Palanteer: A Search Engine for Community Generated Microblogging Data

Ee-Peng Lim and Palakorn Achananuparp

Singapore Management University, Singapore
{eplim,palakorna}@smu.edu.sg

Abstract. Unlike standard web search, people search microblog messages to look for temporally relevant information. Due to the recency nature of microblogs and a massive amount of data generated by users of popular services such as Twitter, it is challenging to design and implement a microblog retrieval system that satisfies the searcher and technical requirements. In this paper, we present a microblog search engine called Palanteer. Palanteer utilizes a unique framework for gathering and searching microblog data by focusing on harvesting community-relevant content. Next, Palanteer uses a timeline-based interface and a word cloud visualization to enable the searchers to explore and make sense of temporally-relevant information. The customizable framework can be used to create search engines for different community of users and microblogging sites.

Keywords: Information retrieval, microblog, twitter, search engine.

1 Introduction

1.1 Motivation

Microblogging sites such as Twitter, Weibo, and Plurk have gained huge popularity in recent years due to their success in bringing information and social networks together. Today, microblogging users include individuals, news media companies, celebrities, politicians, and others. Users can gain access to interesting information in realtime by subscribing to or *following* others' message feeds. Users also depend on a microblogging platform to establish social links with others who may be their offline friends, colleagues, customers, or family members.

In this paper, we will focus mainly on Twitter which is arguably the best known microblogging site. As of March 2012, Twitter has more than 140 million active users and its users generate 340 million messages known as tweets each day[1]. With massive number of messages generated in realtime, it becomes a challenge to effectively search for content in Twitter. Instead of creating a search engine for the entire Twitter network, we therefore decide to develop a search

[1] http://thenextweb.com/socialmedia/2012/03/21/twitter-has-over-140-million-active-users-sending-over-340-million-tweets-a-day/

H.-H. Chen and G. Chowdhury (Eds.): ICADL 2012, LNCS 7634, pp. 239–248, 2012.

engine for user communities that provide a well-defined scope for gathering and searching Twitter data. This search engine is known as **Palanteer**[2].

Even as Palanteer is designed to search only a community's Twitter data, the data volume involved can still be enormous. For instance, from a sample of more than 150K Twitter users located in Singapore, we observed more than half a million tweets getting generated each day and 15 millions tweets each month. Other than coping with the sheer volume of Twitter data, the other challenges to be addressed by Palanteer include: (a) selecting the community relevant content; (b) organizing stream data for easy searching; (c) reducing the impact of noise data in searching; and (d) summarizing the search results for browsing and viewing. Challenge (c) also exists for web search engines. The popular web search engine Google introduces Pagerank as a way to select important web pages as results. Due to limited message content and realtime nature of Twitter data, a different approach is required. Challenges (a), (b), and (d) are unique to Palanteer's design and need to be addressed in this research.

1.2 Overview of Palanteer

Palanteer addresses the aforementioned challenges at both the system and interface design levels. The design goal is to have a system that can be adapted to different user communities while scaling with time. A user community here could be networks of users sharing some common attribute(s) such as location or topics of interests. From our experience, the Palanteer's system design has been shown to work for a number of user communities, including a community of Singapore users and a community of Thailand users. The design also works for different microblogging sites as demonstrated by the Palanteer system developed for Taiwan users using Plurk[3], one of the most popular microblogging sites in Taiwan.

At the interface design level, we create a search interface that is appropriate for searching text stream data for a number of search scenarios, including topic and entity search. The search interface design handles noisy data by presenting the information aggregated from individual users and supporting the navigation at the aggregated level. Similarly, search results can be presented as line charts and word clouds to simplify browsing.

2 Related Work

Due to its vast and diverse user base, Twitter is one of the most popular microblogging services studied by researchers in several disciplines. Early work has explored the properties of social network in Twitter [7,5] and found that Twitter users microblog about their daily life as well as sharing information. Jansen et al.[6] investigated Twitter as an electronic word-of-mouth for sharing brands'

[2] `http://palanteer.sis.smu.edu.sg`
[3] `www.plurk.com`

opinions and sentiments among consumers. Next, several researchers have studied the behavior of Twitter users, e.g., comparing Twitter with traditional news media [8,19], identifying influential Twitter users [18,3,13], etc. Social signals generated by Twitter users have been used in detecting earthquakes [14] and news events [17].

Information retrieval research has examined the information needs of Twitter users [12,16]. Teevan et al.[16] analyzed millions of Twitter and web search queries and found that, unlike web search which is more fact-based, people use Twitter to look for temporally relevant information, such as breaking news and popular trends, and information about people. Not surprisingly, queries related to celebrities are overwhelmingly more frequent in Twitter than web searches. Palanteer is specifically designed to facilitate the consumption of temporal information and well-known named entity queries (e.g. celebrities, leaders, businesses, places, sports, etc.).

Many academic [4,10,15,2,9] and non-academic[4] systems have explored several user interface designs to aggregate, summarize, and visualize the microblogging data. Most of them use a timeline-based visualization to display temporally relevant information. Palanteer follows the same design principle of those prior work by utilizing timeline to summarize the Twitter search results. Furthermore, Palanteer's user interface encourages users to navigate the search results' context through an interactive word cloud. For the users with unspecified information needs, Palanteer also provides a trending topic browsing interface to summarize recently popular topics across multiple categories. Section 4 describes the user interface design in detail. Lastly, unlike other systems, Palanteer filters the Twitter stream by identifying community relevant messages (see Section 3).

3 Crawling User Community Relevant Data

The first challenge of designing Palanteer is to identify members of the target user community so as to crawl the Twitter data generated by its members. Being user community specific, Palanteer can offer a well defined scope for searching information that is particularly useful in business and social search scenarios. This also helps to reduce the amount of data to be handled by the search engine.

Community detection has been an active research topic in network science[11]. This thread of research however focuses on detecting all communities within a network of users, as opposed to finding members of a target user community for search purposes. We hence define a user community to be a group of users meeting three main general criteria below:

- Members of the community share some common attributes such as location and topics of interest which could possibly be identified by some target keywords in the user profile description;
- Members have publicly accessible Twitter data; and
- Members of the community are connected with one another.

[4] `http://archivist.visitmix.com/`

Our approach to constructing a target user community starts from some seed user(s) who satisfy the attribute criteria before crawling their followers and followees so as to identify their communities. We hereby assume that it is easy to get these seed users as there are well known members of the target community(ies). For example, the seed users of the Singapore community can be Singapore political leaders, celebrities, local news media, etc. These users can be easily discovered via Twitter directories such as wefollow.com[5]. If we crawl the one-hop neighbors of the seed users who satisfy the attribute criteria, we will obtain the communities which are the ego networks of the seed users. We can repeat the crawl to get the two-hop members and beyond. By expanding the target user community from the seed users, we can make sure that there are no disconnected users in the community. In practice, we found that two-hop members from seed users can allow us to construct sufficiently useful user communities. This is not a surprise given that online social networks are known to have smaller degree of separation[1].

Algorithm 1 depicts the above crawling strategy. The strategy terminates when the crawl reaches sufficient number of hops (denoted by K) or when there are no more new users added to the user community (denoted by U). Once we have obtained the members of U, Palanteer will start crawling the tweets generated by Twitter users at regular intervals to obtain the latest data before these data are no longer available via Twitter APIs. The choice of interval size depends on a number of factors including timeliness of data in search engine, data limits of Twitter APIs, the number of users, and the rate of data generated by these users.

Algorithm 1. User Community Crawling Strategy

Input: S (seed users), K (number of hops required)
Output: U (members of the target user community)

1: $U \leftarrow S$
2: $F \leftarrow S$
3: $\#hop \leftarrow 0$
4: **repeat**
5: $F' \leftarrow \varnothing$
6: **for each** $u \in F$ **do**
7: Crawl the followers and followees of u and add them to F'
8: Remove from F' any members who do not meet the attribute criteria or do not open their tweet messages to others
9: **end for**
10: $F \leftarrow F' - F$
11: $U \leftarrow U \cup F$
12: $\#hop \leftarrow \#hop + 1$
13: **until** $\#hop = K$ or F is empty

[5] `http://wefollow.com`

Similar to web search engines, Palanteer has to keep up with dynamic changes to the Twitter network. New users may join the user community. Users may update their profile description but we expect this should not happen often. Such updates may potentially remove them from the user community. Users also update their follow links resulting in changes to the composition of user communities. To cope with these changes, we have to regularly perform Algorithm 1 using the same seed users on a daily basis.

4 User Interface Design

There are not many search engines designed for text stream data such as Twitter. Palanteer adopts a unique design that is based on the following design assumptions which are closely related to the challenges of coping with noise data, organizing and navigating stream data.

- *Search users are interested in recent data*: Every tweet message is assigned a timestamp. Search users' attention of older tweets decreases over time as more recent events emerge. Given the limited attention resource each search user has, it is reasonable to expect him or her to focus largely on the recent data. This is consistent with the way users currently consume tweets in reverse chronological order.
- *Search users require contextual guidance to find information*: Regular search users often formulate sparse and short queries, often times due to their ambiguous information needs. To help them refine their queries and navigate the search results, certain contextual guidance shall be provided.
- *Search users only need to make sense of the summarized results*: Due to a large volume of Twitter data and an extremely short and noisy textual content of tweets, standard web search's ranked list may not be ideal for displaying microblog data. To overcome information overload problem, the search results shall be presented in a summarized form so as to help the users easily make sense of the results as a whole.

With the above assumptions, we now describe the design of the Palanteer's search interface in the following subsections.

4.1 Trending Item Directory

Palanteer displays a trending item directory at its homepage to list different types of items that may be interesting to the users. In the context of political user community, search users may be interested in politician names, political party names, and voting constituencies. In the context of popular-music user community, search users may be interested in artist names, their managing agents, CD albums, etc. Instead of deciding on the search keywords, a search user can simply select a trending item for query. This is consistent with our first design assumption that most users care about recent data.

The trending item directory is an approach to give contextual guidance to search users. Items vary their popularity over time as new tweet messages are generated. The types of items included in the directory are different depending on the community's topics of interest. For example, the Singapore community's trending item categories are overseas and local celebrities, sport related items (players, teams, etc.), businesses, places of interest, and public services. Visually, each item is shown in a specific font size to indicate its aggregated frequency in the tweet messages at a specific time window (by default 24 hours). The larger the font size, the more frequent is the item. Showing the currently popular ones at the homepage provides a useful cue for the search users to discover potentially interesting items to be explored. Instead of 24 hours, the search users can select longer duration (e.g., last one week, last two weeks, etc.) recent data to derive the trending items. For those with a specific information need, they can directly submit their queries through the search box.

At present, the directory of items has to be manually created by Panlanteer's operator. The future plan is to automatically determine the popular items mentioned by the user community and classify them into different types.

4.2 Result Summarization and Navigation

Every query, either fired from a trending item or a user-supplied query, is processed and its results are shown in a summarized form to facilitate navigation. In Palanteer, we use a combination of volume line chart and word cloud to summarize the results.

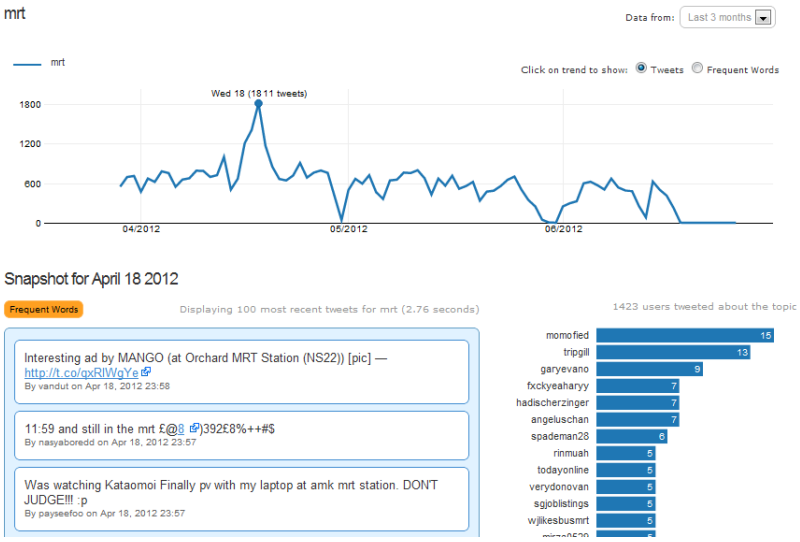

Fig. 1. The "mrt" Query Example

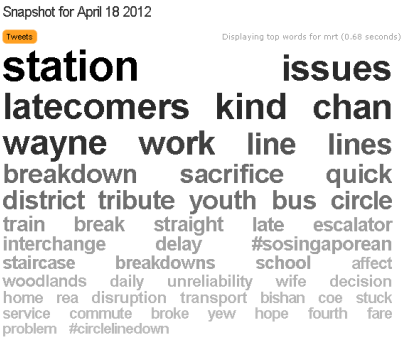

Fig. 2. Word Cloud of Query "mrt"

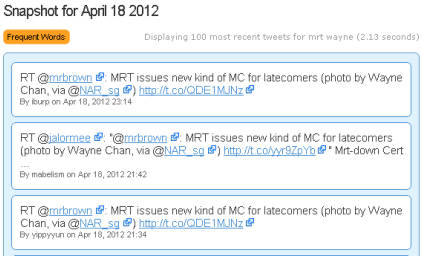

Fig. 3. Some Results of Query "mrt"+"wayne"

Figure 1 depicts how the results of the query keyword "mrt" (name of a mass transit system in Singapore) are summarized in the form of volume line chart over a three-month time period. The line chart gives the search user a context of how popular the query is over time. The dominating low volume data can be treated as white noise that does not receive much attention from the users in the community.

Instead of showing all tweets, Palanteer returns a sample of 100 most recent tweets whenever a user searches a keyword at a specific time point. The small sample size is intended to keep the overhead of reading the tweets to a manageable level. This is consistent with our design assumption of not overloading users with excessive information. On the right of the retrieved tweet listing is a list of users contributing to these tweets along with the number of tweets they contributed. For example, the users "momofied" and "tripgill" contributed 15 and 13 result tweets, respectively. Most of the remaining users contributed only 1 to 9 result tweets each. This helps the search user quickly identify the Twitter users who are more interested in the "mrt" topic.

To provide the contextual knowledge for the above truncated results, a word cloud is used to summarize the words that appear in the entire search results. Figure 2 shows a word cloud visualization of the "mrt" results on April 18, 2012. The font sizes here are also indicative of frequencies of the words used in the results. Again, this helps us to reduce noise and information overload. The word cloud also guides the search users to navigate the data further. For example, Figure 3 shows that some mrt-related event has taken place on April 18. The keywords "station", "issues", "latecomers", etc., emerged as important words in the results.

5 Use Case Examples

In this section, we use a few use case examples to illustrate the usefulness of Palanteer to find interesting tweet content and trends.

5.1 Topic Search

Consider the previous example of the "mrt" query in Figure 1 as a topic search example. If we further navigate the topic search results further, we can quickly drill down into smaller sub-results that give further detailed information about the topic. For example, among the frequent keywords is someone called "wayne" who is intuitively not related to the subway. If one selects "wayne" in the word cloud, we can further refine the query into "mrt"+"wayne" and obtain the results containing both keywords as shown in Figure 3. As the results show, there are many tweet messages mentioning the phrase: "MRT issues new kind of MC for latecomers (photo by Wayne Chan, via @NAR_sg)"[6]. This suggests that there is a viral message that pokes fun at the subway breakdown causing inconvenience to passengers.

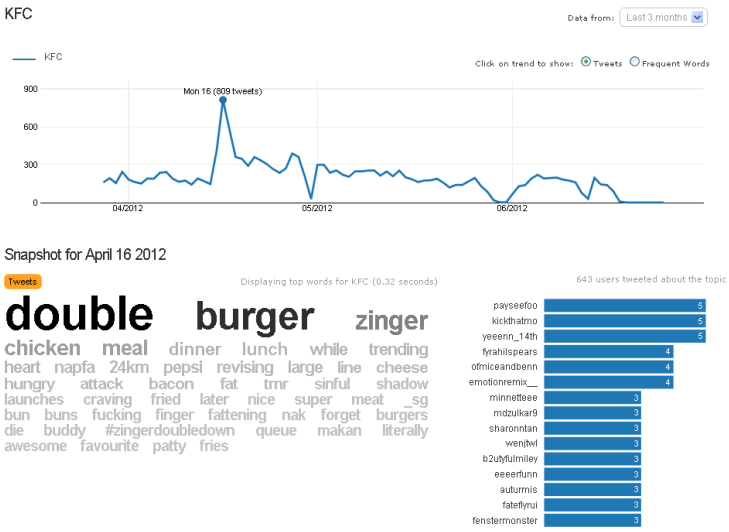

Fig. 4. Results of Query "kfc"

5.2 Entity Search

Figure 4 shows a summary of results from a query involving a fast food chain "KFC" over a three month period. The volume line chart shows a spike on April 16, 2012 corresponding to KFC launching its new "Double Down" burger in Singapore. Many users gave their comments over Twitter about the new product. The search results also show that KFC has enjoyed a uniform attention from the users over the three months. Note that a few dips in volume at some time points may be caused by data gathering problems.

[6] MC is an abbreviation of medical certificate.

Additionally, one can perform a comparative entity search by supplying two named entities in a query. Figure 5 shows a query involving "KFC" and another fast food chain "McDonald". The results show that KFC and McDonald enjoyed similar popularity during a three month period. Compared with KFC, McDonald did not have any obvious spike. This suggests that McDonald has not introduced any new product or promotion activity in the recent past.

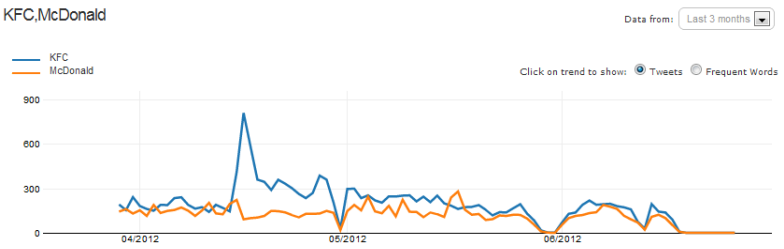

Fig. 5. Results of Queries "kfc" and "mcdonald"

6 Conclusion

The recency nature of microblogging data has motivated our research to design a new search engine called Palanteer. Palanteer adopts a framework that supports search on user community generated microblogging data. The search interface design places a heavy emphasis on the importance of recent data and provides a summarized view of the search results for quick interpretation and navigation.

This search framework is generic and can be used to create search engines for different microblogging sites. For example, apart from the Singapore community edition, we have also developed other editions of Palanteer including Thailand, software development, and Taiwan communities. Each edition is created using the same framework in the original Singapore edition. In addition, the framework is flexible enough such that specific modules, e.g., data crawlers, language-dependent word segmentation component, etc., are customizable.

For the future work, we would like to continue applying the search framework to create more search engines for other communities of microblogging users. We also plan to conduct useability study on Palanteer to identify areas for improvement. There are also many other research topics such as automated extraction of items and assignment of item type labels, collaborative search, and better visualization of search results that can be studied.

Acknowledgement. This work is supported by Singapore's National Research Foundation's research grant, NRF2008IDM-IDM004-036.

References

1. Ahn, Y.Y., Han, S., Kwak, H., Moon, S., Jeong, H.: Analysis of topological characteristics of huge online social networking services. In: Proceedings of WWW 2007, p. 835. ACM Press, New York (2007)
2. Bernstein, M.S., Suh, B., Hong, L., Chen, J., Kairam, S., Chi, E.H.: Eddi: Interactive Topic-based Browsing of Social Status Streams. In: Proceedings of UIST 2010, p. 303. ACM Press, New York (2010)
3. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.P.: Measuring User Influence in Twitter: The Million Follower Fallacy. In: Proceedings of ICWSM 2010, Washington DC, USA (May 2010)
4. Diakopoulos, N.A., Naaman, M., Kivran-Swaine, F.: Diamonds in the Rough: Social Media Visual Analytics for Journalistic Inquiry. In: Proceedings of VAST 2010 (2010)
5. Huberman, B.A., Romero, D.M., Wu, F.: Social networks that matter: Twitter under the microscope. First Monday 14(1) (2009)
6. Jansen, B.J., Zhang, M., Sobel, K., Chowdury, A.: Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology 60(11), 2169–2188 (2009)
7. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding microblogging usage and communities. In: Proceedings of WebKDD/SNA-KDD 2007, pp. 56–65. ACM Press, New York (2007)
8. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: Proceedings of WWW 2010 (2010)
9. Marcus, A., Bernstein, M.S., Badar, O., Karger, D.R., Madden, S., Miller, R.C.: Twitinfo: Aggregating and Visualizing Microblogs for Event Exploration. In: Proceedings of CHI 2011, p. 227. ACM Press, New York (2011)
10. Mathioudakis, M., Koudas, N.: TwitterMonitor: Trend Detection over the Twitter Stream. In: Proceedings of SIGMOD 2010, p. 1155. ACM Press, New York (2010)
11. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E 69(2), 1–16 (2004)
12. Ramage, D., Dumais, S., Liebling, D.: Characterizing Microblogs with Topic Models. In: Proceedings of ICWSM 2010. The AAAI Press, Menlo Park (2010)
13. Romero, D.M., Galuba, W., Asur, S., Huberman, B.A.: Influence and Passivity in Social Media. In: Proceedings of WWW 2011 (2011)
14. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of WWW 2010 (2010)
15. Shamma, D., Kennedy, L., Churchill, E.: Tweetgeist: Can the Twitter Timeline Reveal the Structure of Broadcast Events? In: Proceedings of CSCW 2010 (2010)
16. Teevan, J., Ramage, D., Morris, M.R.: #TwitterSearch. In: Proceedings of WSDM 2011, p. 35. ACM Press, New York (2011)
17. Weng, J., Lee, B.S.: Event Detection in Twitter. In: Proceedings of ICWSM 2011 (2011)
18. Weng, J., Lim, E.P., Jiang, J., He, Q.: TwitterRank: finding topic-sensitive influential twitterers. In: Proceedings of WSDM 2010, pp. 261–270 (2010)
19. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.P., Yan, H., Li, X.: Comparing Twitter and Traditional Media Using Topic Models. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 338–349. Springer, Heidelberg (2011)