論文 / 著書情報
Article / Book Information

題目(和文)	ハッシュタグコミュニティ検出に基づくニュース・トピック関連ソーシャルメディアマイニング
Title(English)	News-Topic Oriented Social Media Mining Based on Hashtag Community Detection
著者(和文)	肖峰
Author(English)	Feng Xiao
出典(和文)	学位:博士(学術), 学位授与機関:東京工業大学, 報告番号:甲第9601号, 授与年月日:2014年6月30日, 学位の種別:課程博士, 審査員:徳田　雄洋,米崎　直樹,佐伯　元司,権藤　克彦,西崎　真也
Citation(English)	Degree:Doctor (Academic), Conferring organization: Tokyo Institute of Technology, Report number:甲第9601号, Conferred date:2014/6/30, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

News-Topic Oriented Social Media Mining Based on Hashtag Community Detection

Thesis submitted

for the

Degree of Doctor of Philosophy

Department of Computer Science
Graduate School of Information Science and Engineering
Tokyo Institute of Technology, Japan

Xiao Feng

Advisor: Prof. Takehiro Tokuda

2014

Contents

1

Introduction

News articles, as a traditional medium for distributing information all over the world, have been increasingly impacted by a new way of information delivery called social media. Social networking services, such as Twitter [1], Facebook [2], and Digg [3], provide plenty of ways for users to share information with others and affect the way of news spreading and news consumption of users. However, there are some problems for users to share contents with others and get valuable information from social networking services, which motivate us to propose new methods to solve these problems.

1.1 Motivation

Microblogging [4] is a new way for users to collect and provide information on the Web. One of the most famous microblogging services is Twitter, which attracts over 200 million active users creating over 400 million messages, called tweets, everyday [5]. Most of these tweets often concern topics of headline news or persistent news [6], making Twitter an important data source for news.

Functions provided by Twitter help users easily share news with others. A user could follow other users who have the same interest with him. He can repost interesting tweets, called retweet, when he would like to share them with his followers. Mention and reply, prefixing user name with @ symbol, are used for purposes such as direct communication with others, or referring to users who are relevant. Hashtags (the # symbol prefixed to a short character string) are widely used by Twitter users to categorize and joint tweets together for a certain topic, and virtual user communities defined by hashtags are created to exchange opinions/interests/comments with others in these communities [7][8]. We refer to a group of users who use the same hashtag in their tweets as hashtag community, and these users are members of the hashtag community.

Although Twitter is a good platform to share news, a recent survey conducted on ordinary users reveals that 85% of users would do a specific keyword search for their interested news topics using news search engine while getting news from social media like Twitter is supplemental for news consumption [9]. In this

thesis, we are interested in the scenario that after a user does a specific keyword search for his interested news topics, which Twitter users would be worth following for him to get tweets from if he wants to get supplemental information about these news topics?

Suppose, for example, a user who is interested in U.S. presidential election sends a query “Obama” to a news search engine to get news articles containing the query word. After finish reading some news articles about the news topic, he wants to know what other users are saying for this news topic in Twitter. However, it is difficult for him to get tweets related to this news topic by sending the same query to Twitter. That is because the tweet has the length limitation of 140 characters. Tweets related to the news topic do not necessarily contain the query word. Another option is to follow other Twitter users to get tweets related to this news topic. However, it is still difficult to find Twitter users worth following. Tweets posted by some Twitter users are valuable and more likely to interest others while tweets posted by other Twitter users, even related to the news topic, are unattractive and more likely to be ignored. Following those users whose tweets are paid close attention to by others would help to get attractive contents and understand why some opinions are popular for the news topic. However, measuring the value of tweets posted by a Twitter user is a non-trivial task. Also, Twitter users could post tweets freely while it is hard to know whether contents of these tweets are reliable or not. Following Twitter users who have high authority on the news topic (e.g. a political journalist reporting the presidential election) would help us get more reliable information. However, for ordinary users, especially users who are novices for the news topic, professionals of the news topic might be unknown to them.

The purpose of our research is to help ordinary users find influential Twitter users worth following for a news topic after they search for the news topic by a keyword (we refer to the keyword as target word in this thesis). Two new methods are proposed to find two types of influential Twitter users for the news topic in which ordinary users are interested. One type of influential Twitter user often posts tweets containing valuable information for the news topic. Their tweets are more likely to interest others (e.g. get retweeted). We refer to this type of Twitter user as content-based influential Twitter user. Following this type of Twitter user could get tweets which are very attractive and help us understand why some opinions for the news topic are popular. The other type of influential Twitter user has high authority on the news topic so that other Twitter users would be more likely to communicate (e.g. mention) with him. We refer to this type of Twitter user as authority-based influential Twitter user. Following this type of Twitter user could get tweets which are reliable because these users have high authority on the news topic. For the news topic of U.S. presidential election, one good example of content-based influential Twitter user is “@PatDollard”, a famous filmmaker in the U.S. who often shares his opinions about the election and attracts many others, especially Republican supporters. One good example of authority-based influential Twitter user is “@andersoncooper”, the Twitter account of a famous American journalist. His tweets for the presidential election are reliable due to his special social position. To find these two types of influential

Twitter users for a news topic, tweets related to the news topic are needed. However, due to the length limitation of tweets, ordinary Information Retrieval methods are no longer effective in collecting tweets related to the news topic. In this thesis, we collect tweets related to the news topic by detecting hashtags which are relevant to the news topic. A hashtag which is often used to share contents about the news topic are considered to be relevant to the news topic. We refer to this hashtag as a news-topic-related hashtag, and the hashtag community defined by this hashtag as news-topic-related hashtag community. Tweets containing news-topic-related hashtags are taken as tweets related to the news topic. Two types of influential Twitter users could be found from users who posted these tweets.

To detect hashtags which are relevant to the news topic, relevance between the news topic and hashtag is measured by the cosine similarity where news topic and hashtag are both represented by vectors. For representing news topic and hashtag, two new methods are proposed to detect characteristic co-occurrence word with the query word or hashtag. Characteristic co-occurrence words are words which provide important information about a certain topic. By using our newly proposed methods, we can detect characteristic co-occurrence words from news articles and tweets to create the news topic vector and hashtag vector. Since Twitter users often use hashtags to categorize and joint tweets for a certain topic, those hashtags highly relevant to the news topic could be recommended to users who want to use hashtags in tweets to joint conversations about the news topic.

1.2 Organization of the Thesis

The organization of the rest of this thesis is as follows. We present related work in the next chapter. In Chapter 3, we introduce a newly proposed method called Probabilistic Inside-Outside Log method (PIOLog) to detect characteristic co-occurrence words with the target word from news topics related to the target word. In Chapter 4, we describe our method to recommend hashtags for news topics in which users are interested and searched by the target word. In Chapter 5, we introduce two new methods to find content-based and authority-based influential Twitter users for a news topic related to the target word. This could help ordinary users find valuable and reliable tweets about the news topic posted by these two types of influential Twitter users. Experimental results and evaluations are described in Chapter 6. In Chapter 7, we make the conclusion with directions for future research.

2

Background and Related Work

In this chapter, background of our researches and related methods are discussed. We also show problems existing in related methods that we are trying to solve in our researches.

2.1 Background

Currently, microblogging provides a new way for users to get information about news topics. Users could find what's happening in the world and the latest evolvement of events from social media which could be even earlier than from news media. Also, microblogging provides not only a platform for sharing news, but also a platform for sharing opinions of users about news topics.

One of the most famous microblogging services is Twitter, which attracts many users in the world sharing news related tweets everyday. However, it is hard for ordinary users to get useful information about his interested news topics and share his opinions/interests/comments widely with others. Twitter users often choose to follow other users who often post tweets about news topics in which they are interested to get information from them. However, it is difficult to choose proper Twitter users to follow since ordinary users have no idea whose tweets are valuable for the news topic and more likely to interest others. Also they do not know whose tweets are trustable for the news topic since Twitter users can near-freely post tweets. Other researchers proposed many methods to find influential Twitter users. However, these methods are either topic-independent, or not considering different relations of users, which are unsuitable in our research.

Twitter provides a function of hashtag for users to join tweets for the same topic. This function could help us collect tweets related to the same news topic to find users worth following. It could also help Twitter users share their tweets with others widely without following each other. However, it is difficult for ordinary users to use proper hashtags in their tweets. Existing methods tried to recommend hashtags for user's newly input tweet. However, the length limitation of tweet seriously affects the effectiveness of existing methods, making

them unsuitable for recommending hashtags. Also, existing term weighting methods (e.g. TF-IDF), which are often used as basic component for detecting topic related hashtags are no longer effective for tweets.

As basic component of our research, detecting characteristic co-occurrence word with the target word provided by ordinary users for a news topic could greatly help to find relevant information about the news topic. There are plenty of methods to detect word co-occurrence from documents. Among these related methods, symmetric method is not suitable for characteristic co-occurrence word detection because one word is a characteristic co-occurrence word with the other word or not depends on the news topic. Other asymmetric methods are often used to detect word collocation, which is a different purpose compared with ours.

2.2 Related Work

Our researches presented here relate to three research fields: co-occurrence word detection, tag recommendation, and finding influential Twitter users.

2.2.1 Co-Occurrence Word Detection

To detect meaningful co-occurrence words, some methods have been proposed and they are mainly classified into two types, symmetric method and asymmetric method. We present one method which is often used in each type.

Jaccard coefficient [10] is one of commonly used symmetric methods for detecting pairs of words co-occurring frequently with each other. It has symmetry property, which means, if a word w_1 co-occurred with another word w_2, the opposite is also true. However, it is not suitable for detecting characteristic co-occurrence words. In the case that w_1 appears in news articles including w_2 and it does not often appear in other news articles, w_1 will be a characteristic co-occurrence word with w_2. On the other hand, if w_2 often appears not only in news articles including w_1 but also in many other news articles, w_2 might not be a characteristic co-occurrence word with w_1. We think methods for detecting characteristic co-occurrence words should be asymmetric to reflect this idea.

Additionally, since Jaccard coefficient is calculated by dividing the number of news articles including both w_1 and w_2 by the number of news articles including both/either w_1 and/or w_2, w_1 and w_2 should co-occur in many news articles which contain w_1 or w_2 to get high Jaccard coefficient score. However, w_1 does not have to appear in many news articles including w_2 to be a characteristic co-occurrence word. Whether w_1 often appears in the others (news articles not including w_2) or not is more important for judging if w_1 is a characteristic co-occurrence word with w_2. Our method could also reflect this idea while Jaccard coefficient can't.

Log Likelihood Ratio (LLR) [10] is another method for word co-occurrence detection which is an asymmetric method. This method set two hypotheses as null hypothesis and alternative hypothesis, assuming that word w_1 is independent

(null hypothesis) or dependent (alternative hypothesis) on the other word w_2, and test these two hypotheses to decide whether we should accept the null hypothesis or reject it. However, LLR is often used to detect word collocation, which is an expression consisting of two or more words that correspond to some conventional way of saying things. However, characteristic co-occurrence words co-occur with the target word due to a specific topic, not as a grammar unit constantly. Additionally, LLR is appropriate for detecting sparse word collocation which co-occurs in a small number of documents. Although this is an advantage of LLR to detect word collocation, it is a big disadvantage to detect characteristic co-occurrence word since characteristic co-occurrence words should co-occur with the target word in many documents.

2.2.2 Tag Recommendation

Hashtag in Twitter is one special type of a more general concept, called tag, which is an important feature for many social networking services. People could create tags with few taxonomic constraints to categorize resources for later browsing, or to mark resources for searching. Many approaches for tag recommendation in social networking services have been proposed recently. They are mainly classified into two classes.

One class of these approaches focuses on the relationship between tags and their associated resources, and recommends tags for a newly added resource. One application of this class is the tag recommendation system for weblog. Figure 1 gives an example of system structure about how to recommend tags for weblog posts. Brooks et al. [11] tried to select words in blog posts that have high TF-IDF scores and used as tags. They found that those tags are more representative than human-assigned ones. Mishne [12] and Sood et al. [13] recommended tags for a new blog post by recommending tags in those old blog posts which have high cosine similarity with the new one. These approaches recommended tags from similar weblog posts by using techniques from Information Retrieval (e.g. TF-IDF). However, these methods are no longer effective in hashtag recommendation because TF-IDF reduces the chance of relevant tweets to be selected since the tweet length is limited and their contents have less information than blog posts [14].

Other approaches exploit tag co-occurrence patterns through a history of tag assignments in a collaborative tagging environment when the resource with which the tag was associated is hard to retrieve, such as audio, video, and image. Figure 2 gives an example of system structure about how to recommend tags for photos in Flickr [15]. Sigurbjornsson et al. [16] recommended tags for each user-defined tag about photos based on tag co-occurrence in Flickr. Wartena et al. [17] proposed another approach to calculate the similarity between tag co-occurrence distribution and the user profile. Tags with high similarity are recommended to the user. Belem et al. [18] extended tag co-occurrence exploiting and consider about terms extracted from other textual features such as title and description. All these approaches are based on two assumptions: tags

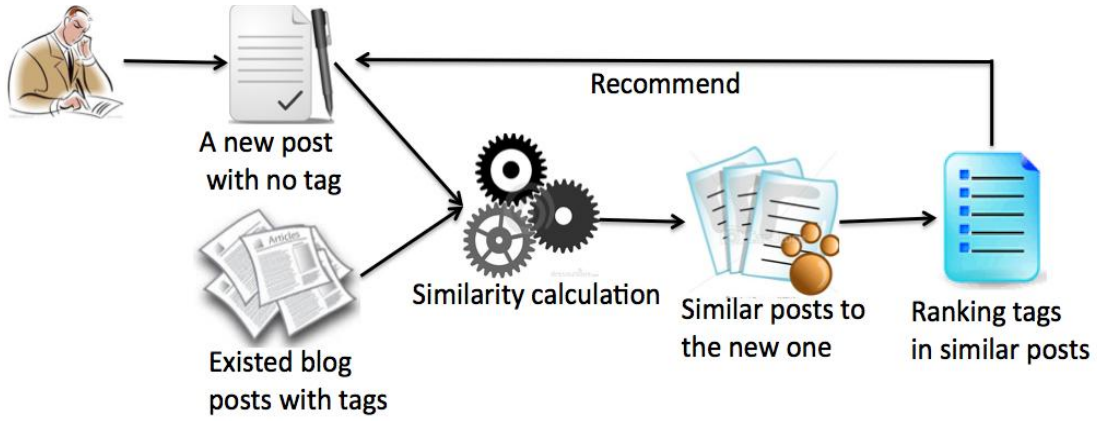


Figure 1. An example of system structure for tag recommendation in weblog

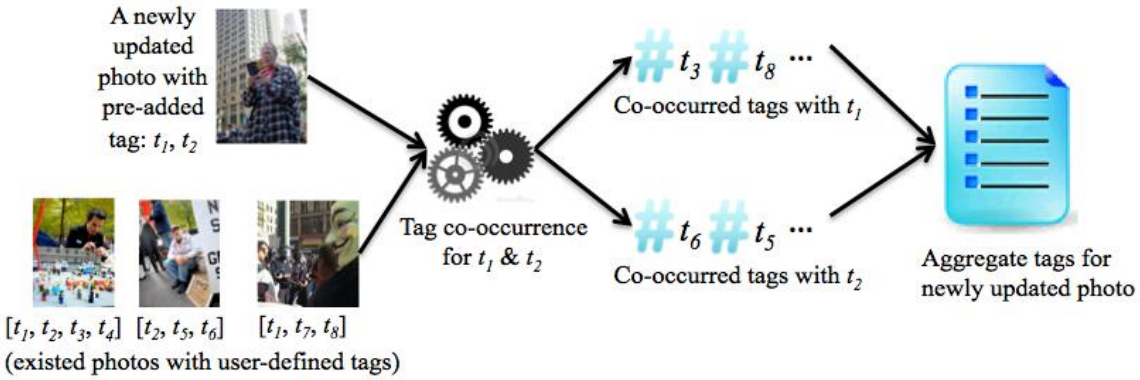


Figure 2. An example of system structure for tag recommendation in Flickr

are assigned to resources beforehand, and most resources have two or more tags. For example, Flickr allows its users to add 75 tags per photo at most. In YouTube the total length of the tag list is limited to 500-character for each video. However, most of tweets in Twitter only contain one or even no hashtag. For example, in all news-related tweets collected on December 20th 2011, 88.6% of tweets contain one or no hashtag. Exploiting tag co-occurrence in tweets becomes impossible due to the limited number of tweets containing two or more hashtags.

Recently, researchers found that hashtags in Twitter play a different role compared to tags in other social networking services. Huang et al. [7] compared user tagging behaviors between Twitter and Delicious. They found out that hashtags in Twitter are used to join discussions on existing topics while in Delicious tags are used to re-access resources. Our approach is based on the conversational nature of hashtags and tries to recommend hashtags to help users join the conversation about the news topic so that users do not need to be "exposed" to too many hashtags.

Approaches for hashtag retrieval/recommendation in Twitter have been proposed while there are still some problems. Lehmann et al. [19] classified hashtags in four classes based on their activity profiles over time. However, our purpose is to detect the relevance between hashtags and a news topic based on the content they relate to. Popularity variation of hashtag in its activity profile could not reflect this relevance. Weng et al. [20] proposed methods for modeling the interestingness of hashtags by studying how hashtags are discussed within and across communities, but they do not correlate hashtags with topics in which users are interested. Correa et al. [21] proposed a new approach for recommending tags for other social networking services such as Flickr and YouTube, using hashtags and terms in tweets. Our approach is different because we correlate Twitter with traditional news media, not other social networking services. Efron [22] and Wagner C. et al. [23] proposed new approaches to retrieve useful hashtags after a keyword is given. However, one keyword may relate to more than one topic and all hashtags related to different topics would be mixed together. Also, ranking hashtags based on their in-degree in [23] would make some general hashtags (e.g. #tech) be ranked higher, which is not helpful. Zangerle et al. [24] proposed a method to recommend hashtags for users' input contents by calculating similarities between newly input tweet and old tweets based on TF-IDF [48]. Hashtags which frequently appear in old tweets with high similarities are recommended. Mazzia et al. [25] proposed use of Bayesian model to estimate probabilities of many hashtags by observing newly input tweet contents. Kywe et al. [26] considered not only newly input tweet contents, but also similarity among users for hashtag recommendation. Their experimental results showed that this method yields better performance than recommendation based only on tweet contents. Although these researches seem to be reasonable, there are still some problems. Firstly, similarities between tweets in researches above simply rely on common words in tweets. However, due to the length limitation, two tweets may refer to the same topic while both of them share no common word. Secondly, TF-IDF is no longer a good choice for short text like

tweets [14]. Due to the huge number of tweets, the IDF part would dominate the final score, assigning too large a score to the word which appears scarcely (e.g. misspelling). Lastly, researches above rely on newly input tweet contents while our purpose is to detect hashtags relevant to a news topic searched by the target word.

Other approaches interweaving traditional news media with social networking services have also been proposed for Topic Detection and Tracking [27, 28], news recommendation [29], and user profile construction [30]. To the best of our knowledge, our approach is the first one trying to recommend hashtags for news topics in which users are interested.

2.2.3 Finding Influential Twitter Users

Finding influential users in social networking services has been focused by researchers recently. Many methods have been proposed for measuring user's influence in Twitter. These methods could be mainly classified into two classes based on user's relation type.

One class of these methods measures user's influence based on user's follow relation. The most intuitive way to measure the user's influence is to count the number of followers a user has. It is based on the assumption that more followers he has, more impact he could make on other users. Another similar measure uses the ratio of the number of user's followers to the number of users he follows. However, follow relation is not a good indicator for user's influence. A Twitter user who has many followers is not necessarily influential [31]. Users could follow a large number of other users, wishing them to follow back for courtesy. Moreover, only considering follow relation ignores the user's interaction with other users. The user whose tweets are ignored by most of his followers has less influence on the others even if many users follow him.

The other class of method measures user's influence based on his interactive activities like mention, reply, and retweet. Cha et al. [31] analyzed three influence measures as mention, retweet, and number of follower independently. They found that the number of user's follower reveals little about his influence. Retweet represents the value of tweet contents, and mention represents user's name value. Other researches combine different user activities to measure the influence. Leavitt et al. [32] defined Twitter user's influence as the potential of a user's action to initiate a further action by other users. They measured user's influence by the ratio of attentions he received (being mentioned, replied, and retweeted) to the number of tweets he posted. Anger and Kittl [33] proposed a new influence measure based on the ratio of user's tweets being retweeted and the ratio of user's followers retweeting his tweets or mentioning him. Hajian and White [34] proposed Influence Rank, a variant of PageRank [51], to quantify user's influence in Twitter. The difference between Influence Rank and PageRank is the way in which the teleportation vector is defined. The teleportation vector in Influence Rank is calculated based on a combination of user's follow, like, comment and retweet activities. Romero et al. [35] proposed

another Influence-Passivity algorithm to measure the influence and passivity of Twitter users based on retweet activity. They proposed methods to define two transition matrices to measure the amount of influence each user accepts/rejects from others. Then HITS algorithm [36] is applied to these two transition matrices to determine the influence of each user (hub score in HITS). Although researches presented above seem to be reasonable, an influential Twitter user in general might not be influential for a specific news topic searched by the target word. Our approach could find those Twitter users who are influential for the news topic in which ordinary users are interested. Also, retweet and mention are used as the same relation type to build user relations in these researches while different purposes of these activities are ignored. We take this difference into account and propose methods to find two types of influential Twitter users based on retweet and mention activities respectively.

Finding topic related influential Twitter users has also been explored. Ye and Wu[37], Bigonha et al. [38] found influential Twitter users for a manually selected topic (Michael Jackson's death and soda brands) based on user's activities like reply, and retweet. However, they ignore the link structure among users. A user should be more influential if he is retweeted/mentioned by other influential Twitter users rather than users with less influence. Noro et al. [39] proposed a new approach to find influential Twitter users related to a query word. However, one query word might correspond to multiple topics and influential Twitter users for different topics would get mixed together. Weng et al. [40] found high follow reciprocity among Singapore Twitter users and proposed TwitterRank method to find influential Twitter users for topics based on users' follow relations. They defined a new transition matrix with teleportation vector, taking into account the number of tweets posted and the topical similarity between users. However, results from [31] contradict the high follow reciprocity after analyzing near-complete data from Twitter. Also, the definition of topic in TwitterRank is different from the definition in our methods. The topic from TwitterRank is distilled by Latent Dirichlet Allocation [41] as a distribution over a fixed vocabulary. Our news topic is defined as a group of news articles published in a period of time (for example: one day) reporting about the same recent event in the world. Cano et al. [42] also proposed Topic-Entity PageRank to find influential Twitter users for both topic and entity. Tweets are categorized into predefined topics by OpenCalais [43]. Then a transition matrix is defined for each topic based on retweet activity. PageRank algorithm [51] is applied to this transition matrix to find influential Twitter users for the topic. However, topics from Topic-Entity PageRank are predefined while our news topics are automatically clustered from news articles. Also, an influential Twitter user for one topic from OpenCalais, for example Politics, might not be always influential for all political issues in the world.

In our research, we measure the relevance between hashtags and news topics searched by the target word. Then hashtag communities defined by hashtags which are relevant to the news topic are created. Content-based and authority-based influential Twitter users for this news topic could be found from these hashtag communities based on user's retweet and mention activities.

3

Detection of Characteristic Co-Occurrence Words from News Articles

In this chapter, a newly proposed method for detecting characteristic co-occurrence words with the target word provided by user is introduced. Target word is the query word provided by the user searching for his interested news topic. Characteristic co-occurrence words are words co-occurring with the target word in news articles for a specific news topic, providing important information for the news topic related to the target word. Detecting characteristic co-occurrence words with the target word could help us quickly understand the contents of the news topic.

3.1 Overview

Rapid growth of the Internet technology has let us access a variety of information easily. Especially, thousands of news articles are provided by various news sites every day. We can quickly find out what has happened or what is going on in the world. However, it is difficult for us to do it just by searching for news articles by keywords (e.g. Google News [44]) or topics (e.g. New York Times Topics [45], Yahoo! News Topics [46]), or following news directories for classifying news articles [47]. People could better understand the topics if we provide what words are linked to the topic.

Suppose, for example, we are interested in a topic on a car maker Toyota. Although we can get news articles including the word “Toyota” by sending the query word “Toyota” to a news search engine, we still need to find out what topics the obtained news articles are reporting. It is a difficult task if the number of news articles got from the search engine is large. If we notice that words “recall”, “accident” and “lawsuit” appear in news articles including “Toyota”, we could guess some problems happened to Toyota cars and Toyota

launched recall. Detecting words co-occurring with a word of interest (target word) will help us find out such information.

In this chapter, we will present our newly proposed methods for detecting characteristic co-occurrence words with a target word. In our method, all news articles published in a certain period of time are clustered into news topics. Then we divide these news articles into two groups: one is a group of news articles including the target word and belonging to the news topic related to the target word (in the case of the example described above, the target word is “Toyota”), and the other is a group of news articles not including the target word or not belong to the news topic related to the target word. Then we compute score of a word co-occurring with the target word in some news articles by counting the number of news articles including the co-occurring word for each of the news article groups. This method will help us find out characteristic co-occurrence words co-occurring with the target word for news topics.

Our characteristic co-occurrence word detection method is based on two assumptions:

- Characteristic co-occurrence word *w* should often co-occur with the target word *t* in news articles. We take it as the Inside Part.

- Characteristic co-occurrence word *w* should not always appear in news articles without the target word *t*. We take it as the Outside Part.

Based on these two assumptions, words often co-occur with the target word in news articles while being less likely to appear in news articles without the target word are taken as characteristic co-occurrence words. However, there are still some problems.

One of the problems is that all of the news articles including the target word *t* do not always deal with the same topic. In the case that there are more than one news topic related to the target word, we may not be able to detect characteristic co-occurrence words properly since co-occurrence words related to different topics will be mixed together.

Another problem is that there are some general words which often co-occur with the target word in news articles regardless of the news topic. For example, “Obama” often co-occur with “White House” and “administration” in news articles. However, if one user wants to search for some news about Obama and use “Obama” as the target word, “White House” and “administration” would give no information about recent news topics of Obama. These general words should be excluded since they often co-occur with the target word regardless of topics and they provide little information about the news topic related to “Obama”.

The organization of this chapter is as follows. We introduce Hierarchical Agglomerative Clustering method to cluster news articles into topics in the next section. In Section 3.3, we propose our method to detect characteristic co-occurrence words with the target word from news articles of the same news topic. Finally we show existing methods to detect word co-occurrence and give qualitative comparison with our newly proposed method.

3.2 Hierarchical Agglomerative Clustering

To solve these problems, clustering is used to group news articles into different news topics. A news topic is a group of news articles published in a period of time (for example: one day), reporting about the same recent event happened in the world. News articles which are reporting for the same news topic are hopefully put into the same topic while news articles relating to different topics get separated. For each news topic related to the target word, we detect characteristic co-occurrence words from them respectively without mixing words from different topics together. Also, news articles in news topics other than the news topic we focus on are treated as “Outside Part” regardless of existence of the target word. This will exclude general words which often co-occur with the target word if news articles including the target word are separated into two or more than two clusters.

Generally there are two types of clustering, partitional and hierarchical [10]. Partitional clustering iteratively partitions the data set into *k* clusters based on a distance function given a predefined *k* value specified by the user. However, partitional clustering is not suitable for clustering news articles in our dataset. One reason is that the value of *k*, the number of clusters in its final result, should be specified before clustering while it is difficult to know how many news topics existing in news articles collected. Also, news articles collected by us are from different news directories like politics, economy, local and so on. The number of news articles from these directories is skewed. News articles from some directory (e.g. local) are much less than from other directories and contents of these news articles are highly different from others. Since partitional clustering often use the mean of news article vectors as the centroid of cluster, outliers will cause negative affection to the result, resulting in undesirable clusters. Finally, partitional clustering is unstable. Choosing different *k* initial seeds will result in different clusters, which makes the result hard to interpret.

The other type of clustering is hierarchical clustering [10]. There are two main types of hierarchical clustering methods. Agglomerative (bottom-up) clustering merges the most similar pair of clusters in each step and finally all news articles are merged into a single cluster. Divisive (top-down) clustering split the whole data set into sub-clusters and each sub-cluster is recursively divided into smaller clusters until only a single news article in each cluster. Since news topics change constantly and dynamically on each day, hierarchical agglomerative clustering is more suitable and popular to cluster news articles.

Each news article is parsed into a bag of words appeared in this news article while word sequence and position are ignored. Then a news article is represented by a word vector that each dimension of this vector corresponds to a separate word appearing in this news article. The value for each

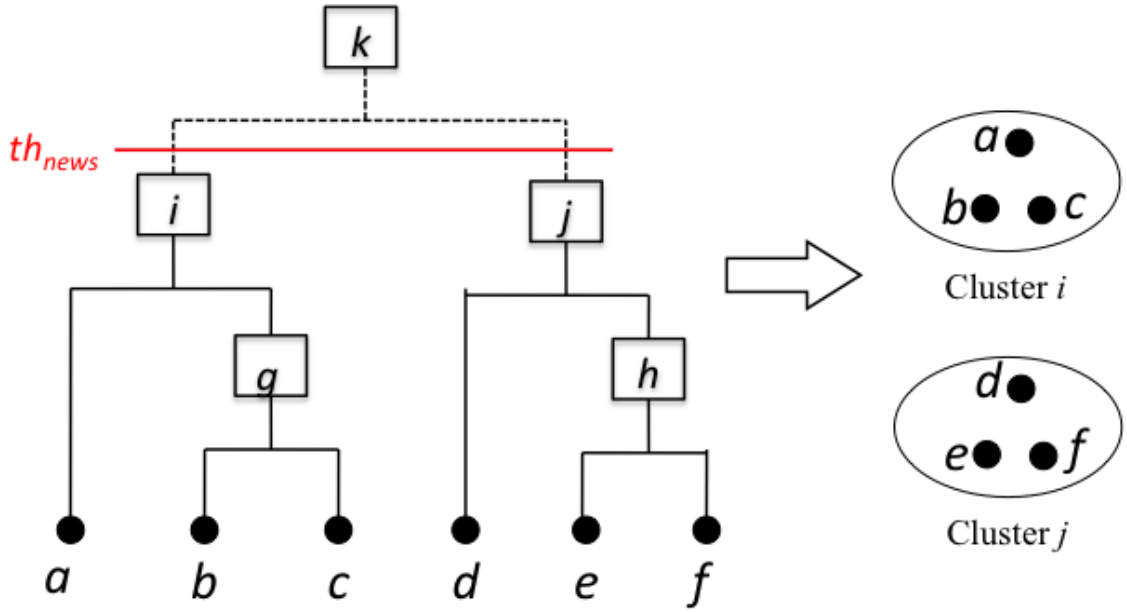


Figure 3. An example of Hierarchical Agglomerative Clustering (HAC)

dimension is calculated based on Term Frequency (TF) and Inverse Document Frequency (IDF) [48]. Then news articles are clustered by hierarchical agglomerative clustering as follows:

Hierarchical Agglomerative Clustering method

1. Make each news article in news article dataset D={doc_1, …, doc_m} as a single cluster and form m clusters C={c_1, …, c_m} where c_i={doc_i}.
2. Calculate all pair-wise similarity between clusters in C.
3. **Repeat**
4. Get the cluster pair whose pair-wise similarity is the maximum value for all cluster pairs.
5. Merge these two clusters to form a new cluster and calculate the centroid vector for representing this new cluster.
6. Calculate the similarity between other clusters and this new cluster.
7. **Until** maximum pair-wise similarity value is less than a predefined threshold (th_{news}), or |C| == 1.

Vector of a cluster is defined as centroid of all news article vectors in the cluster. The similarity of two clusters is calculated by cosine similarity based on Vector Space Model. Figure 3 gives an example of six news articles. At the

bottom of the figure, each news article is taken as a single cluster and there are six clusters {*a*, *b*, *c*, *d*, *e*, *f*}. For the next step, cluster *b* and *c* are merged to form a new cluster *g*. When iterating between line 3 and line 7 of the HAC method, we have fewer and fewer clusters until the maximum similarity among clusters is less than th_{news} or all news articles are clustered into the same cluster. At last, we get two clusters: cluster *i* = {*a*, *b*, *c*} and cluster *j* = {*d*, *e*, *f*}.

3.3 Probabilistic Inside-Outside Log Method

After grouping news articles into clusters by hierarchical agglomerative clustering in the former section, news clusters are taken as news topics related to the target word if at least half of their news articles in the news cluster contain the target word. Then for each news topic related to the target word, we detect characteristic co-occurrence word based on two assumptions proposed in Section 3.1. The whole procedure is shown in Figure 4.

Our Probabilistic Inside-Outside Log method (PIOLog) to detect characteristic co-occurrence word *w* with the target word *t* for a news topic *c* is as follows:

$$\mathrm{PIOLog}(w,t,c) = \log \frac{(1-s_p)\mathrm{P}(w \mid t \wedge c) + s_p}{(1-s_p)\mathrm{P}(w \mid \neg(t \wedge c)) + s_p} \tag{1}$$

$$\mathrm{P}(w \mid t \wedge c) = \frac{df\,(w \wedge t \wedge c)}{df\,(t \wedge c)} \tag{2}$$

$$\mathrm{P}(w \mid \neg(t \wedge c)) = \frac{df\,(w \wedge \neg(t \wedge c))}{df\,(\neg(t \wedge c))} = \frac{df\,(w) - df\,(w \wedge t \wedge c)}{N - df\,(t \wedge c)} \tag{3}$$

where *df* (*w*) is the number of news articles containing the word *w*. *df* (*w*∧*t*∧*c*) is the number of news articles containing both *w* and *t* in the news topic *c*. *df* (¬(*t*∧*c*)) gives the number of news articles not containing *t*, or not in the news topic *c*. *N* is the total number of news articles and s_p is a smoothing parameter ranging from 0 to 1 (Figure 5). *df* (*t*∧*c*) is taken as the Inside Part. Words which often co-occur with *t* in news articles of the topic *c* will get a large score in Equation (2), reflecting the idea of the first assumption. *N* - *df* (*t*∧*c*) is taken as the Outside Part. Words which are less likely to appear in news articles without *t* or unrelated to the topic *c* will get a small score in Equation (3), reflecting the idea of the second assumption. Words whose PIOLog scores calculated in Equation (1) are large would be more likely to be characteristic co-occurrence words with the target word for the news topic.

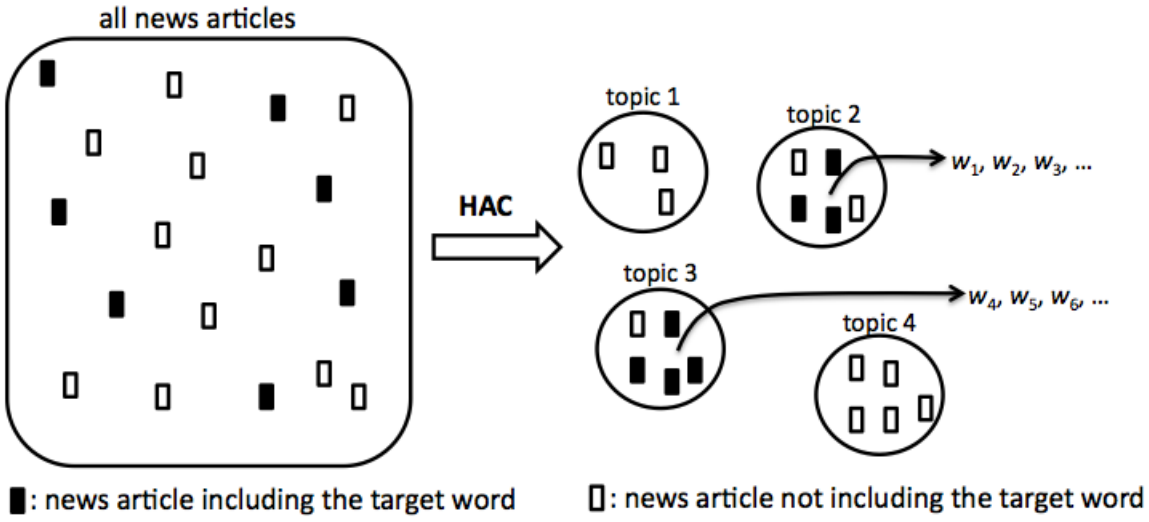


Figure 4. Procedure of characteristic co-occurrence word detection

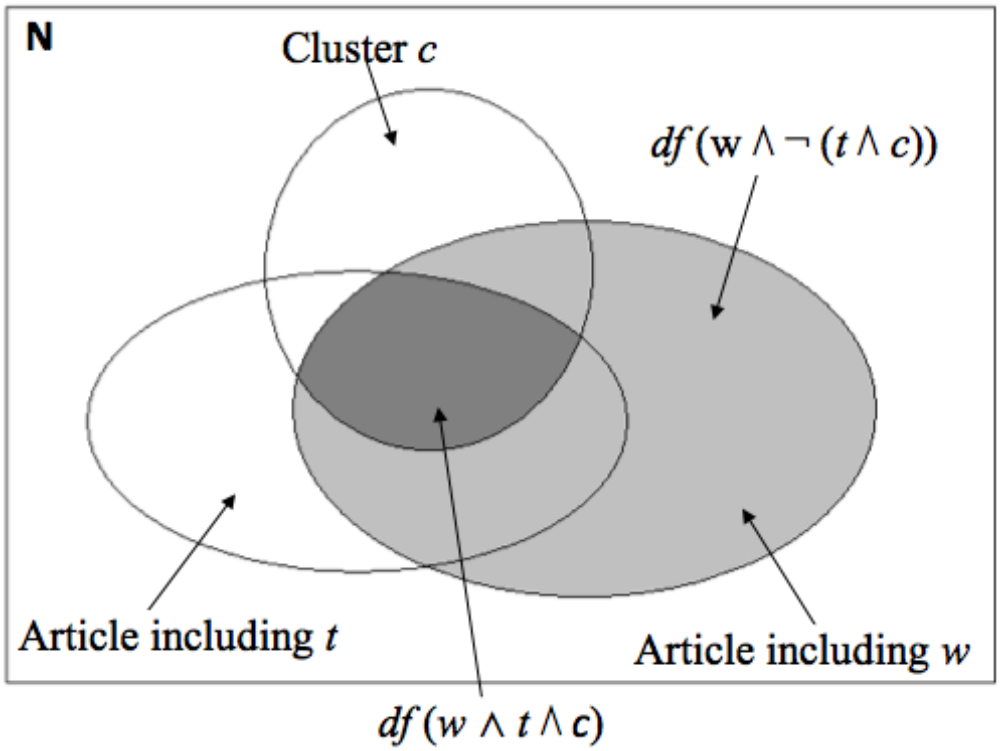


Figure 5. Probabilistic Inside-Outside Log method

3.4 Comparison with Related Methods

There are also many other methods to detect word co-occurrence from documents and widely used for many tasks. They are generally divided into two types: symmetric method and asymmetric method.

Symmetric method measures the co-occurrence between word w_1 and word w_2 in a reciprocal way, which means if w_1 is judged as a word co-occurring with w_2, the opposite is also true. The most representative and widely used symmetric method is the Jaccard method [10], which measures co-occurrence between words as below:

$$Jaccard(D(w_1), D(w_2)) = \frac{|D(w_1) \cap D(w_2)|}{|D(w_1) \cup D(w_2)|} \quad (4)$$

where $D(w_1)$ is the document set whose documents contain w_1. Jaccard method measures word co-occurrence as the size of intersection between $D(w_1)$ and $D(w_2)$ divided by the size of the union of these two sets.

However, Jaccard method is not suitable to detect characteristic co-occurrence words we focus on here. In the case that w_1 appears in news articles including w_2 and it does not often appear in the others, w_1 will be a characteristic co-occurrence word with w_2. On the other hand, if w_2 often appears not only in

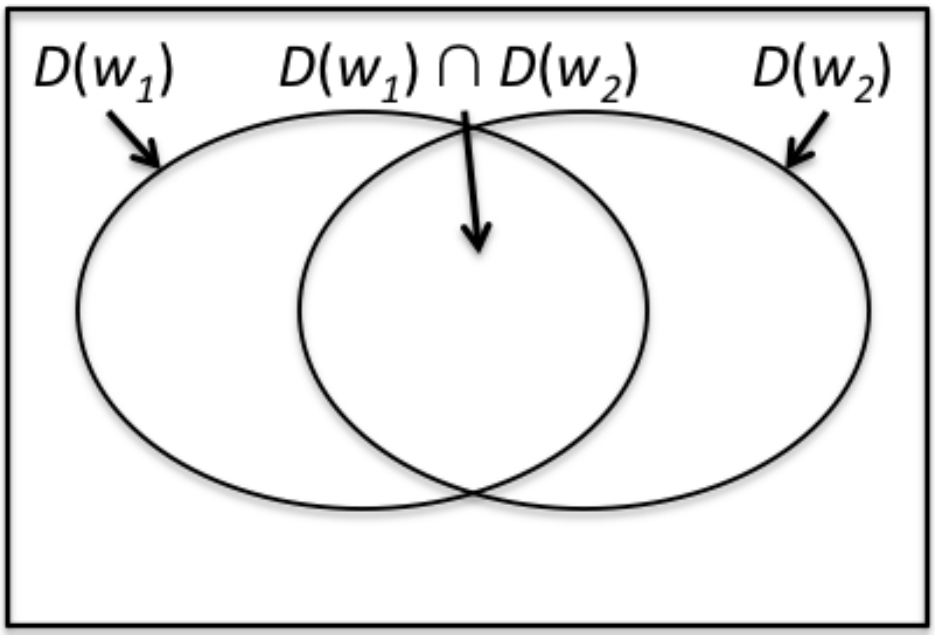


Figure 6. An example of Jaccard method

news articles including w_1 but also in the others, w_2 is not a characteristic co-occurrence word with w_1. Methods for detecting characteristic co-occurrence words should be asymmetric. Our method PIOLog could reflect this idea and Jaccard method can't.

Additionally, since Jaccard index is calculated by dividing the number of news articles including both w_1 and w_2 by the number of news articles including w_1 or w_2, w_1 and w_2 should co-occur in many news articles to get high Jaccard index score. However, w_1 does not have to appear in many news articles including w_2 to be a characteristic co-occurrence word. Whether w_1 often appears in the others (news articles not including w_2) or not is also important for judging if w_1 is a characteristic co-occurrence word with w_2. Our method also takes this idea into account.

Asymmetric method is another type of method to measure word relations from documents. Different with symmetric method, it measures co-occurrence in a nonreciprocal way, which means if word w_1 is judged as a word co-occurring with w_2, the opposite might not be the case. The most representative and widely used asymmetric method is the Log Likelihood Ratio [10], which is a method based on hypothesis testing. This method measures word co-occurrence by whether these two words occur together more often by chance or not. Log Likelihood Ratio method creates two hypotheses for detecting co-occurrence of w1 and w2 as follows:

$$\text{Null hypothesis} \mathrm{H}_0 : P(w_2 \mid w_1) = p = P(w_2 \mid \neg w_1) \tag{5}$$

$$\text{Alternative hypothesis} \mathrm{H}_1 : P(w_2 \mid w_1) = p_1 \neq p_2 = P(w_2 \mid \neg w_1) \tag{6}$$

$$p = \frac{c_2}{N} \quad p_1 = \frac{c_{12}}{c_1} \quad p_2 = \frac{c_2 - c_{12}}{N - c_1} \tag{7}$$

where the null hypothesis assumes that occurrence of w_2 is independent of w_1 and the alternative hypothesis hold a opposite assumption (occurrence of w_2 depend on occurrence of w_1). p, p_1, and p_2 are calculated by maximum likelihood estimate. Here, c_1, c_2, c_{12} are the number of documents containing w_1, w_2, and both w_1 and w_2 respectively. N is the total number of documents in the dataset. Assuming a binomial distribution b(k; n, x), the likelihood of having c_1, c_2, and c_{12} observed in the dataset under the null hypothesis is $L(H_0) = b(c_{12}; c_1, p)b(c_2 - c_{12}; N - c_1, p)$ while under the alternative hypothesis the likelihood should be $L(H_1) = b(c_{12}; c_1, p_1)b(c_2 - c_{12}; N - c_1, p_2)$. Then the Log Likelihood Ratio (LLR) is defined as follow:

$$\mathrm{LLR} = -2\log\lambda = -2\log\frac{\mathrm{L}(\mathrm{H}_0)}{\mathrm{L}(\mathrm{H}_1)} \tag{8}$$

The -2logλ is asymptotically χ^2 distributed. If the LLR score of two words is larger than the critical value for one degree of freedom under a confidence level of α (usually α = 0.005), we can confirm that these two words often co-occur with a confidence of 99.5%.

However, asymmetric methods like Log Likelihood Ratio are still not suitable for detecting characteristic co-occurrence word here because these methods are often used for detecting word collocation, which is a different purpose from ours. Word collocation is an expression consisting of two or more words that corresponds to some conventional way of saying things like the “hot dog”. Our characteristic co-occurrence word detection is trying to find two words often co-occur because both of them are highly related due to a specific news topic, not a conventional way of word using as a grammar unit constantly.

In this section, we give qualitative comparison between related methods and our newly proposed PIOLog method. Quantitative experimental results and comparisons are given in the evaluation section (Section 6.3).

4

News-Topic Oriented Hashtag Recommendation in Twitter

In this chapter, we present a new approach for recommending hashtags to the user who wants to join the conversation for a news topic by using hashtags in his tweets after he/she searches for the news topic by the target word. We use the PIOLog method introduced in the former chapter to detect characteristic co-occurrence words with the target word from news articles, and then create the news topic vector based on these detected words with their PIOLog scores. We also extend these two assumptions of PIOLog method in the former chapter and propose a new method to detect characteristic co-occurrence words with the hashtag from tweets. Hashtag vector is created based on these detected words. Similarities between news topics and hashtags are calculated. Hashtags having high similarity scores with the news topic get recommended. By using these recommended hashtags, users could share their tweets with other users who are interested in the same news topic either, helping them exchange their opinions more easily.

4.1 Overview

News articles, as a traditional medium for distributing information all over the world, have been increasingly impacted by a new way of information delivery called social media. Social networking services, such as Twitter, Facebook, and Digg, provide plenty of ways for users to share information with others. Most of news websites provide Tweet Button [49] in their Web pages to help readers easily share news articles with their followers in Twitter. Retweet function greatly accelerates the spreading speed of information and mention function helps Twitter users exchange information directly with others. Hashtags (the # symbol prefixed to a short string characters) are widely used to categorize and joint tweets together based on a certain topic and make your tweets more easily searchable by other users who have the same interest.

However, it is not easy for Twitter users to use hashtags in their tweets properly when they want to share contents or their opinions/interests/comments for news topics. For many news websites, they do not provide any hashtag in tweets after users click the Tweet Button on their Web pages, which means users' sharing could only be seen by their followers and might not reach far to the others. Other news websites add hashtags automatically while most of them are not for the purpose of helping users share their tweets or too unique. Some of news websites use their formal name (such as "#CNN") as the hashtag in every tweet after clicking the Tweet Button in their news Web pages no matter what the topic of the news article reports. Such kind of hashtag could only help these sites watch the information spreading in Twitter or promote reputation for advertising. Other news websites like Yahoo! Japan News provide hashtags such as "#yjfc_wall_street_protest" when users post tweets from news Web pages reporting protest in Wall Street while such kind of hashtag might only be used by Yahoo! Japan readers and is not widely used by other users.

It is also not easy for Twitter users to create/select proper hashtags by themselves. Users try to create hashtags which they took for granted that these hashtags should be widely used for topics while the truth might be just on the contrary. For example, "# Wall_Street_Protest" might be thought as a meaningful hashtag used in tweets talking about the protest in Wall Street, but we found that no one uses this hashtag in his tweets up to the point of writing this thesis. Users could search for some topic-related keywords and read all those responded tweets to find hashtags that relate to the topic. However, there might be too many hashtags contained in those responded tweets, relating to more than one topic, that users may have no idea which hashtag should be used. If all else fails, users may have to add the # symbol prefixed to each word in their tweets, wishing one of these hashtags could be the one which is widely used by others for the topic in Twitter. However, such a behavior would make tweets hard to read and impolite. The user may be taken as a Twitter spammer.

Our purpose is to recommend hashtags to users who want to join conversation in Twitter about a news topic by using hashtags after they search for the news topic by a target word. In our approach, news topics and hashtags are represented by vectors under the Vector Space Model [50]. News topic vector is created based on characteristic co-occurrence detection in Section 3. Detected words with their PIOLog scores are used to create the new topic vector for representing the news topic. We also extend PIOLog method for hashtags to detect informative words co-occurring with the hashtags in tweets. All these detected words with their scores are used to create hashtag vectors. We refer to the extended method as Probabilistic Inside-Outside Log method for Hashtag (PIOLogH).

Notice that our approach is trying to recommend hashtags which have been created and used in tweets. New hashtag generation is not our goal. Also we are trying to help users who want to share their opinions/interests/comments and join conversations for news topics in Twitter. Other kinds of Twitter users, such as bots, are not considered.

The organization of this chapter is as follows. We show the system structure of our hashtag recommendation in the next section. In Section 4.3, we explain how to create news topic vector based on TF-IDF and our newly proposed PIOLog method. Then in Section 4.4, we explain how to create hashtag vector. Two newly proposed methods (TF-IHF and PIOLogH) to weight terms in hashtag vector are introduced. Finally the approach to recommend hashtags for news topics related to the target word is introduced in Section 4.5.

4.2 System Structure

The whole system structure is shown in Figure 7. In our approach, we first collect news articles and news-related tweets published in a certain period of time concurrently. Then news articles are clustered into topics. News-related tweets containing the same hashtag excluding the tagged screen name of news providers (e.g. #CNN) are concatenated. A vector for representing each hashtag is created. After the target word has been given, news topics which relate to the target word are selected and a vector is created for representing each of the news topic. We calculate the similarity score between each news topic vector and each hashtag vector. Hashtags with high similarity scores are recommended for the news topic.

To represent news topics that relate to the target word, we use our newly proposed Probabilistic Inside-Outside Log (PIOLog) method in Chapter 3 to detect characteristic co-occurrence words from news articles. Words with their scores detected by our PIOLog method are used to create news-topic vectors. We also extend this PIOLog method for hashtags to detect characteristic co-occurrence words co-occurring with hashtags in tweets. Characteristic co-occurrence words with a hashtag from tweets are those words which provide information for the topic the hashtag is often used for. These words could be detected based on assumptions that characteristic co-occurrence words should often co-occur with the hashtag in tweets while they are less likely to co-occur with other hashtags. All these detected words with their scores are used to create hashtag vectors. We refer to the extended method as Probabilistic Inside-Outside Log method for Hashtag (PIOLogH). After vectors of news topics and hashtags are created, we calculate the cosine similarity between their vectors and recommend hashtags which have large cosine similarity scores with the news topic.

4.3 News Topic Vector Creation

A news topic is a group of news articles published in a period of time (for example: one day) reporting about the same recent event in the world. Traditional method for representing the news topic is to define a centroid vector which is calculated by averaging vectors of all news articles in this topic under

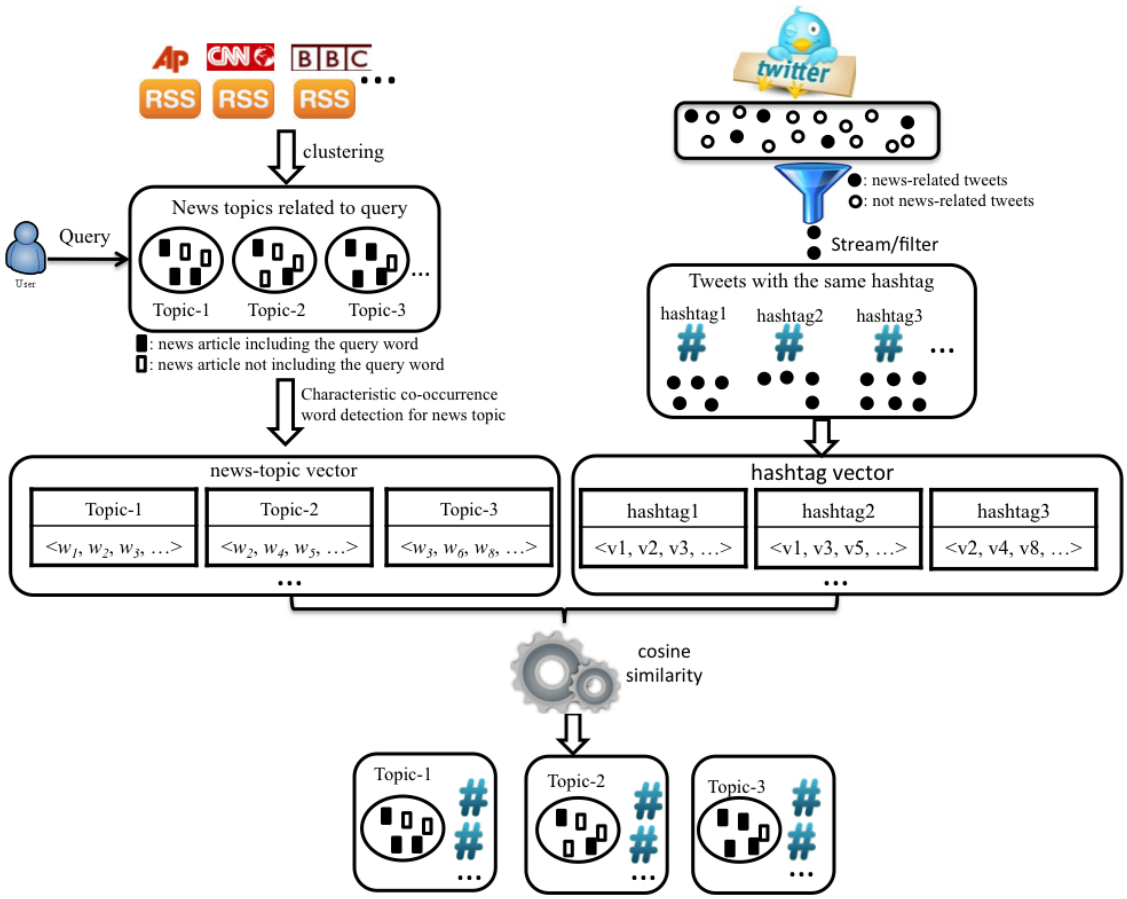


Figure 7. System structure of the news-topic oriented hashtag recommendation

the Vector Space Model [50]. Each vector dimension corresponds to a separate term in news articles and term weights are calculated by the TF-IDF. Although TF-IDF works well in many tasks such as Information Retrieval, it is no longer the best choice for our approach. Firstly, TF-IDF is a query-independent term weighting method, which means the term weight doesn't change no matter what the query is. Secondly, TF-IDF is a topic-independent method. The term which appears in most news articles of a news topic should be weighted higher while TF-IDF could not reflect this idea. At last, even news articles of the same news topic may share many common terms, a news topic may contain thousands of separate terms, which would greatly increase the computation.

In order to solve these problems, we use PIOLog in Chapter 3 for detecting characteristic co- occurrence words of news topics with the target word. Since characteristic co-occurrence words are those words which provide important information for news topics related to the target word, news topic vectors created by these detected words are topic-dependent and query-dependent which could better reflect the meaning of new topics and user's interests.

4.3.1 Term Frequency-Inverse Document Frequency

Traditional way to weight terms in documents is to use term frequency-inverse document frequency (TF-IDF) [48] method. This method measures the importance of word by considering the number of times a word appears in the document and the frequency of the word in all documents. The TF-IDF value of a term would increase if it appears many times in one document and it appears in not too many documents in the corpus. This could help to prevent from giving too large a score to a word which is generally more common than the other words, for example stop words. Detailed calculation of TF-IDF is as follows:

$$\text{TF-IDF}(w,d,D) = \text{TF}(w,d) \times \text{IDF}(w,D) \tag{9}$$

$$\text{TF}(w,d) = \frac{n_{w,d}}{\sum_k n_{k,d}} \tag{10}$$

$$\text{IDF}(w,D) = \log \frac{|D|}{|d_j : w \in d_j| + 1} \tag{11}$$

where w is the term in document d. D is the whole document corpus and $|D|$ gives the number of documents in D. $n_{w,d}$ indicates the number of times w appears in d. $|d_j : w \in d_j|$ gives the number of documents containing w. IF a term w appears many times in d, its TF score would be large. If w appears in less documents of D, its IDF score would be also large, and w is considered to be important for d.

A news topic contains multiple news articles and each news article has a vector whose each dimension corresponds to a separate term in the news article and its value is the TF-IDF score. The news topic vector is created by computing the centroid vector of all vectors of news articles about this topic. Top-n terms/dimensions which have larger TF-IDF values than the rest are kept in the centroid vector. Its creation is as follows:

$$\overrightarrow{centroid}(c,t) = \frac{\overrightarrow{na}(d_1) + \cdots + \overrightarrow{na}(d_{|c|})}{|c|} = < nw_1, \cdots, nw_{|W|} >, c = \{d_1, \cdots, d_{|C|}\} \tag{12}$$

$$\overrightarrow{nv}_{\text{TF-IDF}}(c,t) = < nw_1, \cdots, nw_n > \text{where} \sqrt{(nw_1^2 + \cdots + nw_n^2)} = 1 \tag{13}$$

where $\overrightarrow{na}(d_i)$ is the vector for representing news article d_i. c is the news topic related to the target word t containing multiple news articles as $\{d_1, \ldots, d_{|C|}\}$. nw_j

are distinctive term appearing in news articles of c. $\overrightarrow{centroid}(c,t)$ gives the centroid vector of all vectors of news articles belonging to the same news topic. Top-n terms {nw_1, …, nw_n} in the centroid vector whose TF-IDF scores are larger than the rest are selected to create the news topic vector $\overrightarrow{nv}_{\text{TF-IDF}}(c,t)$ and its value is normalized to make the square sum equal one.

4.3.2 Probabilistic Inside-Outside Log Method

After user provides the target word, news topics related to the target word are selected. Then we use our newly proposed PIOLog method in Section 3.3 to detect characteristic co-occurrence words with the target word for each news topic. Top-n detected words whose PIOLog scores calculated by Equation 1 are larger than the rest are selected to create news topic vector. Each dimension of the vector corresponds to a separate term in news articles and the weight of each dimension is the PIOLog score of that term. The news topic vector is created as follow:

$$\overrightarrow{nv}_{\text{PIOLog}}(c,t) = < nw_1, \cdots, nw_n > \text{where} \sqrt{(nw_1^2 + \cdots + nw_n^2)} = 1 \quad (14)$$

where $\overrightarrow{nv}_{\text{PIOLog}}(c,t)$ is the news topic vector of news topic *c*, which relates to the target word *t*. {nw_1, …, nw_n} are top-n characteristic co-occurrence words with *t* from news articles of this topic. The value of each nw_i is its PIOLog score. Each news topic vector is normalized to make the square sum of all its entries equal one.

4.4 Hashtag Vector Creation

In order to find news-topic oriented hashtags, one intuitive way is to retrieve tweets related to a news topic and recommend commonly used hashtags among these tweets. However, tweet content is limited within 140 characters, which means there is far not enough information in a single tweet to decide whether the tweet relates to the news topic or not. Two tweets may refer to the same topic while both of them share no common word. Also, traditional way as TF-IDF for weighting terms is no longer effective for short text [14] since the number of tweets are too large that the IDF part would dominate the final score while TF part has less affection to the final score.

To solve these problems, we extend our two assumptions in Section 3.1 and propose a new method to detect characteristic co-occurrence words with hashtags. Then all detected words are used to create the hashtag vector for representing each hashtag. These two assumptions for hashtags are described below:

- Characteristic co-occurrence word *w* should often co-occur with the hashtag *ht* in tweets. We take it as the Inside Part.

- Characteristic co-occurrence word *w* should not always appear in tweets without the hashtag *ht*. We take it as the Outside Part.

Based on these two assumptions, we group tweets containing the same hashtag. A hashtag vector is created based on detected words from these tweets. Each dimension of the hashtag vector corresponds to a separate term. To calculate the score for each dimension, we proposed two different methods. One method is term frequency-inverse hashtag frequency (TF-IHF) which is an extended method of TF-IDF. The other method is based on these two assumptions above. We refer to it as the Probabilistic Inside-Outside method for Hashtag (PIOLogH). We describe these two methods in following sections.

4.4.1 Term Frequency-Inverse Hashtag Frequency

This method is a variation of TF-IDF, which considers not only the term frequency in tweets containing the same hashtag, but also the general importance of terms. TF-IHF is calculated as follows:

$$\mathrm{TF\text{-}IHF}(w, ht) = \mathrm{TF}(w, ht) \times \mathrm{IHF}(w, \mathrm{HT}), \mathrm{HT} = \{ht_1, \cdots, ht_i, \cdots\} \quad (15)$$

$$\mathrm{TF}(w, ht) = \frac{n_{w,ht}}{\sum_k n_{k,ht}} \quad (16)$$

$$\mathrm{IHF}(w, \mathrm{HT}) = \log \frac{|\mathrm{HT}|}{|ht_i : \#\mathrm{tweet}(w, ht_i) != 0| + 1} \quad (17)$$

where *w* is the term from tweets containing the hashtag *ht*. HT is the hashtag set containing all hashtags in our tweet set. $n_{w,ht}$ gives the number of times *w* appears in tweets containing *ht*. #tweet(w,ht_i) gives the number of tweets containing both term *w* and the hashtag ht_i. TF(*w,ht*) gives a high value when the term *w* often co-occur with the hashtag *ht* while IHF(*w,HT*) gives a low value when the term also co-occur with many other hashtags since this term might be generally more common than the other words. TF-IHF value ranges from 0 to log|HT|. High value would be reached when *w* frequently appears in tweets containing hashtag *ht* while it is rarely co-occur with other hashtags.

The hashtag vector is created based on top-n words whose TF-IHF scores are larger than the rest. Each dimension of the vector corresponds to a separate

term in tweets containing the same hashtag and the weight of each dimension is the TF-IHF score of that term. The hashtag vector is created as follow:

$$\overrightarrow{hv}_{\text{TF-IHF}}(ht) = < hw_1, \cdots, hw_n > \text{where} \sqrt{(hw_1^2 + \cdots + hw_n^2)} = 1 \quad (18)$$

where $\overrightarrow{hv}_{\text{TF-IHF}}(ht)$ is the hashtag vector for representing hashtag *ht*. {hw_1, …, hw_n} are top-n words whose TF-IHF scores are larger than the rest. The value of each hw_i is the normalized score of the term's TF-IHF score that the square sum of these scores equals one.

However, TF-IHF does not consider about the number of tweets containing both term *w* and hashtag *ht,* which might cause a bias towards terms appearing many times in a few tweets with the hashtag. These terms might get higher TF-IHF scores compared to the others which appear in more tweets with the hashtag but only occur once in each tweet.

4.4.2 Probabilistic Inside-Outside Log Method for Hashtag

To conquer problems in TF-IHF method, we apply those two assumptions proposed at the beginning of this section to detect characteristic co-occurrence words with hashtags from tweets. Our Probabilistic Inside-Outside Log method for hashtags takes those tweets containing hashtag *ht* as the Inside part and tweets containing other hashtags as the Outside part. Terms which often co-occur with hashtag *ht* in tweets of the Inside part while not so often appear in tweets with other hashtags in the Outside part would be taken as the characteristic co-occurrence words with the hashtag and have a high term weight. PIOLogH score will be calculated as follows:

$$\text{PIOLogH}(w, ht) = \log \frac{(1 - s_p)\text{P}(w \mid ht) + s_p}{(1 - s_p)\text{P}(w \mid \neg ht) + s_p} \quad (19)$$

$$\text{P}(w \mid ht) = \frac{\#\text{Tweet}(w \wedge ht)}{\#\text{Tweet}(ht)} \quad (20)$$

$$\text{P}(w \mid \neg ht) = \frac{\#\text{Tweet}(w) - \#\text{Tweet}(w \wedge ht)}{TN - \#\text{Tweet}(ht)} \quad (21)$$

where #Tweet($w \wedge ht$) indicates the number of original tweets containing both *w* and *ht*. Original tweets are tweets posted by users excluding retweeted tweets. Because official retweet function does not allow users to revise tweet contents, so hashtags in retweeted tweets could not reflect original ideas of hashtag usage

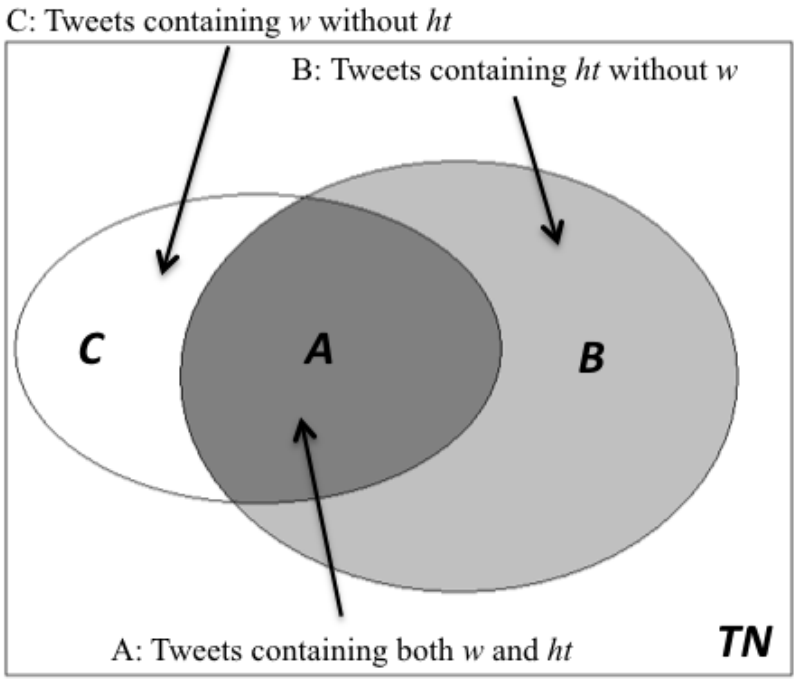


Figure 8. An example about Inside part and Outside part for hashtag *ht*

of users and get excluded here. *TN* is the total number of original tweets containing hashtags in our dataset. #Tweet(*ht*) is taken as the Inside part and words which often co-occur with *ht* in tweets will get a large score in Equation 20, reflecting our first assumption. *TN* - #Tweet(*ht*) is taken as the Outside part and words which are less likely to appear in tweets with other hashtags will get a small score in Equation 21, reflecting our second assumption. Words whose PIOLogH scores calculated by Equation 19 are large would be more likely to be characteristic co-occurrence words with the hashtag. For example, in Figure 8, the Inside part is the region (*A* + *B*) and the Outside part is the region (*TN* – *A* – B). When the region A takes a great portion of region (*A* + *B*) while region C is much smaller than the region (*TN* – *A* – B), *w* would be more likely to be the characteristic co-occurrence word with the hashtag *ht*.

We also create the hashtag vector based on top-n words whose PIOLogH scores are larger than the rest. Each dimension of the vector corresponds to a separate term in tweets containing the same hashtag and the weight of each dimension is its PIOLogH score of that term. The hashtag vector is created as follow:

$$\overrightarrow{hv}_{\mathrm{PIOLogH}}(ht) = < hw_1, \cdots, hw_n > \text{where } \sqrt{(hw_1^2 + \cdots + hw_n^2)} = 1 \qquad (22)$$

where $\overrightarrow{hv}_{\mathrm{PIOLogH}}(ht)$ is the hashtag vector for representing hashtag *ht*. {hw_1, …, hw_n} are top-n words whose PIOLogH scores are larger than the rest. The value of each hw_i is the normalized score of the term's PIOLogH score that the square sum of these scores equals one.

4.5 News-Topic Oriented Hashtag Recommendation

To measure the relevance between the news topic and hashtags, we create the news topic vector and the hashtag vector using top-n characteristic co-occurrence words detected in former sections. Our method to recommend news-topic oriented hashtags is based on two assumptions:

- Tweets containing recommended hashtags should relate to the news topic.
- Recommended hashtags should be widely used by Twitter users when they discuss the news topic.

For the first assumption, when Twitter users are discussing a news topic, some informative words of this news topic would be likely to be used in their tweets. The second assumption means that when one hashtag is widely used for a news topic in Twitter, users would use this hashtag to exchange information about the news topic from different perspectives, which means more informative words of the news topic would be likely to be used in users' tweets. Both of them will result in a high cosine similarity between news topic vector and hashtag vector. Relevance between news topic *c* related to the target word *t* and the hashtag *ht* is calculated as follows:

$$\mathrm{HTRelevance}(ht,c,t) = \cos(\overrightarrow{hv}(ht), \overrightarrow{nv}(c,t)) = \overrightarrow{hv}(ht) \cdot \overrightarrow{nv}(c,t) \quad (23)$$

$$\overrightarrow{hv}(ht) = < hw_1, \cdots, hw_n > \text{ and } \overrightarrow{nv}(c,t) = < nw_1, \cdots, nw_n > \quad (24)$$

where HTRelevance(*ht, c, t*) measures the relevance between *ht* and *c* by calculating the cosine similarity between hashtag vector $\overrightarrow{hv}(ht)$ and news topic vector $\overrightarrow{nv}(c,t)$. Here top-n characteristic co-occurrence words nw_i ($1 \leq i \leq n$) whose TF-IDF or PIOLog scores are larger than the rest are selected to create the news topic vector. Hashtag vector is created in the same way.

We use two different methods (TF-IDF and PIOLog) to weight terms from news articles of the same news topic to create the news topic vector. We also use two different methods (TF-IHF and PIOLogH) to weight terms from tweets containing the same hashtag to create the hashtag vector. Cosine similarity between the news topic vector and the hashtag vector is calculated to measure the relevance between the news topic and the hashtag. Hashtags whose cosine similarity is large get recommended for that news topic. To evaluate recommended hashtags, we ask assessors to evaluate the efficacy of our

recommendation approach. Also, to evaluate the effectiveness of our newly proposed PIOLogH method, we calculate the cosine similarity between the news topic vector and hashtag vector by using different term weight methods, methods which outperform others are considered ranking those topic-specific informative words higher and hashtags recommended by these methods are considered to be more proper for the news topic. Experimental results are shown in Section 6.4.

5

Finding News-Topic Oriented Influential Twitter Users

In this chapter, we propose two new methods to find two types of influential Twitter users for news topics searched by the target word. One type of influential Twitter user often post valuable tweets about the news topic and his tweets are attractive and often retweeted by other users. We refer to this type of user as content-based influential Twitter user. The other type of influential Twitter user has high authority about the news topic due to his reputation or social position. Tweets posted by this type of user are more reliable than other users. We refer to this type of user as authority-based influential Twitter user. Instead of considering user's follow relation which is unsuitable for finding influential Twitter users, we consider about two types of user activity (retweet & mention) with their different motivations. Based on link structures of these two types of activities, we extend PageRank algorithm [51] and proposed two new methods (RetweetRank & MentionRank) to find these two types of influential Twitter users from users who posted tweets about the news topic.

5.1 Overview

Twitter users often post tweets about news topics of what's happening in the world. Although Twitter is a good platform to share news and some suggested that social media lead potential impact on news consumption of ordinary users, a recent survey conducted on ordinary social media users reveals that 92% of users choose to go directly to news websites and 85% of users would do a specific keyword search for their interested news topics. Getting news from social media like Twitter is supplemental for news consumption [9]. However, it is difficult for a user to find those supplemental contents about their interested news topics.

Normally Twitter users often get information by following other Twitter users. However, it is difficult to find users worth following. Tweets posted by content-based influential Twitter users are attractive and valuable while tweets from

others are more likely to be ignored. Finding those content-based influential Twitter users is difficult since measuring the value of tweets is a non-trivial task. Also, Twitter users could post tweets freely while it is hard to know whether these tweets are reliable or not. Tweets posted by authority-based influential Twitter users might be more reliable than tweets posted by others since authority-based influential Twitter users often have high authority on the topic due to their reputation or social position.

To find these two types of influential Twitter users for the news topic, tweets related to the interested news topic of ordinary users are needed. However, due to the length limitation of tweets, traditional Information Retrieval methods are no longer effective in collecting tweets related to the news topic. Here we collect tweets related to the news topic by detecting hashtags in tweets which are relevant to the news topic as we described in Chapter 4. A hashtag which is relevant to the news topic is often used to share contents about the news topic. We refer to this hashtag as news-topic-related hashtag. A group of Twitter users who use this hashtag in their tweets is defined as a news-topic-related hashtag community, and these Twitter users in the community defined by this hashtag are members of this hashtag community. Tweets containing these news-topic-related hashtags are taken as tweets related to the news topic, and two types of influential Twitter users are found from these news-topic-related hashtag communities.

Our approach to find these two types of influential Twitter users for a news topic is based on two assumptions:

- More users a user gets retweeted/mentioned from, more influence the user would have.

- A user has high influence if other users who retweet/mention him are influential.

Based on these two assumptions, we extend the PageRank method and propose RetweetRank and MentionRank to measure the content-based and authority-based influence of Twitter users based on retweet and mention activities. Since retweet represents the value of user's tweet contents, and mention represents user's name value, we consider these two activities respectively to find content-based and authority-based influential Twitter users.

The organization of this chapter is as follows. We show the system structure for finding content-based and authority-based influential Twitter users about a news topic searched by the target word in the next section. Then in Section 5.3 we explain how to detected news-topic-related hashtag communities based on our newly proposed characteristic co-occurrence word detection methods from Chapter 3 and Chapter 4. RetweetRank which is used to find content-based influential Twitter users is presented in Section 5.4. MentionRank which is used to find authority-based influential Twitter users is presented in Section 5.5. In Section 5.6, we explain how to create topic-related teleportation vector in RetweetRank and MentionRank, which makes these two newly proposed

methods be more topic-sensitive. Finally we show how to rank content-based and authority-based influential Twitter users for the news topic in Section 5.7.

5.2 System Structure

The whole system structure of our approach is shown in Figure 9. We first collect news articles and tweets related to news published in a certain period of time concurrently. Then news articles are clustered into topics, and tweets with the same hashtag are grouped together. After a user provides the target word for searching, news topics related to the target word are selected, and news-topic-related hashtags could be detected based on two characteristic co-occurrence word detection methods described in Section 3.3 and Section 4.4. For users in communities defined by these news-topic-related hashtags, two user activity graphs are created. One is retweet graph created based on users' retweet activities, and the other is mention graph created based on users' mention activities. Content-based and authority-based influential Twitter users could be found from these two user activity graphs by using newly proposed RetweetRank and MentionRank methods.

Since a user retweets tweets of others because he is interested in tweet contents while he mentions other users because mentioned users are relevant to the topic he is talking about, we treat user's retweet and mention activities differently and propose RetweetRank and MentionRank to find content-based and authority-based influential Twitter users based on user's retweet and mention activities. Experimental results show that, to find content-based and authority-based influential Twitter users in news-topic-related hashtag communities, RetweetRank and MentionRank outperform other methods using tweet number, in-degree, and PageRank.

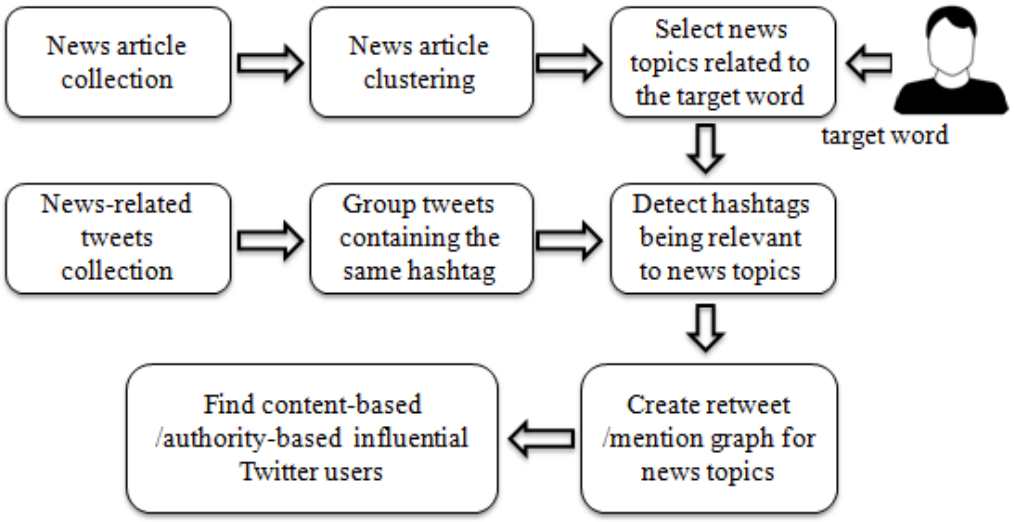


Figure 9. System Structure for finding content-based and authority-based influential Twitter users

Although our study focuses on members of hashtag communities in this research, we believe that influential Twitter users found by our methods are quite helpful. After invented in 2007, hashtags become more and more widely used in Twitter to form conversations about a topic among users globally without following each other. Other conversations formed by functions like reply are restricted by follow relations. A user is less likely to reply to other users who are not followed by him because their tweets will not appear in his Twitter timeline.

5.3 News-Topic-Related Hashtag Community Detection

In this section, we explain how to detected news-topic-related hashtag communities based on our newly proposed characteristic co-occurrence word detection methods from Chapter 3 and Chapter 4. After creating vectors of news topics related to the target word and vectors of hashtags, cosine similarity between the news topic vector and the hashtag vector is calculated. If a hashtag whose cosine similarity with a news topic is larger than a predefined threshold of th_{ht}, this hashtag is taken as the news-topic-related hashtag for the news topic. We refer to the set of these hashtags which are highly relevant to the news topic c as H_c.

After collecting all news-topic-related hashtags in H_c whose cosine similarities with the news topic are larger than the th_{ht}, we create news-topic-related hashtag communities defined by these collected hashtags. All tweets containing any hashtag in H_c are grouped together, and Twitter users who posted these tweets are members of the community defined by the hashtag. Detailed steps are shown in Figure 10. For example, there are three hashtags (hashtag1, hashtag2, and

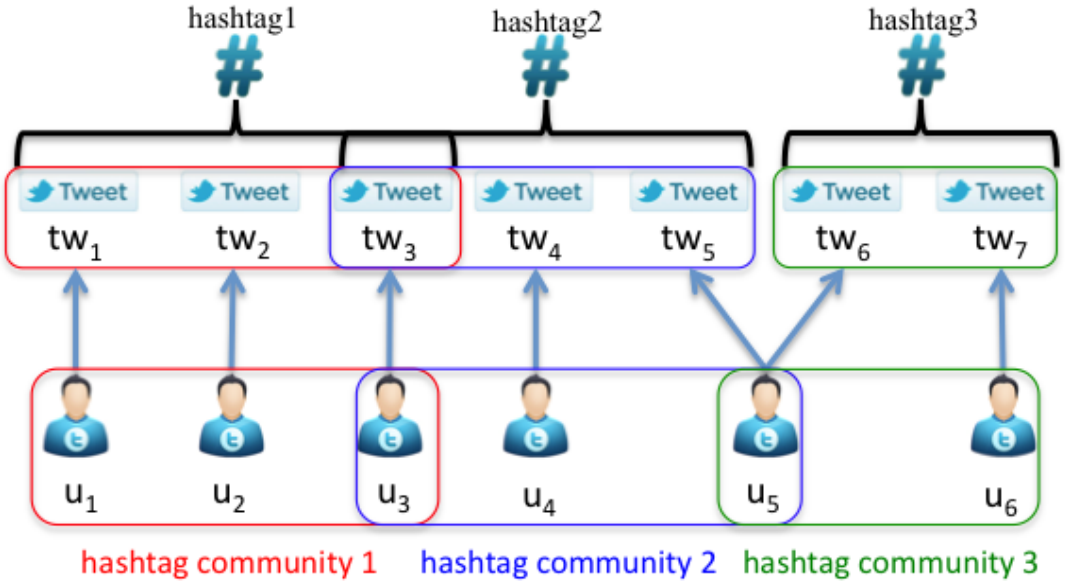


Figure 10. Hashtag Community Creation for News Topic *c*

hashtag3) which are highly relevant to the news topic *c*. Seven tweets (tw_1 – tw_7) posted by six Twitter users (u_1 – u_6) contain at least one of these three hashtags including tweet tw_3 contains hashtag1 and hashtag2. Three hashtag communities are created corresponding to each of these three news-topic-related hashtags, which are shown in red, blue, and green rectangle. As we can observe that one Twitter user may be in more than one hashtag community if he use more than one news-topic-related hashtag in his tweets. For example, user u_5 posted two tweets (tw_5 and tw_6) containing hashtag2 and hashtag3, so he belongs to two communities defined by these two hashtags.

5.4 RetweetRank: Finding Content-based Influential Twitter Users

After creating hashtag communities defined by news-topic-related hashtags in H_c, a directed graph $G_{RT}(V_{RT}, E_{RT})$ is created among these Twitter users based on their retweet activities. We refer to this retweet graph as G_{RT} in this paper. V_{RT} is the vertex set, which contains all Twitter users who retweeted tweets or got retweeted by others in hashtag communities defined by hashtags in H_c. E_{RT} is the edge set. If user u_a retweet a tweet containing $\forall ht \in H_c$ from user u_b, there is an edge between these two users, directing from u_a to u_b. Figure 11 gives an example of G_{RT}.

RetweetRank uses a model of random surfer on G_{RT}. The random surfer follows edges in E_{RT} to visit the next Twitter user based on retweet activities of the former one. The random surfer would also jump to any Twitter user with certain probability even if there is no edge between them. Unlike PageRank whose random surfer visits the next vertex uniformly, the random surfer of RetweetRank visits the next vertex based on user's retweet activities and hashtag preference for *c*. In RetweetRank, the random surfer would be more likely to visit the next user whose tweets containing news-topic-related hashtags are often retweeted by the former user and these two users often use common hashtags for the news topic.

We refer to the transition matrix of RetweetRank for a news topic as A_{RR}. The transition probability from u_a to u_b is calculated as follows:

$$\mathrm{A}_{\mathrm{RR}}(u_a, u_b) = \frac{\#RT(u_a, u_b \mid \forall ht \in \mathrm{H}_c)}{\sum_{u_i \in V_{RT}} \#RT(u_a, u_i \mid \forall ht \in \mathrm{H}_c)} \times \mathrm{HSim}(u_a, u_b) \quad (25)$$

$$\mathrm{HSim}(u_a, u_b) = \overrightarrow{H}(u_a) \cdot \overrightarrow{H}(u_b) \quad (26)$$

$$\overrightarrow{H}(u_i) = <\#\mathrm{Tweet}(u_i, ht_1), \cdots, \#\mathrm{Tweet}(u_i, ht_m) > \text{and } u_i \in \mathrm{V_{RT}}, \mathrm{H}_c = \{ht_1, \cdots, ht_m\} \quad (27)$$

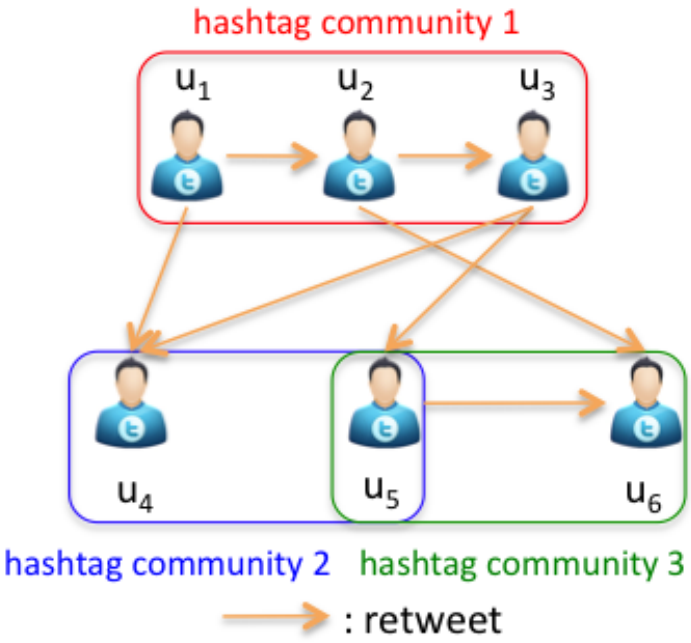


Figure 11. An example of G_{RT}. V_{RT} = {u_1, u_2, u_3, u_4, u_5, u_6}

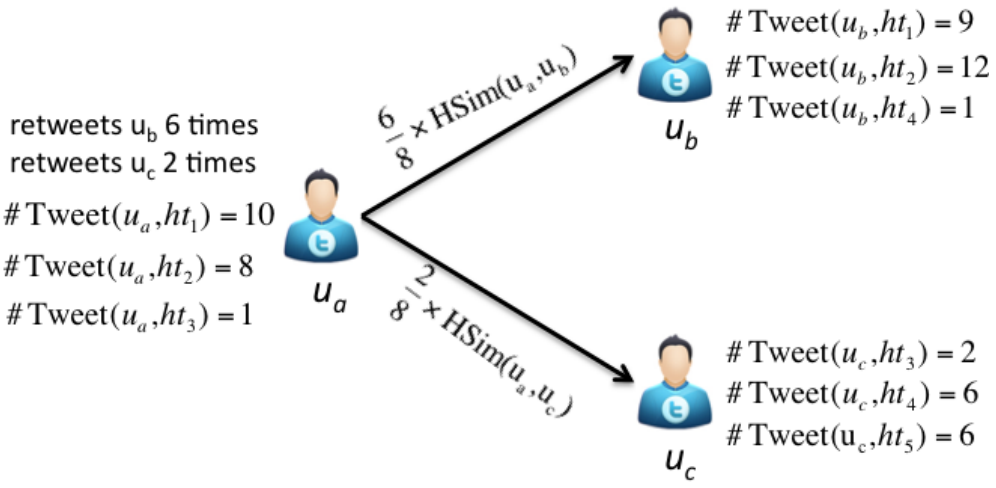


Figure 12. An example of transition probability from u_a to u_b and u_c in A_{RR}

where $\#RT(u_a, u_b \mid \forall ht \in H_c)$ gives the number of tweets u_a retweeted from u_b containing $\forall ht \in H_c$. HSim(u_a, u_b) gives the similarity of hashtag preference between u_a and u_b. $\overrightarrow{H}(u_i)$ is the hashtag preference vector of user u_i in V_{RT}. Each dimension of this vector is #Tweet(u_i, ht_j), which is the normalized number of original tweets containing ht_j posted by u_i. Similar hashtag preference of two users indicates similar interest of them for c. Transition probability between two users in retweet graph is large if one user often retweets tweets containing $\forall ht \in H_c$ from the other user, and they have similar hashtag preference for c. Finally A_{RR} is made to be stochastic so that sum of entries in each row equals one.

Figure 12 gives an example about how to calculate transition probability among Twitter users in G_{RT}. In this example, u_a retweeted more times from u_b than from u_c (6 retweets vs. 2 retweets). This indicates that tweets posted by u_b are more attractive for u_a than those tweets posted by u_c. Also u_a prefers to use news-topic-related hashtags ht_1 and ht_2 in his tweets to share contents about the news topic c and u_b has similar preference of the hashtag usage about c. However, u_c would be more likely to use other hashtags (ht_4 and ht_5) for c. This indicates that u_a and u_b may share common interests for the news topic c while u_c may be different. Defining different transition probabilities between these two pairs of users would help us to find influential Twitter users whose tweet contents are more likely to attract other users. In this example, when the random surfer comes to the vertex of u_a, he is more likely to visit u_b in the next step rather than visiting u_c.

5.5 MentionRank: Finding Authority-based Influential Twitter Users

Similar assumptions are also applied to find authority-based influential Twitter users. A directed graph $G_{MN}(V_{MN},E_{MN})$ is created among users in news-topic-related hashtag communities based on user's mention activities. We refer to mention graph as G_{MN} in this paper. V_{MN} is vertex set. It contains Twitter users who mentioned others or got mentioned in hashtag communities defined by hashtags in H_c. E_{MN} is edge set. If user u_a mention user u_b in his tweets containing $\forall ht \in H_c$, there is an edge between them, directing from u_a to u_b.

MentionRank also uses the random surfer model in G_{MN}. The random surfer in G_{MN} visits the next user following edges between users. He would also jump to any user in G_{MN} with certain probability without any edge. The random surfer of MentionRank would visit the next user based on mention activities of the former one. He would be more likely to visit the next user often mentioned by the former user than other users mentioned in fewer times.

We refer to the transition matrix of MentionRank for a news topic c as A_{MR}. The transition probability from u_a to u_b is defined as follows:

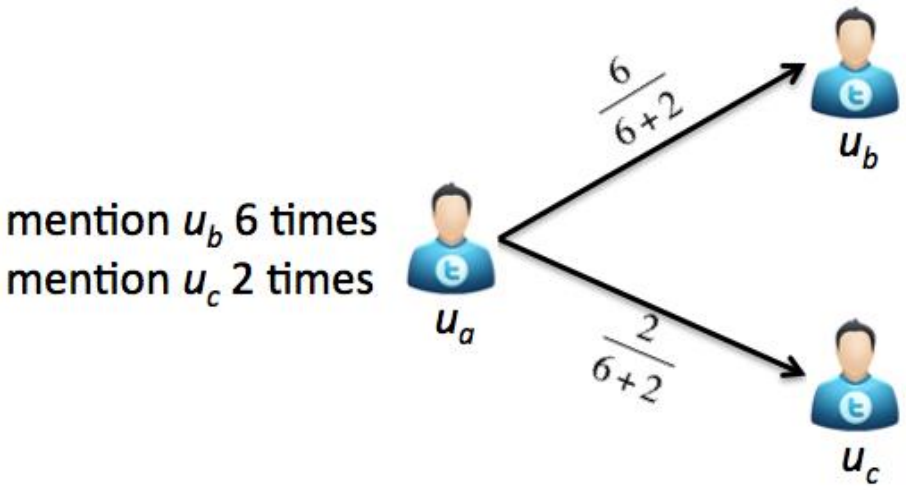


Figure 13. An example of transition probability from u_a to u_b and u_c in A_{MR}

$$A_{MR}(u_a, u_b) = \frac{\#MN(u_a, u_b \mid \forall ht \in H_c)}{\sum_{u_i \in V_{MN}} \#MN(u_a, u_i \mid \forall ht \in H_c)} \quad (28)$$

where $\#MN(u_a, u_b \mid \forall ht \in H_c)$ gives the number of original tweets containing $\forall ht \in H_c$ and mentioning u_b by u_a. Transition probability from u_a to u_b would be large if u_a often mentions u_b in tweets containing any hashtag in H_c while u_a is less likely to mention others. We do not consider hashtag preference of users here because authority-based influential Twitter users are often mentioned by others due to their name value for the topic, not hashtags they use. A_{MR} is also made to be stochastic so that sum of entries in each row equals one.

Figure 13 gives an example about how to calculate transition probability among Twitter users in G_{MN}. In this example, u_a mentioned u_b 6 times in his tweets containing news-topic-related hashtags while u_a only mentioned u_c 2 times in these tweets. This indicates that when talking about the news topic c, u_a would be more willing to contact with u_b rather than u_c. That is to say, the name value of u_b about the news topic c is more valuable than the name value of u_c as u_a believes. So when the random surfer comes to the vertex of u_a, he would be more likely to go from u_a to u_b than from u_a to u_c. More influence scores from u_a would be propagated towards u_b compared with scores propagated towards u_c. In this example, the random surfer from u_a would visit u_b in the next step with the probability of 0.75 while the probability of him to visit u_c from u_a is 0.25.

5.6 Topic-Related Teleportation Vector

To guarantee that the probability distribution in PageRank would converge to a steady state, a teleportation vector is introduced to make the transition matrix be irreducible and aperiodic [51]. Also, in our retweet and mention graph, some pairs of Twitter users retweet/mention each other in a looping manner without retweeting/mentioning others. These user pairs accumulate influence scores without propagating their influence outside. To solve these problems, a teleportation vector is introduced to allow the random surfer to jump to vertices without an edge in certain probability instead of travelling along edges of the graph.

The random surfer in PageRank jumps to any vertex of the graph uniformly while it does not consider about relevance between vertices and the topic. Here we introduce a topic-related teleportation vector for all vertices (users) in retweet and mention graphs considering user's relevance to the topic. It would make the random surfer be more likely to jump to the next user who is highly relevant to the news topic, making the final results more topic-sensitive. A user is highly relevant to the news topic *c* if he often posts tweets about *c* in news-topic-related hashtag communities, and hashtags used by the user are highly relevant to *c*. Since users who are highly relevant to the news topic are interested in this topic, those tweets they retweeted are more valuable than tweets randomly retweeted by others, and Twitter users who are mentioned in their tweets are more likely to have high authority on that news topic compared with Twitter users mentioned for other purposes like getting followed back. This could help to find out content-based and authority-based influential Twitter users effectively and completely. The topic-related teleportation vector $\overrightarrow{TV_c}$ for a news topic *c* is defined as follows:

$$\overrightarrow{TV_c} = < \text{UserRelevance}(u_1, c), \cdots, \text{UserRelevance}(u_n, c) >, \text{and } \mathrm{u_i} \in \mathrm{V_{RT}} \text{ or } \mathrm{u_i} \in \mathrm{V_{MN}} \quad (29)$$

$$\text{UserRelevance}(u_i, c) = \left[\sum_{ht_j \in \mathrm{H}_c} \left(\frac{\#\text{Tweet}(u_i, ht_j)}{\#Tweet(u_i, \mathrm{H}_c)} \times \text{HTRelevance}(ht_j, c, t) \right) \right] \times \log\left[\#\text{Tweet}(u_i, \mathrm{H}_c) + 1 \right] \quad (30)$$

where each dimension of $\overrightarrow{TV_c}$ corresponds to a user in the retweet/mention graph. The value of each dimension UserRelevance(u_i, *c*) measures user's relevance to *c*. HTRelevance(ht_j, *c, t*) is the relevance score between the hashtag ht_j and *c* calculated in Section 4.5. #Tweet(u_i, ht_j) gives the number of original tweets containing $ht_j \in \mathrm{H}_c$ posted by u_i. #Tweet (u_i, H_c) gives the total number of original tweets posted by u_i containing $\forall ht \in \mathrm{H}_c$. A user who is interested in the news topic *c* and often shares contents in news-topic-related hashtag communities would get a large relevance score in his dimension. The random

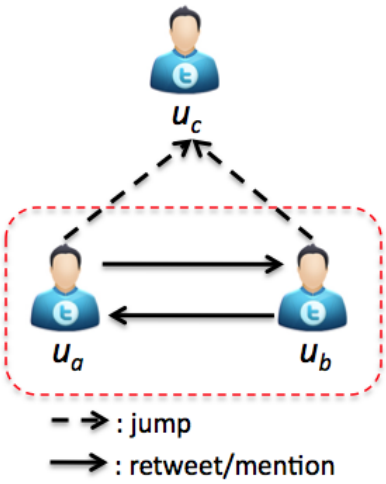


Figure 14. An example of looping manner among users

surfer would be more likely to jump to him. Finally the teleportation vector is normalized to make the sum of dimension values equal one.

Figure 14 gives an example about how the random surfer jumps out of a loop relation between two users. u_a and u_b retweet/mention each other while both of them do not retweet/mention other users. If there is no other edge connecting these two users to others, influence scores of them would be accumulated without propagating to other users, and finally make these two users have large influence scores. The teleportation vector makes the random surfer be able to jump out of this loop by creating edges between these two users and any other users having no connection with u_a and u_b (dashed arrows) to prevent from accumulating influence scores between u_a and u_b. The transition probability of this jump to a third user (for example, u_c) depends on the relevance between the news topic and the user. If u_c is highly relevant to the news topic based on the relevance score computed in Equation 30, the random surfer would be more likely to jump from u_a or u_b to him, and influence scores would be propagated towards those influential users through u_c.

5.7 Ranking Content-based and Authority-based Influential Twitter Users

With the transition matrices for retweet and mention graphs and the topic-related teleportation vector defined, RetweetRank and MentionRank can be calculated by using power iteration method as follows:

$$\text{RetweetRank}: \overrightarrow{RR_c} = d(\mathrm{A}_{\mathrm{RR}})^T \cdot \overrightarrow{RR_c} + (1-d)\overrightarrow{TV_c} \text{ until} \| \overrightarrow{RR_c}(k) - \overrightarrow{RR_c}(k\text{-}1) \| < \varepsilon \quad (31)$$

$$\text{MentionRank}: \overrightarrow{MR_c} = d(\mathrm{A}_{\mathrm{MR}})^T \cdot \overrightarrow{MR_c} + (1-d)\overrightarrow{TV_c} \text{ until} \| \overrightarrow{MR_c}(k) - \overrightarrow{MR_c}(k\text{-}1) \| < \varepsilon \quad (32)$$

where $(A_{RR})^T$ is the transposed transition matrix for retweet graph calculated in Equation 25. $(A_{MR})^T$ is defined in the same way. $\overrightarrow{TV_c}$ is the teleportation vector for the news topic *c* calculated in Equation 29. *d* is the damping factor. Computations for RetweetRank vector $\overrightarrow{RR_c}$ and MentionRank $\overrightarrow{MR_c}$ are done iteratively. These two vectors would converge to stationary probability vectors until 1-norm of the residual vector is less than a predefined threshold ε. Finally, value in each dimension of $\overrightarrow{RR_c}$ or $\overrightarrow{MR_c}$ indicates a user's content-based influence score in retweet graph, or his authority-based influence score in mention graph. Calculations for these two vectors are described as below:

1. $\overrightarrow{RR_c}(0) \leftarrow e/n; \overrightarrow{MR_c}(0) \leftarrow e/n;$
2. $k \leftarrow 1;$
3. Repeat
4. $$\begin{aligned} \overrightarrow{RR_c(k)} &= d(\mathrm{A_{RR}})^T \cdot \overrightarrow{RR_c(k-1)} + (1-d)\overrightarrow{TV_c} \\ \overrightarrow{MR_c(k)} &= d(\mathrm{A_{MR}})^T \cdot \overrightarrow{MR_c(k-1)} + (1-d)\overrightarrow{TV_c} \end{aligned};$$
5. $k \leftarrow k+1;$
6. Until $\| \overrightarrow{RR_c}(k) - \overrightarrow{RR_c}(k-1) \| < \varepsilon; \| \overrightarrow{MR_c}(k) - \overrightarrow{MR_c}(k-1) \| < \varepsilon;$
7. Return $\overrightarrow{RR_c(k)}$ and $\overrightarrow{MR_c(k)}$;

6

Experiments and Evaluation

In this chapter, we firstly give the description about how to collect data including news articles and news related tweets. Secondly we explain setups of our experiments about how to preprocess collected data. Thirdly, preliminary experiments for estimating some parameters are given to show general effectiveness of our methods. Lastly, three experiments with their evaluations are given to prove the effectiveness of our newly proposed methods.

6.1 Description of Dataset

In order to show the effectiveness of our methods, news dataset and news-related tweet dataset are prepared for our experiments. Two crawlers for collecting news articles and news-related tweets are created for preparing these two datasets.

Crawler for collecting news articles is running everyday in our server. It collects newly published news articles written in English from 96 news sites in 21 countries/regions at 2:00, 6:00, 10:00, 14:00, 18:00, and 22:00 everyday. The structure of this crawler is shown in Figure 15. We firstly prepared RSS feeds of these news sites manually. Then the crawler starts 10 threads, each thread would load one news RSS feed and collect news articles from URLs of this feed. After extracting news articles from Web pages, all news articles are stored in MySQL database in our server. However, Web pages of news articles contain not only contents of news, but also other components like news directories, advertisement and so on. Extracting main textual contents from web pages is another research direction called Information Extraction. There are some methods proposed [52] [53] to extract news contents from Web pages. We use the method proposed by Han et al. [52] because their method is effective with quick extracting speed. For each day, about 10,000 news articles are collected and stored in our database.

However, it is not easy to collect news-related tweets because it is difficult to decide whether one tweet relates to a news topic or not due to the length limitation of tweet. Our solution is to manually select 54 active Twitter accounts of news providers and collect tweets containing mentioned/tagged screen name

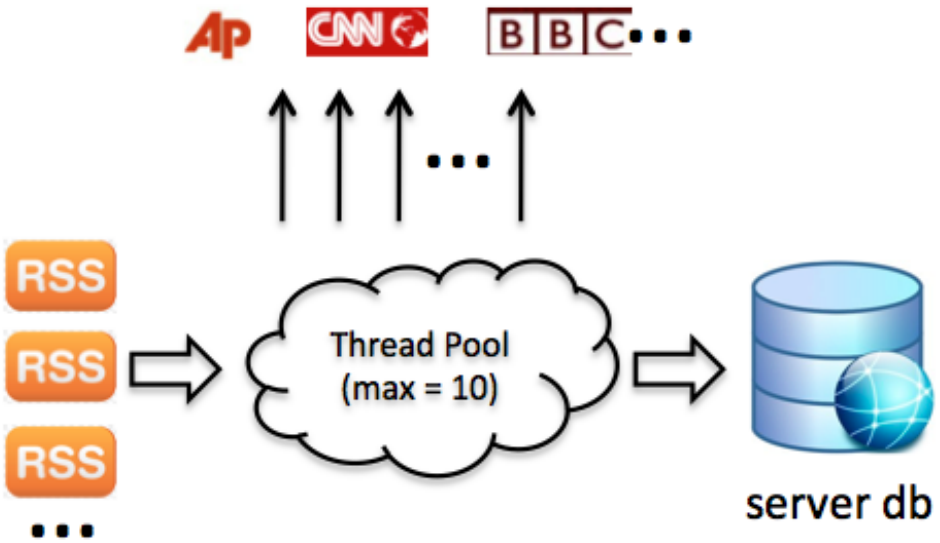


Figure 15. Structure of news crawler

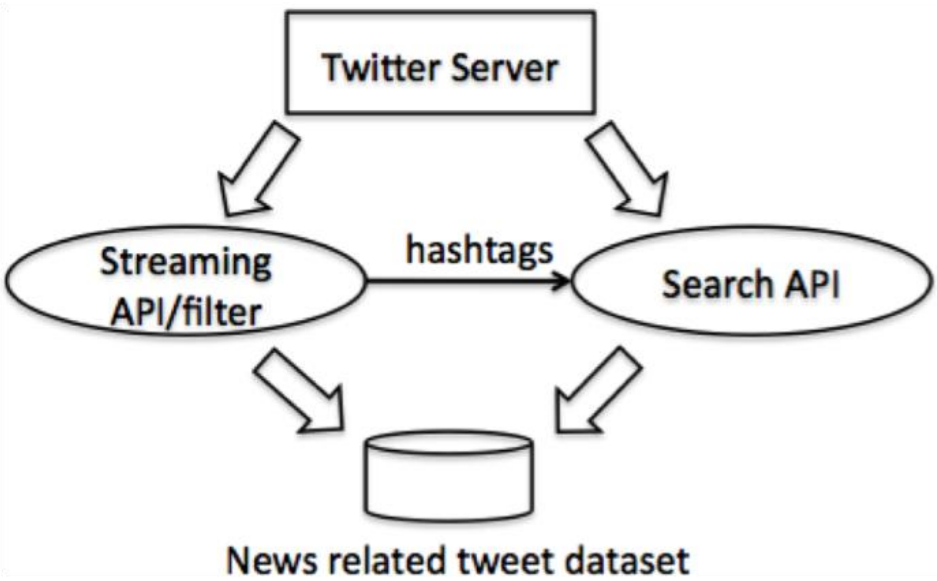


Figure 16. Structure of news-related tweet crawler

of these accounts (e.g. @CNN, #CNN) by using Twitter Streaming API [54]. Then hashtags excluding tagged screen name of these news providers and used in more than 10 collected tweets are selected. These hashtags are used as queries to search for more tweets by using Twitter Search API [55]. At last we combine tweets collected from these two APIs to create the news-related tweet dataset. The whole procedure is shown in Figure 16. We select news articles and news-related tweets collected on October 11th, 2012 for our final experiment. There are 6,868 news articles and 1,496,420 news related tweets collected on this day. Although there might be some other tweets related to news topics, collecting those tweets by using ordinary Information Retrieval technologies is no longer effective.

6.2 Parameter Estimation

In this section, we will discuss how to estimate values of parameters in our system. In order to get rid of overfitting the experimental datasets, parameters need to be estimate in separate datasets to show general effectiveness of our methods. Several preliminary experiments are done on different datasets to estimate values of parameters which can get the optimal results. Then these estimated values are used in our system for experimental datasets to get the final results for evaluation.

Parameters which need to be estimated are listed as follows:

- th_{news}: it is the parameter used in Hierarchical Agglomerative Clustering (HAC, Section 3.2). News articles are clustered in to topics iteratively. The iteration will stop until the maximum pair-wise similarity between any two news clusters is less than th_{news}.

- top-n: n is the number of characteristic co-occurrence words detected from news articles or from tweets. These characteristic co-occurrence words are used to create vectors for representing news topics related to the target word, or hashtags (Section 4.5).

- s_p: it is the smoothing parameter used in our newly proposed PIOLog and PIOLogH methods (Section 3.3 and Section 4.4.2).

- th_{ht}: it is the threshold for detecting news-topic-related hashtag communities (Section 5.3). Hashtags whose relevance scores with the new topic are larger than th_{ht} would be selected with their defined communities. Two types of influential Twitter users for that news topic would be found from these communities.

6.2.1 Parameter Estimation for th_{news}

To estimate the value of th_{news} in HAC, we do a preliminary experiment based on news articles collected on October 9th, 2012. In total, there are 6,322 news articles collected on this day. We set different values of th_{news} in HAC and clustered news articles into news topics. Then we randomly select some news topics and manually check precisions of these news topics with different values of th_{news}. A large value of th_{news} will result in high precisions of news topics while there might be other news articles related to the news topic not clustered into the news topic. That is to say, the recall would be low. Also, a small value of th_{news} will make news topics contain more news articles while some of them are mis-clustered news articles. That is to say the recall is high while the precision will become low. A proper value of th_{news} should be estimated to make both precision and recall high. However, all news articles are unlabeled and it is hard to compute the recall because counting all news articles about a news topic from the whole dataset is difficult. Here, we use the precision as the evaluation metrics while the size of each news topic should be large, containing as many news articles about the news topic as possible.

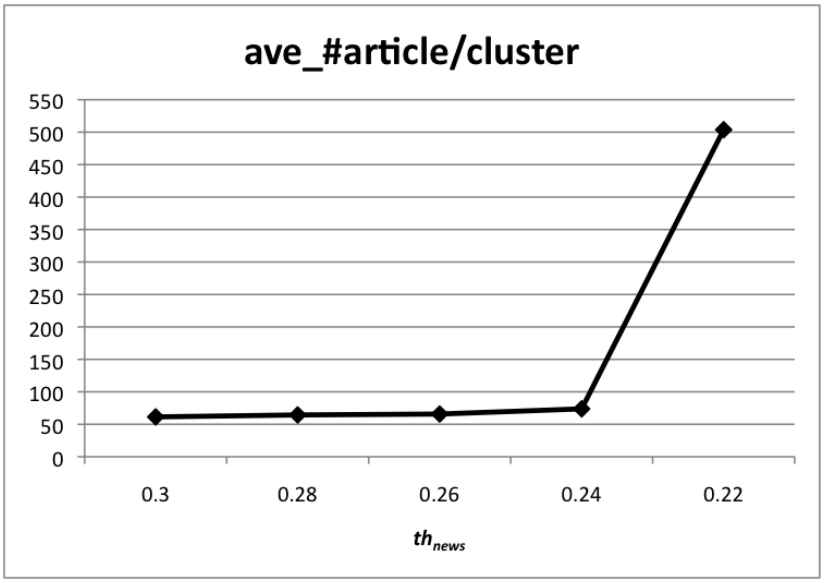
Five news topics (c_1, c_2, c_3, c_4, c_5) on October 9th, 2012 used to compute the precision are described in Table 1. We set the value of th_{news} from 0.2 to 0.3 with the interval of 0.02 and compute the precision values of these five news topics with different th_{news} values. After clustering finished, we manually checked news articles clustered in each news topic. The average number of news articles clustered in each news topic and variations of the average precision value along with the decreasing of th_{news} are shown in Figure 17 (a) and (b).

As we can observe that along with the decreasing of th_{news}, the number of news articles clustered into news topics get increased while precisions of these news topics get decreased. Specifically, precision start to decrease when th_{news} equals 0.24, and when th_{news} is set to 0.22, precisions had a sharp decrease. So the value of th_{news} should be 0.26 or 0.24 where precisions start to decrease between these two values.

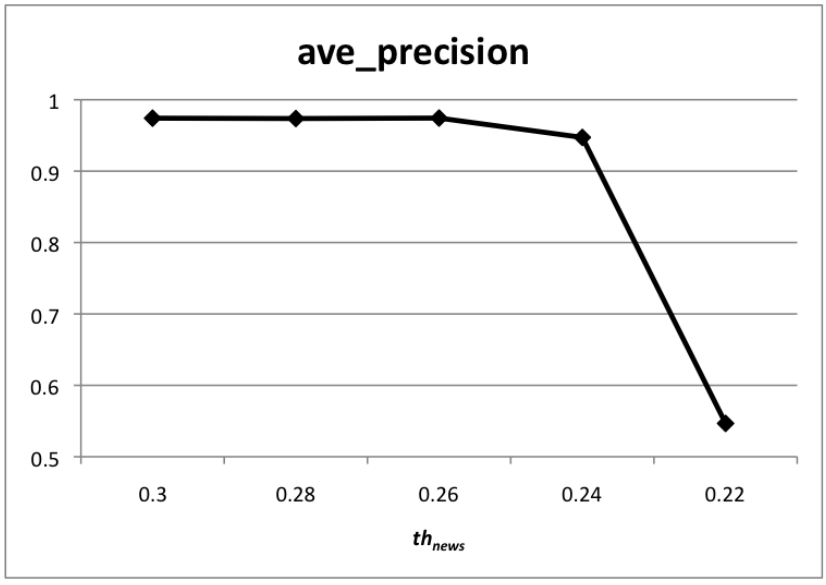
We think that in our research, precision is more important than recall. Notice that in our PIOLog method, false-positive error will cause larger negative affection to our method than false-negative error since df ($t \wedge c$), the size of the Inside part, is always less than (N - df ($t \wedge c$)), the size of the Outside part. Even if

Table 1. Summary of news topics related to the target word

Target word	News topic	Description
Huawei	T_1	Chinese telecom firms face US security threat accusation
North Korea	T_2	North Korea Missiles Warning
Venezuela	T_3	Chavez wins the third re-election in Venezuela
Syria	T_4	Syria crisis reports
Obama	T_5	U.S. presidential election



(a) number of news articles for each news topic



(b) precisions for each news topic

Figure 17. Variation of news article numbers and precisions for each news topic when th_{news} = 0.22, 0.24, 0.26, 0.28, 0.3

a few news articles related to the news topic are not clustered into the news topic (false-negative error), it has less affection to the PIOLog score compared with the affection brought by mis-clustered news articles (false-positive error) which is more serious. In our experiments, th_{news} is set to 0.26.

6.2.2 Parameter Estimation for s_p

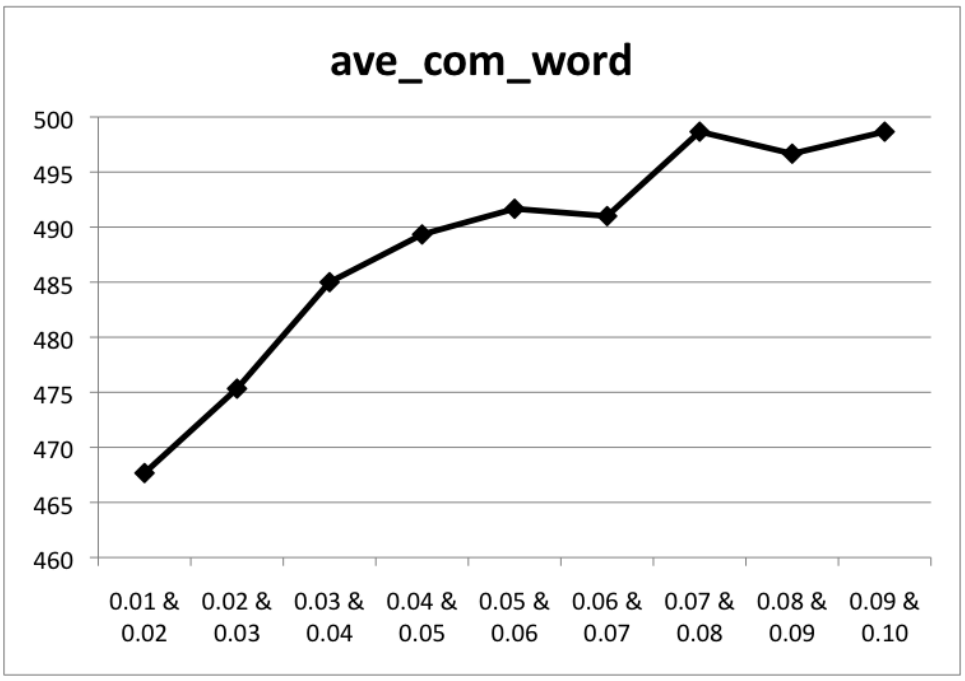
Parameter of s_p in PIOLog and PIOLogH methods is the smoothing parameter to make denominators of Equation 1 and Equation 19 nonzero. Smoothing parameter should be small enough while detected words should be stable, not varying greatly for different values of s_p. Because characteristic co-occurrence words should be words containing important information about the news topic or hashtag, detected words should not vary greatly when s_p makes a small change. Result should be able to reach a stable status because s_p has little affection to the final result.

To estimate the value of s_p in Equation 1 to detect characteristic co-occurrence words with the target word in news topics, we do a preliminary experiment based on news articles collected on March 12th, 2011. There are 6,892 news articles collected on this day. We use HAC to cluster these news articles into news topics with the defined parameter th_{news} estimated in the former section. “earthquake” and “Libya” are used as target words. Three news topics (E_1, E_2, L_1) are selected as news topics related to these two target words, including two of them (E_1 and E_2) relate to the target word of “earthquake”. Detailed descriptions about these three news topics are described in Table 2.

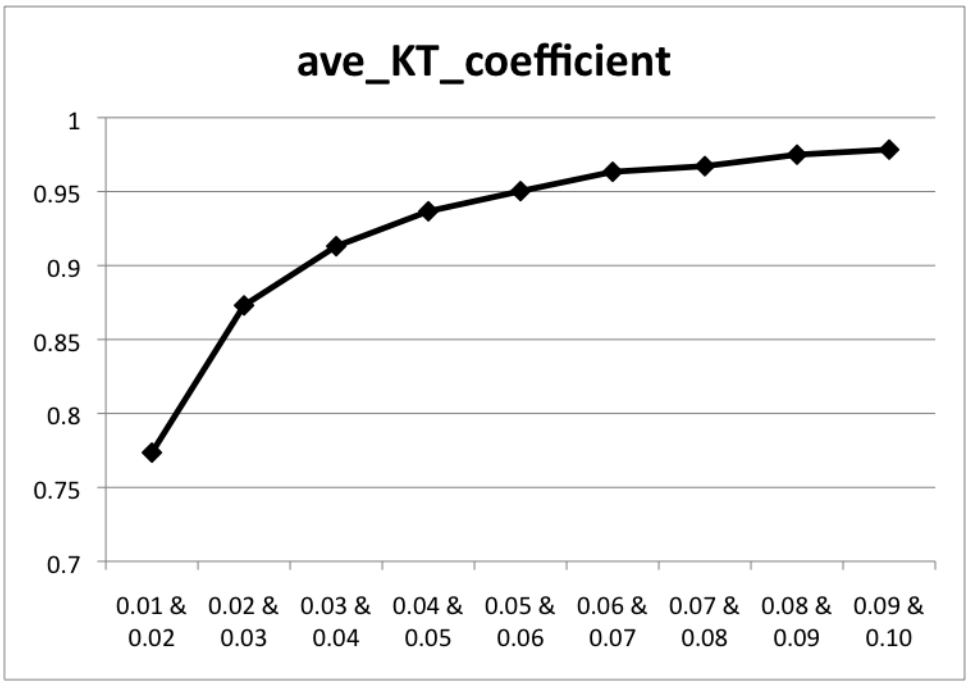
We apply our PIOLog method to each news topic related to the target word and compare detected top-500 words with different s_p values. The value of s_p ranges from 0.01 to 0.1 with the interval of 0.01. We use the number of common words detected by different s_p values and Kendall’s tau coefficient to measure the variation of results for different s_p values. Large number of common words and high score of Kendall’s tau coefficient indicate that results detected by different s_p values do not vary greatly. Figure 18 (a) and (b) show the variation of average common detected words and Kendall’s tau coefficient between detected words by two adjacent s_p values with the interval of 0.01 for three news topics. Each point in x-axis indicates a comparison between detected words with two different s_p values. As we can observe that the number of common words and the Kendall’s tau coefficient do not vary greatly along with the increase of s_p. Also, in these two figures, both curves get increased from starting point, and after

Table 2. Summary of news topics related to the target word

Target word	News topic	Description
earthquake	E_1	Explosion in Fukushima nuclear power
	E_2	Big earthquake hit Japan
Libya	L_1	Libya civil war

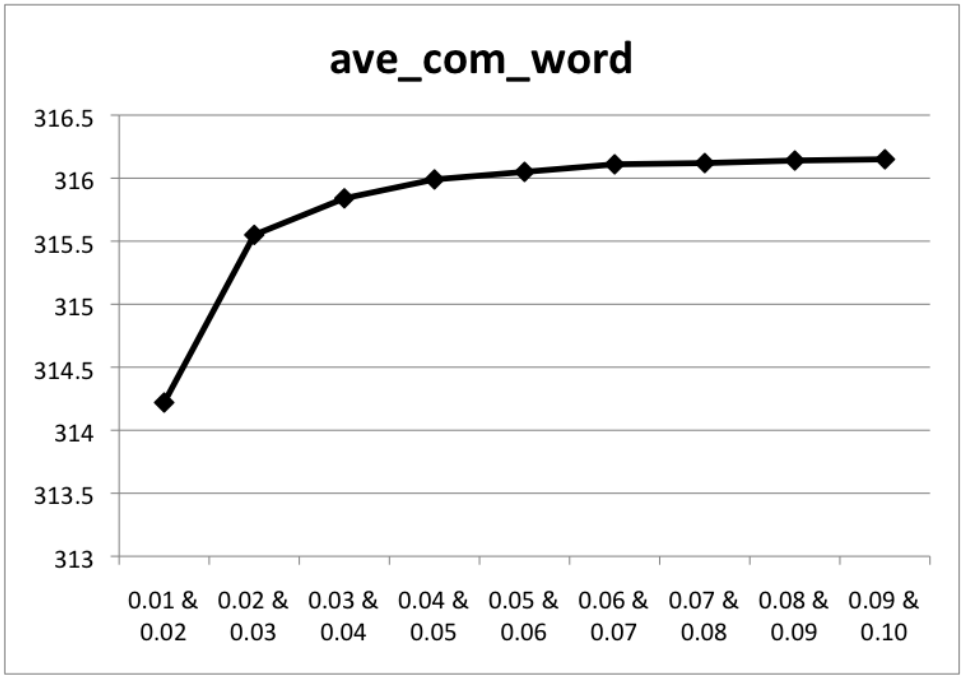


(a). Average number of common words detected by PIOLog with different s_p value

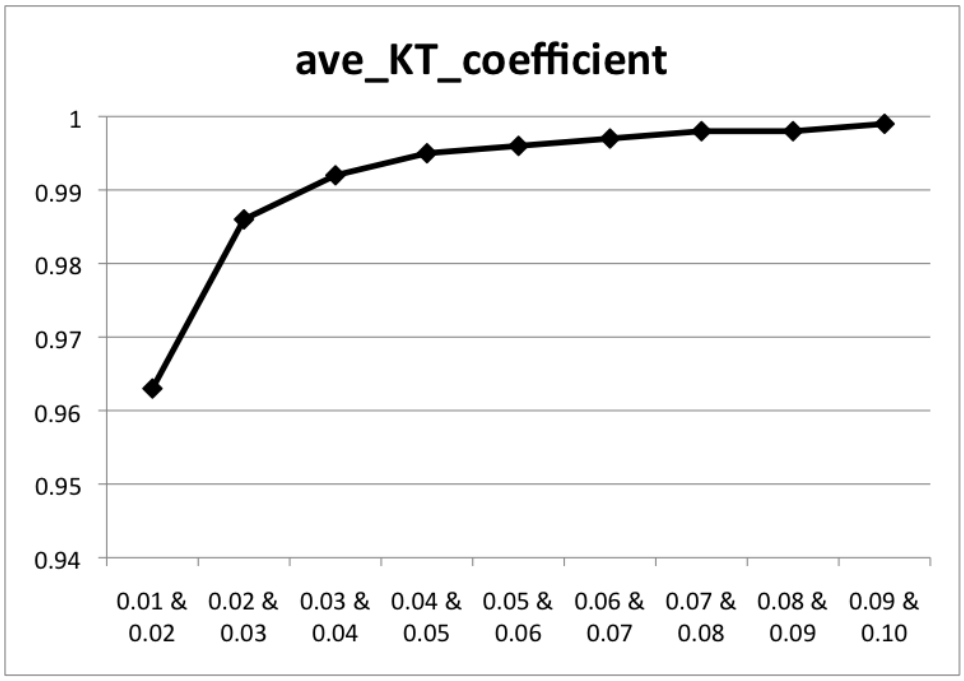


(b). Average Kendall's tau coefficient between detected words by PIOLog with different s_p values

Figure 18. Comparison of PIOLog results with different s_p values



(a). Average number of common words detected by PIOLogH with different s_p values



(b). Average Kendall's tau coefficient between detected words by PIOLogH with different s_p values

Figure 19. Comparison of PIOLogH results with different s_p values

they reach points of “0.04 & 0.05” or “0.05 & 0.06”, these two curves reach a near-stable status with little variation. So we set the s_p in PIOLog to 0.05.

To estimate the value of s_p in PIOLogH to detect characteristic co-occurrence words with hashtags in tweets, we also do a preliminary experiment based on news-related tweets collected on October 9th, 2012. We select hashtags used in more than 50 tweets on this day and there are 1,215 hashtags being selected. Then we apply our PIOLogH method to detect top-500 characteristic co-occurrence words with each hashtag in tweets. Parameter s_p in PIOLogH is set from 0.01 to 0.1 with the interval of 0.01. We measure variation of characteristic co-occurrence words for each hashtag with two adjacent s_p values based on the number of common detected words and the Kendall’s tau coefficient between detected word lists. If detected words by PIOLogH do not change greatly along with the increase of s_p from a certain value, result from PIOLogH reaches to a near-stable status and parameter s_p is set to the value when the near-stable result starts.

Figure 19 (a) and (b) shows the variation of average common detected words and Kendall’s tau coefficient between detected words by two adjacent s_p values with the difference of 0.01 for all hashtags. Each point in x-axis indicates a comparison between detected words with two different s_p values, the same as in Figure 18. As we can observe that the number of common words and the Kendall’s tau coefficient do not vary greatly along with the increase of s_p. Also, in these two figures, both curves get increased from starting point, and after they reach points of “0.04 & 0.05” or “0.05 & 0.06”, these two curves reach a stable status with little variation. So we also set the s_p in PIOLogH to 0.05, the same value as we set in PIOLog.

6.2.3 Parameter Estimation for top-n

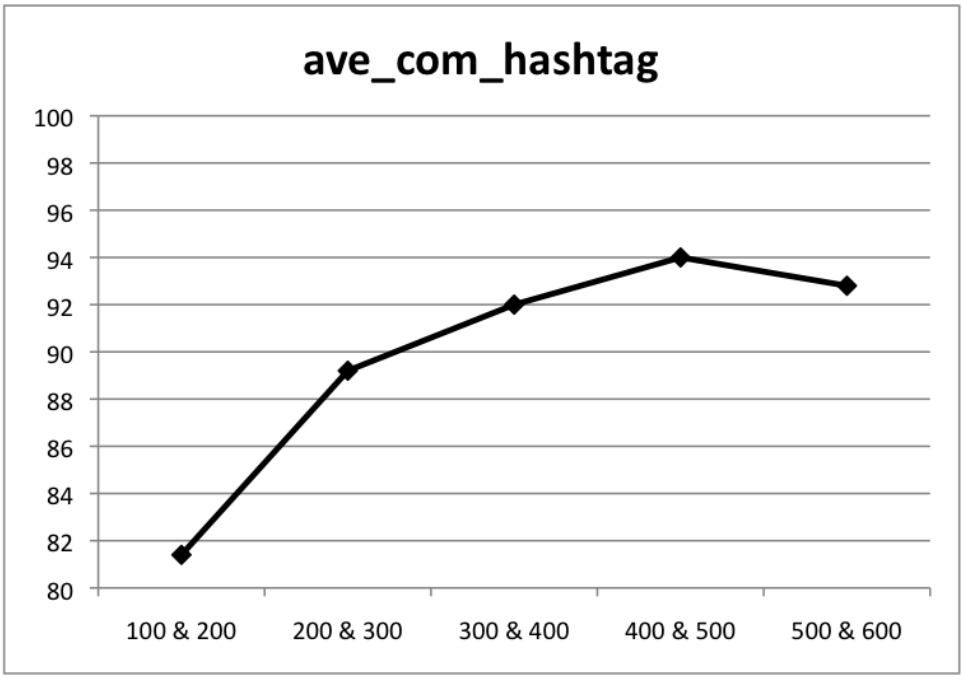
Parameter n is the number of dimension of news topic vector and hashtag vector. We select top-n characteristic co-occurrence words with the target word from news articles related to the same news topic to create the news topic vector. Also top-n characteristic co-occurrence words with a hashtag from tweets containing the hashtag are used to create the hashtag vector. One advantage of selecting top-n words is that it could help to reduce computation. Some news topics contain many news articles with tens of thousands of words, and the dimension of the news topic vector is very large. Storing/processing these large vectors are not only time consuming, but also memory consuming. A more important advantage of using top-n words is that it could help to measure the performance of different characteristic co-occurrence word detection methods. A good detection method could assign large scores to those words which are highly related to the news topic and rank those words higher than the rest. This could help to find hashtags more relevant to the news topic by using the cosine similarity between the news topic vector and the hashtag vector.

We also use news articles and news related tweets collected on October 9th, 2012. We chose “Huawei”, “North Korea”, “Venezuela”, “Syria”, “Obama” as

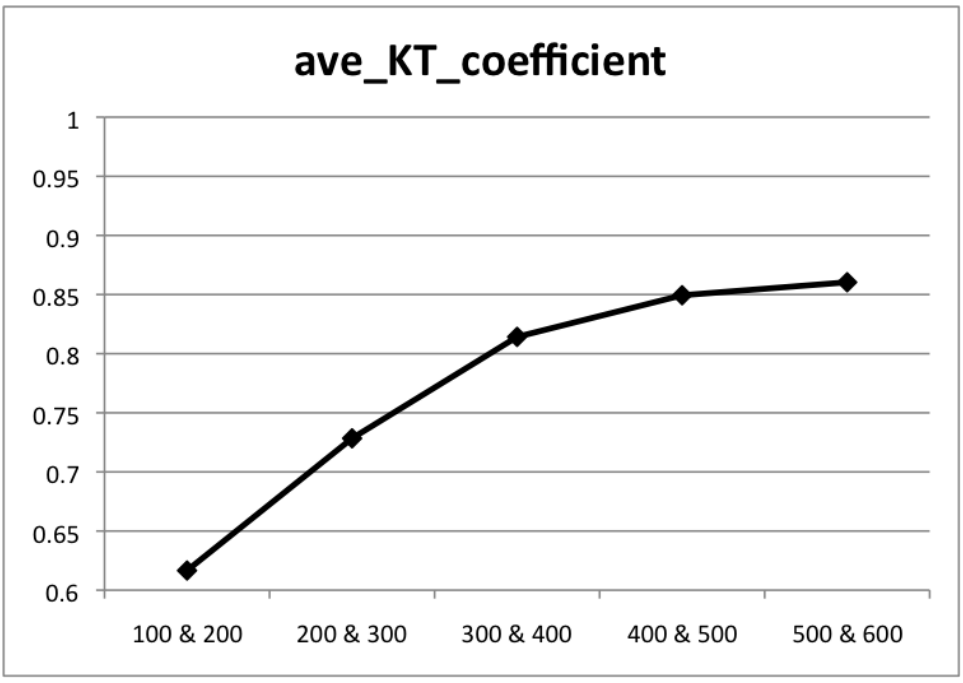
target words and select five news topics related to these target words. Detailed descriptions about these topics are shown in Table 1. Also, hashtags used in more than 50 news related tweets on this day are selected and tweets containing the same hashtag are grouped together. Then PIOLog method is used to detect top-n characteristic co-occurrence words with the target word from news topics to create the news topic vector, and PIOLogH method is used to detect top-n characteristic co-occurrence words with the hashtag from tweets to create the hashtag vector. Hashtags which are highly relevant to the news topic could be detected by using cosine similarity between their vectors.

Hashtags highly relevant to a news topic should be a constant hashtag set with determined ranking order even if dimensions of the news topic vector and hashtag vector continue to increase. That is because words highly relevant to the news topic or hashtag are ranked on top while increasing dimensions could only bring words having little relation with the news topic or hashtag. To determine the value of n, we range its value from 100 to 600 with the interval of 100 and detect top-100 hashtags relevant to each news topic. Then for each value of n, top-100 detected hashtag list is compared with the other top-100 hashtag list with the dimension value of (n - 100). We measure the difference between two hashtag lists based on the number of common hashtags and the Kendall's tau coefficient. When n is set to 100, the start value, not all characteristic co-occurrence words get included in vectors, so when the value of n get increased, detected hashtag list would also vary. However, when the value of n reaches to some value which could detect most of characteristic co-occurrence words, detected hashtags would not vary greatly even if the value of n get increased because newly added dimensions have little affection to the final results since they are not so highly related to the news topic or hashtag compared with those words ranked on top of them.

Figure 20 (a) shows the number of common hashtags among top-100 hashtags detected by different n values with the interval of 100. Each point in x-axis indicates a comparison between top-100 detected hashtags with two adjacent n values. Y-axis gives the number of common hashtags. As we can observe, along with the increase of n from the start point, more common hashtags exist. However, these common hashtags may have different ranking orders. So we use the Kendall's tau coefficient to measure the ranking difference of these common hashtags. Figure 20 (b) shows the Kendall's tau coefficient of common hashtag ranking lists with different n values with the interval of 100. X-axis has the same meaning as in Figure 20 (a). Y-axis gives the Kendall's tau coefficient value. Larger the score is, more similar two ranking list would be. As we can observe, along with the increase of n from the start point, Kendall's tau coefficient is also increased. When the value of n is above 400, we can observe that the number of common hashtags and the Kendall's tau coefficient do not have great variation. Based on these two figures, we set n to 400, which we think would be the proper value.



(a) Average number of common hashtags detected with different n dimensions



(b). Average Kendall's tau coefficient between detected hashtags with different n dimensions

Figure 20. Comparison of top-100 detected hashtags with different n dimensions

Table 3. Summary of news topics related to the target word

Target word	News topic	Description
Republican	R_1	Iowa Republican caucus
	R_2	House Republicans refused to extend payroll tax cut bill
North Korea	NK_1	Power transition after Kim Jong-il's death
	NK_2	World stock market affected by King Jong-il's death
Syria	S_1	Syria allowed observers into the country to end crisis
protester	P_1	Egyptian army started to clear Tahrir Square with force

6.2.4 Parameter Estimation for th_{ht}

Parameter th_{ht} is used to select news-topic-related hashtag communities. If the relevance between the news topic and the hashtag is larger or equal th_{ht}, the hashtag would be taken as the news-topic-related hashtag, and the community defined by the hashtag would be taken as the news-topic-related hashtag communities. Two types of influential Twitter users for the news topic could be found from these hashtag communities.

To estimate the proper value of th_{ht}, we collected 11,820 news articles and 182,979 news related tweets on March 7th, 2012. In this preliminary experiment, after clustering all news articles into news topics, we choose "Republican", "North Korea", "Syria", and "protester" as target words. For each target word, news topics which contain more than half of news articles including the target word are selected as the news topics related to the target word. Detailed descriptions about these news topics are described in Table 3.

We use our newly proposed approach in Section 4.5 to detect hashtags relevant to each news topic related to target words. Values of other parameters are set to estimated values from former sections. To evaluate those detected hashtags from our approach, we asked assessors to judge the relevance of the detected hashtags to each news topic. The whole procedure is shown as below

- 1. Three assessors are asked to read at least ten news articles which are carefully selected from each news topic so that these news articles cover the main contents of the news topic to make them understand the news topic.

- 2. Top-15 hashtags with largest similarities detected by our approach (Section 4.5) are selected for each news topic. Assessors judge the relevance of each hashtag in this list to the news topic on a three-point scale: highly relevant, partially relevant and irrelevant. They can use any tool (TagDef [56] or Google [57]) to help them make the decision.

- 3. For each news topic, hashtags which are judged as highly relevant by at least two assessors are defined as highly relevant hashtags. We also

define relevant hashtags as they should not be judged as irrelevant by any assessors. Notice highly relevant hashtags are a sub-set of relevant hashtags.

The value of th_{ht} ranges from 0.1 to 0.2 with the interval of 0.01. For each value of th_{ht}, we calculate the Mean Average Precision (MAP) of detected hashtags for all news topics related to target words and draw a MAP- th_{ht} curve in Figure 21. For a given value of th_{ht}, the MAP is calculated as follows:

$$\mathrm{MAP}=\frac{\sum_{i=1}^{|C|}\mathrm{AveP}(c_i)}{|C|}, C=\{\mathrm{R}_1,\mathrm{R}_2,\mathrm{NK}_1,\mathrm{NK}_2,\mathrm{S}_1,\mathrm{P}_1\} \tag{33}$$

$$\mathrm{AveP}(c_i)=\frac{\sum_{j=1}^{|\mathrm{HT}|}\mathrm{precision}(j)}{|\mathrm{HT}|} \tag{34}$$

where AveP(c_i) is the average precision of detected hashtags being relevant to the news topic c_i. C is the news topic set related to the target word. precision(j) is the ratio of relevant hashtags for the news topic among top-j detected hashtags. |HT| is the number of hashtags used for evaluation. Here |HT| is the number of hashtags whose relevance scores are larger than or equal th_{ht}.

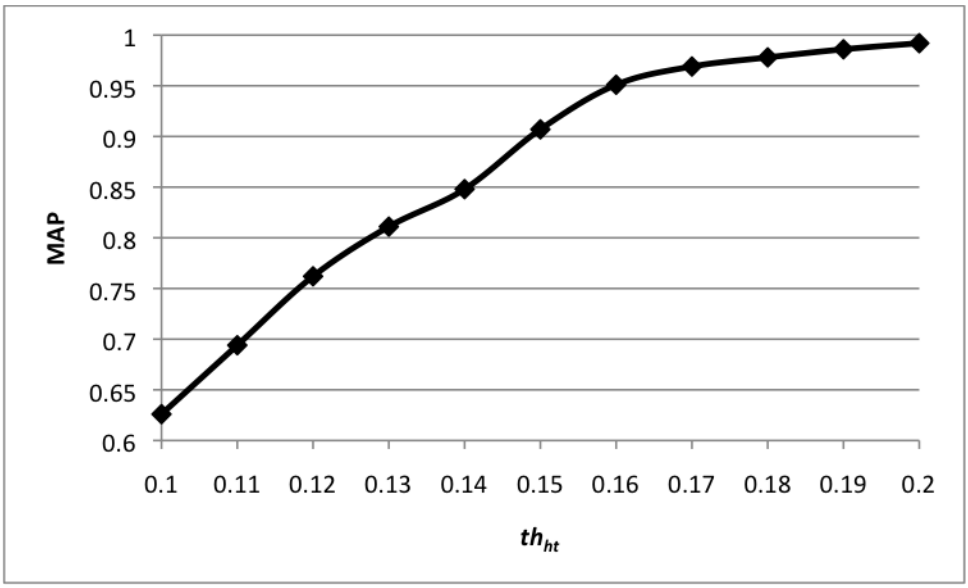


Figure 21. MAP of news topics (R_1, R_2, NK_1, NK_2, S_1, P_1) for different th_{ht} values

Figure 21 shows the MAP of these six news topics for different th_{ht} values. As we can observe that along with the increase of th_{ht}, MAP for all news topics related to target words also increases, which means hashtags whose cosine similarity scores with the news topic is above th_{ht} are more likely to be relevant to the news topic. We can also observe that after the value of th_{ht} reaches to 0.16 or 0.17, the MAP keeps in a near-stable status and does not vary greatly. Here we set the value of th_{ht} to 0.17 and use this value in later experiments.

6.3 Evaluation for Characteristic Co-Occurrence Word Detection from News Articles

In this section, we describe our experiment for detecting characteristic co-occurrence words from news articles and compare our newly proposed PIOLog method with related methods. We also evaluate performance of each method.

6.3.1 Experimental Setup

In order to evaluate the effectiveness of our PIOLog method to detect characteristic co-occurrence word, we use 6,868 news articles collected from news RSS feeds of news providers on October 11th in 2012. All news articles are processed as follows:

- Step1: All news articles are parsed by TreeTagger, a morphological analysis tool [58], and all nouns, proper nouns, foreign words, verbs, and adjectives are picked up.

- Step 2: Named entities such as “White House” are recognized by Stanford Named Entity Recognizer (SNER) [59] and each named entity is treated as a single word.

- Step 3: Each sequence of proper nouns is treated as one word even if it is not recognized as named entity by SNER or separated by preposition. If SNER recognizes one part of the sequence as a named entity, the sequence is separated into the named entity and the remaining part (e.g. “Vice President Joe Biden” is separated into “Vice President” and “Joe Biden”).

After parsing contents of news articles, each news article is represented by a vector under the Vector Space Model [50]. Each dimension of this vector corresponds to a separate word appearing in news articles, and the value of each dimension is calculated by TF-IDF [48]. Then all news articles are clustered

Table 4. Summary of news topics related to the target word

ID	Summary
Target Word = "Obama"	
O_1	U.S. presidential debate
Target Word = "Syria"	
S_1	Syria crisis and conflictions
Target Word = "game"	
G_1	American Major League Baseball news
G_2	News for England football match

Table 5. Detailed information about news topics related to target word

	$df(c)$	$df(t \wedge c)$
O_1	179	164
S_1	86	84
G_1	65	63
G_2	33	25

based on Hierarchical Agglomerative Clustering method described in Section 3.2. News articles related to the same news topic are hopefully grouped into the same cluster with a predefined similarity threshold of 0.26. All news clusters containing at least 20 news articles are selected while other cluster with less than 20 new articles are excluded because these news clusters often relate to local news and noise topics.

We choose three words, "Obama", "Syria", and "game", as target words. A news cluster is taken as a news topic related to the target word if at least half of its news articles contain the target word. Summaries of each news topic related to the target word are described in Table 4. There are four news topics selected and denoted by O_1 (for "Obama"), S_1 (for "Syria"), G_1 and G_2 (for "game"). Notice that there are two news topics related to the target word of "game". Without clustering, characteristic co-occurrence words from these two news topics will be mixed together and it is difficult to separate words for different topics. Detailed information about news topics related to the target word is described in Table 5. *df* (*c*) indicates the number of news articles in the news topic *c*. *df* ($t \wedge c$) indicates the number of news articles containing the target word *t* in the news topic *c*.

6.3.2 Comparison with Related Methods

We compare the PIOLog method with Jaccard coefficient (Jaccard) and Log Likelihood Ratio (LLR) discussed in Section 3.4. However, in order to apply

Jaccard and LLR to our experiments, these two methods should be revised to adapt for the news topic. Jaccard method is extended as follow:

$$Jaccard = \frac{df(w \wedge t \wedge c)}{df(w \vee (t \wedge c))} = \frac{df(w \wedge t \wedge c)}{df(w) + df(t \wedge c) - df(w \wedge t \wedge c)} \quad (35)$$

where *df* ($w \wedge t \wedge c$) is the number of news articles in the news topic *c* and containing both word *w* and target word *t*. *df* ($w \vee (t \wedge c)$) indicates the number of news articles containing *w* or containing *t* and in *c*.

Log Likelihood Ratio is also extended. Here the null hypothesis is set as the occurrence of word *w* is independent of the target word *t* and the news topic *c*. Alternative hypothesis is set as *w* is dependent on *t* and *c*. These two hypotheses are described as follows:

$$\text{Null hypothesis} H_0 : P(w \mid t \wedge c) = p = P(w \mid \neg(t \wedge c)) \quad (36)$$

$$\text{Alternative hypothesis} H_1 : P(w \mid t \wedge c) = p_1 \neq p_2 = P(w \mid \neg(t \wedge c)) \quad (37)$$

$$\text{where } p = \frac{df(w)}{N} \; p_1 = \frac{df(w \wedge t \wedge c)}{df(t \wedge c)} \; p_2 = \frac{df(w) - df(w \wedge t \wedge c)}{N - df(t \wedge c)} \quad (38)$$

LLR assume a binomial distribution b(k; n, x) for each word in news articles, then likelihoods of having *df*(*w*), *df* ($t \wedge c$), and *df* ($w \wedge t \wedge c$) observed under H_0 and H_1 are defined as follows:

$$L(H_0) = b(df(w \wedge t \wedge c); df(t \wedge c), p) \cdot b(df(w) - df(w \wedge t \wedge c); N - df(t \wedge c), p)$$

$$L(H_1) = b(df(w \wedge t \wedge c); df(t \wedge c), p_1) \cdot b(df(w) - df(w \wedge t \wedge c); N - df(t \wedge c), p_2)$$

The Log Likelihood Ratio for news topic *c* is defined as logarithmic value of $L(H_0)$ divided by $L(H_1)$ multiplied by – 2 (Equation 8 in Section 3.4).

Since our purpose is to use characteristic co-occurrence words to recommend hashtags for news topics, we compare the quality of results achieved by the same hashtag recommendation approach in Section 4.5 when input characteristic co-occurrence words are detected by different methods. The method whose detected words are topic-specific and more relevant to the news topic could help to detect hashtags which are more relevant to the news topic.

For each news topic related to the target word, we apply three methods (Jaccard, LLR, and PIOLog) to detect characteristic co-occurrence words. We also prepare news-related tweets on the same day and preprocess their contents. Detailed procedure is described in Section 6.4.1. For each hashtag from news-related tweets, we also apply these methods or its extension (Jaccard, LLR, and PIOLogH) to detect characteristic co-occurrence words with the hashtag from tweets. Top-n words whose scores calculated by these methods are larger than the rest are selected from news articles or tweets.

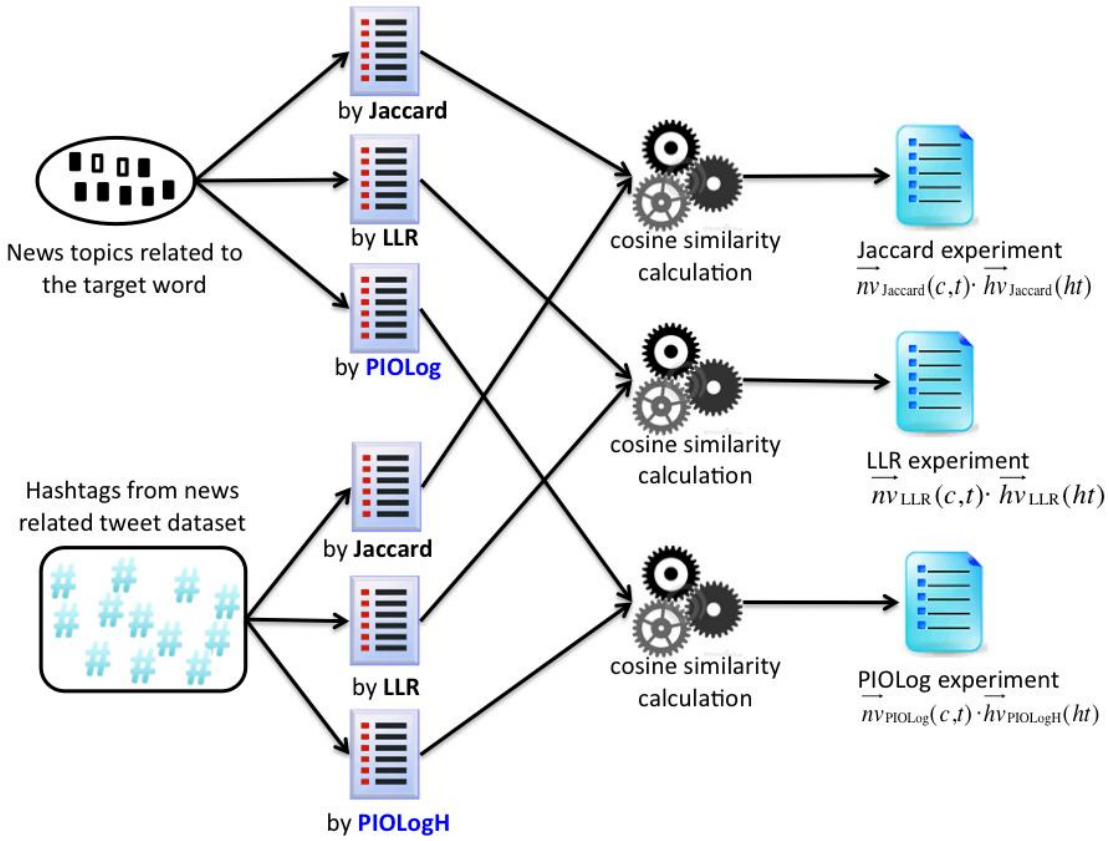


Figure 22. Experiments to evaluate characteristic co-occurrence word detection methods in the context of hashtag recommendation

We prepare three experiments using different methods to detect characteristic co-occurrence words. The experiment is shown in Figure 22.

- PIOLog experiment: PIOLog method is used to detect characteristic co-occurrence words for news topics. Its extension as PIOLogH method (Sec. 4.4.2) is used for hashtags.
- LLR experiment: Log Likelihood Ratio method is used to detect characteristic co-occurrence words for news topics and hashtags.
- Jaccard experiment: Jaccard coefficient is used to detect characteristic co-occurrence words for news topics and hashtags.

6.3.3 Evaluation

We select top-n characteristic co-occurrence words detected by different methods in these three experiments. The value of n is set to 400 as we estimated in Section 6.2.3. Detected words are applied to the same hashtag recommendation system. Method which outperforms others would rank those

topic-specific informative words higher and hashtags recommended by the experiment using the method should be more relevant to the news topic.

After we get recommended hashtags for each news topic from three experiments, we ask two assessors to judge these recommended hashtags are highly relevant, relevant, or irrelevant to the news topic. Detailed procedure is described in Section 6.4.2. Then Precision at Highly Relevance (P@HR) and Precision at Relevance (P@R) are used to evaluate results of these three experiments. P@HR is the ratio of highly relevant hashtags among top-k recommended hashtags and P@R is the ratio of relevant hashtags among top-k recommended hashtags. We range the value of k from 1 to 15 and link up P@HR and P@R values into curves. Experiment who's these two curves locate higher than the rest outperforms and could recommend hashtags more relevant to the news topic. Characteristic co-occurrence word detection method used in this experiment also outperforms other methods.

Figure 23 – Figure 30 show P@HR and P@R curves of four news topics (O_1, S_1, G_1, and G_2) based on top-15 detected hashtags using Jaccard/LLR/PIOLog. As we can observe that curves of PIOLog experiment locate higher than other experiments using Jaccard coefficient and Log Likelihood Ratio. This indicates that our newly proposed PIOLog method and its extension as PIOLogH method are more likely to detect characteristic co-occurrence words than Jaccard coefficient and Log Likelihood Ratio.

As we can also observe, asymmetric methods (LLR, PIOLog) outperform symmetric method (Jaccard) in characteristic co-occurrence word detection. As we have pointed out in Section 3, when we take two words be w_1 and w_2, whether w_1 is a characteristic co-occurrence word with w_2 or not and whether w_2 is a characteristic co-occurrence word with w_1 should be different in general. Asymmetric methods could reflect this idea while symmetric methods can't. This advantage of asymmetric methods for detecting co-occurrence relation has also been found in [16]. Also, w_1 does not need to appear in many news articles containing w_2 to be characteristic co-occurrence word. Whether w_1 often appears in other news articles not including w_2 is also important for judging w_1 is a characteristic co-occurrence word or not.

For these two asymmetric methods, although LLR always performs better than Jaccard, it still can't outperform PIOLog method. LLR considers the appearance of word *w* is independent/dependent of the target word *t* in news topic *c* that seems to be similar to our assumptions of PIOLog. However, LLR is still not suitable to detect characteristic co-occurrence words because LLR is often used to detect word collocation, which is a different purpose compared with ours. Our characteristic co-occurrence word detection is to detect words strongly related to the target words due to a specific news topic, not as a grammar unit constantly.

As we can observe that PIOLog and PIOLogH methods performs better than LLR in large news topics, but they perform very similar in small news topics. For example, there are 179 news articles in O_1 and results of PIOLog experiment outperform LLR experiment in Figure 22 and Figure 23. However, for the news topic of G_2 containing 33 news articles, their results are very similar (Figure 29

and Figure 30). This is because LLR is more appropriate for sparse data. That is to say, words co-occurring with the target word in news topics of small size are more likely to be detected by LLR while in large size of news topics, they are less likely to be detected because LLR prefers those words co-occur with the target word in less news articles of the news topic. Although being appropriate for sparse data is an advantage of LLR to detect word collocation, it is a big disadvantage to detect characteristic co-occurrence word because characteristic co-occurrence words should co-occur with the target word in more news articles of the news topic. This feature of LLR contradicts the definition of characteristic co-occurrence word.

Since characteristic co-occurrence words should be strongly related to the target word about the news topic, users could well understand the news topic if we provide what words are linked to the news topic related to the target word. For every news topic (O_1, S_1, G_1, G_2), top-20 characteristic co-occurrence words are selected from each method (Jaccard, LLR, PIOLog) and two assessors are asked to judge relevance of each characteristic co-occurrence word in the following procedure:

- Step 1. Ten representative news articles are carefully selected from each news topic so that they cover the main content of the topic. They should belong to this news topic, not a mis-clustered article.

- Step 2. The two participants are asked to read these ten news articles to understand the news topic.

- Step 3. All of the characteristic co-occurrence words detected from these three methods are mixed together to form a word list. The two assessors judge each word in the word list on a three-point scale with the meaning described in Table 6.

- Step 4. The final relevance score of each word is calculated by averaging the relevance scores assigned by two participants if the difference of their relevance score is 1 (e.g. in the case that one participant assigned 1 and the other assigned 0). In the case that one participant assigned 0 and the other assigned 2, another participant is asked to judge relevance of the word.

We give the top-20 characteristic co-occurrence words detected by Jaccard, LLR, and PIOLog about four news topics (O_1, S_1, G_1, and G_2) in Table 7 – Table10. For example, Table 7 gives top-20 characteristic co-occurrence words for the news topic of U.S. presidential election (O_1), assessors assign relevance score of 2 to the word “Romney”, one candidate for U.S. president election, since he is one of the main characters of the election and highly related to the topic. Word “Ohio” is assigned a relevance score of 1 because “Ohio” relates to news articles reporting that fierce competition between Obama and Romney happened in Ohio State while it is difficult to foretell who will win. However, there are many

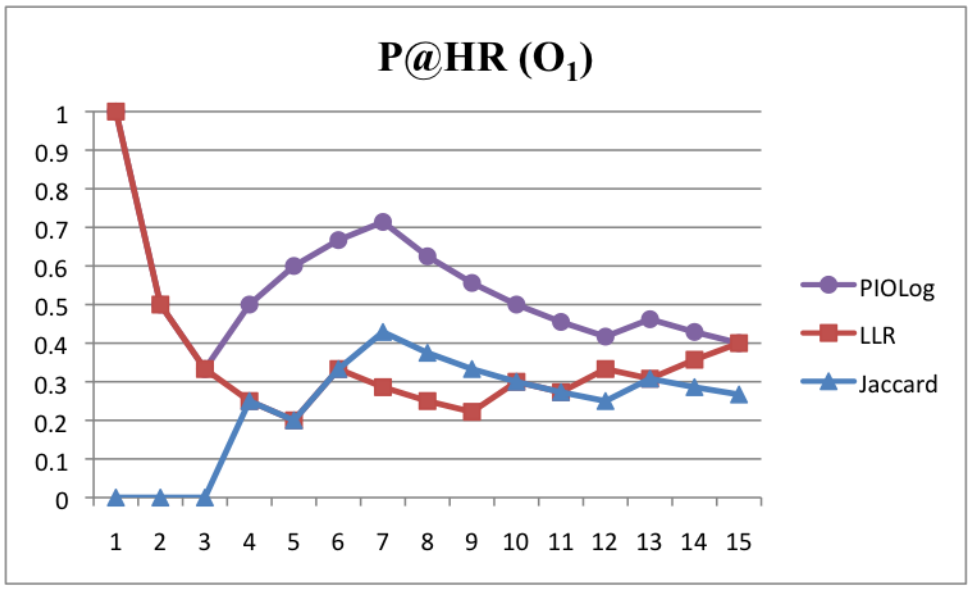


Figure 23. P@HR for O_1

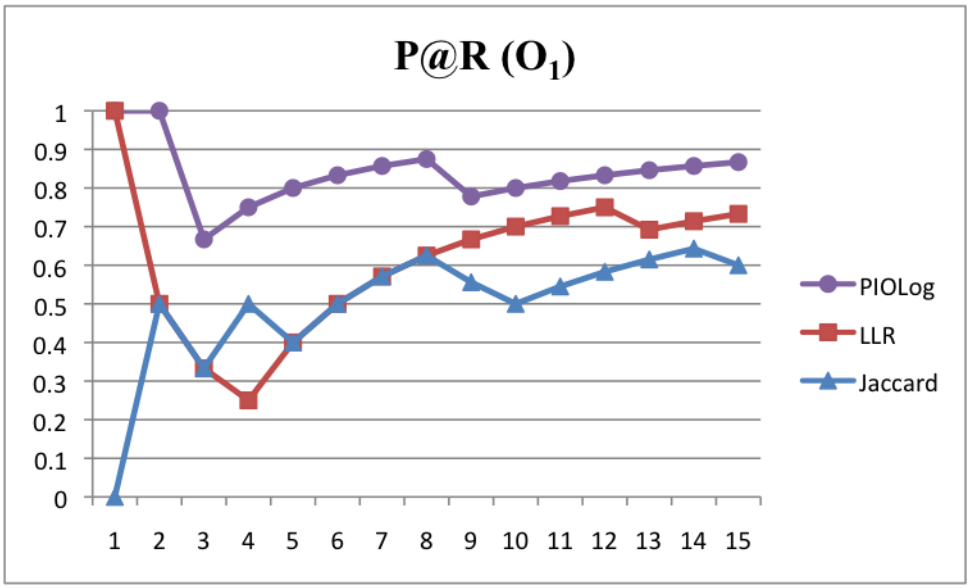


Figure 24. P@R for O_1

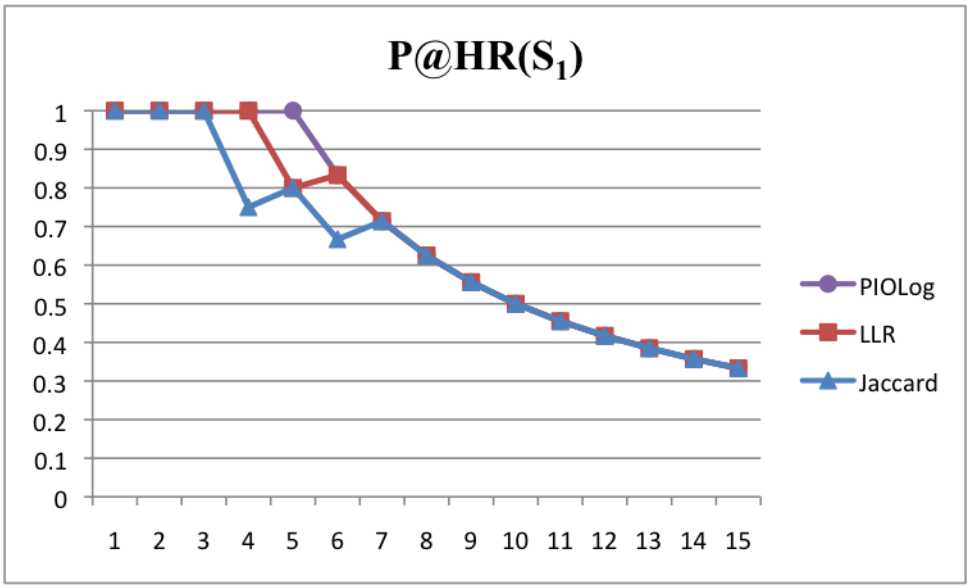


Figure 25. P@HR for S_1

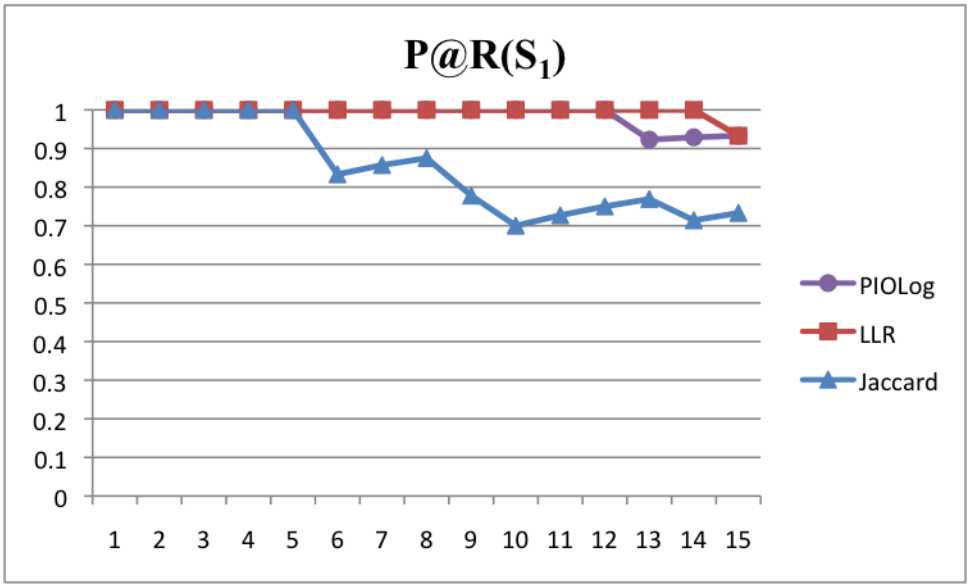


Figure 26. P@R for S_1

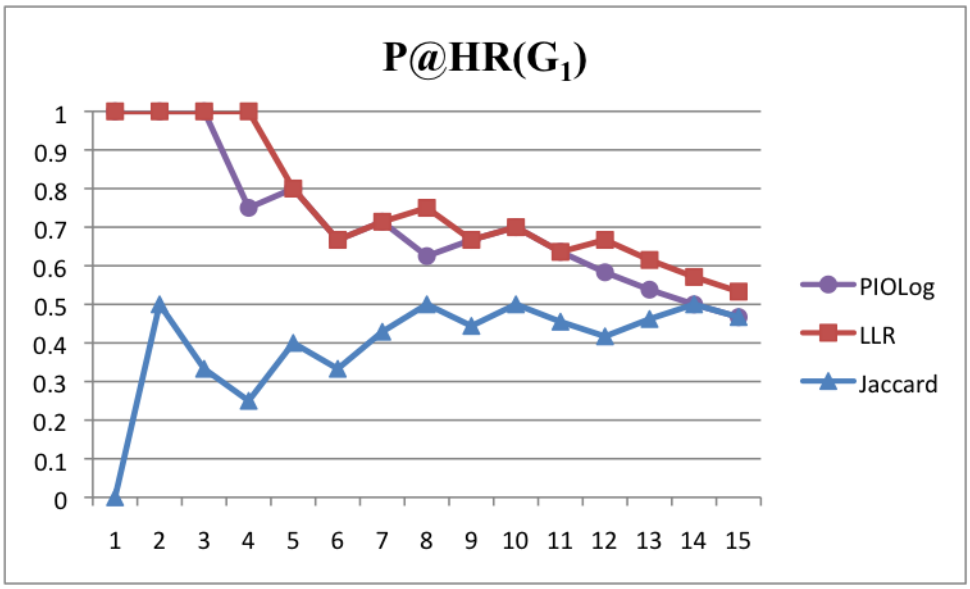


Figure 27. P@HR for G_1

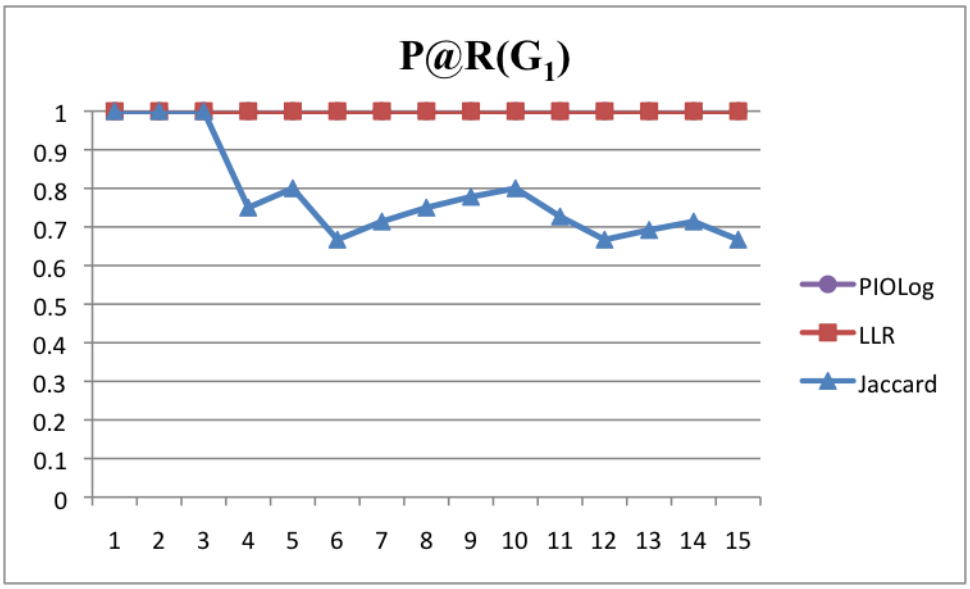


Figure 28. P@R for G_1

(Curves of PIOLog and LLR are overlapped)

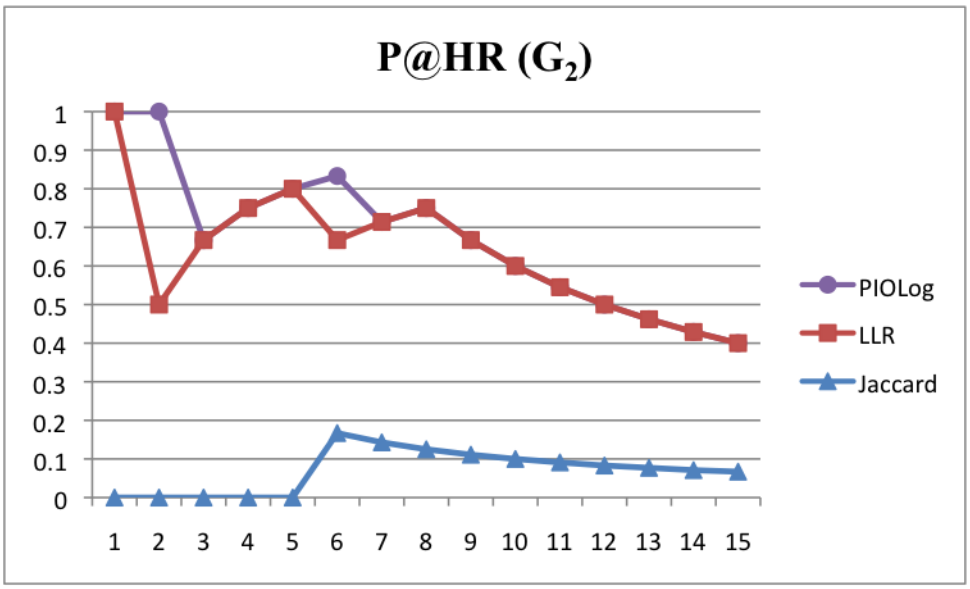


Figure 29. P@HR for G_2

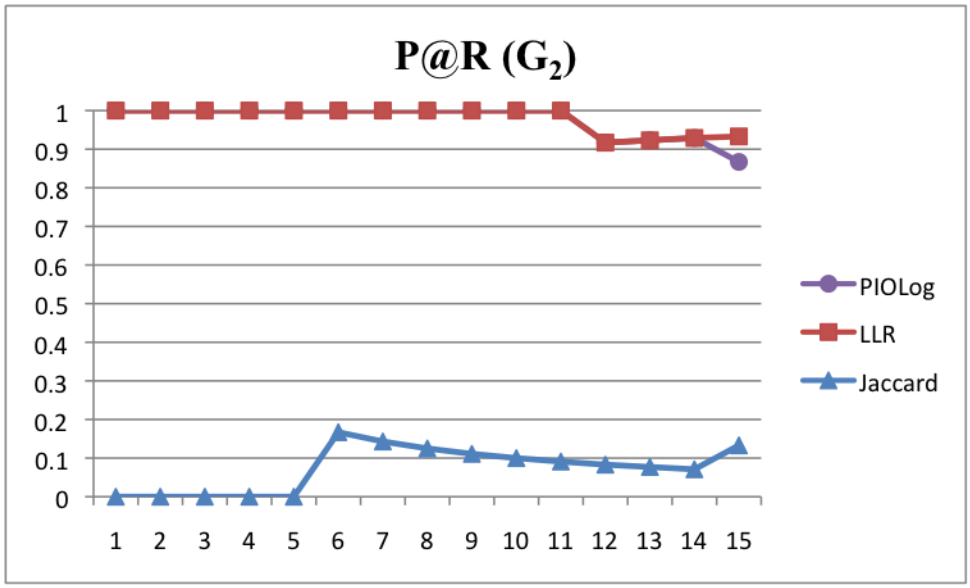


Figure 30. P@R for G_2

Table 6. Relevance scores for characteristic co-occurrence word

Score	Relevance
2	Completely relevant:
1	Partially relevant
0	Completely irrelevant

other news articles reporting the competition happened in other states and discussions for candidate's standpoint. Also, word "campaign" relates to not only presidential election campaign, but also other business campaigns like "Toyota conduct recall campaign in India". So word "campaign" is also assigned the score of 1. Assessors give zero score to the word of "performance". Although this word appeared in news articles of the topic, it is a general word that does not contain important information about the U.S. presidential election while it is also widely used in news articles of other topics like the performance of a player in Giants baseball team, and performance comparison between tablets.

After getting average relevance scores for each detected words, we evaluate effectiveness of these three methods by Discounted Cumulative Gain (DCG) [60]. DCG measures relevance between a list of characteristic co- occurrence words and the news topic related to the target word by considering not only relevance scores but also ranking position of each word in the list. The lower the ranked position of a word, the less relevance it is for the topic. Method whose DCG value is larger than the rest outperforms other methods. DCG is calculated as follows:

$$DCG_{20} = rel_1 + \sum_{i=2}^{20} \frac{rel_i}{\log_2 i} \tag{39}$$

where rel_i is the average relevance score of the i-th ranked word of the method. The DCG_{20} value ranges from 0 to 15.625. We calculate the DCG value of each method based on top-20 characteristic co-occurrence words for each news topic and results are shown in Figure 31.

As we can observe that PIOLog method performs better than other methods in most cases. It proves that PIOLog method is more likely to detect characteristic co-occurrence words strongly related to each news topic, and it will rank those characteristic co-occurrence words higher than the other methods. Jaccard and LLR perform not so well compared to PIOLog. As we explained before, characteristic co-occurrence word should be asymmetric to the target word, and whether a word often appears in other news articles not containing the target word is also important for judging the word is a characteristic co-occurrence word or not. Jaccard method does not satisfy these requirements. For LLR method, since it prefers words appearing in less number of news articles in the news topic, LLR method did not get a high DCG value in this evaluation because people are more sensitive to frequent-appearing words and they do not clearly remember words with low frequency. However, DCG value of PIOLog for news topic G_1 is lower than the value of Jaccard. We examined the co-occurrence word lists of these two methods and found that some words from

Table 7. The top-20 words (t = “Obama”, news topic O_1)

	Jaccard	**LLR**	**PlOLog**
1	Romney	Romney	Romney
2	Mitt Romney	Mitt Romney	Mitt Romney
3	voter	debate	debate
4	debate	presidential	voter
5	Paul Ryan	voter	presidential
6	presidential	candidate	Republican
7	Republican	campaign	candidate
8	Democrats	Republican	poll
9	Ohio	poll	Democrats
10	Biden	Paul Ryan	Paul Ryan
11	Joe Biden	Democrats	Ohio
12	candidate	Ohio	campaign
13	poll	Barack Obama	Barack Obama
14	Ryan	Biden	Biden
15	Barack Obama	Joe Biden	Joe Biden
16	Vice President	President	Ryan
17	Republicans	performance	race
18	abortion	Ryan	Republicans
19	campaign	race	President
20	presidency	election	Vice President

Table 8. The top-20 words (t = “Syria”, news topic S_1)

	Jaccard	**LLR**	**PlOLog**
1	Syrian	Syrian	Syrian
2	Ankara	Turkish	Turkish
3	Damascus	Damascus	Damascus
4	Turkish	Ankara	Ankara
5	Ahmet Davutoglu	Turkey	Turkey
6	Assad	border	jet
7	Turkey	jet	shell
8	artillery	shell	border
9	shell	Ahmet Davutoglu	rebel
10	airspace	Assad	fighter
11	Jordan	weapon	weapon
12	jet	military	Assad
13	rebel	rebel	civilian
14	mortar	fighter	Ahmet Davutoglu
15	Aleppo	conflict	Jordan
16	Bashar al-Assad	civilian	conflict
17	refugee	Jordan	troop
18	fighter	troop	Moscow
19	civilian	artillery	plane
20	border	airspace	regime

Table 9. The top-20 words (t = “game”, news topic G_1)

	Jaccard	**LLR**	**PIOLog**
1	inning	inning	inning
2	best-of-five	best-of-five	best-of-five
3	homer	series	pitch
4	postseason	pitch	homer
5	mound	homer	playoff
6	two-run	playoff	postseason
7	pitcher	postseason	series
8	playoff	two-run	ninth
9	baseman	mound	mound
10	pitch	pitcher	pitcher
11	Cincinnati	hit	score
12	Reds	win	two-run
13	hitter	ninth	World Series
14	ninth	score	Reds
15	World Series	baseman	ace
16	ace	Cincinnati	ball
17	Giants	Reds	Cincinnati
18	Oakland	World Series	out
19	strikeout	hitter	Giants
20	out	ace	baseman

Table 10. The top-20 words (t = “game”, news topic G_2)

	Jaccard	**LLR**	**PIOLog**
1	Wembley	England	England
2	Steven Gerrard	Wembley	Wembley
3	San Marino	San Marino	San Marino
4	Football Association	Football Association	Football Association
5	Manchester City	Steven Gerrard	Chelsea
6	Frank Lampard	Chelsea	World Cup
7	Joe Hart	FA	FA
8	St George s Park	Manchester City	Steven Gerrard
9	Roy Hodgson	Frank Lampard	Manchester City
10	FA	Roy Hodgson	captain
11	Wayne Rooney	Joe Hart	club
12	Hodgson	St George s Park	Poland
13	Hart	World Cup	striker
14	Cole	Wayne Rooney	Roy Hodgson
15	Chelsea	club	Manchester United
16	captaincy	Hodgson	qualifier
17	armband	Poland	football
18	Manchester United	captain	Frank Lampard
19	right-back	Manchester United	Joe Hart
20	Phil Jagielka	Hart	St George s Park

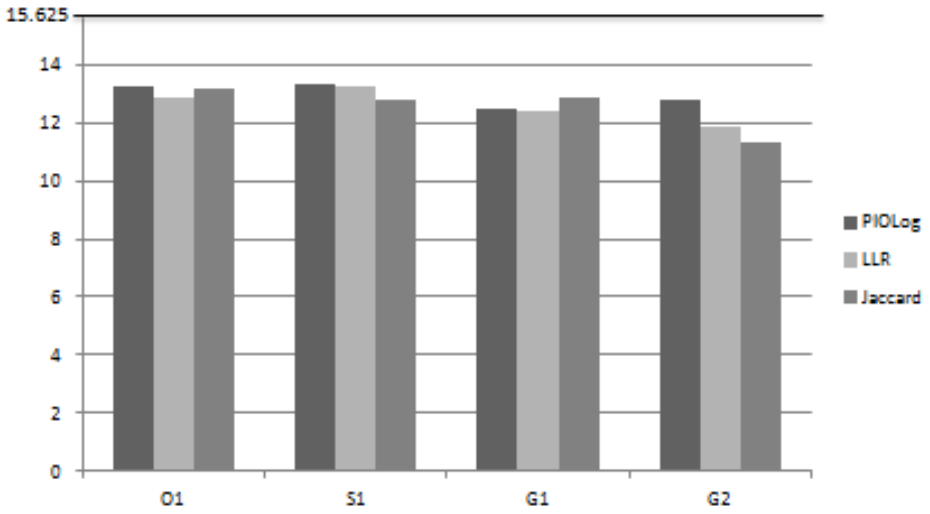


Figure 31. DCG_{20} values for news topic O_1, S_1, G_1, G_2

PIOLog like “ninth”, “score” and “out” are given the relevance score of 0 by some assessors. Assessors might think these words are general words and provide little information for G_1 about baseball match. However, when we examine the contents of news articles, sentences like “The A's were down to their last gasp in the ninth inning against the Detroit Tigers”, “Cardinals shut out Nationals; take 2-1 series lead” give important information for G_1 and “ninth” and “(shut) out” play important roles in these expressions. Assessors might not notice the context of these words and give a low score to them. Considering the context of co-occurrence words in evaluation would make PIOLog show more improvement compared with others.

6.4 Evaluation for News-Topic Oriented Hashtag Recommendation

In this section, we give the experiment for news-topic oriented hashtag recommendation. We compare our method with other related methods and show the outperformance of our method.

6.4.1 Experimental Setup

In order to evaluate recommended hashtags for news topics related to the target word and the effectiveness of our newly proposed PIOLogH method proposed in Chapter 4, we calculate the relevance between the news topic vector and hashtag vector by using different term weight methods. We use 6,868 news

articles crawled from news RSS feeds of 96 news providers in 21 countries/regions on October 11th, 2012 and 1,496,420 news related tweets collected on the same day.

We preprocessed all news articles in the same way as in the former section. All news articles are parsed by TreeTagger [58] and Stanford Named Entity Recognizer (SNER) [59] to extract terms. Then all news articles are clustered into news topics by HAC method. After user provides the target word, news topics related to the target word are selected. To represent the news topic *c* related to the target word *t*, traditional TF-IDF method (Section 4.3.1) and PIOLog method (Section 3.3) are used to weight terms to create a news-topic vector. For the TF-IDF method, we calculated the centroid vector of all news article vectors for this topic, and top-n words in the centroid vector which have higher TF-IDF scores than the rest are selected to create the news-topic vector. We refer to it as $\overrightarrow{nv}_{\text{TF-IDF}}(c,t)$. We also selected top-n words whose term weights are calculated by the PIOLog method, and the news-topic vector is created based on these top-n words. We refer to it as $\overrightarrow{nv}_{\text{PIOLog}}(c,t)$. Words which are informative and highly related to the news topic should be selected in these top-n words with larger term weight than the rest.

News related tweets are also preprocessed. Firstly, non-English tweets from Twitter Search API and tweets having no hashtag or written by non-native English users whose language setting in their Twitter profile is not set to “en” are excluded. Also tweets from those 54 Twitter accounts of news providers are excluded. Secondly, original tweets are selected to create the hashtag vector. Retweeted tweets (we take tweets which begin with “RT @username:” as retweeted tweets.) are not used here because Twitter users are not allowed to revise the tweet contents when they use the official “Retweet” function. Hashtags in these retweeted tweets could not reflect original ideas of Twitter users. Thirdly, for each hashtag in those tweets, we group tweets which contain the same hashtag and parse these tweets by using TreeTagger and SNER while mentions, URLs, and hashtags are excluded. After excluding hashtags used in less than 50 tweets and tagged screen name of news providers, 2,772 hashtags are selected. To represent these hashtags, we use TF-IHF method proposed in Section 4.4.1 to weight terms from tweets containing the same hashtag. Top-n terms whose TF-IHF scores are larger than the rest are selected to create the hashtag vector. We refer to it as $\overrightarrow{hv}_{\text{TF-IHF}}(ht)$ for hashtag *ht*. We also use newly proposed PIOLogH method in Section 4.4.2 to weight terms in the hashtag vector. Top-n terms whose PIOLogH scores are larger than the rest are used in the vector. We refer to it as $\overrightarrow{hv}_{\text{PIOLogH}}(ht)$.

6.4.2 Comparison with Related Methods

We choose the same target word, “Obama”, “Syria”, and “game” with their related news topics (O_1, S_1, G_1, and G_2) described in Table 4. For each news topic, we

set four experiments with different combinations of term weighting methods to calculate similarities between news topics and hashtags (Figure 32).

- Exp. 1: $\overrightarrow{nv}_{\text{TF-IDF}}(c,t) \cdot \overrightarrow{hv}_{\text{TF-IHF}}(ht)$. Term weight for news-topic vector is the TF-IDF score and term weight for hashtag vector is the TF-IHF score.

- Exp. 2: $\overrightarrow{nv}_{\text{TF-IDF}}(c,t) \cdot \overrightarrow{hv}_{\text{PIOLogH}}(ht)$. Term weight for news-topic vector is the TF-IDF score and term weight for hashtag vector is the PIOLogH score.

- Exp. 3: $\overrightarrow{nv}_{\text{PIOLog}}(c,t) \cdot \overrightarrow{hv}_{\text{TF-IHF}}(ht)$. Term weight for news-topic vector is the PIOLog score and term weight for hashtag vector is the TF-IHF score.

- Exp. 4: $\overrightarrow{nv}_{\text{PIOLog}}(c,t) \cdot \overrightarrow{hv}_{\text{PIOLogH}}(ht)$. Term weight for news-topic vector is the PIOLog score and term weight for hashtag vector is the PIOLogH score.

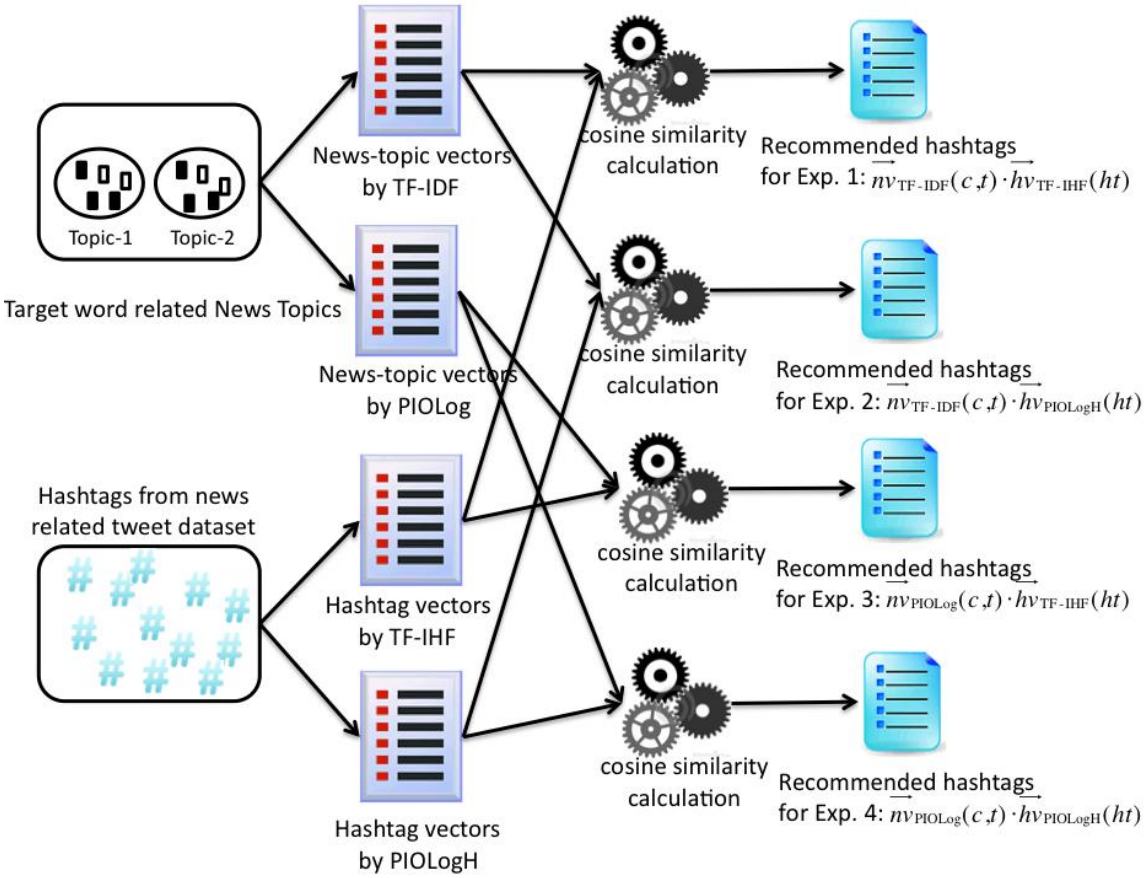


Figure 32. Experiments for hashtag recommendation

In each experiment, top-n words whose term weights are larger than the rest for news topics and hashtags are selected to create vectors with n equals 400. Methods which outperform others are considered ranking those topic-specific informative words higher and hashtags recommended by these methods are considered more relevant to the news topic.

To evaluate recommended hashtags from four experiments, we ask assessors to judge the relevance of the recommended hashtags to each news topic. To help our assessors better understand the news topic, they could scan/search for any information if they need to make a proper decision. The whole procedure is shown as below.

1. Two assessors are asked to read at least ten news articles which are carefully selected for each news topic so that these news articles can cover the main contents of the news topic to make them understand the news topic.

2. Top-15 hashtags with largest similarities recommended by each of four experiments are mixed to form a hashtag list for each news topic. Assessors judge the relevance of each hashtag in this list to the news topic on a three-point scale: highly relevant, partially relevant and irrelevant. They can use any tool (e.g. TagDef [56] or Google [57]) to find definitions for hashtags.

3. For each news topic, hashtags which are judged as highly relevant by two assessors are defined as highly relevant hashtags. We also define relevant hashtags as they should not be judged as irrelevant by any of assessors. Notice that highly relevant hashtags are a sub-set of relevant hashtags.

For example, for the new topic of O_1 about U.S. presidential election (Table 11), “#election2012” is taken as the highly relevant hashtag about the news topic because tweets containing this hashtag mainly relate to the presidential election. However, “#politics” which is often used in tweets about political issues is considered as relevant hashtag since it is not only about O_1, but also other political topics. Hashtags such as “#mostrecent” and “#libyagate” used for other purposes or news topics are judged as irrelevant hashtag.

To evaluate performances of four experiments for these news topics, we use precision as the evaluation metric under two-levels:

- **Precision at highly relevance curve** (**P@HR curve**): each point on this curve indicates the fraction of top-r recommended hashtags that are highly relevant hashtags for the news topic.

- **Prevision at relevance curve** (**P@R curve**): each point on this curve indicates the fraction of top-r recommended hashtags that are relevant hashtags for the news topic.

The value of r ranges from 1 to 15. Experiment whose P@HR and P@R curves locate higher than the rest should be the best one for recommending hashtags for news topics, and term weighting methods used in this experiment outperform other methods for detecting characteristic co-occurrence words. We only consider precision here because recall and F-measure have the same result.

We also use DCG [60] to measure the performance of these four experiments. The method whose DCG score of top-15 hashtags are larger than the rest for these four news topics outperforms.

6.4.3 Evaluation

We select top-n words whose term weights are larger than the rest by using different term weighting methods to create vectors for representing news topics and hashtags. Here n equals 400 estimated in parameter estimation section. Four experiments described in the former section are used to recommend hashtags for four news topics (O_1, S_1, G_1, and G_2). Results for these four news topics with highly relevant and relevant hashtags are described in Table 11 - 14 (all hashtags are written in low case since the hashtag is capitalize insensitive, and "#" symbol is ignored in these tables).

Figure 33 - Figure 40 show the P@HR and P@R curves for four news topics based on top-15 recommended hashtags from four experiments. For example, in Table 11 about news topic O_1, for these top-5 hashtags recommended from Exp. 4, three of them (romneyryan2012, election2012, mitt2012) are judged as highly relevant hashtags and the precision when r = 5 is calculated as 3 divided by 5 equals 0.6 (Figure 33). Also top-5 hashtags recommended from Exp. 1 only contain two highly relevant hashtags (debate2012, 2012election) and its precision for r = 5 is calculated as 2 divided by 5 equals 0.4 (Figure 33). Precision for relevant hashtags is calculated in the same way. For each news topic, precisions at r which ranges from 1 to 15 are calculated to draw P@HR and P@R curves.

As we can observe from results above, P@HR and P@R curves of Exp. 4, which apply our proposed methods based on Inside and Outside assumptions to both news topic (Section 3.3) and hashtags (Section 4.4.2), locates higher than curves of other experiments. This indicates that hashtags recommended by the Exp. 4 are more meaningful than hashtags recommended by other experiments.

Applying our proposed methods only for hashtags or news topics in Exp. 2 and Exp. 3, results show an improvement compared to the Exp. 1 though they still perform not so well compared to Exp. 4. These improvements also show that our Probabilistic Inside-Outside Log methods have positive affection to the recommended hashtags.

We also use Discounted Cumulative Gain to evaluate recommended hashtags from these four experiments. We get the same conclusion. For DCG_{15} values of these four experiments for four news topics (Figure 41 - Figure 44) they show that recommended hashtags from Exp. 4 are more relevant to news topics.

Table 11. Recommended hashtags from four experiments for O_1

Exp. 1	Exp. 2	Exp. 3	Exp. 4
dateline	politics	libyagate	romneyryan2012
libyagate	tcot	joebiden	politics
activismrocks	romneyryan2012	obamaisntworking	tcot
debate2012	teaparty	2012election	election2012
2012election	p2	ia	mitt2012
teamfollback	mitt2012	flipflop	romney
ia	election2012	etchasketch	romney2012
joebiden	romney2012	debate2012	gop
fourmoreyears	gop2012	politicsnation	p2
therealromney	tlot	nobama2012	teaparty
mostrecent	debates	bias	obama
rr2012	debate	fourmoreyears	debate
obamaisntworking	gop	mostrecent	mittromney
nobama2012	obama	connecttheleft	election
etchasketch	obama2012	therealromney	mitt
Highly relevant hashtags (HR)	election2012, mitt2012, romney2012, mittromney, nobama2012, therealromney, romneyryan2012, romney, debate2012, obama2012, 2012election		
Relevant hashtags (R)	2012election, mitt2012, mitt, teaparty, therealromney, debates, debate2012, obama2012, politics, tlot, mittromney, romney, debate, joebiden, obama, obamaisntworking, romney2012, nobama2012, gop2012, romneyryan2012, rr2012, election2012, election, fourmoreyears, gop		

Table 12. Recommended hashtags from four experiments for S_1

Exp. 1	Exp. 2	Exp. 3	Exp. 4
ankara	syria	damascus	syria
deirezzor	turkey	ankara	turkey
damascus	syrian	homs	syrian
idlib	assad	deirezzor	damascus
homs	damascus	idlib	assad
lattakia	fsa	syrian	aleppo
rastan	jordan	hama	idlib
hama	nato	turkish	ankara
aljazeera	ankara	lattakia	homs
sham	middleeast	rastan	fsa
moscow	idlib	sham	russia
turkish	russia	moscow	nato
syrian	worldnews	aljazeera	jordan
plane	aleppo	plane	deirezzor
middleeast	homs	aleppo	turkish
Highly relevant hashtags (HR)	syrian, assad, damascus, turkey, syria		
Relevant hashtags (R)	fsa, syrian, ankara, aleppo, turkish, nato, homs, russia, moscow, assad, deirezzor, turkey, damascus, idlib, syria		

Table 13. Recommended hashtags from four experiments for G_1

Exp. 1	Exp. 2	Exp. 3	Exp. 4
playoffs	mlb	mlbplayoffs	mlb
mlbplayoffs	giants	postseason	nlds
12in12	postseason	tigers	giants
tigers	cardinals	baseball	reds
postseason	reds	playoffs	postseason
stlcards	sfgiants	nlds	cardinals
cards	nlds	12in12	sfgiants
alds	nats	stlcards	nats
reds	nationals	reds	baseball
baseball	baseball	nationals	nationals
orangeoctober	stlcards	alds	stlcards
nationals	playoffs	sfgiants	playoffs
orioles	12in12	cards	tigers
nlds	orangeoctober	giants	orioles
athletics	athletics	orioles	12in12
Highly relevant hashtags (HR)	baseball, mlbplayoffs, postseason, nlds, sfgiants, nationals, giants, mlb		
Relevant hashtags (R)	orangeoctober, stlcards, postseason, alds, nlds, sfgiants, orioles, cardinals, nats, reds, baseball, 12in12, mlbplayoffs, nationals, playoffs, giants, tigers, cards, mlb		

Table 14. Recommended hashtags from four experiments for G_2

Exp. 1	Exp. 2	Exp. 3	Exp. 4
mufc	football	cfc	football
cfc	england	mufc	chelsea
fifa	sport	chelsea	england
safc	lfc	arsenal	mufc
chelsea	chelsea	blamesuarez	cfc
cricket	soccer	fifa	lfc
blamesuarez	mufc	safc	sport
flag	rugby	soccer	nufc
fa	blamesuarez	tdf	soccer
san	fifa	cricket	rugby
arsenal	cfc	natitude	blamesuarez
mostrecent	cricket	ynwa	nats
soccer	fa	wonga	arsenal
scarf	nufc	belgium	fifa
country	safc	fa	postseason
Highly relevant hashtags (HR)	mufc, lfc, football, nufc, cfc, chelsea		
Relevant hashtags (R)	mufc, arsenal, sport, fifa, nufc, rugby, chelsea, england, football, lfc, ynwa, wonga, blamesuarez, soccer, natitude, cfc		

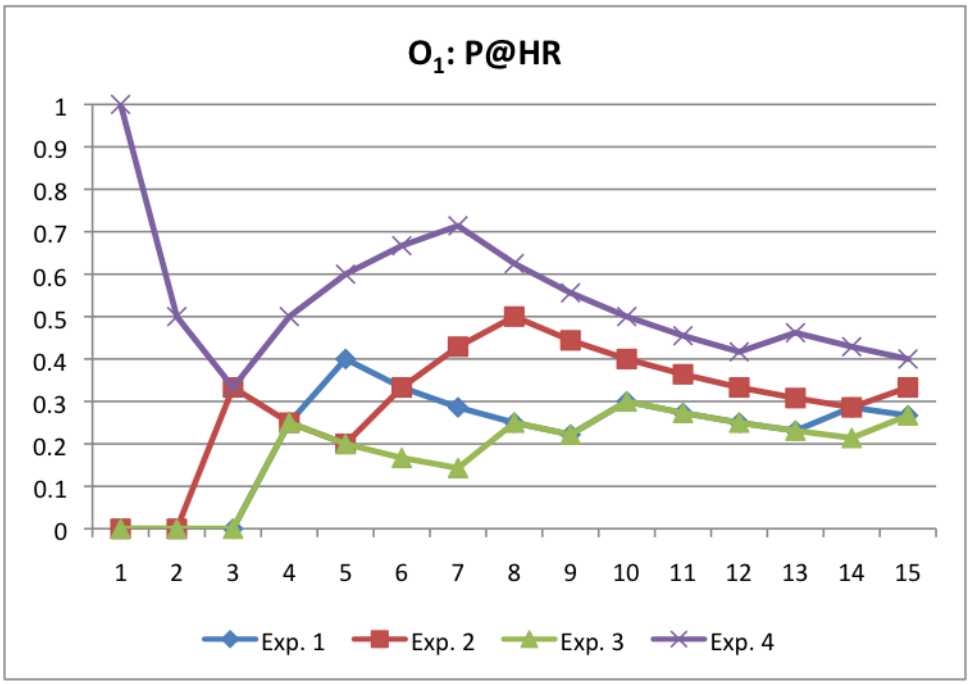


Figure 33. P@HR curve for O_1 with top-15 hashtags

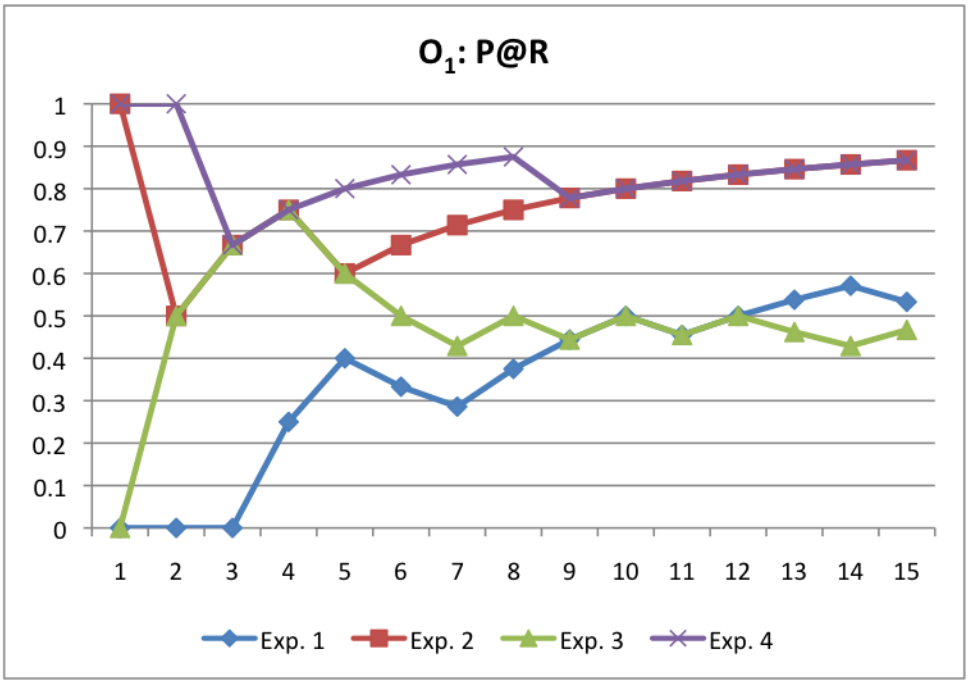


Figure 34. P@R curve for O_1 with top-15 hashtags

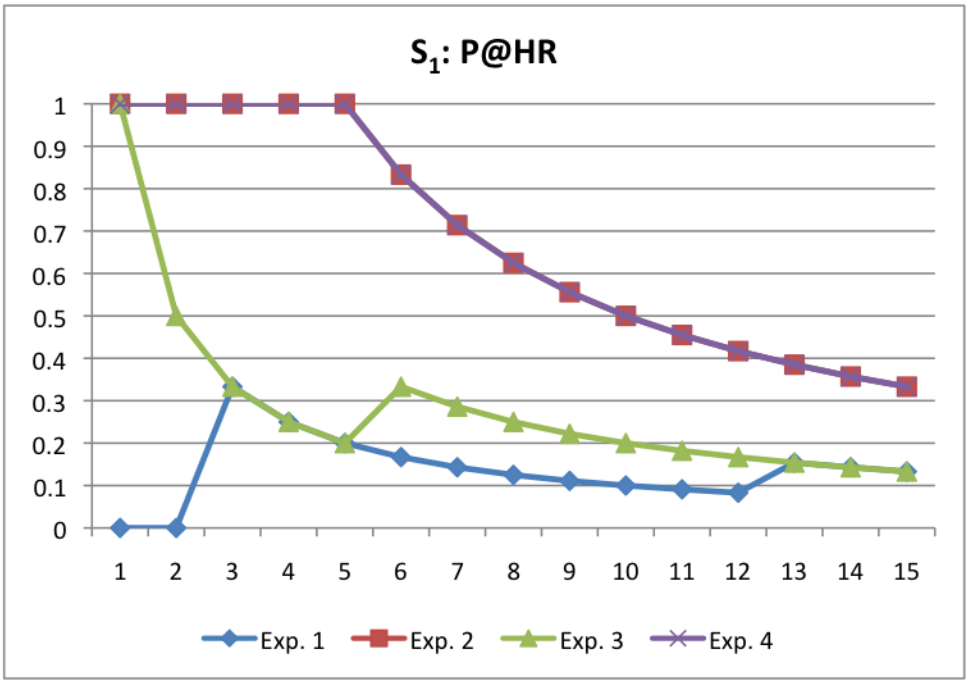


Figure 35. P@HR curve for S_1 with top-15 hashtags

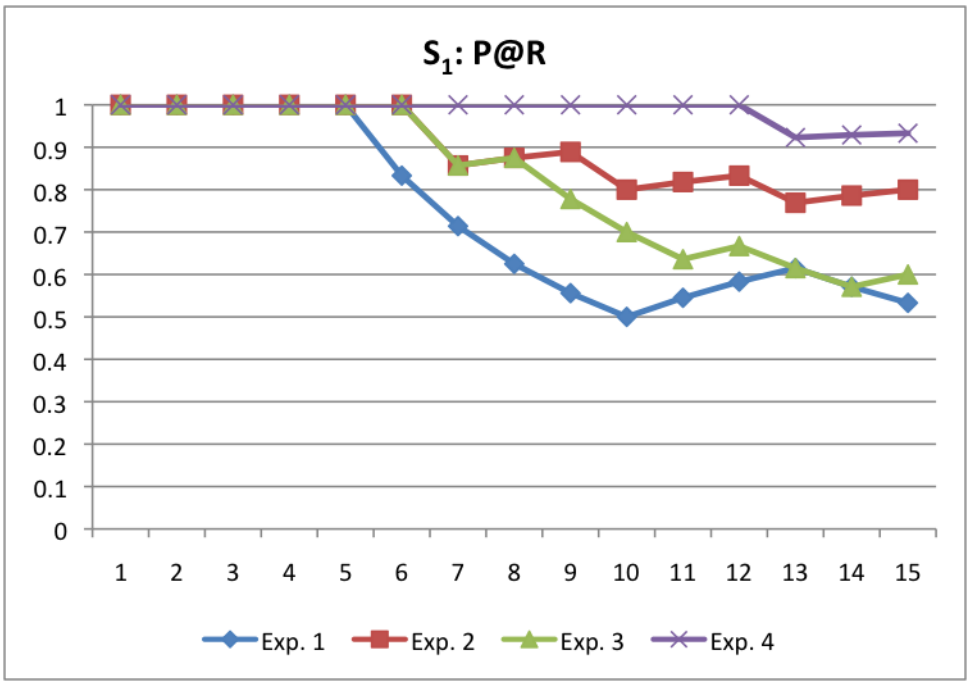


Figure 36. P@R curve for S_1 with top-15 hashtags

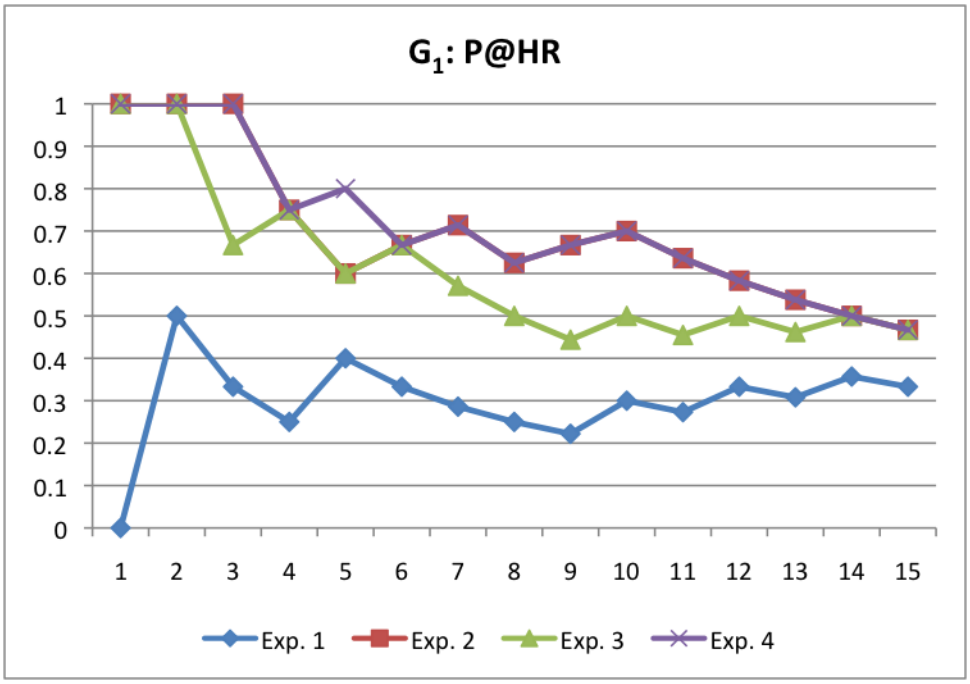


Figure 37. P@HR curve for G_1 with top-15 hashtags

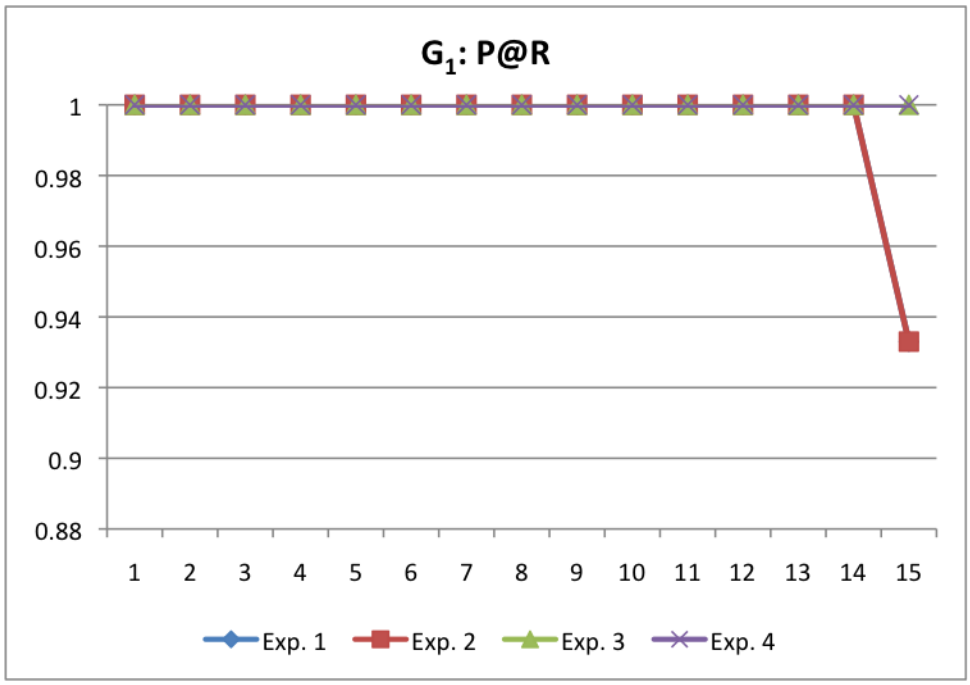


Figure 38. P@R curve for G_1 with top-15 hashtags
(Curves of Exp.1 and Exp. 2, Exp. 3 and Exp. 4 get overlapped here)

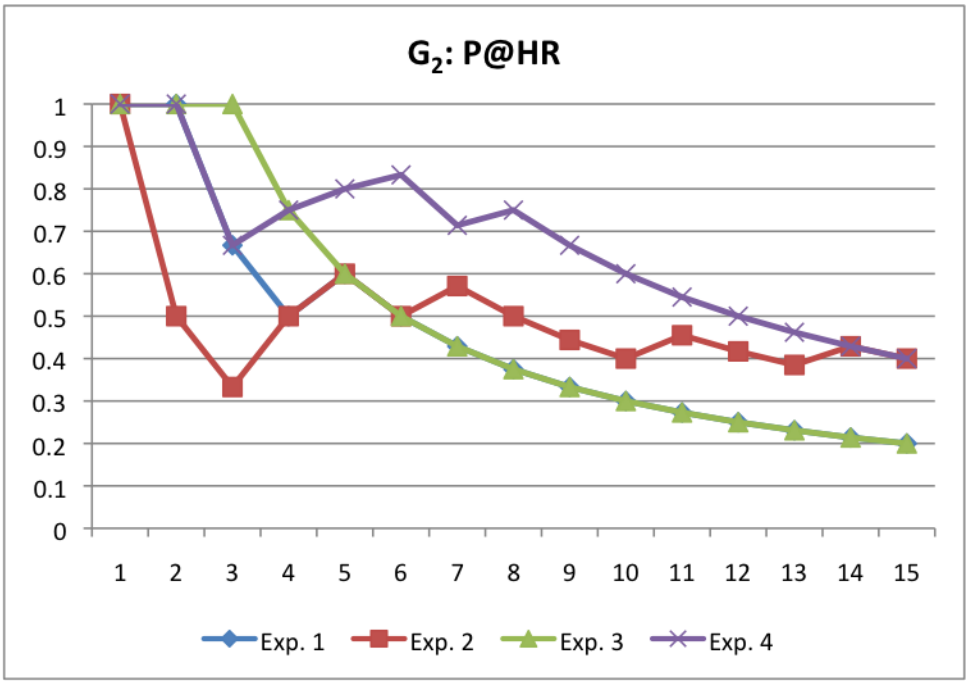


Figure 39. P@HR curve for G_2 with top-15 hashtags

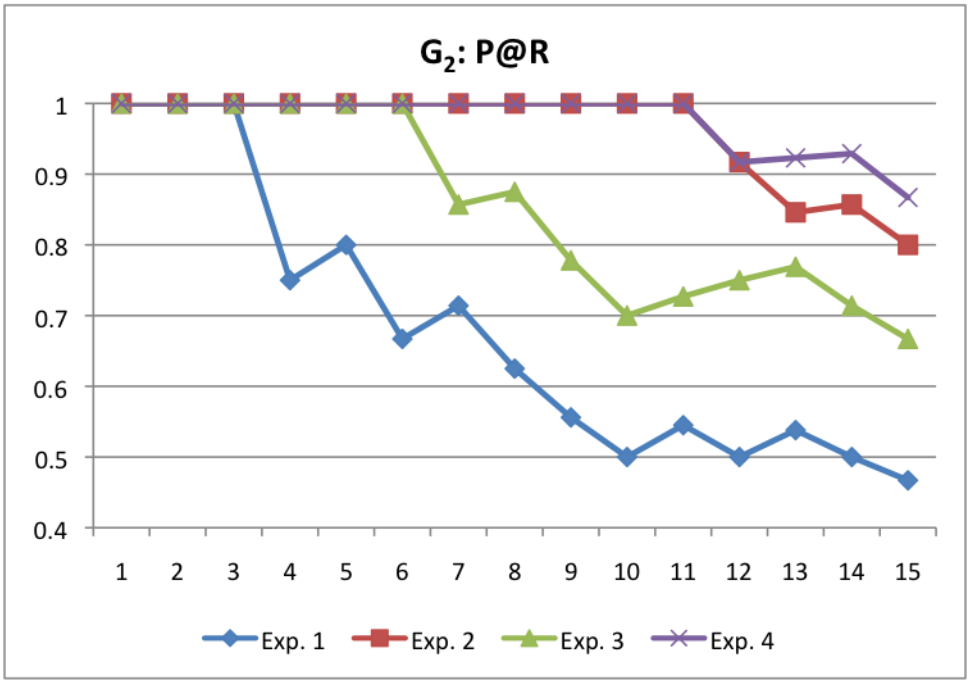


Figure 40. P@R curve for G_2 with top-15 hashtags

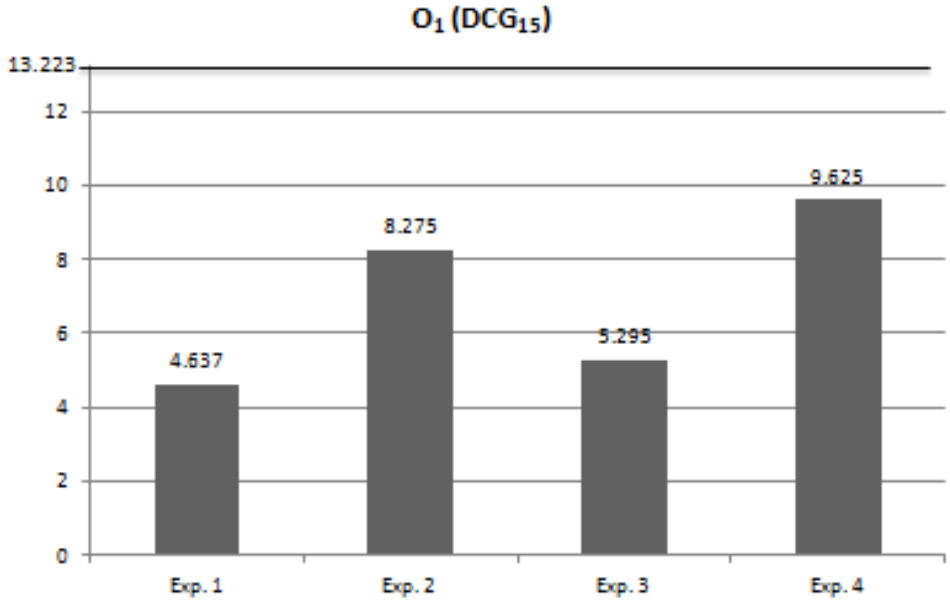


Figure 41. DCG_{15} values for news topic O_1

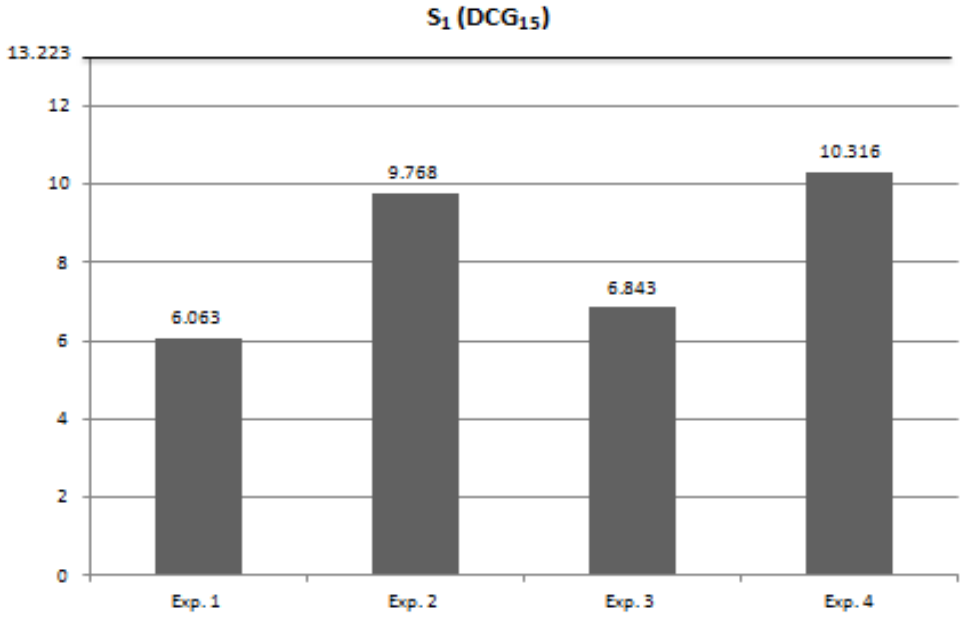


Figure 42. DCG_{15} values for news topic S_1

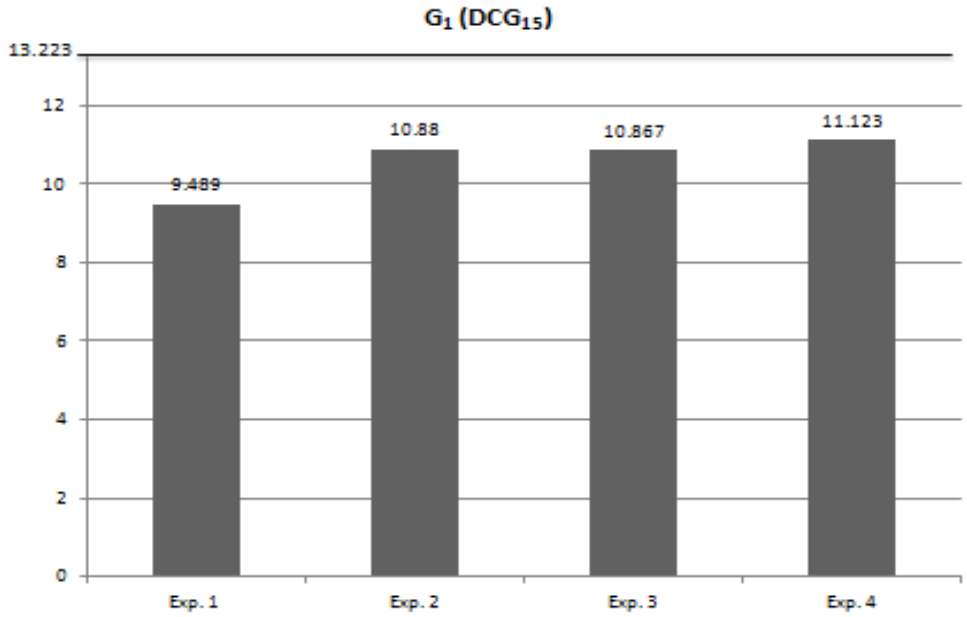


Figure 43. DCG_{15} values for news topic G_1

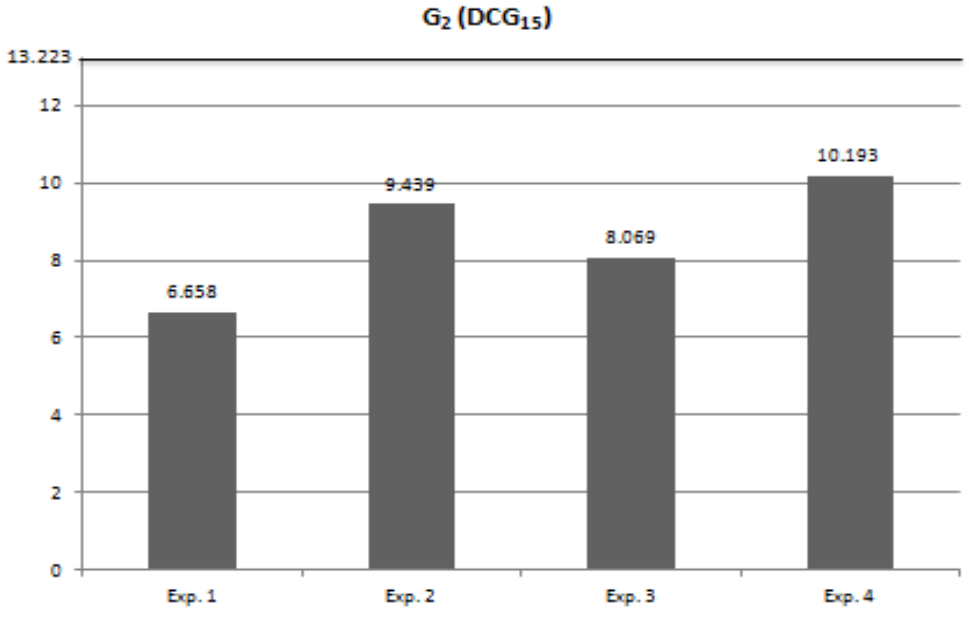


Figure 44. DCG15 values for news topic G_2

Someone may think that an intuitive approach for hashtag recommendation is to simply collect tweets containing the target word and recommend hashtags which are frequently used in these tweets. However, due to the length limitation, a tweet related to the news topic might not contain the target word. Also, tweets containing the target word might not relate to the news topic since the target word may correspond to multiple topics. In order to prove the infeasibility of this intuitive approach and show the problem brought by the tweet length limitation, we prepared two experiments.

- Precision experiment (Figure 45): in this experiment, we randomly select 100 tweets containing the target word and manually check how many of them relate to the news topic.

- Recall experiment (Figure 46): in this experiment, we manually select 100 tweets related to each news topic and check how many of them contain the target word.

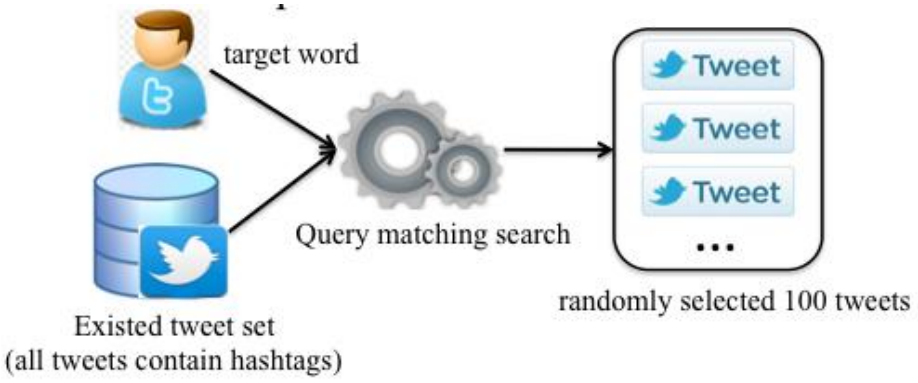


Figure 45. Precision experiment

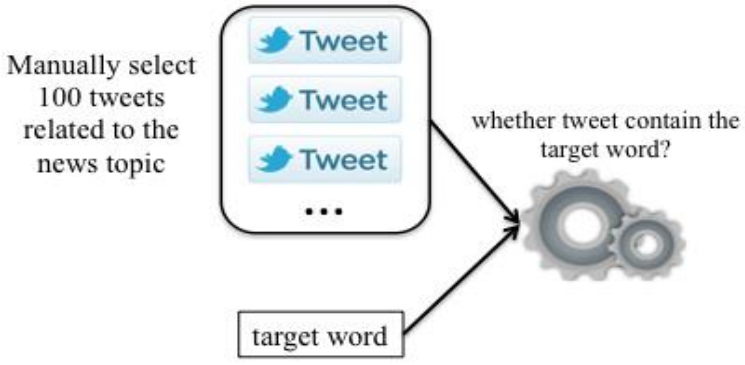


Figure 46. Recall experiment

We did these two experiments for four news topics. Results are shown in Table 15 and Table 16. As we can observe that precisions for these news topics are not high especially for the target word with ambiguous meaning like “game”.

We manually check tweets containing the target word of “game” and find that many tweets related to iPad/iPhone games are selected. Also, the recall for these topics is very low, which means most of tweets related to the news topic do not contain the target word. If we select news topic related tweets by simply using string match of the target word, some selected tweets are not related to the news topic while most tweets related to the news topic would get missed.

To check the affection of low precision/recall to the hashtag recommendation, we prepared a comparative experiment (#Tweet). Figure 47 gives the procedure of this experiment. After user provides the target word, we select all tweets from news-related tweet dataset that contain the target word. Then hashtags from these selected tweets are ranked in decreasing order of the number of tweets containing the hashtag. Hashtags which are frequently used in these tweets get recommended.

We use the same evaluation metrics as described in Section 6.4.2. P@HR and P@R curves are averaged for our four news topics. We compare results of comparative experiment with results of our approach (Exp. 4 in Section 6.4.2). Figure 48 and Figure 49 show evaluation results of these two experiments. As we can observe that our approach whose P@HR and P@R curves locate higher performs better than the comparative experiment.

Table 15. Precision of tweets for each news topic

News topic	#Tweet related to news topic
O_1	71/100
S_1	77/100
G_1	45/100
G_2	5/100

Table 16. Recall of tweets for each news topic

News topic	#Tweet related to news topic
O_1	2/100
S_1	0/100
G_1	14/100
G_2	3/100

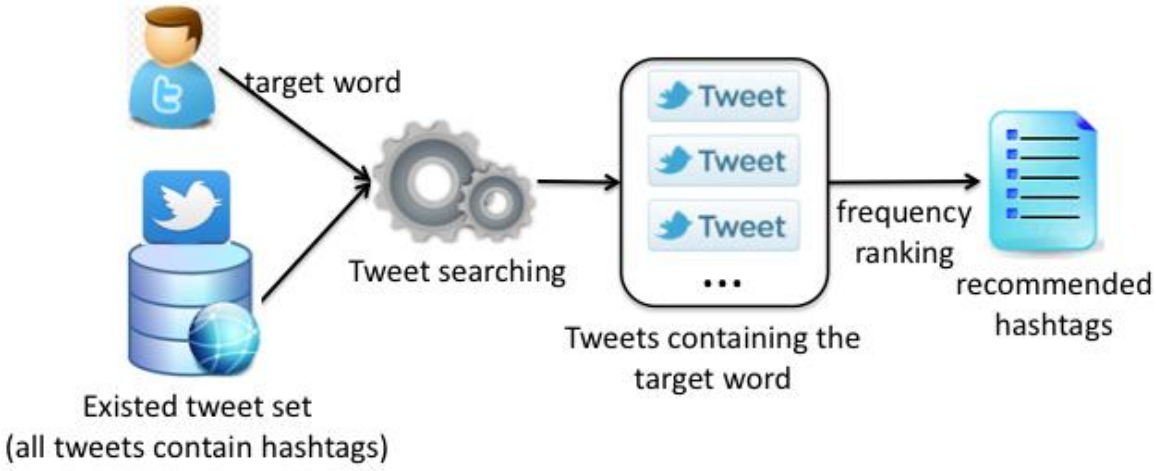


Figure 47. Comparative experiment (#Tweet) for hashtag recommendation

After we manually examined recommended hashtags from these two experiments, we found that there are mainly three reasons why the comparative experiment performs not well. Firstly, since one target word may correspond to more than one news topic, hashtags being relevant to different topics might get mixed together. Secondly, some general hashtags, such as #news, #breakingnews, would get ranked high in comparative experiment. Obviously they are not topic dependent. Lastly, since tweets related to the news topic do not necessarily contain the target word, many tweets related to the news topic are missed.

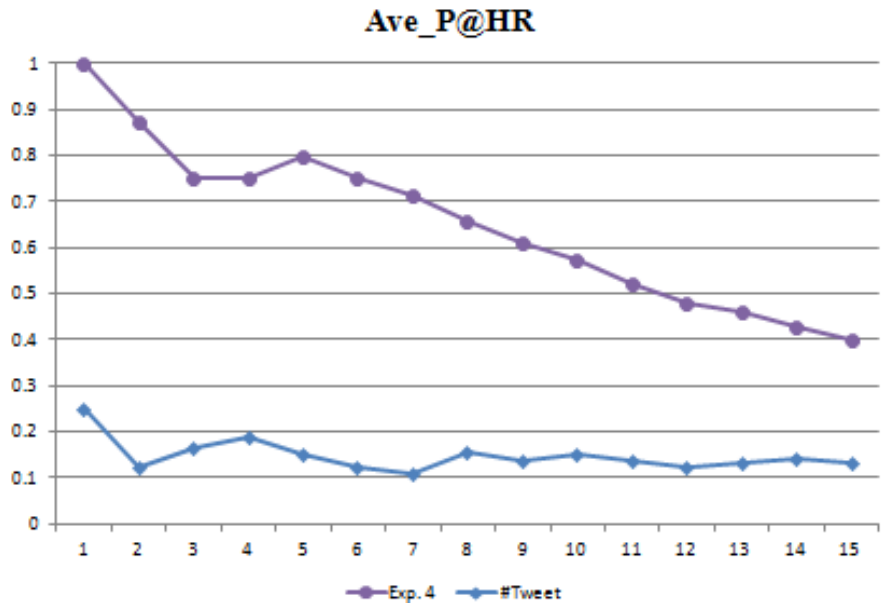


Figure 48. Average P@HR curves of newly proposed approach (Exp. 4) and comparative experiment (#Tweet)

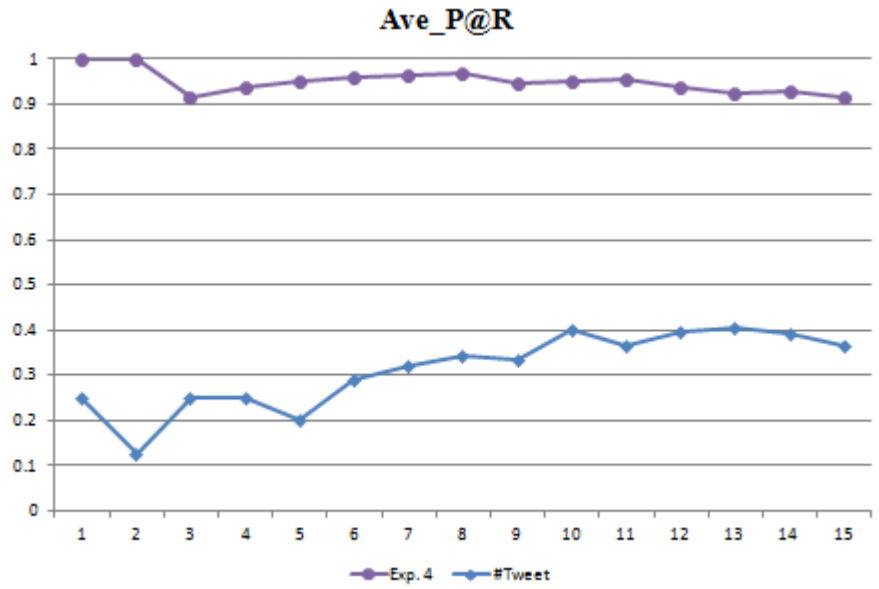


Figure 49. Average P@R curves of newly proposed approach (Exp. 4) and comparative experiment (#Tweet)

6.5 Evaluation for Finding News-Topic Oriented Influential Twitter Users

In this chapter, we give the experiment for finding news-topic oriented influential Twitter users. We compare RetweetRank and MentionRank with other related methods. Experimental results show that our methods outperform others for finding content-based and authority-based influential Twitter users.

6.5.1 Experimental Setup

We collect news articles and news related tweets concurrently on October 11th, 2012 for the experiment. There are 6,868 news articles and 1,496,420 news related tweets collected.

News articles and news related tweets are preprocessed in former sections (Section 6.3.1 and Section 6.4.1). After excluding hashtags used in less than 50 tweets and tagged screen name of news providers, 2,772 hashtags with corresponding hashtag communities are selected. On average, there are 363.32 tweets posted by 235.46 users for each hashtag. Tweets posted by users containing the same hashtag in each hashtag community are parsed into terms by using TreeTagger and SNER while mentions, URLs, and hashtags are excluded. We use the approach described in Section 4.5 to detect hashtags being relevant to the news topic related to the target word. Hashtags whose relevance scores are larger than a predefined threshold th_{ht} are taken as news-topic-related hashtags. Hashtag communities defined by these hashtags are taken as news-topic-related hashtag communities.

Retweet graph G_{RT} and mention graph G_{MN} are created among users in these news-topic-related hashtag communities. Retweeted tweets and tweets containing mentions of other users are selected to create relations among users in G_{RT} and G_{MN}. Topic related teleportation vectors for each news topic are also created based on tweets posted by users in G_{RT} and G_{MN}. The damping factor d is set to 0.85, the same value as in PageRank. The threshold ε for stopping power iteration is set to 0.00005 since it does not affect results too much. We choose “Obama”, “Syria”, and “game” as target words. News topics related to each target word are selected. Summaries of these topics are described in Table 4. There are four news topics selected and denoted by O_1, S_1, G_1, G_2, including two of them (G_1 and G_2) relate to the target word of “game”.

For each news topic, we create retweet graph and mention graph. Table 17 shows the detailed information. $|c|$ gives the number of news articles in the news topic. $|H_c|$ gives the number of news-topic-related hashtag communities. $|V_{RT}|$ and $|V_{MN}|$ show the number of vertices (users) in retweet and mention graphs, and $|E_{RT}|$ and $|E_{MN}|$ show the number of edges in these two graphs. RetweetRank and MentionRank are applied to the retweet and mention graph of

Table 17. Data about retweet and mention graph for news topic *c*

Topic	$\|c\|$	$\|H_c\|$	$\|V_{RT}\|$	$\|E_{RT}\|$	$\|V_{MN}\|$	$\|E_{MN}\|$
O_1	179	112	3691	6924	4307	6560
S_1	86	20	464	996	439	534
G_1	65	24	176	131	403	321
G_2	33	6	80	63	178	147

each news topic. Twitter users whose RetweetRank scores or MentionRank scores are larger than the rest are taken as content-based or authority-based influential users.

6.5.2 Comparison with Related Methods

In this section, we discuss related methods for finding content-based and authority-based influential Twitter users. To find these two types of influential Twitter users, comparison against related methods are conducted. Other methods used to find these two types of influential Twitter users are described as follows:

1. Tweet number. This method measures Twitter user's influence based on the number of tweets posted by the Twitter user containing news-topic-related hashtags. More tweets posted by the user, more influential the user would be.

2. In-degree. This method measures Twitter user's influence based on the number of times the Twitter user gets retweeted/mentioned by others in retweet/mention graph. More times the user gets retweeted/mentioned, more influential the user would be.

3. PageRank [51]. This method measures Twitter user's influence in the retweet graph and mention graph by using PageRank algorithm. However, relevance between users and news topic are ignored in its teleportation vector. User's retweet/mention preferences are also not considered when calculating transition probabilities. Larger the PageRank score of a user has more influential the user would be.

For ease of presentation, RetweetRank and MentionRank are denoted by RR and MR. Method using the number of posted tweets is denoted by TN. In-degree is denoted by IND and PageRank is denoted by PR.

6.5.3 Evaluation

In this section, we show our evaluation results of finding content-based and authority-based influential Twitter users. We also discuss performances of different methods.

To evaluate the effectiveness of our newly proposed RR and MR, we apply TN, IND, PR, and RR to the retweet graph of each news topic to find content-based influential Twitter users. We also apply TN, IND, PR, and MR to mention graph of each news topic to find authority-based influential Twitter users.

To evaluate content-based and authority-based influential Twitter users found by different methods for the news topic *c*, we select top-15 users from each method and ask two assessors to manually assign content-influential score or authority-influential score for each user on a three-point scale. Definitions of the content-influential score and authority-influential score are described in Table 18 and Table 19. Then Discounted Cumulative Gain (DCG) [60] for top-15 users found by each method is calculated as follow:

$$\mathrm{DCG}_{15} = score_1 + \sum_{i=2}^{15} \frac{score_i}{\log_2 i} \tag{40}$$

where $score_i$ is the content-influential score or authority-influential score manually assigned by assessors for the *i*-th Twitter user. DCG considers not only user's influential score, but also their ranking position. Method whose DCG value is larger could rank users often posting valuable tweets or having high authority on the news topic higher, and outperform other methods whose DCG values are small.

We give top-15 content-based influential Twitter users for each news topic detected by TN, ING, PR, and RR in Table 20, 22, 24, and 26. Top-15 authority-based influential Twitter users for each news topic detected by TN, IND, PR, and MR are in Table 21, 23, 25, and 27. Before showing evaluation results, someone may think that tweets posted by authority-based influential Twitter users may also be valuable and get retweeted many times because these tweets are highly trustable. However, as we can observe, there are few users being taken as both content-based and authority-based influential Twitter users. One reason is that most of tweets posted by authority-based influential users often concern about latest evolvement of the news topic while tweets from content-based influential users are more opinionated and more likely to attract other's interests since Twitter users often share opinions on variety of topics and discuss current issue [61].

For example, for the news topic O_1 about U.S. presidential election searched by the target word of "Obama", assessors assign content-influential score or authority-influential score for each ranked user. For content-based influential Twitter users of O_1, @PatDollard is assigned scores of 2 by both assessors. That's because he is a famous Twitter user who often share his opinions about the presidential election, and attract many others who often retweet his tweets,

Table 18. Content-influential score manually assigned for content-based influential Twitter users of news topic *c*

Score	Description
2	The user often posts tweets for *c* while most of them often get retweeted by many users.
1	The user posts many tweets for c while only part of them get retweeted. The user's tweets for *c* get retweeted while his tweets for other topics get retweeted much more times.
0	The user posts tweets unrelated to *c*. The user's tweets for *c* do not interest others.

Table 19. Authority-influential score manually assigned for authority-based influential Twitter users of news topic *c*

Score	Description
2	The user's tweets are highly trustable for *c*. The user is highly relevant to *c* in the real world
1	The user is supported by some users about *c*. The user has high authority for other related topics while he also post tweets for *c*.
0	The user posts tweets unrelated to *c*. The user's tweets are ignored by most of users.

especially Republican supporters. @Norsu2 posted many tweets about the news topic while only some of them get retweeted by a few users. @AlJazeera_Live not only posted tweets about O_1, but also for many other news topics. These two users are assigned content-influential score of 1. @redostoneage posted a huge amount of tweets about O_1 while few of them get retweeted by others. He is more likely to be a robot rather than an ordinary user. @SoccerGrlProbs is a user often posts tweets about women's soccer, which is unrelated to O_1. These two users get the content-influential score of zero.

For authority-based influential Twitter users of O_1, @MittRomney and @PaulRyanVP are both assigned the authority-influential score of 2, since they are accounts of presidential election nominees from Republican Party. @rotolo is a professor at Syracuse University whose major is Information Science. Although his major is different from the news topic, assessors still assign him an authority-influential score of 1 since he has a high social position and his tweets about O_1 are still reliable. Other users who posted unrelated tweets are assigned score of zero.

Evaluation results are shown in Figure 50(a) and Figure 50(b). As we can observe, DCG values of RR and MR for these news topics are larger than the rest in most cases, which means RR and MR outperform other related methods. TN performs the worst compared with other methods, which means the number of tweets posted by the user is not a good indicator for his influence because these tweets might be ignored by his followers. IND seems to be reasonable to measure the influence. However, notices that retweet and mention are often used for campaigns, e.g. marketing campaigns, in Twitter to gain reputation. These retweets/mentions are not suitable for measuring user's influence. Also, IND ignores link structure among users. The link structure of user's retweets/mentions is helpful to find influential users, which has been proved by better performances of PR, RR, and MR. Newly proposed RR and MR outperform PR because both of them consider user's retweet/mention preference for the topic and user's topic relevance. PR ignores these, causing negative affection to its results.

As we can also observe that RR and MR are not always better than the others in some cases. One explanation for this is that due to the rate limit of Twitter API, it is hard to collect all tweets related to news topics. For some news topics, vertices (users) in G_{RT} and G_{MN} are not well connected. For example, in retweet graph of G_1, the average in-degree of vertex is 0.744, which is the lowest in all retweet and mention graphs. This means that users are not well connected. RR does not perform better than IND in this retweet graph. However, RR and MR give better results in other graphs having higher average in-degree per vertex.

Table 20. Top-15 content-based influential Twitter users for O_1

TN	IND	PR	RR
screen_name	**screen_name**	**screen_name**	**screen_name**
AlJazeera_Live	PaulRyanVP	PatDollard	PatDollard
GomerWHoward	PatDollard	LeftsideAnnie	LeftsideAnnie
KD0NHM	redostoneage	PaulRyanVP	PaulRyanVP
preciousliberty	Heritage	redostoneage	NETRetired
Progress2day	preciousliberty	NathanHale1775	maxnrgmike
preciseBlogs	SoccerGrlProbs	DarrellIssa	BlueDuPage
CAFalk	TheDailyEdge	ConNewsNow	NathanHale1775
michaelemlong	JeffersonObama	JeffersonObama	redostoneage
QuisMrEastCoast	edshow	BlueDuPage	Norsu2
GOPPrimary	LeftsideAnnie	edshow	CoffeeBean26
teeocee	PPact	NETRetired	ConNewsNow
Common_Sense4U	TheNewDeal	TheNewDeal	chasepolitics
MrHappy4870	WikiLeaks_GCC	RasmussenPoll	retfado
incognito912	OutFrontCNN	dgjackson	Conservativeind
scarletmonahan	NathanHale1775	preciousliberty	DarrellIssa

Table 21. Top-15 authority-based influential Twitter users for O_1

TN	IND	PR	MR
screen_name	**screen_name**	**screen_name**	**screen_name**
AlJazeera_Live	MittRomney	MittRomney	MittRomney
GomerWHoward	PaulRyanVP	PaulRyanVP	PaulRyanVP
KD0NHM	edshow	MarthaRaddatz	MarthaRaddatz
preciousliberty	140elect	140elect	140elect
Progress2day	MarthaRaddatz	AC360	AC360
CAFalk	mashable	edshow	edshow
michaelemlong	piersmorgan	rotolo	rotolo
QuisMrEastCoast	GOP	InesMergel	InesMergel
GOPPrimary	DarrellIssa	rickklein	andersoncooper
jillherring2	PatDollard	jonkarl	rickklein
teeocee	CentreC	andersoncooper	jonkarl
Common_Sense4U	teeocee	FlakeforSenate	FlakeforSenate
incognito912	AC360	cspan	GOP
1861_again	andersoncooper	FranTownsend	cspan
suspended	FlakeforSenate	GOP	DarrellIssa

Table 22. Top-15 content-based influential Twitter users for S_1

TN	IND	PR	RR
screen_name	**screen_name**	**screen_name**	**screen_name**
SyriaTweetEn	almayadeentv1	almayadeentv1	almayadeentv1
SyrianVideos	AlArabiya_Eng	KadirUstun	KadirUstun
tonierosa	Avaaz	SETADC	SETADC
syriatweet	MARYAMALKHAWAJA	syriancommando	syrianfalcon11
epaulnet	mog7546	WashingtonPoint	What_iiif
mog7546	WashingtonPoint	What_iiif	syriancommando
Visionaryck	rozalinachomsky	syrianfalcon11	WashingtonPoint
LarryMiller2012	Anon_Central	RT_com	RT_com
unaa2011	SyrianSmurf	EatingMyPeaz	CustosDivini
RobotNickk	RT_com	ZeinakhodrAljaz	rozalinachomsky
WashingtonPoint	bbclysedoucet	Mou2amara	SyrianSmurf
zimniya	ZeinakhodrAljaz	Iran	Mou2amara
TaziMorocco	kashafham	samersniper	ZeinakhodrAljaz
GamerOps	AJELive	3arabiSouri	EatingMyPeaz
eman_cipation_	KenRoth	SyrianSmurf	Iran

Table 23. Top-15 authority-based influential Twitter users for S_1

TN	IND	PR	MR
screen_name	**screen_name**	**screen_name**	**screen_name**
SyriaTweetEn	syriatweet	LerisZ	LerisZ
SyrianVideos	ErinBurnett	RT_com	RT_com
syriatweet	RT_com	garretpustay	garretpustay
WorldNews409	Avaaz	ASLANmedia	ASLANmedia
ActivismRocks	AlArabiya_Eng	Ted1733)	Ted1733)
epaulnet	UNICEF	AlArabiya_Eng	AlArabiya_Eng
mog7546	AJStream	HDNER	SyrianSmurf
Visionaryck	ahramonline	AFP	HDNER
unaa2011	AFP	SyrianSmurf	AFP
MENewsflash	AnonOpsSweden	BegForMariee_	KadirUstun
RobotNickk	sebaboerner	jirou_ARX7	ThoroughlyUS
WashingtonPoint	rozalinachomsky	PassySolomon	MokhtarGhazzawi
zimniya	SyrianSmurf	Sneakerpedia	LelikahHoee
TaziMorocco	jordantimes	vihargg	BegForMariee_
eman_cipation_	nytjim	MokhtarGhazzawi	vihargg

Table 24. Top-15 content-based influential Twitter users for G_1

TN	IND	PR	RR
screen_name	**screen_name**	**screen_name**	**screen_name**
HockeyNews247	JalenRose	JalenRose	JalenRose
OlympiaEthiopia	stl_baseball	HunterEstes1	stl_baseball
Miss_Placed_	Nationals	mandie11184	mandie11184
drgridlock	abc7newsBayArea	BigCountry125	BigCountry125
RealCashew	MLBNetwork	papi_chulo_96	HunterEstes1
_iTyra	ksdknews	stl_baseball	papi_chulo_96
M4llyM0u53	OccupyOakland	MLBNetwork	MLBNetwork
WashTimesSports	nbcwashington	Nationals	Nationals
SkyKerstein	dgoold	SkyKerstein	SkyKerstein
HarryElephante	LukeRussert	TFlow	MLBONFOX
acomak	MLBONFOX	federalbaseball	TFlow
MontcoCourtNews	PostSports	MLBONFOX	federalbaseball
CKBLUEPRINT	drgridlock	Toribelle4	PLAYLOUNGENYC
JMNats	DDOTDC	OccupyOakland	Toribelle4
masnKolko	ZuckermanCSN	keewee447	abc7newsBayArea

Table 25. Top-15 authority-based influential Twitter users for G_1

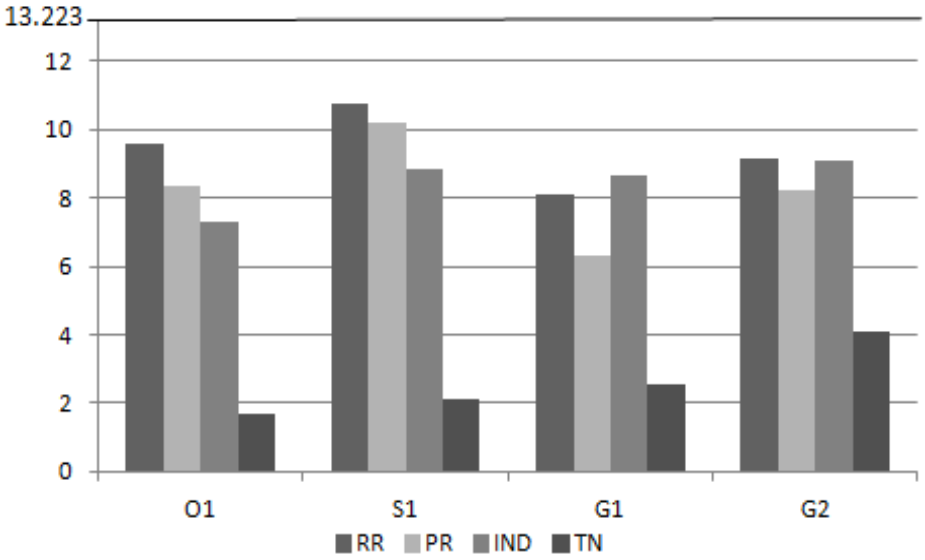
TN	IND	PR	MR
screen_name	**screen_name**	**screen_name**	**screen_name**
BridgettColling	Nationals	Nationals	Nationals
Miss_Placed_	educationweek	TheNatsBlog	stl_baseball
kingkaps7	JalenRose	stl_baseball	Toribelle4
drgridlock	stl_baseball	BarryEnright45	valliant306
bfentress	TheNatsBlog	Toribelle4	BarryEnright45
baseballchickie	MLBNetwork	EBJunkies	TheNatsBlog
tracytran	pandora_radio	JalenRose	cannonjw
SportsRadioApps	abc7newsBayArea	bluto79	danvaline
SkyKerstein	SkyKerstein	valliant306	JalenRose
MotownFans	PostSports	danvaline	EBJunkies
ArminRosen	wmata	baseballchickie	Windog0101
Moii_Rivass	drgridlock	gogobot!	bluto79
Ballsdc	nbcwashington	Digi_Sports	npatenaude2000
HarryElephante	slackadjuster	sports20	DaveBedford81
RRodriguez661	tracytran	1DarrahNicole	melknepp

Table 26. Top-15 content-based influential Twitter users for G_2

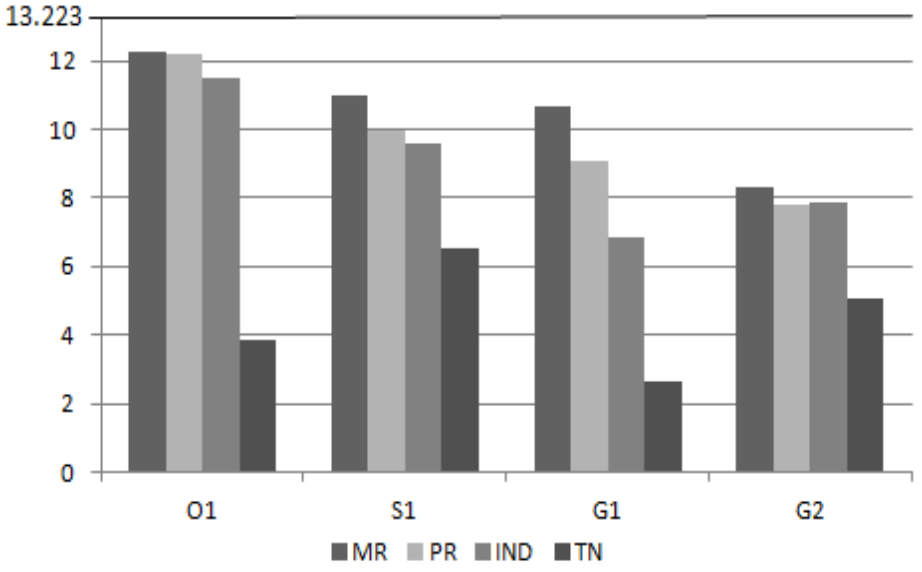
TN	IND	PR	RR
screen_name	**screen_name**	**screen_name**	**screen_name**
hunter_troll	VauxhallEngland	VauxhallEngland	VauxhallEngland
redfootball08	10thMar1905	LFC	LFC
SuryanovSky	LiverpooIFCNews	dlovett32	dlovett32
TheRedmenTV	dlovett32	redfootball08	redfootball08
redwesade	LFC	MarkBAllen	MarkBAllen
beatberlusconi	guardian_sport	LiverpooIFCNews	LiverpooIFCNews
danholling	talkSPORT	TheRedmenTV	TheRedmenTV
keswickbro	Karlton81	10thMar1905	10thMar1905
porkpackerpete	Liddellpool	Liddellpool	Liddellpool
SaghirM	Alreet_John	guardian_sport	guardian_sport
LFC_Since_1892	bitter_sweet140	Karlton81	Karlton81
talkSPORT	TheRedmenTV	jjlsmith7	ablondebabe
VauxhallEngland	LFC_Since_1892	Alreet_John	talkSPORT
LiverpoolWays	LiverpoolWays	himml3r	sandlotsportsco
antniwhite83	jjlsmith7	LFC_Since_1892	Alreet_John

Table 27. Top-15 authority-based influential Twitter users for G_2

TN	ING	PR	MR
screen_name	**screen_name**	**screen_name**	**screen_name**
Being_Brendan	LFC	ChelseaChadder	ChelseaChadder
hunter_troll	GotSaga	SitiNurmalaa.	SitiNurmalaa.
redfootball08	guardian_sport	LFC	LFC
knoller2	follow_b_army	guardian_sport	voicesfrmthesky
ChelseaStats	ChelseaStats	amidgley1982	KubraFatima
Alviandhika_	LiverpooIFCNews	Froners	SibsMacd
TheRedmenTV	dlovett32	voicesfrmthesky	MsThatoM
5BigEars	Karlton81	SibsMacd	guardian_sport
KOPWATCH	talkSPORT	MattHoltGBBox	lee_drysdale07
kenpootCFC	Now__Football	KubraFatima	Decwhite
CFC_FORLIFE	VauxhallEngland	ckeogh1971	amidgley1982
nickzytotre	Nicoleseaton_x	MsThatoM	Froners
Korean_Kop	KOPWATCH	lee_drysdale07	MattHoltGBBox
follow_b_army	ChelseaChadder	DJSpoony	ckeogh1971
Scotty_N_92	SkywalkerHD	ebbynf	DJSpoony



(a). DCG_{15} for content-based influential Twitter users



(b). DCG_{15} for authority-based influential Twitter users

Figure 50. DCG for top-15 influential Twitter users about four news topics

7

Conclusion

In this thesis, we presented our methods for finding two types of influential Twitter users about news topics in which users are interested and searched by the target word based on a new approach to detect news-topic-related hashtags. As basic components, Probabilistic Inside-Outside Log method which is used to detect characteristic co-occurrence words with the target word from news topics and Probabilistic Inside-Outside Log method for Hashtag method which is used to detect characteristic co-occurrence words with the hashtag from tweets have been proposed.

For characteristic co-occurrence word detection from news articles, we proposed PIOLog method which can detect characteristic co-occurrence words for each news topic related to the target word. News articles are clustered into topics in advance. Words which often co-occur with the target word in news articles of the news topic and are less likely to appear in other news articles are detected by PIOLog method. Experimental results showed that our PIOLog method is more likely to detect characteristic co-occurrence words with the target word for news topics since our method is asymmetric and topic-dependent.

For hashtag recommendation about news topics related to the target word, we proposed a new approach to recommend hashtags for news topics in which users are interested and searched by the target word. We applied our PIOLog method to news topics to create news topic vectors. We also extended the PIOLog method and proposed PIOLogH method to detect/weight characteristic co-occurrence words with the hashtag from tweets. Hashtag vectors are created based on PIOLogH method. Cosine similarity between news topic vectors and hashtag vectors is calculated to measure the relevance between news topics and hashtags. Hashtags with high similarity scores get recommended. Experimental results showed the effectiveness of our approach. PIOLog and PIOLogH methods outperform other methods like TFIDF.

For finding content-based and authority-based influential Twitter users, we proposed RetweetRank and MentionRank to find these two types of influential Twitter users from hashtag communities which are relevant to news topics searched by the target word. News-topic-related hashtags with their defined hashtag communities are detected based on PIOLog and PIOLogH methods. For users in those news-topic-related hashtag communities, RetweetRank and

MentionRank are applied to find content-based and authority-based influential Twitter users for the news topic. Experimental results showed that RetweetRank and MentionRank could find these two types of influential Twitter users from those news-topic-related hashtag communities and outperform other related methods.

In the future, we are planning to improve our data collection method. More tweets related to news topics should be collected within the rate limit of Twitter API. Also, not only news topics, but also other topics discussed by Twitter users should be considered. After we manually checked contents of some tweets posted by influential Twitter users, we found that tweets posted by some users, especially content-based influential Twitter users, are highly opinionated and strongly supported/opposed by other Twitter users. Mining opinions from tweets posted by influential Twitter users could help us understand why some opinions are popular and widely accepted, which is another research direction we are considering.

Acknowledgement

I would like to express my sincere gratitude to my supervisor Professor Takehiro Tokuda for his continuous guidance, support and help for my Ph. D study and research. His guidance helped me a lot in all my time of research and writing this thesis. I would be benefited from things which I have leant from my professor in my whole life. My sincere gratitude also goes to Assistant Professor Tomoya Noro for his great help, advices and comments for my research.

I would also like to thank Mr. Hao Han, Mr. Bin Liu, Mrs. Junxia Guo and other members of Tokuda Laboratory for their warmest regards and constant support for my research and life in Japan.

My sincerely thanks also goes to my wife Mrs. Chunji Li and my families. Help and support from all of you encourage me and support me spiritually throughout all my life.

Finally, I thank Japanese Monbukagakusho:MEXT for the scholarship to support my research and life in Japan.

Bibliography

[1] Twitter. https://twitter.com/

[2] Facebook. https://www.facebook.com/

[3] Digg. http://digg.com/

[4] Microblogging. http://en.wikipedia.org/wiki/Microblogging

[5] Celebrating #Twitter7. http://blog.twitter.com/2013/03/celebrating-twitter7.html

[6] Kwak, H., Lee, C., Park, H. and Moon, S., What is Twitter, a Social Network or a News Media? In *Proceedings of the 19th International World Wide Web Conference*, pp. 591-600, Raleigh, North Carolina, USA. ACM, 2010.

[7] Huang, J., Thornton, K. M., and Efthimiadis, E. N., Conversational Tagging in Twitter. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, pp. 173-178, Toronto, Ontario, Canada. ACM, 2010.

[8] Yang, L., Sun, T., Zhang, M., and Mei, Q., We Know What @You #Tag: Does the Dual Role Affect Hashtag Adoption? In *Proceedings of the 21st International Conference on World Wide Web*, pp. 261-270, Lyon, France. ACM, 2012.

[9] What Facebook and Twitter Mean for News. http://stateofthemedia.org/files/2012/03/Facebook-and-Twitter-Topline.pdf

[10] Christopher D. Manning and Hinrigh Schutze. Foundations of Statistical Natural Language Processing. MIT Press, 1999.

[11] Brooks, C.H., Montanez, N.: Improved Annotation of the Blogosphere via Autotagging and Hierarchical Clustering. In *Proceedings of the 15th International Conference on World Wide Web*, Edinburgh, UK. 2006.

[12] Mishne, G.: AutoTag: A Collaborative Approach to Automated Tag Assignment for Web- log Posts. In *Proceedings of the 15th International Conference on World Wide Web*, Edinburgh, UK. 2006.

[13] Sood, S.C., Owsley, S.H., Hammond, K.J., Birnbaum, L.: TagAssist: Automatic Tag Suggestion for Blog Posts. In *International Conference on Weblogs and Social Media*. 2007.

[14] Singhal, A., Buckley, C., Mitra, M.: Pivoted Document Length Normalization. In *19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 21–29. ACM, 1996.

[15] Flickr. http://www.flickr.com/.

[16] Sigurbjornsson, B., van Zwol, R.: Flickr Tag Recommendation based on Collective Know- ledge. In *Proceedings of the 17th International Conference on World Wide Web*, Beijing, China. 2008.

[17] Wartena, C., Brussee, R., Wibbels, M.: Using Tag Co-Occurrence for Recommendation. In *Proceedings of International Conference on Intelligent System Design and Application (ISDA 2009)*, Pisa, Italy. November 2009.

[18] Belem, F., Martins, E., Pontes, T., Almeida, J., Goncalces, M.: Associative Tag Recommendation Exploiting Multiple Textual Features. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Beijing, China. July 2011.

[19] Lehmann, J., Gonçalves, B., Ramasco, J. J. and Cattuto, C., Dynamical Classes of Collective Attention in Twitter. In *Proceedings of the 21st International Conference on World Wide Web*, pp. 251-260, Lyon, France. ACM, 2012.

[20] Weng, J., Lim, E.-P., He, Q., Leung, C.W.-K.: What Do People Want in Microblogs? Measuring Interestingness of Hashtags in Twitter. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, ICDM 2010, pp. 1121–1126. 2010.

[21] Correa, D., Sureka, A.: Mining Tweets for Tag Recommendation on Social Media. In *Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents*, SMUC 2011, Glasgow, Scotland, UK. 2011.

[22] Efron, M.: Hashtag Retrieval in a Microblogging Environment. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM SIGIR 2010, Geneva, Switzerland. 2010.

[23] Wagner, C., Strohmaier, M.: The Wisdom in Tweetonomies: Acquiring Latent Conceptual Structures from Social Awareness Streams. In *Proceedings of the 3rd International Semantic Search Workshop*, pp. 6. ACM, 2010.

[24] Zangerle, E., Gassler, W., Specht, G.: Recommending #-Tags in Twitter. In *Proceedings of the Workshop on Semantic Adaptive Social Web*, UMAP 2011, Gerona, Spain. 2011.

[25] Mazzia, A. and Juett, J., Suggesting Hashtags on Twitter. EECS 545 (Machine Learning) Course Project Report: http://www-personal.umich.edu/~amazzia/pubs/545-final.pdf

[26] Kywe, S.M., Hoang, T-A., Lim, E-P. and Zhu, F., On Recommending Hashtags in Twitter Networks. In *Proceedings of the 4th International Conference on Social Information*, page 337-350, 2012.

[27] Phuvipadawat, S., Murata, T.: Detecting a Multi-Level Content Similarity from Microblogs Based on Community Structures and Named Entities. *Journal of Emerging Technologies in Web Intelligence* 3(1). February 2011.

[28] Sankaranarayanan, J., Samet, H., Heitler, B.E., Lieberman, M.D., Sperling, J.: TwitterStand: News in Tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information System*, ACM GIS, Seattle, WA, USA. November 2009.

[29] Phelan, O., McCarthy, K., Smyth, B.: Using Twitter to Recommend Real-Time Topical News. In *Proceedings of the 3rd ACM Conference on Recommender Systems*, ACM RecSys, New York, NY, USA. October 2009.

[30] Abel, F., Gao, Q., Houben, G.-J., Tao, K.: U-Sem: Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web. In *Proceedings of International Workshop on Usage Analysis and the Web of Data*, USEWOD 2011, Hyderabad, India. 2011.

[31] Cha, M., Haddadi, H., Benevenuto, F. and Gummadi, K. P., Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proceedings of International AAAI Conference on Weblogs and Social Media*, Washington DC, USA. AAAI Press, 2010.

[32] Leavitt, A., Burchard, E., Fisher, D. and Gilbert, S., The Influentials: New Approaches for Analyzing Influence on Twitter. Web Ecology Project. http://www.webecologyproject.org/wp-content/uploads/2009/09/influence-report-final.pdf. 2009.

[33] Anger, I. and Kittl, C., Measuring Influence on Twitter. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, Graz, Austria. ACM, 2011.

[34] Hajian, B. and White, T., Modelling Influence in a Social Network: Metrics and Evaluation. In *IEEE Third International Conference on Social Computing*, pp. 497-500, Boston, MA, USA. IEEE, 2011.

[35] Romero, M. D., Galuba, W., Asur, S. and Huberman, A. B., Influence and Passivity in Social Media. In *ECML/PKDD*, pp. 18-33, volume 6913 of Lecture Notes in Computer Science, Springer, 2011.

[36] Kleinberg, J. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 1999, 46(5): p. 604-632.

[37] Ye, S. and Wu, F., Measuring Message Propagation and Social Influence on Twitter.com. In *Proceedings of the 2nd International Conference on Social Informatics*, pp. 216-231, Laxenburg, Austria. 2010.

[38] Bigonha, C., Cardoso, T. N. C., Moro, M. M., Almeida, V. A. F. and Goncalves, M. A., Detecting Evangelists and Detractors on Twitter. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, Belo Horizonte, Brazil. 2010.

[39] Noro, T., Ru, F., Xiao, F. and Tokuda, T., Twitter User Rank Using Keyword Search. In *The 22nd European-Japanese Conference on Information Modelling and Knowledge Bases*, pp. 31-48, Prague, Czech Republic. 2012.

[40] Weng, J., Lim, E. P., Jiang, J. and He, Q., TwitterRank: Finding Topic-sensitive Influential Twitterers. In *Proceedings of the Third International Conference on Web Search and Data Mining*, pp. 261-270, New York, New York, USA. 2010.

[41] Blei, D. M., Ng, A. Y. and Jordan, M. I., Latent Dirichlet Allocation. In *Journal of Machine Learning Research*, Vol. 3, pp. 993-1022,2003.

[42] Cano, A. E., Mazumdar, D., and Ciravegna, F., Social Influence Analysis in Microblogging Platforms - A topic-Sensitive based Approach. *Special Issue on the Semantics of Microposts. Semantic Web Journal*. 2013.

[43] OpenCalais. http://www.opencalais.com/

[44] Google News. https://news.google.com/

[45] New York Times Topics. http://topics.nytimes.com/

[46] Yahoo! News Topics. http://news.yahoo.com/topics/

[47] Noro, T., Liu, B., Nakagawa, Y., Han, H., and Tokuda, T., A news index system for global comparisons of many major topics on the earth. In *Information Modelling and Knowledge Bases XX*, page 194-211, 2009.

[48] Jones, K. S., A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, Vol. 28, pp. 11-21. 1972.

[49] Tweet Button. https://dev.twitter.com/docs/tweet-button

[50] Salton, G., Wong, A. and Yang, C.S., A Vector Space Model for Automatic Indexing. *Communications of the ACM*, Vol. 18, Issue 11, pp. 613-620. 1975.

[51] Brin, S. and Page, L., The Anatomy of A Large-scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems* 30(1-7): 107–117. 1998.

[52] Han, H., Noro, T. and Tokuda, Takehiro., An Automatic Web News Article Contents Extraction System Based on RSS Feeds, *Journal of Web Engineering*, Vol.8, No.3, pp.268-284. Sep. 2009.

[53] Kohlschutter, C., Fankhauser, P. and Nejdl, W., Boilerplate Detection Using Shallow Text Features. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining* (WSDM'10), page 441-450, New York, New York, USA. 2010.

[54] Twitter POST statuses/filter.
https://dev.twitter.com/docs/api/1.1/post/statuses/filter

[55] Using the Twitter Search API. https://dev.twitter.com/docs/using-search

[56] #TagDef. http://tagdef.com/

[57] Google. https://www.google.com/

[58] Schmid, H., Probabilistic part-of-speech tagging using decision trees. In *First International Conference on New Methods in Natural Language Processing*, page 44-49, 1994.

[59] Finkel, J.R., Grenager, T. and Manning, C., Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, page 363-370, 2005.

[60] Jarvelin, K. and Kekalainen, J., Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4): 422-446, 2002.

[61] Park A. and Paroubek P., Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of the International Conference on Language Resources and Evaluation* (LREC'10). Valletta, Malta. 2010.