Grouping of Hashtags using Co-relating the Occurrence in Microblogs

Ambresh Bhadra Shetty
Assistant Professor,
Dept. of Studies in Computer Applications (MCA),
Visvesvaraya Technological University,
Centre for PG studies, Kalaburagi
ambresh.bhadrashetty@gmail.com

Megharani S Ambalgi
Student, MCA VI Semester,
Dept. of Studies in Computer Applications (MCA),
Visvesvaraya Technological University,
Centre for PG studies, Kalaburagi
megharanis10694@gmail.com

Abstract: In this paper we introduce a new topic model to understand the chaotic micro blogging environment by using hashtag graphs. Hash tag is symbol (#) used on social media and micro blogging sites. A word or phrase is preceded with hashtag;hashtag isgenerally used forhighlighting the topic.HGTM: Hash tag Graph based Topic model exploresan elemental proposal that uses the aspects which are suppressed by means of the hash tags that are added through the user HGTM discovers word semantic relations even if words are not correlated within aspecific tweet, hash tags are used as keywords to find the semantic relations. HGTM has the capability to Handel the sparseness and noise problem in tweets.

Keywords:*Micro blog, HGTM, LDA, Sparseness of short text*

*****

I. INTRODUCTION

Tweeter is one of the area of micro blogging Jack Dorsey, Noah Glass,Biz Stone, EvanWilliams are the founder of tweeter, tweeter is launched in July 15 2006 written in Java,JavaScript, Scala, Ruby, Tweeter had more than 319 million monthly active user, tweeter is Famous for its short text messages the length of text should not be more than 140 characters tweeter is founded in march 21 2006 in tweeter the one who get registered can post tweets,[1] in tweeter people discuss about current activities. The hash-tags are the kind of subject matter signs.

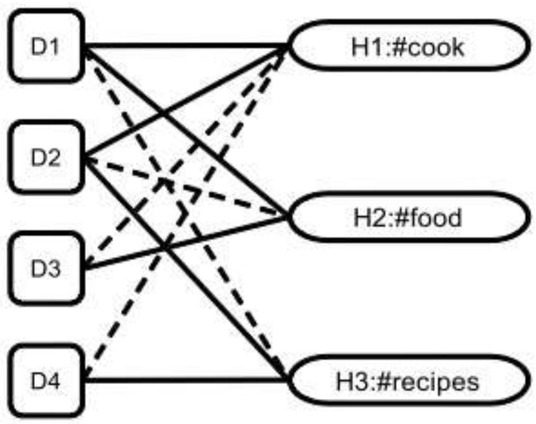

Fig 1: Explicit relationship.

There are two types of relations [1] are there one is explicit relationship and the other one is potential relationship. In the fig 1 the red lines shows the co-occurrencerelationship.Taking an example, where the users are into debate on the subject as regards to the cricket world-cup for the year 2014 were they include the tweets like hash tags as "#Mexico Vs Brazil" in D1 , D2, D3 tweet in that order; [4]in which the tweets are connected through the semantic link that has the hash-tag in analogous form. Now when "#Mexico Vs Brazil" and "#Ochoa" co-occur with the particular tweet D2 and D1 then this precise co-occurrence point out the related area under discussion upon the tweets surrounded by one among the two hash tags.

II. LITERATURE SURVEY

In the year 2010, D.Ramage,D.Liebling made a study on micro blogs.[2]the problem is still primarily focused on their social graph .it present a supervised learning model that maps the content of tweeter feed to subscribe [2]It argues that best representation of textual content on tweeter ,improving methods for following new user and topic .

In the year 2010 S.vieweg, A.L.Hughes, starbrid and L.palen he made a study on Analyzing micro blogging post [1]improving situational awareness in emergency situation through automatic methods.

In the year 2013 ,s.li and R.Pan made a study on latent Dirichlet allocation ,[3]it discover the statistical distribution of the topic model[3] this analyze a probabilistic approach for mining semi structured documents.

ISSN: 2321-8169
687 – 689

III.PROBLEM DEFINITION

If the user wants to find information related to his topic in tweeter then without HGTM it is difficult,because tweeter contains huge collections of tweets.Characterizing the contents of documents is a major problem addressed in informational retrieval and statistical natural language processing tweets contain limited words furthermore, the usage of informational language enlarge the size of dictionary.

The below diagram is for graphical model representation

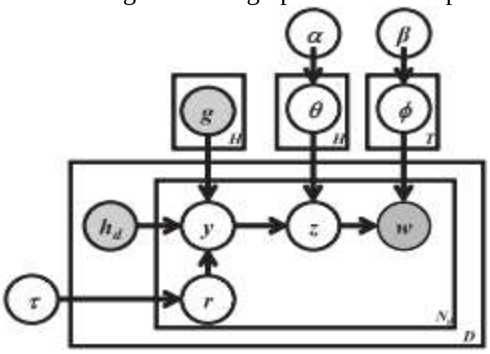

Fig 1: Graphical model of Hgtm.

Where y indicates the tag assignment of current word and alpha and beta are hyper parameters.

V. ARCHITECTURE DIAGRAM

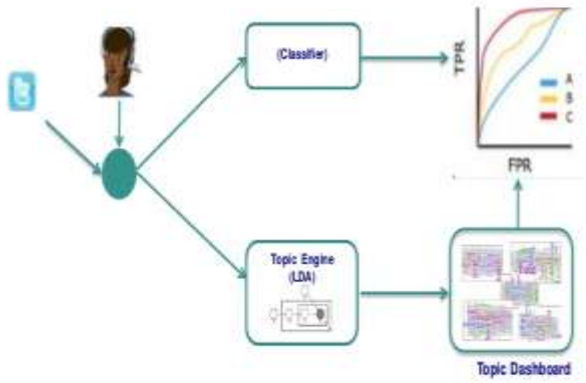

Figure 2: Architecture diagram

In the above diagram first user has to post the tweets including hashtag. Then the related hash tags are classified using latent Dirichlet allocation. After that those title are at the topic dashboard.

Following are the examples

Topics	Hashtag
IDIOMS (126):	#ihate, #cantcandidateyou, #followback
POLITICAL (39)	#Jan25, #tcot, #glennbeck, #obama, #hcr
TECHNOLOGY (57)	#nikeplus, #teamautism, #amwriting
SPORTS (42)	#golf, #yankees, #nhl, #cricket, #lakers
MOVIES (32)	#lost, #glennbeck, #bones, #newmoon
CELEBRITY (4)	#mj, #brazilwantsjb, #regis, #iwantpeterfacinelli
Books(3)	#bio ,#chemistry ,#maths

Figures 3: Table that contain the hashtag and the topics of those.

VI. IMPLEMENTATION

Below is the Gibbs sampling algorithm

Input: topic number T, hashtag graph G, iteration times NN, a, b, t, word sequence w, hashtag sequence h; Output: Q, f; Initialization: randomly initialize the hashtag assignments y and topic assignments z for all words;

1: for ii=1: NN do

2: for d= 1: Nd do

3: for I =1: Nd do

4: Draw Yd (I) ~ Uni (h)

5: Draw r ~Bern (α)

6: if r=1 then

7: yd= yd (I)

8: else

9: Draw yd (I) ~ Multi (norm (g(y (I))

10: end if

11: Draw a topic z (di) ~Multi (∑ydi)

12: Update c (Z)

13: end for

14: end for

15: Calculate Θ as, equation (9)

16: End For

In the first step of the Gibbs sampling algorithm three variables is predefined values [4] alpha, beta. Hashtag h=1 and for each topic t=1, for document d=1.then we need to draw its length later make initial hash tag assignment. From the matrix if r=0 then it is potential related if It r=1then it is explicit relationship.

ISSN: 2321-8169
687 – 689

VII. RESULTS

The expected results of the projects are given below:

Figure 4: Search based on Hash tags

This results we get Figure(4) after entering the keywords hash tag, which user wanted to know the related messages and he can read those tweets Figure(4) and as well user can also have the option to rate the tweets.

Figure 5: Semantic Relationship

This output shows the semanticcoo relationship between similar tweets Figure(5)here the user come to know that how much his keyword hash tag is related with other tweets.

VIII. CONCLUSION

This work argues that better representations of textual content on Twitter. A topic novel model to connect semantically related words in the micro blogs. At earliest period the hash tag tables are initiated in the proposed arrangement as the inadequately organized particulars with the intention of the effectual forming of the tweet depending upon the lexical that will embrace mutually the sparseness plus the noise sort of tweets as well.

REFERENCE

[1] D.Ramage,S.Dumais,D.Liebling,"Characterizing micro blogs with topic models,," in Proc. Int. AAAI Conf. Weblogs Soc. Media, 2010, vol. 5, no. 4, pp. 130–137.

[2] Y Chinch Amir and Ts Chua "Emerging topic detection for organizations from micro blogs "in proc.36th Int., 2013, pp.43-42

[3] S. Li, J. Li, and R. Pan, "Tag-weighted topic model for mining semi-structured documents," in Proc. 23rd Int. Joint Conf. Artif. 2013, pp. 2855–2861.

[4] Y. Wang, J. Liu, J. Qu, Y. Huang, J. Chen, and X. Feng, "Hashtag graph based topic model for tweet mining," in Proc. Int. Conf. Data Mining, Dec. 2014, pp. 1025–1030.

[5] D. M. Romero, B. Meeder, and J. Kleinberg, "Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter," in Proc. 20th Int. Conf. World Wide Web, 2011, pp. 695–704.

[6] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in Proc. Human Language Technol.: Linguistics, 2010, pp. 100–108.

[7] D. Bile and j Lafferty "Correlated topic model", in pro neural Sits, 2006, vol.18, p.147

[8] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled LDA: A supervised topic model for credit attribution in multilabeled corpora," in Proc. Conf. Empirical Methods Natural Language Process.: Volume 1, 2009, pp. 248–256. [24] D. Ramage, S. Dumais, and D. Liebli