International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-2 Issue-6, August 2013

A Relative Study on Search Results Clustering Algorithms - K-means, Suffix Tree and LINGO

R.Mahalakshmi, V.Lakshmi Praba

Abstract- The performance of the web search engines could be improved by properly clustering the search result documents.. Most of the users are not able to give the appropriate query to get what exactly they wanted to retrieve. So the search engine will retrieve a massive list of data , which are ranked by the page rank algorithm(7) or relevancy algorithm or human judgment algorithm. The user will always find himself with the unrelated information related to the search due to the ambiguity in the query by the user. Evaluating the performance of a clustering algorithm is not as trivial as counting the number of errors or the precision and recall of a supervised classification algorithm In this paper a comparative analysis is done on three common search results of clustering algorithms to study the performance enhancement in the web search engine. If we effectively organize the web documents through the proper means of clustering techniques, we could definitely increase the performance of the search engines ..

A systematic evaluation of the three clustering algorithms viz., Suffix tree clustering Lingo, and K-Means using multiple test collections and evaluation measures . It turns out that STC works well, when one wants to get a quick overview of documents relevant to distinct subtopics, whereas clustering is more useful when one is interested in retrieving multiple documents relevant to each subtopic.

Index Terms: Keywords: Information retrieval,Search engines, clustering, STC, Lingo, K-Means.

I. INTRODUCTION

The existing search engines always come out with a long list of results for the given query and they are ranked by their relevance to the same query. Information retrieval and ranking functions are vital to the search engines. The organization and presentation of the results is also vital and could significantly affect the utility of the search engines[i]. A vast literature survey on page ranking and retrieval are being made by the researchers. But, there is relatively very little research that had been done on how to improve the effectiveness of search result organization[1][14]. The general concepts of the search engines are to focus upon the words that they find on a web page rather than the meaning of the words. In search result clustering, it is meant that the documents were returned in response to a query. The default presentation of search results in information retrieval is a simple list. Users scan the list from top to bottom until they have found the information they are looking for. Instead, in the case of clusters similar documents appear together. It is often easier to scan a few coherent groups than many individual documents in disarray. This is particularly useful if a search term has different word senses. Clustering of web search results is an attempt to organize the results into a number of thematic groups in the manner a web directory does it. This approach, however, differs from the human-made directories in many aspects. First of all, only documents that match the query are considered while building the topical groups. Clustering is thus preformed *after* the documents matching the query are identified. Consequently, the set of thematic categories is not fixed – they are created dynamically depending on the actual documents found in the results. Secondly, as the clustering interface is part of a search engine, the assignment of documents to groups must be done efficiently and on-line. For this reason it is difficult to download the full text of each document from the Web. Clustering ought to be performed based solely on the snippets returned by the search service.

Current search engines are still facing the lexical ambiguity issue. Ambiguity is often the consequences of the low number of query words entered on average by web users[8] The search engines retrieve thousands of for a typical query, Only top ranked pages would be viewed by the users, possibly missing the additional relevant information that might be appearing in the body of the retrieved document. The inherent ambiguity in interpreting a word/phrase in the absence of its context means that large percentage of the returned result could be irrelevant to the user.[1][4]

In recent years, Web clustering engines [7] have been proposed as a solution to the issue of lessening lexical ambiguity in Web Information Retrieval. These systems group search results, by providing a cluster for each specific aspect (i.e., meaning) of the input query. Users can then select the cluster(s) and the pages therein that contain the best answer their query needs. However, many Web clustering engines group search results on the basis of their lexical similarity, and therefore suffer from synonymy (same query expressed with different words) and polysemy (different user needs expressed with the same word).

This paper is organized in six sections Section 1 gives Introduction Section 2 The related Study, Section 3 deals the Approaches of existing Algorithms , Section 4 briefs the Analysis of existing Algorithms , Section 5 shows tables and graphs ,and Sec 6 presents the conclusion and the future work.

II. RELATED STUDY

The Most common approach in this task is Ranked List presentation. The ranked list is by far the most commonly used presentation interface nowadays. In this model the documents retrieved in response to a query are sorted out according to the relevance of the query – most relevant first. A single entry in the ranked list usually consists of the title of the document, its URL and a short excerpt from it called a snippet. The other approach is using Web directories such as the Open Directory Project. They provide categorization for the classification of Web pages.

Manuscript published on 30 August 2013.
* Correspondence Author (s)
R. Mahalakshmi, R, Research Scholar, M.S. University Tirunelveli, Tamil NAdu, India.
Dr. V. Lakshmi Praba, Assistant Professor, Sivaganga Women's College, Madurai, Tamil Nadu, India.

Given a query, search results are organized by category. This approach has three main weaknesses: first, it is static, thus it needs to be updated manually to cover new pages; second, it is unable to cover large portion of the Web; third, Web pages are classified based on common categories. This latter feature of Web directories makes it difficult to distinguish between instances of the same kind. While methods for the automatic organization of Web documents have been proposed and some problems have been tackled effectively[ii] [2], these approaches are usually supervised and still suffer from a reliance on a predefined taxonomy of categories.

Different techniques of research direction consisting of associating explicit semantics like WSD- word Sense Disambiguation, Lateral Semantic Indexing(LSI) , and Transduction based .clustering are also used in the process. However, contrasting results have been reported on the benefits of these techniques:

The next approach to clear the query ambiguity is search result clustering. Given a query, a flat list of text snippets returned from commonly-available search engine is clustered using some notion of textual similarity. Search result clustering approach can be classified as data-centric or description-centric [7]. The former focus more on the problem of data clustering than on presenting the results to the user. A pioneering example is Scatter/Gather [4,6], which divides the dataset into a small number of clusters and, after the selection of a group, performs clustering again and proceeds iteratively. Developments in this approach have been proposed which improve on cluster quality and retrieval performance [17]. Other data-centric approaches use agglomerative hierarchical clustering [21], rough sets [22] or exploit link information [5][47]. Instead description-centric approaches are, more focused on the description of Clustering Web Search Results with Maximum Spanning Trees [9] to produce for each cluster of search results. Among the most popular and successful approaches are those based on suffix trees and K-means clustering algorithms. Other methods are based on formal concept analysis [5][8], singular value decomposition [5][30], spectral clustering [5][11] and graph connectivity measures [5] [14].

Diversification is another research topic dealing with the issue of query ambiguity. Its aim is to reorder the top search results using a criteria that maximize their diversity. Similarity functions have been used to measure the diversity among the documents and between document and query [7]. Other techniques include the use of conditional probabilities to determine which document is most different from higher-ranking ones [9] or use affinity ranking [4], based on topic variance and coverage. More recently, an algorithm called Essential Pages [7] had been proposed to reduce information redundancy and return. The use of hidden topics had been proposed to identify query meanings [19] Clustering can also speed up search. In such cases, we could compute the similarity of the query to every document, but this is slow. The cluster hypothesis offers an alternative viz., finding the clusters that are closest to the query and only consider those documents for analysis. Within this much smaller set, we can compute similarities exhaustively and rank documents in the usual way. Since there are many fewer clusters than documents, finding the closest cluster is fast; and the documents matching a query are all similar to each other, they tend to be in the same clusters. While this algorithm is inexact, the expected decrease in search quality is small.

III. SEARCH RESULT CLUSTERING ALGORITHMS

Web search result clustering is typically performed in three steps:

1. Given a query *q*, a search engine is used to retrieve a list of results
R = (r1, . . . , *rn*);
2. A clustering *C* = (*C*0, *C*1, . . ., *Cm*) of the results in R is obtained by means of a
clustering algorithm;
3. The clusters in *C* are optionally labeled with an appropriate algorithm
for visualization purposes.

3.1 Suffix Tree Clustering:

The Suffix Tree Clustering (STC) algorithm groups the input texts according to the identical phrases they share [1]. The rationale behind such approach is that phrases, compared to single keywords, have greater descriptive power. This results from their ability to retain the relationships of proximity and order between words. A great advantage of STC is that phrases are used both to discover and to describe the resulting groups. The Suffix Tree Clustering algorithm works in two main phases: base cluster discovery phase and base cluster merging phase. In the first phase a generalized suffix tree of all texts' sentences is built up using words as basic elements. After all the sentences are processed, the tree nodes contain information about the documents in which particular phrases appear. Using that information documents that share the same phrase are grouped into base clusters of which only those whose score exceeds a predefined Minimal Base Cluster Score are retained. In the second phase of the algorithm, a graph representing relationships between the discovered base clusters is built based on their similarity and on the value of the Merge Threshold. Base clusters belonging to coherent sub graphs of that graph. They are merged into final clusters.

A clear advantage of Suffix Tree Clustering is that it uses phrases to provide concise and meaningful descriptions of groups. However, as noted in [12] STC's thresholds play a significant role in the process of cluster formation, and they turn out particularly difficult to tune. Also, STC's phrase pruning heuristic tends to remove longer high-quality phrases, leaving only the less informative and shorter ones. Finally, as pointed out in [12], if a document does not include any of the extracted phrases it will not be included in the results although it may still be relevant.

3.2 K Means Clustering Algorithm

This classic approach attempts to adopt the well-known clustering algorithms, originally
designed for numerical data, such as Hierarchical Agglomerative Clustering (HAC) or K-
means, to the data of textual type. The algorithms require that for every two objects in
the input collection a similarity measure be defined. The measure, which is usually calculated as a single numerical value, represents the "distance" between these objects that are "close" to each other in this sense will be placed in the same cluster. Thus, to use the classic algorithms for the new type of data, the measure of similarity between two texts must be defined.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Many different approaches to this problem have been proposed [20]After the distance measure between two texts has been chosen, classic clustering algorithms are ready be used on textual data. We briefly describe below the main ideas behind these two most popular classic approaches: Hierarchic Agglomerative Clustering and Kmeans. **For more details regarding the algorithms refer to and [ ].** K-means is an iterative algorithm in which clusters are built around *k* central points12 called centroids. The algorithm starts with a random set of centroids and assigns each object to its closest centroid. Then, repeatedly, for each group, based on its members, a new central point (new centroid) is calculated and object assignments to their closest centroids are changed if necessary. The algorithm finishes when no object reassignments are needed or when certain amount of time elapses.

3.3 Lingo

The general idea behind LINGO is to first find meaningful descriptions of clusters, and then, based on the descriptions, determine their content. Similar technique seems to be used by Vivisimo [9] and Carrot 3.6.1[9]– for search results clustering engines, In this approach, the careful selection of cluster labels is crucial The algorithm must ensure that both the labels differ significantly from each other and at the same time cover most. It is possible to find such labels using the Vector Space Model along with the Latent Semantic Indexing(LSI) technique. To assign documents to the already labeled groups LINGO could use the Latent Semantic Indexing in the setting for which it was originally designed viz., given a query – retrieve the best matching documents. When a cluster label is fed into the LSI as a query, result contents of the cluster will be returned. This approach should take advantage of the LSI's ability to capture high-order semantic dependencies in the input collection. In this way not only would documents that contain the cluster label be retrieved, but also the documents in which the same concept is expressed without using the exact phrase. In web search results clustering, however, the effect of semantic retrieval is sharply diminished by the small size of the input web snippets. This, in turn, severely affects the precision of cluster content assignment.

To become a full-featured clustering algorithm, the process of finding cluster labels and contents must be preceded by some preprocessing of the input collection. This stage should encompass text filtering, document's language recognition, stemming and stop words identification. It is also recommended that post-processing of the resulting clusters be performed to eliminate groups with identical contents and to merge the overlapping ones. Many more research studies are needed to have an efficient search result clustering engines.

IV. ANALYSIS OF EXISTING ALGORITHMS

Test Sets. Experiments on two datasets were conducted and the results were presented below.

– AMBIENT (AMBIguousENTries) is a data set which contains 44 ambiguous queries 2. and eTools web Search It is a new data set that allows open queries . The aim was to study the behavior of Web search algorithms on queries of meanings of different lengths. The AMBIENT dataset is composed mostly of single-word queries. eTools web Search provides dozens of queries of length 2&3 . This can increase recall since a group of documents with high mutual similarity is often relevant as a whole

Parameters.

The common parameter is the maximum number of clusters(N). We experimentally set this value to 20 , Table 1 presents the details of the data source and keywords. Table 2 brings out the analysis of the three algorithms based on Table 1 and other parameters are given in Table3..

4.1 Evaluation Criteria

There are four external criteria for evaluating clustering quality viz., ***purity, Normalized mutual information, Rand index* and *F measure***. we briefly discuss them below.

4.1.1 Purity

Purity is a simple and transparent evaluation measure. To compute this external quality measure, each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned named entities and dividing by *N*. It is expressed formally as:

$$purity\ \Omega,\mathbb{C} = 1N \max j|wk \cap cj|k$$

4.1.2 Random Index

The *Rand index* ( ) measures the percentage of decisions that are accurate. Two documents can be assigned to the same cluster if they are similar. A true positive (TP) decision assigns two similar documents to the same cluster, a true negative (TN) decision assigns two dissimilar documents to different clusters. There are two types of errors that are committed. A (FP) decision assigns two dissimilar documents to the same cluster. A (FN) decision assigns two similar documents to different clusters.

$$\mathrm{RI} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}$$

The Rand index gives equal weight to false positives and false negatives. Separating similar documents is sometimes worse than putting pairs of dissimilar documents in the same cluster.

4.1.3 F-Measure

We can use the *F measure* which measures to penalize false negatives more strongly than false positives by selecting a value $\beta > 1$, thus giving more weight to recall.

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \qquad R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

Based on the Documents in the clusters ,we have calculated F measure having $\beta = 1$ and r $\beta = 5$. We have also used the external cluster quality measure called purity to evaluate quality of the clusters. For the accuracy of a label assigned to a cluster there is no guidance and the decision is made based on the evaluator's subjective judgment

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

V. ANALYSIS AND GRAPHS

Fig1.0 presents the data source and keyword with base cluster Count..

Fig1.0

data source	E Tool search
Keyword	system
cluster base count	20

Fig2.0 Presents the various constraints of Lingo, STC and K- Means clustering algorithms for the Keyword System, Data source-e-tools search.

Fig 2.0

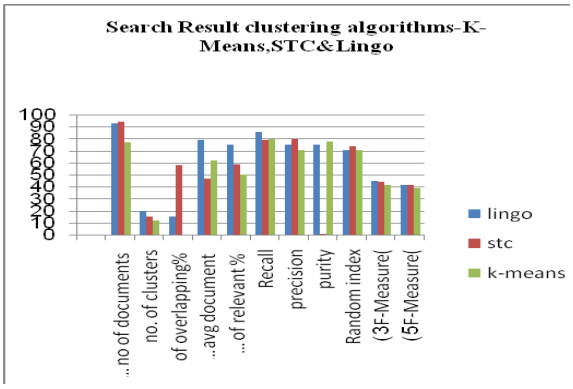
Parameters	LINGO	STC	K-MEANS
No Of Documents Retrieved	93	94	77
No. Of Clusters	20	15	7
% Of Overlapping	15.05	58	0
Avg Document Relevancy To Clusters	78.79	46.39	45.08
% Of Relevant Labels	75	58.33	50
Recall	85.37	79.27	79.41
Precision	75.27	79.79	70.13
purity	75.26882	0.740532	77.92208
Random index	70.34	73.78	70.52
F-Measure(3)	44.44444	44.18162	41.37931
F-Measure(5)	41.6	41.354	38.73103

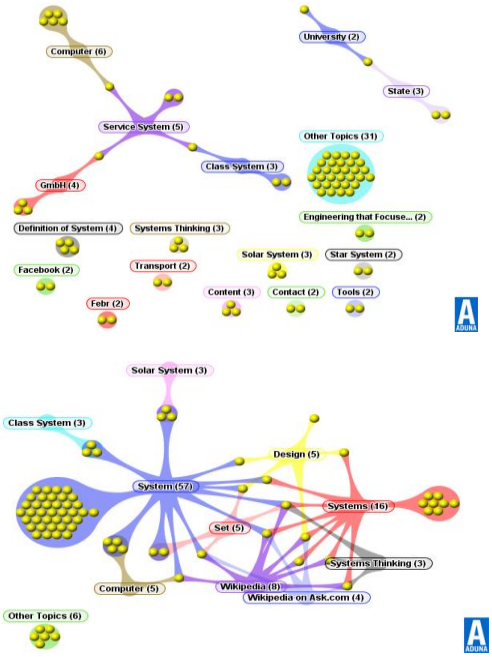
Fig3.0

Other Parameters	Lingo	STC	K-Means
Cluster Merging Threshold	0.7	0.6	
Label Assignment	Simple	Simple	-
Title Word Boost	2	2	2
Factorization Method(Default)	Non Negative Factorization ED Factory	-	Non Negative Factorization ED Factory
Factorization Quality	High	High	HIGH
Term Weighting(Default)	Log Tf Idf Term Weighting	-	Linear Tf Idf Term Weighting
Word Frequency Threshold	0.9	1	1
Maximum Words Per Label	1	1	1
Overlapping	0.6	0.5	Nil

VI. GRAPHS

The Graph- I below shows the Performance of the three search result clustering algorithms for the Key word System using the e-tools Search dataset





Aduna Cluster Map Visualizations of K-means &STC for the key word System

Fig 4.0 presents the data source and keyword with the base cluster Count..

Data Source	**Ambinet**
Keyword	Monte Carlo
Cluster Base Count	20

Fig 4.0

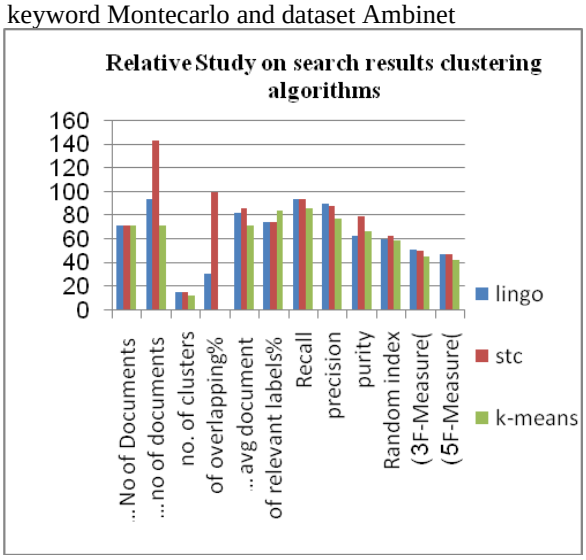
The Fig 5.0 explains the various constraints of Lingo, STC and K- means for the keyword Monte carlo and dataset Ambinet

Parameters	**LINGO**	**STC**	**K-MEANS**
No Of Documents Retrieved	93	94	77
No. Of Clusters	20	15	7
% Of Overlapping	15.05	58	0
Avg Document Relevancy To Clusters	78.79	46.39	45.08

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

% Of Relevant Labels	75	58.33	50
Recall	94.11764706	94.48819	86.15385
precision	90.42553191	88.19444	77.77778
purity	62.76595745	79.16667	66.66667
Random index	60.09333333	63.05747	59.09722
F-Measure(3)	51.24147545	50.68495	45.41768
F-Measure(5)	47.96202102	47.44111	42.51095

Graph II presents the parameters in a diagrammatic form for the keyword Montecarlo and dataset Ambinet



VII. CONCLUSION

Each algorithm has its own merits and demerits, Lingo produces high cluster diversity, the Small outliers are highlighted well, In STC and K-means algorithms the small outliers are rarely highlighted .In Lingo the number of clusters produced are more when compared to other two algorithms..With respect to the cluster labels , in LINGO they are descriptive but lengthy , not very descriptive in K-Means ,but in STC cluster labels are small but very appropriate. The Scalability is high in STC compared to Lingo and K-Means. Other features of K-Means clustering are Running time: O(KN) (K = number of clusters) ,Fixed threshold ,Order dependent. Features of STC are Overlapping clusters, Non-exhaustive ,Linear time, and High precision.

6.1 Future Work

The lingo could be improved to generate well formed long descriptive labels, and STC can be further developed to consider small outliers and K-Means can be augmented to have overlapping clusters. And in all the three algorithms, reducing the "**others**" cluster options can be tried which otherwise not able to label appropriately by the algorithms..Most frequent / most highly weighted / most discriminative terms in each cluster chosen often gives highly unsuggestive results. Subclass extractors could be given more importance,look for instances of a subclass rather than the super class can be observed . E.g. it is easier to find people described as "physicists", "biologists", "chemists" etc. rather than "scientists.

REFERENCES

1. Oren Zamir and Oren Etzioni. *Document Clustering: A Feasibility Demonstration*.Proceedings of the 19th International ACM SIGIR Conference on Research and Development of Information Retrieval, 1998, pp 46-54.
2. Oren Zamir and Oren Etzioni. *Grouper: A Dynamic Clustering Interface to Web Search Results*. WWW8/Computer Networks, Amsterdam, Netherlands, 1999.
3. Oren E. Zamir. *Clustering Web Documents: A Phrase-Based Method for Grouping Search Engine Results*. Doctoral Dissertation, University of Washington, 1999
4. Scatter/gather a cluster based approach to browsing large document collections.Douglassr cutting,David R.Karger ,Jan O Pederson,15 annual International SIGIR 92,ACM 0-89791-542-0912/0006/0318
5. Antonio Di Marco and Roberto Navigli, Clustering Web Search Results with Maximum Spanning Trees **other publication details**
6. Ke,W., Sugimoto, C.R., Mostafa, J.: Dynamicity vs. effectiveness: studying online clustering for scatter/gather. In: Proc. of SIGIR 2009, MA, USA, 2009, pp. 19–26 7.Carpineto, C., Osinski, S., Romano, G.,Weiss, D.: A survey of web clustering engines. ACM Computing Surveys 41(3), 2009, pp. 1–38
8. Kamvar, M., Baluja, S.: A large scale study of wireless search behavior: Google mobile search. In: Proc. of CHI 2006, New York, NY, USA, 2006, pp. 701–709
9. Osinski, S., Weiss, D.: A concept-driven algorithm for clustering search results. IEEE Intelligent Systems 20(3), 2005, 48–54
10. Sanderson, M.: Ambiguous queries: test collections need more sense. In: Proc. of SIGIR 2008, Singapore, 2008, pp. 499–506
35. Schutze, H.: Automatic word sense discrimination. Computational Linguistics 24(1), 1998,p. 97–124
11. Chen, J., Za¨ıane, O.R., Goebel, R.: An unsupervised approach to cluster web search results based on word sense communities. In: Proc. ofWI-IAT 2008, Sydney, Australia, (2008),. pp. 725–729
12. Zhang, X., Hu, X., Zhou, X.: A comparative evaluation of different link types on enhancing document clustering. In: Proc. of SIGIR 2008, Singapore, 2008,. pp. 555–562
13. iBoogie – meta search engine with automatic document clustering. http://www.iboogie.tv/.14. Inducing word senses to improve web search result clustering [2] Robert Navigli and Giuseppe Crisafulli department of Informatics,Rome,Proceedings osf the 2012 Conference on EMpherical Methods in Natural Language Processing,Pg 116-126 MIT,USA OCT9-11 2010 @)ACL
15. Incremental document clustering for webpage classification: In this paper they.Proposed a new feature extraction mechanism,and introduced a tree structure called DC Tree for incremental and hierarchical web document clustering.Which is less sensitive to the document insertion order.[5]
16. Incremental document clustering for webpage classification,Wai-Chiu Wong and Ada Wai-Chee Fu,Dept of Computer science and Engineering,The Chineese University Hong Kong July 1,2000.
17. A New algorithm for clustering search results Giansalvatore Mecca, Salvartore Raunich Alessandro Papppalardo Department of Mathematics and Informatics ,university of Basilicata,Potenza,ItalyApril 3,2007[6]
18. Clustering and Diversifying Web Search Results with Graph-Based Word Sense Induction Antonio Di Marco_Sapienza University of Rome Roberto Navigli Dipartimento di Informatica, Sapienza Università di Roma, Via Salaria, 113, 00198 Roma Italy.
19. ClusteringWeb-Search Results Using Transduction-Based Relevance ModelLurong Xiao and Edward Hung Department of Computing, The Hong Kong Polytechnic University, Hong Kong fcslxiao,csehungg@comp.polyu.edu.hk
20. Design Trade-Offs for Search Engine Caching RICARDO BAEZA-YATES, ARISTIDES GIONIS, FLAVIO P. JUNQUEIRA, VANESSA MURDOCK, and VASSILIS PLACHOURAS Yahoo! Research and FABRIZIO SILVESTRI ISTI – CNRACM Transactions on the Web, Vol. 2, No. 4, Article 20, Publication date: October 2008.
21. Navigli, R., Crisafulli, G.: Inducing word senses to improve web search result clustering. In:Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP), Boston, USA, (2010), pp. 116–126
22. Ngo, C.L., Nguyen, H.S.: A method of web search result clustering based on rough sets. In: Proc. of WI 2005, Compi`egne, France, (2005), pp. 673–679

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)