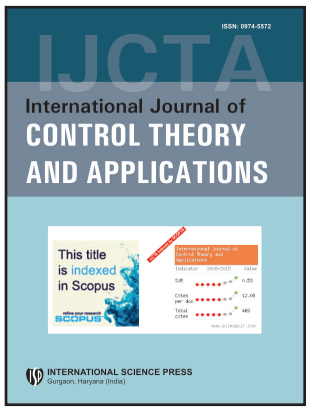

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 9 • Number 43 • 2016

Evaluation of Meta Search Engines

Naresh Kumar[a], Sonali[b] and Shivani Gupta[c]

[a-c]Department of Computer Science & Engineering, MSIT, Janakpuri, New Delhi, India. Email: [a]narsumsaini@gmail.com; [b]sonali850099@gmail.com; [c]gupta.shivani1296@gmail.com

Abstract: It is very challenging to manage and present the available information to user in an effective way. A number of researches using MSE have been done to overcome these limitations but no immense feasible solution has been found. So this paper emphasizes on study and problems associated with available literatures as well as existing online MSE's. Moreover this paper also presents some feasible solutions to address the stated problems.

Keywords: Search Engine, Meta Search Engine, Clustering, Ranking, Precision, Recall.

1. INTRODUCTION

Internet is developing rapidly and almost everyone uses internet. So, it becomes a very important part of life [1]. Number of web pages are growing briskly [2] with each passing day. So it becomes very troublesome for the users to find relevant information which satisfy their needs and interests. The search engine (SE) is a very efficient tool designed for users so that they can get relevant information [3]. But it may also be possible that while browsing search engine, it may not return relevant results because any SE can cover only 16% of the entire web [4]. So the remaining 84% part of web remains unreached which may have desired results required by the users. At this stage it becomes very difficult to find desire information due to information overload [5]. A study of [6] shows that each SE have different coverage area. So it becomes beneficial to use multiple SEs on a single interface which is called as Meta Search Engine (MSE). To represent the results in an effective way MSE uses clustering of results. That is why the main focus of this paper is to introduce the problems related with MSE clustering. Clustering is a technique that groups similar documents in clusters. Clustering algorithm has been considered very useful technique for extracting information from large amount of datasets [7]. Many clustering algorithms like K-means, fuzzy c-means, hierarchical and density based etc. have been introduced to organize the results effectively. By using clustering, user can identify the required group at a glance by looking at the cluster labels. Clustering has various applications which are text mining, data mining, machine learning, image processing and bioinformatics [8].

The rest of paper is organized as; section II describes available literature on MSEs and MSEs available online to search information from www. Section III describes the general architecture of MSE and available

clustering techniques. Section IV addresses the problems of clustering based on literature study and existing MSE whereas section V results the feasible solutions to address the problems. Section VI presents conclusion and future scope of the paper.

2. EXISTING WORK

An extensive literature survey has been done on MSEs available in literatures as well as available MSEs online. So, this section is categorized in two parts which are as described below.

A. Literature Study

In [1], the application of a data mining tool WEKA was presented and authors applied K-Means clustering to find clusters from large data sets. For clustering, data is collected from various websites' source code. Clusters are formed on the basis of title length, number of keywords, URL length and number of back-links and in-links of the websites. The main feature of the proposed system is that if more data are collected from websites, then user can get more and relevant results. At the end, authors claimed that by using this method > 60000 back-links, < 50 title length, > 3 title keyword, < 25 URL length and >200 in-links can be found which is good for optimization.

A system based on personalized search called DUCFF was proposed in [3] which represents dominant user context filtering framework. The user profile, user interest, user search history, semantic knowledge, user short-term information, user session, content interest and ontology of metadata were used for personalization. The method consists of initial study, data collection, data analysis and confirmatory study was performed to measure the success of proposed framework. Five queries-handicap, cookie, crocodile, car and games were used to test the proposed framework. Results shows that 6 groups of DUC(Dominant User Context) were found for keyword handicap, 5 for cookie, 5 for crocodile, 6 for car and 8 for games in data collection.

The significant characteristics of the proposed framework were as follows:

i) Increased relevance of the suggestions by providing personalized results.

ii) The method uses semantic knowledge of keywords.

iii) User context was used for formation of clusters.

At the end, authors affirmed the proposed system can match the recommended results somewhere in between 79 to 100 percent.

Authors of [5] proposed a weblog extraction with fuzzy classification method that uses folksonomies and fuzzy clustering algorithm. It was used for results of related relevant terms. Fuzzy clustering algorithm is extension of traditional set theory i.e. an element may or may not belong to a set based on binary terms. The main parts of fuzzy weblog extraction are interface, query engine, MSE, aggregated documents. Authors used OLED, OEL, LCD and LED terms in four different weblogs. By using Boolean search; user can find weblog containing OLED only. But by using proposed fuzzy search, user can find search terms as well as related terms. Hence two weblogs will be given to user in two different folders; one folder is labeled with OLED and other with LED.

A new, data density based approach called DDC was suggested in [7]. The number of clusters was automatically determined by the proposed method. For each data sample, the offline computation was significantly reduced by using RDE.

The quality features of the proposed system were:

i) Since the number of calculations required to calculate the density has been minimized greatly, DDC can be efficiently applied to Big Data.

ii) The main advantage of DDC lies in the fact that it does not require the pre-defining of thresholds means it varies situation to situation.

iii) Since the final cluster radius was adjusted based on the actual cluster data spread, hence it is less sensitive to the initial radii user input.

Authors of [9] introduced a new method to produce clusters of web documents which extract terms from URL, Title Tag and Meta Tag. Authors selected these parts of a web document because they contain keywords which provide major information contained in that webpage. K-means clustering algorithm was used for the formation of clusters.

Further the experimental results demonstrated that the proposed system produced clusters of closely related documents

The highlights of the suggested method were:

i) The method results into minimum intra-cluster distance and maximum inter-cluster distance.

ii) The vectors produced by the proposed method were not too long and also it took less processing time.

iii) High quality clusters were generated since the terms in URL, Title Tag and Meta Tag have greater significance in providing identification for the web page.

A new framework for cluster generation of Non-Text User Generated Content (UGC), such as videos and images, using keyword similarity was proposed in [10]. In the proposed framework, the search results with similar content were segregated into small groups. After processing the metadata of the search results, the proposed algorithm returned all the possible relationships of related content. Markov Clustering Algorithm was used to cluster the web results.

Experimental analysis revealed that the overall efficiency of the proposed methodology was between 50%-70%;

The distinctive attributes of the suggested system were:

i) The desired results could be found easily and quickly.

ii) Computation of search results and clustering of similar content was done automatically, which leads to results optimization and higher accuracy.

An intelligent cluster search engine called ICSE was proposed in [11] which uses a fast tree based search algorithm and Meta-Directory tree for knowledge base. The main modules of ICSE include MSE, meta-directory tree, web pages clustering and topic generation module. Twenty queries were used to test the proposed system which has average access time 507.54ms.

The characteristics of ICSE were as follows:

i) More relevant results can be obtained because system combines the results from MSE and information from directories like Google, ODP and Yahoo. So it provides the benefits of MSE and directory on a single interface.

ii) This system can handle large data sets since it has access to MSE as well as to directory.

iii) Since the system uses Meta-Directory tree, so computation time for clustering reduces significantly.

iv) When the query is executed repeatedly the response time is almost instantaneous because the system uses cache results from database.

At the end, authors put forth that proposed system is 5 to 20 times faster than other existing system.

An image meta search engine called ITSE was put forward in [12].The basic idea of the system is to greatly improve the coverage of images by combining the power of several standard SE. The proposed MSE adopted a method of vectorization and introduced HACM clustering techniques on images search. Further the authors optimized the results by using a genetic algorithm. A more significant and restricted set of images was generated by the MSE and represented as the final result for an end user.

The experimental results showed that ITSE achieves 80% accuracy after classification of web results.

The significant aspects of the established procedure were:

i) ITSE had adopted HACM clustering and genetic algorithm to classify and organize the search results, thereby providing a rich set of images to the user.

ii) ITSE search engine returned more relevant results to the query terms.

Authors used content mining technique and proposed a system [13] called WISE which uses hierarchical soft clustering of webpages to organize the search result. The document processing was performed based on concept extraction and phrase extraction. The results were organized into clusters by using an algorithm; named as PoBOC soft clustering algorithm.

The significance of WISE was stated below:

i) It ignores irrelevant documents and emphasizes on relevant documents only.

ii) It allows a single document to be in more than one cluster to organize results into a hierarchy of concepts and a classical labeling process.

iii) Algorithm uses phrase extraction to define key concepts that represents web document.

iv) It can represent the documents semantically for further analysis.

At the end results shows the correctness of cluster generation, quality of labels, unambiguity of concept and independence of language.

Authors of [14] came up with a new ranking method to cluster the results of a search server and define a descriptor for each cluster. The proposed approach emphasized the fact that the cluster descriptors provide a more precise representation of the document space.

The prominent characteristics of the suggested method were:

i) The relevance of a searching server to a user query term is estimated more accurately and effectively if the web documents covered by a SE are clustered and represented by cluster descriptors.

ii) Cluster-based server selection method is highly appropriate for the large-scale meta-search systems.

To characterize the web results with acceptable accuracy, it is not necessary to cluster the web documents in the index of a large number of clusters. Instead, it is just sufficient to describe a moderate number of clusters. This results in reduction in the clustering cost as well as reduction in the storage spaceauthor.

B. Study of Existing Meta Search Engine

This subsection includes the study of Yippy, eTools.ch, Carrot2,qksearch and iBoogie MSEs which are very popularly used in searching. This study includes only those MSEs which are using clustering technique only and problem in these clusters are based only on the experience and problems faced by the authors of this paper.

i) Yippy (formerly known as Clusty) isa research based MSE developed by Vivisimo [17]. It uses Ask.com, Gigablast, Live, NY Times, ODP, Shopzilla, Yahoo news and Yahoo stock. It has different tabs to search jobs, news, US govt., images and blogs.

ii) Carrot2 is an open source, text mining and clustering MSE tool which uses Lingo Clustering and K-Means Clustering Algorithm to cluster the search results [18]. It fetches its results from Google, Bing, eTools MSE, Lucene, SOLR and etc. As per our experience the relevancy of results is high.

iii) iBoogie [19] is a document clustering MSE with customizable search type tabs. It uses all the web and MSN for their result generation. It allows user to add their sources such as specific library, scientific, government and educational searches to their database.

iv) The SEs used by OpenText[20] are Google, Yahoo, Bing, Ask, Wikipedia and Open Directory. It displays the number of results and time taken to fetch these results from each source. It improves productivity of user and search breadth also increases. It perform grouping of result on the basis of content or location.

v) qksearch calls itself a three in one MSE which performs clustering search, blend search and split search [21].

vi) eTools.ch is a Swiss MSE which emphasizes on enhancing the search results by producing relevant clusters and provides full privacy [22]. It queries 17 search engines which are Ask, BaseBing, DeuSu, DuckDuckGo, Exalead, Faroo, Fastbot, Goo, Google, Moose, Search, Tiger, Webliste, Wikipedia, Yahoo and Yandex.

3. META SEARCH ENGINE AND CLUSTERING TECHNIQUES

Meta search engine is a search tool (as shown in Figure 1) which makes use of other SEs to produce effective and relevant results for a user query term. The coverage area of popular SE's such as Google,Yahoo! And Alta Vista is low and results are also low in precision, Therefore MSEs are introduced. User types a query in UI provided by MSE which submits query to several other SEs such as Google, Bing, Yahoo! and Alta vista and results returned by these SEs are fused to form clusters and these clusters are returned to user as results [4]. Advantage of using MSE is that there is an amplification of coverage area. MSE uses indexing of other SE to maintain results [4], form clusters and provide results to users in an innovative way with relevant labels [5]. It is a useful approach for user if he wants to gain overview of the topic or to extract quick answers. . Instead of these advantages, sometimes it may also be possible that results returned by SEs are more relevant than that of MSEs. MSE are not able to translate query syntax completely and do not interact with large search engines for their result formation.

MSE uses several other search engines for their results [4]. In turn SE visits number of websites and create their own database to store different results that is known as indexing. So MSEs take the advantage of other available SEs. But different MSEs use different algorithms to arrange and present the results of SEs. Clustering is a technique that is used by MSE to present their results so that user obtains relevant results in excellent form. Some MSE also uses different ranking algorithms so that user receives outstanding results without any complication [4]. Currently available MSEs are Dogpile, Sputtr, Clusty, ixquick, Mamma and etc.

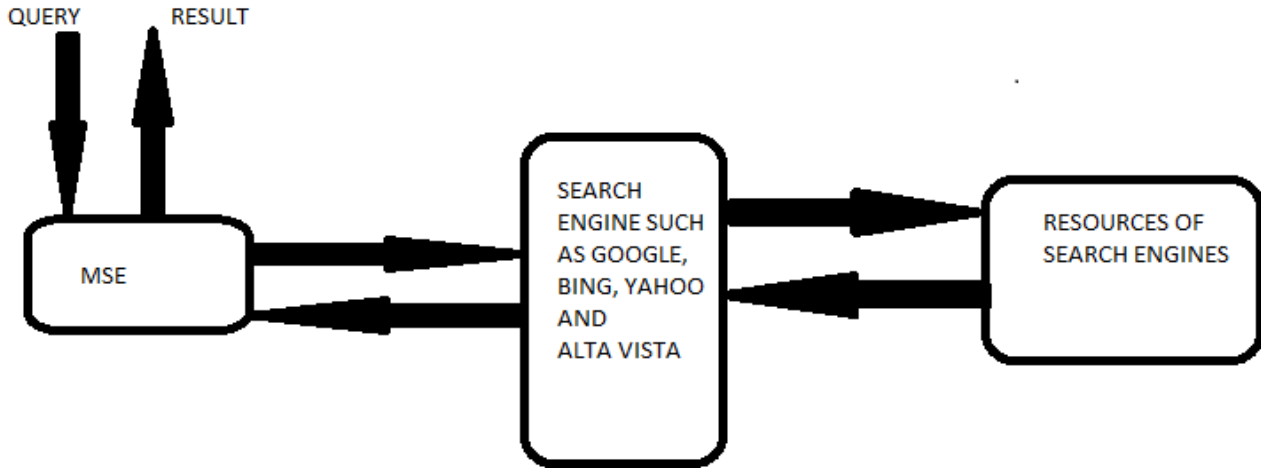

Figure 1: General architecture of Meta Search Engine

The following are the parameters that affect the efficiency of MSEs [16]:

i) Number of SE: Quality of the results is not evaluated by the number of SE used but quality of results returned by these SE's.

ii) Relevancy of results: Any result is beneficial if it is relevant to the user and he receives quality and precise information of his interest.

iii) Clustering: clustering is a technique in which similar documents are grouped together to form clusters.

iv) Ranking: It is a technique in which links of higher priority occurs first than that of low priority.

v) Re-Ranking: In it links are ranked once again by the search engines so that user can get leading links at the top of resultant list.

Dynamic clustering: Formation of clusters is based on the similarity between query and documents.

Cluster Analysis is the process of categorizing a collection of objects in a way such that the similar objects are in the same group [6]. These formed groups of similar data are known as clusters which differ in characteristics. It is used as a data mining tool to group web results into relevant clusters so that end users can easily identify their required data by looking at the cluster labels [9]. Following two criteria should be satisfied to achieve effective clustering as stated in:

i) Web results should be similar to documents in the same cluster i.e. intra cluster distance should be minimum.

ii) Web results should be dissimilar to documents in the other clusters i.e. inter cluster distance should be maximum.

The following are the methods of clustering [8]:

1. Partitioning Method: In Partitioning Method, a database of 'n' web documents is partitioned into 'k' clusters and $k \leq n$. These partitions satisfy the following requirements:

   (i) At least one search result must belong to each cluster.

   (ii) Each resultant document must be contained in a single cluster only.

2. Hierarchal Method: This method involves decomposition of the document space in a hierarchical manner. It is a rigid method i.e. once splitting or merging is done, it can't be undone. It involves the following two approaches:

    i) Agglomerative Approach: It is bottom-up approach in which similar documents or groups are merged together. This step continues until all the groups are merged into.

    ii) Divisive approach: It is a top-down approach in which a cluster is broken down into small clusters. This step continues until each cluster holds only one document

3. Density Based Method: This method is based on the concept of density of documents. The regions of higher density are defined as clusters. Documents in low density areas are usually considered as outliers and are used to keep the clusters separate.

4. Grid Based Method: In this method, the document space together forms a grid structure with finite number of cells. It processes the results at a very fast rate.

5. Constraint Based Method: This method performs clustering by fusing the user outlook with the characteristics of the formed clusters. It allows the user to enhance the clustering results according to their requirements.

6. Model Based Method: It is a robust clustering method in which each cluster is hypothesized by a model to determine the best fit of documents for a given model. It automatically computes the number of clusters depending upon the standard statistics, considering outlier and noise.

4. CHALLENGES ASSOCIATED WITH MSES

The following are the challenges experienced by the authors while literature study and study of existing meta search engines:

i) Accurate and Authentic identification of user context and user behavior is needed because formulation of framework as proposed in [3] is done on the basis of this information. FAQ are adopted to collect information about user context and behavior and since it is questionnaire survey it may not contain accurate information. Further the information should be organized in such a manner so that it matches to a particular context; it is a crucial problem of [3].

ii) The number of clusters defined by users in fuzzy clustering algorithm is a bad practice because user does not have exact approximation of search results returned and number of clusters to be formed for them [5]. It uses HITS algorithm which is obsolete nowadays.

iii) In [7], DDC technique is slow and not suitable for small dimensional data sets.

iv) The major challenges of [9] are as:

    i) Clustering is performed without using the knowledge of semantic relationship between the web documents which may lead to the formation of irrelevant clusters.

    ii) The system depends on the user to define the number of clusters. This poses a problem as user is unaware of the number of clusters to be formed to cover all the search results.

    iii) The proposed method extracted terms from URL, Title Tag, Meta Tag to formulate the results into clusters. However, it may be possible that these parts of a web page may not cover all the keywords required to describe a web page completely.

v) The suggested algorithm performs stemming as one of its steps which generally leads to the production of meaningless words.

v) Group Clustering stage in [10] leads to the overlapping groups in a cluster, which affects the originality of a group.

vi) Topic generation module in [11] assumes that words at the beginning and ending are more important than the content available in middle of any webpage but for some cases this is not always true. Further it does not support the concept of ontology (i.e., semantic knowledge is not used to analyze keywords) and web 2.0. Moreover it uses offline processing to generate and sustain meta directory tree. Therefore it takes more time for human being to renovate these directories. Hence it may not provide up-to-date information available on internet to users.

vii) Though ITSE in [12] there is enhancement in the accuracy of the formed clusters, however it takes more time when compared to other search engines i.e. it is slow for high dimensional data set. Moreover it may not be easy to identify the correct number of clusters by the dendogram formed in the last step of HACM.

viii) Extraction method in [13] selects relevant web pages on the basis of retrieval of absolute URLs and URL whose number of occurrences beat average relevance. The problem is that it may not cover all the relevant documents and satisfy user needs. It uses PoBOC algorithm which allows a single document to be present in more than one cluster. Therefore it does not satisfy the identity and unity of clusters.

ix) Overlapping clusters are generated by [17] i.e. a web document may be present in more than one cluster. This affects the originality of a cluster. Moreover, cluster labels are not descriptive enough to provide quality information about the results contained in a cluster. Furthermore, ranking of clusters has not been done efficiently and effectively. As a result, relevancy of results is low and also it contains too many sponsored results.

x) A search result in [18] may be present in more than one cluster which causes the formation of poor quality overlapping clusters. It may also be possible that a resultant cluster may contain a single document more than one time due to which redundancy increases within the cluster.

xi) Literature [19] generates overlapping clusters. As a result of which relevancy of results is low.

xii) A cluster generated by [20] may contain the same documents multiple times which leads to redundant results and makes it less memory efficient. Furthermore, clusters lack in relevant documents. Hence, [20] fails to generate high quality and pure clusters.

xiii) The major problem of [21] is that for every query term, it displays the same message "Our server could not serve your request. Please try again later." So, it is useless and hence wastage of search time at all.

xiv) A cluster may contain the same web document more than once which causes the generation of unnecessary and irrelevant results by [22]. Besides this, labels of few clusters are not self-descriptive. Moreover, a web document may be assigned to more than one cluster which affects the quality of the clusters by generating overlapping clusters.

5. FEASIBLE SOLUTIONS TO CHALLENGES OF MSES

Possible feasible solutions are also presented by the authors of this paper which are as follows:

i) In [3], FAQs can also be based on case studies as it may provide user experience in his search domain which will enhance the accuracy of DUC filtering framework.

ii) Instead of using HITS algorithm in [5], a data mining tool such as WEKA can be used to identify the number of inlinks and backlinks.

iii) K-means algorithm in [9] and FCM algorithm in [5] can be replaced by identity based clustering algorithm in order to eliminate the dependency on user to define the number of clusters.

iv) Work in [9] can be extended by considering semantic relation between the documents which possibly will lead to the generation of more accurate clusters.

v) By introducing the concept of web 2.0 to the system proposed in [11], web technology will be improved which eases the process of communication and increases the safety of shared information.

vi) Topic generation module in [11] can be made effective if semantic analysis is performed for all the keywords of the webpage, whether they appear at beginning, middle or at the end of any webpage. Semantic analysis is performed to examine how keywords in a webpage related to each other. Therefore by using this technique topic generation module can be improved to generate better clusters with informative labels.

vii) In [12], HACM algorithm can be substituted by subtractive clustering algorithm or density based clustering algorithm so that less time is consumed to process the vast document space. This may also lead to the easy identification of the number of clusters to be generated.

6. CONCLUSION AND FUTURE SCOPE

In today's world it is impossible to control the amount of data being created and uploaded on web. MSEs can be used to manage this huge data. However, it is always challenging to develop an MSE that produces absolutely effective results. This paper is based on literature study and existing MSEs. The main objective of this paper is to reduce the problems of clustering and authors found various problems and presented some feasible solutions to improve these techniques of clustering. Currently, authors are working on the implementation of an effective clustering based MSE which may provide much more relevant results to the user when compared to the existing ones.

REFERENCES

[1] M. Jindal and N. Kharb, "K-means Clustering Technique on Search Engine Dataset using Data Mining Tool," *International Journal of Information and Computation Technology*, Vol. 3, issue 6, pp. 505-510, ISSN: 0974-2239, 2013.

[2] The WorldWideWebSize website [Online]. Available: http://www.worldwidewebsize.com/

[3] N. A. Kadir, A. M. Lokman and A. Ahmad, "Dominant User Context (DUC) Filtering Framework for Web Personalized Search," *IEEE Symposium on Wireless Technology and Applications*, pp. 280-285, ISSN: 978-1-4673-2210-2, 2012. DOI: 10.1109/ISWTA.2012.6373862.

[4] N. Kumar and R. Nath, "A Meta Search Engine Approach for Organizing Web Search Results using Ranking and Clustering," *International Journal of Computer*, Vol. 10, issue 1, pp. 1-7, ISSN: 2307-4531, 2013.

[5] E. Portmann, "Weblog Extraction with Fuzzy Classification Methods," *Second International Conference on the Applications of Digital Information and Web Technologies*, pp. 411-416, ISSN: 978-1-4244-4457-1, 2009. DOI: 10.1109/ICADIWT.2009.5273930.

[6] N. Kumar, R. Nath and P. Kherwa, "An Automated Framework based on Three Level Scoring Method to Choose Best Search Engine in a Particular," *International Journal of Computer Applications*, Vol. 83, issue 14, pp. 42-48, ISSN 0975 – 8887, 2013.

[7] R. Hyde and P. Angelov," Data Density Based Clustering" *Computational Intelligence, 14th UK Workshop*, ISSN: 978-1-4799-5538-1, 2014. DOI: 10.1109/UKCI.2014.6930157.

[8] L.V. Bijuraj, "Clustering and its Applications," *Proceedings of National Conference on New Horizons in IT NCNHIT*, pp. 169-172, ISBN: 978-93-82338-79-6, 2013.

[9] S. Poomagal and Dr. T. Hamsapriya," K-means for Search Results clustering using URL and Tag contents," *International Conference on Process Automation, Control and Computing*, pp. 1-7, ISSN: 978-1-61284-764-1, 2011. DOI: 10.1109/PACC.2011.5978906.

[10] P.Juasiripukdee, L. Wiyartanti and L. Kim," Clustering Search Results of Non-text User Generated Content," *Digital Information Management, Fifth International Conference*, pp. 95-100, ISSN: 978-1-4244-7571-1,2010. DOI: 10.1109/ICDIM.2010.5663293.

[11] C.W. Tsai, K. Huang, M. C. Chiang, and C. S. Yang, "A Fast Tree-Based Search Algorithm for Cluster Search Engine," *Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics San Antonio, TX, USA*, pp. 1603-1608, ISSN: 978-1-4244-2794-9, 2009. DOI: 10.1109/ICSMC.2009.5346100.

[12] H. J. Li and J. K.Wang," Precise Image Retrieval on the Web with a Clustering and Results Optimization," *International Conference on Wavelet Analysis and Pattern Recognition, Beijing*, Vol. 1, pp.188-193, ISSN: 1-4244-1066-5, 2007. DOI: 10.1109/ICWAPR.2007.4420661.

[13] R. Campos, G. Dias and C. Nunes, "WISE: Hierarchical Soft Clustering of Web Page Search Results based on Web Content Mining Techniques," *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 301-304,ISSN: 0-7695-2747-7, 2006. DOI: 10.1109/WI.2006.201.

[14] Y. Shen and D. Lee," A Meta-search Method Reinforced by Cluster Descriptors," *Web Information Systems Engineering, Second International Conference*, Vol. 1, pp.125-132, ISSN: 0-7695-1393-X, 2002. DOI:10.1109/WISE.2001.996473.

[15] J. C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms," *Plenum Press, New York*, 1981. (http://link.springer.com/book/10.1007%2F978-1-4757-0450-1)

[16] S. Ball and R. Nath, "To Evaluate the Performance of Meta Search Engines: A Comparative Study," *Journal of Technology and Engineering Sciences*, Vol. 1, issue 1, pp.70-78, 2009.

[17] The yippy website [Online]. Available: http://www.yippy.com/

[18] The carrot2 website [Online]. Available: http://search.carrot2.org/stable/search

[19] The iboogie website [Online]. Available: http://www.iboogie.com/

[20] The opentext website [Online]. Available: http://www.opentext.com/

[21] The qksearch website [Online]. Available: http://qksearch.com/

[22] The etools website [Online]. Available: https://www.etools.ch/