Acquiring the Best Page using Query Term Synonym Combination

Lobo L. M. R. J
Research Scholar, Department of C.S.E
SGGS Institute of Engg & Techology, Nanded
Associate Professor, Department of C.S.E
Walchand Institute of Technology, Solapur, India

R. S. Bichkar
Professor, Department of E & TC
G. H. Raisoni College of Engg & Management,
Pune, India

ABSTRACT
A large amount of information is available on the web. Generating relevant information from the web for a user has become a question of great concern. This information is made available to the user by search engines as per the query given by a user. Search engines return pages depending on a ranking algorithm based on links, to and from the page and on how popular a page is, with respect to the hits received by users. In most cases the pages returned are too many and irrelevant. It is therefore necessary to use a technique that caters to relevance of the page returned with respect to the query fired by a user. A system associated with use of synonyms of terms in the query has proved to be useful. A combination of such synonyms fired to a search engine has returned relevant information pages. In most cases it has also generated a better page than an individual search engine.

The present paper reflects on how synonyms of terms generated from the query are crossed over and fired to the search engine to generate more relevant semantic associated pages. These pages are then tested against the pages returned by the individual search engine with use of the original query using a standard page ranking tool. The pages are also examined for relevance to specific users and usefulness of content to a specific domain. The pages are examined for their positions using ranking tools, trustworthiness tools and intent drifting. It is found that the pages returned using the method of combining synonyms of terms of the user query are placed at better ranking positions. An analysis of the pages returned also indicates relevance to user, usefulness of content to specific domains and possibility of intent drift.

General Terms
Data Mining, Genetic Algorithms, Web search, Soft computing.

Keywords
Best Page, relevance, users' interest, synonyms, ranking.

1. INTRODUCTION
Without the use of internet life has almost become impossible. The internet is needed to surf and browse sites to get us desired and updated information. Information that is relevant and available in a flicker of time is more appreciated by a user. These demands for best web pages to be identified not only in terms of links associated with them as presented by Brin and Page [1], or using linkage information as in Hou and Zhang [2] and giving a rating to these pages in the form of a rank based on the probability of citations of a page, a damping factor and the number of links leaving a page but by going a step ahead to examine the basic features of a page [3].

An important fact in these scenarios is that the information retrieval should be based upon relevance and proper representation as per the requirements of a specific user [4].

Once such pages of good quality are sampled out, a method is required to compare between similar pages. This may involve using indexing, links and agent search techniques [5]. Jaccard scores could be used to compare similarity between two pages [6]. A best first search algorithm could then be used to identify the best pages. Genetic algorithms have a very close resemblance to an adaptive web search [7].

A users' visit to a page is an important parameter to be considered [8]. This was considered as a measure of popularity of a page. Using machine learning methods to generate features by using a co-occurrence matrix analysis [9] and to classify web pages automatically gained importance. Here web pages were set on constructed decision trees which determine appropriate category for each web page. Here consider parameters like error rate, precision and recall for evaluation of a page.

The web content designed nowadays stresses on being very user friendly. It is for human reading and is supposed to be relevant to a user. The semantic web has to provide structured content by adding annotation tools that are available [10]. Annotation is the concept to associate semantics with a file [11]. Usually when data is in image form it is captured by a camera, it gets stored with the filename as a numbered image. To perform search with such filename and to retrieve those image files is tedious. To improve the searching technique semantic file annotation is implemented [12] which annotates the image and retrieves the required file. XML format can also be viewed as a browsing list on the mobile screen. At the same time, it also allows users to edit or refresh the meta-data at any time [13]. A personalized search is one of such examples where the web search experience is improved by generating the returned list according to the modified user search contents [14], [15].

The use of synonyms reduces irrelevant search. It actually does not cause intent drifting every time. Synonym discovery is context sensitive. The operations we use on synonyms are different from stem specifications. Users are not sure of how to phrase queries to be fired to search engines to return desired results and hence using synonyms of the terms in the queries prove useful. Sometimes pools of synonyms may be created and sampled synonyms may be fired. Synonyms are provided along with vocabulary in some systems. A dictionary of synonyms may sometimes be plugged into a search engine to improve the quality of search results. Synonyms help to prepare good indexing logs and search reports. Synonyms are used for schema matching. All these characteristics of synonyms motivate to devise a method using synonym combinations in order to generate relevant results at a higher ranked position.

The method presented here for searching the best page is based on the synonyms of the terms (words) used in the query

given as input to the system. The basic procedure adopted is initially to take a query from the user. The query is then split into its subsequent terms (strings). The terms are passed one by one to a dictionary. Tables of all the synonyms of terms are then created and these elements are passed to the search engine parser. The search result is then generated. The elements in the table are then crossed to form new elements choosing a random crossover points and exchange the result query. Changing specific table elements if they would result in a better query is sometimes done. These operations can be continued till all combination of input terms are dealt with or the results required by the user are achieved. Some of the things implemented into the method are a directory based search mechanism of previous searches, an adaptive mechanism of machine learning by users inputs and a Jaccard score for ranking web pages. In directory based search mechanism, the database of the previous searches is maintained and the directory is shown to the user if it matches the string of previous searches.

The pages returned by this method are checked for relevance to particular users, categories of people and the ranking given by different page-ranking tools. It is observed that the best pages generated by this method are more relevant to a particular user, more relevant to desired categories and in most cases are ranked better to the best pages returned by individual popular search engines. The analysis done also indicates conditions under which an intent drift is possible for a grammatical combination of words.

The subsequent sections of this paper include related work, a methodology and experimental setup with results generated for some experiments performed.

2. RELATED WORK

It is reported that Netscape uses web-page content analysis, usage pattern information and linkage analysis to determine the best page. Hyper-link analysis is extracted by parsing the web page codes. Hyper-links reflect semantic judgements that are objective and independent of synonymy of words used for page ranking in Google and relevant page finding. Two problems are encountered one construction of page source related to a given page and second establishing an effective algorithm to find relevant pages from the page source. The page source should have relatively small size (number of pages) and should be rich in relevant information.

Kleinberg's algorithm used hyperlink induced topic search. These algorithms find authority pages. They must be constructed properly, and hence source must be constructed appropriately to avoid topic drift. Dean and Hithmenzinger's algorithm used a page source that consisted of sibling pages of a given page. The algorithm was based on co-citation analysis and similarity between a page and given page is measured by number of common parent pages, named co-citation degree.

Pun and Lochovsky used cohesiveness for finding high quality web pages. They have defined a distance matrix to measure the closeness in the ontology. The matrix is used to calculate the cohesiveness of a page as the total distances of the entire concept in it. Cho et al. defined page quality as the probability of link creation by a new visitor.

Hou and Zhang [2] have used linkage information for effectively finding relevant web pages. Bharathi and Venkatesan [16] showed that when a user inputs a simple keyword query to a search engine it returned results with low precision, which is due to the irrelevance and low recall, due to the inability to index all the information available on the web. Here synonyms of the query term was used so that from the retrieved documents of the data set the correlated semantic terms of the specified query term was identified and finally more similar documents were ranked based on semantic correlation similarity. This improved the accuracy of the retrieved relevant documents without much increasing time. Choudhary [17] used search with synonyms as a challenging problem for web search, as it could easily cause intent drifting since synonym discovery is context sensitive. High quality synonyms have the same or nearly the same meaning only in some senses. If we simply replace them in search queries in all occurrences, it is very easy to trigger search intent drifting.

Madhu et al. [18] have used typically domain specific knowledge. Roul and Sahay [19] brought about a method of finding the synonyms of frequent words in the Word-Net database, and adding the synonyms to the pool of frequent terms that comprise the cluster label candidates. The detection of synonyms helped in grouping together snippets that contained different but synonymous words that would otherwise have not been grouped together using the original Lingo algorithm. Beel et al. [20] boldly remarked that to their knowledge, none of the major academic search engines currently considers synonyms. Google Scholar does not index text in figures and tables inserted as raster/bitmap graphics, but it does index text in vector graphics.

Vasnik et al. [21] described a searching scheme with a specific keyword eventuating to unsatisfactory, but with its synonym to appropriate results exploiting only one semantic relation, such as synonym was not effective, so it was better that a combination of semantic relations to be used. Beel and Gipp [22] concluded that in all analysed full texts, the search terms that were used occurred at least once in the text. Accordingly, it can be assumed that Google Scholar abides strictly to an article text and does not consider synonyms. Since Google Scholar does not consider synonyms, users should think carefully about the terms they search for. Otherwise they could miss out on relevant documents. This may be considered an additional overhead.

Wei et al. [23] have verified that search with synonyms was a challenging problem for web search, as it can easily cause intent drifting. Hliaoutakis et al. [24] have shown that term similarity was computed by matching synonyms, term neighbourhoods, and term features. Li [25] has remarked that one could discover synonyms, extract new concepts, and build a thesaurus. Sudhakar et al. [26] indicated their observation that every root word is considered for Dictionary construction and a dictionary is built with synonyms for the user query every result page keywords and content words were pre-processed and compared against the dictionary. Cui et al. [27] showed that from a thesaurus constructed, one will be able to obtain synonyms or related terms given a user query. Thus, these related terms can be used for supplementing user original queries. Chakrabarti et al. [28] show how the Easy-Ask system supports a wide variety of features such as approximate word matching, word stemming, synonyms and other word associations. Tang et al. [29] showed existing linguistically-related methods find either synonyms or other linguistic-related words from thesaurus, or find words frequently co-occurring with the query keywords.

Most successful sites emphasize the important of interaction and service quality. Hasan and Abuelrub [30] examined the Chinese Websites of the Alexa ranking. Their findings of content analysis indicated that most firms use the marketing functions distributed evenly in service quality dimensions.

3. METHODOLOGY

The work performed by Bharathi and Venkatesan was a motivation for this work. They used a measure called the F measure of quality. The external quality measure combines the precision and recall ideas from information retrieval. The higher F measure is the higher accuracy of cluster. Here, each cluster was treated as if it were the result of a query and each class is treated as if it were the desired set of documents for a query. Then, the recall and precision of that cluster for each given class was computed. More specifically, for cluster j and class i

$$\text{Recall (i, j)} = \frac{\text{nij}}{\text{ni}} \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$$\text{Precision (i, j)} = \frac{\text{nij}}{\text{nj}} \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Where, nij is the number of members of class i in cluster j, nj is the number of members of cluster j and ni is the number of members of class i. The F measure of cluster j and class I is then given by

$$\text{F (i, j)} = \frac{(2 * \text{Precision(i,j)} * \text{Recall(i,j)})}{(\text{Precision(i,j)} + \text{Recall(i,j)})} \ldots\ldots\ldots\ldots (3)$$

The overall value for the F measure was computed by taking the weighted average of all values for the F measure as given by the following

$$\text{Fc} = \sum_i \frac{ni}{n} * \max \text{F (i, j)} \ldots\ldots\ldots\ldots .(4)$$

Where n is the total number of documents. Higher the value of F-measure better is the cluster quality. Precision P is defined as the proportion of retrieved documents that are relevant i.e. $\frac{Ra}{A}$. Recall is defined as the proportion of relevant documents that are retrieved i.e. $\frac{Ra}{R}$. A is the number of retrieved documents. R is the number of relevant documents. Ra is the number of retrieved relevant documents.

The method developed for this work [31] was based on getting more relevant documents based on word synonyms. An extension to this methodology is presented in this paper. The main objective to be considered in this scenario is to get the best page. Here relevance of the page to a particular user, a category based on type of useful information returned, a ranking given by standard tools indicating popularity of returned page, trustworthiness of a page determined by an expert on-line committee and type of Query based on grammatical word combination is of importance to be implemented.

The methodology followed is to initially consider an input query formed of terms. The query is such selected that it may be very vague or specific to a particular user category or a class of utility depending on the type of information returned. It may be a combination of grammar based words using a combination of different parts of speech or words any synonyms that will really be useful to a user.

This query is then broken into meaningful terms which are the constituent words forming the query. The words are then sent to a popular dictionary, Thesaurus. The dictionary returns synonyms *S1, S2 …, Sm*. The synonyms are listed in a tabular form. Each column of the table enlists the synonyms for a term in the respective row for the term. The words in the query determine total number of terms to be considered. These rows are now crossed at word completion positions to develop combination of synonyms to form new queries. These queries are then sent to a search engine like Google, Yahoo, Bing, etc.

The fetched pages from the search engine are stored and tested by Jaccard score for their fitness. The logic used to find the best pages is fist to input home-pages returned from the search engine and their linked home-pages are saved in H = $\{h_1, h_2, \ldots h_k\}$. Initialize the count for home-pages k to 1. Now the best homepage has to have the highest Jaccards score among all the homepages. It is then stored as the output_k. The Jaccard score is computed as

$$JS_{\text{links}}(h_i) = \frac{1}{N}\sum_{j=i}^{N} JS\ links\ (\text{input}_{i,}, h_i) \ldots\ldots (5)$$

Where, $JS\ links$ (input_j, h_i) represents the Jaccard score between input_j and hi based on links. Similarly for indexing of pages the Jaccard score is computed as

$$JS_{\text{index}}(h_i) = \frac{1}{N}\sum_{j=i}^{N} JS\ index\ (\text{input}_j, h_i) \ldots\ldots\ldots (6)$$

The Jaccard score for hi is then computed as a sum of the Jaccard scores of that for links and index for a homepage. Fetch the best homepage and add all its linked home-pages to H and increase k by 1.Repeat this procedure till all output home-pages are obtained. All fit pages are accepted and their ranking is displayed. Now all the pages returned are stored in a file. The original query term is then passed to search engines Google, Yahoo and Bing and the pages returned are also stored in a file.

The approach has a time complexity of O (m . n) for a query consisting of two terms , where m is the number of synonyms of first term and n is the number of synonyms of second term, which are comparisons in the crossing of a worst case data instance. For a query of two terms where the same number of maximum synonyms are returned by the dictionary say m the worst case complexity is O (m^2). In general for a query on n terms and m synonyms returned for each term by the dictionary the worst case complexity is O (m^n).

The analysis for these stored home-pages are then carried out.

3.1 Analysis based on relevance to a user:

The home pages returned for a vague or specific query are categorised as per the relevance they will have for different categories of users like a computer professional, a non-technical person, a novice, a sports person or a student. The percentage average number of pages returned by each search engine for each user is calculated as

$$\left(\sum_{j=1}^{N} \frac{Xj}{N}\right) * 100 \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (7)$$

Where, Xj is the number of pages returned by the search engine for a particular user and N is the number of total pages returned for the query term synonym combination. This acts as an estimate of relevance to a user.

3.2 Analysis based on type of useful information returned with varied content:

The home-pages here are analysed for whether information returned is of social network, database/knowledge base, product related, article/blog, book/publication/software, lifestyle etc and a browsing History may be recorded. The percentage average number of pages returned by each search engine for each category separated by its processed content is calculated as

$$\left(\sum_{i=1}^{N} \frac{Xi}{N}\right) * 100 \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (8)$$

Where, Xi is the number of pages returned by the search engine for each category separated by its content and N is the number of total pages returned for the query term synonym

combination. This can be used as an estimate to detect relevant content.

3.3 Analysis based on Ranking Tools:

Here ranking tools like Alexa Rank, WooRank and Google rank are used to check the popularity of the page with standards. A common estimate would be really useful which would combine the effect of ranks given for a page by different ranking tools and assigning weight-ages to these tools as per their importance.

The weighted sum calculated for giving an estimate of relevance of the page is given as

$$\sum_{i=1}^{N} Ri * Wi \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (9)$$

Where, Ri denotes the ranks given to a page by different ranking tools and Wi denotes the weight's given to the different ranking tools.

3.4 Analysis based on Trustworthiness of a returned Page:

Here tools like Web of Trust (WOT) are used, which tells which websites can be trusted from those available on the internet. This gives a user internet safety. WOT is a safe tool that provides website ratings and reviews. WOT secures a user against scams, malware, rogue web stores and dangerous links. If a poor reputation site, based on user ratings is detected, WOT shows a warning.

3.5 Analysis based on building queries by a intelligent combination of Grammar based words:

Here queries may be built with adjective-noun, Adverb-noun, verb-noun, etc. combination and some words and synonyms for which this will be really useful are identified.

4. EXPERIMENTAL SETUP AND RESULTS

The system is developed in JAVA. The Thesaurus dictionary is used to generate the synonyms of the words. JSOUP is used to fetch synonyms from the dictionary which is in Cascaded Style Sheet (CSS) format. JSOUP is an open-source Java library of methods designed to extract and manipulate data stored in HTML documents. It uses CSS and J-query-like methods for extracting and manipulating files. Terms (words) in the query are separated and the synonyms of terms are found and placed in tables. Combination of synonyms are generated by crossing rows from table at a specific word separation points and fed to the Parser of the search engine. The search engine returns pages. These pages are ranked using a Jaccard score.

It is clear that all the existing systems use search engines to extract pages. There are number of search engines available and these search engines use synonyms to retain context for the delivery of required content. However, by actually giving a query to three well known search engine and our system it is seen that an analysis can be done on this pages to rate them to be best based on different criteria. The results of each criterion are presented her

4.1 Results of analysis based on relevance to a user:

In the first experiment we have considered some vague queries like ’difficult level’, ’crisp summary’ and ’simple gifts’ and some specific queries like ’Foreign key’ and ’shinning stars’. The home pages returned by Google, Yahoo, Bing and presented system were stored and analysed as per the relevance to different categories of users like a computer professional, a non-technical person, a novice, a sports person or a student and the observations for query ‘difficult level’ are shown in Table 1.

Table 1: Results of analysis based on relevance to a user

Search Engine	Total results	Cat 1: Social Nets	*Cat 2: Dbase /K*base	Cat3: Product	Cat 4: Articles/blogs	Cat 5: Books/Pub
Google	363	212 (80.6%)	37 (10.2%)	06 (2.28%)	85 (23.41%)	23 (6.33%)
Yahoo	89	69 (77.5%)	05 (5.69%)	00 (0%)	14 (15.73%)	01 (1.12%)
Bing	499	382 (76.5%)	39 (7.81%)	02 (0.04%)	57 (11.42%)	19 (3.80%)
Presented System	182	64 (35.2%)	27 (14.83%)	15 (8.24%)	64 (25.16%)	12 (6.5%)

It is observed that the query ‘difficulty level’ was more relevant to a non technical person and a student. Query ’crisp summary’ was more relevant to a non technical person, query ’simple gifts’ was more relevant to a non technical person and was least relevant to a novice or a sportsman. Query ‘foreign key’, being a specific concept in databases was more relevant to a computer professional.

This is concluded because the average percentage number of pages returned by all search engines and our system is more than for other users. The query ‘shinning stars’ is most relevant to a non technical person. It is also observed that the system gives more relevance of specific user as compared to the other.

4.2 Results of analysis based on type of useful information returned with varied content:

In the second experiment we have considered the same vague and specific queries and the home-pages returned are analysed for whether information returned is of social network, database/knowledge-base, product related, article/blog, book/publication/software, lifestyle etc. and a browsing history may be recorded. The observations are shown in Table 2.

It is observed that the vague query ‘difficult level’ returns pages which have more useful information about social networks and information is also through articles and blogs. It has less content which is published in books/publications and product based information is negligible.

Table 2 : Results of analysis based on type of useful information returned with varied content

Search Engine	Total results	*User 1 : Comp* Prof	User 2 : Non Tech	User 3 : Novice	User 4 : Sportsman	User 5 : Student
Google	368	46 (12.5%)	138 (37.5%)	45 (12.2%)	57 (15.4%)	*82 (22.2%)*
Yahoo	89	05 (5.61%)	*36 (40.4%)*	04 (4.49%)	22 (24.7%)	22 (24.7%)
Bing	499	22 *(4.4%)*	275 (55.1%)	05 (1.00%)	91 (18.2%)	106 (21.2%)
presented System	182	33 *(18.1%)*	103 (56.5%)	04 (2.19%)	44 (24.1%)	31 (17.0%)

4.3 Results of analysis based on Ranking Tools:

Here the best page returned by three search engines and presented method is fed to Alexa ranker, Google ranker and WooRanker. The results of these rankers are shown in Table 3, Table 4 and Table 5. The weighted average of these ranks for few good pages also serves as an estimate or some statistical analysis like mean is done for the ranks.

It is observed that for the query 'crisp summary' Alexa ranks the best page returned by presented implementation higher at global level and regional level. Google ranker returns a higher ranking for the best page returned for presented implementation whereas WooRanker returns for the best page generated by presented implementation, a rank comparable to Google and Yahoo.

Table 3: Alexa Ranking for best pages returned for the Query

Query: Crisp summary	Google	Yahoo	Bing	Presented Implementation
Global Rank	501	501	817823	8
Popularity at	United States	United States	United States	India
Regional Rank	385	385	137493	1
Backlinks	36,161	36,161	83	28940

Table 4: Google Ranking for best pages returned for the Query

Query: Difficult level	Google	Yahoo	Bing	Presented Implementation
Page Rank for best page	9	9	9	7

Table 5: Woo Ranking for best pages returned for the query

Query: Shinning Stars	Google	Yahoo	Bing	Presented Implementation
Page Rank for best page	77.4	77.4	34.8	77.4

4.4 Results of analysis based on Trustworthiness of a returned Page:

WOT Returns almost all the first pages returned by presented system as dark green or green donuts for all queries indicating Excellent or good and the pages are trustworthy and secure. Further below are found some pages that are tagged by yellow donuts indicating pages are unsatisfactory. As we go still further we encounter light red and dark red donuts indicating pages are poor and very poor quality. In some cases no rating is given to a page which is not in the database of WOT. Thus WOT returns a page reputation.

4.5 Results of analysis based on building queries by a intelligent combination of Grammar based words:

We have taken queries with adjective-noun, adverb-noun, verb-noun, etc. combination and try to identify some words and synonyms for which this will be really useful to detect intent drifting. When considering building queries by a intelligent combination of grammar based words it is observed that for the same thesaurus-based based synonym replacement for an original query like 'well drainage' and a new query with synonyms used 'wells drain' there is no intent drift but for an original query 'cell phone' and a new query with synonyms used 'cell earpiece' there is an intent drift.

5. CONCLUSION

The results achieved after implementing the synonym based search system adopting the directory search mechanism has saved time in visiting previously visited URLs for their information. The adaptive mechanism gives results as desired by the user. The results contain more useful information with varied content.

Some conclusions are drawn from the experiments conducted on the 'best' pages found in the presented system.

- With respect to relevance to a user, the statistical comparison between users to which the pages may be relevant shows that the returned pages are more relevant to a particular user since the percentage average number of pages returned by each search engine for a particular user calculated is greater.
- With respect to type of useful information returned with varied content to a user it is observed that this content is more valid to a user for whom the percentage average number of pages returned by each search engine for each category separated by its processed content calculated is larger.
- Analysis based on Ranking Tools reveals that a page which has the weighted sum (rank and weight given to ranking tool) calculated for giving an estimate of relevance of the page is high and the page is more relevant.
- With respect to Trustworthiness of a returned Page a page is given to a tool like WOT. The software computes the measure of trust the rating users have in websites, combined with data from, among others and tells about how trustworthy the page is.
- With respect to Intent drift testing through building queries by a intelligent combination of Grammar based words it is observed that word combinations that will lead to an intent drift and those combinations that return the same intent even if their words are replaced by other grammatical categories are distinctly identified.

As a future scope an implementation generating clusters of best pages could be devised. More efficient classifiers could be added in the experimental set-up which would take care of new pages generated dynamically daily since, enormous pages are added to the web on regular basis.

6. ACKNOWLEDGMENT

The authors would like to thank their affiliated institutions for all the support given to them during their research and preparing of this publication.

7. REFERENCES

[1] S. Brin and L. Page, "The anatomy of a large-scale hyper-textual web search engine," in Computer Networks and ISDN Systems. Elsevier Science Publishers B. V., 1998, pp. 107–117.

[2] J. Hou and Y. Zhang, "Effectively finding relevant web pages from linkage information," in IEEE Transactions

On Knowledge And Data Engineering, Vol. 15, No. 4, 2003.

[3] M. J. Martin-Bautista and M.-A. Vila, "A survey of genetic feature selection in mining issues," in Proceedings of the 1999 Congress on Evolutionary Computation, CEC 99, Vol 3, IEEE, 1999, pp. 1314–1321.

[4] M. J. Martin-Bautista, M.-A. Vila, and H. L. Larsen, "Building adaptive user profiles by a genetic fuzzy classifier with feature selection." IEEE, 2000, pp. 308–312.

[5] H. Chela, Y.-M. Chung, M. Lamsey, C. C. Yang, P.-C. Ma, and J. Yen, "Intelligent spider for internet searching," in Thirtieth IEEE Hawaii International Conference on System Sciences, 1997.

[6] N. Tomca, "A flexible tool for jaccard score evaluation," in B. Sc. Thesis, University of Belgrade, Belgrade, Serbia, Yugoslavia, 1997.

[7] Mahbub, "Genetic algorithm in adaptive web search," in Filed under Research, 2007.

[8] M.Richardson, A. Prakash, and E. Brill, "Beyond page rank: Machine learning for static ranking," in Proceedings of the International Conference on World Wide Web, 2006.

[9] M. Tsukada, T. Washio, and H. Motoda, "Automatic web-page classification by using machine learning methods," in Web Intelligence: Research and Development, LNAI. Springer-Verlag, 2001, pp. 303–313.

[10] F. Ciravegna, A. Lavelli, D. Petrelli, and F. Pianesi, "The geppetto development environment - version 2.1 - user manual," Tech. Rep., 1997.

[11] M. B. Jadhav and B. M. Patil, "File annotation and sharing on low end devices in pan," in IJCA Journal Volume 83 - Number 13, 2013.

[12] J. B. Filho and J. Gensel, "A contextual annotation-based access control model for pervasive environments," in Second International Workshop on Security and Privacy in Spontaneous Interaction and Mobile Phone Use (IWSSI/SPMU), 2010.

[13] C. A. N. Soules and G. R. Ganger, "A contextual annotation-based access control model for pervasive environment : why can't i find my files? New methods for automating attribute assignments," in Proceedings of the 9th conference on Hot Topics in Operating Systems HOTOS'03 - Volume 9 Pages 20-20, 2003.

[14] M. J. Carman, M. Baillie, and F. Crestani, "Tag data and personalized information retrieval," in In Proceedings of the CIKM workshop on Search in social media. ACM, 2008, pp. 27–34.

[15] R. Jschke, R. Marinho, A. Hotho, L. Schmidt-thieme, and G. Stumme, "Tag recommendations in folks anomies," in PKDD, LNAI 4702,Springer, 2007, pp. 506–514.

[16] G. Bharathi and D. Venkatesan, "Improving information retrieval using document clusters and semantic synonym extraction," in Journal of Theoretical and Applied Information Technology February 2012. Vol. 36 No.2 ISSN: 1992-8645), 2012.

[17] P. Choudhary, "A comparative analysis of various web search engines," in International Journal of Computing and Business Research (IJCBR) ISSN (Online): 2229-6166 Vol. 3 Issue 2, 2012.

[18] G. Madhu, "Intelligent semantic web search engines: A brief survey," in International journal of Web & Semantic Technology (IJWesT) Vol.2, No.1, 2011.

[19] R. K. Roul and S. K. Sahay, "An effective information retrieval for ambiguous query," in arXiv : 1204.1406v1 [cs.IR], 2012.

[20] J. Beel and E. Wilde, "Academic search engine optimization (aseo): Optimizing scholarly literature for google scholar & co." in Journal of Scholarly Publishing, 41 (2):, 2010, pp. 176–190.

[21] H. Ishkewy and H. Harb, "Iswse: Islamic semantic web search engine," in International Journal of Computer Applications (0975 8887) Volume 112 No. 5, 2015.

[22] J. Beel and B. Gipp, "Google scholars ranking algorithm: An introductory overview," in 12th International Conference on Sciento-metrics and Informatics (ISSI09), volume 1,Rio de Janeiro (Brazil), 2009, pp. 230–241.

[23] X. Wei, F. Peng, H. Tseng, Y. Lu, X. Wang, and B. Dumoulin, "Search with synonyms: Problems and solutions," in coling 2010: Poster Volume, Beijing, 2010, pp. 1318–1326.

[24] A. Hliaoutakis, G. Varelas, E. Voutsakis, E. G. M. Petrakis, and E. Milios, "Information retrieval by semantic similarity," in Intern. Journal on Semantic Web and Information Systems (IJSWIS), 3(3):5573, July/Sept. 2006. Special Issue of Multimedia Semantics, 2006.

[25] Y. Li, "Search with synonyms: Problems and solutions," in IEEE Internet Computing 1089-7801/98, 1998.

[26] G. P.Sudhakar and R. Kumar, "Content based ranking for search engines," in International Multi-Conference of Engineers and Computer Scientists Vol I, 2012.

[27] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma, "Query expansion by mining user logs," in IEEE TRANSACTION ON KNOWLEDGE AND DATA ENGINEERING, vol. 15, no. 4, pp. 829–839, 2003.

[28] K. Chakrabarti, M. Ortega, K. Porkaew, and S. Mehrotra, "Query refinement in similarity retrieval systems," in Bulletin of the Technical Committee on Data Engineering Vol. 24 No. 3 IEEE Computer Society, 2001.

[29] X. Tang, K. Liu, J. Cui, S. Member, F. Wen, and X. Wang, "Intent search: capturing user intention for one-click internet image search," in IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 34, No. 7, 2012.

[30] L. Hasan and E. Abuelrub, "Assessing the quality of web sites," in INFOCOMP Vol 7 DOI: 10.1016/j.aci.2009.03.001, 2008.

[31] L. M. R. J Lobo and R. S Bichkar, "Finding the best page using synonyms," in International Journal of Computer Applications, Volume 65 No.8, 2013.