A Graph Based Similarity Measure (GBSM) for Finding the Semantic Relation between the Words in Microblogs

K.Suguna
Research Scholar, Bharathiar University,
Assistant Professor, Department of Computer Applications,
Dr.N.G.P Arts and Science College, Coimbatore-641046,India.

Dr. K.Nandhini
Assistant Professor, PG & Research Department of Computer Science,
Chikkanna Government Arts College, Tirupur-641602, India.

Orcid: 0000-0003-3423-8319

Abstract-

This paper introduce a new graphics model for micro blogs to understand the mostly inference topics. The twitter deals with more than hundred billions of tweets per day, so it is not an easy task to find the recurrent topics. Twitter are the one of the leading social networking site where the people can share their arousing and opinion. The tweets are very petite text, noisy and unstructured. Tweets are constantly screening up with rich user-generated[1]. The constructs the semantic relationships to each other and also provides a way to connect the semantically related and co-occurred word. The proposed method Graph Based Similarity Measure (GBSM) define a topic model to avoid the sparseness of short text and noisy. This paper examines the problem of overlapping in nodes and edges of the Graph-Based Topic Model(HGTM).

Keywords: graph, Short text, Similarity Measure.

INTRODUCTION

Twitter is the most significant social networking service that spreads the information's worldwide. Twitter to be a rich and vast resource for data on the web. The content of the twitter should be well recognized because the people are using this as a platform to expose their reaction at instantaneous occasion. The text contents in tweets are diverse in nature, so it is necessary to predict the information which is frequently tagged. Most of the sentiment analysis is conducted on tweets with traditional algorithms. There are three key reasons to propose a new model :(i) the severe sparsity problem, (ii) models are designed for flat texts without structure,(iii)wide range of information conflicts with the assumption. The weakly-supervised information provided by s can build direct semantic relations between tweets so that the words in tweets have more complex topical relationships than in normal text.

RELATED WORKS

Notations and Definitions

A graph is an undirected graph, denoted as *G =(V,E)*, where nodes V are s from the dictionary *fhgh=1:H* and edges *E = f(h, h`)g* are obtained from co-occurrence relations between s in the explicit relationship. The edge $e_{hh`}$ is weighted based on the association weight between h and h`.

Removing of non-english character

Algorithm1:NaïveBayesclassifier (Landid.py)

```
Langid/train/Nbtrain.py

Shutil.rmtree(outdir)

Except NameError:

        Pass

Except OSError

//clean up non-english characters.

defsetup_pass_tokenizer, b_dirs,sample_count, sample_size, term_freq, line_level);
```

LDA(Latent Dirichlet Allocation) Algorithm

LDA represents documents as mixtures of topics that create words with certain probabilities. The documents are provided in such a way that,

(i) Find the number of words *N* in the document,

(ii) Find the fixed set of topics *K*,

(iii) Generate each word W_i in the document

(a) Pick a topic

(b) Generate the word using topic

(c) For each topic compute and Generate each word *wi* in the document by:

- First picking a topic (the multinomial distribution that you sampled above, for example, you might select the food topic with 1/3 probability and the cute animals topic with 2/3 probability).
- Using the topic to generate the word itself (according to the topic's multinomial distribution). For example, if "food" is selected topic, it might generate the word "broccoli" with 30% probability.
- LDA then tries to backtrack from the documents to find a set of topics that are likely to have generated the collection.
- And for each topic *t,* compute 1) *p*(topic *t* | document *d*) = the proportion of words in document d that are currently assigned to topic t, and 2) p(word *w* | topic *t*) = the proportion of assignments to topic *t* over all documents that come from this word w. Reassign *w* a new topic, where choose topic t with probability p(topic t | document *d*) * *p(word w | topic t)*
- After repeating the previous step a large number of times, you'll eventually reach a roughly steady state where your assignments are appealing in high-quality. So use these assignments to estimate the topic mixtures of each document and the words associated to each topic

BTM(Biterm Topic Model)

1. Randomly assign topic uniformly to each biterm b1.

2. Reset Topic assignment of b1.

3. Reassign topic k to biterm b1.

Algorithm 2: Gibbs sampling algorithm for BTM

```
Input: the number of topics $K$, hyperparameters$\alpha, \beta$,
biterm set $B$
Output: multinomial parameter $\boldsymbol{\varphi}$ and $\boldsymbol{\theta}$
    initialize topic assignments randomly for all the biterms
        for $iter = 1$ to $Niter$ do
            for $b \in B$ do
                draw$z_b$from $P(z|\mathbf{z}-b,B, \alpha, \beta)$
                update$n_z$, $n_{wi}$
```

GBSM (Graph Based Similarity Measure) Algorithm over HGTM.

The generative process for HGTMis given by the following

Algorithm 3: Graph Based Similarity Measure Algorithm

```
1 T, a, b, t are predefined
2 For each of the s h = 1 : H, draw $Ø_h$_ Dir(à)
3 For each of the topics t = 1 : T, draw $Ø_t$ ~Dir(ß)
4 For each of the documents d =1 : D, draw its length $N_d$,
  given a  set $h_d$ referred to the document d
5 For each word $w_{di}$, i= 1 : $N_d$
    1)  draw an initial hash tag assignment $y^1_{di\sim}$  Uni($h_d$)
    2)  draw r ~ Bern(P)
    3)  if r =1, draw a hash tag assignment $y^1_{di}$ = y1$_{di}$,
    4)  if r =0,draw a hash tag assignment $qy^1_{di}$
        ~Multi(norm(gy1 $_{di}$ )
    4)  draw a topic assignment
    5)  draw a word assignment
```

EXPERIMENTAL RESULTS

A probabilistic generative model describes the process of generating a semi-structured tweet collection with weakly-supervised information from graphs.

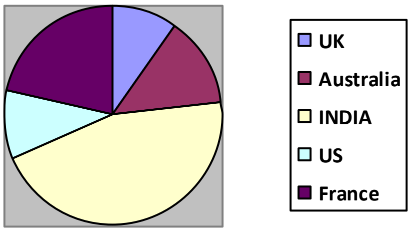
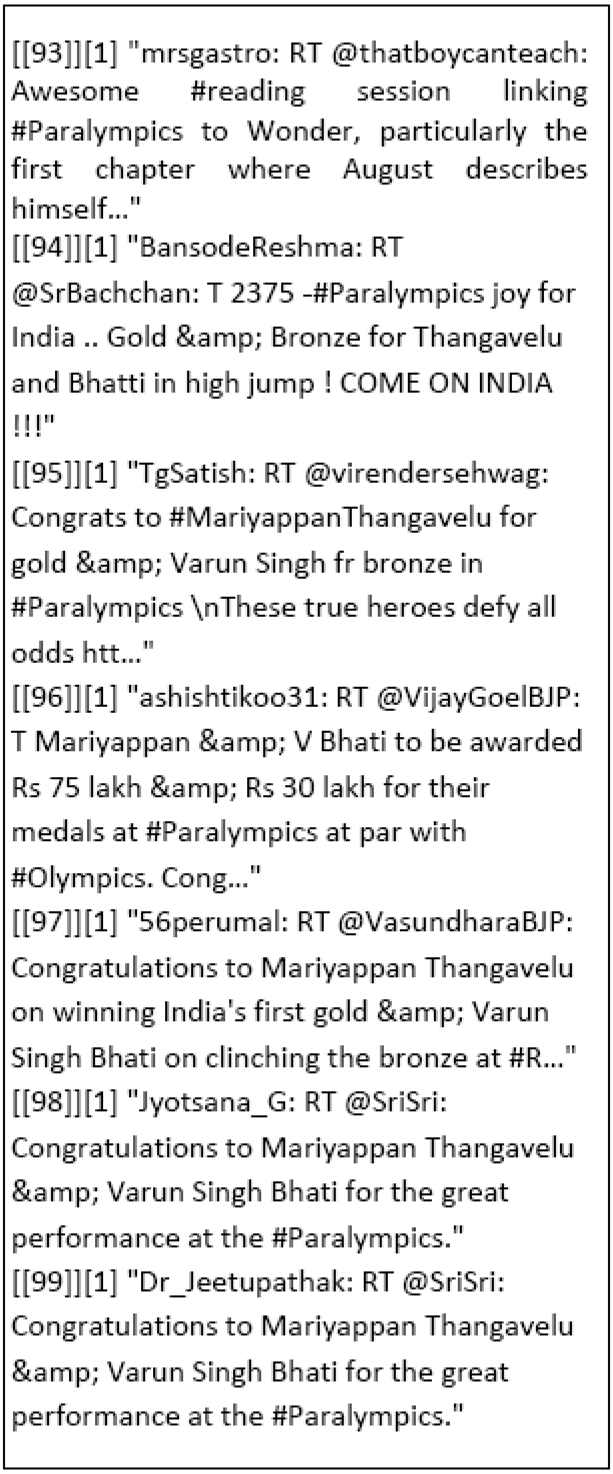


Figure 4.1: Piechart for country wise representation of tweets



[[93]][1] "mrsgastro: RT @thatboycanteach: Awesome #reading session linking #Paralympics to Wonder, particularly the first chapter where August describes himself..."
[[94]][1] "BansodeReshma: RT @SrBachchan: T 2375 -#Paralympics joy for India .. Gold &amp; Bronze for Thangavelu and Bhatti in high jump ! COME ON INDIA !!!"
[[95]][1] "TgSatish: RT @virendersehwag: Congrats to #MariyappanThangavelu for gold &amp; Varun Singh fr bronze in #Paralympics \nThese true heroes defy all odds htt..."
[[96]][1] "ashishtikoo31: RT @VijayGoelBJP: T Mariyappan &amp; V Bhati to be awarded Rs 75 lakh &amp; Rs 30 lakh for their medals at #Paralympics at par with #Olympics. Cong..."
[[97]][1] "56perumal: RT @VasundharaBJP: Congratulations to Mariyappan Thangavelu on winning India's first gold &amp; Varun Singh Bhati on clinching the bronze at #R..."
[[98]][1] "Jyotsana_G: RT @SriSri: Congratulations to Mariyappan Thangavelu &amp; Varun Singh Bhati for the great performance at the #Paralympics."
[[99]][1] "Dr_Jeetupathak: RT @SriSri: Congratulations to Mariyappan Thangavelu &amp; Varun Singh Bhati for the great performance at the #Paralympics."

Figure 4.2: sample data collected from tweets on Paralympics.

CONCLUSIONS

This paper presents GBSM that describe the hash tag relation graphs as weakly-supervised information for tweet semantic modeling. It demonstrates that hash tag graphs contain reliable information to bridge semantically-related words in sparse short texts. GBSM can enhance semantic relations between tweets. And also reduces the noise to compare with the HGTM and produces the hyper parameters.

REFERENCES

[1] Yuan Wang, Jie Liu, Yalou Huang, and Xia Feng,” Using Hashtag Graph-Based Topic Modelto Connect Semantically-Related Words Without Co-Occurrence in Microblogs” IEEE Transactions On Knowledge And Data Engineering, VOL. 28, NO. 7, JULY 2016.

[2] Vijayan and Jayasudha j. Greeshmas, A survey on web pre-fetching and web caching techniques in a mobile environment, cs& it-cscp 2012

[3] Ashok Kumar D.Loraine Charlet Annie M.C., “web log mining using K-Apriori Algorithm”, volume 41, March -2012.

[4] Indlakasthuri , M.A. Ranjit Kumar K. SudheerBabu,Dr.S.SaiSatyanarayana Reddy, An Advance Testimony for Weblog Prefetching Data Mining,IJARCSSE, 2012.

[5] Harish Kumar and Anil Kumar,”Clustering Algorithm Employ in Web Usage Mining: An Overview”, INDIA Com publication, Edition 2011.

[6] B.Santhosh Kumar, K.V. Rukmani,” Implementation of Web Usage Mining Using Apriori and FP-Growth Algorithms”, volume: 01, Issue: 06, Pages: 400-404(2010).

[7] M. Khattak, A. M. Khan, sungyoung lee*, andyoung-koo lee, Analyzing Association Rule Mining and Clustering on sales day Data with XLMiner and Weka

[8] RajanChattamvelli, “Data Mining Methods”, Narosa publications, Edition 2009.

[9] Jiawei Han, Ian Pei, Yiwen Tin, Runying Mao, “Mining Frequent Pattern without Candidate Generation: A Frequent Pattern Tree Approach”, Volume-8.

[10] Jian Pei, Jiawei Han, BehzadMortazavi-asl, Hua Zhu, “Mining Access Pattern Efficient from Web Logs”

[11] J.Han and Kamber,”Data Mining: Concepts and Techniques”, Morgan Kaufman Publishers, 2000.

[12] B. Gomathi,sakthivel”Implementing Fusion to Improve the Efficiency of Information Retrieval Using Clustering and Map Reduction”springer ,2016

[13] Chen Wang, Mingsheng Hong, Jian Pei, Haofeng Zhou, Wei Wang and Baile Shi ,“Efficient Pattern Growth Method For Frequent Tree Pattern Mining”,Springer publication,2002.

[14] Jiaweihan, “Mining Frequent Pattern Without Candidate Generation:A Frequent-Pattern Tree Approach, Springer Publication,2004.

[15] Cristobal Romero, Sebastian Ventura, Amelia Zafra, Paul de Bra,Applying web usage mining for personalizing hyperlinks in web based adaptive educational systems-,Elsevier-2009.

[16] Virtual education environments and web mining - TuncaySevindik, NecmiDemirkeser, Zafer Comert,Elsevier-2010

[17] Graph based new approach for frequent pattern mining-AnuragChoubey, Ravindra Patel, J.L.Rana-IJCSIT vol-4,issue 1-2012.

[18] Attias,” A variational Bayesian framework for graphical models”.In *Advances in Neural Information Processing Systems 12*, 2000.

[19] Marco Lui and Timothy Baldwin“langid.py: An Off-the-shelf Language Identification Tool”, Republic of Korea, 8-14 July 2012