques like:-

- Euclidean Distance Method
- Naïve Based Classification

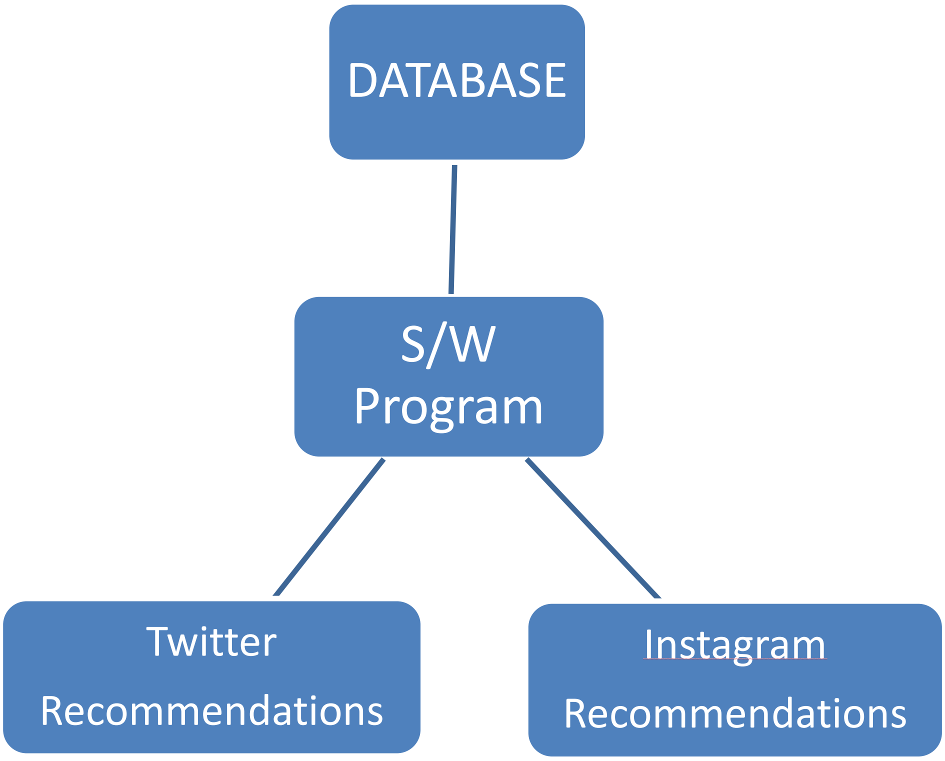


Figure 5

Phase 3: Accumulation and Final Output

In this phase we finally accumulate all the new and old related hashtags and put them together on a web portal.

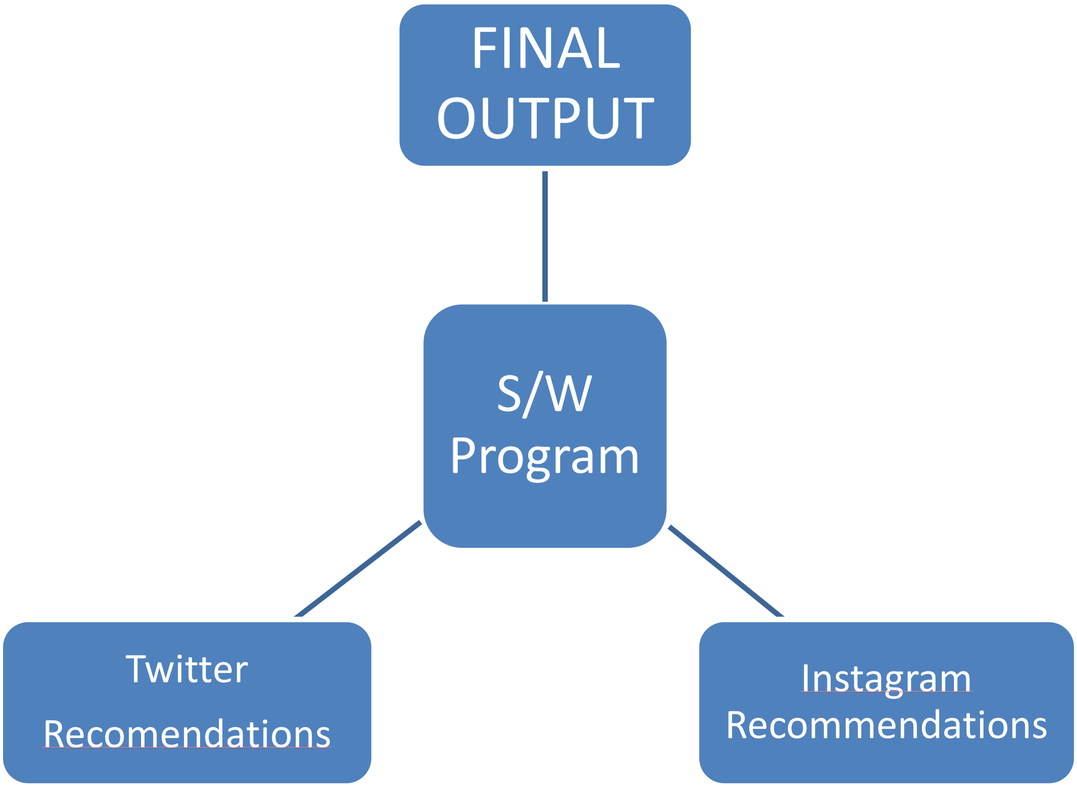


Figure 6

Chapter 2: Literature Survey

Summary of Research Papers

Title of Papers	Text-Based Twitter User Geological Location
Authors	Bo Han, Paul Cook and Timothy Baldwin
Year of Publication	2012
Publishing Details	Proceedings of the 24th International Conference on Computational Linguistics
Summary	Geographical location is vital to geospatial applications like local search and event detection . Hashtags can be predicted on the basis of geological location of a user. There are various geological references in a text (e.g. gazetteer terms, dialectal words) that are indicative of its author's location , using these references we can predict the geological location and accordingly its accompanying hashtags.

Title of Papers	Analyzing and predicting viral Tweets
Authors	Maximilian Jenders, Gjergji Kasneci and Felix Naumann
Year of Publication	2011
Publishing Details	Proceedings of the 20th international conference companion on World wide web.
Summary	In Virality of tweets, we gave deep analysis of “obvious” and “latent” tweets and user features with respect to their impact on the spread of tweets. For reliable prediction of viral tweets it is not enough to consider structural, content based or sentimental aspects in isolation. Rather, a combination of features covering all these aspects and a Learning model that avoids simplifying independent assumptions that are key to high prediction quality .An extensive analysis on this hypothesis is done.Twitter is one popular web application nowadays. Twitter allows users to use “Hash tags” to classify their tweets. It is called a micro-blog because people can post short, quasi-public messages up to 140 characters in length.

Title Of Papers	What’s in a Hashtag?
Authors	Oren Tsur and Ari Rappoport
Year of Publication	2011
Publishing Details	Proceedings of the 13th international conference on World Wide Web
Summary	Current social media research mainly focuses on temporal trends of the information ow and on the topology of the social graph that facilitates the propagation of information.In this paper we study the effect of the content of the idea on the information propagation. We present an efficient hybrid approach based on a linear regression for predicting the spread of an idea in a given time frame. We show that a combination of content features with temporal and topological features minimizes prediction error.Predicting the spread of ideas in online communities is an interesting task from both commercial and psychological perspectives. Traditional approaches model the propagation of ideas in social media by analyzing the topology of the social graph. In this work we took a hybrid approach to predicting spread of ideas according to their content as well as the topology of the social graph.

Title of Papers	On Predicting Twitter Trend: Important Factors and Models
Authors	Peng Zhang, Xufei Wang and Baoxin Li
Year Of Publication	2011
Publishing Details	In Proceedings of the 20th international conference on World Wide Web.
Summary	In this paper, we study two basic problems in information trend prediction, i.e. important factors and appropriate models. We designed features of different trend factors from both tweet content and network context. We also investigate model categories as the combination of two basic properties, i.e. (non)-linearity and (non)-state-space. Experiments on large Twitter dataset lead to the following observations. Both content and context factors will help trend prediction. However, node context factors of user's behavior on trend, e.g. trend stimulus and activeness, are most relevant. As for the prediction model, non-linear models are significantly better than their linear peers,which may mainly due to the complex information diffusion process in large social network. State-space can help to improve prediction but only on a slight degree.

Title of Papers	Twitter Hash Tag Prediction Algorithm
Authors	Tianxi Li & Yu Wu
Year Of Publication	2010
Publishing Details	In Proceedings of the 19th international conference on World Wide Web
Summary	Hash tag prediction is different from normal texts classification. Here we don’t know how many clusters we need to find. In addition, the tag set changes so frequently that it is almost impossible to effectively carry out classification or clustering, since a new tag would force us to establish a new class and a new classification rule. Our intuition is: if we can measure the correlation between various tweets as the mathematical metric we can treat the collected tweets as points in a high dimensional space, and construct a network by the latent space model. An intuitive way to solve this problem is to use Euclidean distance between points as the measurement of their similarity.

Title Of Papers	Suggesting Hashtags on Twitter
Authors	Allie Mazzia, and James Juett
Year of Publication	2011
Publishing Details	Proceedings of the 13th international conference on World Wide Web
Summary	As micro-blogging sites, like Twitter, continue to grow in popularity, we are presented with the problem of how to effectively categorize and search for posts. Looking specifically at Twitter, we see that users may categorize their posts using *hashtags*, and any word or phrase may be used as the category. Attempting to search for tweets about Facebook, a user would need to try many different hashtags, like *#Facebook*, *#FB*, *#Facebook.com*, or *#Zuckerberg*. To combat this, we propose, implement and evaluate a tool for suggesting relevant hashtags to a user, given a tweet.Initial analyses suggest our dataset is rich enough to extract informative distributions of words for many hashtags that will facilitate a naive Bayes model for hashtag recommendation given a query post..

Title Of Papers	Spatio-Temporal Meme Prediction: Learning What Hashtags Will Be Popular Where
Authors	Krishna Y. Kamath and James Caverlee
Year of Publication	2011
Publishing Details	Proceedings of the 13th international conference on World Wide Web
Summary	In this paper, we tackle the problem of predicting what online memes will be popular in what locations. Specifically, we develop data-driven approaches building on the global footprint of 755 million geo-tagged hashtags spread via Twitter. Our proposed methods model the geo-spatial propagation of online information spread to identify which hashtags will become popular in specific locations. Concretely, we develop a novel reinforcement learning approach that incrementally updates the best geo-spatial model. In experiments, we find that the proposed method outperforms alternative linear regression based methods..

Title Of Papers	On the Role of Conductance, Geography and Topology in Predicting Hashtag Virality
Authors	Siddharth Bora, Harvineet Singh, Anirban Sen, Amitabha Bagchi, and Parag Singla
Year of Publication	2011
Publishing Details	Proceedings of the 13th international conference on World Wide Web
Summary	We focus on three aspects of the early spread of a hashtag in order to predict whether it will go viral: the network properties of the subset of users tweeting the hashtag, its geographical properties, an d, mostimportantly, its conductance-related properties. One of our significant contributions is to discover the critical role played by the conductance based features for the successful prediction of virality. More specif-ically, we show that the first derivative of the conductance gives an early indication of whether the hashtag is going to go viral or not. We present a detailed experimental evaluation of the e_ect of our various categories of features on the virality prediction task. When compared to the baselines and the state of the art techniques proposed in the literature our feature set is able to achieve signi_cantly better accuracy on a large dataset of 7.7 million users and all their tweets over a period of month, as well as on existing datasets..

Title Of Papers	Predicting Bursts and Popularity of Hashtain Real-Time
Authors	Shoubin Kong, Qiaozhu Mei, Ling Feng1, Fei Ye and Zhe Zhao
Year of Publication	2011
Publishing Details	Proceedings of the 13th international conference on World Wide Web
Summary	Hashtags have been widely used to annotate topics in tweets(short posts on Twitter.com). In this paper, we study the problems of real-time prediction of bursting hashtags. Will a hashtag burst in the near future? If it will, how early can we predict it, and how popular will it become? Based on empirical analysis of data collected from Twitter, we propose solutions to these challenging problems. The performance of different features and possible solutions are evaluated.

Title Of Papers	On the Real-time Prediction Problems of Bursting Hashtags in Twitter
Authors	Shoubin Kong, Qiaozhu Mei, Ling Feng, Zhe Zhao, and Fei Ye3
Year of Publication	2011
Publishing Details	Proceedings of the 13th international conference on World Wide Web
Summary	Hundreds of thousands of hashtags are generated every day on Twitter. Only a few become bursting topics. Among the few, only some can be predicted in real-time. In this paper, we take the initia- tive to conduct a systematic study of a series of challenging real-time prediction problems of bursting hashtags. Which hashtags will become bursting? If they do, when will the burst happen? How long will theyremain active? And how soon will they fade away? Based on empiri-cal analysis of real data from Twitter, we provide insightful statistics to answer these questions, which span over the entire lifecycles of hashtags.

Title Of Papers	Using Topic Models for Twitter Hashtag Recommendation
Authors	Fréderic Godin, Viktor Slavkovikj and Wesley De Neve
Year of Publication	2011
Publishing Details	Proceedings of the 13th international conference on World Wide Web
Summary	Since the introduction of microblogging services, there has been a continuous growth of short-text social networking on the Internet. With the generation of large amounts of microposts, there is a need for effective categorization and search of the data. Twitter, one of the largest microblogging sites, allows users to make use of hashtags to categorize their posts. However, the majority of tweets do not contain tags,which hinders the quality of the search results. In this paper, we propose a novel method for unsupervised and contentbased hashtag recommendation for tweets. Our approach relies on Latent Dirichlet Allocation (LDA) to model the underlying topic assignment of language classified tweets. The advantage of our approach is the use of a topic distribution to recommend general hashtags..

Title Of Papers	Effective Tweet Contextualization with Hashtags Performance Prediction and Multi-Document Summarization.
Authors	Romain Deveaud and Florian Boudin
Year of Publication	2011
Publishing Details	Proceedings of the Second International Conference on Human Language Technology Research,
Summary	In this paper we describe our participation in the INEX 2013 tweet Contextualization track and present our contributions. Our approach is the same as last year, and is composed of three main components: preprocessing, Wikipedia articles retrieval and multi-document summarization.We however took advantage of a larger use of hashtags in the topics and used them to enhance the retrieval of relevant Wikipediaarticles. We also took advantage of the training examples from last year which allowed us to learn the weights of each sentence selection feature.Two of our submitted runs achieved the two best informativeness results, while our generated contexts where almost as readable as those of the most readable system.

Title Of Papers	Connecting Tweets to Explicit Topics
Authors	Wei Feng , Jianyong Wang
Year of Publication	2011
Publishing Details	Proceedings of the 13th international conference on World Wide Web
Summary	Current social media research mainly focuses on temporal trends of the information ow and on the topology of the social graph that facilitates the propagation of information.In this paper we study the effect of the content of the idea on the information propagation. We present an efficient hybrid approach based on a linear regression for predicting the spread of an idea in a given time frame. We show that a combination of content features with temporal and topological features minimizes prediction error.Predicting the spread of ideas in online communities is an interesting task from both commercial and psychological perspectives. Traditional approaches model the propagation of ideas in social media by analyzing the topology of the social graph. In this work we took a hybrid approach to predicting spread of ideas according to their content as well as the topology of the social graph.

Title Of Papers	Time-aware Personalised Hash Tag Recommendation on Social Media
Authors	Qi Zhang, Yeyun Gong, Xuyang Sun,and Xuanjing Huang
Year of Publication	2010
Publishing Details	Proceedings of the 21st ACM conference on Hypertext and hypermedia,
Summary	The task of recommending hashtags for microblogs has been received considerable attention in recent years, and many applications can reap enormous benefits from it. Various approaches have been proposed to study the problem from different aspects. However, the impacts of temporal and personal factors have rarely been considered in the existing methods. In this paper, we propose a novel method that extends the translation based model and incorporates the temporal and personal factors. To overcome the limitation of only being able to recommend hashtags that exist in the training data of the existing methods, the proposed method also incorporates extraction strategies into it. The results of experiments on the data collected from real world microblogging services by crawling demonstrate that the proposed method outperforms state-of-the-art methods that do not consider these aspects.

Title Of Papers	On Recommending Hashtags in Twitter Networks
Authors	Tuan Anh Hoang, Ee Peng LIM and Feida ZHU
Year of Publication	2011
Publishing Details	International Workshop on Social Web Mining
Summary	Twitter network is currently overwhelmed by massive amount of tweets generated by its users. To effectively organize and search tweets,users have to depend on appropriate hashtags inserted into tweets. We begin our research on hashtags by rest analyzing a Twitter dataset generated by more than 150,000 Singapore users over a three-month period. Among several interesting findings about hashtag usage by this user com-munity, we have found a consistent and significant use of new hashtags on a daily basis. This suggests that most hashtags have very short life span.We further propose a novel hashtag recommendation method based on collaborative iterating and the method recommends hashtags found in the previous month's data. Our method considers both user preferences and tweet content in selecting hashtags to be recommended. Our paper also proposes a personalized hashtagrecommendation method that considers both target user preferences and target tweet content. Given a user and a tweet, our method selects the top most similar users and top most similar tweets. Hashtags are then selected from the most similar tweets and users and assigned some ranking scores..

Title Of Papers	The Unpredictability of Emotional Hashtags in Twitter.
Authors	Florian Kunneman_, Christine Liebrecht and Antal van den Bosch
Year of Publication	2013
Publishing Details	Computational Linguistics and Intelligent Text Processing
Summary	Hashtags in Twitter posts may carry different semantic payloads. Their dual form (word and label) may serve to categorize the tweet, but may also add content to the message, or strengthen it. Some hashtags are related to emotions. Potentially, Twitter offers a vast amount of data to exploit for the construction of computational models able to detect certain sentiments or emotions in unseen tweets. Yet, in the typical scenario of applying supervised machine learning classifiers,some annotation effort will be required to label sentiments and emotions reliably. Currently there are two main approaches to labeling tweets. In our experiments we showed that machine learning classifiers can be relatively successful both in predicting the hashtag with tweets which were indeed tagged with them, and classifying tweets without the hashtag as exhibiting the emotion denoted by the hashtag, for two of the four fully analysed hashtags: #zinin and #fml. In contrast, the classifier of the hashtag #geenzin was only able to re-link tweets that are stripped from the target hashtag with this hashtag, but failed to capture the complex emotion behind the hashtag. The performance of the #omg classifier lags behind in both tasks.

Title Of Papers	On Predicting Popularity Of Newly Predicting Hash Tags In Twitter
Authors	Zongyang Ma, Aixin Sun, and Gao Cong
Year of Publication	2011
Publishing Details	Proceedings of the 13th international conference on World Wide Web
Summary	Because of Twitter's popularity and the viral nature of information dissemination on Twitter, predicting whichTwitter topics will become popular in the near future becomes a task of considerable economic importance.Many Twitter topics are annotated by hashtags. In this article, we propose methods to predict the popularity of new hashtags on Twitter by formulating the problem as a classification task. We use five standard classification models (i.e., Naïve bayes, *k*-nearest neighbors, decision trees, support vector machines, and logistic regression)for prediction. The main challenge is the identification of effective features for describing new hashtags. We extract 7 content features from a hashtag string and the collection of tweets containing the hashtag and 11 contextual features from the social graph formed by users who have adopted the hashtag. We conducted experiments on a Twitter data set consisting of 31 million tweets from 2 million Singapore-based users. The experimental results show that the standard classifiers using the extracted features significantly outperform the baseline methods that do not use these features. We propose methods to predict the hashtag popularity of new topics on Twitter by formulating the problem as a classification task and evaluating three baseline methods and five classification methods.

Chapter 3: System Development

Currently, Twitter has not implemented any hashtag recommendation system which suggests appropriate hashtags for the users' tweets. In the research literature, there are works related to hashtag recommendation and hashtag prediction.

We found two hashtag recommendation approaches that are relevant and both of them use *only* tweet content. They will be described below in greater detail. Hashtag prediction refers to predicting the hashtag to be used by a user in the future.

Preliminary analysis of hashtags usage in a Twitter data collection obtained by a set of search queries shows that 86% of unique hashtags are used less than five times within 3,209,281 tweets with hashtags. The five most popular hashtags (#jobs, #nowplaying, #zodiacfacts, #news and #fb) appear in 8% of all tweets with hashtags. In other words, a few popular hashtags are used intensively while most of the other hashtags are used very sparsely. The paper also finds out the use of hashtags by spammers (e.g. assigning 17 hashtags to a single spammed tweet).

Zangerle et al. proposed a hashtag recommendation system that retrieves a set of tweets similar to a user given tweet. Similarity score is calculated by TFIDF scheme. Then, the hashtags are extracted from the retrieved similar tweets and are ranked using one of the proposed score functions:
(a) OverallPopularityRank score: number of hashtag occurrences in the whole dataset;

(b)RecommendationPopularityRank score: number of hashtag occurrences in the retrieved similar tweet dataset; or

(c) SimilarityRank score: similarity score of the most similar tweets containing the hastag.

Experiments showed that SimilarityRank score is the best among them and the performance of the recommendation system is the best when only five hashtags are recommended.

3.1Design and Development

The main aim of the project is:-

To present an interface that would provide a platform to view the related posts (from twitter) and pictures (from instagram) on a particular topic together at one place.

And to achieve this objective the following design and development methodologies are followed.

Requirement Analysis Models

Through this phase the requirements were found more in detailed manner by examining their detailed boundary conditions & exceptional cases.

Data Flow Diagram:

These diagrams depict the flow of data from one point of the system to another point. It mainly consists of three parts divided on the basis of its level:

a) *Level '0' Data Flow Diagram (Context Diagram).*
b) *Level '1' DFD.*

Level 0 DFD/Context Diagram:

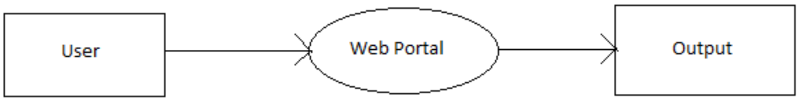


Figure 7

Level 1 DFD:

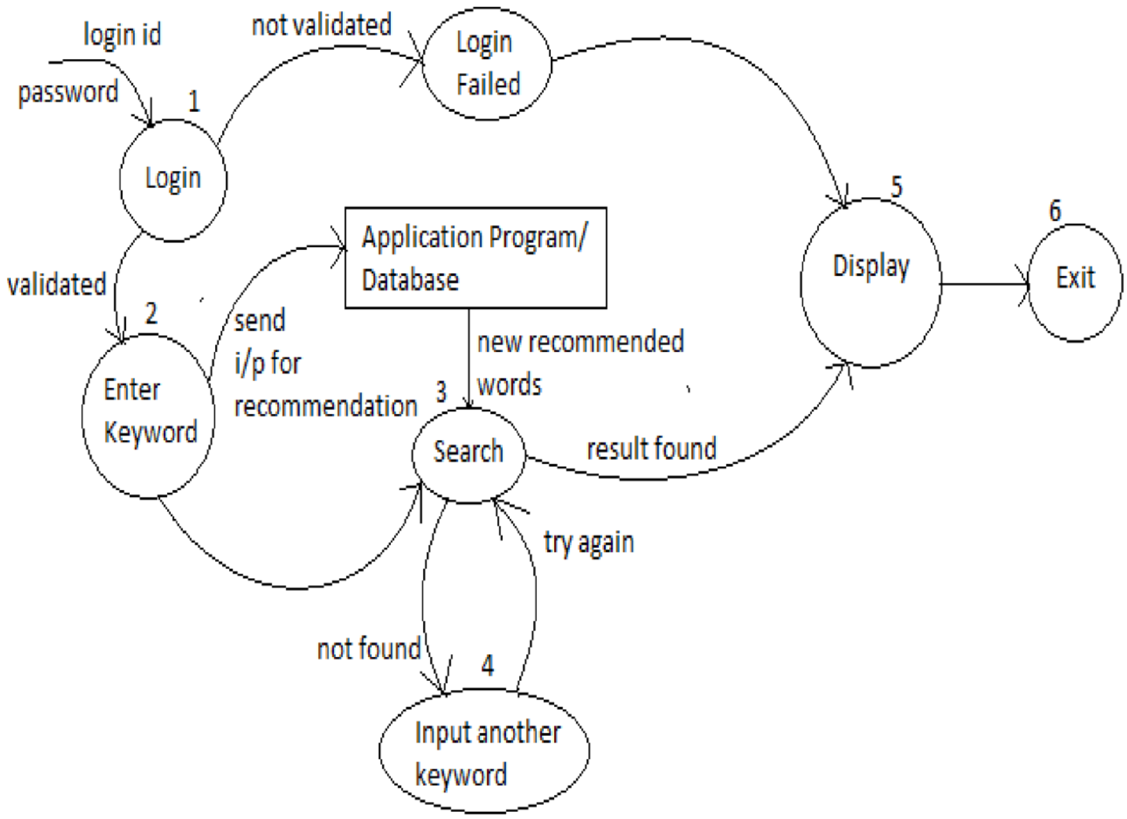


Figure 8

Behaviour Modelling:

It tells about how system behaves & how users use it. In this one must use ‘Use-Cases’ & ‘Use-Case Description’ for which steps have to be developed.

Sequence Diagram:

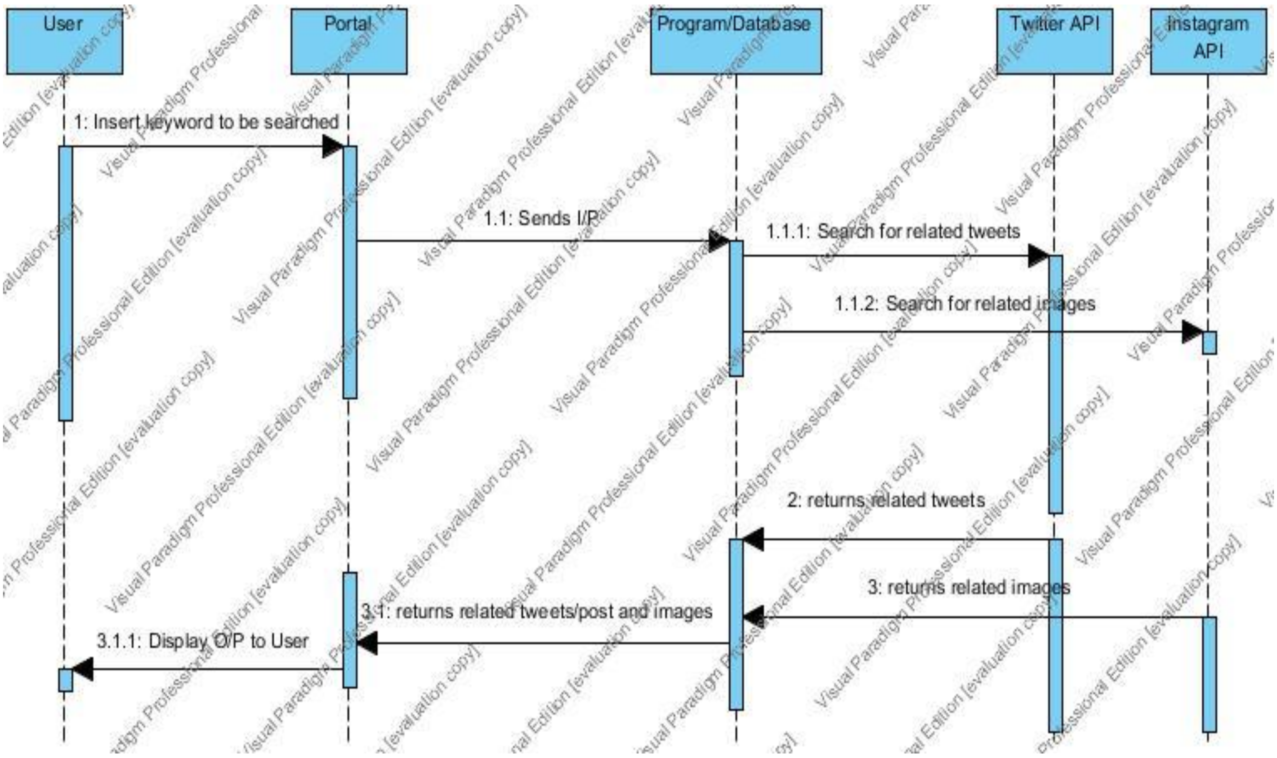

User
Portal
Program/Database
Twitter API
Instagram API
1: Insert keyword to be searched
1.1: Sends I/P
1.1.1: Search for related tweets
1.1.2: Search for related images
2: returns related tweets
3: returns related images
3.1: returns related tweets/post and images
3.1.1: Display O/P to User

Figure 9

3.2Development Methodologies

NAIVE BAYES METHOD

In the experiments, the Twitter dataset used is first cleaned by removing micro-memes and spams. Micro-memes are detected by identifying tweets which use the same hashtags but are very dissimilar. Spams are filtered by removing users who have too many tweets using the same hashtag. The Bayes model used in this paper is represented by the following formula.

$p(Cij|x1, ... , xn) = p(Ci)\ p(x1|Ci)...p(Ci)p(xn|Ci)\ /\ p(x1...xn)$

where *Ci* represents the *ith* hashtag and *x*1,..., *xn* represent the words. $p(Cij|x1, ... , xn)$ is the probability of using hashtag *Ci* given the words that the user generates and the hashtags with the highest probabilities are recommended to the user. $p(Ci)$ is the ratio of the number of times hashtag *Ci* is used to the total number of tweets with hashtags. $p(x1jCi):::p(xnjCi)$ is calculated from the existing data of tweets.

A complete analysis of hashtag usage in the entire Twitter network is not possible as such a dataset is not publicly available. Most researchers in the past chose to analyze Twitter data collected using some forms of data sampling on the stream of Twitter data returned by the APIs provided by the company. Inevitably, the analysis results will be biased by the query relevant tweets.

The analysis aims to answer the following research questions:
(a) How often are hashtags used in tweets?
(b) How many hashtags do we expect in a tweet?
(c) How familiar are users in using hashtags?
(d) Do the hashtags assigned already appear in earlier tweets?

Providing answers to the above questions will give a good understanding of the hashtag usage patterns of a user community and their changes over time.

Twitter data generated by more than 150,000 Singapore users who are identified by the location field in their user profiles. Every user is at least directly or indirectly connected to a small set of carefully selected seed users so as to prevent spammers to be included. The seed users are popular political bloggers, commentators, election candidates and news media during Singapore Election 2011. Since election is a big socio-political event, we believe that we cover the majority of Singapore Twitter users. All tweets of these Singapore users on a daily basis have been crawled. In this manner, we are assured that almost all tweets from this user community have been completely downloaded for our study. Table 1 shows the important statistics found in this dataset. There are more 65,000 users who have written some original tweets during the three-month period. The remaining users (nearly 60% of total user population) do not write original tweets. They could perform retweeting or simply reading tweets from others. The dataset also contains nearly 450,000 distinct hashtags and 45M original tweets.

EUCLIDEAN DISTANCE METHOD

The hash tag (the # sign followed by a phrase to a tweet, for example #superbowl) is probably the most important function of Twitter search, and the most used The hash tag enables Twitter users to create searchable subject groups and so to be able to navigate the hypertext structures of the whole site. The power of the hash tag is that it creates very specific sets of content. If you want to know what other people think of the superbowl that just came on you can find it easier by searching for the hash tag than by searching for something similar in a normal search engine. Every day, many new hash tags are formed and this process can happen right before your eyes-heck. The frequent creation of new tags makes the prediction of tags challenging. This motivates us to develop the following method.

METHOD

Theory

An intuitive way to solve this problem is to use Euclidean distance between points as the measurement of their similarity. We developed our theory based on this distance. Since in a Euclidean Space, the distance is equivalent to the norm of a vector, we will focus our discussion on norms.

Let **u**1,**u**2 · **u***p* be the standard bases (with unit norm) of a *p*-dimensional Euclidean Space. Then for any vector **v** with coordinates (*x1, x2, ..., xp-1, xp*), we have

$$v = \sum x_i u_i$$

Then the Euclidean norm of vector **v** is given by

$$||\mathbf{v}||^2 = \mathbf{v.v} = \sum x_i u_i . \sum x_i u_i = \sum x_i x_j u_i u_j$$

where · represents the inner product operation defined in the Euclidean Space. Clearly, if we assume $u_i . u_j = 0$ that is, u_i and u_j are orthogonal, whenever $i \neq j$, the Euclidean norm equals to

$$||\mathbf{v}||^2 = \sum x_i^{\ 2}$$

In our problem, the bases are the words in the dictionary. The preliminary assumption for Euclidean distance is that the bases are orthogonal to each other, that is, the words in dictionary are uncorrelated, which is against common sense. Therefore, we need to perform some transformation to capture this correlation.

In Equation (1), as u_i and u_j are unit vectors, their inner product is actually the cosine of the angle between them. Thus we can rewrite (1) in a matrix form as:

$$||\mathbf{v}||^2 = (x_1 \quad \dots \quad x_p) \begin{pmatrix} \cos\Theta_{11} \dots \cos\Theta_{1p} \\ . \quad\quad . \quad\quad . \\ \cos\Theta_{p1} \dots \cos\Theta_{pp} \end{pmatrix} \begin{pmatrix} x_{i1} \\ . \\ x_p \end{pmatrix} = XMX^T \qquad (2)$$

where *cos* · *ii* = 1, *i*=1, …, *p*, and x = (*x1, x2, ..., xp-1, xp*).

Now we try to find the angle between each pair of terms in the dictionary and then calculate the matrix *M* . Notice that *M* is clearly a symmetric and non-negative definite matrix. If we decompose M in the way

$$M = CC^T \qquad (3)$$

then (2) becomes

$$||\mathbf{v}||^2 = XCC^T X^T = X\sim X\sim^T \qquad (4)$$

where *X~* = *X C* . So the norm can be seen as the Euclidean norm of the transformed coordinates. Here we take (3) as the Eigen value decomposition of *M*, so *X.* could be the coordinates of vector **v** in a new coordinate system where axes are orthogonal to each other. Please note that we can use any other decomposition in the form of (3) to get the same norm in computation, even when *C* is not a square matrix.

With this property, the computation becomes applicable.

Estimate the Cosine Matrix

First, we construct the preliminary weighted matrix, say *H*, by using the WordNet to initialize the semantic correlation among words from the dictionary. If two words t_i, t_j are similar to each other, and they both appear in one Tweet, we add positive weights for both words. This process can be expressed as

$$\hat{x}_i = x_i + \sum_{j \neq i}^{p} \rho_{ij} x_j$$

where ρ_{ij} ϵ(0,1) , equals to one when , t_i, t_j are similar words and zero otherwise. Here we take the same positive number ρ for all ρ_{ij} ϵ(0,1) , and if $\rho_{ij} > 0$ so is ρ_{ij} .Then we can construct the symmetric matrix H as:

$$\mathrm{H} = \begin{pmatrix} 1 & \cdots & \rho_{1p} \\ \vdots & \ddots & \vdots \\ \rho_{p1} & \cdots & 1 \end{pmatrix} \qquad \hat{X} = X\mathrm{H} \quad (6)$$

In the second step, we get m tweets, say X_1 ... X_m , and transform them by (4) to get $X_{1^\wedge}$... $X_{m^\wedge}$ Then by these data, we use cosine similarity in variable analysis to construct matrix M . Set the text matrix as the $m \times p$ matrix :

$$\Omega = (\hat{X}_1^T \quad \cdots \quad \hat{X}_m^T)^T = \begin{pmatrix} \hat{x}_{11} & \cdots & \hat{x}_{1p} \\ \vdots & \ddots & \vdots \\ \hat{x}_{m1} & \cdots & \hat{x}_{mp} \end{pmatrix}. \qquad (7)$$

We would estimate the cosine between the ith and jth terms as

$$\cos\theta_{ij} = \hat{X}_{\bullet i}^T \hat{X}_{\bullet j} \Big/ \left\|\hat{X}_{\bullet i}\right\| \left\|\hat{X}_{\bullet j}\right\| = \frac{\sum_{k=1}^{m} \hat{x}_{ki}\hat{x}_{kj}}{\sqrt{\sum_{k=1}^{m} \hat{x}_{ki}^2 \sum_{k=1}^{m} \hat{x}_{kj}^2}} \qquad (8)$$

The distance estimate obtained from formula (5) is with a better mathematical explanation. Note that since our data is represented as frequency, all the elements of the matrix Ω would be non-negative. So the cosine estimated in this way can only be non-negative. Therefore, all angles between words are cute or right angles. In this way, all words tend to be similar to each other in some degree. This may well incorporate the similarity elements, but might also be vulnerable to noise. In the following, we give a modified estimate which also includes the possibility of obtuse angle and takes dissimilarity into consideration, which is also the sample correlation in statistics,

$$\cos\theta_{ij} = \frac{\sum_{k=1}^{m}(\hat{x}_{ki} - \hat{x}_{\bullet i})(\hat{x}_{kj} - \hat{x}_{\bullet j})}{\sqrt{\sum_{k=1}^{m}(\hat{x}_{ki} - \hat{x}_{\bullet i})^2 \sum_{k=1}^{m}(\hat{x}_{kj} - \hat{x}_{\bullet j})^2}} \qquad (9)$$

Since the distance from formula (8) was named as Ontology Based Distance (OBD), here we call the distance in formula (9) centralized Ontology Based Distance (COBD). We will discuss the pros and cons of the two methods in experiments. In the following sub section, we will make another adjustment to the method.

Normalization

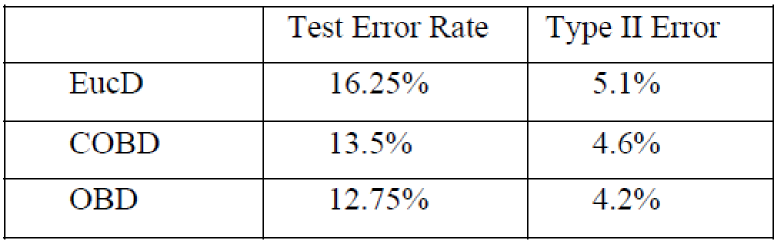
Note that the various scales of vectors may still cause us some problem. Consider a special case where $X^{\sim}_{1} = (1,0,0)$, $X^{\sim}_{2} = (10,0,0)$, $X^{\sim}_{3} = (0,0,1)$. Obviously, $X^{\sim}_{1}$ and $X^{\sim}_{2}$ should have high similarity value between them. But in this case, the distance between $X^{\sim}_{1}$ and $X^{\sim}_{3}$ is much smaller. To make our method more reasonable, before we compute the distance between transformed points, we need to rescale their distances to the original point as 1. And then we measure the Euclidean distance between normalized points.

Prediction of Tags

Finally, we predict tags based on the distance. One intuitive way is to simply select the tag of the closest tweet. In this case, it may be unwise to simply pick the closest tweet's tag, since that is not resistant to noise. To increase the accuracy, we collect a few closest tweets, and make the prediction based on tag ratios. Specifically, we will collect *n* initial closest tweets at first (*n* usually ranges from 4 to 6). Then from this point, we will keep adding tweets while check a certain tag has become dominate. If there is a tag with a ratio higher than 50%, we will choose this tag as our primary predicted tag. Since in some cases tags have very similar meanings (such as #government vs. #election), sometimes we will also pick a secondary tag to predict.

3.3 EVALUATIONS

To compare the performances of various distances discussed above, we use a test dataset consist of 400 tweets that are not included in the sample set we used to estimate matrix. There are 4 different tags. We first process the OBD on a dataset with 665 tweets that are not in our test set, choose the best performance ρ (=0.2) and use it for both OBD and COBD. The table below shows the test result for Euclidean Distance (EucD), OBD and COBD.



	Test Error Rate	Type II Error
EucD	16.25%	5.1%
COBD	13.5%	4.6%
OBD	12.75%	4.2%

Table1: The test error rate and type II error for three distances. Type II error is the rate we assign a wrong tag to a particular tweet.

Both OBD and COBD outperform EucD, and OBD is the best one. If we see the data for different tags (not provided here for concise), we would find COBD is the most stable one, while EucD is far more unstable. But the disadvantage of COBD lies in computation. We need to estimate the cosine matrix M to construct the distance, which involves computation for matrices with tens of thousands rows and columns. It won’t be a big problem for OBD since the matrices are sparse. But in COBD, the matrix becomes non-sparse, so we need many decompositions and transformations of matrices to make the computation applicable. Given their close performances, OBD is more practical in application, while the COBD is a better model theoretically.

The top picture in Figure 1 shows the COBD from other tweets to a random selected tweet. Different colors represent tweets with different tags. It can be seen that most of the tweets are very close to the 1.4142 distance boundary, and the majority of points falling in the circle are from the correct tag group. This indicates that tweets with different topics are projected onto orthogonal axes. The right plot illustrates the distance distribution. The lighter the color is, the shorter the corresponding distance is. Since the tweets are sorted by tags, we see that the distance within each group appears to be shorter, as shown by the light rectangles along the diagonal.

In Figure2, different colors represent what tag cluster the tweets belong to. A link will be added between a pair of nodes when they are near enough. In addition, the deeper color the line is, the higher the similarity value is. As we can see, the lines appear to be very dense among each tag cluster, and sparse between tweets with different tags. It indicates that tweets with the same tag cluster are near on average.

Due to the vagueness of many tweets, the correct rate of more than 86% is actually very high. Apart from the accuracy, our method has other advantages:

(1) The whole system is easy to store (we only need to store the *C* matrix in Equation (3).

(2) It is easy to update when dictionary changes (only needs to compute an extra column and add it back to original matrix).

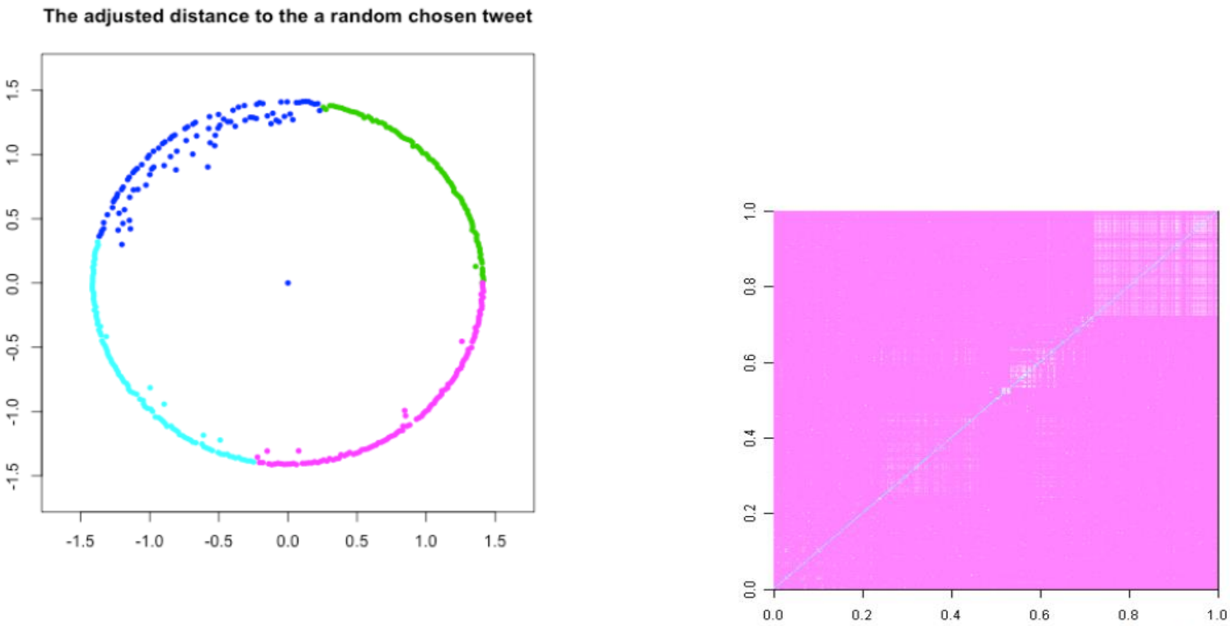


Figure 1 Distance to one point and the distribution of sample distance matrix

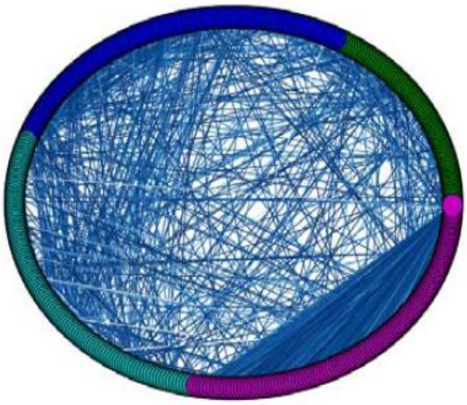


Figure 2 Prediction Visualization

Figure 10

(3) It won’t lose power when the topics trend changeswith time, and it can work with personal elements and settings, which makes it more flexible (since we can set the algorithm to only consider the distance of the objective tweet to certain subset of other tweets, so elements like location, time, etc can be incorporated.

(4) In addition, the distance provides us with the possibility to transform the twitter system and even other text systems into social networks by latent space approach. So we can use traditional social network methods to discuss the properties of such systems.

Chapter 4: Performance analysis

Hashtag Usage Analysis

There are substantial fraction of users (about 39%) using hashtags in their original tweets, and very small fraction of original tweets containing hashtags (<8%) as shown in Table 1. This suggests that many users know how to use hashtags but very few actually tweet a lot using hashtags. Figure 1 shows that the fraction of users using hashtags and the fraction of tweets containing hashtag over the three-month period remain very stable for this user community.

We define *tweet popularity* of a hashtag by the number of tweets containing the hashtag. We show the scatterplot of tweet popularity of hashtags in Figure (a). Each point in the figure represents the number of hashtags with the same tweet popularity. The distribution is power law-like showing that most hashtags appear in one tweet each and very few tweets enjoy very high tweet popularity. In a similar way, we define *user popularity* of a hashtag by the number

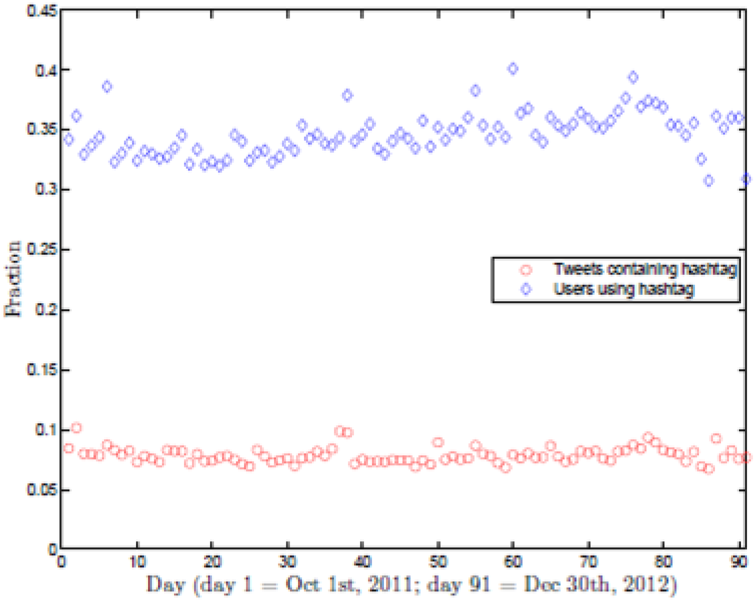


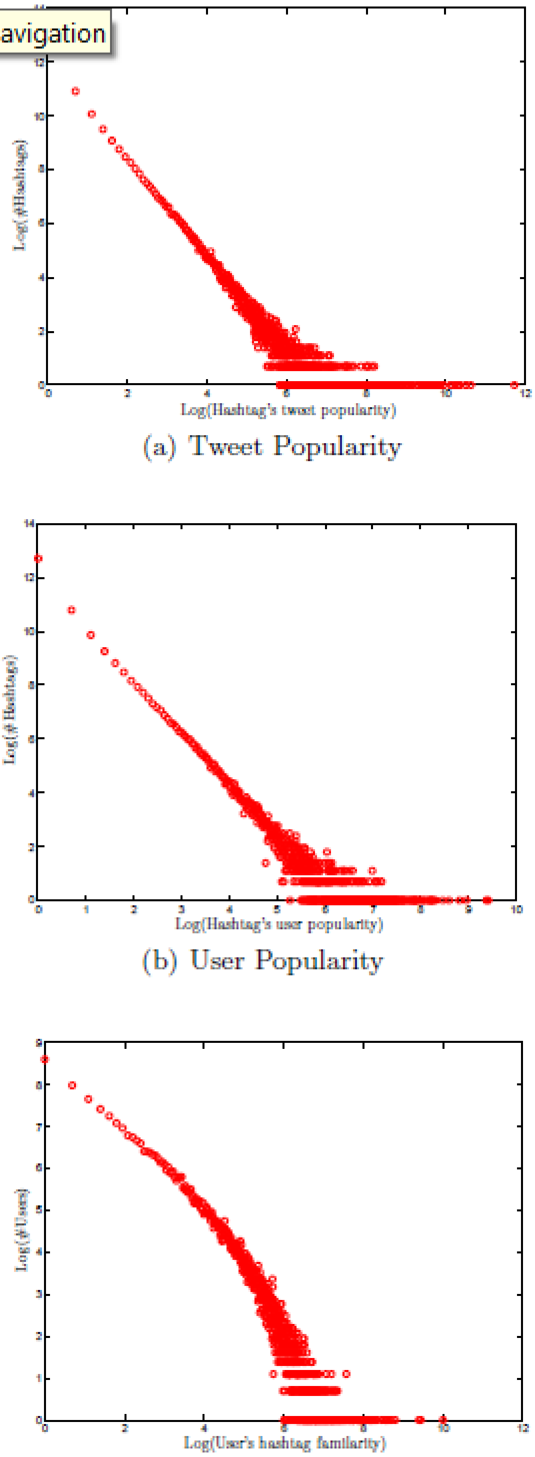
Figure 11

of users using the hashtag. Figure (b) shows that the user popularity distribution of hashtags also follows the power law distribution. This suggests that only a few hashtags enjoy high popularity while most hashtags are used by a single user.

Next, we analyze how frequently users write tweets with hashtags. As shown in Figure (c), most users write only one tweet containing hashtag(s) during the observed period. Very few users write many tweets that contain hashtags. Finally, we found out most tweets with hashtag(s) contain only one hashtag as shown in Figure . There are very few tweets containing more than one hashtag. This is not a surprise given the short tweet length.

Finally, we want to know if the hashtags are new as users assign them to tweets. Unfortunately, the verification of new hashtags is very costly and may not be viable due to the lack of all historical twitter data. We therefore introduce the definition of \fresh hashtag". A hashtag is said to be fresh to a user community if it has not been used by any user in the community in the last *k* months. This definition constrains the freshness verification to only *k* previous months of data generated by a user community. To reduce the verification cost, we have *k* = 1 in our current study.

Figure 4 depicts the fraction of fresh hashtags, the fraction of tweets containing fresh hashtags and the fraction of users using fresh hashtags for each day. It is interesting to find 40% fresh hashtags are introduced each day. This suggests that another 40% hashtags are replaced each day. The life expectancy of many hashtags are therefore very short. Less than 30% of tweets contain fresh hashtags and around 40% of users use fresh hashtags each day. These observations lead us to believe that hashtag recommendation is an important task as it helps



(a) Tweet Popularity

(b) User Popularity

(c) User Hashtag Familiarity

Figure 12

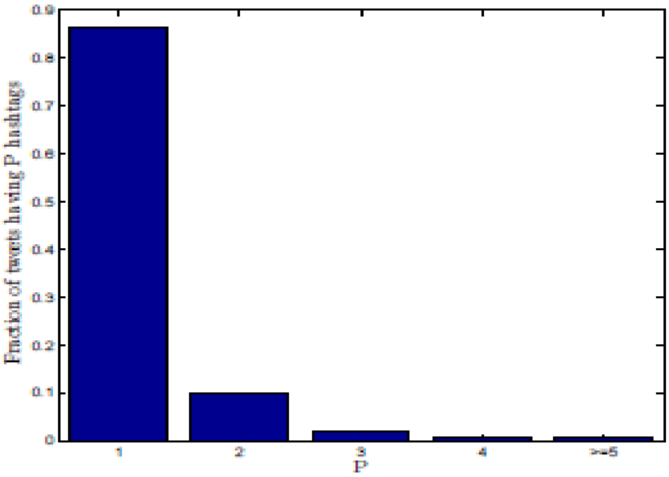
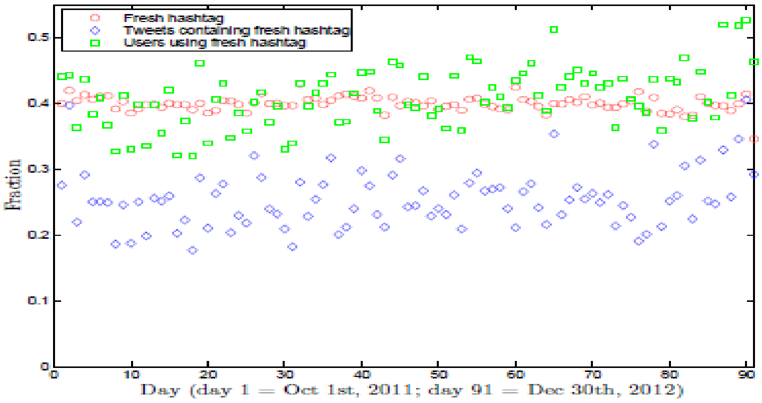


Figure 13

users to adopt more hashtags and makes their tweets easily searchable by other relevant users. The recommendation should also involve recent past data so as to recommend fresher hashtags.



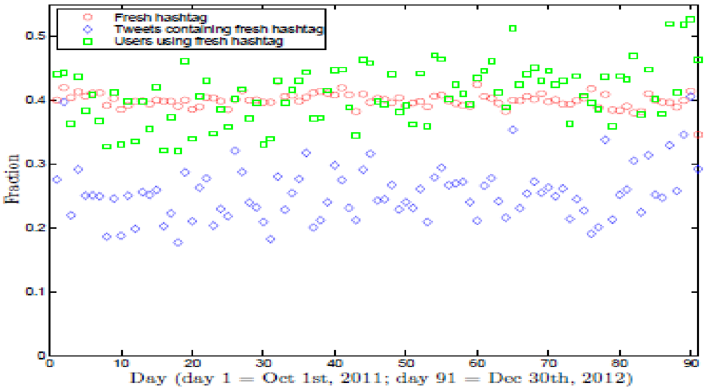


Figure 14

Finding similar user-tweet pairs involves three subtasks: (a) selecting hashtags from users with preferences similar to the target user, (b) selecting hashtags from tweets that are similar to the target tweet, and (c) deriving ranking scores for the selected hashtags. In both subtasks (a) and (b), we adopt a TF-IDF scheme to find similar users and tweets as described below.

Selecting hashtags from similar users. We represent a user by her preference weights for each hashtag in our hashtag dictionary *H*. Formally, a user *uj* is represented by a weight vector:

$$u_j = \{w_{1j}, w_{2j}, w_{3j}, \ldots, w_{i|H|}\}$$

where *wij* is the preference weight of user *uj* towards hashtag *hi* and can be defined by the TF-IDF scheme.

$$w_{ij} = TF_{ij}.IDF_i$$

$$TF_{ij} = \frac{Freq_{ij}}{Max_j}, IDF_i = log\left(\frac{N_u}{n_i}\right)$$

where *Freqij* = usage frequency of hashtag *hi* by user *uj* , *Maxj* = maximum hashtag usage frequency by *uj* , *Nu* = total number of users, and *ni* = number of users who use *hi* before.

The intuition of *TFij* is that if a user uses a hashtag a lot, more preference weight is given to the hashtag. At the same time, this weight is normalized by the maximum hashtag frequency of the user. *IDFi* assigns higher weight to a hashtag if the latter is rarely used by other users.

Given a target user *u* and another user *ui*, we can measure the cosine similarity between them as follows.

$$Sim(u, u_i) = \frac{u \cdot u_i}{||u|| \cdot ||u_i||}$$

The users are ranked by similarity score and the most similar *X* users are selected. Let *TopXUsers*(*u*) denote the *X* users most similar to *u*, and *Hashtags*(*ui*) be the set of hashtags previously used by *ui*. We combine the hashtags from these top-*X* users to be our candidate hashtag set *HTofUsers*(*u*).

$$HTofUsers(u) = \cup_{u_i \in TopXUsers(u)} Hashtags(u_i)$$

Selecting hashtags from similar tweets. In a similar manner, we represent a tweet *tk* can be represented by a weighted vector of words in a word vocabulary *W*.

$$t_k = \{w_{k1}, w_{k2}, w_{k3}, \dots, w_{k|W|}\}$$

where

$$w_{kl} = TF_{kl}.IDF_l$$

$$TF_{kl} = \frac{Freq_{kl}}{Max_k}, IDF_l = log\left(\frac{N_t}{n_l}\right)$$

frequency in *tk*, *Nt* = total number of tweets, and *nl* = number tweets in which *wl* appears.

The similarity score between the target tweet *t* and another tweet *tk* is defined by:

$$Sim(t, t_k) = \frac{t \cdot t_k}{||t|| \cdot ||t_k||}$$

We now select the top-*Y* tweets most similar to the target tweet *t*, denoted by *TopY Tweets*(*t*). Let *Hashtags*(*tk*) denote the set of hashtags in tweet *tk*. We derive a second set of candidate hashtags *HTofTweets*(*t*) from *TopY Tweets*(*t*) as follows.

$$HTofTweets(t) = \cup_{t_k \in TopYTweets(t)} Hashtags(t_k)$$

Ranking candidate hashtags. The candidate hashtags to be recommended for the target user *u* and tweet *t* can be obtained by the union of hashtags from top-*X* similar users and top-*Y* similar tweets.

$$SuggestedHashtags(u,t) = HTofUsers(u) \cup HTofTweets(t)$$

After that, hashtags in *SuggestedHashtags*(*u; t*) are ranked by frequency. The hashtag frequency is defined by adding the number of times the hashtag is used by top-*X* users with the number of times it appears in top-*Y* tweets. Finally, the top ranked hashtags are finally recommended to the user *u*.

To evaluate our hashtag recommendation method, we conduct experiments using the tweets generated by Singapore users in November and December of 2011. Tweets that do not contain hashtags are removed from the dataset. The remaining dataset in November contains 2,264,801 tweets and 37,617 unique users and is used as training data. To evaluate the recommendation results, we randomlyselected 5606 original tweets from the December data with authors in the training set. These tweets form our target tweet set. The hashtags actually used in the target tweets serve as the ground truth. Since the hashtags to be recommended are from November, we expect that they are still relatively fresh.

Since other previous methods recommend hashtags purely based on similar tweets, our experiment varies the number of similar users (i.e., *X*) used in our method. When *X* = 0, our method will recommend only hashtags from similar tweets. We also want to evaluate the dfferent number of similar tweets *Y* usedin recommendation.

For each target user-tweet pair, we consider the top *five* and top *ten* recommended hashtags and measure the performance of our method using hit rate as defined below.

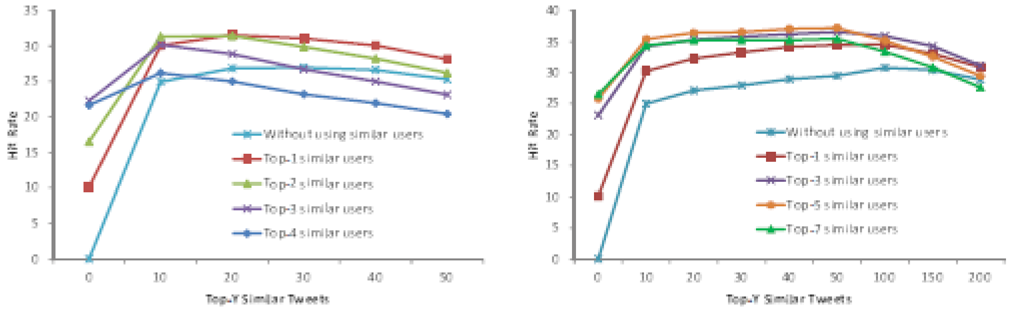
Hit Rate =Number of Hits/Number of Target User-Tweet Pairs (1)

A hit occurs when the recommended hashtags for a target tweet *t* include at least one of the ground truth hashtags. Although multiple hashtags may be used in a target tweet, such cases are rare. Hence, it is reasonable to use the above hit rate measure.

We use Apache Lucene2 to derive the similarity scores and retrieve the hashtags of the top-*X* similar users and hashtags of the top-*Y* similar tweets as Lucene is very e±cient in such computation and retrieval.

Figure (a) below shows the hit rate (in percentage) of top five recommended hashtags. We vary the number of top similar tweets *Y* used from 0 to 50, and measure the performance of our method with top *X* = 0 to 4 similar users. The figure shows that as we increase the number of similar tweets from 0 to 10, the hit rate improves significantly. The improvement beyond 10 similar tweets is however very small or even negative. We can also observe that considering top 1 to 3 similar users can help to further improve the hit rate when the number of similar tweets are small, i.e., 10 and 20. The improvement percentage of recommendation using top 1 similar user over recommendation without similar user at *Y* = 10 is about 20%. Our method performs best with hit rate = 31.56% when *X* = 1 and *Y* = 20.

Figure (b) below shows the hit rate (in percentage) of top ten recommended hashtags. We vary the number of top similar tweets *Y* used from 0 to 200, and measure the performance of our method with top *X* = 0*;* 1*;* 3*;* 5 and 7 similar users. On the whole, the hit rate has improved as we recommend more hashtags. Again, most significant improvement in hit rate occurs between *Y* = 0 and *Y* = 10. Beyond *Y* = 10, the improvement is small. On the other hand, using similar users is almost always better than not using similar users. The improvement margin of recommendation using top 1 similar user over recommendation without similar user at *Y* = 10, i.e., 21%, is similar to that observed for top 5 Recommended hashtags. This time, our method performs best with hit rate = 37.19% when *X* = 5 and *Y* = 50.



(a) Hit Rate (%) for Five Recommended Hashtags

(b) Hit Rate (%) for Ten Recommended Hashtags

Figure 15

Conclusion

As micro-blogging sites, like Twitter, continue to grow in popularity, we are presented with the problem of how to effectively categorize and search for posts. To cope with the volume of information shared daily, twiter has introduced *hashtags*, keywords prefaced with “#”, to help users categorize and search for tweets. To help Twitter users more easily incorporate hashtags into their posts, we have proposed an automatic hashtag prediction tool that, when given a keyword, will return a short list of relevant hashtags as suggestions.

Our task of hashtag prediction is divided into the various categories due to the viral and dense nature of these hashtags like geological location of a user or temporal and viral nature etc.

Further users can select a particular hashtag and can see relevant textual posts and images from Instagram on a same platform i.e. our web portal. From our current work, we are hopeful about the success of our hashtag suggestion tool. This tool is important as most tweets do not carry hashtags and most hashtags do not have long life span. We also observe that the usage patterns are stable over the period. The method that considers both target user preferences and target tweet content. Given a user and a tweet, our method selects the top most similar users and top most similar tweets. Hashtags are then selected from the most similar tweets and users and assigned some ranking scores. Experiment results show that using user preferences and tweet content will give us better recommendation than just using tweet content alone.

We can further divide hashtags into different categories, e.g., by freshness or by topic, and study their recommendation accuracies. So far, the methods we have mentioned are based on simple collaborative filtering. More sophisticated methods such as matrix factorization can also be used in the future.

References, IEEE Format

Research Papers

[1] Roman Dovgopol and Matt Nohelty, ”Twitter Hash tag Recommendation” in *University of Minnesota.*

[2] Daniel Preot¸iuc-Pietro and Trevor Cohn(et al.),” A temporal model of text periodicities using Gaussian Processes” in *University of Sheffield in* Portobello Street Sheffield.

[3] Maximilian Jenders, Gjergji Kasneci and Felix Naumann(et al.) “Analyzing and Predicting Viral Tweets” *in Hasso Plattner Institute in Potsdam,* Germany.

[4] Wei Feng and Jianyong Wang ,” Connecting Tweets to Explicit Topics” in *Tsinghua University in* Beijing, China.

[5] Romain Deveaud and Florian Boudin,” Effective Tweet Contextualization with Hashtags Performance Prediction and Multi-Document Summarization” in *University of Avignon.*

[6] Fréderic Godin (et al.), “Using Topic Models for Twitter Hashtag Recommendation” in *Ghent University, Ghent, Belgium.*

[7] Oren Tsur and Ari Rappoport, “Content based Prediction of the Spread of Ideas in Microblogging Communities” in *The Hebrew University.*

[8] Su Mon Kywe (et al.), “On Recommending Hashtags in Twitter Networks” in *Singapore Management University.*

[9] Shoubin Kong (et al.), “On the Real-time Prediction Problems of Bursting Hashtags in Twitter”.

[10] Shoubin Kong(et al.) , “Predicting Bursts and Popularity of Hashtags in Real-Time” in *Tsinghua University.*

[11] Siddharth Bora(et al.), “On the Role of Conductance, Geography and Topology in Predicting Hashtag Virality” in *Indian Institute of Technology, Delhi.*

[12] JasonWeston(et al.) ,” Semantic Embeddings from Hashtags” .

[13] Krishna Y. Kamath and James Caverlee ,” Spatio-Temporal Meme Prediction:Learning What Hashtags Will Be Popular Where” in *Texas A&M University*.

[14] Allie Mazzia and James Juett, “Suggesting Hashtags on Twitter” in *University of Michigan.*

[15] Bo Han (et al.), “Text-Based Twitter User Geolocation Prediction” in *The University of Melbourne.*

[16] Lisa Posch(et al.),” Meaning as Collective Use: Predicting Semantic Hashtag Categories on Twitter” in *Graz University of Technology.*

[17] Zongyang Ma (et al.),” On Predicting the Popularity of Newly Emerging Hashtags in Twitter”
in *Nanyang Technological University.*

[18] Tuan Anh Hoang (et al.) , “On Recommending Hashtags in Twitter Networks” in *Singapore Management University.*

[19] Florian Kunneman(et al.), “The (Un)Predictability of Emotional Hashtags in Twitter” in *Radboud University Nijmegen.*

[20] Qi Zhang (et al.) ,” Time-aware Personalized Hashtag Recommendation on Social Media” in *Fudan University*.

[21] Peng Zhang (et al.),” On Predicting Twitter Trend: Important Factors and Models” in *Arizona State University*.

[22] Tianxi Li and Yu Wu ,” Twitter Hash Tag Prediction Algorithm” in *Stanford University*.