Journal of Theoretical and Applied Information Technology
30th June 2013. Vol. 52 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

REDUCING REPLICA OF USER QUERY CLUSTER-CONTENT AND SUB-HYPERLINKS IN THE SEARCH ENGINE LOG BASED USER PROFILE

P.SRINIVASAN, K.BATRI

Centre for Advanced Research, Department of Computer Science and Engineering

Muthayammal Engineering College, Rasipuram-637408, Tamilnadu, India

E-mail: srinimenaka@gmail.com , krishnan.batri@gmail.com

ABSTRACT

Most important visible component of the internet contains millions of web pages waiting to present information on an amazing collection of topics. The search engine has played an increasingly important role. Nowadays, the search engine based string matching has inbuilt troubles like a low accuracy, inadequate individual support, duplicates document and replica of hyperlinks in the user profile and so on. This work introduced a method against previously proposed personalized query clustering method by other authors. Experimental results show that a profile captures and utilizes both user's replica and non-replicated cluster-content and links. The non replica of hyper, sub-hyperlinks and cluster-content perform the best of among results. An important result from the experiments is that profiles with replica of links can increase the separation between similar and different queries. The separation provides a clear threshold value for a LINGO clustering algorithm to terminate and improve the overall quality of the resulting query clusters and hyper, sub-hyper links to improve search engine performance through user profile.

Keywords–*search engine, user profile, replica, hyperlinks, sub-hyperlinks*

1. INTRODUCTION

Many commercial search engines return roughly the same results for the same inquiry, regardless of the user's real interest. Since queries submitted to search engines tend to be short and imprecise, they are not likely to be able to express the user's accurate needs. For example, a farmer may use the query "beans" to find information about growing delicious beans, while JAVA programmer may use the same query to find information about JAVA beans. Since users are usually unwilling to explicitly provide their replication [12]. Due to the extra manual effort, recent research has focused on the automatic learning of user replica from user's search histories or browsed digital documents and the development of customized systems based user queries. The various user profiling strategies are involved in engine personalization and it is observed by the following problems in existing strategies. Most personalization methods focused on the creation of one single profile for a user and applied the same profile to all the user's queries. We believe that different queries from a user should be handled differently because a user's preferences may vary regarding queries. For example, a user who prefers information about fruit on the query "orange" may prefer the information about JAVA beans for the query "beans." Personalization strategies employed a single large user profile for each user in the personalization process. Existing click through-based user profiling strategies can be categorized into document based and concept-based approaches [10]. They both assume that user clicks can be used to infer user's interests, although their implication methods and the outcomes of the inference are different. Document-based profiling methods try to estimate users' document replication [12,14]. On the other hand, concept-based profiling methods aim to derive topics or concepts that users are highly interested in [4]. While there are document-based methods that consider both users' positive and negative preferences, to the best of user's knowledge, there are no concept-based methods that are considered both positive and negative preferences in deriving user's topical interests.

Search engine searches the internet based on important words and required information [3]. They keep an index of the words they find and

Journal of Theoretical and Applied Information Technology
30th June 2013. Vol. 52 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

where they find them. They allow users to look forwards or combinations of words found that index. Search engine will index hundreds of millions of pages and respond to tens of millions of queries per day. During the retrieval process search engine will provide many result based on page rank and user profile logs [5,6,7]. This work focuses more on the replication of cluster-content, hyper and sub-hyperlinks are eliminated and recover the user profiling identical links in the user logs. It will improve the performance of the search engine user profile. The main advantage of a clustered solution is automatically recovered from failure that is recovery without user intrusion. Semantic web search is predominantly taking places while searching content in the web.

LINGO algorithm is chosen as suitable for the work [1]. Various areas of research have been explored in relation to search results clustering. In terms of clustering algorithms, the Suffix Tree Clustering algorithm introduced the use of recurring phrases in document snippets as a method of identifying similarity between documents [14]. Lingo uses a similar approach for labeling clusters [1,8]. Many algorithms have been proposed that extend or improve on the suffix tree clustering approach [12]. The related areas of research include new approaches for identifying document similarity, based on the standard *term frequency–inverse document frequency (tf-idf)* weight formula [7,11]. The *metasearch engines*, which combine results from various search engines before clustering the results [2,3]. Supervised clustering algorithms have been proposed that use learning to improve the cluster label generation process [12]. A general comparison of various clustering approaches can be found in data-centric system.

2. SYSTEM ARCHITECTURE AND WORK FLOW

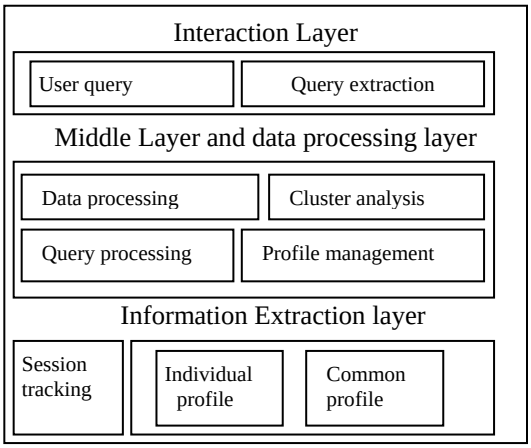

Figure 1. System Architecture

1. **Interaction layer**: Interaction layer is mainly about data user interaction, this layer can be divided into two independent parts the module stores and accesses common profiles and individual profiles with User query and Query extraction. **2. Middle layer:** Normally, the clustering analysis module assembles the results from sort module, it generates category labels, and then the results are put in a proper category to provide to users [5,3]. Middle layer can be roughly divided into four modules Data preprocessing, Cluster analysis, Query processing and Profile management. It will interact the intermediate transaction processing. **3. Information extraction:** The function of session tracking interactive layer is to provide users an easy-to-use interface to user profiles. Individual and Common user profile created and it is involved to calculate the weight of the word and match with user keyword content. The result value is stored into a Database.

Journal of Theoretical and Applied Information Technology
30th June 2013. Vol. 52 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

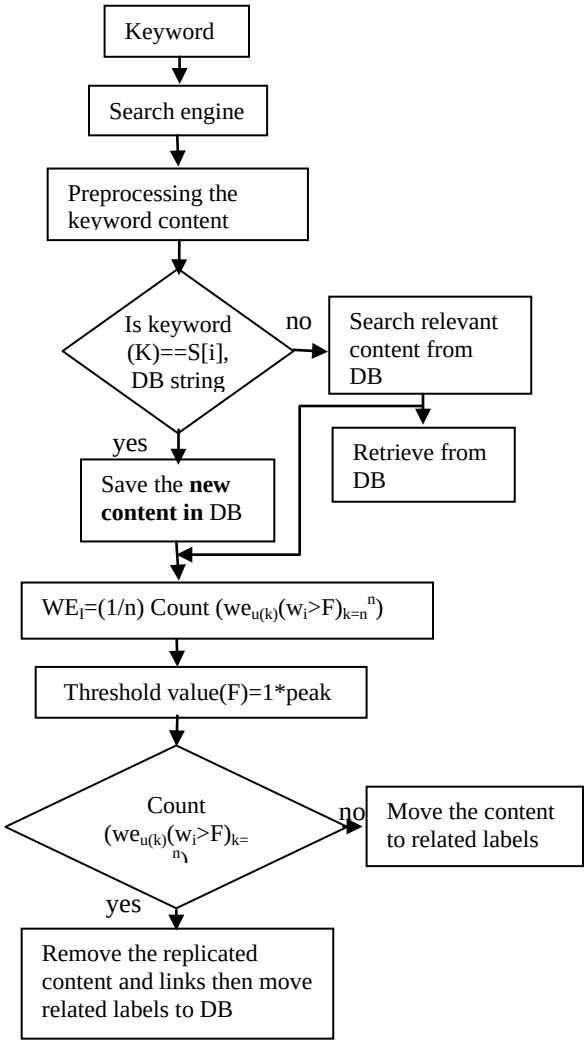

Figure 2. Functional frame work flow graph

3. ALGORITHM

/** Phase 1: Preprocessing */
for each document in the user profile for DB
{
do text filtering;
identify the document's language and keyword content;
apply stemming;
mark stop words;
}
/** Phase 2: Feature extraction */
discover frequent cluster-content and hyper, sub-hyperlinks;
/** Phase 3: Cluster label induction */
use LSI to discover abstract concepts;
for each abstract concept
{
Find matching cluster-content and hyper, sub-hyperlinks;
}
compare similar cluster labels;
remove replicated cluster-content and hyper, sub-hyperlinks and user queries
Store the content and hyper, sub-hyperlinks in to DB

4. CLUSTER ANALYSIS IN THE SEARCH RESULTS

Lingo [1,11] is particularly suited to solving the problem of search result clustering. Unlike most other algorithms, it first attempts to discover cluster-content and hyperlinks for future clusters and only then proceeds to assigning each cluster with matching documents of DB. This reversed process, compared to other search results clustering algorithms, allows Lingo to partially avoid the trap of verbally unexplainable clusters. The hyperlinks have verbally unexplainable clusters so that LINGO is more suitable for the cluster analysis.

4.1 PREPROCESSING

4.1.1 Text filtering

In the text filtering step, all terms that are useless or would introduce noise in cluster labels
are removed from the input documents. Among such terms are:
• HTML tags (e.g. <table>) and entities (e.g. &) • non-letter characters such as "$", "%" or "#" (except white spaces and sentence markers such as '.', '?' or '!'). Note that at this stage the stop-words are not removed from the input documents. Additionally, words that appear in snippet titles are marked in order to increase their weighting further phases of clustering [5].

4.1.2 Language identification

Before proceeding with stemming and stop words marking, for each input document separately, LINGO tries to recognize its language. In this way, for each snippet, appropriate stemming algorithm and stop list can be selected [9]. This step is immensely important for two main reasons. First of all, it may be inconvenient for the users to choose manually the appropriate language version of the clustering algorithm. With the automatic language recognition there is no need for the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

users to trouble with such choice. Secondly, for many queries it is unreasonable to stick to one language only as documents in a mixture of different languages may be returned.

4.1.3 Stemming

In this step, Twenty five required stemmer is available, inflection suffixes and prefixes are removed from each term appearing in the input collection. It guarantees that all inflected forms of a term are treated as one single term, which increases their descriptive power. LINGO is the best algorithm for using English and Polish stemming. This work follows the free Java implementation of Porter Stemmer along with LINGO algorithm [9].

4.1.4 Stop words removal

Although stemming alone does not present any descriptive value, stop words may help to understand or disambiguate the meaning of content (compare: "he is a great man and he is a man of great"). That is why we have decided to retain them in the input documents, only adding appropriate keywords. This will enable the further phases of the algorithm to e.g. filter out cluster-content and hyper, sub-hyperlinks ending with a stop word or prevent the indexing stop words at all.

4.1.5 Feature extraction

In LINGO, the data on which the normal clustering algorithms work is built of words, rather than single letters as in the examples. In this way, as a result, terms and keywords along with the number of their occurrences are returned. In the final step of the feature extraction keyword content, terms and hyper, sub-hyperlinks that exceed the term frequency threshold value are chosen. We have empirically established that for best accuracy of clustering, the value of the threshold value should fall within the range between 0 and 1. The fairly low values of these boundaries result from relatively small sizes of the input documents (i.e. snippets). A summary of all parameters of LINGO and its default values are used.

5. USER PROFILING TECHNIQUES

The aim of user profile modeling is to describe user's action of browsing website and searching information. The profiling process will provide knowledge and information best meeting users' needs. User profile consists of common user profile and individual user profile [3,4,8].

5.1 Common User Profile

Need of describing common action of a group while individual user profile only describes individual user's action. This hierarchy framework could be further divided into individual user profile and common user profile. When a new user registers, he should select a group to which he belongs, his individual user profile inherits some will attribute from the group's common profile. Common user profile is defined as CP = (CI, WL, IP) and CI = (GID, NAME, DE, GID) GID is the unique ID of this group. NAME is the current group's name while DE is some other description information of the current group. WL = $((W_1,WE_1), (W_2,WE_2),\ldots\ldots\ldots)$ Wi denotes a word in the category label hierarchy, WEi denotes the current word's weight, and its definition is shown in the following formula:

$$WEi = \frac{1}{n}\text{count}(we_{u(k)}(wi > F)_{k=1}^{n}) \text{ - - - - -(1)}$$

$(wi > F)_{k=1}^{n})$ $\text{count}(we_{u(k)}(wi > F)_{k=1}^{n})$ are the number of times in which situations that the weight of word Wi is greater than threshold value F. F may be given a proper value like 0.05, 0.1, …1

5.2 Individual User Profile

Individual user profile is defined as UP = (UI, P, UPL) and UI = (UID,UN,UD) P = <CP,NIP>UID is the unique ID of an individual user; UN is the current user's name.

While UD is some other description information of the current user. CP is the pointer pointing to the common user profile of the group to which the individual user belongs, and NIP points to the next node of the user profile link list in the same group. UPL= ((UW1,UPW1,UWE1), (UW2, UPW2, UWE2,,) ,……) UWi denotes a word. UPWi denotes the category label that the word belongs to. UWEi is the weight of the

Journal of Theoretical and Applied Information Technology
30th June 2013. Vol. 52 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

current word UWEi can get a negative value while WEi can only get a plus value. Assuming that a user has done m times searching, and this user has clicked n websites in a special search. UWEi is calculated in the way shown in the formula bellow

$$UWEi = \frac{1}{m} \frac{\sum_{k=1}^{n} Cik}{\max\{\sum_{k=1}^{n} Cjk,\ j = 1,2,.....\}} - - - (2)$$

C_{ik} is the number of times that the number of "i" word emerges in the number of "k" website $\sum_{k=1}^{n} C_{ik}$ is a formula to count the number of times that the number of "i" word emerges in all the n websites. $\max\{\sum_{k=1}^{n} C_{jk}, j = 1,2, \ldots\ldots\}$ denotes the max number of times of all words. The formula above may describe well a keyword importance in all searches. Finally, remove the replica of cluster-content, hyper, sub-hyperlinks and store in to the data base.

6. EXPERIMENTAL RESULTS AND DISCUSSIONS

A precise user profile can significantly improve a search engine's performance by identifying the information needs for individual users. The techniques make use of click-through data to extract from Web-snippets to build concept-based user profiles automatically. The user profiling strategies were evaluated and compared with the existing personalized query clustering method. Initially, interaction between users can be mined from the concept-based user profiles to perform collaborative filtering. This allows users with the same interests to share their profiles. Subsequently, the existing user profiles can be used to predict the intent of unseen queries, such that when a user submits a new query, personalization can benefit the unseen query. Soon after, the concept-based user profiles can be integrated into the ranking algorithms of a search engine so that search results can be ranked according to individual user's interests.

The following results works on the user profile and improves the performance of search engine. Before and after removal of replica in the user profile are shown in terms of sub-hyperlinks and cluster-content received from the database.

Table 1: Before removal of replica in cluster-content, hyper and sub-hyperlinks with threshold value 0.05

Keywords	URL Clicks	Total number of user clicks		
		Cluster-content	Sub-hyper links	Hyper links
Apple	Stock	60	30	4
Search	Engine	57	75	5
Laptop	Information	32	30	2
Bioluminescence	Version	34	35	4
Web	Inc	55	30	5
Lingo	Algorithm	35	30	3
Engineering	College	40	25	-
iPod	Info	30	35	-
Cluster	Content	54	33	-
query processing	Query	45	35	4

Table 2: After removal of replica in cluster-content, hyper and sub-hyperlinks with threshold value 0.05

Keywords	URL Clicks	Total number of user clicks		
		Cluster-content	Sub- hyper links	Hyper links
Apple	Stock	43	36	4
Search	Engine	40	81	5
Laptop	Information	15	36	2
Bioluminescence	Version	17	41	4
Web	Inc	38	36	5
Lingo	Algorithm	18	36	3
Engineering	College	23	31	-
iPod	Info	13	41	-
Cluster	Content	37	39	-
query processing	Query	28	41	4

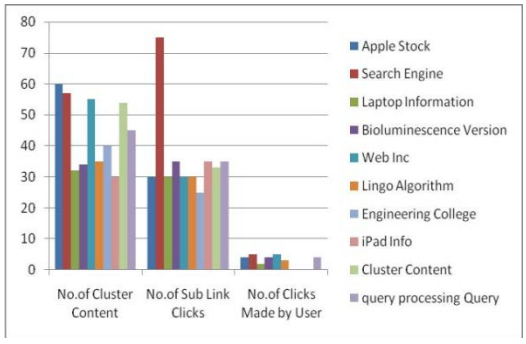

Figure 3: Before removal of replica with threshold value=0.05

Journal of Theoretical and Applied Information Technology
30th June 2013. Vol. 52 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

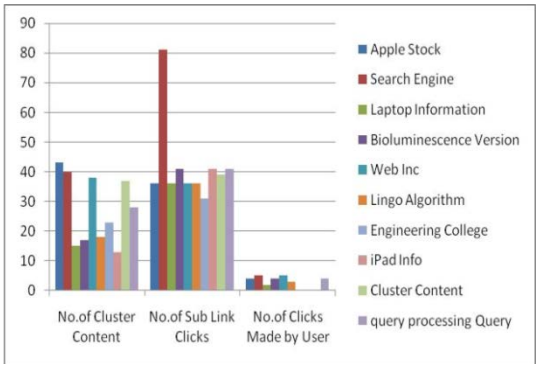

Figure 4: After removal of replica with threshold value 0.05

The keywords and hyperlinks have some replicated links and cluster-contents are unlikely to be found in the data base. Unlike the majority of the links and the keyword contents that are used in the search engine log data base is having replication, which comprise of carefully edited keyword and links to reflect a specific table 1 and Table 2. Before and after the removal of replica of keyword content and links are shown in the table with threshold value value 0.05. Figures 3 and 4 clearly describe the total number of cluster-content, hyper and sub hyper links before and after the removal of the replica are not reply any remarkable changes.

Table 3: Before removal of replica in cluster-content, hyper and sub-hyperlinks with threshold value 0.1

Keywords	URL Clicks	Total number of user clicks		
		Cluster-content	Sub- hyper links	Hyper links
Apple	Stock	68	40	4
Search	Engine	62	70	5
Laptop	nformation	48	32	2
Bioluminescence	Version	42	35	4
Web	nc	63	25	5
Lingo	Algorithm	70	25	3
Engineering	College	72	27	-
Pod	nfo	41	35	-
Cluster	Content	54	32	-
uery processing	Query	46	36	4

Table 4: After removal of replica in cluster-content, hyper and sub-hyperlinks with threshold value 0.1

Keywords	URL Clicks	Total number of user clicks		
		Cluster-content	Sub- hyper links	Hyper links
Apple	Stock	17	55	4
Search	Engine	20	110	5
Laptop	nformation	13	60	2
Bioluminescence	Version	20	45	4
Web	nc	17	30	5
Lingo	Algorithm	20	50	3
Engineering	College	20	40	-
Pod	nfo	17	45	-
Cluster	Content	23	35	-
uery processing	Query	17	50	4

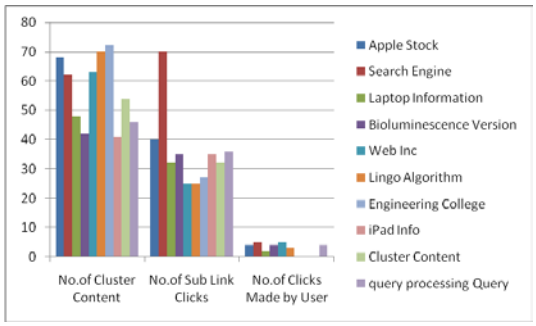

Figure 5: Before removal of replica with threshold value 0.1

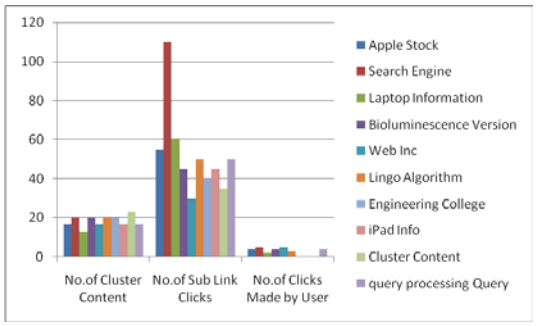

Figure 6: After removal of replica with threshold value 0.1

Tables 3 and Table 4 show the result with threshold value value 0.1. The observation made by the Figure 5 and Figure 6 illustrates the number of sub-hyper links click is been improved compare with the threshold value value 0.05 and other cluster-content reduced remarkably. It is understood that, it has some replica.

Journal of Theoretical and Applied Information Technology
30th June 2013. Vol. 52 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 5: Before removal of replica in cluster-content, hyper and sub-hyperlinks with threshold value 0.5

Keywords	URL Clicks	Total number of user clicks		
		Cluster content	Sub- hyper links	Hyper links
Apple	Stock	72	42	4
Search	Engine	65	73	5
Laptop	nformation	50	34	2
Bioluminescence	Version	42	35	4
Web	nc	63	25	5
Lingo	Algorithm	70	25	3
Engineering	College	72	27	-
Pod	nfo	41	35	-
Cluster	Content	54	32	-
uery processing	Query	46	36	4

Table 6: After removal of replica in cluster-content, hyper and sub-hyperlinks with threshold value 0.5

Keywords	URL Clicks	Total number of user clicks		
		Cluster content	Sub- hyper links	Hyper links
Apple	Stock	2	255	4
Search	Engine	3	180	5
Laptop	nformation	0	0	2
Bioluminescence	Version	3	0	4
Web	nc	2	0	5
Lingo	Algorithm	2	145	3
Engineering	College	1	0	-
Pod	nfo	2	0	-
Cluster	Content	1	0	-
uery processing	Query	1	0	4

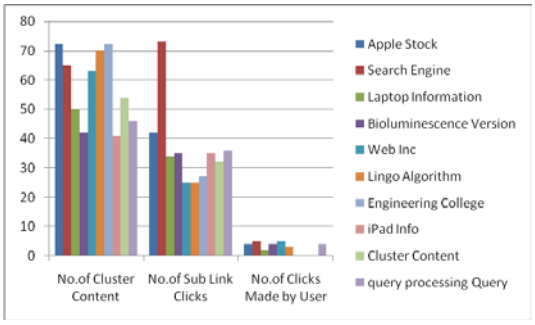

Figure 7: Before removal of replica with threshold value 0.5

Table 5 and Table 6 also show that the total number of cluster-content is being reduced subsequent threshold value value 0.5. The threshold value value 1 will be the peak value that returns more sub-hyperlinks and few cluster-content. From the result it is observed that when the threshold value value increased the number of cluster-content is reduced and the number of sub-hyperlinks is increased. Subsequently sub-hyperlinks are increased.

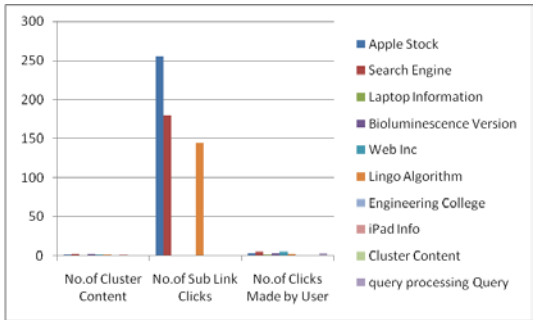

Figure 8: After removal of replica with threshold value 0.5

Table 7: Before removal of replica in cluster-content, hyper and sub-hyperlinks with threshold value 1

Key words	URL Clicks	Total number of user clicks		
		Cluster content	Sub- hyper links	Hyper links
Apple	Stock	72	42	4
Search	Engine	65	73	5
Laptop	nformation	50	34	2
Bioluminescence	Version	42	35	4
Web	nc	63	25	5
Lingo	Algorithm	70	25	3
Engineering	College	72	27	-
Pod	nfo	41	35	-
Cluster	Content	54	32	-
uery processing	Query	46	36	4

Table 8: After removal of replica in cluster-content, hyper and sub-hyperlinks with threshold value 1

Key words	URL Clicks	Total number of user clicks		
		Cluster content	Sub- hyper links	Hyper links
Apple	Stock	1	500	4
Search	Engine	1	500	5
Laptop	nformation	0	0	2
Bioluminescence	Version	1	450	4
Web	nc	0	0	5
Lingo	Algorithm	1	350	3
Engineering	College	1	400	-

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Pod	nfo	0	0	-
Cluster	Content	0	0	-
uery processing	Query	0	0	4

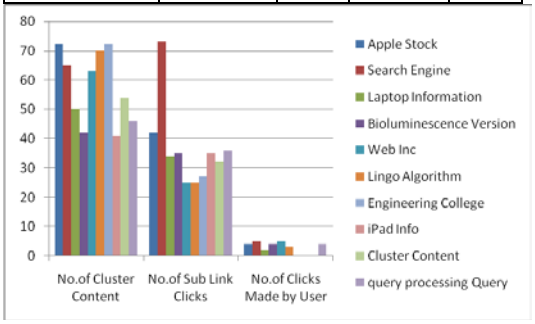

Figure 9: Before removal of replica with threshold value 1

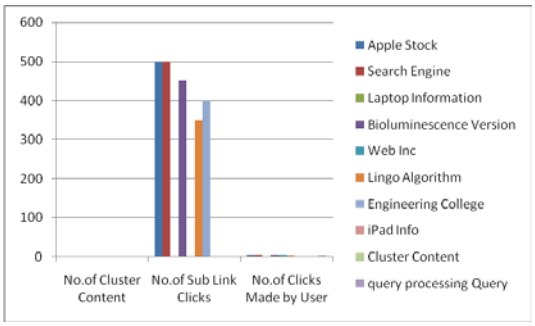

Figure 10: After removal of replica with threshold value 1

Table 7 and Table 8 show that the peak value of the maximum threshold value value 1. From the Figure 9 and Figure 10 we understood that, the sub-hyperlinks shows peak maximum value, consequently, Number of cluster-content is almost near to null.

Let's look into the profile mechanism every time a user submits a query request, system arranges and assembles the searched results according to user profile to provide user well-organized information. A special search result containing websites about "Keywords" has already been added into the common user profile and these websites will be arranged towards the front of the result website list. On the other hand, websites relevant to "Cluster-content and sub-hyperlinks" also be arranged at a proper location in all websites relevant to "keyword" according to the weight of the word in individual user profile.

In the profile management the capacity of individual and common user profile shouldn't be too large, if the capacity is too large, search engine's speed will be seriously slowed down. To avoid the slow down, we will remove the replica of the cluster-content and hyperlinks and store into the common user profile. If the user required more links and cluster-content to be viewed, they are recommended to use the threshold value value 0.05 or 0.1. If the user required indistinguishable cluster-content and hyperlinks then sub-hyperlinks, they may use threshold value 0.5,1.

7 CONCLUSION

This paper has made some investigations about removing replica of hyper, sub-hyper links and keyword cluster-content. More number of cluster-contents are replicated however, at the same time the number of sub-hyper links is being increased. Sub-hyper link does not offer much replication. The user profile log replica free access has been improved reasonably in the threshold value 0.1 than 0.05. Consequently, the threshold value value 0.5 and 1 will produce replica free user profiles at the same time the sub-hyperlinks are identical. So, that it will increase instead of decreasing in number in the results. It provides the most relevant document and the quality of searching has improved.

REFERENCES

[1] Ahmed Sameh, Amar Kadray: Semantic Web Search Results Clustering Using Lingo and Word Net. Prince Sultan University, Dept. of Computer Sc. & Info. Sys The American University in Cairo, *International Journal of Research and Reviews in Computer Science* Vol. 1, No. 2, 2010, pp.71-76.

[2] Liu. B.: Web Data Mining-Exploring Hyperlinks, Contents, and Usage Data. Department of Computer Science, University of Illinois, Chicago, USA. *A Book from Springer Series on Data-Centric Systems and Applications, Springer-Verlag Berlin Heidelberg*, 2011, pp.17-594.

[3] Kenneth, Wai-Ting Leung, Dik Lun Lee "Deriving Concept-Based User Profiles from Search Engine Logs", Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong. *IEEE Transactions on Knowledge and Data*, Vol. 22, No. 7, 2010, pp.969-982.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[4] Geraci, Filippo, Marco Pellegrini, Marco Maggini, and Fabrizio Sebastiani. "Cluster Generation and Cluster Labeling for Web Snippets". *13th International Conference, SPIRE 2006: Springer Berlin Heidelberg*, 2006, pp 25-36.

[5] Jie Yuan, Xinzhong Zhu, Jianmin Zhao, Huiying Xu "An Individual WEB Search Framework Based on User Profile and ClusteringAnalysis", *First IEEE International Conference on Ubi-Media Computing 2008*, 2008, pp 106-112.

[6] Jianshuang Deng, Qilun Zheng, Hong Peng, and Weiwei Deng. "Keywords Clustering Based on the Search Engine", *Computer Science*.Vol 03, 2007.

[7] Kenneth Wai-Ting Leung and Dik Lun Lee "Deriving Concept-Based User Profiles from Search Engine Logs" *IEEE Transactions on Knowledge and Data Engineering*, Vol. 22, NO. 7, July 2010.

[8] Osinski, Stanislaw, and Dawid Weiss. "Conceptual Clustering Using Lingo Algorithm: Evaluation on Open Directory Project Data." *Proceedings of the International IIS: IIPWMÂ´04 Conference*, 2004, pp 369-378.

[9] Oikonomakou, Nora, and Michalis Vazirgiannis. "A Review of Web Document Clustering Approaches" *Data Mining and Knowledge Discovery Handbook*, 2010.

[10] Q. Tan, X. Chai, W. Ng, and D. Lee, "Applying Co-training to Click through Data for Search Engine Adaptation," *Conference Proceedings on Database Systems for Advanced Applications (DASFAA): Springer-Verlag Berlin Heridelberg2004*, 2004, pp 519-533.

[11] Stanis law Osi´nski, Jerzy Stefanowski, Dawid Weiss(2004)LingoSearch Results Clustering Algorithm Based on Singular Value Decomposition, Institute of Computing Science, Pozna´n University of Technology, ul. Piotrowo, Pozna´n, Poland, *Procedings of the International IIS:IIP, Springer-Verlag Berlin Heidelberg*-2004, pp.380–389.

[12] Tommy W. S. Chow and M. K. M. Rahman "Multilayer SOM With Tree-Structured Data for Efficient Document Retrieval and Plagiarism Detection" *IEEE transactions on Neural Networks 20*, 2009,pp 1385–1402.

[13] Wei, ZHANG, XU Baowen, ZHANG Weifeng, and XU Junling. "ISTC: A New Method for Clustering Search Results." Wuhan University Journal of Natural Sciences, Vol 04, Pages 501-04, 2008.

[14] Wei Jiang, Mummoorthy Murugesan, Chris Clifton, Luo Si "Similar Document Detection with Limited Information Disclosure" *Proceedings of the 24th International Conference on Data Engineering: IEEE* ,2008,pp 7-12.