A Survey of Information Retrieval on Microblog

Sindur Patel
Charotar University of
Science &Technology,
Changa, Gujarat
India

Nirav Bhatt
Charotar University of
Science &Technology,
Changa, Gujarat
India

Chandni Shah
Charotar University of
Science &Technology,
Changa, Gujarat
India

ABSTRACT

Twitter is most popular microblogging site. It provide us with real time data. This article Provide survey of techniques for retrieving information from twitter stream. This techniques aim is finding real world and most relevant information with respect to the query. For retrieve most relevant information used query expansion techniques. Twitter data contain large amount of information. Information rank retrieval techniques find important data and gives the final score to that information with respect to user interest profile

Keywords

Real time data, relevance information, microblog, twitter stream.

1. INTRODUCTION

1.1 Introduction

Microblog is a broadcast medium that allows user to post short and frequent message [8]. It’s a new communication medium compared with traditional data, micro blogging has gained increased attention among user, organization, research scholars in different disciplines.

Twitter is currently fastest growing micro blogging services, with more than 140 or 150 million users producing over 400 or 500 million tweets per day [8]. It’s an unable to user update status or tweets, no more than 140 characters to a networks of followers using various communication service. Tweets size are limited, Twitter is updated millions of time a day by user all over the world[8], and its data varies hugely based on user interest and behaviors. So twitter data have large amount of information scaling from latest news, events etc.

Twitter Provides timely information of any event. Observing, keeping and analyzing this content of user-generated data can yield new unprecedented important information, which not available from traditional media [8]. Tweets do the live reporting of event [13] means finding the information what people are talking away from some conferences, debates, sporting events etc.

This Article provides a survey of retrieve information from microblog using various techniques and gives the most relevance score.

1.2 Challenges

Twitter have reported everything from daily life story to real word event. Millions of tweet updated so people have no time to visualize all those tweet.

A major problem is there is no any restriction to post a tweet, update information or status so many people provide false, incorrect information about some events. Large number of spellings and grammar error, and the use of not a proper sentence structure and mixed language so people can’t distinguish important data from unused data. Not all tweets are relevant to the user query or interest profile.

One way communication. Twitter often acts as a one-way communication platform. Twitter used by celebrities, TV shows, companies and websites to simply get the word out. It is not used for relationship building.

2. MICROBLOG REAL TIME FILTERING

A user has interest in real time topic, events and they wants to stay up to date in that topic using stream of microblog posts [2].

Main goal of real time filtering is monitoring stream of microblog with respect to user interest profile (query) and find relevance score.

Provide interesting content to user by [1, 2]

1. Push notification on a mobile phone: Interest content might be shown to the user through mobile phone notification.
2. Periodic email digest: user interested profile might be aggregated into an email form that periodically sent to a user [1]. In that case a user could read a longer story about content.

2.1 System Overview

In this section introduce system architecture for retrieve tweets and do the scoring of tweets based on query.

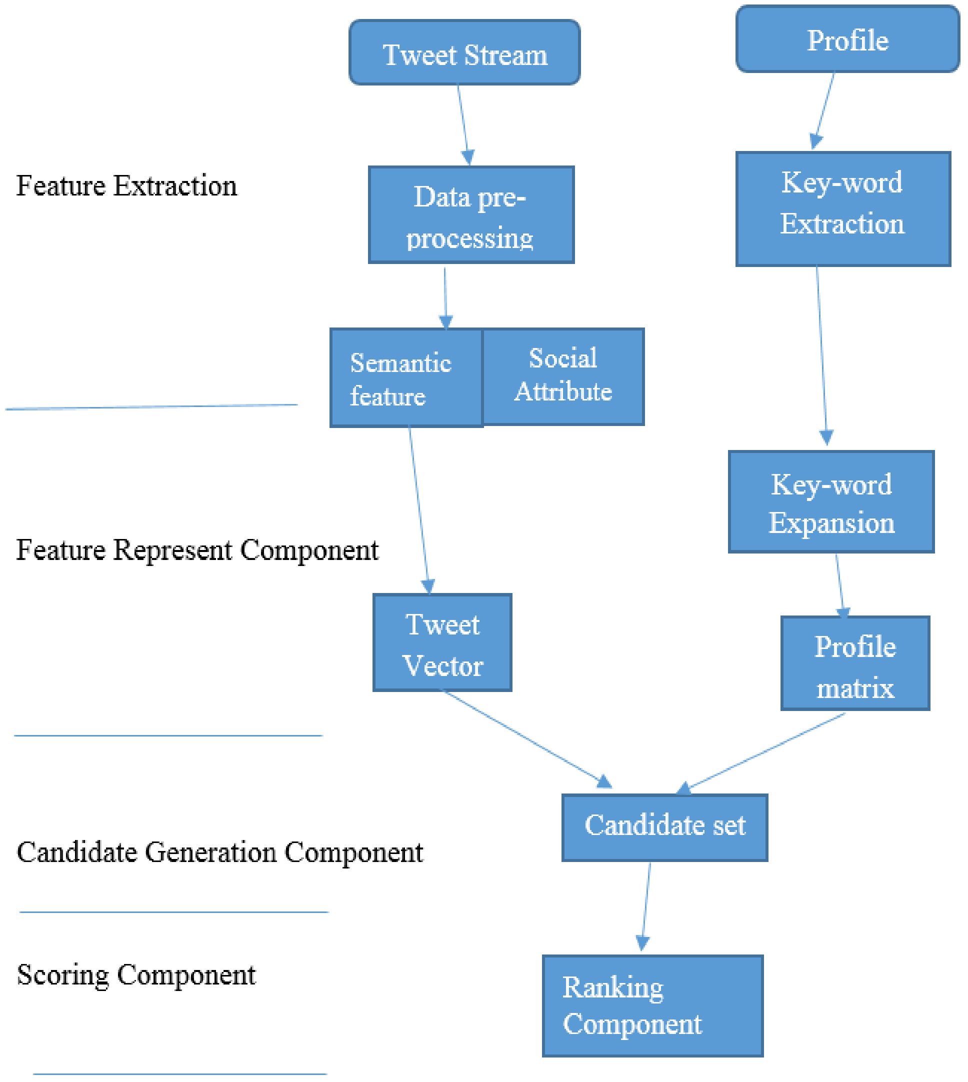

Fig 1: System Architecture

System contain four component [2].

I. Feature Extraction Component: It extract features from twitter based on TREC-API (Stream API and Rest API). After obtaining twitter stream we apply preprocessing and filtering to reduce tweets we need to process.

II. Feature Representation Component: It represent and expand semantic feature by different expansion techniques.

After extracting tweet we need to represent those feature in proper format so it is suitable to calculate relevance score between tweet and profile.

III. Candidate Generation Component: We classify tweets into the most relevant profile or remove it directly if it does not match any profile.

IV. Scoring and pushing component: By the semantic feature (consider verbs and nouns in tweet text) and social media attributes we got two score semantic (Ci) and quality (Qi) so final score Si = CiQi.

3. QUERY EXPANSION

The query provided by the user is not in a structured and that is incomplete. So then we need to expand that query and do the correct for the better relevance information.

Main problem in retrieval is that query is short and unable to accurately describe user’s information needs. So solution of this problem is query Expansion [4, 9].

For exemplar, rather than writing “earth quake”, people may instead use the word “quake” or directly include a hashtag such as “#eq” in message [4]. It is not a realistic to manually recognize the whole set of related keyword and categorization tag (i.e., hashtags) for each user profile.

Simple methods of query expansion,

- Finding exactly or nearly the same as another word of query words, and discovering for the synonym as well
- Finding all the various morphological forms of words be caused by each word in the search query[10]

- Fixing grammatical mistake, free from spelling error and automatically finding for the corrected form or forward for consideration it in the results
- Re-weighting the terms in the original query[10]

4. RELEVANCE MODEL

The topic represent as a triple of a title that contains few keywords, description that summarizes the topic in one sentences and a narrative that consist of a paragraph that gives the more details. Based on this we retrieve relative or relevance data.

4.1 Tf-Idf Score

Which calculates the similarity between a tweet and a profile in vector space model with TFIDF weight of terms [1]. Tweets and profiles can be expressed as vectors.

TFIDF is the product of two statistics[17]

- Term Frequency
- Inverse document frequency

4.1.1 Term Frequency (TF)

TF measure the number of times a term occurs in a document. But all documents are different in length so on a large document the frequency of the terms will be much higher compare to small document. So we require to normalize the document based on its size[17].

For that divide the TF by total number of terms.

Term frequency of word i in document j is

$tf_{ij} = f_{ij} / Ntj,$

Here f_{ij} = Frequency of term i in document j.

Ntj = Total Term in Doc 1

4.1.2 Inverse document frequency

In TF all term are consider equally important. But such a term like the, is, that etc. is not important and may seems lots of times so we require a way to weigh down the effect of more time appearing or occurring terms [17]. Also the words that happen less in the documents can be closer.

So Inverse document frequency is

$idf_i = 1 + \log_2 (N / df_i)$

Here df_i = document frequency of term i

N: total number of documents

The tf-idf weight of a particular word is the yield of its tf weight and its idf weight.

$W_{ij} = tf_{ij}\, idf_i = tf_{ij} \log_2 (N / df_i)$

Score of a document with respect to a query:

$Score(q,d) = \sum tf\text{-}idft,d$

4.2 BM25

BM25 is a bag of words ranking function[16] that ranks a information based on the user interest profile or query words appearing in each document information[1,2].

Developed in the context of the Okapi system in London University.

BM25 formula contains many parameter which need to be tuned from relevance assessment [15].

Given a user interest profile P, containing keywords $p_{1.}$., pn the BM25 score of a document D is

$$\text{Score (D, P)} = \sum_{i=1}^{n} \text{IDF}(pi). \frac{f(pi,D)(k1+1)}{f(p1,D)+k1(1-b+b.|D|/avgdl\ )}$$

Where f(Pi, D) is pi's term frequency in the document D, |D| is the length of the document D in words, and avgdl is the average document length in the text collection from which documents are drawn[19]. k1 and b are default parameters, usually chosen, in absence of an advanced optimization, as k1 ∈ [1.2, 2.0] and b ∈ [0.5, 0.8][19]. IDF (qi) is the IDF weight of the query term qi[19].

4.3 Language Model

Main Goal Of language Model is assign a probability of a sentence or sequence of words [18]. If sequence of word is $w_1, w_2, w_3 \ldots w_n$ then

$$P(W) = P(w_1, w_2, w_3 \ldots w_n)$$

We can also find probability of upcoming word based on previous word

$$P(w_n \mid w_1, w_2, w_3 \ldots w_{n-1})$$

Here, P(W) and $P(w_n \mid w_1, w_2, w_3 \ldots w_{n-1})$ is language model.

Rank each documents by the probability of particular document given a query [18].

P (D|Q) = P (Q|D) P(D)/P(Q)

(Here P (Q|D) = □P(qi|D) and
P(qi|D) = fqi, D/|D|)

5. CONCLUSION

In this paper survey the research in the area of information retrieval on microblog. Its aims finding most relevance information respect to user query and provide real-word occurrence. As a fast growing microblog site, Twitter provides us to valuable and real content of any real-word event. Using query expansion technique we can retrieve better result. This article display major challenges of twitter and describe different technology which used in different paper for expand query and then retrieve tweet. And give the relevance score.

6. FUTURE WORK

The research can be continued in use a language model and word2vec both for query expansion and providing more training to the system so we can get more accurate result and provide highly relevance tweet to the user.

7. REFERENCES

[1] Xiang Zhu, Jiuming Huang, Sheng Zhu, Ming Chen, Chenlu Zhang, Li Zhenzhen, Huang Dongchuan, Zhao Chengliang, Aiping Li, Yan Jia. NUDTSNA at TREC 2015 Microblog Track: A Live Retrieval System Framework for Social Network based on Semantic Expansion and Quality Model.

[2] Mossaab Bagdouri, Douglas W.Oard. CLIP at TREC 2015: Microblog and LiveQA.

[3] Luchen Tan Adam Roegiest Charles L.A. Clarke. University of Waterloo at TREC 2015 Microblog --*Track.

[4] Runwei Qiang, Feifan Fan, Chao Lv, Jianwu Yang. Knowledge-based Query Expansion in Real-Time Microblog Search

[5] S. Kumar, F. Morstatter, H. Liu. Twitter Data Analytics, published in Springer 2014.

[6] http://webtrends.about.com/od/twitter/a/why_twitter_uses_for_twitter.htm

[7] Aggrawal, C. C. 2011. An introduction to social network data analytics. In Social Network Data Analytics. Edited by C. C. AGGAR-WAL. Springer: New York, pp. 1–15

[8] Farzindar Atefeh and Wael Khreich: A Survey of techniques for event detection in twitter. NLP Technologies Inc., Montreal, QC, Canada

[9] Cher Han Lau, YueFeng Li, Dian Tjondronegoro, Microblog retrieval using topical features & query expansion, Queensland University of Technology

[10] Ashish Kankaria, Query Expansion techniques, Indian Institute of Techno-logy Bombay, Mumbai

[11] Donald Metzler, Congxing Cai, Eduard Hovy, Structured Event Retrieval over Microblog Archives, University of Southern California.

[12] Jimmy Lin, 1 Miles Efron 2, Yulu Wang, 3, Garrick Sherman 2, and Ellen Voorhees, Over view of the TREC-2015 Microblog Track, 1 University of Waterloo, 2 University of Illinois, Urbana-Champaign, 3 University of Maryland, College Park, 4 NIST

[13] http://webtrends.about.com/od/twitter/a/why_twitter_uses_for_twitter.htm

[14] Iadh- Ounis 1, Craig Macdonald 1, Jimmy Lin 2 , 3, Ian Soboroff 4 ∗ Over view of the TREC-2011 Microblog Track, 1 University of Glasgow, Glasgow, UK 2 Twitter, San Francisco, CA, USA 2 University of Maryland, College Park, MD, USA 4 NIST, Gaithersburg, MD, USA.

[15] Krysta M. Svore, Christopher J. C. Burges A Machine Learning Approach for Improved BM25 Retrieval, Microsoft Research One Microsoft Way Redmond, WA 98052

[16] Ben He and Iadh Ounis, Term Frequency Normalisation Tuning for BM25 and DFR Models, Department of Computing Science United Kingdom.

[17] https://janav.wordpress.com/2013/10/27/tf-idf-and-cosine-similarity/

[18] Book of Information Retrieval by Christopher D.Manning, Prabhakar Raghavan and Hinrich Schütze

[19] https://en.wikipedia.org/wiki/Okapi_BM25.