Search | Microblogging | Trend

Technology Reports

Real-time Tweet Search System

*The service provided by Twitter[*1] Inc. has achieved global recognition as a new form of media where users can publish instant updates (called "tweets") on their current circumstances. To further enhance the convenience of mobile devices, NTT DOCOMO released a tweet search service in August 2011. To make this service easy to use even by people who are unfamiliar with Twitter, we developed a mechanism that introduces tweets by mixing them into the ordinary search results when users search for celebrities or events, and a technique for analyzing trends on Twitter in real time. In this article, we introduce the tweet search service and the technologies that support it.*

Service & Solution Development Department
Daisuke Torii
Yasuhiro Yokoi
Yuji Mori
Smart Communication Service Department
Yuko Kon

1. Introduction

Twitter has achieved global recognition as a new medium where users can publish instant updates on their current circumstances, and NTT DOCOMO is working with Twitter Inc. to implement new services that further enhance the convenience of mobile devices.

Most users of social networking services (SNSs) such as Twitter and Facebook[*2] are highly computer-literate and accustomed to using online services.

However, NTT DOCOMO has a broad range of users and some users have little or no familiarity with SNS.

Therefore, we have designed our services so that even such users can enjoy Twitter content.

We have thus compiled a database of words associated with celebrities and events, and we have developed a search mechanism that displays related tweets with the regular search results for queries containing these words. Furthermore, to provide content that can be enjoyed more by users with an interest in tweets, we have also developed a real-time trend analysis technique to extract words, celebrities, images and hashtags[*3] that are currently topical among Japanese users.

In this article, we describe the tweet search technique and trend analysis technique, and we present the search service that was deployed in August 2011 to i-mode and smartphone users.

2. Tweet Retrieval Using Direct-box

2.1 Service Overview

The direct-box is the part of the screen that displays information suited to the user's needs — e.g., weather information, the meanings of words and movie information — that matches the keywords entered into the search box (**Figure 1**(a)) when searching using an i-mode terminal or smartphone.

In this tweet search system, tweets

*1 **Twitter**: A registered trademark of Twitter Inc. in the United States and other countries.
*2 **Facebook**: A registered trademark of Facebook, Inc.

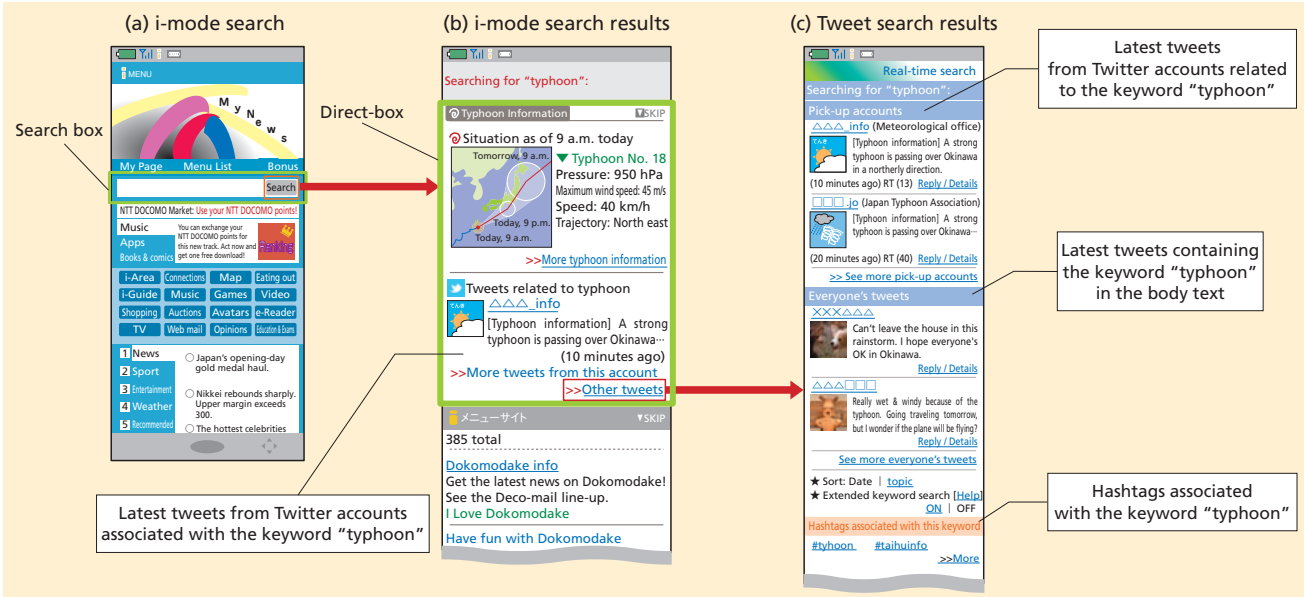

Figure 1 Overview of the tweet search service (i-mode version)

associated with keywords (celebrities or events) entered into the search box are displayed as content in the direct-box (fig. 1(b)). By introducing Twitter accounts associated with the search keywords, it becomes easy for users with no experience of Twitter to access real-time Twitter content in routine Web searches.

Furthermore, by clicking the "More" links inside the direct-box, users can browse tweets from Twitter accounts derived from the keywords (hereinafter referred to as "Pick-up accounts"), tweets containing the keywords in the body text (hereinafter referred to as "Everyone's tweets"), and hashtags associated with the keywords (fig. 1(c)).

These tweets can not only be displayed in the order in which they were submitted, but can also be ordered by popularity based on the amount of attention each tweet has received.

Also, since there are some tweets containing language that is unfamiliar in Japan and words that are unsuitable for minors (stopwords) and tweets containing character codes that cannot be displayed on terminal devices, these tweets are filtered from the search results.

2.2 Tweet Search System

Figure 2 shows an overview of the tweet search system. Since tweet searches require real-time performance unlike conventional search engines, tweets obtained from Twitter are registered to a tweet search server where they can be made searchable in as short a time as possible.

The process flow when the search engine receives a query is described below. A search query from a user terminal (fig. 2(1)) is first processed by a search receipt server (fig. 2(2)). To decide whether or not to display related tweets in the direct-box, this server uses a celebrity/event DB to judge the suitability of the search query (fig. 2(3)). If the query is judged to be suitable, a tweet search request is issued to the tweet search server (fig. 2(4)). In the tweet search server, tweets are retrieved based on the search request (fig. 2(5)). For example, if a search request contains the name of a celebrity, the latest tweets will be retrieved from the celebrity's Twitter account. Also, if there is a word that expresses an event — such as "typhoon", for example — then suitable tweets will be retrieved

*3 **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).

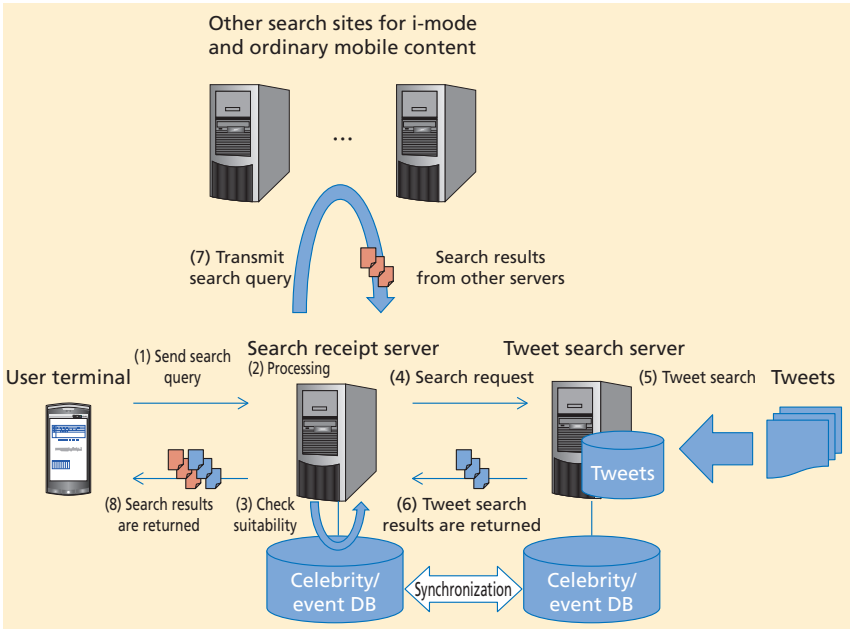

Figure 2 Overview of the tweet search system

from (multiple) accounts related to typhoons. The search results are returned to the search receipt server in a predetermined format (fig. 2(6)). On the other hand, when a search query from a user is also sent to another search server (fig. 2(7)), a search result screen is compiled together with the tweet search results (e.g., fig. 1(b)), and is sent back to the user terminal (fig. 2(8)).

When the user requests additional information by selecting an option such as "See more tweets" from the tweet search results displayed in the direct-box, or when a search is performed from the search box in the tweet search results, the abovementioned search receipt server is bypassed and the request is issued directly to the tweet search server. To display pick-up accounts for celebrities and events, the tweet search system has a celebrity/event database synchronized with the search receipt server. When a search request has been received, the system queries this database to decide whether or not to display pick-up accounts as the search results, and when pick-up accounts are found in the database, it searches the tweets of the corresponding accounts and displays them as pick-up accounts. Keyword searches are also performed in tweets of ordinary accounts in addition to pick-up accounts, and the search results are displayed as "Everyone's tweets" (e.g., fig. 1(c)).

3. Trend Search System

3.1 Service Overview

In addition to tweet searches, the system also provides trend searches (**Figure 3**(b)) whereby topics of discussion on Twitter are provided for the further enjoyment of users.

In the trend search site, the results of tweet analysis on Twitter are used to provide the following four functions:

- Popular tweet images

  Displays a collection of images that are attracting attention on Twitter. It is also possible to browse tweets relating to each image (fig. 3(c)).
- Popular celebrities

  Displays a ranking of celebrities that are attracting attention on Twitter. It is also possible to browse tweets relating to each celebrity (fig. 3(d)).
- Popular keywords

  Displays keywords that are attracting attention on Twitter (hereinafter referred to as "hot keywords"). It is also possible to browse tweets containing each hot keyword in the body text.
- Popular hashtags

  Displays hashtags that are attracting attention on Twitter. It is also possible to browse tweets containing each hashtag.

Since these hot keywords are appropriately applied to the abovementioned direct-box trigger conditions, related tweets are displayed as the search results when searching for hot keywords. This provides the user with fur-

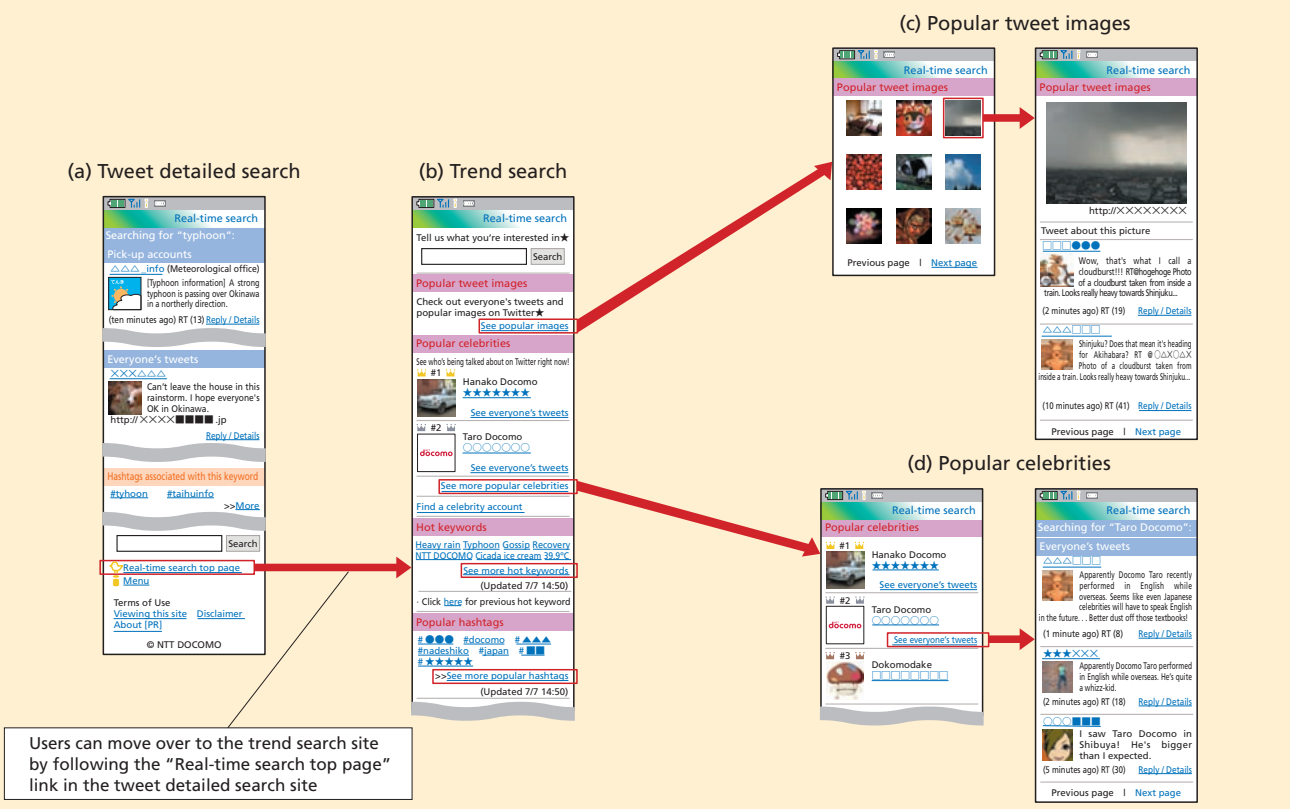

Figure 3 Overview of the trend search service

ther real-time information about what is happening in the world.

3.2 Trend Analysis System

Figure 4 shows a system configuration of the trend analysis system. To present information that is currently trending, tweet analysis is performed in real time.

The tweet analysis process flow is discussed below. The four functions introduced in Section 3.1 are described as a series of processes.

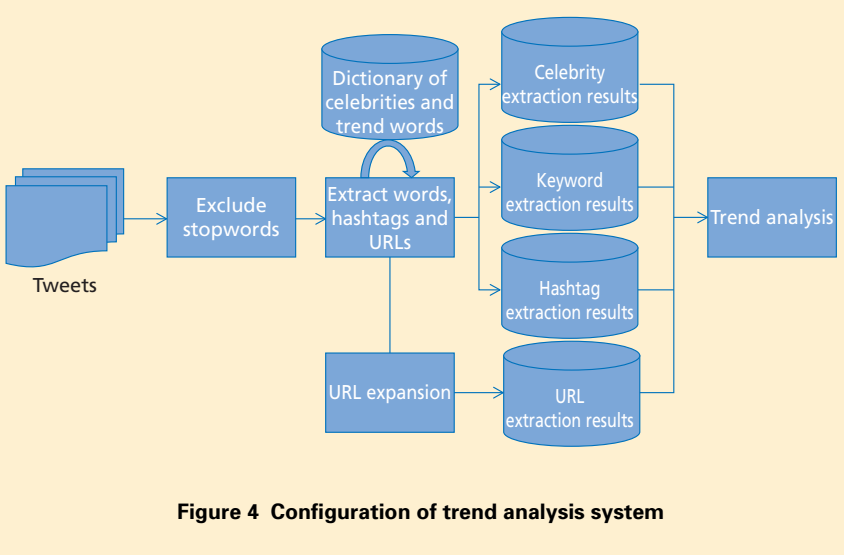

Figure 4 Configuration of trend analysis system

First, tweets containing stopwords that should not be shown to underage people are excluded from the scope of the analysis. Next, the words, hashtags and URLs are extracted from the tweets to be analyzed. This word extraction is performed using a prepared dictionary

of celebrities and trends designed for Japanese users. To ensure compatibility with the wide-ranging forms of expression used in tweets, this dictionary includes not only the official names of celebrities but also nicknames, and a collection of new words obtained by crawling the Web. Using the dictionary of celebrities and trend words, candidates for popular images, celebrities, keywords and hashtags are extracted via tweet analysis.

The extraction is performed according to the following process: (1) Popular celebrity candidates are stored in the dictionary as bundles comprising their official names and various other names to account for variations in how they are referred to by Twitter users. (2) Hashtags are extracted by extracting the text between a hash character ("#") and the following space, and no dictionary is used. (3) URLs are extracted by searching tweets for URLs that reference image sharing services. Since Twitter has a 140 character limit, URL shortening services are often used to make the URLs in tweets shorter. Since there are a number of different providers of URL shortening services, there may be more than one short URL pointing to the same image. Shortened URLs are therefore first expanded to the original URLs.

After the extraction, a trend score for each candidate is calculated from the temporal transitions of the term frequency. Trends are analyzed by periodic aggregation of the words, hashtags and URLs extracted at the abovementioned steps. This enables us to provide images, celebrity names, keywords and hashtags that are currently popular.

4. Conclusion

In this article, we have described the real-time tweet search service and system developed for NTT DOCOMO users. To facilitate use of the service by people who are unfamiliar with Twitter, we have developed a system that features tweets in the ordinary search results when users search for celebrities or events, and a real-time trend analysis technique for Twitter. Our service has already been released to i-mode and smartphone users, allowing it to be put to practical use.

In the future, by linking up with location information and other forms of media, we will develop new services and technologies that enable the instant delivery of up-to-date information in a way that is easily understood, even by users who are unfamiliar with Twitter.