Reviewed Paper

Volume 3 / Issue 4 / December 2015

International Journal of Informative & Futuristic Research | ISSN: 2347-1697

Segment-Based Real Time Event Extraction From Twitter Using Named Entity Recognition & Microsoft Web N-Gram Services

Paper ID	IJIFR/ V3/ E4/ 077	Page No.	1476-1480	Subject Area	Computer Engineering
Keywords	Hybridseg, Twitter, Information Retrieval, Social Media				

1st	Chetan G. Puri	Student, Master of Engineering Department Of Computer Engineering Sir Visvesvaraya Institute of Technology, Sinnar, Nashik-Maharashtra.
2nd	S. M. Rokade	Head & Associate Professor, Department Of Computer Engineering Sir Visvesvaraya Institute of Technology, Sinnar, Nashik-Maharashtra.

Abstract

As we are now get addicted with social media, it means whatever we are doing it should send or reach within a minute, also now from last decade social media is being a part of life of all college fellows and from IT peoples, now every person has a multimedia phone set, in that most popular symbols are nothing but twitter , facebook, and what's app etc, lot many social media apps are available. Twitter has attracted billions of users to share and use most up-to-date information, news, also simultaneously it results large volumes of data produced every day, in fact every couple of hours. As for handling this large amount of information data many applications in Information Retrieval (IR) available. In this paper, we are just gathering some information and making a survey on this huge amount of data also propose a framework for tweet segmentation in some special mode, called HybridSeg. I this scenario by splitting tweets into meaningful different segments, because of systematic segments we are easily extract information whenever we need. We formalize the problem of summarizing event-tweets and give a solution based on learning the hidden state representation of the event..

1. Introduction

Twitter, as a new type of social media, has seen tremendous growth in last decade. It has attracted great interests from both industry and youth. Many private and public organizations have been use social media like twitter but also same time they use to monitor Twitter stream to collect and understand users' and customers opinions about the organizations. In short a sort of survey also

ISSN: 2347-1697
International Journal of Informative & Futuristic Research (IJIFR)
Volume - 3, Issue -4, December 2015
Continuous 28th Edition, Page No.:1476-1480

made by twitter. For example, the criterion could be a region so that users' opinions from that particular region are collected and monitored; it could also be one or more predefined keywords so that opinions about some particular services can be monitored. There is also an emerging need for early crisis detection and response with such target stream. For example, a sai info-tech company is interested in automatically discovering any new named entities in a targeted stream it creates for the company and its products, By doing this, the company is able to acquire first-hand information about the crisis and make early response. Such applications require a good named entity recognition (NER) system for Twitter, which is the focus of this paper. To extract information from this large volume of tweets generated by Twitter's millions of users, Named Entity Recognition (NER), which is the focus of this work, is already being used by researchers. NER can be basically defined as identifying and categorizing certain type of data in a certain type of text.

2. Related Work

Finding Named Entities In this study, the idea of segmenting a tweet text into a set of phrases, each of which appears more than random occurrence is adopted. Therefore, a corpus serving this purpose in Turkish is needed. To this aim, TS Corpus, which indexes Wikipedia articles and also Tweets, is used. In the proposed solution, TS Corpus is used for gathering statistical information for various segmentation combinations by means of a dynamic programming algorithm. While collecting statistical information for segment combinations, tweet collection of TS Corpus is also used while computing probability of a segment is a valid named entity, which is different from the previous studies. The knowledge base that is constructed using Turkish Wikipedia dump is used to validate the candidate named entities.

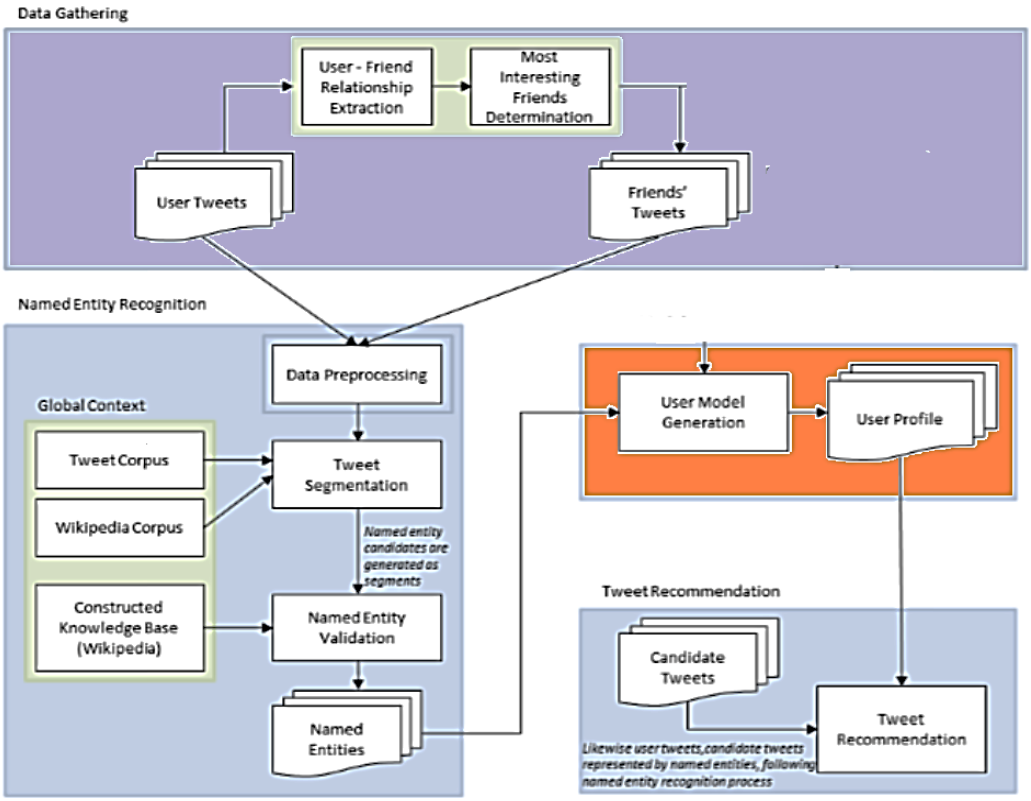

Figure 1: Model architecture of Tweets recording

ISSN: 2347-1697
International Journal of Informative & Futuristic Research (IJIFR)
Volume - 3, Issue -4, December 2015
Continuous 28th Edition, Page No.:1476-1480

Segmentation constitutes the core part of named entity recognition method. The aim here is to split a tweet into consecutive segments. Each segment contains at least one word. For the optimal segmentation, the following objective function is used, where F is the stickiness function, t is an individual tweet, and s represents a segment.

arg max s1...sn

F(t)=n X i=1

F(si) ……(1)

Although the term stickiness is generally used for expressing tendency of a user to stay longer on a web page by a user, Li et. al defined it as the metric of a word group to be seen together in documents frequently, or not and it is instead of generating all possible segmentations and compute their stickiness, dynamic programming algorithm described in is adapted to this study to compute stickiness values evidently. The algorithm basically segments the longer segment, which can be tweet itself, into two segments and evaluates the stickiness of the resultant segments recursively. More formally, given any segment

s = w1w2...wn ,

adjacent binary segmentations

s1 = w1...wj and s2 = wj +1...wn i

s obtained by satisfying the following equation.

Arg max s1,s2

F(s)=F(s1)+F(s2) ……(2)

Thus far, tweets are segmented making use of the stickiness function. In the result of this phase, tweet segments, which are candidate named entities, are obtained. These candidate named entities have to be validated whether they are real named entities or not, so that they can be used as an indicator of the user's interest. For this purpose, as explained before, Wikipedia is chosen as a reference for a segment to be a named entity, and a graph-based knowledge based on Wikipedia is constructed.

Tweet Recommendation

To determine whether a tweet is interesting or not is achieved by comparing NE representation of the tweet with the generated user interest model. This comparison results in a ranking of candidate tweets. As the first step, candidate tweets are processed to obtain their NE representations. C represents the frequency count of a named entity for a user, n represents the count of friends included in the user interest model, RR represents the relative ranking score of a followed, and U represents the user himself. With the same approach, the final score of all of the named entities appearing in the user interest model is calculated.

SCNE = n X i=1

RRi ·Ci + RRU ·CU

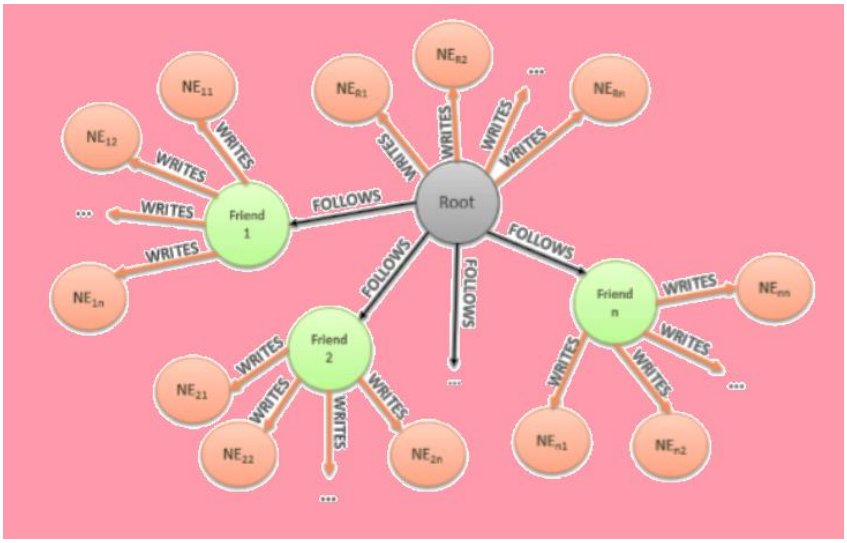

Figure 2 : Records of Tweets

ISSN: 2347-1697
International Journal of Informative & Futuristic Research (IJIFR)
Volume - 3, Issue -4, December 2015
Continuous 28th Edition, Page No.:1476-1480

3. Experimental Exercise

To evaluate the system from recommendation point of view, two types of datasets as candidate tweets for recommendation and two types of user groups to recommend tweets are formed. The first dataset of candidate tweets, GNRL, is a general dataset containing 100 tweets crawled from news- papers' Twitter accounts. The second dataset, PSNL is a personal dataset containing 100 tweets that are crawled from the followers of followers of the selected users. There are 10 users volunteered for this experiment where half of them are active Twitter users, whereas the other half is inactive Twitter users. Active Users are the users that use Twitter frequently, have retweeting and mentioning habits, and update followed list when necessary where Inactive Users do not post, retweet, or mention often, and do not update followers list frequently. Volunteered users are categorized on the basis of the information they provided about their Twitter usage habits.

For each user, user interest model is constructed under SCP measure on Wikipedia Corpus along with length normalization for stickiness function, which gives the best results ac- cording to the validation experiments. In addition, the best NT and NF values are experimentally obtained, therefore 20 followers and 10 tweets of each followed are included in the model.

Table 1: Outcomes after experiment

		Classification Acc %		Ranking Acc %	
		GNRL		PSNL	
Active users	User 1	68	64	0.778	0.907
	User 2	66	61	0.698	0.766
	User 3	62	57	0.760	0.781
	User 4	71	71	0.720	0.814
	User 5	71	66	0.601	0.678
Average AU		67.79	63.45	0.713	0.790
Inactive Users	User 6	74	48	0.521	0.613
	User 7	42	38	0.572	0.651
	User 8	39	37	0.432	0.471
	User 9	43	37	0.321	0.301
	User 10	49	48	0.566	0.514
Average IU		43.41		41.73	
Average Overall		54.09		0.623	

4. Conclusions And Future Work

In this paper we conclude that twitter is an now growing brand of social media as also a new type of social media, has attracted great interests from both industry and youth ie. students. Many private and public organizations are also using twitter. We present the HybridSeg framework which segments tweets into phrases called special segments using context. Tweet segmentation helps to preserve the same meaning of tweets, which sequent benefits many stream applications, e.g. named entity recognition. We identify two directions for our future research.

5. References

[1] "Twitter Company Info". Twitter. February 6, 2015. Retrieved May 2, 2014.

[2] Humble, Charles (July 4, 2011). "Twitter Shifting More Code to JVM, Citing Performance and Encapsulation As Primary Drivers". InfoQ. Retrieved January 15, 2013.
[3] Gomes, Lee (June 7, 2014). "Twitter Search Is Now 3x Faster". Blogger.[dead link]
[4] "Twitter.com Site Info". Alexa Internet. Retrieved December 12, 2015.
[5] "Twitter MAU Were 302M For Q1, Up 18% YoY - Twitter (NYSE:TWTR) | Benzinga]". April 28, 2015. Retrieved May 2, 2015.
[6] Arrington, Michael (July 15, 2006). "Odeo Releases Twttr". TechCrunch. AOL. Retrieved September 18, 2010.
[7] "Twitter via SMS FAQ" Retrieved April 13, 2012.
[8] "About Twitter" Retrieved April 24, 2014.
[9] Twitter (March 21, 2012). "Twitter turns six". Twitter.
[10] "Twitter Passed 500M Users In June 2012, 140M Of Them In US; Jakarta 'Biggest Tweeting' City". TechCrunch. July 30, 2012.
[11] Twitter Search Team (May 31, 2011). "The Engineering Behind Twitter's New Search Experience". Twitter Engineering Blog. Twitter. Retrieved June 7, 2014.
[12] "Twitter turns six" Twitter.com, March 21, 2012. Retrieved December 18, 2012.
[13] "Top Sites". Alexa Internet. Retrieved May 13, 2013.
[14] D'Monte, Leslie (April 29, 2009). "Swine Flu's Tweet Tweet Causes Online Flutter". Business Standard. Retrieved February 4, 2011. Also known as the 'SMS of the internet', Twitter is a free social networking service
[15] Broder, Andrei Z.; Glassman, Steven C.; Manasse, Mark S.; Zweig, Geoffrey (1997). "Syntactic clustering of the web". Computer Networks and ISDN Systems **29** (8): 1157–1166. doi:10.1016/s0169-7552(97)00031-7.
[16] Alex Franz and Thorsten Brants (2006). "All Our N-gram are Belong to You". Google Research Blog. Retrieved 2011-12-16.
[17] Ted Dunning (1994). "Statistical Identification of Language". New Mexico State University. Technical Report MCCS 94-273
[18] Soffer, A (1997). "Image categorization using texture features". Proceedings of the Fourth International Conference on **1** (233): 237. doi:10.1109/ICDAR.1997.619847.
[19] Tomović, Andrija; Janičić, Predrag; Kešelj, Vlado (2006). "n-Gram-based classification and unsupervised hierarchical clustering of genome sequences". Computer Methods and Programs in Biomedicine **81** (2): 137–153. doi:10.1016/j.cmpb.2005.11.007.
[20] Wołk, K.; Marasek, K.; Glinkowski, W. (2015). "Telemedicine as a special case of Machine Translation". Computerized Medical Imaging and Graphics.
[21] Wołk K., Marasek K. (2014). Polish-English Speech Statistical Machine Translation Systems for the IWSLT 2014. Proceedings of the 11th International Workshop on Spoken Language Translation. Tahoe Lake, USA.
[22] Carterette, Ben; Can, Fazli (2005). "Comparing inverted files and signature files for searching a large lexicon". Information Processing and Management **41** (3): 613–633. doi:10.1016/j.ipm.2003.12.003.
[23] www.google.com
[24] [www.wikipedia.in
[25] Chenliang Li, Aixin Sun, Jianshu Weng, and Qi He "Tweet Segmentation and its Application to Named Entity Recognition" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, SUBMISSION 2013
[26] Deepayan Chakrabarti and Kunal Punera "Event Summarization using Tweets" Yahoo! Research 701 1st Avenue Sunnyvale, CA 94089
[27] Alan Ritter, Sam Clark, Mausam and Oren Etzioni "Named Entity Recognition in Tweets: An Experimental Study"Computer Science and Engineering University of Washington Seattle, WA 98125, USA
[28] Deniz Karatay & Pinar Karagoz "User Interest Modeling in Twitter with Named Entity Recognition", 5th Workshop on Making Sense of Microposts
[29] Chenliang Li, Jianshu Weng, Qi He , Yuxia Yao, Anwitaman Datta, Aixin Sun, and Bu-Sung Lee1 "TwiNER: Named Entity Recognition in Targeted Twitter Stream" SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.