Using Twitter Data to Improve News Results on Search Engine

Abraham Krisnanda Santoso
Informatics / Computer Science
Institut Teknologi Bandung
Bandung, Indonesia
abraham.krisnanda@gmail.com

Gusti Ayu Putri Saptawati
Informatics / Computer Science
Institut Teknologi Bandung
Bandung, Indonesia
putri@informatika.org

Abstract— **Web search results often consist of web results and news results. News results are displayed based on measurement of click-through Rate (CTR), a comparison between clicks and views of content. In other words, the more clicks obtained by content, the probability of that content appear, as news results will be higher. The CTR measurement is not effective for recent news due to recent news only have few clicks. On the other hand, a micro-blogging platform Twitter has short, real-time, wide coverage of news. In this paper, we use Twitter data to improve news results, so the recent news can have higher probability to appear on news results.**

Keywords—news results; CTR; clicks; Twitter; recent news

I. INTRODUCTION

An effective portal web search engine should be able to fulfill many type queries, such as 'recency sensitive queries', which refers to queries where documents are expected to be relevant and fresh [1]. Due to high demand on online news searching, many of major portal web search offer dedicated news tab [2,3]. The example of this web search results is shown in Figure 1 by which news results were displayed along with the web results, However, most of search engines still encounter two significant problems. First problem is due to crawler limitation and politeness policies, so web crawler is not allowed to fetch more than one page at a time from a particular web server. Consequently the news results couldn't be fresh, although relevant. The second is the fact that many features for document ranking such as PageRank, aggregate click, popularity, could fail in representing fresh documents. This failure is because fresh documents may have very few links and clicks [1].

Nowadays, Twitter is very popular means of communication, especially for sharing real-world events starting from widely known events (e.g. president election, popular artist concert) to small or local events (e.g. accident, heavy rains, earthquake) [1,2], reflecting these events as happen. For this reason, the content of Twitter is potential for real-time detection of real-world events, including early detection of interest in these events [1,2]. Several research have studied specific type of event identification in Twitter such as news event [4], earthquake, [5]. Recently, a research has focused on improving ranking fresh URLs on search engine [1].

Based on those previous researches, we conclude that Twitter data is particularly useful as source content of recent news. In this paper, we propose a method to use Twitter Data as source of recent news and ranking function to improve performance of search engine to provide the most recent news. To do so, our work focuses only on displaying the most recent news, assuming that search engine has the ability to detect query intent.

This paper is organized as follows: in Section 2 we describe related relevant works in Ranking Process of (i) search engine, (ii) news ranking, (iii) CTR prediction, and (iv) Twitter. Moreover, section 3 describes several researches, which have studied Twitter data to obtain breaking news from Twitter users. We then discuss our proposed approach in Section 4, and system development in Section 5. Section 6 discusses our experiment while section 7 concludes, our work.

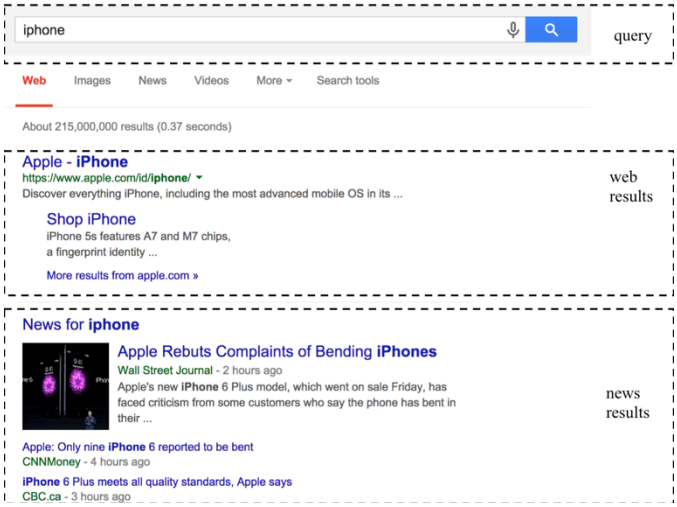

Figure 1: Web results and news results

978-1-4799-7996-7/14/$31.00 ©2014 IEEE

II. Related Works in Ranking Process of Search Engine

A. Ranking Process

Search engine is intended to return relevant documents / results related to user query. By which relevance means topical relevance and user relevance [6]. Topical relevance is when the document results are on the same topic with query. User relevance means the document results are recent, or written in the same language where the user comes from. To return the most relevant documents to its users, search engine needs a ranking process.

Ranking process is consists of five aspects, documents, features, queries, retrieval functions and scores [6]. The search results are ranked based on its score. It is calculated by ranking function R, which is based on features similarity, as in

$$R(Q,D) = \sum_i g_i(Q) f_i(D) \quad (1)$$

Where $g_i(Q)$ gives scores for every feature from query Q and $f_i(D)$ gives scores for every feature from document. So, document score is the result of multiplying both features scores from query and document. Figure 2 describes the calculation ranking function calculation.

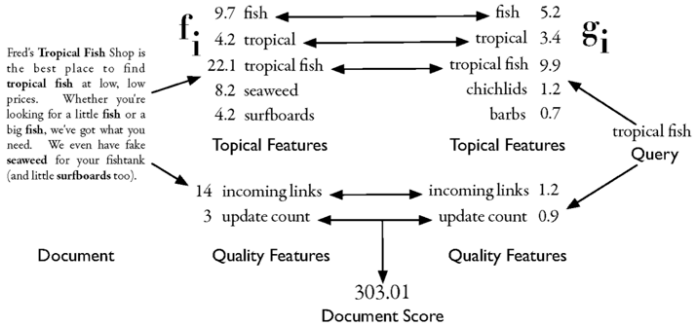

Figure 2: Ranking Function Calculation [6]

The ranking process starts its work by extracting and weighting every related keyword / index term from document and query. Besides, the ranking process also weights quality features, such as how many incoming links to the document and how many updates since first published. Then,, based on features' similarity the search engine would return the most topical relevant documents..

B. Estimation of Click-through Rate (CTR)

News results are consisted of newsworthy documents, which are not only topical relevant, but also user relevant and recent. [2] studied that there is a strong correlation between newsworthiness of contents and its click-through rate. For this reason, it click-through-rate (CTR) is one measure of the relevance to the search query.

Click-through rate (CTR) is the ratio of click to display on document, like ads or news results. In other words, the more clicks obtained by a document, the CTR of that document become higher. Search engine displays news results, which has the highest CTR. Unless the recent news has appeared on web results before, there is no data to be measured for recent news. To solve CTR measurement problem on recent news, several research have focused on predicting CTR score for recent news by using trend of query [1] and trend of news topic [3]. [1] research predicts the CTR score based on intensity of query submitted by users. [3] research predicts CTR based on three corpora, i.e. (i) news articles, (ii) blog posts, and (iii) Wikipedia. It detects 'spike' or new trend by monitoring blog posts, and then searches the news in news articles. Wikipedia is used for comparison corpus, means if the query terms are salient in Wikipedia but not salient in other corpora, then the query is more likely a general information query, instead of news-related query.

C. Improving Recency Ranking using Twitter Data

Dong *et al*. use Twitter Data to improve ranking for recent documents in search engine [1]. They crawl URLs from *tweets* and its features to compute score for *recent* / *fresh* documents with impoverished link and click information. They use Twitter Data just for ranking of the web results. The result shows the improvement both relevance-based and fresh-based metrics for recent queries.

III. Related Works in Using Twitter as News Media

Twitter could be described as technology that is likely to capture and transmit the sum total of all human experiences of the moment [7]. Twitter is described as a medium of information diffusion [8] and Twitter users could influence others to spread valuable tweets [9]. Empirical research shows that users retweet tweets because the event is important or interesting or valuable to share [8]. Twitter users more likely talk about topic from headline news and responds to fresh news [8]. Moreover, Twitter also covers news more than newswire and in some topics, like sports, natural disasters, business, entertainment, and technology. Consequently, Twitter could provide fresh reports/news before newswire [10].

IV. The Proposed Method

We propose a method to crawl content rapidly and rank content with impoverished link by using Twitter Data. The difference between our method and [1] is the document collection / corpus of the search engine. Instead of using the document collection of search engine, we use Twitter Data as source of news content.

A. Twitter Data as Document Collection

Displaying recent news as news results bears two main challenges. First challenge is about crawler limitation because of politeness policies. To overcome this challenge, search engines should make more web crawler to crawl web pages/ This solution will be limited to "robots.txt" file. Therefore, tweet from Twitter users is potential source of news content since it is not limited by politeness policies, real-time, and has coverage as wide as web pages.

Twitter with its wide coverage, real-time stream of news, and represents human social network has potential for being solutions as described in section above. Gathering data from Twitter is not limited by politeness policies, real-time, and has coverage as wide as web page. Even Twitter has some limitations on its API per September 2014, like total API calling per hour, it's sufficient enough to crawl most of tweets.

Most of newswire has official accounts which usually posts tweets at the same time with news update on its website.

B. Twitter Data as Ranking Function

Twitter has short, real-time, and wide coverage of news. When the news is published on a Twitter account, in the same time all of his / her followers can see the tweet. If the followers think the tweet is worth to share, they will share it by using retweet to their followers. Since the number of readers of the original tweet could grow exponentially in the short time, this implies Twitter could gain feedbacks faster than search engine to gather feedbacks for CTR score.

CTR scoring also takes time to gather feedbacks based on user clicks on search engine. Some research focused on prediction of CTR score of some contents like in [1,3]. However, CTR prediction could still severe risk if the search engines fail to predict CTR score, *zero recall problem* [1], which is the content will never appear on news results. To overcome this problem, we propose using Twitter retweet instead of CTR as one measure of the recency of search query. This is based on analysis that user query for news is considered as information retrieval process [11]. Therefore, when user find the relevant information, he / she will click the result, and the CTR score for that content will increase. On the other side, Twitter user who find a valuable tweet, she/he will do retweet, so the number of retweet would increase as more user retweet the valuable tweet. Hence, we conclude that retweet is equivalent to user clik. Figure 3 describes the equivalence between user click and user retweet. In our work, we use only credible and popular Indonesia newswire accounts, since they have an enormous amount of followers. As a result, it is easy and fast to collect feedback (retweet).

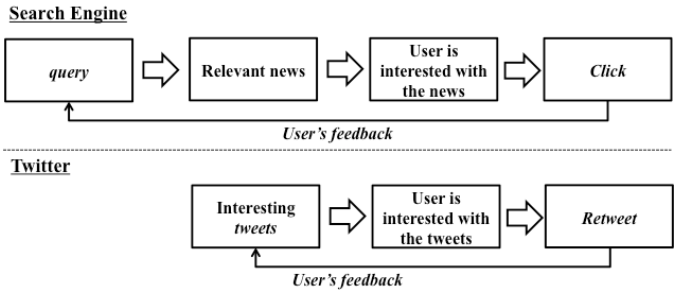

Figure 3: The equivalence between user clicks and retweets

V. System Development

Search engine architecture is divided into two processes, indexing process and query process [6]. Indexing process is building structure which make searching possible on search engine. Query process is using the structure that has been built on indexing process.

A. System Architecture of Indexing Process

The proposed method was implemented by using Apache Solr. Apache Solr is open-source enterprise search engine. We use Apache Solr because Apache Solr has provided most of core components we need to do the experiment as we focus on text acquisition and ranking process. Moreover, since Apache Solr does not provide crawler, we need to develop a crawler fetch data from Twitter API GET statuses/user_timeline. The crawler fetch data only from Indonesian newswire most popular and credible accounts (@detikcom, @kompascom, @korantempo, @Metro_TV, @RadioElshinta, @SINDOnews, @okezonenews, @republikaonline, @Beritasatu) every 2 minutes. We choose those accounts due to some reasons, i.e. (i) they have a lot of followers, (ii) they tweet frequently, and (iii) their followers tend to retweet their tweets quickly, so we can get valid feedbacks from their followers.

Only some portion metadata of those data is stored on MongoDB, because of not all metadata is useful for this case. Table 1 shows the detailed stored metadata. All of stored metadata were loaded to Solr using Solr's REST-like Web Services. Then, Solr managed the data automatically. Figure 4 summarize the illustration of system architecture.

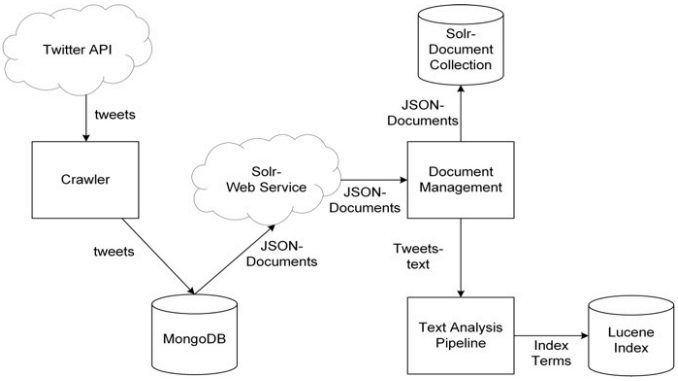

Figure 4: System Architecture

TABLE I
STORED METADATA

Metadata	Description
created_at	UTC time when tweet was released.
id	Unique identifier for tweet.
retweet_count	Number of retweet.
screen_name	Author of tweet.
text	Content of tweet.
link	URL linked to news article.

B. System Architecture of Query Process

To modify Solr's ranking function, we use Solr's *function query*. Our *function query* is shown in Figure 5 and the explanation of our *function query* is shown in Table 2.

```
q=_query_:"{!func}scale(query($keywords),0
,100)" AND
_query_:"{!func}recip(ms(NOW/HOUR,created_
at),3.16e-11,1,1)"
AND
_query_:"{!func}scale(retweet_count,0,100)
"
```

Figure 5: Our *Function Query*

TABLE II
EXPLANATION OF FUNCTION QUERY

Part of *Function Query*	**Description**
`_query_: "{!func}scale (query($keywords),0,100)"`	Weighting based on textual keyword relevance (tf-idf) with scale from 0 – 100.
`_query_: "scale({!func} recip(ms(NOW/HOUR, created_at),3.16e-11,1,1),0,100)"`	Weighting based on calculation with recip function with scale 0 – 100 on time when tweet was released (created_at). The more recent the document, the weight will be closer to 100.
`_query_: "{!func}scale (retweet_count,0,100)"`	Weighting based on retweet score with scale 0 – 100.

VI. EXPERIMENT AND RESULTS

To evaluate the effectiveness of our ranking process, we conducted user test [12], instead of using test collection [13] to measure the recency of news results. Specifically, we applied *convenience-sampling* technique to choose participants [12]. Each participant was provided 10 most recent and popular news on a specific date. The news was taken from most popular national newswire, Detik.com and Kompas.com at (17:57). Before the experiment started, all participants read the news and assessed them (in scale 1 (unfamiliar) to 5 (very familiar)) describing how they are familiar with. The experiment was executed in 2 systems, i.e. commercial system and our system.

For each topic, each participant entered 3 queries and executed in both system. Then, we picked 3 highest news results for each query from each system. To evaluate the relevance metric, each participant assessed the relevance of those news results with scale from 1 to 5. Meanwhile, fresh metric was measured by counting number of news results that did not occur in each system. The result of the experiment was summarized in Table 3..

TABLE III
EXPLANATION OF FUNCTION QUERY

Respondent	**Time of evaluation**	$\bar{x}NDCG_3$ **our system**	$\bar{x}NDCG_3$ **commercial search engine**	**nMissing our system**	**nMissing commercial search engine**
1	17:57–19:03	0.99893 (+ 0.277 %)	0.99616	13	29
2	19:07–20:03	0.98225 (- 1.352 %)	0.99554	21	40
3	20:19–21:05	0.98289 (- 0.581 %)	0.98861	22	15
Mean		0.98751 (- 0.607 %)	0.99351	18.67	28.333

Regarding relevance metric, it is clear that our system has similar relevance level with the commercial one (i.e. mean difference for both system is 0.607%). For fresh metric, the performance of our system outperforms the commercial search engine during the first and second time of evaluation. However, during the third (last) time of evaluation, the performance of commercial search engine is better than our system. This results support our analysis that for very recent news (less than 2 hours), Twitter data is appropriate source for better ranking process. On the other hand, ranking process based on CTR prediction would display news result accurately since that news have enough clicks during previous time period.

VII. CONCLUSIONS

Based on our analysis and experiment, we can conclude that using Twitter data as *document collection* could give *news results* that have similar relevance level with *news results* on commercial search engine. Moreover, retweet count on Twitter Data is appropriate predictor to display *most recent news* (timley is less than two hours) as news *results*. On the other hand, the commercial search engine CTR measurement would effective to display less recent news result.

REFERENCES

[1] Dong, Anlei, Ruiqiang Zhang, Pranam Kolari, Jing Bai, Fernando Diaz, Yi Chang, Zhaohui Zheng, and Hongyuan Zha. "Time is of the essence: improving recency ranking using twitter data." In Proceedings of the 19th international conference on World wide web, pp. 331-340. ACM, 2010.

[2] Diaz, Fernando. "Integration of news content into web results." In Proceedings of the Second ACM International Conference on Web Search and Data Mining, pp. 182-191. ACM, 2009.

[3] König, Arnd Christian, Michael Gamon, and Qiang Wu. "Click-through prediction for news queries." In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp. 347-354. ACM, 2009.

[4] Becker, Hila, Mor Naaman, and Luis Gravano. "Beyond Trending Topics: Real-World Event Identification on Twitter." ICWSM 11 (2011): 438-441.

[5] Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. "Earthquake shakes Twitter users: real-time event detection by social sensors." In Proceedings of the 19th international conference on World wide web, pp. 851-860. ACM, 2010.

[6] Croft, W. Bruce, Donald Metzler, and Trevor Strohman. Search engines: Information retrieval in practice. Reading: Addison-Wesley, 2010.

[7] Sankaranarayanan, Jagan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. "Twitterstand: news in tweets." In Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems, pp. 42-51. ACM, 2009.

[8] Kwak, Haewoon, Changhyun Lee, Hosung Park, and Sue Moon. "What is Twitter, a social network or a news media?." In Proceedings of the 19th international conference on World wide web, pp. 591-600. ACM, 2010.

[9] Cha, Meeyoung, Hamed Haddadi, Fabricio Benevenuto, and P. Krishna Gummadi. "Measuring User Influence in Twitter: The Million Follower Fallacy."ICWSM 10 (2010): 10-17.

[10] Petrovic, Sasa, Miles Osborne, Richard McCreadie, Craig Macdonald, Iadh Ounis, and Luke Shrimpton. "Can Twitter replace Newswire for breaking news?." In ICWSM. 2013.

[11] Rose, Daniel E., and Danny Levinson. "Understanding user goals in web search." In Proceedings of the 13th international conference on World Wide Web, pp. 13-19. ACM, 2004.

[12] Kelly, Diane. "Methods for evaluating interactive information retrieval systems with users." Foundations and Trends in Information Retrieval 3, no. 1—2 (2009): 1-224.

[13] Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze.Introduction to information retrieval. Vol. 1. Cambridge: Cambridge university press, 2008.