International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 07 | July -2017 www.irjet.net p-ISSN: 2395-0072

Mining query log to suggest competitive keyphrases for sponsored search via improved topic model using ITCK method

Yeetika Dhingra[1], Dr. R.K. Chauhan[2]

[1]M.Tech Scholar, Dept. Of Computer Science and Applications, Kurukshetra University, Haryana, India
[2]Professor, Dept. Of Computer Science and Applications, Kurukshetra University, Haryana, India

---

Abstract - *The study has introduced the concept of query log which maintains the information regarding the user intent. Mining search log is a popular task in suggesting long tail keyphrases for the advertisers bid on specific keywords in search engine auction process to place their advertisements on search engine result page. The sponsored result is generated considering the keywords typed by the user. Here the keyphrases are derived specifying the seed keyword, based on the methodology of hidden topic retrieval using the improved topic model. ITCK method is proposed to provide topic modeling based suggested key phrases. The experiment is being performed on AOL search engine query log. The experiments have been conducted and it has been proved that the proposed work performs better than existing one.*

Key Words: Sponsored search, Query log, Topic modeling, Keyword generation, LDA.

1.INTRODUCTION

Sponsored search is the biggest key factor in terms of generating revenues using potential customers. When a user types query on search engine, two results are mainly drawn: organic and sponsored results. Both results work very differently as the organic result listing is based on retrieval of valuable information but sponsored searching is based on the auction process conducted by the search engine. In this process, the bidding mechanism is performed by which advertisers bid on specific keywords so that it matches with the query posted by user through search engine and some candidate advertisements are selected. The ad with highest bidding will be displayed and the advertisers meant to be paid highly.

For keyword generation in sponsored search, the process of mining query log is used. Search Log mining is a data mining process which aims at extracting useful information for different user behavior models. Query log is basically a file which is maintained by the search engine server. The log file typically consists of a record related to the query requested by user and the results delivered by the search engine. It is used to draw a relationship between user and the search engine. Mining search log is a fast emerging trend which is applied in different areas of information storage and retrieval. Basically it is used to extract the intention behind the user requested query. It is a kind of process used to extract user behavior which can be applied to various platforms [1] [2]. The focus here is to establish a relationship between what user searches and what needs to be relevant. Query log is defined as a set, containing Qi = {query, count} where, query refers to the keywords submitted by the user in search engine query box and count refers to searched volume related to the query.

A keyword suggestion method has been proposed which is applied in context of seed key terms which are selected randomly. Seed terms are short and ambiguous and therefore these are categorized topically. The keyword suggestion method is a great support to search engine advertising in which the advertisers bid on these expanded forms and generate revenues. It is used to target potential customers. Suppose we use the seed term ''Colgate'', there are many keywords which co-exists with the seed term in the query log therefore, a co-occurrence relationship is maintained to generate candidate keywords related to seed key term. The methods like synonym based approach; conceptual graph construction and concept hierarchy are previously used for keyword generation [3][4][5]. These methods lack in generating long tail keyphrases for the problem. Therefore, ITCK method is proposed to overcome the failures in existing approaches. ITCK is an improved topic based competitive keyphrase suggestion method. The keyword suggestion model that has been designed, is based on a machine learning approach. The model consists, a three step procedure: candidate keyword generation using association rule mining, topic modeling approach (LDA) and improvement over LDA approach. The proposed method is based on graphical model which is used to develop a correlation between the seed and its candidates. The query log is represented in Table1:

Table -1: Query log representation

Query	Count
q.kw	q.vol
Colgate	36
University Colgate	16
Pamolive Colgate	20
Teeth colgate	8

Here q.kw is query keyword and q.vol is the searched volume related to query.

2. PROPOSED METHOD

In this section, an improved topic modeling method has been introduced that is ITCK method which performs better than the existing method for providing keyword suggestion to the advertisers. The framework is illustrated with the help of Fig 1:

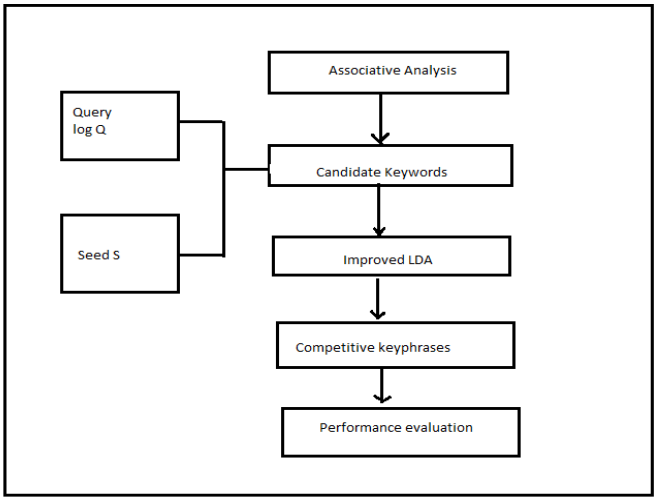

Fig 1: ITCK: improved topic based competitive keyphrase suggestion method

The keyword suggestion method used to explore seed terms which are drawn randomly from the query log to perform two major steps. One is to generate candidate keywords through associative analysis and other one is to suggest topic based competitive keyphrases using factor graph modeling.

2.1 Candidate generation

Given a query log Q where each query q is associated with two terms: q.kw which is the query keyword and the other is q.vol which refers to the searched volume associated with the keyword. There are two definitions related to the query keyword which is explained as:
1)Volume of keyword-Given a query log Q and keyword k, volume of keyword k refers to the count that how many times query occurs within the query log which is denoted by k.vol.
k.vol=aggregated volume of the queries having k

2)Associative keyword- Given a keyword k in query log within which " a " can be co-exists which is the associative word related to keyword k.

Basically in query log dataset there are more than one associative keywords related to keyword k , so a new term is introduced here, that is k.AK having all keywords related to k. For a given seed word s, an associative keyword set is represented with notation, s.AK. Furthermore, for each keyword a in s.AK, aggregate a.Ak for all a belongs to s.Ak and lastly, candidate keywords are generated which are represented using following notation:

s.cand={c1,c2,c3……..cn}

The problem is illustrated using an example, given a keyword colgate i.e k and its associatives are k.Ak. A set is formed i.e. S.Ak which consists of all associatives related to k. Now again, association analysis is performed for each word a in s.Ak and then candidates keywords like parents spokesperson are derived. To generate candidates, association rule mining is performed. The steps are explained using the following algorithm:

Algorithm 1: Candidate generation
Input: Query log dataset Q and seed word s
Output: candidate keywords i.e s.cand={c1,c2….}
1)Initialize destination list=empty
2)for each seed s
3)for i=1 to length(log)
4)if ql contains s// ql is query log
Then add to destination list
5)for testing phase initialize seed s
6)Initialize final list =empty
7)Call function f1(log ,seed)
8)Store result to s.Ak and update to corp
9)s.Ak=find_associative_keyword(Q,s)//method to find associative keyword
10)for each keyword a in s.Ak do
11)a.Ak=find_associative_keyword(Q,a)//double association
12)Aggregate a.Ak for all a belongs to s.Ak
13)s.cand are generated
For each query q' perform following steps:
14)for each q' in Q
15)if q'contains seed s
16)Add q' to query list

2.2 Topic based competitive keyphrase suggestion

1)LDA-This is the next step regarding the procedure, to apply LDA on the candidates keywords generated. The motive is to provide topic based competitive keyword suggestion. LDA is latent dirichlet allocation which is an unsupervised approach of machine learning used to derive the latent topics associated with the document or large datasets [6]. It is a generative process which works on probability distribution of words. Here for a large dataset every topic is distributed over words.

Suppose we have a dataset D associated with topics T with probability of theta and each topic is distributed over words with probability phi. So for each word in dataset D, a metric Z is sampled with multinomial distribution theta

related with dataset and a word w from multinomial distribution phi associated with topic metric z, sampled consecutively. LDA model is represented graphically in Fig 2. Here the lines represent the relationship between the variables and its two main parameters are phi and theta. The gibbs sampling is used to train the LDA model and parameters values alpha and beta i.e. 0.1 and 0.01, are set. The value of alpha and beta are known previously. And t is the no of topics. Here w is no. of distinct words. The result constitutes of three matrixes namely dt wt and z:
dt-which consists of document and no of topics ,no of times a word appear in dataset referred to topic t
wt- Used to find unique terms in the dataset
z- Form a row column where row represents no of topics and column represents no of words

Gibbs Sampling- To train the lda model , gibbs sampling is referred which is used to find the best topic for the word that appears in the document. This is being estimated with the help of following formula:
p (Z_{d,n} = t) \propto (C_t^d + \alpha) \times\frac{C_t^w + \beta}{C_t + V\beta}
here d-document
w-word appear in a document
C_t^d-no of times topic t arrived in document
C_t-no of times topic t arrived in corpus
Alpha-used for topic distribution
Beta-used for word distribution
It can be represented graphically as:

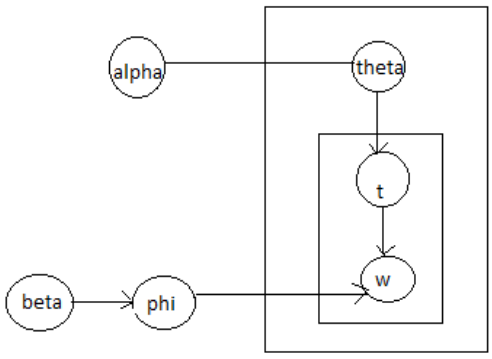

Fig 2: LDA on query log

Algorithm 2: Basic LDA (TCK method)
Input: candidate keywords, s.cand
Output: Topic modeling based keywords
For each candidate c,
1)initializing the parameters phi, theta and t
T*alpha;
W*beta;
2)Reading the log file to extract unique terms
Wt matrix is formed
3)Estimation method(training_LDA-gibbs sampling) // Training phase of algorithm
For each topic 1 to t
Draw word distributions (word_index)
Update phi
For each document 1 to d
Draw topic distribution (topic_index)
Update theta
end
update Z matrix i.e.z[d][n].topic=newtopic;
Where row represents no of topics and column represents no of words
4)prediction_method(testing _LDA) // Testing phase of algorithm
For each seed s testing is performed
If model contains (seed word)
Select the topic
Else
randomly select the topic

Drawbacks of TCK method and Emergence of ITCK method:

The TCK method explained in [7] increases the complexity of the task so in order to overcome the complexity of TCK and to improve the relevancy of keyword generation, the other approach i.e. ITCK which is improved version of existing one is used, in which we combine the LDA with Alias method and metropolis hasting sampling.

2)Alias LDA-It is a fast metropolis hasting algorithm which is used to allow topic sampling and it reuses the concept of alias table with many tokens.
It is divided into two parts:
1)Alias method- Alias method is used to sample from discrete probability distribution, by drawing random distribution it requires less operations. It is used to calculate the word proposal by the equation of topic probability i.e. directly proportional to topic/doc proposal + topic/word proposal. The process is divided into two parts-first is to create an alias table and other is to do sampling.
2)Metropolis hasting sampling-It is used for acceptance and rejection of the proposal. It is based on joint distribution in which all values are sampled in one succession.

Algorithm3:AliasLDA(ITCK Method)- It uses the concept of alias table used to create word proposal and Metropolis hasting which is used to accept or reject the word proposal generated by alias method.

Input: candidate keywords, s.cand
Output: Topic modeling based keywords
For each candidate c,
1)initializing the parameters phi, theta and t
T*alpha;
W*beta;
2)Reading the log file to extract unique terms
Wt matrix is formed

3)Estimation method(training_LDA) // Training phase of algorithm
For each document d
For each word w do
Proposal=flipcoin();
a)init b=aliastable(v,t) //generate alias table
b)for each iteration i=1 to n
c)Compute probability of word in a document
d)update reference distribution s
e)sample=SMH(Probablitytable,Aliastable)//stationary metropolis hasting used to draw new topic
f)end for loop
g)now call SMH(pi,s,n)
h)for x=1 to n do
i)if uniform_rand(1)<min(1,t) where t =p(r)s(i)/p(i)s(r) then// acceptance or rejection of proposal
j)i=r
k)end if loop
l)end for loop
m)return i
n)End
update Z matrix i.e.z[d][n].topic=newtopic;
4)prediction_method(testing _LDA) // Testing phase of algorithm
For each seed s testing is performed
If model contains (seed word)
Select the topic
Else randomly select the topic
After applying ITCK method on candidates the competitive keywords are generated and these competitive keywords are find in query list and the queries that contain competitive keywords are returned as competitive key phrases.

2.3 Performance Evaluation

To measure the effectiveness of the proposed work, F1 measure metric is used which is a common method used in information retrieval systems. Precision is used to measure the accuracy of the keywords that are derived using the results. Another metric which is used, is the Recall that is used to measure the power of proposed work in order to retrieve the relevant keywords. Both the method of precision and recall are biased. The F-measure is basically used to compare the ITCK (improved topic based competitive keyphrase suggestion) method with the previously proposed methods. Following measures are used to evaluate the proposed work:

1)Precision-ratio of relevant instances among retrieved instances
2)Recall-ratio of relevant instances that have retrieved over total relevant
3)F1 measure-harmonic mean of precision and recall
4)Evaluation Time-It is the time taken by algorithm to complete the prediction process. Each algorithm takes different time, and it is a measure of task or time complexity. Lesser the time complexity of algorithm, better are the results.

3 EXPERIMENTAL RESULTS

3.1Dataset

The experiment is performed on the AOL query log dataset (2006) then the dataset is being preprocessed using the filtering function to extract two terms query and counter as described in Table 1. Seed keywords are tested on the dataset using the online shopping category. The seed keywords which are tested namely:
1)Colgate
2)Laptop
3)Fanta
4)i-reader
5)Dove

3.2 Evaluation methodology

To demonstrate the effectiveness of ITCK method, criteria of relevancy is fulfilled using the threshold value which is being decided on the basis of occurrence of seed keyword in the query. The criterion is discussed with the help of following rule:
If occurrence of (seed keyword) in query>2.
Then it is said to be relevant.

3.3 Measures

To effectively measure the performance of the new derived method, three metrics are used namely:
1) Precision
2) Recall
3) F-measure
For a seed keyword s having keywords as k= {k1,k2,…kn}
Ei is the set of relevant keywords in set ki, the formula for above metrics is given below as:

Precision=Ei/Ki=relevant keyword intersection retrieved keyword/retrieved keyword

Recall=Ei/Ej=relevant keyword intersection retrieved keyword/relevant keyword

F1 measure=2 * Precision* Recall/ Precision + Recall

3.4 Comparative Results using Tables

Here, the ITCK method is proposed which is compared with existing TCK method.

1)TCK method- It is a basic method given by Dandan Qiao [7],to generate keywords for sponsored search advertising process. The method is called topic based competitive keyword suggestion which is based on a machine learning

approach. On the basis of query logs, the method used to explore the indirect associations between keywords and extracts the hidden topic information to identify the competitive keywords. It helps advertisers not only broaden the choices of keywords but also to carry out a competitive strategy for search advertising. The TCK method is based on applying LDA algorithm to the candidate keywords generated using a sample seed key term.

2)ITCK method- It is the new proposed method which is an improvement over the TCK method. It is called; improved topic based competitive keyphrase suggestion method. The method is designed to overcome the failures that exist in the basic TCK method and also to enhance the search engine auction process for sponsored search advertisements. The ITCK model is based on applying an improved LDA algorithm to the candidate keywords generated using association rule mining. This section used to conduct experiments which conclude that the proposed method performs better than the existing one. There are some failures in the TCK method, which are overcome and described as follows:

1)ITCK method is providing a better accuracy and relevancy in terms of keywords that are generated.

2)The proposed method is used to decrease the time complexity as compared to the TCK method.

3)The ITCK method is optimized by using an enhanced LDA algorithm which gives better results as compared to the TCK method. The enhanced algorithm consists of merging of two concepts: alias tables and metropolis hasting sampling technique.

4)A filtering data model is used for the query log which is used to remove all noises and disturbances that occur in the base model.

Table -2: Comparative analysis of TCK and ITCK method

Algorithm	Precision	Recall	F-measure
LDA with gibbs sampling (TCK)	0.803333333	0.879562044	0.839721254
Alias LDA (ITCK)	0.886666667	0.97080292	0.926829268

Table -3: Comparative analysis of evaluation time of TCK and ITCK method

Algorithm	Execution Time(in milli sec)
LDA with gibbs sampling(TCK)	610
Alias LDA(ITCK)	245

4 COMPARING RESULTS USING GRAPHS

1)Precision- It is the fraction of relevant instances among the retrieved instances. It is used in proposed work to measure the relevancy of the keywords that are generated using ITCK method.

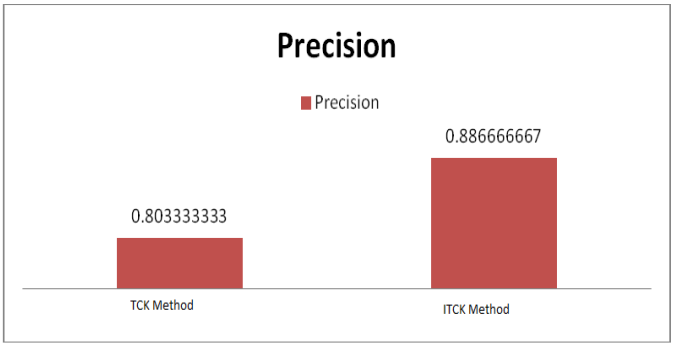

Chart -1: Precision graph

2)Recall- It is the fraction of relevant instances that have been retrieved over total instances. It is used to measure the relevancy of the keywords that are generated using proposed method.

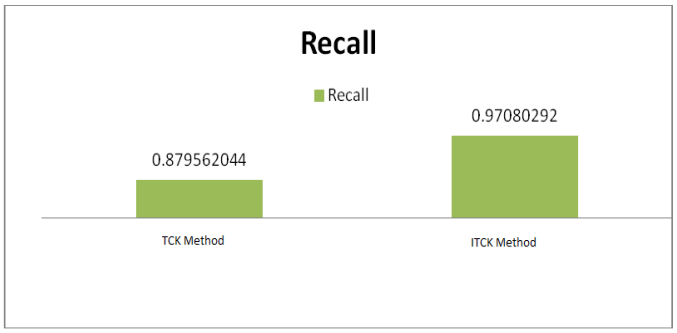

Chart -2: Recall graph

3)F-measure- It is a measure of accuracy. It is used to consider both precision and recall for the experiment to compute the score. It can be interpreted as an average of precision and recall where the F1 measure reaches its best having value 1 and worst value at 0.

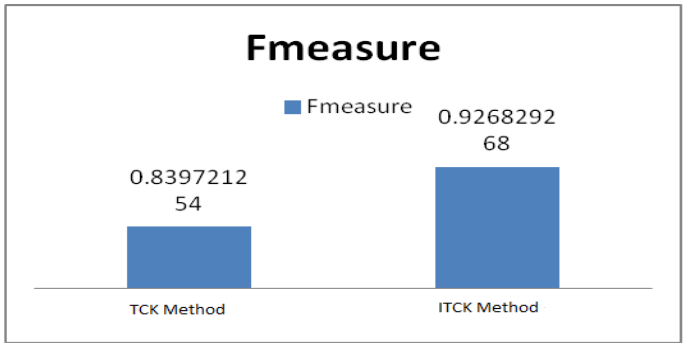

Chart -3: F1-Measure graph

5 CONCLUSION

The paper proposed a topic modeling based competitive keyphrase suggestion for sponsored search advertising. The experiments have been conducted to prove that proposed method performs better than existing one. The existing method is upgraded to reduce the sampling complexity of LDA with gibbs sampling algorithm. Furthermore, the method is used to provide more relevant and accurate keywords for the advertisers participating in auction process conducted by the search engine. It gives a new direction to the area of search engine advertising.

References

[1] J. R. Wen, J. Y. Nie and H. J Zhang," Query clustering using user logs", ACM transactions on information systems, vol. 20,no.1,2002.

[2] Z Zhang and O. Nasraoui, " Mining search engine query logs for social filtering based query recommendations", Applied soft computing,2008.

[3] L. Sarmento, P. Trezentos, J. P. Gonçalves, E. Oliveira, "Inferring local synonyms for improving keyword suggestion in an on-line advertisement system", Proceedings of the 3rd International Workshop on Data Mining and Audience Intelligence for Advertising, 2009, pp. 37–45.

[4] H. Amiri, A. AleAhmad, M. Rahgozar, F. Oroumchian "Keyword suggestion using conceptual graph construction from wikipedia rich documents", International Conference on Information and Knowledge Engineering, California, USA, 2008.

[5] Y. Chen, G. Xue, Y. Yu, "Advertising keyword suggestion based on concept hierarchy", Proceedings of the 2008 International Conference on Web Search and Data Mining, Stanford, CA, USA, 2008, pp. 251–260.

[6] David. M. Blei , Andrew y. Ng, "Latent dirichlet allocation", journal of machine learning research, 2003.

[7] D.Qiao, et al., "Finding competitive keywords from query logs to enhance search engine advertising", Proceedings of Information and Management, 2016.