Search Real-time Tweet

Technology Reports

Development of Real-time Search Services Offering Daily-life Information

The SNS provided by Twitter, Inc. is popular as a medium for reporting on events around the world in real-time. NTT DOCOMO is offering search services that provide useful information (daily-life information) regarding railway operational conditions, TV programs and landmarks, from the large volume of tweets. In this article, we introduce these search services and describe the search technology behind them.

Service & Solution Development Department

Daisuke Torii
Hayato Akatsuka
Keiichi Ochiai
Kousuke Kadono

1. Introduction

Twitter[*1] has become established as a Social Networking Service (SNS) for submitting and sharing large numbers of messages about events occurring around the world, and huge numbers of tweets[*2] are shared every day. Much of the information shared is personal and very local to individual users, but much information useful in daily life is also included. However, it is not always an easy task to choose suitable search keywords for obtaining the desired information.

As such, NTT DOCOMO provides different types of search services that enable them to easily monitor useful tweets in daily life. We have developed search services that are highly compatible with Twitter, especially for tweets related to railway operational conditions, television programs and landmarks[*3]. Tweets related to railway operational conditions and landmarks are useful as on-the-scene reports of conditions. Tweets related to TV programs are useful for the viewers to share their thoughts and impressions of the programs in real-time, which can make viewing more enjoyable.

What is important for these searches is selecting tweets that are relevant to each domain (railway operational conditions, TV programs or landmarks) and indexing them in real-time. The accuracy of selecting tweets related to the domain is an especially significant issue for these services. Our approach is to select relevant tweets by matching tweet text with a domain-specific dictionary (e.g.: a train-line or a landmark dictionary). To improve accuracy further, we use additional techniques tailored to each domain. In this article, we introduce each search service and describe the search technologies behind them.

2. Railway Operational Condition Tweet Search Service

2.1 Service Overview

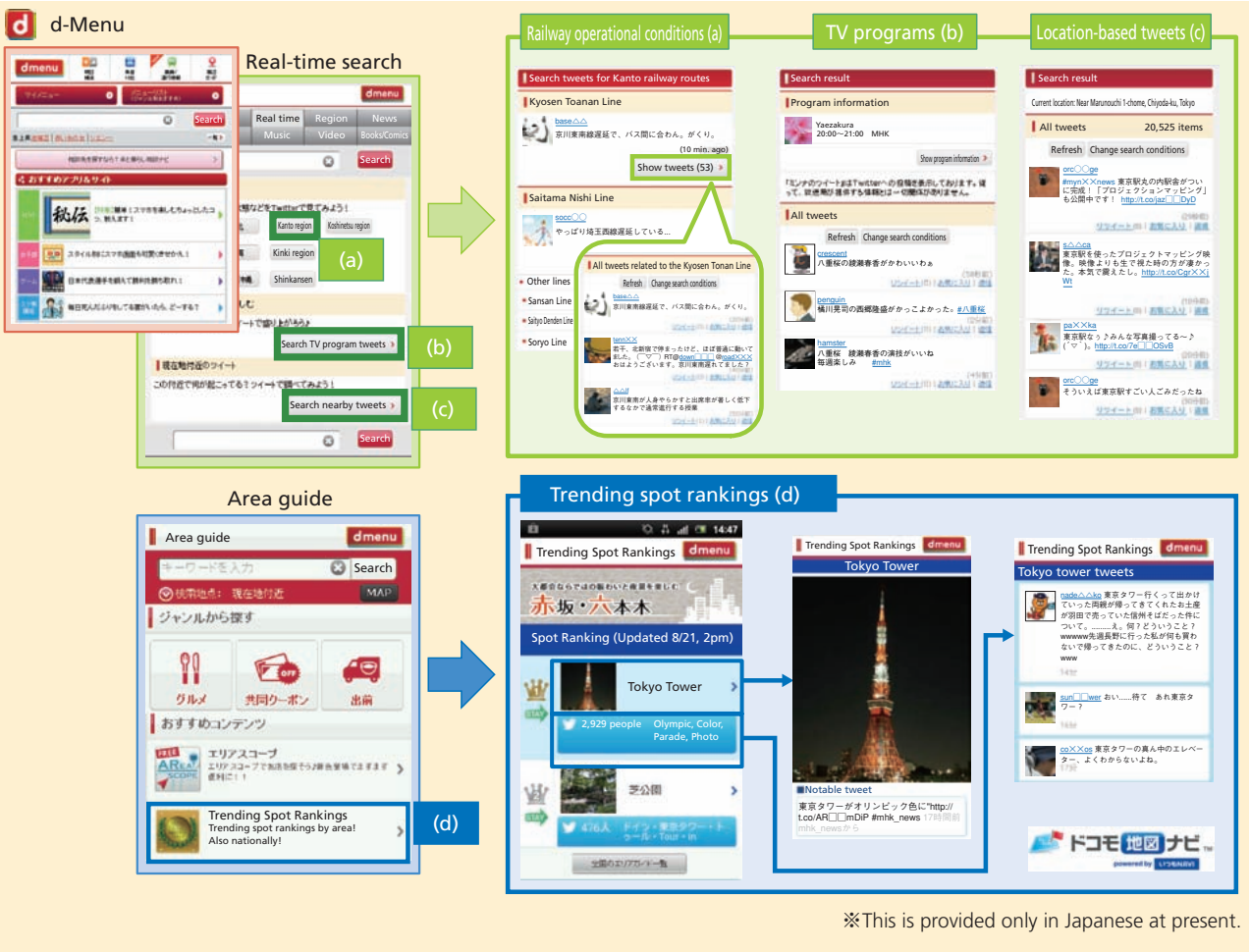
This service allows users to monitor tweets related to railway operational conditions, or so called "train tweets." Train tweets comprise tweets that include information about railway service delays, suspensions, and other railway operational issues throughout

*1 **Twitter**: The name Twitter, the logo and the Twitter bird are registered trademarks of Twitter, Inc. in the U.S.A. and other countries.
*2 **Tweet**: A term used to refer to entries on the micro-blogging service provided by Twitter Inc.

Japan. This search system detects railway problems in real time, and ranks railway routes based on the degree of exposure in Twitter due to the problems. **Figure 1**(a) shows the Web page of the real-time search service. The Web page has a section marked "Train-related tweets," followed by buttons for each region in Japan. The user is allowed to select a specific region in order to view a list of railway routes which currently have service problems. As an example, when the user clicks the "Kanto" button, a page with a list of railway routes in the Kanto region is displayed. This listing is in order of the attention they are receiving in Twitter. In this page, the most recent tweets for the two highest-ranked railway routes are also shown (See Kyosen Tonan Line and Saitama Nishi Line). To monitor train tweets for a specific railway route, the user simply selects the desired railway route by clicking the "View tweets" button.

2.2 Train-Tweet Processing System

An overview of processing for train tweets is shown in **Figure 2**. For the train tweet search, the railway routes that are receiving much attention in Twitter are extracted by counting the number of train tweets in real time. Tweets are processed by (1) mapping tweets to railway routes, (2) extracting tweets related to railway operational conditions to produce train tweets, (3) registering train tweets in the search



※This is provided only in Japanese at present.

Figure 1 Overview of real-time search service providing daily-life information

*3 **Landmark**: A structure that is famous or symbolic of a given location.

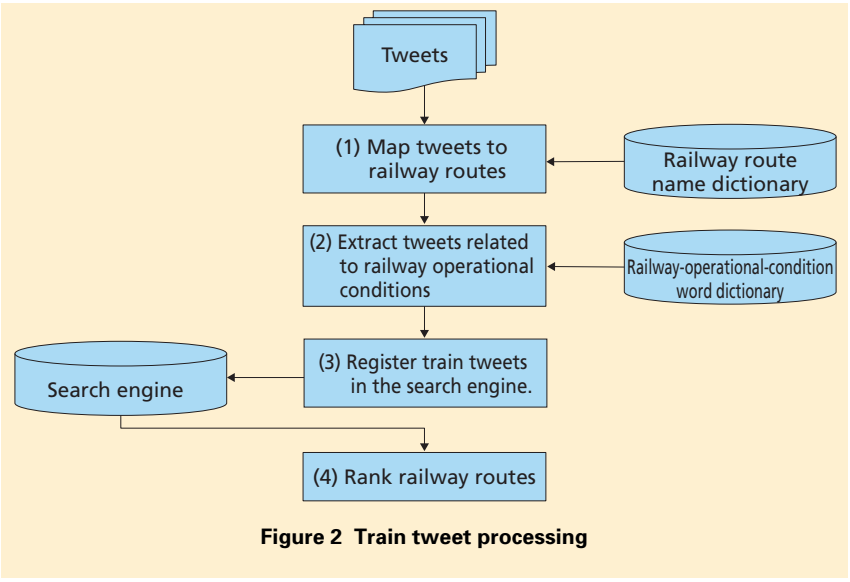


Figure 2 Train tweet processing

engine and (4) ranking railway routes by counting train tweets.

In order to map tweets to railway routes, tweets that include names of railway routes are detected on the server using a specific dictionary which contains a list of names of railway routes and variations in these names. Later steps are skipped for tweets that do not include any railway route names. To extract tweets related to railway operational conditions from the remaining tweets, words in the remaining tweets are compared with words related to railway operational conditions that are pre-registered in a dictionary. Only tweets which include railway-operational-condition words are stored as train tweets in the search engine. Finally, railway routes are ranked by counting train tweets for each railway route and sorting by the numbers of train tweets. Railway routes with a number of train tweets above a certain threshold are displayed to users. The processes of mapping tweets to railway routes, extracting train tweets, and registering train tweets in the search engine are completed in several seconds. This allows users to monitor railway operational conditions for each railway in real time. The ranking of railway routes is performed every few minutes as a batch process.

3. TV Program Tweet Search Service

3.1 Service Overview

This service allows users to monitor "TV tweets" which are tweets that include impressions and comments about TV programs being broadcast currently (Fig. 1(b)). This search system infers tweets that reference TV programs and groups these TV tweets by TV program to provide a social viewing[*4] experience to the users.

The home page of the TV tweets search system displays a list of programs that are broadcast nation-wide in Japan. Since most users are interested in such programs, they can easily access TV tweets belonging to these nation-wide broadcast programs by clicking the "View" button located beside each program on the home page. The search service also provides accesses to TV tweets for local TV programs. Users can first select the specific area in which a local TV program is broadcast, and then select the desired local TV program to view TV tweets. TV tweets are not only associated with programs that are currently broadcasting, but also associated with future programs. The search system tracks the programs scheduled up to one week in the future.

The search system applies two methods for associating tweets to TV programs. The first method is to infer hashtags[*5] closely associated with a broadcast program. Thus, any tweets which include such a hashtag (or "program-specific hashtags") can be easily assigned to a specific TV program. The second method is to extract words closely associated with a broadcast program dynamically, and to assign tweets which include the extracted words ("program-specific characteristic words") to a specific TV program. A list of TV tweets for a specific TV program can be sorted on the page in the

*4 **Social viewing**: A practice in which multiple users share his/her viewing experience of the same event or broadcasted program through social media.

*5 **Hashtag**: A function whereby placing a hash character ("#") at the beginning of a word in a tweet makes it easier for other users to find other tweets on the same subject (e.g., #earthquake).

order of posting time or number of retweets.

3.2 Real-time Selection of TV Tweets

Figure 3 shows the real-time extraction process for TV tweets. Since the broadcasting channel broadcasts TV programs sequentially, the search system must distinguish, in real time, between tweets about the currently broadcasting program and tweets about the program broadcast previously on the same channel. The system must also distinguish between TV tweets associated with the specific TV program and those associated with other currently broadcast TV programs. This complication is solved by two tasks which run in parallel. The first task is to extract program-specific characteristic words and program-specific hashtags. The second task is to extract TV tweets using the hashtags and the characteristic words, relate TV tweets to TV program and register them in the search engine.

The process for extracting program-specific characteristic words and program-specific hashtags is summarized below. First, a list of currently broadcast TV programs, including a unique program ID for each TV station, are obtained from the TV program database (Fig. 3(1)). The TV program database stores TV program meta data, which is updated in real time. Meta data for each TV program contains a program ID as well as general information about the TV program, such as the title, the cast, start time and end time. The set of tweets for a fixed period of time in the past are extracted from the tweet database (Fig. 3(2)). The tweet database contains all tweet data. Old tweets are not useful for extracting program-specific characteristic words and program-specific hashtags, so only recent tweets are used. Tweets which include hashtags of the TV station are extracted from the tweets extracted in (2) in order to create the set of TV program tweets (Fig. 3(3)). If multiple TV programs with different IDs are logically the same TV program, they are merged into one TV program. Hence, each logically

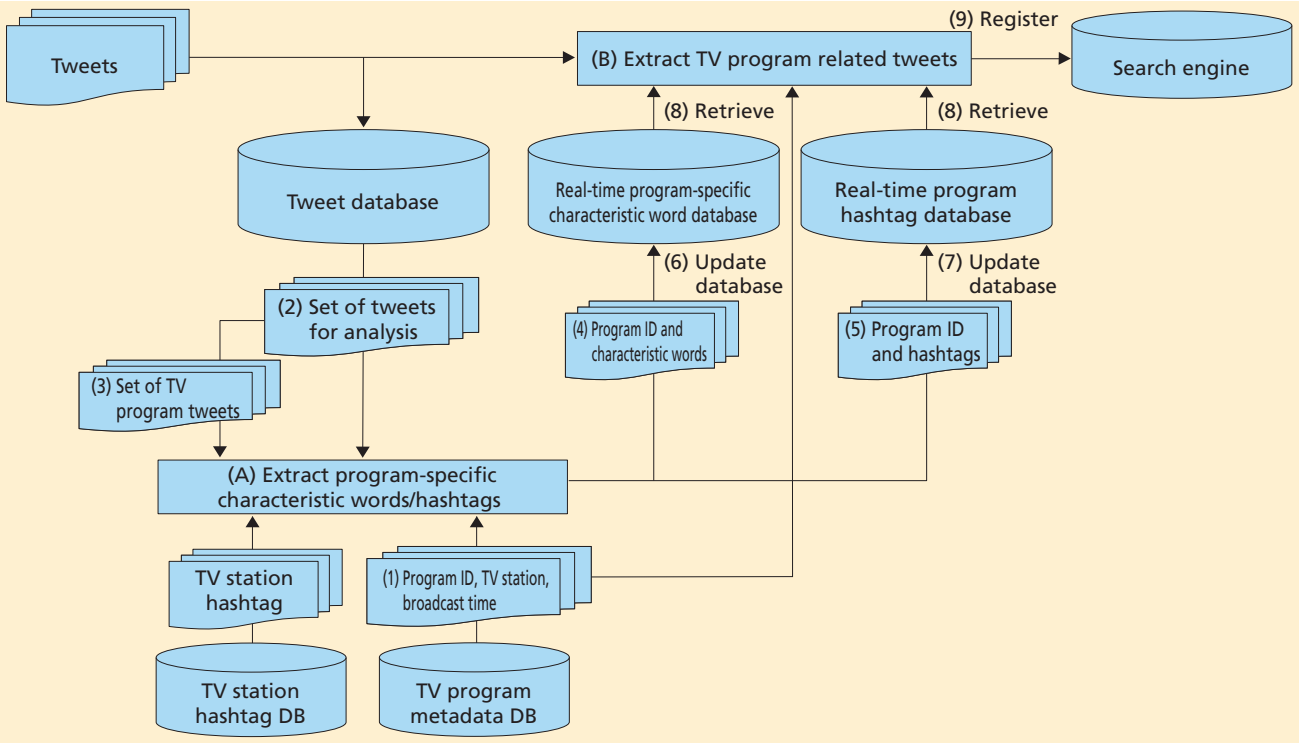


Figure 3 Real-time selection of TV-program-related tweets

unique TV program is associated with tweets that include TV station hashtags. Using the tweets extracted in (3), program-specific characteristic words are extracted for each TV program (Fig. 3(4)). Each characteristic word is assigned with the program ID. Characteristic words are usually informative enough to describe the TV program. Characteristic words for each hashtag are extracted from the tweet dataset in (2) (Fig. 3(5)). The characteristic words and hashtags are compared between TV programs. If the similarity score is above a certain threshold, then the hashtags are selected as hashtags for the TV program. The above processes for extracting program-specific characteristic words and program-specific hashtags are run in real time and kept up-to-date(Fig. 3(6)(7)).

The process for extracting TV tweets using the hashtags and characteristic words is summarized below. Tweets which include the program-specific characteristic words in (4) or the program-specific hashtags in (5) are assigned with corresponding TV program and program ID in real time (Fig. 3((8)). Finally, tweets in (8) are registered in the search engine (Fig. 3(9)).

4. Search for Tweets Related to Landmarks

4.1 Service Overview

Two types of service for searching tweets related to landmarks are provided: "Location-based tweet search" and "Trending Spot Ranking".

Location-based Tweet Search is a search service that allows the user to search tweets referencing nearby train stations and landmarks based on the user's current position (Fig. 1(c)). By selecting the "Search tweets near current location" link in the figure, the user can search nearby tweets. Using the real-time nature of Twitter, this service enables users to search "hot" information nearby, as it is happening. Users can also transition to the Trending Spot Ranking, as discussed below, from the search results.

"Trending Spot Ranking" is a service that provides per-area rankings for sight-seeing spots that are being mentioned on Twitter (Fig. 1(d)). Users can view rankings of sight-seeing spots near their current location or in an area that they select. In the details page for a landmark selected from the ranking page, trending tweets are selected from tweets for that landmark, and relevant keywords, and other basic information such as the address and business hours are displayed. The details pages are also linked to docomo map navi™*6 services such as the map application and Gotochi Guide (local guide application), to detailed views of tweets for the landmark, and to searches using the landmark name and related keywords.

4.2 Location-based Tweet Search System

To implement location-based tweet search, place names included in tweets must be extracted. However, when selecting keywords to express locations, names of places such as train stations and landmarks could also refer to other locations or non-geographical names (e.g. names of people), and this must be considered. For example, Maruyama Koen (Maruyama Park) is the name of two different parks, one in Sapporo and one in Kyoto. Another example is Matsushima. This is the place name of one of the famous "Three Views of Japan" in Miyagi Prefecture, but is also a family name. This sort of ambiguity in identifying places must be eliminated, and is handled in the tweet registration process described below.

The location-based tweet search system process for registering tweets in the search engine is shown in **Figure 4**. The names and locations of train stations and landmarks are maintained within the system as Point Of Interest (POI) data. The search engine decides whether a train station or landmark name is included in the tweet, creates an association between the tweet and the location, and registers it. To disambiguate the location name as discussed above, tweets are processed as described in existing research [1]. Specifically, a disambiguation is made based on the assumption that an ambiguous place name and the nearby place names tend to co-occur*7 in the same document (e.g.: Maruyama Koen and Kyoto occur in the same tweet). In

*6 **docomo map navi™**: "docomo map navi" and its service logo are trademarks of NTT DOCOMO Inc.

*7 **Co-occur**: When a particular word happens to appear with another particular word in the same document.

this way, nearby place names are associated with POI data and stored. This process, linking tweets to POI, is done in real time by the system when the tweet data is received.

Next, we describe the structure of the search system, shown in **Figure 5**. When a user searches tweets, the mobile terminal sends its location (latitude, longitude) to the server. Then, based on this information, a search for nearby tweets is done within a predetermined radius. At this stage, if a set minimum number of tweets are not found, the search radius is increased and the search is repeated. This ensures that the search result contains a set number of tweets. However if the radius frequently needs to be increased in this way, search response time can increase, so search radius values are pre-stored in the system as a basis for searching at the latitude and longitude of each landmark, to ensure that the set number of tweets can be obtained. The server receives the user's latitude and longitude, finds the landmark closest to the user, and sends a query to the search engine using the stored search radius. This enables high-speed searches guaranteeing a number of tweets, independent the location.

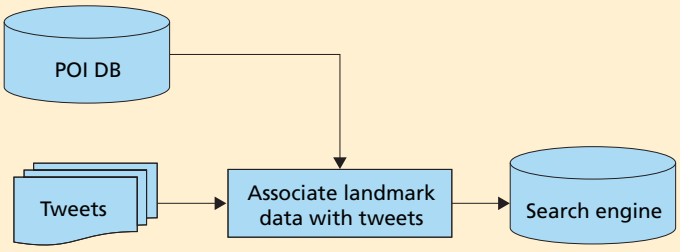


Figure 4 Tweet registration for location-based tweet search

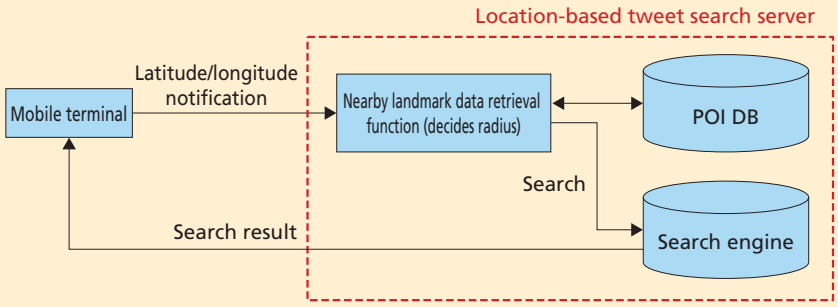


Figure 5 Location-based tweet search system structure

4.3 Processing Tweets for Trending Spot Ranking

Tweets related to landmarks are processed for the trending spot rankings as shown in **Figure 6**. As with the location-based tweet search system described in section 4.2, when initially associating landmark information with tweets, place names may also be used to mean different locations or to mean something other than a place name, so some ambiguity must be eliminated. Here as well, landmark information and tweets are associated in real time. Tweets are associated with landmarks, and the number of tweets for each landmark are counted. Landmarks are associated with the areas which are used by Gotochi Guide, and a per-area ranking

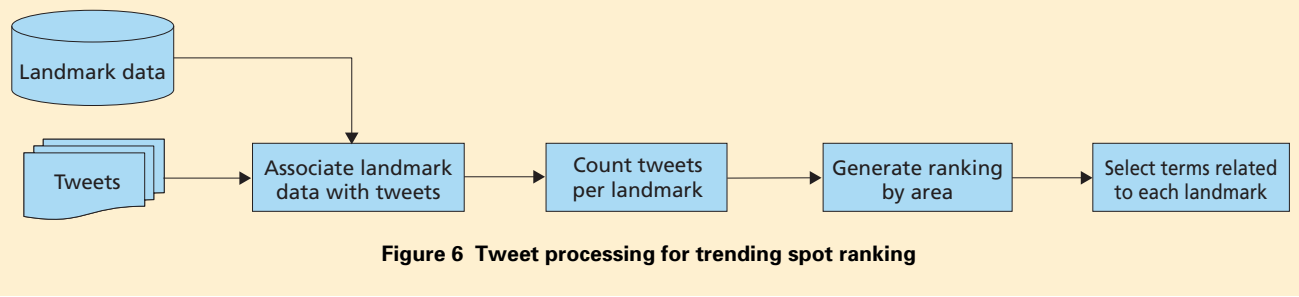


Figure 6 Tweet processing for trending spot ranking

is generated by sorting the landmarks in descending order of the number of tweets. Then, for the landmarks that rank high in each area, related words are selected from the landmark-related tweets, based on co-occurrences of the words with landmark names and on the rarity of the words themselves. Processing to generate per-area rankings and select related words is done on tweet data accumulated over regular intervals.

5. Conclusion

In this article, we have introduced real-time search services for tweets related to useful topics in daily life, especially corresponding to railway operational conditions, TV programs and geographic landmarks. We have also described the search technologies supporting these services. With this development, we have focused on search technology that can accurately select tweets relevant to the domains of each service, and on search interfaces that allow the desired information to be found easily. These technologies make our search services more convenient than a simple keyword search.

In the future, we will continue our research and development to improve the accuracy of our related-tweet selection technologies, and to provide convenient services that push time-sensitive information using real-time data.

REFERENCE

[1] E. Amitay, N. Har'El, R. Sivan and A. Soffer: "Web-a-Where: Geotagging Web Content," SIGIR '04 Proc. of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 273-280, 2004.