Preliminary Research on Thesaurus-Based Query Expansion for Twitter Data Extraction

Vidya Nakade
University of Alabama
Tuscaloosa, Alabama
vnakade@crimson.ua.edu

Aibek Musaev*
University of Alabama
Tuscaloosa, Alabama
aibek@cs.ua.edu

Travis Atkison
University of Alabama
Tuscaloosa, Alabama
atkison@cs.ua.edu

ABSTRACT

With the increasing popularity of microblogging and social media platforms like Twitter, researchers are trying to make use of the massive amount of user-created data to explore new applications/tools. Success of research in data science is highly dependent on the amount and type of data collected. For this effort, a thesaurus-based query expansion technique from information retrieval will be used to extract additional Twitter data. Though there has been research in this general area, our effort concentrates on applying a thesaurus-based query expansion for Twitter retrieval. Experiments are performed to collect Twitter data using the proposed approach for query terms related to disaster situations like hurricanes and shootings. We observed an increase of 32% in tweets received for the Hurricane Harvey event, and a 131% increase in the volume of tweets for a query related to the Vegas shooting incidence using the thesaurus-based query expansion approach.

KEYWORDS

Thesaurus, Query Expansion, Information Retrieval, Twitter Retrieval

ACM Reference Format:
Vidya Nakade, Aibek Musaev, and Travis Atkison. 2018. Preliminary Research on Thesaurus-Based Query Expansion for Twitter Data Extraction. In *ACM SE '18: ACM SE '18: Southeast Conference, March 29–31, 2018, Richmond, KY, USA.* ACM, New York, NY, USA, Article 4, 4 pages. https://doi.org/10.1145/3190645.3190694

1 INTRODUCTION

The use of Twitter has greatly increased over the past decade. There have been several research efforts that use Twitter data for multiple purposes including healthcare [6], politics and poll predictions [16, 21], sentiment analysis in emergencies [4], developing useful applications for detection and/or prediction of disastrous situations [8, 9, 13, 15]. There are several reasons that Twitter data is popular among researchers; it is available for free, is in real time, and in large quantity. The properties which make Twitter data so interesting for research are its scale, immediacy, and availability. Currently, Twitter receives 500 million pieces of data (text, audio, video) in uploads everyday. People are increasingly using social media to interact, share their updates or pass along news. The urge to post one's latest updates on social media right now has grown, and thus Twitter is seeing more discussions/posts about latest happenings from around the world. Twitter provides limited access to its data for free through a public API which makes it easy for researchers to perform data analysis.

Twitter data makes a powerful resource when dealing with life critical situations like natural disasters, such as hurricanes, earthquakes, and landslides. In disastrous situations, people seek help from social media platforms like Twitter because they have limited/no other means of communication with the outside world. This kind of data is very critical for disaster response divisions of government, and thus, collecting every tweet coming from people in the affected area is highly important. The data is not only useful for disaster management and to provide alerts [9], but also for critical studies like analyzing the tweets to derive some insights about the monetary damages, the sentiment of people, and disaster response surveys. Some of the research goes far beyond basic analysis to build tools to detect and/or predict disasters like earthquakes [8, 15] and landslides [13].

Methodologies from information retrieval can be used to fetch more effective, more relevant data pieces efficiently. Information retrieval hinges on finding relevant, but also more and more data for a given query. To achieve this objective, relevance feedback is implemented, either explicitly or implicitly. This feedback is used to perform query expansion, which is a term used for modifying/updating/adding to the original user query using alternative terms. The objective of query expansion is to capture the user's intent thoroughly, or to simply produce a query which is more likely to retrieve the relevant information. One of them is the thesaurus-based query expansion method. In thesaurus-based query expansion, a global resource is used to modify the query or add to the query provided by the user. In other words, the technique uses synonymous words fetched from a thesaurus resource to try various query terms and retrieve more relevant information. The resource used is independent of the query term provided, and is generally present online to access. There are multiple online-thesaurus sources available, such as thesaurus.com and WordNet.

There are multiple research efforts that apply query expansion techniques to Twitter data, but none of them apply thesaurus-based query expansion. Section 2 discusses the existing literature. Section 3 discusses the application query expansion which uses the thesaurus resource, and Section 5 presents the results showing effective and quantitative improvements in retrieved Twitter data.

*Corresponding author.

ACM SE '18, March 29–31, 2018, Richmond, KY, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5696-1/18/03…$15.00
https://doi.org/10.1145/3190645.3190694

2 BACKGROUND

There are multiple research efforts in the area on using query expansion for Twitter data collection. Zhu et al. [23] constructed a real-time personalized Twitter search system that re-ranks search results of data based on the user's preferences and interests. Since user profile terms tend to suffer from the vocabulary mismatch problem with respect to tweet terms, a search engine based query expansion is applied to the profile terms to alleviate this issue. Zingla et al. [24] aimed to deal with short, ambiguous queries by expanding with semantically related terms extracted from Wikipedia, DBpedia, and a text-mining technique that yields terms that are commonly associated with terms in the original query, e.g. stadium would yield football. Our work is similar to that of Zingla et al. as we propose to find more and related tweets by expanding generally used short queries on Twitter.

Sodanil and Ketmaneechairat [19] investigate the effect of the number of added terms in a query expansion and experiment with having experts choose five "subjective" terms from the ten most frequent terms of an initial retrieval to add to the initial query. To overcome the challenge of vocabulary mismatch in Twitter search, Qiang et al. [14] utilize the knowledge base Freebase to generate expansion terms that are conceptually related to the terms found in the query. They argue that the topical structure of Freebase makes it more suitable for such an application than sources like Wikipedia, which contains impractically long and detailed articles, and WordNet, which is limited to synonyms. Since recency is valued in microblogging, they also incorporate temporal evidence to promote more recent tweets [14].

For improving Twitter search and to retrieve more relevant tweets, Efron et al. [7] suggest expanding the documents (tweets). Their research objective is to be able to improve search effectiveness of query on short text corpora [7]. Though it sounds relatively close to our idea, the objective and the study are different from ours. In their research, Massoudi et al. [12] tried query expansion based on the highest frequency terms occurring in top n terms from the database of retrieved results terms, but this approach can face the problem of bias towards most frequently occurring terms; and may be leaving out some important tweets, which may not necessarily contain these terms [12]. In some of the recent work by Albishre et al. [3] used pseudo-relevance feedback for microblog retrieval by using lexical and topical evidence from pseudo-feedback with respect to the original query. Our approach is different in several ways from the state of the art. We propose and verify our claim for life-critical situations of disasters; man-made or natural. In such scenarios, every piece of information is crucial; as the data collected is to be analyzed for making decisions about disaster management and emergency communication.

3 RESEARCH

The free access to tweets provided by the Twitter API is limited by the amount of time in the past for which tweets can be accessed, along with the partial data-sharing constraint. Twitter shares only a subset of tweets through its public API, up to 7 days old. To overcome this limitation, data was collected from an online resource named discovertext.com. Discover text is a text analytic software developed by Shulman [18]. Several articles, blogs and theses in literature use discovertext.com [2, 5, 17, 20].

It is common for any Twitter data analysis study to include a filtering stage. Researchers have in fact implemented several stages of filtering before conducting any analysis on the collected tweets [13]. As the aim for this project is to investigate the application of a thesaurus for expanding short Twitter queries and to observe if there is some increase in the volume of tweets received, filtering was not considered in the scope of this project. A few query terms were selected, then the data was pre-filtered on several restrictions to ensure the data received was about the query terms only. For experiments, the tweets corpus was collected from discovertext.com, and the natural disaster Hurricane Harvey was chosen for the query. Two pre-filters were applied to the Twitter data from discovertext.com as follows:

(1) Tweets containing the term "Harvey" in hashtags
(2) Tweets created for the duration of 17 Aug to 3 Sept

The first filter made sure that only tweets containing the term "Harvey" in hashtags were received, which is one of the query terms. The data collected was also pre-filtered based on the date range for which Hurricane Harvey was active, 17th August 2017 to 3rd September 2017. With these pre-filters, a tweet corpus with around 35,000 tweets was collected. On this collected corpus, a search was conducted for the query term "Hurricane", which yielded 5,152 tweets. This made sure only the tweets created in the above-mentioned date range and containing the term "Harvey" in hashtag and containing the word "Hurricane" in body were collected. Thus, only relevant tweets were collected. The query would look like

(Harvey) (Hurricane) (since: August 17, 2017 till: September 03, 2017)

For query expansion, thesaurus.com was used as the global thesaurus source [11]. It is a digital thesaurus source provided by dictionary.com. Thesaurus.com also provides millions of English definitions, translations, audio pronunciations, example sentences and origins of words. When searching for the synonyms for the term "Hurricane", terms like "tornado, gale, storm, twister, cyclone, monsoon" were found. Using these keywords, tweets were filtered from the downloaded corpus to find 1,644 unique tweets. As per the filters, the query term with expansion would look like:

(Harvey) (Hurricane OR Tornado OR gale OR storm OR twister OR cyclone OR monsoon) (since: August 17, 2017 till: September 03, 2017).

4 VALIDATION OF CONCEPT

There exists a time-range constraint imposed by the Twitter API that disables the retrieval of tweets older than a week. To bypass this, a python package written by Jefferson Henrique called "Get Old Tweets" was used. It works around this by instead retrieving tweets through the Twitter Advanced Search interface, which is readily accessible on the Twitter website using a browser[1] and does not have the aforementioned limitations, by imitating the Twitter

[1] https://twitter.com/search-advanced?lang=en

URL to retrieve tweets depending on the provided criteria. With this package, the thesaurus-based query expansion can be implemented in terms of Twitter Search keyword operators: parenthesizing and OR. The quickest way to understand how this works is by looking at an example. Consider the following search query input into Twitter Search:

("coffee is" OR "computers are" OR "apples are") amazing

The Twitter Search engine will return tweets containing at least one of these sets of words:

(1) ("coffee", "is", "amazing")
(2) ("computers", "are", "amazing")
(3) ("apples", "are", "amazing")

To clarify the query expansion using a thesaurus, instead of simply appending synonyms to the original search query, which would have the opposite effect of restricting our search and thus result in fewer retrieved tweets, we replace query terms with parenthesized OR expressions like the one above that indicate acceptable alternatives to the expanded terms. In practice, not every word will need to be expanded (names for example) and thus a reasonable implementation of thesaurus-based query expansion should provide a means for indicating that a term should not be expanded. Terms that should be expanded are denoted as expandable terms. The procedure for expanding a Twitter search query is stated as:

(1) For each expandable term x, retrieve a set S of synonyms and then add x to S. For example, if x = "fantastic", then S = "fantastic", "insane", "imaginative", "unbelievable", "singular", "incredible".
(2) Convert S to a parenthesized OR expression by forming a list of the elements of S delimited by ORs and then surrounding this list with parenthesis. To deal with multi-worded terms and synonyms, encapsulate them in quotation marks. Thus, S in the previous example is converted to the following string: ("fantastic" OR "insane" OR "imaginative" OR "unbelievable" OR "singular" OR "incredible")

This implementation uses the PyDictionary package to retrieve a list of synonyms from thesaurus.com[2]. One important thing to note is that this procedure assumes that the user wants at least one word from each parenthesized expression. Indeed, one is certain to acquire far more data by instead having the looser requirement of matching zero or more words from a given set of words. Such sets are denoted as optional sets. It is a simple modification to specify zero or more words by taking the union of all optional sets constructed in step 1. The query expansion procedure is then applied to the following query, which is an example of what a researcher analyzing social media data related to the Las Vegas shooting might use:

vegas shooting gunman police

which expands to:

("vegas") ("shooting" OR "gunfire" OR "firing" OR "blasting" OR "gunning" OR "discharging") ("gunman" OR "assassin" OR "hit man" OR "sniper" OR "killer" OR "triggerman") ("police" OR "detective" OR "force" OR "law enforcement" OR "man" OR "corps")

Table 1: Results of Thesaurus-based Query Expansion

Query Terms	Tweets Without Expansion	Tweets With Expansion	% Increase in Number Tweets
Hurricane Harvey	5,152	6,796	31.90
Shooting	304	703	131.25

The search was restricted to the day of the shooting: October 2, 2017, UTC-07:00. Ultimately, 304 tweets were received without expansion and 703 with expansion, a 131.25% increase in the retrieved tweets. All tweets were found to be relevant in this case after manual classification, so the usual metrics like signal-to-noise ratio are not applicable. If the sets of words corresponding to each search term are made optional (as defined in the previous section), an immeasurable number of tweets are received for both searches with and without expansion because the Twitter server will halt the retrieval after processing a large number of tweets (on the order of a hundred thousand tweets). This is perhaps a safety measure to prevent abusing the service. As a result, the search used in this research was kept more restricted to enable comparing the quantity of data retrieved with and without using query expansion. This is justified since the goal is not to retrieve a corpus of tweets as large as possible but instead to demonstrate that using expansion will result in more tweets.

5 RESULTS

The results obtained by applying thesaurus-based query modifications are promising. This effort applied thesaurus-based query expansion for two disasters: Hurricane Harvey and the Las Vegas shooting. These specific events were chosen for our experiments and evaluation because these are life critical situations and every tweet can be important.

Table 1 shows the query terms used to conduct the study and the results that were produced. As described above, query terms like the name of the place or hurricane were not expanded. Only terms describing the phenomenon like shooting, hurricane or gunman, etc. were expanded. As seen in the table, when an expanded query for term "hurricane" was used, 6,779 tweets were gather as opposed to the 5,112 tweets returned without query expansion. The 6,779 tweets includes tweets extracted without query expansion, and hence the new tweets extracted by using expansion are 6,796-5,152=1,644. This is a gain of: (6,796-5,152)/5,152*100 = 31.90%. The increase in tweets of 32% can be very important in disaster situations because the disaster management [10] and emergency communication and knowledge management [1, 4, 22] services are increasingly depending upon social media for communication or discovering cries for help from affected people.

The data for query expansion for the other query "Vegas Shooting" validates the results obtained. When the script was run for the query as described in previous sections, 304 tweets were received without expansion and 703 with expansion. As discussed before, the 703 tweets are inclusive of tweets without expansion, i.e., tweets with original query term; therefore, there were 399 more tweets retrieved using query expansion. This provides a gain of (703-304)/304*100 = 131% more tweets compared to using single

[2] https://pypi.python.org/pypi/PyDictionary/1.3.4

query term. In both scenarios, an increase in relevant tweets was observed.

6 CONCLUSION

In this age of a social media connected world, the data posted on these platforms is tremendous and contains important information which can be analyzed and studied to find interesting insights. Twitter, being one of the most popular microblogging sites, provides scientists with a huge amount of text-based data; however, extracting the relevant data before starting out any analysis still poses a challenge. With the application of the information retrieval concept of using thesaurus-based query expansion for retrieving more data, researchers can have more data than using a single query term. With these experiments, it was proven that a significant amount of additional tweets can be retrieved when using synonymous terms as suggested by the online thesaurus source thesaurus.com. This claim was tested on two disaster situations: Hurricane Harvey which was a natural disaster and the Las Vegas shooting which can be considered a man-made disaster. For Harvey, the thesaurus query expansion retrieved almost 32% more relevant tweets. The result was verified by applying the idea for fetching tweets using the thesaurus query expansion to the Vegas shooting; 131% more tweets were retrieved with expansion.

In disaster events like hurricanes, fires, explosions, shootings or terrorist attacks, retrieving more and more tweets possible for analysis is a very important and challenging task because these are life critical situations and every tweet can be crucial for the further study. The thesaurus query expansion approach discussed in this paper is one way to tackle this situation. As a future path, thesaurus-based query expansion can be combined with existing approaches from literature as discussed in previous sections to compare the percentage increase in tweets with/without our approach.

REFERENCES

[1] Babak Abedin, Abdul Babar, and Alireza Abbasi. 2014. Characterization of the use of social media in natural disasters: a systematic review. In *Big Data and Cloud Computing (BdCloud), 2014 IEEE Fourth International Conference on*. IEEE, 449–454.

[2] Wasim Ahmed. 2017. Using Twitter as a data source: an overview of social media research tools (updated for 2017). *Impact of Social Sciences Blog* (2017).

[3] Khaled Albishre, Yuefeng Li, and Yue Xu. 2017. Effective pseudo-relevance for Microblog retrieval. In *Proceedings of the Australasian Computer Science Week Multiconference*. ACM, 51.

[4] Ghazaleh Beigi, Xia Hu, Ross Maciejewski, and Huan Liu. 2016. An overview of sentiment analysis in social media and its applications in disaster relief. In *Sentiment Analysis and Ontology Engineering*. Springer, 313–340.

[5] Yngvil Nesheim Beyer. 2012. Using discovertext for large scale twitter harvesting. *Microform & Digitization Review* 41, 3-4 (2012), 121–125.

[6] Cynthia Chew and Gunther Eysenbach. 2010. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. *PloS one* 5, 11 (2010), e14118.

[7] Miles Efron, Peter Organisciak, and Katrina Fenlon. 2012. Improving retrieval of short texts through document expansion. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 911–920.

[8] Michelle Guy, Paul Earle, Chris Ostrum, Kenny Gruchalla, and Scott Horvath. 2010. Integration and dissemination of citizen reported and seismically derived earthquake information via social network technologies. *Advances in intelligent data analysis IX* (2010), 42–53.

[9] Nathan O Hodas, Greg Ver Steeg, Joshua Harrison, Satish Chikkagoudar, Eric Bell, and Courtney D Corley. 2015. Disentangling the lexicons of disaster response in twitter. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 1201–1204.

[10] Cheng-Min Huang, Edward Chan, and Adnan A Hyder. 2010. Web 2.0 and internet social networking: A new tool for disaster management?-lessons from taiwan. *BMC medical informatics and decision making* 10, 1 (2010), 57.

[11] B. Kariger and D. Fierro. 1995. Thesaurus.com. - By Dictionary.com, The world's leading digital resource for dictionary and thesaurus source. (1995). http://en.wikipedia.org/w/index.php?title=Comorbidity&oldid=524649802

[12] Kamran Massoudi, Manos Tsagkias, Maarten De Rijke, and Wouter Weerkamp. 2011. Incorporating query expansion and quality indicators in searching microblog posts. *Advances in information retrieval* (2011), 362–367.

[13] Aibek Musaev, De Wang, and Calton Pu. 2014. LITMUS: Landslide detection by integrating multiple sources.. In *ISCRAM*.

[14] Runwei Qiang, Feifan Fan, Chao Lv, and Jianwu Yang. 2015. Knowledge-based query expansion in real-time microblog search. *arXiv preprint arXiv:1503.03961* (2015).

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

[16] Erik Tjong Kim Sang and Johan Bos. 2012. Predicting the 2011 dutch senate election results with twitter. In *Proceedings of the workshop on semantic analysis in social media*. Association for Computational Linguistics, 53–60.

[17] Ashwin Satyanarayan, Bk Sarthak Das, and Divya Krishnan. [n. d.]. Analyzing Advertisements on Twitter during Valentine's Month. ([n. d.]).

[18] Stuart Shulman. 2011. DiscoverText: Software training to unlock the power of text. In *Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times*. ACM, 373–373.

[19] Maleerat Sodanil and Hathairat Ketmaneechairat. 2013. Information retrieval experiment on subjective words query expansion. In *Information and Communication Technology (ICoICT), 2013 International Conference of*. IEEE, 161–165.

[20] Zachary C Steinert-Threlkeld. 2017. Longitudinal Network Centrality Using Incomplete Data. *Political Analysis* (2017), 1–21.

[21] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. *Icwsm* 10, 1 (2010), 178–185.

[22] Dave Yates and Scott Paquette. 2011. Emergency knowledge management and social media technologies: A case study of the 2010 Haitian earthquake. *International journal of information management* 31, 1 (2011), 6–13.

[23] Xiang Zhu, Jiuming Huang, Bin Zhou, Aiping Li, and Yan Jia. 2017. Real-time personalized twitter search based on semantic expansion and quality model. *Neurocomputing* (2017).

[24] Meriem Amina Zingla, Latiri Chiraz, and Yahya Slimani. 2016. Short Query Expansion for Microblog Retrieval. *Procedia Computer Science* 96 (2016), 225–234.