RANKING LOCATION-DEPENDENT KEYWORDS FROM MICROBLOGS

Satoshi Ikeda, Nobuharu Kami and Takashi Yoshikawa
System Platforms Research Laboratories, NEC Corporation, 1753 Shimonumabe, Nakahara, Kawasaki, Kanagawa, Japan

Keywords: TF-IDF, Context Awareness, Keyword Ranking.

Abstract: The spread of microblogging services, such as Twitter, has made it possible to extract location-dependent context such as keywords specific to a geographical region, with fine granularity. The results of content analysis of microblogging services are affected by users who post excessive messages. In addition, because geographical granularity of users' interests differs, it is preferable to support multiple levels of granularity for usability. Thus, we propose a ranking method of location-dependent keywords based on a term frequency-inverse document frequency method, which takes into account diversity of information sources and supports multiple zoom levels of geographical areas by approximation. We evaluated our ranking method with a real dataset from Twitter and showed its effectiveness. We also describe a prototype implementation of a system using our ranking method.

1 INTRODUCTION

Microblogging services, such as Twitter, have been spreading worldwide in recent years. A characteristic of users frequently updating their status has made it possible to obtain users' context information with fine granularity in real time.

Accordingly, user content from microblogging services has been widely leveraged as a research object of user context (Tumasjan, Sprenger, Sandner and Welpe, 2010; Chen, Nairn, Nelson, Bernstein and Chi, 2010). In particular, context based on geographical location information, *location-dependent context*, is notable since users' activities are closely related to their location. Moreover, the popularization of smartphones equipped with GPS receivers has encouraged research on location-dependent context. In fact, many studies analyzing content from microblogging services based on geographical location information have been conducted (Arakawa, Tagshira and Fukuda, 2010; Sakaki, Okazaki and Matsuo, 2010).

One way to express location-dependent context is to extract and rank *location-dependent keywords*, which characterize a geographical region by analyzing content from microblogging services. In content analysis of microblogging services, the impact of advertisement messages must be taken into consideration. Since microblogging services are also used as advertisement tools, malicious users may try to compromise the analysis result by posting a large number of messages. Though spam filtering with a Bayesian filter (Sahami, Dumais, Heckerman and Horvitz, 1998) would be considered effective to determine whether a message is an advertisement or not, it requires training using sample messages in advance. Moreover, it is difficult to determine what should be filtered out as malicious since an advertisement message may be useful if it contains context related to its location. Hence, it is not adequate to filter out all advertisement messages. Additionally, the granularity of geographical areas with which location-dependent context is analyzed is important. For instance, an application that analyzes and visualizes location-dependent keywords should use geographical areas with appropriate granularity based on the zoom level of a map because sizes of geographical areas in which users are interested may differ. Therefore, it is preferable to support multiple levels of granularity with a method that can reduce computational cost and database size, because a naive approach must calculate and store ranking scores for all zoom levels.

We propose a ranking method of location-dependent keywords, which enhances a term frequency-invert document frequency (TF-IDF) method (Salton and Buckley, 1988). Diversity of information sources is taken into consideration with our method. Specifically, by penalizing keywords with low diversity of users, the impact of excessive

Ikeda S., Kami N. and Yoshikawa T..
RANKING LOCATION-DEPENDENT KEYWORDS FROM MICROBLOGS.
DOI: 10.5220/0003905206070615
In *Proceedings of the 8th International Conference on Web Information Systems and Technologies* (WEBIST-2012), pages 607-615
ISBN: 978-989-8565-08-2

repeating of messages by a few users is mitigated. Additionally, our ranking method supports multiple zoom levels of geographical areas by TF-IDF approximation. With this approximation, we need not to calculate and store TF and DF values for all zoom levels because values for only several zoom levels are stored and TF-IDF values for the intermediate zoom levels are interpolated. From evaluations using an actual dataset from Twitter, we show the effectiveness of user diversity and the accuracy of TF-IDF approximation. We also introduce a prototype implementation of a system using our ranking method.

The rest of this paper is organized as follows. Section 2 introduces our ranking method of location-dependent keywords. Section 3 describes the prototype implementation of a system using our ranking method. Section 4 discusses the evaluation of the effect of user diversity and approximation used in our ranking method. Related work is discussed in Section 5, and Section 6 gives the conclusions.

2 RANKING METHOD

In this section, we introduce our location-dependent keyword ranking method that enhances a TF-IDF method. In our method, diversity of information sources is utilized to suppress the impact of loud users. Moreover, our method supports multiple zoom levels of geographical areas by TF-IDF approximation.

2.1 Application of TF-IDF

The basic idea for extracting location context is to apply a TF-IDF method to ranking location-dependent keywords.

In contrast to original TF-IDF for determining how important a word is to a document in a collection of documents, the purpose of our approach is to determine how important a word is to a geographical area. For this purpose, we regard a collection of messages posted in a geographical area as a document of the TF-IDF method. A geographical point, latitude and longitude, tagged in a message is converted into a location label representing an area the point is located in. An area represented by a location label is one cell of a square grid on the Mercator projection of the earth. The size of an area is determined by zoom level. Specifically, a single cell covers almost the whole earth at zoom level 0 and is divided into four cells for each additional zoom level.

To label geographical points, we use tile coordinates, which are used in the Google Maps API (Google Inc., 2009) and reference a specific tile on a map at a specific zoom level. The tile coordinates (x, y) at zoom level z are determined from a geographical point with latitude φ and longitude λ as follows:

$$x = \left\lfloor 2^z \cdot \frac{\pi + \lambda}{2\pi} \right\rfloor \quad (-\pi \leq \lambda < \pi)$$

$$y = \left\lfloor 2^z \cdot \frac{\pi - \ln(\tan\varphi + \sec\varphi)}{2\pi} \right\rfloor \quad (-\varphi_0 < \varphi \leq \varphi_0)$$

where φ_0 is latitude φ such that $\tan\varphi + \sec\varphi = e^{\pi}$, that is, approximately 1.484 rad (approx. 85.05 degrees). For instance, tile coordinates $(0, 0)$ at zoom level 0 includes almost the whole earth.

We consider a TF-IDF-based keyword ranking method for each area whose location is expressed in the tile coordinate system. A TF-IDF value for a word w in an area with a location label l of tile coordinates (x, y) at a specific zoom level is calculated as follows:

$$\mathit{tfidf}_{w,l} = \mathit{tf}_{w,l} \cdot \mathit{idf}_w \tag{1}$$

where $\mathit{tf}_{w,l}$ is the number of occurrences of w in l. The inverse document frequency idf_w is defined as:

$$\mathit{idf}_w = \log_2 \frac{N}{n_w} \tag{2}$$

where n_w is the number of areas where w occurs at least once.

To accurately rank location-dependent keywords, we need to pay attention to a definition of N, which is simply the number of all areas in the case of the original TF-IDF method. Some words are given an unexpectedly high ranking score especially when a zoom level is high in which the minimum and maximum DF values tend to be close if we simply select the number of all areas for N. If we instead take the number of "active" areas that accommodate at least one word, this unexpected ranking problem is mitigated since we can exaggerate the difference in the location-dependence of words by enlarging the difference between the minimum and maximum DF values. Yet, there is still room for improvement and we propose to use $\max_w(n_w)$ instead of these selections of N. Figure 1 plots these selections of N against various zoom levels for a dataset. If the number of all areas is selected as N at zoom level 16, the IDF value of a word with the maximum DF value is about 9.3 and the IDF value of a word that occurs at a single area is about 23.0. This means that, for least and most location-dependent keywords w_L and w_M, w_L is ranked higher than w_M if w_L occurs only thrice of w_M. On the other hand, if the number of active areas is selected as N, the IDF values respectively change to 2.5 and 16.2.

By adopting the maximum DF value as N, IDF values for the least location-dependent keywords are zero. Consequently, a least location-dependent keyword is always ranked lowest.

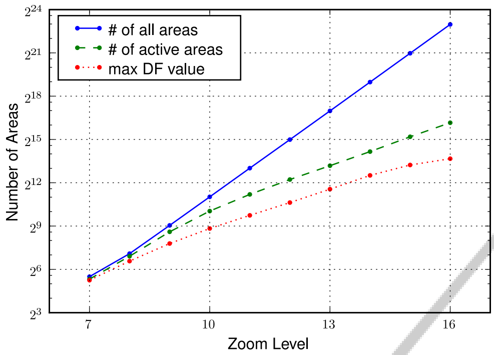


Figure 1: Selection of N in TF-IDF.

2.2 User Diversity Weighting

Microblogging services are also leveraged for commercial advertisements and announcements from public institutions. For instance, some restaurants provide time-limited coupons and some fire departments announce information on responses to 911 calls.

In such cases, the frequency of posts from such services is relatively high on a constant basis such as tens of posts a day. Moreover, since these services usually use message templates, the messages have a strong tendency to include specific words related to the services or the users. As a consequence, TF values of such user specific keywords tend to increase, as well as TF-IDF values. This indicates that malicious users can easily juggle keyword ranking simply by posting a large number of messages with target keywords.

To prevent such user-specific keywords from becoming too influential, we introduce using a diversity index of each keyword, which is a measure that represents how many users equally originate a keyword, for penalizing the ranking scores of those keywords.

There are several diversity indices such as Shannon's diversity index and Simpson's diversity index. We use Simpson's due to its simple definition. The user diversity index for a word w in an area l is defined as follows:

$$D_{w,l} = 1 - \sum_{u \in U} \left(\frac{n_{w,l,u}}{n_{w,l}} \right)^2$$

where U is a set of users, $n_{w,l}$ is the number of occurrences of w in l, and the user term frequency $n_{w,l,u}$ is the number of occurrences of w from a user u in l.

We use the user diversity index to lower the ranking of keywords from malicious users. This index ranges from zero to one and approaches one when a word is uniformly posted by many users. Because of these characteristics, the ranking of keywords from malicious users is lowered simply by multiplying TF-IDF values by the user diversity index. We define diversity-weighted TF-IDF (DTF-IDF) as follows:

$$dtfidf_{w,l} = D_{w,l} \cdot tfidf_{w,l}.$$

2.3 Zoom Support

It is preferable to provide rankings for each zoom level for usability since the granularity of interest may differ among users. Because a change in zoom level alters the geographical partitioning, the simplest solution is to calculate and store TF-IDF values for each zoom level. However, this requires a huge database whose size is approximately proportional to the number of TF entries. Moreover, the computational cost is also roughly proportional to these entries. Figure 2 shows the number of TF entries for each zoom level for a dataset of actual tweets collected from Twitter, which is described in detail in Section 4. If we maintain TF tables at each zoom level from 7 to 16, we need to keep about 15 million TF entries. Additionally, calculation of the user diversity index in real time needs to maintain the number of occurrences of words at all areas for each user. The total database size required for the dataset was about 1.3 GB.

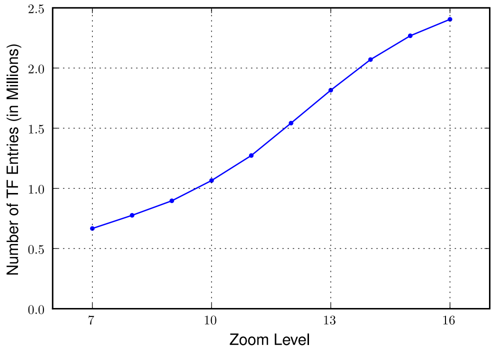


Figure 2: Number of TF entries at each zoom level.

To reduce the database size, we introduce an approximation approach for calculating TF-IDF values. Instead of maintaining calculated results for all zoom levels, calculated results are stored for several zoom levels to the database and TF-IDF values for the omitted zoom levels are approximated from the stored zoom levels. This approximation approach also has another advantage in that it supports various sizes of a target area. An approximated TF-IDF value for an area consisting of 3×3 sub-areas, for example, can be calculated with this approximation approach even if partitioning with 3 × 3 sub-areas is not supported.

2.3.1 IDF Approximation

For approximating TF-IDF values, IDF values for the omitted zoom levels must be estimated. This can be accomplished by interpolating DF values.

In general, location dependency of keywords varies according to the size of the target areas. For instance, if the size is large enough, a name of a nationwide chain store would have least location-dependency (IDF value) since the name would be posted in almost all areas. In contrast, in a small area, the name would have a relatively high dependency since it would be posted within a narrow range near the stores. Hence, we cannot simply use an IDF value for other zoom levels available in a database as the value for the target zoom level. Therefore, appropriate DF values should be used for TF-IDF approximation.

While one might think variations in DF values for all words have similar tendency to the number of areas, this is not the case. Figure 3 shows variations in DF values of some words and the number of active areas. The DF values monotonically increase as the zoom level increases. The shapes of the curves, however, differ from each other including "# of active areas". The DF value of "aid response" increases quickly from zoom level 13 to 16. In contrast, the DF value of "san francisco" increases gradually in this range.

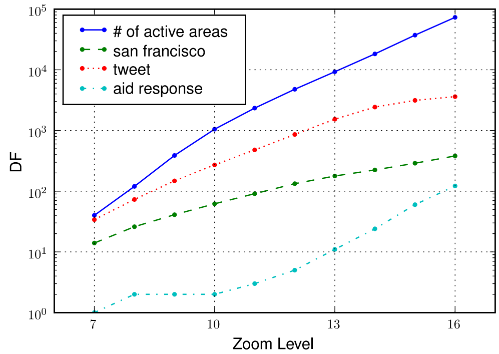


Figure 3: DF values for some words for each level. Line with "# of active areas" is number of areas having at least one word.

The tendency seen in Figure 3 is that the segments that result from splitting the curves several ways are roughly approximated by straight lines in linear-log space. Hence, some interpolation approaches in linear-log space are used to yield good approximation results. With linear interpolation in linear-log space, we store DF values for K zoom levels $\{\xi_1, \xi_2, \ldots, \xi_K\}$ with $\xi_i < \xi_{i+1}$. The DF value df_z at zoom level $z \in (\xi_i, \xi_{i+1})$ is interpolated by the following equation:

$$df_z = df_{\xi_i}\left(\frac{df_{\xi_{i+1}}}{df_{\xi_i}}\right)^{\frac{z-\xi_i}{\xi_{i+1}-\xi_i}}.$$

A maximum DF value used for calculating IDF values is also approximated by this equation.

2.3.2 TF Approximation

Precise TF values are calculated from those at a larger zoom level by just summing up TF values in all sub-areas included in a target area for each word. The user diversity indices are also calculated by aggregating user term frequencies in all sub-areas.

However, the computational cost of calculating TF-IDF values from TF and DF tables is not negligible since aggregation is required per query. Figure 4 shows the number of words in the top 250 areas in decreasing order for zoom levels 7, 10, 13, and 16.

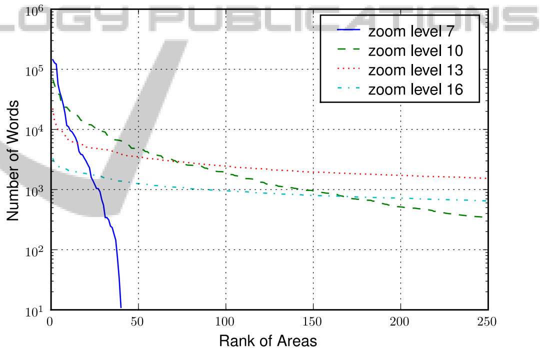


Figure 4: Number of words in top 250 areas in decreasing order.

There were 144,251 words in the area with the largest number of words at zoom level 7. If the naive approach is taken, it requires aggregation for all the words per query for accuracy. Because of this, it is not realistic to calculate precise values for such areas with a large amount of words.

Hence, approximation is taken for TF aggregation to reduce the computational cost in areas with thousands of words. Our approximation approach limits the number of words taken from each sub-area, that is, for some integer k, only the words with top k TF values in each sub-area are taken for aggregation and the others are ignored.

While this approximation may yield TF-IDF rankings with less accuracy, the choice of k can improve accuracy. Undoubtedly, a large enough k results in precise TF values, while it increases the computational cost. The effect on rankings of this approximation is evaluated in Section 4.3.

3 IMPLEMENTATION

With the ranking method described above, we implemented a ranking system for location-dependent keywords. In this section, we briefly introduce our ranking system.

3.1 Architecture

The system architecture is shown in Figure 5. Our system collects geotagged tweets from Twitter using the Streaming API (Twitter Inc., 2010), making it possible to collect tweets within a specific geographical area in real time.

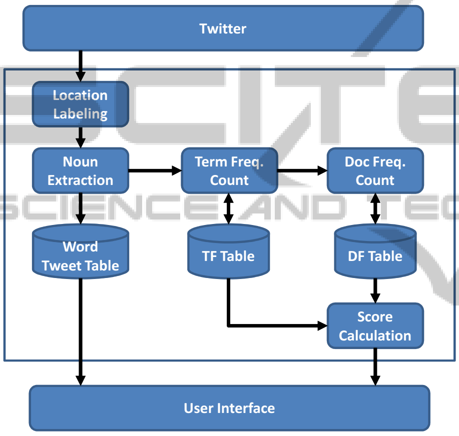


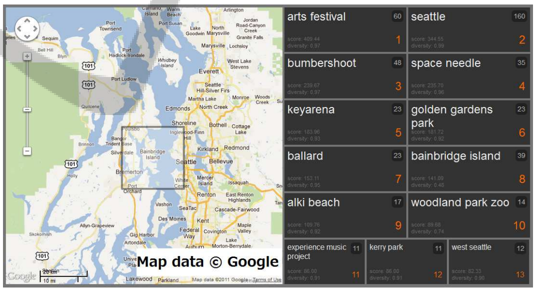
Figure 5: Architecture of our ranking system.

When the system receives a tweet via the Streaming API, the geographical point (latitude, longitude) of the tweet is converted to a location label (tile coordinates) which represents a geographical area. Next, nouns (or noun phrases) in the tweet are extracted as candidate keywords. We use a part-of-speech (POS) tagger for noun extraction. To show tweets that contain such nouns, the mapping between a tweet and the nouns are stored in a database. Then, the number of occurrences of nouns are updated and stored into the TF table. At the same time, user term frequencies are also stored for calculating user diversity indices. If a noun is found to be the first occurrence in the area in the TF counting process, the DF value of the noun is also updated and stored in the DF table. Thus, the TF and DF tables are kept updated in real time. A ranking is generated per request by calculating DTF-IDF scores using the approximation described in Section 2.

On a machine with Intel Xeon 3.1 GHz and 4 GB memory, the response time per request is shorter than 300 milliseconds even for areas with a large number of words if the database is in memory. In contrast, it is shorter than 100 milliseconds for rural areas with not many words. The response time is considered to be improved by caching generated rankings for areas that requests concentrate on.

3.2 User Interface

Our system ranks keywords depending on a specified geographical area by analyzing tweets with geographical information. Figure 6 shows a user interface of the ranking system. When a user clicks at a point on the map, a keyword ranking in an area that contains the clicked point is displayed on the right side. While only the top 13 keywords are shown in Figure 6, the ranking area is scrollable and the top 100 rankings are calculated by default. A user can look at tweets that contain a keyword by selecting it if he or she wants to know why the keyword is rated highly. The zoom level of tile coordinates used to specify a target area is basically one zoom level larger than that of the map. In other words, the size of a target area varies by the zoom level of the map. This makes it possible for users to specify a target area with granularity of interest.


arts festival 60
score: 409.44
diversity: 0.97
1

seattle 160
score: 344.55
diversity: 0.99
2

bumbershoot 48
score: 239.67
diversity: 0.97
3

space needle 35
score: 235.70
diversity: 0.96
4

keyarena 23
score: 183.96
diversity: 0.93
5

golden gardens park 23
score: 181.72
diversity: 0.92
6

ballard 23
score: 153.11
diversity: 0.95
7

bainbridge island 39
score: 141.09
diversity: 0.48
8

alki beach 17
score: 109.76
diversity: 0.92
9

woodland park zoo 14
score: 89.68
diversity: 0.74
10

experience music project 11
score: 86.00
diversity: 0.91
11

kerry park 11
score: 86.00
diversity: 0.91
12

west seattle 12
score: 82.33
diversity: 0.90
13

Figure 6: User interface.

Although the prototype system simply provides a keyword ranking, the ranking can be used also for keyword recommendation or completion. When one inputs a text with a smartphone out the door especially at a location that he or she visit for the first time, its content should relate closely his or her geographical context. By recommendation or completion of location-dependent keywords, he or she would be able to notice the excitement at the location which is usually passed over unnoticed.

4 EVALUATION

We obtained 866,420 geotagged tweets via the Twitter Streaming API from Sept. 3 to Sept. 16, 2011. The geographical area covering the west coast of the United States was specified as the query parameter.

Table 1: Example results of user diversity weighting.

	TF-IDF		DTF-IDF	
rank	*keyword*	*diversity*	*keyword*	*diversity*
1	arts festival	0.97	arts festival	0.97
2	seattle	0.99	seattle	0.99
3	bainbridge island	0.48	bumbershoot	0.97
4	aid response	0.00	space needle	0.96
5	bumbershoot	0.97	keyarena	0.93
6	space needle	0.96	golden gardens park	0.92
7	keyarena	0.93	ballard	0.95
8	golden gardens park	0.92	bainbridge island	0.48
9	new listing	0.00	alki beach	0.92
10	e 21	0.00	woodland park zoo	0.74

The south-west and north-east corners of the area were respectively (32.0, -125.0) and (49.0, -114.0).

We evaluated the effect of the user diversity weighting, errors in IDF approximation, and the effect of approximation on keyword rankings using this dataset.

4.1 User Diversity Weighting

Table 1 lists the example results of the effect of user diversity weighting. These are top ten rankings of location-dependent keywords with and without user diversity weighting in the area with tile coordinates (163, 357), a section of Seattle, WA, at zoom level 10.

The TF-IDF ranking without user diversity weighting contained keywords of ‘aid response’, ‘new listing’ and ‘e 21’. Of these keywords, ‘aid response’ and ‘e 21’ were a part of announcements related to 911 calls posted by one user, and ‘new listing’ was a part of advertisements from a real-estate agency. Though they are indeed location-dependent in the sense that they have large IDF values, the ranks are considered to be overrated due to one user’s massive tweets.

With user diversity weighting, the ranking is adjusted so that the impact of loud users are mitigated. The keywords above were filtered out by the user diversity index of 0.00. At the same time, keywords posted by many users can maintain higher ranks. For instance, the keyword ‘bumbershoot’, a name of a festival, which was posted by 40 users, had a user diversity index of 0.97 and its rank rose from 5th to 3rd.

4.2 IDF Approximation

Cubic interpolation in linear-log space also promises a better outcome than linear interpolation since it yields smooth and continuous curves that linear interpolation can not yield. In this section, we compare the linear interpolation of DF values described in Section 2.3.1 with cubic interpolation. The comparison results show that the two interpolations of DF values in linear-log space did not show significant differences in errors of approximated IDF values.

There were 432,132 words in the dataset, and 374,992 words of them had the same DF values at zoom levels 7 and 16. Since such words yield no errors by the interpolations, these words were excluded from the evaluation. We evaluated the error in IDF values of the remaining 57,140 words. We used DF values at zoom levels 7, 10, 13 and 16 and approximated IDF values at the six intermediate levels.

Errors of approximated IDF values are summarized in Table 2. The errors are classified by the range of precise IDF values (*precise IDF*). In both interpolation methods, the errors tend to decrease as IDF values decrease. The fact of low errors for words with small IDF values indicates that approximation can keep location-independent words as they are.

Another finding is that there is no significant difference in both Rooted Mean Squared Errors (RMSEs) and Mean Absolute Errors (MAEs) between linear and cubic interpolations. In both methods, RMSE and MAE were respectively about 0.24 and 0.19.

From the evaluation results, the approximation by cubic interpolation had no significant advantages compared with linear interpolation. In contrast to linear interpolation requiring only two zoom levels, cubic interpolation requires at least four zoom levels. This would increase database access and computational cost. For this reason, we chose linear interpolation for approximation.

4.3 Effect on Rankings

As mentioned above, for ranking location-dependent keywords, it is not necessarily required to calculate precise TF-IDF values for all candidate words. The important thing is to provide precise ranking against perfect ranking which is obtained from precise TF-IDF calculations.

Table 2: Comparison of approximation errors.

	Linear		Cubic		
precise IDF	RMSE	MAE	RMSE	MAE	rate (%)
< 2.0	0.082	0.065	0.090	0.071	1.3
< 4.0	0.150	0.122	0.154	0.123	8.9
< 6.0	0.235	0.185	0.225	0.178	19.3
< 8.0	0.257	0.182	0.270	0.196	24.9
$8.0 \leq$	0.261	0.213	0.255	0.209	45.6
ALL	0.245	0.190	0.244	0.190	

For evaluation of rankings by the approximation described in Section 2.3.2, we used the normalized discounted cumulative gain nDCG$_k$ (Järvelin and Kekäläinen, 2002). nDCG$_k$ for a top-k ranking is defined as:

$$\mathrm{nDCG}_k = \frac{\mathrm{DCG}_k}{i\mathrm{DCG}_k}$$

where $\mathrm{DCG}_k = R_1 + \sum_{i=2}^{k}(R_i/log_2 i)$. R_i is the relevance value of a word at rank i, which takes a large value if the word strongly depends on the location, and iDCG$_k$ is DCG$_k$ for perfect ranking. From this definition, nDCG$_k$ is such a value that the more similar a ranking is to an ideal one, the closer it is to one. Thus, if nDCG$_k$ for an approximated ranking is close to one, the approximation is considered to be highly accurate. We use precise TF-IDF values of a word at rank i as relevance R_i since a word with a higher TF-IDF value is considered to be strongly dependent on a geographical area.

In the evaluation below, the results in Figures 7, 8 and 9 are the average nDCG$_k$ values of the top ten areas with a large number of words at a specified zoom level.

Figure 7 shows the evaluation results for TF approximation at zoom level 8 using TF values at the base zoom level of 10. The x-axis represents the number of TF entries used in the approximation. If it is 1,600, for instance, the top 100 TF entries for each sub-area are collected since there are 16 sub-areas to aggregate. In this evaluation, we used the precise IDF values without approximation and ranked by TF-IDF instead of DTF-IDF values, without user diversity weighting, to evaluate the effect of TF approximation.

The number of entries per sub-area to achieve a highly accurate ranking is not large. There were about 275,000 words on average in the top ten areas at zoom level 8. For $k = 10$, 25, and 100, respectively, about only 70, 140, and 310 TF entries per sub-area were needed to achieve nDCG$_k$ over 0.99. Since the average number of words per area is less than that at smaller zoom levels, TF entries required to achieve this accuracy are considered to be less than these results for a zoom level larger than 8.

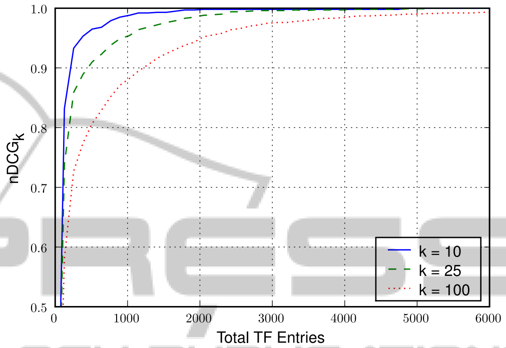


Figure 7: Evaluation of TF approximation at zoom level 8 by aggregating TF entries at base zoom level 10.

The differences among base zoom levels at which TF values were used to approximate TF values at zoom level 8 are shown in Figure 8. Similar to the evaluation above, we used TF-IDF rankings with precise IDF values and the approximated TF values without user diversity weighting. For $z_b = 8$, approximation was performed simply by ignoring words with low TF values. For base zoom levels $z_b = 9$ and $z_b = 10$, there were respectively 2×2 and 4×4 sub-areas to aggregate.

The results show that the total number of TF entries required for equivalent accuracy is not proportional to the number of sub-areas. For nDCG$_{100} \geq 0.99$, the required total entries doubled or tripled when the base zoom level increases by one while the number of sub-areas increases to four times.

In the next evaluation, the effect of IDF approximation is taken into consideration. Figure 9 shows nDCG$_k$ values with and without IDF approximation.

With IDF approximation, the nDCG$_k$ takes small values compared to the case without approximation regardless of k. This is because the errors in IDF approximation make perfect ranking impossible. To be more precise, in a TF-IDF result ranking with IDF approximation, since the approximated TF-IDF values are not necessarily the same as the precise ones, it is impossible to generate a perfect ranking even if all TF entries for each sub-area are aggregated. However, our approximation is considered to have good

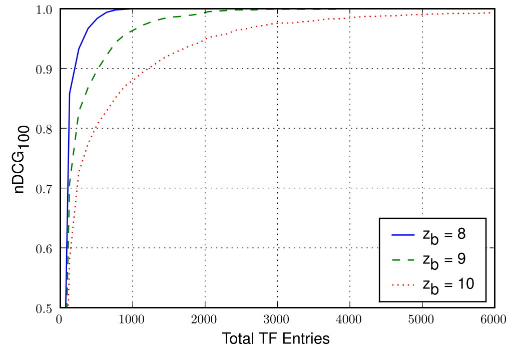


Figure 8: Evaluation of approximation for different base zoom levels z_b.

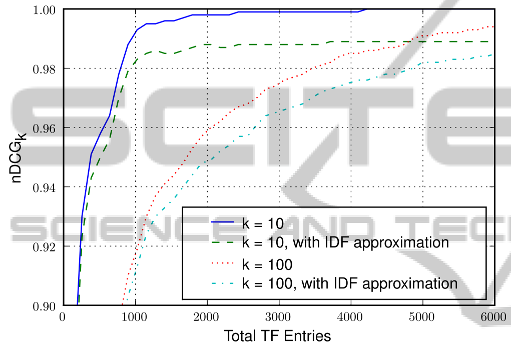


Figure 9: Effect of IDF approximation.

accuracy since the differences in nDCG_k between the cases with and without IDF approximation are less than 0.02.

5 RELATED WORK

Many studies have been conducted for analyzing content in microblogging services and leveraged geographical location context. Arakawa et al. (2010) introduced an extraction method of location-dependent keywords by extracting grids with high density of specific keywords using breadth-first search. It needs to specify a target keyword to calculate its dependency and is suitable for detecting locations the keyword depends on. On the other hand, it is not suitable for ranking keywords that depend on a specific location. TwitterStands (Sankaranarayanan, Samet, Teitler, Lieberman and Sperling, 2009) is a news processing system that analyzes tweets and detects late breaking news with a geographic focus. The geographic focus, which is determined from tweet text and metadata by *geotagging*, is calculated by ranking the geographic locations in a topic cluster. The approach determines the geographic focus after topic clustering. Therefore, the geographic focus might concentrate in areas where many tweets are posted. Sakaki et al. (2010) proposed an event detection scheme using a Kalman filter in real time and estimated earthquake epicenters and typhoon trajectories from Twitter data. Tweets with pre-defined keywords are regarded as sensor data of a target event. Mei, Liu, Su and Zhai (2006) proposed a probabilistic mixture model and analyzed spatiotemporal theme patterns on weblogs (not on microblogs) with the model. Since the granularity of a location is not flexible, parameter estimation of the model must be reperformed when the granularity is changed. Thus, support of multiple levels of geographical granularity has not been discussed sufficiently.

TF-IDF methods are widely used in analysing tweets. TwitterStands uses TF-IDF for weighting important words to cluster topics. Eddi (Bernstein et al., 2010) assigns topics to a tweet using TF-IDF-style key terms obtained from search results of nouns in the tweet. Chen et al. (2010) studied URL recommendation on Twitter. In their approach, the recommendation is made based on a user profile which is a TF-IDF vector generated from his/her tweets. A set of tweets from a user is regarded as a document; in contrast, a set of tweets in an area is regarded as a document in our approach. Our user diversity weighting is applicable for recommendation to mitigate the impact of malicious users.

Diversity of users in a geographical area is useful also for purposes other than mitigating the impact of loud users. Cranshaw, Toch, Hong, Kittur and Sadeh (2010) examined connection between an online social network and the location traces of its users. They showed that users who visit a location with high diversity tend to have more connections in the social network. Toch et al. (2010) showed that users appear more comfortable sharing their presence at locations with high diversity.

6 CONCLUSIONS

We proposed a location-dependent keyword ranking method for microblogging services, which adopts a TF-IDF method to geographical location context, and described the prototype implementation of a keyword ranking system. Our ranking method penalizes keywords with low user diversity and supports multiple zoom levels of geographical granularity by using TF-IDF approximation. The evaluation results showed that user diversity weighting is effective in mitigating the effect of excessive posts from a few users and approximation can yield a highly accurate ranking in terms of similarity to precise TF-IDF ranking without approximation. We plan to extend our method for spatiotemporal analysis so that it can track trends and the spread of location-dependent keywords.

REFERENCES

Arakawa., Y., Tagashira., S., and Fukuda, A. (2010). Extraction of Location Dependent Words from Twitter Logs (in Japanese). *IPSJ SIG Technical Reports*, 2010-MBL-55(10):1–6.

Bernstein, M., Suh, B., Hong, L., Chen, J., Kairam, S., and Chi, E. (2010). Eddi: Interactive Topic-based Browsing of Social Status Streams. In *UIST 2010: 23rd Annual ACM Symposium on User Interface Software and Technology*, pages 303–312.

Chen, J., Nairn, R., Nelson, L., Bernstein, M., and Chi, E. (2010). Short and Tweet: Experiments on Recommending Content from Information Streams. In *CHI 2010: 28th International Conference on Human Factors in Computing Systems*, pages 1185–1194.

Cranshaw, J., Toch, E., Hong, J., Kittur, A., and Sadeh, N. (2010). Bridging the Gap between Physical Location and Online Social Networks. In *UbiComp 2010: 12th ACM International Conference on Ubiquitous Computing*, pages 119–128.

Google Inc. (2009). Google MAPs JavaScript API V3. Retrieved October, 2011, from http://code.google.com/intl/en/apis/maps/documentation/javascript/.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated Gain-based Evaluation of IR Techniques. *ACM Transactions on Information Systems*, 20(4):422–446.

Mei, Q., Liu, C., Su, H., and Zhai, C. (2006). A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs. In *WWW 2006: 15th International Conference on World Wide Web*, pages 533–542.

Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E. (1998). A Bayesian Approach to Filtering Junk E-mail. In *AAAI-98 Workshop on Learning for Text Categorization*.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In *WWW 2010: 19th International Conference on World Wide Web*, pages 851–860.

Salton, G. and Buckley, C. (1988). Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, 24(5):513–523.

Sankaranarayanan, J., Samet, H., Teitler, B., Lieberman, M., and Sperling, J. (2009). Twitterstand: News in Tweets. In *ACM SIGSPATIAL GIS 2009: 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 42–51.

Toch, E., Cranshaw, J., Drielsma, P., Tsai, J., Kelley, P., Springfield, J., Cranor, L., Hong, J., and Sadeh, N. (2010). Empirical Models of Privacy in Location Sharing. In *UbiComp 2010: 12th ACM International Conference on Ubiquitous Computing*, pages 129–138.

Tumasjan, A., Sprenger, T., Sandner, P., and Welpe, I. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *ICWSM 2010: 4th International AAAI Conference on Weblogs and Social Media*, pages 178–185.

Twitter Inc. (2010). Streaming API | Twitter Developers. Retrieved October, 2011, from https://dev.twitter.com/docs/streaming-api.