An Algorithmic Query Refinement Model based on Query Classification

S. Ruban1* and S. Behin Sam2

1Research and Development Center, Bharathiar University, Coimbatore - 641046, Tamil Nadu, India; rub2kin@yahoo.com
2Department of Computer Science, Rajeswari Vedachalam Arts and Science College, Chengalpattu - 603001, Tamil Nadu, India; behinsam@gmail.com

Abstract

Objectives: This work focuses on developing an algorithmic refinement model for performing query refinement to improve the original query by adding more relevant candidate terms. The work has been tested in real time web environment and the results are also provided. Methods and Analysis: Query handling is a vibrant area in the field of Information Retrieval. Experiments reveal that formulating a query plays a vital role in generating relevant results. Since most of the users of web environment are naïve, query formulation cannot be expected to be effective always. This challenge could be overcome by the process of Query Refinement. Though there are many approaches put forward in the literature, in this work we propose a query refinement method based on the classification of queries and generating candidate terms from ontology and Thesaurus. Findings: We used the TREC 2014 web queries for evaluation and calculated Precision and Recall. The evaluation has been done on the Real time web environment which includes the most commonly used search engines like Google, Yahoo and Ask. The results have been compared and we could find a significant increase in both the precision and Recall. We did calculate the similarity measure using the Jaccard Similarity Coefficient. Since the Web environment is highly dynamic, the results tend to change depending upon the timing of execution of queries. Applications/Improvement: Addition of more relevant candidate terms helps to improve the accuracy of search applications. There has been no perfect retrieval system available till today and there is always a scope of improvement. The result of this work reveals an increase in the accuracy of the retrieval system. We tend to analyze the results further, by comparing its impact in most of the available search applications.

Keywords: Information Retrieval, Ontology, Query Classification, Query Refinement, Semantic Web

1. Introduction

The search engines are the widely used applications in the web environment, since it helps to locate information for a user information need. The user information need is normally termed as query. The ability and the way the query is treated by the search engines greatly impact the results they give the user. One of the core components of any Information retrieval system is its query handling which greatly determines the accuracy of its results. Since most of the users are naïve, query formulation cannot be expected to be effective always. Studies in query handling suggest that this problem can be overcome by transforming the initial query in a more meaningful query by adding more meaningful terms or transforming the initial query into another responsible query by replacing its initial terms by relevant terms that can generate more relevant results. The initial query is normally termed as seed query. The process of adding or removing or refining the terms in the initial query is called as Query Refinement. This can also be called as query reformulation or query transformation or query modification.

There is no perfect retrieval system available till today. A perfect retrieval system always gives the correct rel-
An Algorithmic Query Refinement Model based on Query Classification

2. Classification of Web queries

Web query classification fall under two important categories: i) Manual query classification and ii) Automatic query classification. This section splits the related work according to the focus of each work into either one of these.

2.1 Manual Query Classification

Few of the initial works in this area are based on search logs. Earlier works suggest the classification of web user queries into navigational, informational and transactional queries. Based on the survey reports, the work suggests that approximately 36% of queries were transactional in nature, 26% were navigational in nature and estimated 73% of queries were informational in nature. The above classification is based on user intent.

A further study in classification related to E-commerce suggests that web queries related to e-commerce are 12% to 24% of the various web search engine transaction logs. A work based on Searcher’s query classified search queries based on the links that were clicked by the user and the subsequent queries. Their work revealed that more than 60% of the queries were informational in nature.

2.2 Automatic Query Classification

Initial works in this area of automatic query classification attempted to classify web queries either as a homepage or as a topic. Later a study initiated further classified web search queries automatically into informational and navigational queries. The focus of commercialization in launching web queries was also studied. The authors reported that almost 38% of web queries have commercial intention. Later both supervised and unsupervised learning were used to classify web queries into informational queries and non-informational queries. The following section lists out the various works that is done in the field of Query Reformulation.

3. Query Refinement

Many works done in the fields of Query expansion uses different techniques and prove they improve the precision of the retrieval. In query expansion one of the very important issues is the source of the expansion terms. In general query expansion techniques are normally divided into i) Automatic Query expansion techniques, ii) Manual Query Expansion techniques and iii) Interactive query expansion techniques.

In the initial study conducted to expand query focused on collecting expansion terms from the associ-
ated query logs which improved the precision. Further study in this area also suggested adding candidate terms from Wikipedia which is considered as one of the largest collection of information repository in the web. A study by few researchers has explored the idea of collecting the candidate terms for expansion directly from the web.

Few works have also used some other resources like word net as a source for expansion terms. Word Net is one of the widely popular Domain Independent ontology that can be a repository for good expansion terms, but however earlier works suggested that the seed queries expanded using word Net did not much impact in the retrieval, they had a very little improvement. But, however later works revealed that word Net used in the correct sense can help in better results. In this work, we propose a hybrid but an effective way of using word Net and Thesaurus for the query reformulation. The distinguishing features that differentiate our work with other earlier works are the following:

- We classify the query based on our algorithm, and determine whether query reformulation will be beneficial or not.
- We select the terms not just from word Net but also from the Thesaurus, and those terms that occur in both word Net and Thesaurus get more weight age than the other terms for expansion.
- The similarity score between the seed terms and the candidate terms are measured using the Jaccard similarity measure. We found Jaccard similarity more beneficial. However there are also some other similarity measures that is used for the same such as Dice coefficient and cosine coefficient etc.

This work also outperforms the different already available search applications. The main contribution in this work is the use of classification of queries which determines the need of query reformulation or not. Though query expansion is always aimed to improve the results, research also shows that improper usage of candidate terms for expansion may decrease the precision as well. Though there has been some work in the area of query classification to address the way search engines treat the search queries, and is being used very much in the search engine optimization (SEO), we extend this idea by viewing the data that is already available in the transaction log, if the queries comes under the category of the Informational queries, then we suggest adding more candidate expansion terms from our proposed methodology, and our work shows the increase in the precision, whereas if the queries addressed to the system are used only for the purpose of navigation or transaction, then there is no need of adding any candidate terms, because our earlier works reveal they do not have an commendable increase in the precision.

Few more works on query reformulation used wordnet to generate the candidate expansion terms based on synonyms of the seed words available in the query. They used their methods with the CACM collections and it yielded them an improvement of over 7% over the original unexpanded queries. One more approach on Lexical resources also used wordnet to generate the candidate expansion terms for the query terms. They evaluated their method with TREC9, TREC10 and TREC12 queries and were able to get very good results compared to the other results reported earlier. Few combination based techniques that has been used to improve the retrieval effectiveness have also been proposed by different researchers over the period of time. A hybrid approach suggested used a very different approach for query expansion; they generated candidate expansion terms from different sources such as Wikipedia and clue web. Together with that they also used query log which was part of the Microsoft 2006 RFP dataset from MSN and Google N-grams which is available from Linguistic Data Consortium Catalog. Few more studies in this regard also proposed a query reformulation method which is different from the earlier studies, where they combine a distribution based method with an association based method; they used the earlier method to refine the set of candidate terms generated by the other method. Another similar strategy that was suggested used the Pseudo relevance feedback to generate the candidate terms and the refinements of these candidate terms are done using a classifier. The classifier is already trained to differentiate between the terms considered to be useful or not useful. Further study in this field used co-occurrence and distribution methods. One method is used for term selection and the other method is used for weighting.

In this present study we extend our earlier work by first proposing a new way of classification of queries, and then combining this method with our hybrid query reformulation method which involves word net and the Thesaurus thus improving the overall performance.
4. Materials and Methods

4.1 Query Classification

One of the very important problem that may arise when we add candidate terms to the seed terms (initial terms) of the user query. Though it is known, that the normal size of a web queries are short and in particular they range from 2 to 5 words. In the process of improving the query by adding more relevant terms to it, there is also a chance of adding unnecessary words that may cause the query lose its original meaning. Hence in our work, we overcome this problem by query classification where the user logs are used to do the classification. In the first section we will explain the algorithm that we have proposed.

Algorithm 4.1.1 Query Classifier

\begin{algorithm}
\caption{Query Classifier \{ AnonId, Query, QueryTime, ItemRank, clickURL, SeedQuery\}}
\begin{algorithmic}
\STATE 1. If clickURL := NULL {
\STATE 2. Query\_Sim := 0;}
\STATE 3. Else if SeedQuery := Query {
\STATE 4. If clickURL contains SeedQuery;
\STATE 5. { URLString := clickURL;}
\STATE 6. Query\_Sim := 1; }
\STATE 7. Query\_Sim := 0;
\STATE 8. If Query\_Sim := 0 {
\STATE 9. ExpandQuery := 1; }
\STATE 10. ExpandQuery :=0;
\STATE 11. } Return ExpandQuery;
\end{algorithmic}
\end{algorithm}

We use the query classifier to classify the user queries into two categories i) Informational queries and ii) Non informational queries. Our previous works\[19,20\] that were done in this regard reveal that query expansion helps to improve the informational queries whereas the queries that aimed at doing some transaction or queries that aimed to reach certain sites (Using web search engine to navigate) comes under the category of Non informational queries. Our previous study reveal that query expansion do not improve the non informational queries or does not have a significant impact in improving the precision of those queries. It was also studied that most of the Non informational queries consists of terms that are relevant to the website they want to access. We use this algorithm on an AOL Data Set\[21\]. This collection consists of 20 million web queries that were collected from almost 650000 users on a timeline of 3 months. These are real data that are used to study query reformulation.

Algorithm 4.1.2 Candidate Similarity Computation

\begin{algorithm}
\caption{TermsSim \{ Query,t,u,Onto, Thes\}}
\begin{algorithmic}
\STATE 1. Sim\_Term := Null;
\STATE 2. For each term t in Onto {
\STATE 3. For each term u in Thes {
\STATE 4. If t := u {
\STATE 5. Add t to Sim\_Term
\STATE 6.} Return Sim\_Term;
\STATE 7. For each term t in Sim\_Term {
\STATE 8. Compute Jaccard similarity(Query,t); }
\end{algorithmic}
\end{algorithm}

The main contribution of this work is the process of finding out the candidate terms. Though there has been many ways that is proposed in literature. In this work, we generate the candidate terms from two sources wordnet which is one of the widely available lexical resources which is widely known as Domain independent ontology and also the Thesaurus in our case we use the collection provided by the BigHugeLabs. We list out all the possible candidate terms from both of this resources, and then to make the user selecting the right candidate terms we compute the Jaccard similarity between the seed query and each candidate terms that is generated and lists out the details with the score to the user. Though there are many similarity measures that are available, the similarity measure using Jaccard coefficient\[22\] performs well in comparing with the different candidate terms. The Jaccard similarity coefficient is as follows:

\[
Jaccard \ sim(A,B) = \frac{P(A \cap B)}{P(A \cup B)}
\]

Being an interactive query expansion model this method helps the user makes the wise choice of candidate terms that he would wish to add with the seed query that was initially given.

Algorithm 4.1.3 RefineQuery

\begin{algorithm}
\caption{Refine Query \{ QueryClassifier, Query, TermSim\}}
\begin{algorithmic}
\STATE 1. {Query\_Weight := 0;Select – Term := 0;
\STATE 2. Query Classifier \{ AnonId, Query, QueryTime, ItemRank, clickURL, SeedQuery\};
\end{algorithmic}
\end{algorithm}
3. If Expand Query := 1;
4. \{StopWordRemoved Query := StopwordRemover(Query);\}
5. StemmedQuery := stemming(StopWordRemovedQuery);
6. For Each Term t in StemmedQuery {
7. Generate synsets for t from WordNet;
8. Onto += Synset(t);\}
9. For Each Term t in StemmedQuery {
10. Generate RelevantWords for t from Thesaurus;
11. Thes += RelevantWords(t);\}
12. Terms Sim (Query,t,u, Onto, Thes);
13. For Each Term w in Sim_Term {
14. Display “ Terms” to the User;
15. If term == “Selected” then {Select_Term += Term;};
16. Else{Discard “terms”;
17. Query_weight := StemmedQuery;\}
18. For Each Term S in SelectedQuery {
19. Query_weight += s; \}
20. Search_API (Query_weight);\}

Though generating candidate terms and place of selection of these terms holds vital importance in this query refinement process. We first identify the set of terms that are closely related to each query keyword and we calculate the jaccard coefficient value with each of these possible expansion terms with the initial seed search request. Finally in the end, those suggestions with the highest value are kept. In order to overcome any noisy suggestions we limited the number of candidate terms to be added as 3.

5. Results and Discussion

5.1 Experimental Setup
We evaluated our algorithms with set of queries, and we focused the experiments on single-user machines. The set of queries are based on TREC 2014 web queries. We tested all these queries in the real web environment. Table 1 lists out the TREC 2014 web queries that we are using our experiment.

There were a total of 50 queries in the TREC 2014 web queries. But we used the first 20 queries for our evaluation. The query length of these experimental queries ranged from 2 to 5 terms. They were subjected to stop word removal, stemming and then the parsing was done using Stanford parser, following which the relevant candidate terms for the initial keywords were generated from the ontology and the Thesaurus, which were then subjected to the similarity measure evaluation, and latter the three candidate terms with the more relevance were added to the existing query and hence the new query was formulated which will then be passed to the search API. For example the first query “Identifying Spider bites” generated the following candidate terms from the ontology and Thesaurus such as “Wanderer”, “bit”, “morse1”, “bit”, “sting”, “insect bite”, “collation”, “snack”, “pungency”, “sharpness”, “raciness”, “chomp”, “Seize with teeth”, “burn”, “Prick”. Almost 15 words were generated, of which we are going to add 3 more relevant words that may improve the precision. The jaccard coefficient value between the initial query and the term “Wanderer” is 0.651, “bit” is 0.409 and “sting” is 0.393, all other candidate terms have low values than this. Hence we refine the initial query “identifying spider bites” into the following query “wanderer bite sting identifying spider bites or identifying spider bites” In our experiments we have evaluated using the most commonly used search engines like Google, yahoo

| SL No | TREC Web Queries                  |
|-------|-----------------------------------|
| 251   | identifying spider bites          |
| 252   | history of orcas island           |
| 253   | tooth abscess                     |
| 254   | barrett’s esophagus               |
| 255   | teddy bears                       |
| 256   | patron saint of mental illness    |
| 257   | holes by louis sachar             |
| 258   | hip roof                          |
| 259   | carpenter bee                     |
| 260   | the american revolutionary        |
| 261   | folk remedies sore throat          |
| 262   | balding cure                      |
| 263   | evidence for evolution            |
| 264   | tribe formerly living in Alabama  |
| 265   | F5 tornado                        |
| 266   | symptoms of heart attack          |
| 267   | feliz navidad lyrics              |
| 268   | benefits of running               |
| 269   | marshall county schools           |
| 270   | sun tzu                           |
and Ask. Our experiments reveal adding candidate terms as already explained increases the precision value and thus increases the accuracy of the overall search results. Table 2 lists out the refined queries using the algorithms explained.

### Table 2. Refined Queries used for Evaluation

| Sl no | Refined queries                                                                 |
|-------|---------------------------------------------------------------------------------|
| 1     | Wanderer Bite sting identifying spider bites or identifying spider bites.        |
| 2     | Account story history orcas island or history orcas island.                     |
| 3     | Abscess tooth abscess or tooth abscess.                                         |
| 4     | Esophagus oesophagus gullet barrett or barrett.                                 |
| 5     | Bear deliver stand teddy bears or teddy bears.                                  |
| 6     | Sponsor supporter patron saint mental or patron saint mental.                   |
| 7     | Hollow trap hole joseph holes Louis sachar or Loles Louis sachar.               |
| 8     | Pelvis pelvis hip roof or hip roof.                                             |
| 9     | Bee carpenter bee or carpenter bee.                                             |
| 10    | Revolutionist radical American revolutionary or American revolutionary.          |
| 11    | Remedy cure repair folk remedies sore throat or folk remedies sore throat.       |
| 12    | Remedy cure therapeutic balding cure or balding cure.                           |
| 13    | Prove show tell evidence evolution or evidence evolution.                       |
| 14    | Folk federation kin tribe living Alabama or tribe living Alabama.               |
| 15    | Twister crack F5 tornado or F5 tornado.                                         |
| 16    | Attack attempt symptoms heart attack or symptoms heart attack.                  |
| 17    | Words language feliz navidad lyrics or feliz navidad lyrics.                    |
| 18    | Welfare profit gain benefits running or benefits running.                      |
| 19    | School cultivate marshal county schools or marshal county schools.              |
| 20    | Dominicus solarize sun tzu or sun tzu.                                          |

### Table 3. Initial query and no of relevant pages reported in the Web environment

| Sl no | Google | Yahoo | Ask  |
|-------|--------|-------|------|
| 1     | 86     | 62    | 28   |
| 2     | 87     | 45    | 26   |
| 3     | 92     | 88    | 35   |
| 4     | 93     | 68    | 37   |
| 5     | 83     | 45    | 6    |
| 6     | 82     | 68    | 13   |
| 7     | 84     | 81    | 36   |
| 8     | 90     | 68    | 19   |
| 9     | 89     | 95    | 40   |
| 10    | 96     | 77    | 23   |
| 11    | 91     | 80    | 25   |
| 12    | 92     | 89    | 31   |
| 13    | 90     | 75    | 28   |
| 14    | 90     | 68    | 9    |
| 15    | 87     | 82    | 19   |
| 16    | 88     | 73    | 22   |
| 17    | 86     | 87    | 24   |
| 18    | 91     | 96    | 47   |
| 19    | 94     | 80    | 20   |
| 20    | 84     | 90    | 40   |

### Table 4. Refined queries and no of relevant pages reported in the Web environment

| Sl no | Refined query in Google | Refined query in Yahoo | Refined query in Ask |
|-------|-------------------------|------------------------|---------------------|
| 1     | 93                      | 97                     | 70                  |
| 2     | 94                      | 54                     | 30                  |
| 3     | 100                     | 98                     | 60                  |
| 4     | 100                     | 70                     | 50                  |
| 5     | 90                      | 50                     | 40                  |
| 6     | 91                      | 80                     | 20                  |
| 7     | 89                      | 89                     | 42                  |
| 8     | 94                      | 70                     | 53                  |
| 9     | 100                     | 100                    | 87                  |
| 10    | 98                      | 90                     | 39                  |
| 11    | 93                      | 88                     | 60                  |

### 5.2 Experimental Results

We tested all the above queries in the real time web environment and the results are as follows. Table 3 displays the results achieved in the single user machines in the time interval of 6 months from Nov 2016 to April 2017. The
Table 4 illustrates the results of refined query in Google, Yahoo and Ask. First 100 links were evaluated for relevancy. The experiment was repeated on all the search engines.

5.3 Experimental Analysis

Normally Information Retrieval systems are evaluated using Precision and Recall measure. The performance of our system is evaluated by comparing the precision and recall with Google, Yahoo and Ask for the twenty sample queries that has been refined above. Each original query and the refined query were evaluated in all the above search engines and the results have been analyzed. Figure 1, Figure 2, Figure 3 represent the individual comparison of the precision of the 20 queries separately.

Figure 1. Comparison of Precision in Google.

Figure 2. Comparison of Precision in Yahoo.

Figure 3. Comparison of Precision in Ask.

Figure 4, Figure 5, Figure 6 represent the individual comparison of the recall of the 20 queries separately.

Figure 4. Comparison of Recall in Google.

Figure 5. Comparison of Recall in Yahoo.

Figure 6. Comparison of Recall in Ask.
The Mean Precision and the Mean Recall value are also calculated and they show a considerable impact in the accuracy. Figure 7 and Figure 8 shows the analysis of the results.

![Comparison of Mean Precision](image1)

**Figure 7.** Comparison of Mean Precision.

![Comparison of Mean Recall](image2)

**Figure 8.** Comparison of Mean Recall.

The evaluation of precision before and after query refinement revealed that in Google there is a considerable change of 6.55% and in Yahoo the change was 9.20% and in Ask the change has been significant 19.6%.

Recall is used to evaluate the coverage of the Information Retrieval system. The evaluation of Recall before and after query refinement revealed that in Google there is a considerable change of 6.80% and in Yahoo the change was 9.60% and in Ask the change has been significant 20.6%. However since the web environment is very dynamic, the results tend to change depending upon the time of execution of the query, however our experiments reveal a significant change in the accuracy of results with respect to the precision and recall for the three widely used search applications Google, Yahoo and Ask.

### 6. Conclusion

We proposed an algorithmic model to refine the web search queries using the classification of queries by automatically classifying the queries into informational queries and non-informational queries. The informational queries were further refined by suggesting and adding more candidate terms from ontology and thesaurus and further measuring the similarity using the Jaccard similarity. In this context, the paper includes the following contribution:

- A classification algorithm that classifies the queries into informational queries and non-informational queries.
- A hybrid approach to Generate candidate terms from ontology and thesaurus.
- Computing similarity measures using the Jaccard similarity measure.
- A thorough real time analysis in the web environment with the widely used search engines.

We are currently performing investigations to use other similarity measures, and improving the precision of the individual query.

### 7. References

1. Broder A. A Taxonomy of Web search. Proceedings in the SIGIR Forum. 2002; 36(2):3–10. Crossref
2. Spink A, Jansen BJ. Web search: Public searching of the Web. New York. 2004; 42(1):248–63.
3. Rose DE, Levinson D. In Understanding user goals in Web search. Proceedings in the World Wide Web Conference, New York, NY, USA. 2004. p. 13–9. Crossref
4. Kang I, Kim G. In Query type classification for Web document retrieval. Proceedings in the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada. 2004 Aug. p. 64–71.
5. Lee U, Liu Z, Cho J. In Automatic identification of user goals in Web search. Proceedings in the World Wide Web Conference, Chiba, Japan. 2005. p. 391–401. Crossref
6. Dai HK, Nie Z, Wang L, Zhao L, Wen JR, Li Y. In Detecting online commercial intention (OCI). Proceedings in World Wide Web Conference, Edinburgh, Scotland. 2006. p. 829–37. PMID:16835901
7. Baeza-Yates R, Calderon-Benavides L, Gonzalez C. In The intention behind Web queries. Proceedings in String Processing and Information Retrieval, Glasgow, Scotland. 2006 Oct; 98–109.
8. Cui H, Wen JR, Nie JY, Ma WY. Probabilistic query expansion using query logs. Proceedings of the 11th International Conference on World Wide Web, New York, NY, USA: ACM. 2002. p. 325–32. Crossref
9. Li Y, Luk WPR, Ho KSE, Chung FLK. Improving weak ad-hoc queries using wikipedia as external corpus. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA: ACM. 2007. p. 797–8.
10. Bendersky M, Metzler D, Croft WB. Parameterized concept weighting in verbose queries. Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2011. p. 605–14. Crossref
11. Voorhees EM. Query expansion using lexical-semantic relations. Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1994. p. 61–9. Crossref
12. Fang H. Re-examination of query expansion using lexical resources. Proceedings of ACL-08: HLT. 2008; 139–47.
13. Zhang J, Deng B, Li X. Concept based query expansion using word net. Proceedings of the International e-Conference on Advanced Science and Technology. 2009. p. 52–5.
14. Liu S, Liu F, Yu C, Meng W. An effective approach to document retrieval via utilizing word net and recognizing phrases. Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2004. p. 266–72.
15. Bendersky M, Metzler D, Croft WB. Effective query formulation with multiple information sources. Proceedings of the fifth ACM International Conference on Web search and data mining. 2012. p. 443–52. Crossref
16. Pal D, Mitra M, Datta K. Query expansion using term distribution and term association. CoRR, abs/1303.0667. 2013; 19.
17. Cao G, Nie JY, Gao J, Robertson S. Selecting good expansion terms for pseudo-relevance feedback. Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2008. p. 243–50. Crossref
18. Pérez-Aguera JR, Araujo L. Comparing and combining methods for automatic query expansion. Proceedings of CICLING 2008, Advances in Natural Language Processing and Applications. 2010; 33:177–88.
19. Ruban, Sam B, Kumar R. Determining the Impact of Domain Independent Ontology over Transactional user Queries. Indian Journal of Science and Technology. 2015 Sep; 8(24):1–6. Crossref
20. Ruban S, Sam B. A Comparative Study and Performance Analysis of Wordnet Over Transactional and Informational Queries. International Journal of Applied Engineering Research. 2015; 10(70):1–6.
21. Pass G, Chowdhury A, Torgeson C. A Picture of Search the First International Conference on Scalable Information Systems, Hong Kong. 2006 Jun; 75(1):29–41.
22. Suphakit N. Using of Jaccard coefficient for keywords similarity. Proceedings of the International Multi Conference of Engineers and Computer Scientists, Hong Kong. 2013; 1:1–5.