Journal of Theoretical and Applied Information Technology
20th July 2014. Vol. 65 No.2

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

IDENTIFYING A THRESHOLD CHOICE FOR THE SEARCH ENGINE USERS TO REDUCE THE INFORMATION OVERLOAD USING LINK BASED REPLICA REMOVAL IN PERSONALIZED SEARCH ENGINE USER PROFILE

P.SRINIVASAN AND K.BATRI
Centre for Advanced Research, Department of Computer Science and Engineering
Muthayammal Engineering College, Rasipuram-637408, Tamilnadu, India
PSNA College of Engineering and Technology, Dindigul, Tamilnadu,India
E-mail: srinimenaka@gmail.com , krishnan.batri@gmail.com
Mobile:91+9894048461,91+9789680969

ABSTRACT

The search engine information over load reduction has been focused and presented in the paper. The replica of cluster content, hyperlinks and sub-hyperlinks in the personalized user profile have been removed to reduce the information overload. This has been carried out through a personalized search results against existing query cluster's method. It has been found that the user profile with replication has more information overload than non-replicated profile. The experimental result performed by Lingo algorithm with few threshold values after finding the word weight of all content and pre-processing. The result findings confirm that the negative impact on cluster content and positive impact over sub-hyperlinks at the same time and no impact on hyperlinks. In order to measure the performance , few choices of the threshold result findings have been confirmed for the better results. The student t-test is used for analysis, and it confirms that the variations before and after removal of cluster and link based replica content in all the threshold levels.

Keywords: *Cluster Content Sub-Hyperlinks Keyword User Profile Replica Hyperlinks Overload Search Engine*

1 INTRODUCTION

The World Wide Web contains millions of web pages with information on different topics. The usage of web pages is increasing day-by-day for different purposes. This demand is rapidly increasing in recent years due to the availability of more search engines. In fact, the search engine plays a vital role to match the string given by the user with accuracy and speed. Many researchers have developed and reported various techniques to improve the performance of search engines. This work focused more on personalized search results in user profile to reduce search engine information overload.

Ahmed Sameh and Amar Kadray [1] modified the Lingo algorithm to find the synonyms of frequent words in the WordNet database. Earlier, Lingo worked with group of snippets that contain different keywords. Even so, synonymous words have not been grouped together using the original Lingo algorithm. The work added the synonyms to the pool of frequent terms that comprise the cluster label candidates. Christos Makris [2] found a search, based on ranking of the local set of categories that comprise a user's search profile. An algorithm was proposed to utilize web page categories in personalize search results. The work based on user-based experiment accordingly the proposed solution is efficient.

Supervised clustering algorithms were projected by Christoph F. Eick et al [3]. Their work improves the cluster label and its generation process. The Supervised clustering deviates from traditional clustering in that it is applied on classified examples with the objective of identifying clusters. It has high probability density with respect to a single class. Supervised clustering tries to keep the number of clusters low, and also it would be merged into one cluster without compromising the class purity while reducing the number of clusters.

Doug Beeferman and Adam Berger [4] developed a technique for mining and collected the user transaction with an internet search engine to discover clusters of similar URLs. It was reported that the agglomerative clustering algorithm performed better while using actual content of the queries or URLs and also only for click through

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

data. Filippo Geraci [5] performed clustering by of a fast version of the furthest-point-first algorithm for metric *k*-center clustering.

Gloria Bordogna et al. [6] described an iterative three main phase (1) The results of a Web search performed by the user. (2) Clusters were ranked, based on a personalized balance of their content-similarity. (3) From each cluster, a disambiguated query that highlighted the main contents of the cluster was generated. In such a way, the unused query was potentially capable of retrieve unused documents.

A technique was found to invoke semantic similarity in the clustering process by Jongkol Janruang and Sumanta Guha [7]. The proposed data structure to represent suffixes of a single string, called a Semantic Suffix Net (SSN). A generalized semantic suffix net was created to represent suffixes of a set of strings by partially combined nets. Semantic similarity was used for string matching. Judit Bar-Ilan [8] analyzed a number of measures that compared the ranking of search engine results. Five queries were applied for different search engines and found the measures. The results were ranked and compared.

Kenneth Wai-Ting Leung et al. [9] have introduced an effective approach for the first time which captures the user's conceptual preferences in order to provide personalized query suggestions. The online techniques to extract the concepts from the web-snippets using two phases personalized agglomerative clustering algorithm and also to generate personalized query clusters. The authors have evaluated the effectiveness of Google's middleware for collecting click through data to conduct experimental evaluation. In 2010, a new method was proposed against their previously reported personalized query clustering method [10]. Experimental results showed that profiles which capture and utilize both users' positive and negative preferences perform better than the earlier preferences. An important result from the experiments provided a clear threshold for an agglomerative clustering algorithm to terminate and improve the overall quality of the resulting query clusters.

Google's search engine also analyzes page content. However, instead of simply scanning for page-based text which can be manipulated by site publishers through meta-tags, Google's technology analyzes the full content of a page and factors in fonts, subdivisions and the precise location of each word. Google also analyzes the content of neighboring web pages to ensure the results returned are the most relevant to a user's query. Yahoo! Search crawls the web using Yahoo! Slurp every 2–4 weeks and automatically finds new content for indexing. If the pages that are already in their index link to a site, this site will be considered for inclusion in the next update of the index. Yahoo! Search ranked the result according to their relevance to a particular query by analyzing the web page text, title and description accuracy as well as its source, associated links, and other unique document characteristics [11].

Markus Muhr [12] integrated a hierarchical information and parent-child relations, in the cluster labeling process. This work adapted standard labeling approaches, namely Maximum Term Frequency, Chi-square Test and Information Gain to increase labeling accuracy. Mingli FENG et al. [13] introduced the personalized user-query semantic clustering approach. It did not meet the personalized demand of users concerning the interests and professional backgrounds. This work focused on query's semantic ambiguity, search process of general search engine.For every specific user, three relationships have been maintained such as query contents, click sequence and selected documents.

Nadine Hochstetler and Dirk Lewandowski [14] investigated the composition of search engine results in pages. Their work defined an organic result, advertisements, shortcuts and rare queries. The sample queries passed through major search engines and count how often the different elements are used by the individual engines. The results shown and compared based on some sample elements.

Oikonomakou [15] classified the document into three categories i.e text-based, a link-based and hybrid. Furthermore, a comparison made by based on various facets, algorithm features, functionality and matching the document based clustering. Phiradit Worawitphinyo et al. [16] used ranking based clusters by Suffix Tree Clustering (STC) baseline and similarity measures for comparing clusters. The Hierarchical Agglomerative Clustering and STC was combined (STHAC) to improve the cluster merging process, and it achieved 16% more than the original STC.

Stanis law Osi´nski et al. [17, 18, 19] found online automatic result clustering problem for grouping the similar documents in a search hit list for search engine. The work evaluated the Lingo algorithm for open data project and compared the clusters acquired from Lingo to the expected set of database categories mixed in the input. The authors emphasized the cluster description quality for the algorithm and described a method for algebraic

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

transformations of the term-document matrix and frequent phrase extraction using suffix arrays. Lingo algorithm combined common phrase discovery and latent semantic indexing techniques to separate search results into meaningful groups by the same authors.

Stanislaw Osinski [20] discussed the approximate matrix factorizations, and organized document summaries returned by a search engine into meaningful thematic categories. The work compared six different factorizations (SVD, NMF, LNMF, STC, Tolerance Rough Set Clustering (TRC) and K-Means/Concept Decomposition). Q.Tang et al. [21] suggested that a ranking and Co-training Framework. It takes the click through data containing the items in the search result that were clicked on by a user as an input, and generated adaptive rankers as an output. It demonstrated RSCF algorithm produced better higher- ranking results than the standard higher-ranking SVM algorithm.

Utku Irmak [22] focused on improving the overall quality of user-centric entity detection systems, concept extraction technique, which relies on search engine query logs. An estimation of relevancy for feature space was represented by a novel approach in the given context. A large-scale user-centric entity detection system was utilized by a click through data obtained from search engine and also extracted the rank.

Meta search engine is to increase quantity and quality of a search result [23]. The work used services from Google and Yahoo!, and it gave an overview of techniques employed by these search engines to retrieve and rank their search result. Google claims to be a fully automated search engine. Software known as "spiders" is used to crawls the web on the regular basis to find sites to be added to user index. To perform the search, Google combines the measures of overall importance and query specific relevance of a page, to ensure the reliable result appears on the first position analyzed by Wiratna S et al [23].

Xiaofei He and Pradhuman Jhala [24] used click information and semantic relationship between queries. Received information incorporated into a learning system. The work analyzed the practical loss on labeled queries. It preserved the semantic relationship between queries and suggesting that the regularized regression algorithm should be used to search click information effectively for query classification. Zhiyong Zhang and Olfa Nasraoui [25] have combined two methods; (1) Model search engine user's sequential search behavior, (2) Consecutive search behavior in the client-side query refinement. The query refinement process was exploited to learn useful information that helps to generate related queries. By combining these methods, the traditional text or content based similarity method to compensate for the shortness of query sessions for real query log data.

1.1 Organization of the paper

This work consists of seven chapters. The introduction and related work carried out by other authors in the field is discussed in the chapter 1. The second chapter comprises the system architecture and profiling techniques. Cluster analysis and pre processing was carried out in the third chapter. The fourth chapter uses the implementation of profile, and result and discussions are shown in the chapter 5. Result analysis made by using student paired t-test in the chapter 6. The work concluded in chapter 7.

This work focused on search engine overload reduction through user profile. The user profile modified through replica removal based on various threshold values. Earlier, pre-processing has been done using Lingo algorithm and used in the proposed algorithm to remove the replica. The result carried out by experiments, and it is tested using student paired t-test.

2. SYSTEM ARCHITECTURE AND PROFILING TECHNIQUES

The system architecture and its components are given In Figure 1. It consists of three layers namely Interaction layer, Middle layer with data processing layer and Information extraction layer. The first layer performs query processing and extracting query from the search engine. The second layer processes the data and clusters in the profile and manages the profile with effective analysis. Normally, the clustering analysis layer assembles the results from sort of module, category labels and the results are placed into a proper category of users. The third layer maintains two profile i.e individual and common user profiles (IP,CP) with session tracking. The function of session tracking interactive layer is to provide an easy-to-use interface to user profiles after that an individual and common user profile is created, and it is involved in reducing the information overloading.

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

2.1 Cluster analysis

Lingo is particularly suited to solve the problem of search result clustering. Unlike most other algorithms, it is the first attempt to discover the cluster content and hyperlinks for future clusters. In this reversed process, compared with other search results clustering algorithms, Lingo allows, partially avoid the trap and verbally unexplainable clusters. The hyperlinks have verbally unexplainable clusters so that Lingo Algorithm is more suitable in the cluster analysis [1, 3, 17, 18].

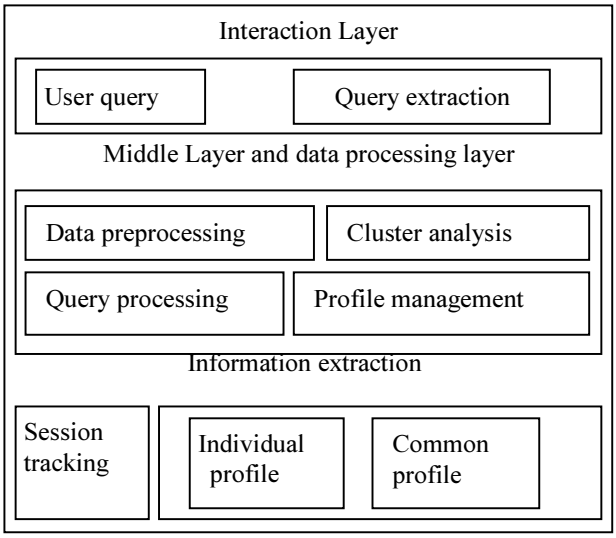

Figure 1: System Architecture With Basic Layers

2.2 Preprocessing

In the text filtering step, all terms that are useless or bearing the noise in cluster labels are removed from the input documents. Among such terms are: 1. HTML tags (e.g. <table>) and entities (e.g. &) 2. 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 keyword titles are marked in order to increase their weight [9, 10, 13, 15]. Before proceeding with stemming and stop words marking, each input document separately identified from database and transferred to Lingo algorithm. Lingo tries to recognize its language. In this way, for each keyword, appropriate stemming algorithm and stop list can be selected.

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 users' interaction. Secondly, for many queries, it is unreasonable to stick to one language as documents are a mixture of different languages. Accordingly, document clustering and click based hyperlinks also can be processed by Lingo [11].

Phase 1 Stemming**:** In this step, Twenty five stemming algorithms are 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. At 25, in the present implementation of Lingo algorithm supports English and Polish stemming. This work uses a free Java implementation of Porter stemmer and imports the input to Lingo [1].

Phase 2 Stop word's removal**:** Although stemming alone does not present any descriptive value, stop words may help to understand or disambiguate the meaning of content. So, that the work retains them in the input documents, only adding appropriate keywords. This will enable the further phases of the algorithm to filter out cluster content and hyper, sub-hyperlinks ending in a stop word or prevent the indexing stop words at all [11].

2.3 Feature extraction

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

2.3 Algorithm

/** Phase 1: receiving input from search engine */
Step 1: Enter keyword or web snippets to search engine

Step 2: get the input from click through based method
/** Phase 2: preprocessing using Lingo*/
Step 1: check the keyword K with existing DB string S[i]

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

Step 2: find out the matching cluster content, hyperlinks and sub-hyperlinks for user profile.

/** Phase 3: feature extraction and cluster label induction using Lingo */

Step 1: calculate word weight using the formula

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

Step 2: Inducting the label using Lingo algorithm with threshold value F.

Step 3: check the threshold limit from 0 to 1.

/** Phase 4: removal of replica */

Step 1: Remove the replica of cluster content, hyperlinks, sub-hyperlinks from the DB.

Step 2. Fix the threshold according to the requirement of the identical information.

Step 3. Get the identical output threshold value 0.05, 0.1, 0.75, 1.

3.4 Flow Graph for user profiling

The data flow for user profiling is shown in Figure 2. The user input keyword has submitted to the search engine for extracting information concerning to the keyword. Preprocessing has been performed after receiving the click information of hyperlinks, sub-hyperlinks and cluster content from the search engine. Received keyword compared with existing user database DB string. If the comparison is true, the algorithm search is relevant content from the user profile and put it into the database else save the new content into the database.After receiving content, the word weight WE have been calculated through Lingo algorithm for further analysis. The threshold value F is limited to 0 to 1. The next step compares the threshold value up to1 in various levels of replica removal and stores the value into the database. Finally, all replica freed content moves to the user profile.

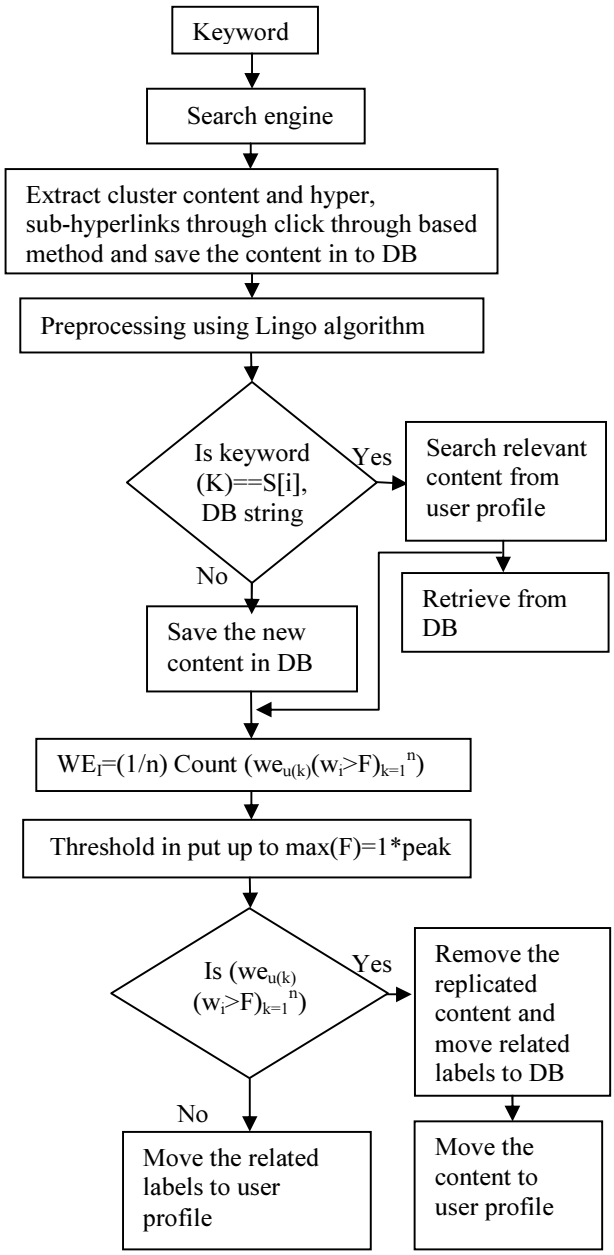

Figure 2: Flow Graph For User Profiling

4. IMPLEMENTATION OF USER PROFILE

4.1 User profiling techniques

The aim of user profile modeling is to describe user's action of browsing and searching. The profiling process will provide knowledge and information to the best of user's need. User profile consists of common user profile and individual user profile [9,10].

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

4.2 Common user profile

In the 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/she should select a group to which he/she belongs. The users' individual profile inherits the 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 about the current group. WL = ((W_1,WE_1), (W_2,WE_2),……..). Wi denotes a word in the category label hierarchy, WEi denotes the current word's weight, and its definition is given in the formula (1).

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

$(wi > F)_{k=1}^{n})$ $count(we_{u(k)} (wi > F)_{k=1}^{n})$ are the number of times in which the weight of the word Wi is greater than threshold value F. Its range starts from 0 to 1.

4.3 Individual user profile

IP is the head node of linked list consisting of all individual users. It is used in situations in which system access individual user profile. Individual user profile is defined as UP = (UI, P, UPL) and UI = (UID, UN, UD) P = <CP, NIP>UID. The unique ID of an individual user; UN is the present user's name while UD is some other description information of the present 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 linked 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 current word UWEi can get a negative value while WEi can only get a positive value. Individual user profile can be managed by user. The 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 (2).

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

C_{ik} is the number of times that the number of "i" word emerges in the number of "k" website $\sum_{k=1}^{n} Cik$ is a formula to count the number of times that the number of "i" word emerges in all the n websites. $\max\{\sum_{k=1}^{n} Cjk,\ j = 1,2,.....\}$. It denotes the maximum number of times that the keywords can be extracted from the search engine. The above formula may describe the keyword importance. Finally, replica-free cluster content, hyper, sub-hyperlinks are stored into the database.

5. RESULT AND DISCUSSIONS

5.1 Input data preparation

Every user click on the hyperlink added in to the database must be accompanied by a short description of a resource it points to. The short descriptions should serve as a substitute for cluster content, hyperlinks and sub-hyperlinks returned from a search engine's response of the user keywords. In the database, N number of click information being kept through click through method. The data sets are provided to Lingo for preprocessing as of user profile from the database DB string. Sample data sets have been taken in the replication removal process. The exact choice of categories for the10 data sets will be selected for the experiments and every category contains meaningful cluster content of each document and inside hyperlinks, sub-hyperlinks has a few schematic words. The Lingo has the flexibility to carry-out structured and unstructured data sets. The final choice of categories was connected to three abstract subjects namely cluster content, hyperlinks and sub-hyperlinks.

5.2 Evaluation measure

Similarities based on query contents, and user clicks represent two different points of view. In general, content-based measures tend to cluster around queries with the same or similar terms. Feedback-based measures tend to cluster queries related to the same or similar topics. Since, user information needs may be partially captured by both query texts and relevant documents. Our approach is based on two criteria: one is on the queries themselves, and the other on user clicks. The first

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

criterion is similar to those used in traditional approaches to document clustering methods based on keywords. The proposition formulates it as the criteria 1 and 2.

Criteria 1 (using keyword contents): If two keywords contain the identical or like terms, they denote the identical or like information needs. Obviously, the longer keywords are more reliable in the criteria 1. However, users often submit short keywords to search engines. A typical keyword on the web usually contains one or two words. In many cases, there is not enough information to deduce users' information needs correctly. Therefore, the second criterion is used as a complement, and it is similar to the intuition underlying keyword clustering in Information Retrieval (IR).

Criteria 2 (using hyperlinks click): If two keywords lead in the selection of the same document (which denote the hyperlinks click), then they are similar. Hyperlinks click are comparable to user relevance feedback in a traditional

IR environment, except that hyperlinks click represent an implicit and not always valid for relevance judgments. The ttwo criteria has been used to get more advantages. The first criterion is to group together keywords of similar compositions. The second criteria have also been used in order to cluster of user keywords' content and number of clicks made by users. However, in this work, the user clicks were used and the combination of document and luster contents are used to determine the similarity.

5.3 Experimental results

The experimental results are shown in the Tables. The variations of cluster content, sub-hyperlinks and hyperlinks are shown before and after removing the replica in the shown Figures. It is observed that, the benefit of using click through information is much more significant than using user query or keyword. There is some significant improvement in sub-hyperlinks with all keywords. It is observed that positive impact on sub-hyperlinks will produce more information, and it has fewer replicas than cluster content. The hyperlinks have no impact from the output. It is observed that, the hyperlinks do not have any replication, and also it is distinct each other.

At the same time, the cluster content have some variation with all threshold values ranging from 0 to 1. It has more replicas at lower threshold value than its counterpart. The user can fix the threshold value depends on their information requirement. If the user needs the distinct value of the content, they can use higher threshold value else lower value produces all search information. However, after removal of replica such as improvement is really not beneficial for the users who have already seen these relevant documents. Consequently, every session a user can view brand new information through this brand new profiling technique. To see how much improvement have achieved by improving the user profile, this work identified user profile improvement of unseen relevant documents in the sub-hyperlinks more than those other fields.

Table 1: Before And After Removal Of Replica For The Threshold Value Of 0.05

Keywords	URL Clicks	Before removal of replica			After removal of replica		
		No. of. Cluster content	Sub- hyper links -click	Hyper links-click	No. of. Cluster content	Sub-hyper links-click	Hyper links-click
Apple	Stock	60	30	4	43	36	4
Search	Engine	57	75	5	40	81	5
Laptop	Information	32	30	2	15	36	2
Bioluminescence	Version	34	35	4	17	41	4
Web	Inc	55	30	5	38	36	5
Lingo	Algorithm	35	30	3	18	36	3
Engineering	College	40	25	0	23	31	0
iPod	Info	30	35	0	13	41	0
Cluster	Content	54	33	0	37	39	0
query processing	Query	45	35	4	28	41	4

Keyword and URL clicks

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

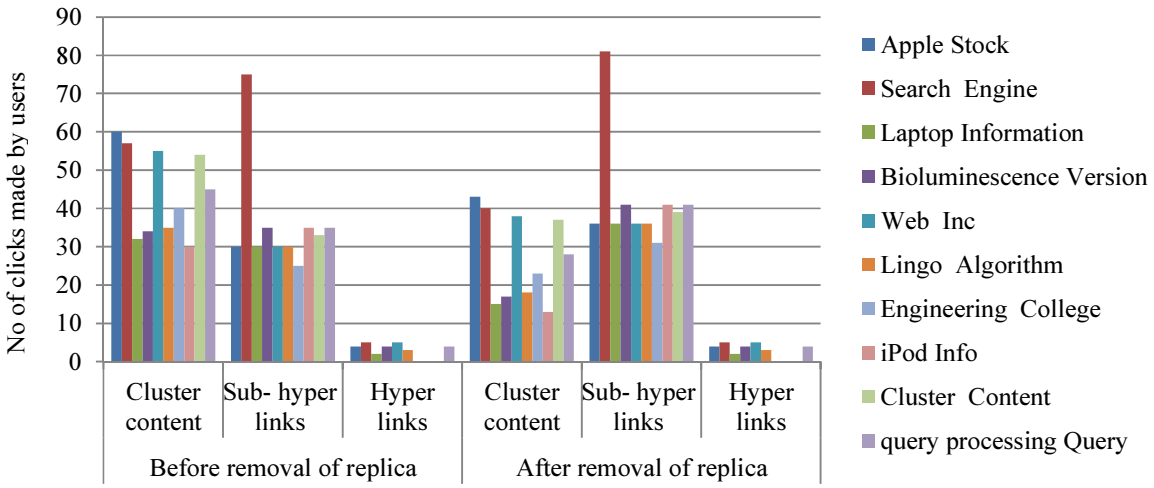

Figure 3: Variations of keyword, URL clicks and clicks made by user for the threshold 0.05.

The algorithm shows the output in Table 1 for the threshold value 0.05. The number of cluster content of the keyword "apple" in the URL stock produced before removing the replica is 60 and after removal, it is reduced into 43. Subsequently, all the keywords produce the identical negative impact on cluster content. The sub-hyperlinks produce positive impact on the identical threshold 0.05 and the identical keyword "apple" from 30 to 36 clicks. *From the result, the work produced better results for cluster content and unconstructive result for sub-hyperlinks for all 10 categories of sample keywords and URL clicks.* An analysis made from the result shows that the cluster content has more replica and sub-hyperlink has more information about the keywords. There are some variations in the keywords "search, Laptop, Bioluminescence, Web, Lingo, Engineering, iPod, Cluster and query processing" in all levels of the session. The deviation from the number of clicks made by the user against keywords, and URL clicks has shown in the Figure 3. In this level, minimum level of replication will be removed and it is used as extra relevant information for the required users.

Table 2: Before And After Removal Of Replica For The Threshold Value Of 0.1

Keywords	URL Clicks	Before removal of replica			After removal of replica		
		No. of. Cluster content	Sub- hyper links -click	Hyper links click	No. of. Cluster content	Sub- hyper links -click	Hyper links-click
Apple	Stock	68	40	4	17	55	4
Search	Engine	62	70	5	20	110	5
Laptop	Information	48	32	2	13	60	2
Bioluminescence	Version	42	35	4	20	45	4
Web	Inc	63	25	5	17	30	5
Lingo	Algorithm	70	25	3	20	50	3
Engineering	College	72	27	0	20	40	0
iPod	Info	41	35	0	17	45	0
Cluster	Content	54	32	0	23	35	0
query processing	Query	46	36	4	17	50	4

Keyword and URL clicks

Journal of Theoretical and Applied Information Technology
20th July 2014. Vol. 65 No.2

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

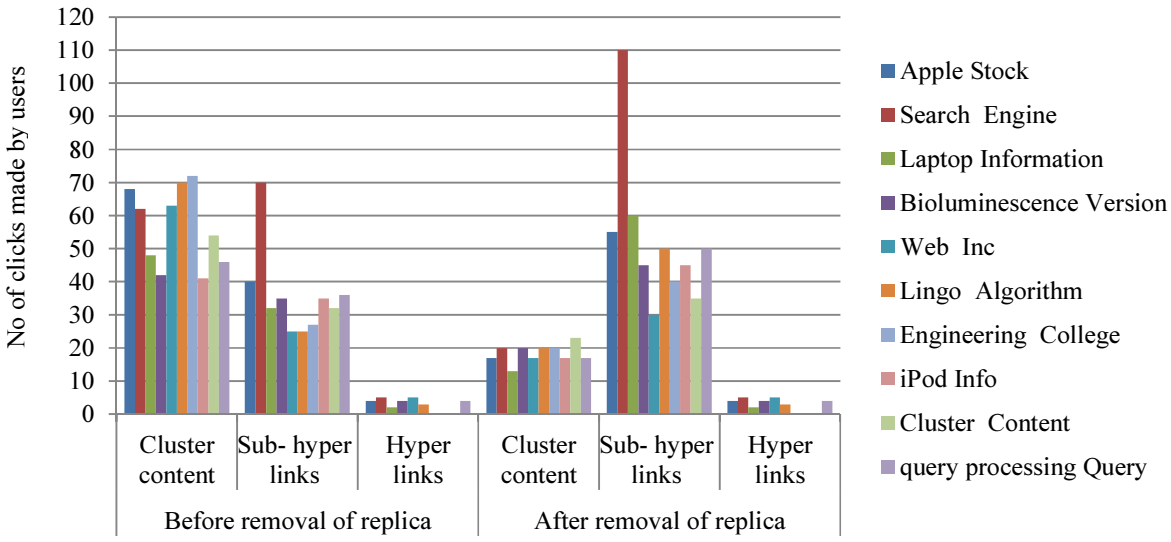

Figure 4: Variations Of Keyword, URL Clicks And Clicks Made By User For The Threshold 0.1

The algorithm shows output in the Table 2 for the threshold value 0.1. The number of cluster content of the keyword "search" in the URL Engine produced before removing the replica is 62 and after removal, it is reduced to in 20. Subsequently, all keywords produce the identical negative impact on cluster content. The sub-hyperlinks produce positive impact at the identical threshold 0.1. From the result, the work produced the better results fo and URL clicks.

An analysis made from the result shown that the cluster content has been little more identical than the threshold value 0.05. There are some variations in the keywords "Apple, Laptop, Bioluminescence, Web, Lingo, Engineering, iPod, Cluster and query processing"in all levels of the session. See, Figure 4 All the variations depend on the keyword, which the user preferably used in the particular session.

Table 3: Before And After Removal Of Replica For The Threshold Value Of 0.75

Keywords	URL Clicks	Before removal of replica			After removal of replica		
		No of. Cluster content	Sub- hyper links -click	Hyper links-click	No of. Cluster content	Sub- hyper links -click	Hyper links-click
Apple	Stock	72	42	4	2	255	4
Search	Engine	65	73	5	3	180	5
Laptop	Information	50	34	2	0	0	2
Bioluminescence	Version	42	35	4	3	0	4
Web	Inc	63	25	5	2	0	5
Lingo	Algorithm	70	25	3	2	145	3
Engineering	College	72	27	0	1	0	0
iPod	Info	41	35	0	2	0	0
Cluster	Content	54	32	0	1	0	0
query processing	Query	46	36	4	1	0	4

Journal of Theoretical and Applied Information Technology
20th July 2014. Vol. 65 No.2

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

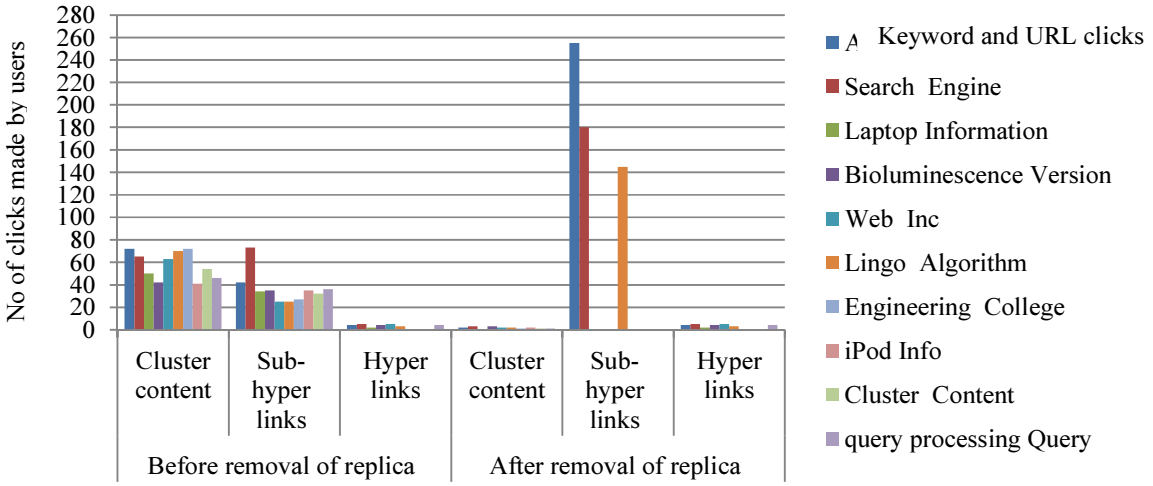

Figure 5: Variations Of Keyword, URL Clicks And Clicks Made By User For The Threshold 0.75.

The algorithm shows output in the Table 3 for the threshold value 0.75. The keyword "laptop" and its URL click information does not return any value for either the cluster content or sub-hyperlinks. Consequently, this level of threshold 0.75, concerns the output for the user both before and after removal of replica is none. The cluster content has the more off-putting impact with few outputs which is more identical being shown in Table 3.

The sub-hyperlinks have revisited few huge values of the outputs shown in Table 3. An analysis made from the result of the enormous sub-hyperlinks gives more related and identical information. The Replicated information has been removed from the user profile that is why it shows zero in Table 3. Figure 5 differentiates the sub-hyperlinks and cluster content variation clearly.

Table 4: Before And After Removal Of Replica For The Threshold Value Of 1

Keywords	URL Clicks	Before removal of replica			After removal of replica		
		No. of. Cluster content	Sub-hyper links-click	Hyper links-click	No. of. Cluster content	Sub- hyper links -click	Hyper links-click
Apple	Stock	72	42	4	1	500	4
Search	Engine	65	73	5	1	500	5
Laptop	Information	50	34	2	0	0	2
Bioluminescence	Version	42	35	4	1	450	4
Web	Inc	63	25	5	0	0	5
Lingo	Algorithm	70	25	3	1	350	3
Engineering	College	72	27	0	1	400	0
iPod	Info	41	35	0	0	0	0
Cluster	Content	54	32	0	0	0	0
query processing	Query	46	36	4	0	0	4

Journal of Theoretical and Applied Information Technology
20th July 2014. Vol. 65 No.2

ISSN: 1992-8645 www.jatit.org

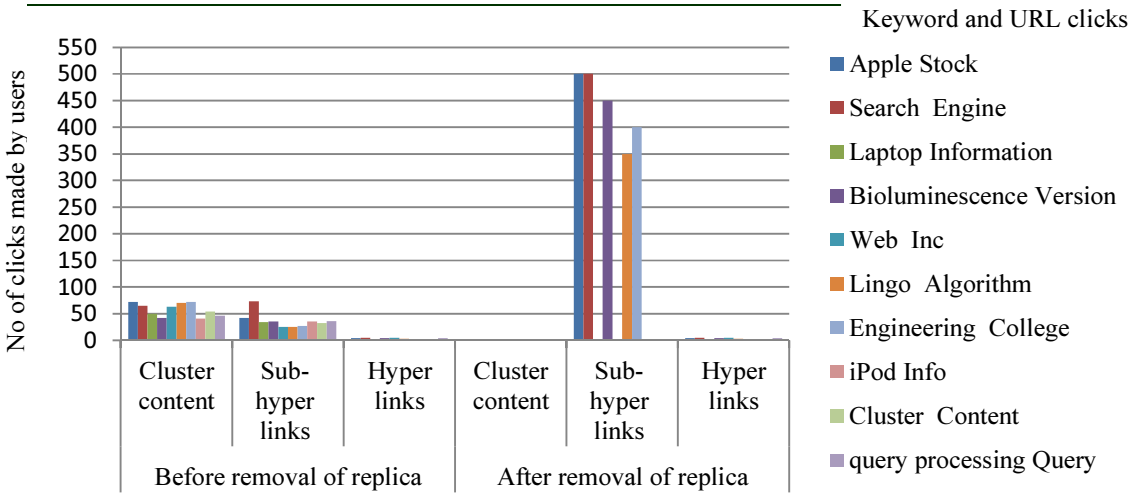

Figure 6: Variations Of Keyword, URL Clicks And Clicks Made By User For The Threshold 1.

Few keywords return zero value for the threshold value 1. Earlier, the maximum limitation of the word weight defined in the Lingo was 1. Even so, still some keywords "apple, search, bioluminescence, lingo, engineering" have produced some identical value shown in Table 4. *It shows that the few values moving beyond the limitations can be moved to user profile directly because, that cannot be reduced further.*

The identical value having more disguisable, as a result that there is no need for further reduction. It means that no replication in the keywords "laptop, web, iPod, cluster, query processing". Figure 6 shows the variation of the keywords and user clicks for threshold value 1. This level of threshold value is useful for the users who are seeing the identical value. It never produces any replicated values. All are more identical.

6 STUDENT PAIRED T-TEST ANALYSIS

The statistical analysis has been done for the cluster content and sub-hyperlinks to confirm the variations. The deviation of cluster content in the mean value validates that all the cluster content has some replica. It has been removed in all threshold values, and it shows in the Table 5. In the case of standard deviation the threshold value 0.05 has fewer deviations than all other threshold values. The sub-hyperlinks have some constructive impact on all the mean value in all thresholds that is also shown in Table 5. The standard deviation also has some extra deviation in all threshold values.

The t-value proven that the variations grant several positive impacts on every level of analysis. So, it has been proven that all the cluster content has some replica and sub-hyperlinks have a few positive impacts. The hyperlink does not reply to any impact on any levels. Therefore, it has no replica at any levels of threshold.

It is observed that, the Table 5 constructed for the t-value analysis, *B- Before removal of replica and *A- After removal of replica against the keywords and URL clicks for the threshold values 0.05, 0.1, 0.75, 1. For the t-value analysis, the Null and Alternate hypothesis have been assumed as H_0 and H_1 respectively, for the population mean μ_0, μ_1 with 18 degrees of freedom.

Null hypothesis H_0 : There is no significant difference in the population mean $\mu_0 = \mu_1$.

Alternate hypothesis H_1: There are some significant changes in the population mean $\mu_0 \neq \mu_1$.

Null hypothesis H_0 is rejected at the 95% confidence level.

By accepting the alternate hypothesis, it is observed that there is some significant difference between the cluster content and sub-hyperlinks in all the threshold value. There is some replica in cluster content and more stupidity in sub-hyperlinks. The replica removal carries out more on cluster content than sub-hyperlinks.

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

Table 5 :Student Paired T-Test For Statistical Analysis.

Keywords	URL Clicks	Cluster content		Sub-hyper links-click		Cluster content		Sub-hyper links-click		Cluster content		Sub-hyper links-click		Cluster content		Sub-hyper links-click		Hyper links-click
		*B	*A	*B	*A	*B	*A	*B	*A	*B	*A	*B	*A	*B	*A	*B	*A	*B*A
		threshold 0.05				threshold 0.1				threshold 0.75				threshold 1				All
Apple	Stock	60	43	30	36	68	17	40	55	72	2	42	255	72	1	42	500	No change
Search	Engine	57	40	75	81	62	20	70	110	65	3	73	180	65	1	73	500	
Laptop	Information	32	15	30	36	48	13	32	60	50	0	34	0	50	0	34	0	
Bio luminescence	Version	34	17	35	41	42	20	35	45	42	3	35	0	42	1	35	450	
Web	Inc	55	38	30	36	63	17	25	30	63	2	25	0	63	0	25	0	
Lingo	Algorithm	35	18	30	36	70	20	25	50	70	2	25	145	70	1	25	350	
Engineering	College	40	23	25	31	72	20	27	40	72	1	27	0	72	1	27	400	
iPod	Info	30	13	35	41	41	17	35	45	41	2	35	0	41	0	35	0	
Cluster	Content	54	37	33	39	54	23	32	35	54	1	32	0	54	0	32	0	
query processing	Query	45	28	35	41	46	17	36	50	46	1	36	0	46	0	36	0	
Mean value		44.2	27.2	35.8	41.8	56.6	18.4	35.7	52.0	57.5	1.7	36.4	58.0	57.5	0.5	36.4	220.0	
Standard deviation		11.5	11.5	14.1	14.1	11.9	2.8	13.0	22.3	12.4	0.9	13.9	97.1	12.4	0.5	13.9	236.0	
Abs(t-Value)		3.3	3.3	0.9	0.9	9.9	9.9	2.0	2.0	14.2	14.2	0.7	0.7	14.6	14.6	2.5	2.5	
df		18		18		18		18		18	18	18	18	18	18	18	18	
Assume Null hypothesis<		0.0039		0.36		0.0001		0.061		0.0001		0.5		0.0001		0.024		

7 CONCLUSIONS

This work carried out supervised clustering techniques for cluster content, hyperlinks and sub-hyperlinks. Lingo executes more on clustering the content received from the user database. An individual user profile concert improves through replica removal of its content. The threshold value 0.05, 0.1, 0.75,1 have chosen for various ranges of result findings, and actual results are shown in the Tables. Cluster content has been decreased for every threshold value and increased sub-hyperlinks simultaneously. There is no impact on hyperlinks in the experimental result since; it does not have any replica. The experimental result found some changes in before and after removal of replica, it has been proven by a student paired t-test exemplifies in the Table 5. From the result findings, this work improves the performance of personalized individual user profile and also reduces the information overload in the search engine. This work demonstrated only on the individual user profiling approach with existing sessions. If there is no replica in exiting individual user profile, it is understood that the information exceeds the threshold value 1. The work can be extended with more sessions with constructive variations. This can be automated with user interests, a choice of threshold values.

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] Hristos Makris, Yannis Panagis, Evangelos Sakkopoulos, Athanasios Tsakalidis: Category ranking for personalized search, Research Academic Computer Technology Institute, Internet and Multimedia Technologies Research Unit 5, N. Kazantzaki Str., 26504 Patras, Greece, *Journal of Data & Knowledge Engineering*, 2005 Elsevier B.V, pp.109-125

[3] Christoph F. Eick, Nidal Zeidat, Zhenghong Zhao: Supervised Clustering – Algorithms and Benefits. Department of Computer Science, University of Houston Houston, TX, 77204, *USA Technical Report Number* UH-CS-05-10, 2005

[4] Doug Beeferman, Adam Berger: Agglomerative clustering of a search engine query log. School of Computer Science, *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM New York, NY, USA, 2010,pp.407-416

[5] Filippo Geraci, Marco Pellegrini, Marco Maggini, Fabrizio Sebastiani: Cluster Generation and Cluster Labeling for Web Snippets. A Fast and Accurate Hierarchical Solution. *SPIRE 2006, LNCS 4209*. Berlin Heidelberg: Springer-Verlag,2006, pp.25-36

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

[6] Gloria Bordogna, Alessandro Campi, Giuseppe Psaila, Stefania Ronchi. Disambiguated query suggestions and personalized content-similarity and novelty ranking of clustered results to optimize web searches. *Information Processing & Management An International Journal*, 2011 Elsevier Ltd. Vol 48, Issue 3,2011, pp.419–437

[7] Jongkol Janruang, Sumanta Guha: Applying Semantic Suffix Net to Suffix Tree Clustering. Computer Science & Information Management Program, Asian Institute of Technology, *3rd IEEE Conference on Data Mining and Optimization (DMO)*, Pathumthani, Thailand, 2011,pp.146-152

[8] Judit Bar-Ilan, Mazlita Mat-Hassan, Mark Levene: "Methods for comparing rankings of search engine results", Department of Information Science, Bar-Ilan University, *52900 Ramat-Gan, Israel, Computer Networks*, 2005 Elsevier B.V, 2006,pp.1448–1463

[9] Kenneth Wai-Ting Leung, Wilfred Ng, Dik Lun Lee: Personalized Concept-Based Clustering of Search Engine Queries. Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Water Bay, Kowloon, Hong Kong. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 20, No. 11, 2008,pp.1505-1518

[10] 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

[11] 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

[12] Markus Muhr, Roman Kern, Michael Granitzer: Analysis of Structural Relationships for Hierarchical Cluster Labeling, *SIGIR'10*, Geneva, Switzerland, ACM, 2010

[13] Mingli FENG, Yajun DU Mingjun FENG Yingyu WANG: Personalized user-query semantic clustering using search click information. School of Mathematics & Computer Engineering Xihua University Chengdu, *IEEE International Conference on Management and Service Science, MASS '09Sichuan*, China, 2010,pp.1-4

[14] Nadine Hochstetler, Dirk Lewandowski: What users see – Structures in search engine results pages, Department for Research and Development, Karlsruhe, Germany, *Information Sciences An International Journal*, 2009 Elsevier Inc, Vol 179, Issue 12,2009, pp.1796–1812

[15] Oikonomakou, Nora, Michalis Vazirgiannis: A Review of Web Document Clustering Approaches. Dept. of Informatics, Athens University of Economics & Business, Patision 76, Greece, *Data Mining and Knowledge Discovery Handbook*. Springer Science &Business media, Part 6,2010, pp.931-948

[16] Phiradit Worawitphinyo, Xiaoying Gao, Shahida Jabeen: Improving Suffix Tree Clustering with New Ranking and Similarity Measures. School of Engineering and Computer Science, Victoria University of Wellington, New Zealand, Springer-Verlag Berlin Heidelberg, *ADMA 2011*, pp.55–68

[17] Stanis law Osi´nski and Dawid Weiss: Conceptual Clustering Using Lingo Algorithm Evaluation on Open Directory Project Data. 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.370–379

[18] Stanis law Osi´nski, Jerzy Stefanowski, Dawid Weiss: Lingo: Search 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

[19] Stanislaw Osi´nski, DawidWeiss: A Concept- Driven Algorithm for Clustering Search Results. Pozna´n University of Technology, *IEEE Intelligent Systems*, Vol.05,2005, pp.48-54

[20] Stanislaw Osinski: Improving Quality of Search Results Clustering with Approximate Matrix Factorisations. *Poznan Supercomputing and Networking Center, Poznan*, Poland, Springer-Verlag Berlin Heidelberg-2006, pp.167-178

[21] Tan .Q, X. Chai, W. Ng, D. Lee: Applying Co-training to Click through Data for Search Engine Adaptation. *Conference Proceedings of Database Systems for Advanced Applications (DASFAA), IEEE*,2004, pp.519-532

[22] Utku Irmak, Vadim von Brzeski, Reiner Kraft: Contextual Ranking of Keywords Using Click Data Yahoo! Inc". 701 First Avenue Sunnyvale, CA USA, *IEEE International Conference on Data Engineering*,2009, pp.457-468

[23] Wiratna S. Wiguna,Juan J. Fern´andez-T´ebar, Ana Garc´ıa-Serrano:Using Fuzzy Model for Combining and Re-ranking Search Result from Different Information Sources to Build Metasearch Engine, *RIMMEL project*, funded by the Spanish Research Council, (2006)

[24] Xiaofei He, Pradhuman Jhala: Regularized query classification using search click information, Yahoo! Research Labs, 3333 Empire Avenue, Burbank, CA 91504, USA, *Pattern Recognition 41*, 2008- Elsevier Ltd,2008, pp.2283-2288

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

[25] Zhiyong Zhang, Olfa Nasraoui: Mining search engine query logs for social filtering-based query recommendation, Knowledge Discovery and Web Mining Lab, Department of Computer Engineering and Computer Science, University of Louisville, Louisville, USA, *Journal of Applied Soft Computing*, 2008 Elsevier B.V,Vol.8, 2008,pp.1326-1334