Exploring Search Behaviour in Microblogs

Anirban Chakraborty
ADAPT Centre, School of Computer Science & Statistics, Trinity College Dublin, Ireland
anirban.chakraborty@adaptcentre.ie

Microblogging sites, like Twitter, have become a common means to share information, participate in topical debate, and express opinions about events or entities. However, the way people search for information in microblogs, and interact with the bearers of this information, is still under-explored, especially when it comes to looking for opinions for a decision-making process. We propose a task-based user study to investigate the search behaviour of users when looking for opinions. We capture users' clicks and ratings and create a query log dataset. The analysis of this data can provide insights about what elements of a tweet are deemed relevant while searching for opinionated information.

Microblog, User Study, Query Log

1. INTRODUCTION

The importance of social media has grown rapidly in recent years. People often search Twitter and other social media to gather opinions, thoughts, comments and discussions about a wide variety of topics, from President Trump's latest online outburst to demonetisation in India. However, the limited length of tweets, the common use of slang and emoji, and the widespread use of URLs, hinder traditional Web search systems from helping the user find relevant information. To this aim, the Microblog track Ounis et al. (2011) was introduced in TREC 2011 to examine search and evaluation methods in microblogging environments like Twitter. However, the investigation of user search behaviour in specific scenario(s), particularly when searching for opinions, still remains a less thoroughly explored area. In this paper, we propose a task-based study of users' search behaviour in microblogs in a specific search scenario, namely opinion search. We collected a dataset of tweets, and set the participants tasks where they had to search through this collection to find opinions which satisfy a specific goal. We then stored information about the users' interactions with the search tool in order to examine their behaviour when performing the tasks. We will also investigate if the analysis of this search behaviour can become a useful source of information to better understand user context. We record search behaviour like query reformulations and link clicks, in addition to the explicit expression of relevance by means of starred tweets, in order to gauge the difficulty associated with the retrieval of opinions from Twitter and to capture patterns and evidence from social media that

Figure 1: *Tweet containing review-like data*

can be exploited in other tasks, like query expansion, item recommendations, or user modelling. We define the experimental scenario in detail, and explain how we capture contextual information related to entities/events, in Section 4. We are interested in Twitter data for the unique nature of tweets. People tend to be more spontaneous and immediate on Twitter. They express relatively unfiltered opinions by giving immediate voice to their daily experience. Tweets can often take the form of a "review", giving an insight into the tweeter's opinion about what they are interacting with in the real world, be it a restaurant, hotel, TV show, celebrity etc. Tweets like the one in Figure 1 provide potentially valuable information about the hotels mentioned and their relationship to the Eiffel Tower, which is different from other on-line reviews (ratings/comments).

The main contribution of this paper is designing a task-based user study to generate a query log of user search behaviour in microblogs, specific to the task of opinion finding. This type of data is not publicly available from platforms such as Twitter. It may also be possible to capture useful contextual information about the entities/events mentioned in tweets, which could be of further use in research areas such as context-aware

recommendation. Tweets can also embed location and time information, providing a further source of context.

The rest of this paper is organized as follows: In Section 2 we discuss the related work. We describe the data collection procedure in Section 3, the proposed user study is explained in Section 4, while we analyse the collected query log data in Section 5 and outline some future research directions in Section 6.

2. RELATED WORK

Exploring search behaviour in Web documents has been well studied. In a user study Teevan et al. (2004) have investigated the orienteering search behaviour of user in emails, files and on the Web. Aula et al. (2010) explored how user search behaviour changes as search becomes more difficult. On the other hand, the short length of microblogs, and the immediate nature of this communication means that it poses serious challenges for the effective retrieval of pertinent information. Teevan et al. (2011) showed how users' search behaviour differs on Twitter when compared to Web search. In a general comparative study, they noticed that queries on Twitter are significantly shorter (number of words) than those issued on general Web search engines, although they contain longer words. We have designed a task-based user study to generate query logs of search behaviour in microblogs by setting up a controlled environment where users search for opinions in tweets in order to satisfy some given criteria. While Teevan et al. have conducted a large-scale, general comparative study of microblog search behaviour versus general web search, our proposed study is focused on examining user search behaviour in the specific task of finding opinions related to entities in microblogs.

3. DATA COLLECTION

We have collected a dataset of tweets by using the Twitter public streaming API. A random sample of the public Twitter stream was collected over a 16 day period (16 February-3 March) in 2017. After filtering, we retained English tweets only. We named our collection of 10 million English tweets "RandomTweets2017". Each tweet is represented as a JSON object that contains all the fields returned by the Twitter API[1]. A field can be an integer (such as the unique "*id*", the "*favorite_count*" or "*retweet_count*" of the tweet), a string (such as the UTF-8 "*text*" of the actual status update), a boolean variable (such as the "*verified*" status of an account) or even another JSON object (e.g. information about another user who originally tweeted the "*text*"). We indexed the corpus using Apache Lucene 5.4.0. The EnglishAnalyzer is used to parse the *text* fields of tweets. Table 1 shows the statistics of the indexed "RandomTweets2017" collection.

[1]Defined at `https://dev.twitter.com/overview/api/tweets`

Information	Value
No. of documents	10,097,460
No. of unique terms	7,406,152

Table 1: *Indexed Collection Statistics*

4. USER STUDY

As outlined in Section 1, the primary focus of this paper is to explore user search behaviour in microblogs, when searching for tweets containing opinions related to an entity or event. We have conducted an initial task-based user study, with 15 users, on our "RandomTweets2017" collection to examine how a user constructs and adapts their search queries to find tweets that are relevant to a specific information need. This initial study consisted of a single task. We captured users' clicks and analysed the captured query log in Section 5. As future work, we plan to conduct a larger, more scientifically rigorous study, in which we will give a set of different tasks to each user and capture their queries and clicks. More information on these tasks can be found in Section 4.2.

4.1. Search Application

We have developed "tweetsearch"[2], an online application that offers search capabilities over the "RandomTweets2017" collection. The system uses a CombSUM Shaw et al. (1994), which is a combination of three retrieval models (implemented in Lucene): Language Model Jelinek-Mercer (with the smoothing parameter λ=0.6); Language Model Dirichlet; and BM25. When a user performs a query in *tweetsearch*, a set of tweets is retrieved and displayed on screen, as shown in Figure 2. We display only the tweet text, without any images or videos, since in our initial study we are only interested in analysing the interaction of the user with the text content of a tweet. In our full-study we plan to include images/videos that are associated with the tweet texts as they may carry information that would aid the user in making a decision. We need to capture the relevance of tweets retrieved for each search task assigned to the user. In order to achieve this, a star button is displayed beside every tweet in order to allow the user to explicitly

[2]Available at `http://anistudy.adaptcentre.ie/tweetsearch/home.jsp`

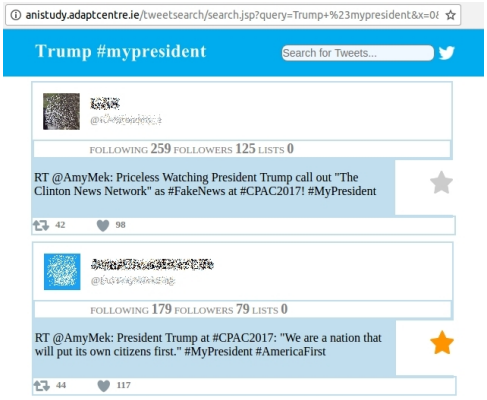

Figure 2: *"tweetsearch" interface*

mark those which they perceive to be relevant for the task at hand. Every tweet is presented along with the name, screen name and profile image of the user. We call this set up our *Baseline Group* (BG). However, Twitter offers other content that can be displayed and which may impact the perceived relevance of a tweet. We have grouped these visual elements into three groups that can be added to the Baseline Group in order to understand their role in the assessment. *Group 1* (G1) shows Retweet and Favourite counts. *Group 2* (G2) shows the Followers, Following and Lists counts, while *Group 3* (G3) displays the Verified account badge of the user. "*tweetsearch*" presents search results to users using four combinations of these groups: Appearance 1 (BG only), Appearance 2 (BG+G1), Appearance 3 (BG+G1+G2) and Appearance 4 (BG+G1+G2+G3). In our initial user trial user study we have evaluated only Appearance 4. In the planned full user study, each user will perform multiple tasks, with each task requiring them to use a different appearance. In such a way we can analyse the impact of each group of Twitter elements on each task, and thus on the search experience. It is our intuition that visual elements, such as Retweet or Favourite count, can bias the user perception of relevance since they are an indicator of the importance or popularity of the tweet or the tweeter. Indeed, Vosecky et al. (2014) showed that retweets and favourites can be exploited as a source of implicit relevance feedback.

4.2. Task and Click Log

Consider the following two example tasks:

i. Search for and "Star" tweets that show positive sentiment toward U.S. President Donald Trump.

ii. You are travelling to Barcelona this summer with your spouse. Search for and "Star" tweets that give you valuable information about places to stay, visit or spend time - for instance Hotels, places, beaches, etc.

We will define an extended set of tasks like task (i) or (ii) for use in our planned, full user study. Each task consists of a specific scenario, where the user is required to search for opinions present in tweets. In task (i), relevant tweets should contain information related to President Trump that is positive in tone. This could be related to his decisions, his activities or just his persona in general. In an unbalanced dataset, with content strongly polarised toward a specific sentiment, the search for positive or negative utterances could represent a real challenge for the user. For instance, if the corpus contains a huge number of tweets which convey neutral or negative sentiment towards President Trump, then this task will become more difficult.

In task (ii) the relevant tweets should contain useful information about places of interest for tourists. The whole task identifies a scenario where the city, the trip type and the location describe the contexts. We could set an even more specific task by specifying a place name (e.g. Las Ramblas) or entity (e.g. Mandarin Oriental Hotel). Then, the relevant tweets will be able to capture information about that specific entity in the given context (such as "in winter", "with family", "weekend break" etc.). Once data about the search behaviour has been collected, an analysis will be performed to derive useful statistics about the search task, such as the average query length, average number of queries in a session etc. In our trial study, we recruited 15 users who were given only task (i). We instructed each user to mark at least 10 relevant tweets for the task. We made sure that at least 10 tweets exist in our corpus that satisfy the task criteria. When a user marks a tweet as relevant, *tweetsearch* captures that click and stores it in the query log. If a tweet contains a URL(s), we also capture any URL click, independent from the starring. Each entry of the generated query log contains 6 fields: a unique user ID, a unique tweet ID, rank of the tweet in the rank list, query text, relevance (1: for starred tweet; 0: otherwise) and number of URL clicks. Note that, *Tweet_ID*, *rank*, *Relevance* and *Click_URL* can be 0, if a user does not find any relevant information.

5. QUERY LOG ANALYSIS

Our plan is to do an extensive user study with a minimum of 100 users, each provided with different tasks that require them to identify some meaningful information about an entity. A user session is defined as the time required to complete one task. In this paper, we limit our analysis to the trial user study conducted on task (i). Our analysis of the query log has produced some interesting statistics, shown in Table 2. The average rank of the starred (relevant) tweets (32.38), which is very low when compared

Information	Value
Avg. #queries in a session	5.27
Avg. query length	2.39
Avg. rank of the starred tweets	32.38
Avg. #starred tweets for a query	4.35
Avg. #tweets with clicked URL(s) for a query	0.57
Number of users	15

Table 2: *Query Log Statistics*

with a traditional Information Retrieval scenario, if taken at face value. However, it should be noted that this was a deliberately difficult task, with relatively few relevant tweets - the dataset contains many more tweets with negative sentiment toward President Trump than positive. It also may be an indication of the challenges faced by IR systems when searching over very short documents. These results will be explored in detail when more data is collected as part of the full user study. We have noticed two different search behaviours among users: 1) The "reformulators": users that frequently reformulate the query when no relevant document is retrieved in the top 10; and 2) The "investigators": users that explore the long tail of results and then mark lower ranked documents as well. Then, the high range of the rank values makes the mean rank low. We noticed that the perception of relevance differs between users. A user may judge a tweet like "Trump made journalism great again" as positive, while someone else may find it sarcastic and discard it as not relevant. Generally, tweets like "Trump saving taxpayers money by not filling 'unnecessary' administration positions. Trim the fat! #CommonSense #MAGA https://t.co/GRr38NQMzE" and "RT @Missyblueblue: LOVE this man Trump....Obummer was a taqiyya talker...Trump is a dynamic doer. https://t.co/L5dvBcRf2Z" were considered positive. While the number of starred tweets with clicked URLs is insignificant in our preliminary study, and the majority of users did not click on URLs, there is also a case where the tweet text itself does not contain any indication of sentiment, but a link to an external document (such as an image, video clip, or news article) has pertinent information. We were able to identify such a case by following the rates of the only zealous user that clicked on each URL.

6. CONCLUSION & FUTURE DIRECTIONS

In this paper we have proposed a task-based user study, and based upon our trial study, we generated a query log that reveals the task specific user search behaviour in microblogs. We have seen that our user study can capture important information about an entity/event. This initial user study examined how a user behaves when performing searches in the presence of a difficult task, such as those that involve judging opinions and sentiments, or searching for very specific topics. We believe that the interesting indicative results from this initial study make this area worth of further research. We are in the process of extending our user study in order to address the following research questions. Does a difficult search task increase the number of query reformulations? Does the tweet-specific information provided in the User Interface impact on the relevance assessments for a tweet? Moreover, it will also be interesting to study how useful or informative the opinions present in tweets are. Can the opinions be used to reliably predict the rating of an item? Can these opinions be used to improve or supplement user modelling? We may also consider to utilise the images or video streams that are present in a tweet, in addition to the tweet text. For instance, if someone tweets an image of a great view of the Eiffel tower from a hotel balcony, it indirectly indicates that this hotel could be of interest for someone looking for an accommodation nearby the Eiffel tower with good views. We think this kind of information can be valuable for generating recommendation lists and we plan to leverage the sentiment associated with such tweets in recommender systems.

ACKNOWLEDGEMENTS

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

REFERENCES

Aula, A., R. M. Khan, and Z. Guan (2010). How does search behavior change as search becomes more difficult? CHI '10, pp. 35–44.

Ounis, I., C. Macdonald, J. Lin, and I. Soboroff (2011). Overview of the trec-2011 microblog track. In *In Proceedings of TREC 2011*.

Shaw, J. A., E. A. Fox, J. A. Shaw, and E. A. Fox (1994). Combination of multiple searches. In *The Second Text REtrieval Conference*, pp. 243–252.

Teevan, J., C. Alvarado, M. S. Ackerman, and D. R. Karger (2004). The perfect search engine is not enough: A study of orienteering behavior in directed search. CHI '04, pp. 415–422.

Teevan, J., D. Ramage, and M. R. Morris (2011). #twittersearch: A comparison of microblog search and web search. WSDM '11, pp. 35–44.

Vosecky, J., K. W.-T. Leung, and W. Ng (2014). Collaborative personalized twitter search with topic-language models. SIGIR '14, pp. 53–62.