Wikipedia-based Topic Clustering for Microblogs

Tan Xu
College of Information Studies
University of Maryland, College Park, MD 20742
tanx@umd.edu

Douglas W. Oard
College of Information Studies and UMIACS
University of Maryland, College Park, MD 20742
oard@umd.edu

ABSTRACT
Microblogging has become a primary channel by which people not only share information, but also search for information. However, microblog search results are most often displayed by simple criteria such as creation time or author. A review of the literature suggests that clustering by topic may be useful, but short posts offer limited scope for clustering using lexical evidence alone. This paper therefore presents an approach to topical clustering based on augmenting lexical evidence with the use of Wikipedia as an external source of evidence for topical similarity. The main idea is to link terms in microblog posts to Wikipedia pages and then to leverage Wikipedia's link structure to estimate semantic similarity, Results show statistically significant relative improvements of about 3% in cluster purity using a relatively small (7500-post, 5-topic) Twitter test collection. Linking terms in microblog posts to Wikipedia pages is also shown to offer a useful basis for cluster labeling.

Keywords
Microblog search, topic detection, topic clustering, cluster labeling, Wikipedia.

INTRODUCTION
Although a relatively new phenomenon, microblogging has grown explosively in recent years. According to Twitter, by September 14, 2010 (4 years after it was first launched), there were 175 million registered users sending 95 million tweets per day (tweets are short snippets of text written by twitter users).[1] This rapid adoption has generated interest in gathering information from microblogging about real-time news and about opinions on specific topics. That interest, in turn, has led to a proliferation of microblog search services from both microblogging service providers such as Twitter and general-purpose search engines such as Bing and Google.

However, compared with traditional document retrieval and Web search, microblog search is still in its infancy. In a typical microblog search scenario using Twitter, around 1500 tweets that contain the query terms will be returned, ranked by their creation time. Although other presentation formats are also available (e.g., ordering results by author popularity, or by hyperlinks referenced), presentation formats optimized for topic monitoring are not yet widely available. The goal of this paper is to explore the potential for topical organization of microblog search results.

This is a challenging problem because microblog posts are short, so traditional topical clustering techniques based on lexical overlap (use of the same words) are necessarily weak. The usual approach is to draw on some external source to enrich the available evidence for topical similarity; in this paper we look to Wikipedia for that external evidence. This, in turn, raises the challenge of how to minimize the adverse effects of the ambiguity that is naturally present in Wikipedia; we leverage ideas first introduced in the INEX link-the-wiki track in which the most consistent possible set of links from each post are sought. These links also turn out to be useful as a basis for cluster labeling. At this stage our work is focused exclusively on automatically creating and labeling topical clusters, and our evaluations are constructed using either a somewhat artificial 5-topic test collection that facilitates automatic scoring at moderate scale or on a hand-coded subset of that collection that supports evaluation with a natural range of subtopic variation. Usability evaluation with actual users is left for future work.

The remainder of this paper is organized as follows. The Microblog Search section provides a brief review of the aspects of information seeking behavior, information retrieval, and social computing that help to characterize microblog search and to motivate the potential value of topically-organized result presentation. The Related Work section then reviews recent related work on discerning topics from microblogs. The Method Design describes our Wikipedia-based microblog topical clustering method, and the Experiment section presents the design of our experiments and our results. The Conclusions and Future Work section summarizes the results, identifies key limitations, and describes some next steps for this work.

MICROBLOG SEARCH
In this section, through an analysis of literature, microblog search behavior is described. By focusing on the question of when to stop a search, we are able to illuminate one limitation of common microblog search result presentations,

ASIST 2011, October 9–13, 2011, New Orleans, LA, USA.

[1] http://twitter.com/about

thus motivating our interest in topical clustering of microblog posts.

Microblog Search Behavior

The information in microblogs grows fast, updates frequently, and covers a wide range of topics. This makes microblogging an important resource from which people could extract knowledge that they need or want. By conducting user studies and analyzing user queries issued to Twitter, Teevan, Ramage and Morris (2011) summarized the types of information that people are mostly interested in knowing from microblogs, as shown in Table 1.

Types of Information	Explanations
Timely information	Breaking news, current events, real-time reporting, friends' daily activities
Social information	People with specific interests, information or microblogs of a specific user or group, and people's overall opinions on a particular topic
Topical information	Similar to traditional Web search, people also search for information of specific interest on Twitter

Table 1. Information People Search Microblogs for

In an effort to satisfy the above information needs, several types of information seeking behaviors have been observed. For instance, a user may "follow" other people who have specific interests, or they may ask questions in the form of their own microblog posts. By analyzing manually coded tweets, Naaman et al. (2010) found that 5% of all tweets are posted for the purpose of asking questions of the poster's followers.

However, as with the World Wide Web, the volume of microblog posting generates a need to be able to find interesting information efficiently. Therefore, searching is an important information seeking strategy in the microblogosphere. Applying Wilson's nested model of information seeking (Wilson, 1999), a user's microblogging information seeking behavior is a subset of their microblogging information behavior, particularly concerned with the variety of methods they employ to discover, and gain access to microblogs; and a user's microbloging search behavior is a subset of their microblogging information seeking behavior, particularly concerned with the interactions between users and computer-based search systems for microblog posts.

In traditional Web search, according to Broder (2002), people's search intentions can be grouped into three classes: navigational (i.e., find a specific Web site), informational (i.e., find information of a specific topic), and transactional (i.e., perform Web-mediated activities). In traditional blog search, Mishne and de Rijke (2006) found that searchers exhibit a somewhat different range of intents. First of all, the majority of blog queries are informational; secondly, the informational intent can be further divided into two classes: tracking references to named entities, and identifying blogs or posts that focus on a specific topic. In microblog search, searchers seem to have similar intents as in traditional blog search. In the Teevan, Ramage and Morris (2011) study, query logs were used to observe cross-corpus searching between Twitter and the Web, with users trying to find specific microblogs and Web pages on a particular topic. They also observed extensive reuse of the same queries—56% of Twitter queries are issued more than once by the same user, which, according to a qualitative analysis, is for the purpose of monitoring topics over time. An additional interest of microblog searchers, which is not observed in traditional blog search, is the intent to get an overall view of what other people are saying about some specific topics, which can only be achieved by somehow summarizing multiple microblog posts.

When to Stop a Search

As pointed by O'Day and Jeffries (1993), one of the fundamental issues faced by a searcher is to determine when to stop a search (either by terminating completely or by starting some new search). This is, of course, also an important question facing microblog searchers. In O'Day and Jeffries's work (1993), they define four triggers that can lead a new search, and 3 stopping conditions. Bates (1979) takes a different approach, using a cost-benefit analysis to characterize the decision about whether to stop a search. The underlying assumption is that searchers will make a decision that maximizes expected utility: if stopping yields higher expected utility than continuing, the searcher will stop.

When considering the informational intent of topic monitoring in microblogs, using search result presentations that ignore topical relationships (e.g., ordering results by author or creation time) results in at best limited support for the decision about when to stop a search. Because to satisfy this intention, (1) searchers need to read as many microblog posts as they can; and (2) the topics varied within posts, which is difficult for searchers to identify these topics and filter microblogs on a specific one. Thus, in this situation, applying O'Day and Jeffries's theory, unless a searcher changes his/her original interest or encounters inhibiting factors, the search task cannot be completed or satisfied by the result representations. By using Bates' theory, a searcher will stop a search task only because the cost of continuing searching is anticipated to be too high, but not because they are benefitted enough from search results.

For this purpose, it would be valuable to provide tools to allow searchers to automatically thread together topically related microblog posts. Specifically, the goal can be described as: given a collection of microblog posts $D = \{d_1, \ldots, d_n\}$, the system can obtain topic groups $T = \{(\delta_1, T_1), \ldots, (\delta_m, T_m)\}$, where each group T_i represents a collection of microblog posts concerning on similar subject of themes ($T_i \subseteq D$) and δ_i is the description of group T_i, which consists of a set of terms as labels.

RELATED WORK

Discerning topics from microblogs has begun to receive attention recently. Chen, et al. (2010) built tweet recommendation systems, exploring several approaches, one of which was topic based. They decide whether an incoming tweet will interest a user depending on whether the topic of this tweet is relevant to the topic model established for the user according to his/her posting history. The comparison method employed by this study was based on the Vector Space Model (VSM), and terms were weighted using TF-IDF. TwitterRank was another system that relied on the topics of tweets (Weng, et al., 2010). The goal of TwitterRank was to identify influential microbloggers. They first used Latent Dirichlet Allocation (LDA) to build "topic models" for each author based on all tweets posted by that author. Then, they compared queries with each author's topic model to find the most "relevant" author. Similar approaches to identify topics of microblogs were also adopted by Pal and Counts (2011) and by Ramage, Dumais, and Liebling (2010).

Those approaches all focused on how best to represent content using only evidence that can be found directly in that content or generalizes automatically from that content ("topic" is an overloaded term – in "topic modeling" what is actually modeled is a set of more abstract conceptual representations that are used compositionally to represent a document). Microblog posts pose challenges for these kinds of techniques, however, because (1) shorter content reduces the scope for textual analysis because of fewer contextual clues (Phan et al., 2008), and (2) the common use of informal language tends to increase data sparsity.

In an effort to address these limitations, another trend has been to complement this sort of "collection-internal" modeling by also leveraging some external source(s) of information. Twopics was a system that used relationships between concepts defined by using disambiguation to link entity references found in tweets to Wikipedia (Michelson & Macskassy, 2010). The goal of Twopics was to develop a "topic profile" for a particular Twitter user. Another example of using external information was the work of Cataldi, Caro, and Schifanella (2010) in whch they first calculated author authority by using PageRank on social relationships, and then used that author authority to assign each term an importance weight. The goal of their work was to find the most important topics that were discussed in tweets during a specified time interval.

In this paper, a Wikipedia is used as the external resource, but in a manner different from Michelson and Macskassy (2010). Wikipedia is an attractive choice because it contains over 3.7 million English articles[2] that are densely structured through inter-wiki links that can be used to characterize topical relationships in a manner that complements lexical evidence.

METHOD DESIGN

Actually, topic detection has been an old problem in natural language processing and information retrieval (Allan, 2002). The task can be defined as grouping together stories that discuss the same event.[3] A common solution for the problem is to use machine learning approach. And according to the different ways of machine learning, detailed methods can be divided into three categories: supervised, unsupervised, and semi-supervised. In this study, since we assume that there is no prior knowledge regarding what features can be used and how important the features are, an unsupervised topic detection approach is adopted. The overall process can be described as following: first of all, useful features are extracted and how such features could imply topical relationship between microblog posts are defined; then, by combining these features, between-microblog distances are calculated; thirdly, by applying certain clustering method, microblogs can be clustered into topical groups; and lastly, for each topical cluster, selecting representative terms to label the topic.

Terminology

For convenience in describing the methods and algorithms in the following sections, some basic terms are defined in this section.

Token (w): a token w is defined as a basic unit of discrete data which is separated by space and punctuations, and contains only alphabetic letters. Functional characters used in microblogs are exceptions. For example, hashtags used in Twitter are also considered as tokens, but with the initial hash symbol being removed. Detailed preprocessing procedures for tokenization is described in the Evaluation Data section.

Feature Term (t): given a microblog post, a feature term refers to a single token (except stop-words) or a group of tokens, such as phrase, that are used as a single unit to designate a meaning. In this study, for feature term consists of multiple tokens, it can only be identified if it is used at least once in Wikipedia as an anchor text of a link.

Document ($d=\{t_1,\ldots,t_n\}$): a document is a single microblog post, which is comprised of a sequence of feature terms.

Topic ($T = \{\delta, \{d_i,\ldots,d_j\}\}$): a topic is a set of documents $\{d_i,\ldots,d_j\} \subseteq D$ concerning a particular subject or theme, and it is represented by a set of feature terms $\delta = \{t_m,\ldots,t_n\}$ that can representatively describe the concepts of a topic.

Feature Term Extraction

There are multiple ways to capture feature terms from a document. In this study, feature terms are identified by using Wikipedia. For terms composed of multiple tokens, only the ones that were used in Wikipedia at least once as an anchor text for a link are recognized as feature terms. There are other alternative possibilities of using Wikipedia to extract feature terms, such as using the ones used as

[2] http://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

[3] http://www.itl.nist.gov/iad/mig/tests/tdt/tasks/detect.html

Wikipedia article titles. The reason for this choice is because (1) using a term as an anchor text means that the term can represent an important concept so that Wikipedia editors selected it to provide additional explanation for readers; and (2) given a concept (represented by a Wikipedia article), there could be plenty various concept mentions (anchor texts) that link to it, which offers large amount of ground truth for mapping from a term in a microblog post to a Wikipedia article. For feature terms that contain only one token, regardless of whether they are used as anchor texts or not in Wikipedia, they are used as feature terms unless stop-words. If overlapping between feature terms happened, the longest one is identified.

Linking Terms to Wikipedia

After extracting feature terms, in this section, the method for identifying concepts from feature terms by using Wikipedia is introduced.

In Wikipedia, an article can be viewed as a description of a concept, which is expressed by the article's title. However, due to language ambiguity (i.e., polysemy, synonymy) and variations of language usage, not every term in microblog post can find its corresponding Wikipedia article by simply conducting word matching with article titles. Therefore, the essential issue is to choose an appropriate linking destination for a term. By its nature, this task can be viewed as a typical document retrieval problem (Huang et al., 2008). On the other hand, by using contextual features, this task can also be formulated as a classification problem (Mihalcea & Csomai, 2007; Milne & Witten, 2008). In this study, this second approach is adopted. Wikipedia's linking history and a term's textual context information are used to disambiguate the term meaning.

First of all, given a term, a list of candidate Wikipedia articles is collected based on the Wikipedia's linking history for the term used as an anchor text. For example, Table 2 lists all articles in Wikipedia that term "atomic" was linked to. The linking probability refers to the prior probability of an article used as a destination for a term in Wikipedia. Although for the purpose of disambiguating term meaning in microblogs, the ideal evidence should come from linking history from microblogs to Wikipedia, however, since the absence of such information, Wikipedia prior linking history is used as a substitute.

Anchor Term	Target Wikipedia Article Title	Linking Probability	Overlapping rate given an Example Local Context
atomic	Atom	0.239	0.167
atomic	Atomic (song)	0.198	0.083
atomic	Atomic Bomberman	0.106	0.083
atomic	…	…	…
atomic	Nuclear power	0.015	0.583
atomic	…	…	…

Table 2. Wikipedia Linking History for Term "Atomic"

Then, an overlap rate between a term's local textual context and the content of a candidate Wikipedia article is calculated to decide which article explains the meaning of a term. Here, because microblog posts are very short, local textual context is composed of all tokens within the same microblog post of the ambiguous term. Take the term "atomic" as an illustration, in a microblog post {China, France cautious on nuke energy: PARIS (AP)--Japan's nuclear crisis reverberated in atomic power-friendly}, its local context is {energy, nuke, Paris, nuclear, crisis, France, Japan, cautious, China, reverberated, power, friendly}. When calculating the overlap rate, all tokens are stemmed by using the Porter's stemming algorithm (Porter, 1980). And the overlap rate is normalized by the total number of tokens in local context, so that to get the value ranges from 0 to 1, which could be further used to compute a final disambiguation confidence score as defined in Equation 1. The basis for using context overlap rate is the lexical cohesion theory as described in (Halliday & Hanson, 1976), which suggests that textual context can help to interpret a term's meaning. Therefore, a higher overlap rate suggests more strongly that a Wikipedia article explains the concept of a term well. Take the above term "atomic" as an example, in the example microblog context, the Wikipedia article "Nuclear power" with the highest overlap rate is a best choice of concept description, despite that this article has a relatively low linking probability (0.015).

Because linking probability represents a general characteristic of a term within Wikipedia and context overlap rate characterizes the specialty of a term in a given microblog context, an equation is build to positively relate these two types of evidence to the selection of Wikipedia article. There are several functions satisfying this requirement, such as linear combination or log-linear combination of independent variables. In this study, the simple linear combination is applied to combine these two types of evidence to find the most appropriate linking destination, as defined in Equation 1:

$$Confidence(Wiki_i \mid t_j, d_k) =$$

$$\lambda * linkProb_{(Wiki_i, t_j)} + (1-\lambda) * OverlapRate_{(Wiki_i, d_k)} \quad (1)$$

where λ is arbitrarily chosen to 0.5 in this study. The Wikipedia article that maximizes this confidence score is chosen for a term t_j given document d_k as its context.

When disambiguating term meaning by using Wikipedia, because inappropriate article could also be identified as a concept of a term. Therefore, by learning from a small set composed of 103 terms, which are extracted from 9 tweets, a cutoff value for disambiguation confidence is set up. Only

when the disambiguation confidence value is equal or above 0.3, will the Wikipedia article be used as a destination for a term. This threshold is decided by the following process: (1) linking these 103 terms to Wikipedia given the 9 tweets context; (2) providing manual judgment for the term disambiguation results; (3) Calculating the overall disambiguation precision ($\frac{\text{Correct disambiguated terms}}{\text{All terms in the set}}$) and the recall rate ($\frac{\text{Terms that can be disambiguated}}{\text{All terms in the set}}$) given a disambiguation confidence threshold; and (4) Choosing the threshold that maximizes the F-measure of disambiguation precision and recall rate ($\frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$), which suggests an optimal disambiguation result given these two criteria.

Document Distance Measurement

As an important step for preparing microblog clustering, the method to measure distance between microblogs is described in this section. Because concepts in microblogs can be identified in Wikipedia, semantic relationship between concepts is used. Many similar efforts have been made to use external knowledge base to infer distance between documents (Phan et al., 2008; Michelson & Macskassy, 2010; Genc, Sakamoto & Nickerson, 2011).

Given two concepts, which are discriminated to two Wikipedia articles from two feature terms, their semantic distance can be calculated as described in Milne and Witten's work (2008). The assumption is that two articles will be related if they are linked by the same third article:

$$Distance(a,b) = \begin{cases} \frac{log(max(|A|,|B|)) - \log(|A \cap B|)}{log(|W|) - \log(\min(|A|,|B|))} & if\ |A \cap B| \geq 1 \\ 1 & otherwise \end{cases} \quad (2)$$

where a and b are the two articles (representing two concepts) of interest, A and B represent the sets of all Wikipedia articles that link to a and b respectively, and W is the total number of articles in Wikipedia, which is 5,230,947 in the version used by this paper (20101011 enwiki dump). The range of $Distance(a,b)$ is [0, 1], with 1 meaning no semantic relationship between the two concepts, and 0 meaning the two concepts are the same in meaning.

Because a threshold for term disambiguation confidence is used, many feature terms cannot be used to calculate semantic relationship. However, they are still important lexical evidence when measuring document distance. A cosine similarity function for two documents can be defined as Equation 3:

$$cos(d_i, d_j) = \frac{\sum_{t \in d_i \cap d_j} w(t,d_i) * w(t,d_j)}{\sqrt{\sum_{t \in d_i} w(t,d_i)^2} * \sqrt{\sum_{t \in d_j} w(t,d_j)^2}} \quad (3)$$

where, $w(t,d)$ is the weight of a term t in a document d, which is calculated by using TF-IDF weight.

Because semantic similarity between two terms represents their similarity in meaning, a term t_i can be viewed as a variation of another term t_j.Therefore, with a transformation effort, these two terms can be considered as an overlap between two documents, as described in Equation 4:

$$w(t_i, d_i) * w(t_j, d_j) * Sim(t_i, t_j) \quad (4)$$

where $Sim(t_i, t_j)$ =*1-Distance(t_i, t_j)*. Hence, a lexical overlap between two documents is a special case with $t_i = t_j$, which means *Distance*(t_i, t_j) = 0, and $Sim(t_i, t_j)$ =1.

Therefore, terms' semantic distance can be combined with the traditional cosine similarity into a final Semantic Cosine Similarity function for measuring document distance as Equation 5:

$$cos(d_i, d_j) = \frac{\sum_{t_i \in d_i, t_j \in d_j} \frac{Sim(t_i,t_j)}{\sum_{t_j \in d_j} Sim(t_i,t_j)} * w(t_i,d_i) * \frac{Sim(t_i,t_j)}{\sum_{t_i \in d_i} Sim(t_i,t_j)} * w(t_j,d_j) * Sim(t_i,t_j)}{\sqrt{\sum_{t \in d_i} w(t,d_i)^2} * \sqrt{\sum_{t \in d_j} w(t,d_j)^2}} \quad (5)$$

where component $\frac{Sim(t_i,t_j)}{\sum_{t_j \in d_j} Sim(t_i,t_j)}$ is used to adjust the $w(t_i, d_i)$ for t_j, because t_i can be transformed to multiple terms in d_j, and therefore needs to be normalized so that all the transformations are originated from one term; component $\frac{Sim(t_i,t_j)}{\sum_{t_i \in d_i} Sim(t_i,t_j)}$ is used for the same purpose. By normalizing these two term weights, the range of $Scos(d_i, d_j)$ is [0,1], therefore, this score is comparable with traditional $cos(d_i, d_j)$. If either t_i or t_j cannot be disambiguated to a Wikipedia page, then

$$Sim(t_i, t_j) = \begin{cases} 1 & if\ t_i = t_j \\ 0 & otherwise \end{cases} \quad (6)$$

Cluster Analysis

By applying the method proposed above, given a collection of documents, a semantic cosine similarity matrix between documents can be obtained. Each value of this matrix ranges from 0 to 1, with 0 meaning two documents have no similarity, and 1 meaning two documents are equal. For clustering purpose, a distance matrix is firstly created by using (*1-Semantic Cosine Similarity Score*) for each elements in the similarity matrix.

Agglomerative and divisive are two general types of clustering algorithms. Agglomerative algorithms cluster documents in a bottom-up manner, and divisive algorithms cluster document in a top-down manner. In this study, to reduce computational complexity, k-means clustering algorithm is selected, which belongs to divisive clustering. It begins by randomly generating a chosen number of 'k' clusters, and then by calculating the distance between each data point and clusters' center, reassigns each data point to a nearest cluster until the assignment of data points to clusters stop changing (MacQueen, 1967). The objective of k-means clustering is to produce clusters with a minimal sum of squares of Euclidean distance from documents to

their cluster centers, which can be calculated by the sum of squares within groups (WSS) as defined in Equation 7.

$$WSS = \sum_{j=1}^{g} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 \quad (7)$$

where, n is the number of elements in a group, g is the total number of groups, y_{ij} is the i^{th} element in group j, and $\bar{y}_j$ is the mean or centroid of documents in group j. A smaller value of WSS indicates tighter clusters.

To perform k-means clustering, the first important work is to find a best value of 'k'—optimal cluster cardinality. For this purpose, the "elbow method" is employed, which can be traced back to the work of Thorndike (1953).

Topic Labeling

Another important challenge for topic detection is to label the detected topic groups. Some earlier topic discovery systems such as Scatter/Gather (Cutting et al., 1992) and Suffix Tree Clustering (STC) (Zamir et al., 1997) use the most frequent words or word sequences to describe the detected topics in the corresponding clusters. Later work in hierarchical topical clustering attempts to describe topics by means of a cluster hierarchy. Recently, a concept of frequent term set (i.e., a set of terms grouped according to co-occurrence) is employed by several researchers to obtain both the documents on a topic and its description (Fung et al., 2003; Li et al., 2008). In this study, since terms in microblogs can be linked to Wikipedia articles, which disclose concepts of terms, selecting the most frequently used concepts to label a topic reveals another possible solution.

Given a topic group $T = \{d_i, \ldots, d_j\}$, a bag of feature terms can be obtained. This proposed approach first computes pairwise semantic distance of terms by using Equation 8, then clusters these terms into concept groups by using k-means clustering method.

$$Semantic\ Distance(t_i, t_j|T) = \lambda * Distance(t_i, t_j) + (1 - \lambda) * \left(1 - CooccurrenceRate(t_i, t_j|T)\right) \quad (8)$$

where λ is arbitrarily set to 0.5 in this study;

$$CooccurrenceRate(t_i, t_j|T) = \frac{Number\ of\ documents\ contain\ both\ t_i, t_j\ within\ T}{Total\ number\ of\ documents\ in\ T} \quad (9)$$

Here, two terms are assumed semantically related if they occur in the same microblog post. Due to the length constraint on microblogs, microbloggers often selectively choose the term in a single post, and the co-occurrence of terms by chance is rare and can be ignored in this study. $Distance(t_i, t_j)$ is computed by Equation 2. If either t_i *or* t_j has no link to Wikipedia, then

$$\text{Distance}(t_i, t_j) = \begin{cases} 1 & if\ t_i = t_j \\ 0 & otherwise \end{cases} \quad (10)$$

The assumption for Equation 8 is that $Distance(t_i, t_j)$ computed from Wikipedia indicates a general semantic distance between t_i *and* t_j and $(1 - CooccurrenceRate(t_i, t_j|T))$ indicates their semantic distance within a specific topic group. Therefore, a simple linear combination of these two types of distances is applied to infer its positive relationship with the final between-term distance, which is used for term clustering.

By applying Equation 11, terms with top scores are then selected as labels for the topic group. In Equation 11, $df(t_i|T)$ denotes document frequency of term t_i within T, and $CentroidDistance(t_i|T)$ means the Euclidean distance from term t_i to the centroid of the cluster T, which is defined as the average of all points in the cluster—that is the arithmetic mean over all the points in the cluster on each dimension. The shorter this distance is, the closer this term is to one of the major concepts of T, which means the better choice it is to represent the cluster.

$$LabelingScore(t_i|T) = df(t_i|T) * \left(1 - CentroidDistance(t_i|T)\right) \quad (11)$$

EXPERIMENT

In this section, a series of experiments are designed to evaluate the effectiveness of the methods designed in section 4. All the statistical significance tests in this section are based on Pearson's Chi-square test at a 95% confidence level (Pearson, 1900).

Evaluation Data

By selecting 5 queries from the trending words in Twitter, evaluation data composed of 7,500 tweets was collected, with each tweet categorized by the corresponding query term. For each query term, 1,500 tweets were collected, as shown in Table 3.

Query Term	Number of Tweets	ID
Japan Earthquake	1500	E
Japan Nuclear	1500	N
Libya	1500	L
Jennifer Lopez	1500	J
Boy Meets World	1500	B

Table 3. Test Data Set Description

Because original tweets are noisy and contain informal language usage, the following preprocessing steps are performed: (1) decode HTML entities into UTF-8, such as decode "<" to "<"; (2) convert all characters into UTF-8; (3) filter out any embedded URLs, which are characterized by space-isolated strings starting with "http://" or "https://"; (4) remove stop-words; (5) remove functional characters and tokens, such as Twitter account names, which are space-isolated strings that start with "@", Twitter slangs (e.g., "RT"), and initial mark "#" for trends; (7) remove headings, such as "NEWS"; (8) remove non-alphabetical tokens, such as digits; (9) remove punctuation; (10) stemming by using Porter's stemming algorithm (Porter, 1980).

Evaluation Measurement

Given the task of microblog topic detection, two major parts of the techniques are evaluated: microblog clustering and topic labeling

Evaluation of Microblog Clustering

To validate the quality of induced clusters, the following measures are adopted:

(1) An optimal cluster cardinality calculated as described in section of *cluster analysis*. Since a heuristic based approach is used to calculate the value of k for k-means clustering, it indicates an optimal WSS, which means tighter clusters are generated. However, since this criterion alone does not necessarily translate into a good clustering quality, the other evaluation measures are also used.

(2) Purity is a simple and transparent clustering evaluation measurement, which can be computed as the total number of correctly clustered documents divided by the size of the collection. It is used to compare different clustering algorithms when they induce equal numbers of clusters. When evaluating a specific cluster, its topic is determined by the class of the plurality of documents. Formally, purity is defined as Equation 12:

$$Purity(W, C) = \frac{1}{N}\sum_k max_j |\omega_k \cap c_j| \quad (12)$$

where, W is set of clusters generated by the clustering algorithm $\{\omega_1, \dots, \omega_k\}$, C is a set of manually created clusters $\{c_1, \dots, c_j\}$, and N is the total number of documents.

However, there are two drawbacks when using purity: (1) it cannot trade off the clustering quality against the number of clusters; and (2) it cannot penalize the error of assigning similar documents to different clusters well. Therefore micro- and macro-averaged F1 measures are also used in clustering evaluation.

(3) Micro- and Macro-averaged F1measure

Given a system induced cluster ω_k and a manually labeled cluster c_j:

$$Precision(\omega_k, c_j) = \frac{|\omega_k \cap c_j|}{|\omega_k|} \quad (13) \qquad Recall(\omega_k, c_j) = \frac{|\omega_k \cap c_j|}{|c_j|} \quad (14)$$

$$F1(\omega_k, c_j) = \frac{2 * Precision(\omega_k, c_j) * Recall(\omega_k, c_j)}{Precision(\omega_k, c_j) + Recall(\omega_k, c_j)} = \frac{2 * |\omega_k \cap c_j|}{|\omega_k| + |c_j|} \quad (15)$$

When computing the overall F1 for all the induced clusters, micro-averaged F1 gives equal weight to every document, whereas macro-averaged F1 gives equal weight to every topic, as calculated by the following formulas:

$$MacroAveraged\ F1 = \frac{1}{N_{topics}}\sum_{j=1}^{N_{topics}} arg\ max_k\ F1(\omega_k, c_j) \quad (16)$$

$$MicroAveraged\ F1 = \frac{2*\sum_{j=1}^{N_{topics}} arg\ max_k |\omega_k \cap c_j|}{\sum_{j=1}^{N_{topics}} |c_j| + \sum_{j=1}^{N_{topics}} arg\ max_k |\omega_k|} \quad (17)$$

where, N_{topics} is the total number of manually created topics.

Topic Labeling Evaluation

To evaluate the topic labeling methods, micro- and macro-averaged F1 are also used. In order to remove the error caused by clustering, labeling methods are assessed by using manually created clusters. Then, by using above equations, ω_k represents a system induced label set, and c_j represents a manually created label set. Because when manually create topic labels, coders do not know the boundaries of feature terms that are generated by the machine, to minimize the errors caused by this inconsistency of label selection, words in the labels are used for comparing. Therefore, all labeling terms are tokenized and stemmed before evaluation.

Result Analysis

Effect of Document Distance Measurement

The first experiment tests the effectiveness of the proposed semantic cosine document distance measurement (SCS). For comparison, a baseline of cosine document distance measurement (CS) is set up. To avoid the effect caused by the size of the evaluation data, five test sets with size of {1500, 3000, 4500, 6000, 7500} are created by randomly select tweets from the evaluation data. After preprocessing, the tweets left for each set are {1500, 2989, 4391, 5845, 7316}.

(1) Figure 1 shows optimal cluster cardinality on each test set by using the k-means clustering with semantic cosine document distance measurement. As shown in Figure 1, only when the test set size is 2989 and 7316, the optimal cluster cardinality equals to 5, which is known as the correct number of general topics in the data, the other results are still close to 5. The reason that more topics are detected is because when sampling documents from the evaluation data, enough documents for certain sub-topics are extracted. And the fact that no optimal cluster cardinality is under 5 can be viewed as an evidence of the effectiveness of the method. For the following evaluations, k is fixed to 5.

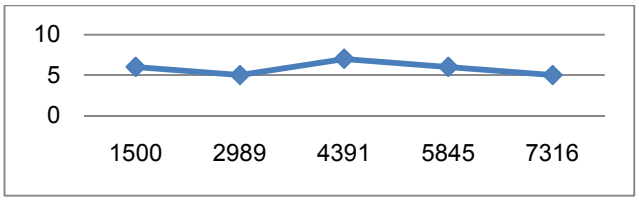

Figure 1. Optimal Cluster Cardinality for Test Sets

(2) Table 4 shows clustering purities for different document distance measurements on each test set by using the k-means clustering, where a "*" in the second column indicates a statistically significant difference from the first column (for the rest parts of result analysis, "*" is used for the same purpose). First, except the test set with size 2989, it can be appreciate that the proposed SCS statistically significantly outperforms CS. Second, except the test set with size 1500, the test set size has no statistically significant effect on the clustering quality.

Test Sets	CS	SCS
1500	0.327	0.368*
2989	0.284	0.301
4391	0.282	0.306*
5845	0.283	0.304*
7316	0.286	0.310*

Table 4. Clustering Purity of Different Document Distance Measurements on Test Sets

(3) Micro- and macro-averaged F1 scores for different document distance measurements are reported in Table 5. Only for the test sets with size of 1500 and 2989, SCS statistically significantly outperforms CS. One additional finding is that micro- and macro-averaged F1 values for the test sets are correlated, which suggests that there is no effect to the clustering quality caused by the topic group size.

Test Sets	Evaluation	CS	SCS
1500	**Micro-F1**	0.308	0.358*
	Macro-F1	0.231	0.289*
2989	**Micro-F1**	0.267	0.285
	Macro-F1	0.186	0.219*
4391	**Micro-F1**	0.275	0.266
	Macro-F1	0.205	0.214
5845	**Micro-F1**	0.271	0.276
	Macro-F1	0.202	0.209
7316	**Micro-F1**	0.272	0.280
	Macro-F1	0.205	0.212

Table 5. Micro- and Macro-averaged F1 Values for Different Document Measurements on Test Sets

Clustering Evaluation on the Japan_Nuclear Test Set

In the above experiment, the evaluation data is artificially created by pooling tweets retrieved by very different queries, which only simulate the microblogging users' general browsing situation, where the clustered microblogs are varied in topics. In order to simulate a typical microblog search scenario, in this experiment, the effectiveness of the proposed method is only evaluated on one of the five topical group in the evaluation data—Japan Nuclear.

To test on this set, topics are firstly manually detected and labeled to provide ground truth. Since it is time-consuming work to detect topics for the whole collection, a randomly selected sample set consists of 445 tweets are used, which can represent the whole collection with a confidence interval of 3.9 at a 95% confidence level. Because in this sample set, 6 of the tweets are written in Japanese, and one contains only a meaningless term "FWD", the final number of tweets used for manual topic detection is 438. Totally 53 topics were identified by the author on this sample set. The top five topics (ranked according to the number of microblogs contained in the topic group) are listed in Table 9. When creating this topic list, the policies for identifying a topic were defined as: (1) each tweet is assumed to discuss only one topic, and (2) a new topic is appended to the list if it is not in the current list.

Then, another random sample set is created, which consists of 486 tweets, excluding tweets completely written in non-alphabetical letters or containing only meaningless terms. Two graduate students were invited and trained to assign the author's manually detected topics to tweets in this sample set. Note that this second sample set contains 127 overlapping tweets with the first sample set, which are used to validate the quality of the coding. Overall, 39 topics were assigned, with 13 tweets marked as no topic available by both coders. For the 127 overlapping tweets, coder I gets 97.6% agreement with the manual topic assignment in sample one, and coder II gets a 98.4% agreement. Totally, there are 29 disagreements between the two coders in sample two, which means a 94% agreement. Thus, the Cohen's Kappa coefficient between the two coders is 0.936. The final evaluation set is then created by using only the agreed tweets in sample set II, which contains 443 tweets covering 39 topics.

The document distance measurement to be evaluated is: semantic cosine similarity. Cosine similarity is used as a baseline. When manually set k = 39, the micro- and macro-averaged F1 values for each method are shown in Table 6. It can be observed that: (1) since several topics contain only one tweet, which is difficult to be separated from other tweets, the macro-F1 values for all three methods are significant lower than the micro-F1 values; and (2) although the proposed semantic cosine similarity is numerically larger than the cosine similarity slightly, it is not statistically significant on this test set.

	CS	SCS
Macro-F1	0.214	0.247
Micro-F1	0.442	0.465

Table 6. Macro- and Micro-Averaged F1 Values on Sample Japan_Nuclear Test Set

Effect of Topic Labeling Method

When comparing different topic labeling methods, the sample I test set of Japan_Nuclear collection is used. Because only 39 of the 53 manually detected topics are confirmed in the sample II set by human coders, totally 430 tweets, which cover these 39 topics, are used for evaluating topic labeling methods. The ground truth label for each topic is created by three graduate students collaboratively.

In this study, for each topic group, it was required to use no more than 5 labels.

The baseline is labeling using the most frequent terms. For both the frequent term method and our frequent concept method, if the system needs to select from candidate labeling terms with equal computation scores, an alphabetical order from A to Z is applied for selection. Since all tweets in this test set contain terms "Japan" and "Nuclear", these two terms are ignored in the evaluation if they are selected as labels. The macro- and micro-averaged F1 values for these two methods are shown in Table 7.

	Frequent Term	Frequent Concept
Macro-F1	0.317	0.416*
Micro-F1	0.320	0.419*

Table 7. Macro- and Micro-F1 Values for Different Topic Labeling Methods

As shown in Table 7, the proposed frequent concept topic labeling method statistically significantly outperforms the frequent term method. And for the top 5 largest topic groups, the labels generated by the frequent-concept method are shown in Table 8:

ID	Manually Selected Labels	Frequent Concept Topic Labels
1	end, nuclear crisis, plan, TEPCO, work	Japan, nuclear crisis, operator, TEPCO, Tokyo
2	alter, Chernobyl, nuclear crisis, top level, upgrade	Chernobyl, Japan, nuclear accident, nuclear crisis, nuclear level
3	Fukushima, nuclear plant, radioactive, sea, water	Fukushima, Japan, nuclear plant, radioactive, water
4	compensation, demand, evacuees, Fukushima, leave	compensation, demand, evacuee, Fukushima, Japan
5	Japan, nuclear health, study, watch, WHO	eye, Japan, health, world health organization, year

Table 8. Topic Labeling for Top 5 Largest Clusters (Labels are ordered alphabetically)

CONCLUSIONS AND FUTURE WORK

This study was inspired by real-life experience with the limitations of current microblog search result presentation. A review of the literatures suggested that effective topic clustering could be helpful. To mitigate the limitations of lexical clustering for short microblog posts, a microblog clustering method that leverages Wikipedia as an additional source of similarity evidence was proposed and experimentally evaluated. Relatively small but statistically significant improvements in cluster purity (over computing lexical similarity using the cosine measure) were obtained, and other measures and other evaluation settings provide confirmatory results. Large and statistically significant improvements were also shown to result from leveraging Wikipedia links for topic labeling.

As with any experimental study, practical factors resulted in a number of important limitations that might be addressed in future work. Most obviously, first, many other external sources of evidence might also be tried; examples include authors' social networks, post creation time and location, or content expansion using Web links contained in the microblog posts. Second, when making use of an external resource, two risks are involved. One is linking inappropriately, and the other is using the link in a way that hurts rather than helps. A nuanced analysis of these two errors could help to improve our present design. Third, several parameters were set arbitrarily in this study, and some attention to parameter tuning could yield improvements.

Looking ahead, an obvious next step is to put the resulting clusters in front of actual users, initially perhaps in a structured user study, but ultimately as a part of a deployed system. Techniques of this type may also prove to be useful for other short-text clustering settings, as might be encountered on YouTube, Flickr or eBay, thus opening additional directions for exploration. Wikipedia tells us that a journey of a thousand li begins with a single step,[4] and in this paper we have taken our first step in that journey.

ACKNOWLEDGMENTS

We are grateful to Jordan Boyd-Graber, Man Huang, Christina Pikas, Yan Qu and You Zheng for their comments on earlier versions of this paper and (in 3 cases) for help with annotation of the evaluation data.

REFERENCES

Allan, James. (2002). Introduction to Topic Detection and Tracking. The Kluwer International Series on Information Retrieval, James Allan, (Ed.), 1-16. Kluwer Academic Publishers.

Bates, M. J. (1979). Information Search Tactics. Journal of the American Society for Information Science, 30(4), 205-214. doi: 10.1002/asi.4630300406.

Broder, A. (2002). A Taxonomy of Web Search. SIGIR Forum, 36(2), 3-10. ACM. doi: 10.1145/792550.792552.

Cataldi, M., Caro, L. D., & Schifanella, C. (2010). Emerging Topic Detection on Twitter based on Temporal and Social Terms Evaluation. In Proceedings of the Tenth International Workshop on Multimedia Data Mining (MDMKDD '10).New York, NY, USA: ACM. doi: 10.1145/1814245.1814249.

Chen, J., Nairn, R., Nelson, L., Bernstein, M., & Chi, E. H. (2010). Short and Tweet: Experiments on Recommending Content from Information Streams. Scanning, 1185-1194. ACM. doi: 10.1145/1753326.1753503.

[4] En.wikipedia.org/wiki/Li_(unit)

Cutting, D. R., Karger, D. R., Pedersen, J. O., & Tukey, J. W. (1992). Scatter/Gather: a Cluster-based Approach to Browsing Large Document Collections. In Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, A. M. Pejtersen, & E. A. Fox (Eds.), 318-329. N. Belkin, P. Ingwersen: ACM. doi: 10.1145/133160.133214.

Fung, B. C. M., Wang, K., & Ester, M. (2003). Hierarchical Document Clustering Using Frequent Itemsets. In Proceedings of the SIAM International Conference on Data Mining, 30(5), 59–70. SIAM.

Genc, Y., Sakamoto, Y., & Nickerson, J. V. (2011). Discovering Context: Classifying Tweets through a Semantic Transform based on Wikipedia. Human Computer Interaction International. Springer.

Halliday M & Hansan R. (1976). Cohesion in English. London: Longman Group.

Huang, D., Xu, Y., Trotman, A., & Geva, S. (2008). Overview of INEX 2007 Link the Wiki Track. *Focused Access to XML Documents*. Springer Berlin Heidelberg. doi: 10.1007/978-3-540-85902-4_32.

Li, Y., Chung, S., & Holt, J. (2008). Text Document Clustering based on Frequent Word Meaning Wequences. Data & Knowledge Engineering, 64(1), 381-404. doi: 10.1016/j.datak.2007.08.001.

MacQueen, J. B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, L. M. L. Cam & J. Neyman (Eds.), 281-297. University of California Press.

Michelson, M., & Macskassy, S. A. (2010). Discovering Users ' Topics of Interest on Twitter : A First Look. In Proceedings of the fourth workshop on Analytics for noisy unstructured text data (AND '10), D. Lopresti, C. Ringlstetter, S. Roy, K. Schulz, & L. V. Subramaniam (Eds.), 73-80. New York, NY, USA: ACM. doi: 10.1145/1871840.1871852.

Mihalcea, R., & Csomai, A. (2007). Wikify!: Linking Documents to Encyclopedic Knowledge. In Proceedings of the 16th ACM Conference on Information and Knowledge Management, 233-242. ACM. doi: 10.1145/1321440.1321475.

Milne, D., & Witten, I. H. (2008). Learning to Link with Wikipedia. In Proceeding of the 17th ACM Conference on Information and Knowledge Management CIKM 08, 509. ACM. doi: 10.1145/1458082.1458150.

Mishne, G., & De Rijke, M. (2006). A Study of Blog Search. In Proceedings of the European Conference on Information Retrieval Research ECIR, 3936, 289-301. Springer. doi: 10.1007/11735106_26.

Naaman, M., Boase, J., & Lai, C. (2010). Is it Really About Me? Message Content in Social Awareness Streams. In ACM 2010 Conference on Computer Supported Cooperative Work, M. Ac (Eds.), 189-192. ACM. doi: 10.1145/1718918.1718953.

O'Day, V. L., & Jeffries, R. (1993). Orienteering in an Information Landscape: How Information Seekers Get from Here to There. In Proceedings of the INTERACT '93 and CHI '93 conference on Human factors in computing systems (CHI '93), 438-445. New York, NY, USA: ACM. doi: 10.1145/169059.169365.

Pal, A., & Counts, S. (2011). Identifying Topical Authorities in Microblogs. In Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, 45-54. New York, NY, USA: ACM. doi: 10.1145/1935826.1935843.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine, Series 5 50 (302): 157–175. doi:10.1080/14786440009463897

Phan, X.-H., Nguyen, L.-M., & Horiguchi, S. (2008). Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-Scale Data Collections. In Proceeding of the 17th international conference on World Wide Web WWW 08, 91. ACM. doi: 10.1145/1367497.1367510.

Porter, M. F. (1980). An algorithm for suffix stripping. Program, K. S. Jones & P. Willet, (Eds.), 14(3), 130-137. doi: 10.1108/00330330610681286.

Ramage, D., Dumais, S., & Liebling, D. (2010). Characterizing Microblogs with Topic Models. International AAAI Conference on Weblogs and Social Media 2010. The AAAI Press. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/ viewPDFInterstitial/1528/1846.

Teevan, J., Ramage, D., & Morris, M. R. (2011). TwitterSearch : A Comparison of Microblog Search and Web Search. WSDM'11, 35-44. Hong Kong: ACM.

Thorndike, R. (1953). Who Belong in the Family? Psychometrika, 18, 267-276.

Weng, J., Lim, E.-P., Jiang, J., & He, Q. (2010). Twitterrank: Finding Topic-Sensitive Influential Twitterers. In Proceedings of the Third ACM International Conference on Web Search and Data Mining, 261-270. ACM. doi: 10.1145/1718487.1718520.

Wilson, T. D. (1999). Models in Information Behaviour Research. Journal of Documentation, 55(3), 249-270. doi: 10.1108/EUM0000000007145.

Zamir, O., Etzioni, O., Madani, O., & Karp, R. M. (1997). Fast and Intuitive Clustering of Web Documents. In Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, D. Heckerman, H. Mannila, D. Pregibon, & R. Uthurusamy (Eds.), 287-290. AAAI Press.