Finding the Best Page using Synonyms

Lobo L.M.R.J
Research Scholar, Department of C.S.E
SGGS Institute of Engg & Techology,
Nanded

R.S.Bichkar, PhD.
Professor, Department of Eln & TC
G.H.Raisoni Institute of Engg & Management
Pune

ABSTRACT
Rating a page to be a best one, based only on Page Ranking algorithm of Brin and Page would be insufficient. This method relied totally on Link information alone. However, due to application of Soft Computing in Data Mining and Knowledge Discovery, machines were made more effective, additional features of a Page involving its indexing, terms used, capitalizations, anchor texts, hit information, etc. were considered. The classification problem helped to induce this to a great extent. The complexity of dealing with a large number of web pages on the net made researchers to think of solutions dealing with sampling pages randomly and then making an analysis of the features of these pages. Soft Computing techniques were used for analysis of the features of the page. These techniques involved Genetic Algorithms, Neural Networks, Fuzzy Logic and Rough sets. User' profiles of pages were created from the retrieved ones. Good and bad Pages were categorised on the basis of the terms they contained and these profiles were preserved for further reference. Pages were compared with each other for their similarity using Jaccard score and Best First search algorithm with developed software agents. Adaptive methods were used. Such methods were close to the concept of Genetic algorithm applications. The frequency at which a user visited web pages was also considered as a parameter of interest. Techniques to generate features of pages using co-occurance analysis were developed and web pages were classified based on machine learning. A good method of rating a page provided benefits like relevance, efficiency and indirectly on a crawl priority of a search engine which was more preferred. The web content designed as on date is for human reading and not typically tractable for machines. The semantic web had to provide structured content by adding annotations. Tools were made available to do these conversions. User-generated metadata that expresses a user taste and interest was used to personalize information to an individual user. Specifically, a machine learning method that analyzed a corpus of tagged content was to be used to find hidden topics. It then used these learned topics to select content that matched a users' interest, thus returning best relevant information pages.
Even though Google scholar does not use synonyms and is strict to article text for searching a document, the use of synonyms reduce irrelevant search, causes intent drifting but synonym discovery is context sensitive these features motivate the use of synonyms to expediate the search and to rank relevant documents at a higher position. Google and Wordnet use synonyms but no documentation mentions using combination of synonyms for a term to generate a better relevant search,

The present paper will concentrate on presenting a developed search technique to find a best page based on synonyms. The technique is based on the concept of adaptive search using synonyms of a search keyword extracted from a dictionary. These synonyms are then combined in different sets and given to a search engine which will return most relevant documents required by the user at a higher ranking.

General Terms
Data Mining, Genetic Algorithms, Search, Soft computing.

Keywords
Best Page, relevance, users' interest, synonyms, metadata

1. INTRODUCTION
Surfing and browsing the net has almost become an inevitable task in today's world. A great demand is now concentrated on the fact of having relevant information required by the user extracted within fractions of seconds. 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] and giving a rating to these Pages in the form of a Page 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 [2]. This feature analysis considers information about each page in the form of an index, the terms contained in a page and the number of times they are repeated, the occurrence of words in a page including its position, font and type specific information along with the page Hits. This feature selection must be done before classification of a page. The results of the feature selection process resulting from Machine Learning and Knowledge Discovery is applied to the Data Mining and Knowledge Discovery. Here irrelevant features are identified and separated.
The web has a large number of pages involved. Classification and categorization tasks now become important. These tasks result in application of traditional mine tasks to information access and retrieval in relation to the web users needs. Soft Computing techniques now come into the picture. Using Genetic Algorithms, Neural computing, Fuzzy Logic and Rough sets along with collaborative filtering would tend to give sufficiently good results. 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 [3].
Once such pages of good quality are sampled out, we require a method to compare between similar pages. This may involve using indexing and agent search techniques [4]. Jacquards' scores could be used to compare similarity between two pages [5]. A Best First search algorithm

could then be used to identify the best pages. The results achieved are better than index search engines, in terms of local search, reading time of search and quality of documents found.

Genetic algorithms have a very close resemblance to an adaptive web search [6]. An adaptive web search is a smart technique to find what you want using a web search engine. After making an initial guess it is needed to adapt the keywords given an input to the search engine till the user gets the relevant information page termed as the best page. Genetic Algorithm is an evolutionary process based on Darvin's theory of existence and survival of the fittest concept. It is a simulation technique that uses a formal approach to simulate solutions which are approximate for a problem. The steps used for getting a best page are shown in Figure 1.

- Start with initial guesses which are random.
- Search for these keywords.
- Select acceptable page from search results and mark keywords from it.
- Do until results are approximately relevant.
- Stop if search takes too long or best page is achieved.

Figure 1: Steps for getting best page

Since this technique depends on initial guess, which may sometimes produce unrelated keywords, the searched results may be bad. It also depends on judgment of accepting the result. Very early generated results and no final results would both be frustrating. If all results are accepted too many searches with all unrelated results would have to be analyzed.

A user's visit to a page was an important parameter to be considered [7]. This was considered as a measure of popularity of a page. This parameter along with other static features based on anchor text and domain characteristics resulted in getting best pages. Arranging best pages in a hierarchical sequence and categorizing them became popular. Using Machine Learning methods to generate features by using a co-occurrence matrix analysis [8] 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 we may 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 [9]. Annotation is the concept to associate semantics with a file [10]. Technology has rapidly advanced and come with the huge storage capacity and many advanced features. Usually when data is in image form it is captured by a camera, it gets stored with the filename as image 001, image 002, etc. To perform search with such filename and to retrieve those image files is tedious. To improve the searching technique semantic file annotation is implemented [11] which annotates the image and retrieves the required file. This technique extracts the basic image attributes from the underlying file system of the device, and uses attribute information as the annotation tags for the corresponding file and parse it using k-XML to store in XML structure.The images are annotated automatically on a device. In addition to the basic attributes, additional keywords can also be added to annotate any image. The XML document is then searched for the required field to retrieve any image in search. The framework provides a variety of options to search for a required file on the device itself or even on the other connected devices, if authorized. The stored meta-data of files in an Extensible Mark-up Language (XML) format, 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 [12]. There are many social media sites developed nowadays, such as Flickr and del.icio.us. These sites allow a user to upload content and annotate it with descriptive labels known as tags, join special-interest groups, etc. This builds a large-scale user generated meta-data that not only facilitate users in sharing and organizing multimedia content, but provide useful information to improve media retrieval and management. 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. [13], [14].

The use of Synonyms reduces irrelevant search. It causes intent drifting. 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. New concepts are extracted with the use of synonyms. Google uses context of synonyms to compare two strings. The similarity is returned as a median score between these searches and uses a synonym operator. A searching scheme with specific keywords may evaluate to be unsatisfactory until exact keywords are entered. Searching and recommendation of multimedia information is mainly based on synonyms. Synonyms help to prepare good indexing logs and search reports. Synonyms are used for schema matching. All these characteristics of synonyms motivate us to devise a method using synonym combinations in order to generate relevant results at a higher ranked position.

The method we have developed for searching the best page is based on the synonyms of the words used in the Query given as input to the system. The basic procedure we have adopted is initially to take a query from the user. The query is then split into its token (strings). We then pass the tokens one by one to a dictionary. Tables of all the synonyms of tokens are then created and these elements are passed to the search engine parser. The search result is then generated. We then cross elements in a table to form new elements choosing a random crossover points and exchange the result query. Change specific table elements if they would result in a better query. 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 we are implementing onto the search engine are

1. A directory based search mechanism of previous searches.
2. An Adaptive mechanism of machine learning by users inputs.
3. A Jaccard's score of ranking web pages

In directory based search mechanism, we have kept the database of the previous searches and show a directory to

the user it matches the string of previous searches. This will enable the user to perfectly determine the string he wants to evaluate or search from a list of previously searched terms. The adaptive mechanism of machine learning by users inputs will allow the ranker to rank various sites on basis of ranks allocated by the user and the quality of content, thereby improving the quality of searches every time.

2. RELATED WORK

G .Bharathi and D.Venkatesan [15] remarked that a user inputs a simple keyword query to a search engine but 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 were used so that from the retrieved documents of the dataset the correlated semantic terms of the specified query term were identified and finally more similar documents were ranked based on semantic correlation similarity. This improves the accuracy of the retrieved relevant documents without much increasing time.

Pooja Choudhary [16] 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. .

G.Madhu, Dr.A.Govardhan and Dr.T.V.Rajinikanth [17] have used typically domain specific knowledge. They commented that users don't include all potential Synonyms and variations in the query, actually user have a problem and aren't sure how to phrase a query to be fired to a search engine to get desired results.

Ahmed Sameh and Amar Kadray[18], brings about a method of finding the synonyms of frequent words in the WordNet database,and adding the synonyms to the pool of frequent terms that comprise the cluster label candidates. The detection of synonyms helps in grouping together snippets that contain different but synonymous words that would otherwise have not been grouped together using the original Lingo algorithm.

Joeran Beel, Bela Gipp, and Erik Wilde [19] boldly remark 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. It is also known that neither synonyms nor PDF metadata are considered.

Nandkishor Vasnik, Shriya Sahu and Devshri Roy[20] 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 Hypernym or Synonym is not effective, so it was better that a combination of semantic relations to be used.

Jöran Beel and Bela Gipp [21] concluded that in all analyzed 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's 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.

Xing Wei, Fuchun Peng, Huishin Tseng, Yumao Lu, Xuerui Wang and Benoit Dumoulin [22] have verified that Search with synonyms was a challenging problem for Web search, as it can easily cause intent drifting.

Angelos Hliaoutakis, Giannis Varelas, Epimenidis Voutsakis, Euripides G. M. Petrakis and Evangelos Milios [23] have shown that Term similarity was computed by matching synonyms, term neighborhoods, and term features.

Yanhong Li [24] have remarked that one could discover synonyms, extract new concepts, and build a thesaurus.

P.Sudhakar, G.Poonkuzhali and R.Kishore Kumar [25] indicated their observation that every root words are 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. If a match is found then a particular weight is awarded to each word. Finally, the total relevancy of the particular link against user request is computed by summarizing all the weights of the keyword and content words. The page which contains total relevancy value nearest to 1 are ranked as first page and 0 are ranked as last page.

Hang Cui, Ji-Rong Wen, Jian-Yun Nie, and Wei-Ying Ma [26] 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 users' original queries.

Kaushik Chakrabarti, Michael Ortega, Kriengkrai Porkaew and Sharad Mehrotra[27] in " Query Refinement in Similarity Retrieval Systems" show how the EasyAsk system supports a wide variety of features such as approximate word matching, word stemming, synonyms and other word associations .It also recognized phrases, and supported comparisons which it translated into appropriate SQL conditions. Padding the query with synonyms model was easily extendible to allow scaling down of node weights to account for approximate match or synonyms. They also stated that Synonyms were particularly useful in the context of matching metadata.

Xiaoou Tang, Ke Liu, Jingyu Cui, Fang Wen and Xiaogang Wang [28] 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.

3. METHODOLOGY

The method developed was based on getting more relevant documents based on word synonyms.

Objective - Given an input query consisting of terms. We wished to select the best page (having highest Ranking according to the document containing large number. of terms from the query).

Algorithm –

1. Break the user Query into a number of meaningful terms
2. Send the terms to the dictionary to fetch the Synonyms. The dictionary used here was Merriam-Webster, originally known as the G. & C. An American Dictionary of the English Language (1828).
3. The synonyms were listed in a tabular form as below each element in this table is called a term.

Synonym 1	Synonym2	…	Synonym m

The words in the query determined the total number of terms to be considered.

4. Crossover of term elements in the table is carried out to obtain the best combination. This is done at random position as per the terms in the table.

Example for Query – Stress Testing

Synonyms for stress are Pressure, strain and tension which are inserted in the table.

Pressure	Strain	Tension

There is no synonym for testing, so testing is inserted into the second table.

Testing

Now a combination is carried out between terms of the two tables. This is shown below.

Tension	Testing

Strain	Testing

Pressure	Testing

5. Now we insert those terms not selected in crossover. It is optional.
6. These new generated combination terms are sent to the parser of a search engine (SE).
7. Best Pages are generated.

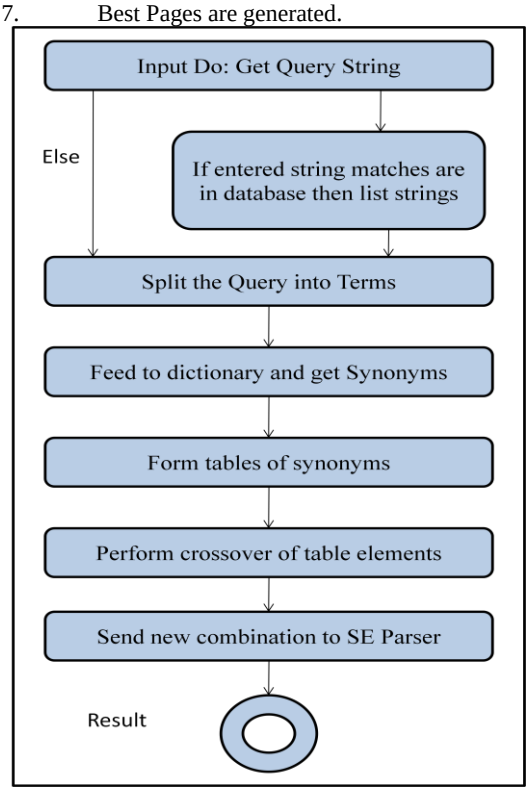

Figure 2: Method used to extract best page

The representation of the method used is as shown in Figure 2.

4. EXPERIMENTAL SETUP AND RESULTS

The system is developed in JAVA. The Marian Webster dictionary is used to generate the synonyms of the words. After the separation of words in the query and finding the synonyms of words and placing them in tables, combination of words are generated and fed to the Parser of the search engine.

The output for the Query "Good result" is as shown in the Figure 3.

A comparative study with the existing systems project drawbacks listed below

- Low precision due to irrelevant pages being returned and low recall due to in availability to index all information available on the web.
- Using similarity measures which were usually unpredictable unless all synonyms and variations.
- Building unnecessary synonym pools consuming a large amount of memory.

This shows that more relevant pages are returned when we use a crossover combination of synonyms of terms in the Query. No similarity measure is directly calculated but all returned pages are ranked to have higher rank to more relevant content pages. We have to deal with one crossover Query at an instant; hence less amount of memory is required at an instant of time. Some of the comparisons between existing systems and the developed system are as show in Table 1.

Parameter	Existing Systems	System using combination of crossover of synonyms of Query terms
Synonym pool	Yes	No
Intent drifting fear	more	less
Relevance of pages returned	less	more
Content	Less concentrated	More concentrated
Speed of search	less	more
Use of memory	more	less
Crossing of synchronyms	yes	more
No of pages returned	more	less

Table 1: Comparison with existing systems.

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 four well known search engine and our system that uses crossover of the synonyms of the terms used in the query given by the user it is seen that the number of results (pages) returned by the latter system developed by us is considerably few as shown in Table 2. This is due to removal of redundencies in pages previously visited.

It is clear from the table that large number of pages is returned by the four search engines but only relevant

results are retained for the system we have developed. High ranking is given to more content pages. These pages contained relevant information which is required by a user. This reduces the search space and time required to retrieve relevant information required by a user with more content. These pages are then ranked at top positions.

Query ↓	Search Engines				System using combination of crossover of synonyms of Query terms
	Google	Yahoo	AltaVista	Bing	
Good day	4910 Million	150 Million	807 Million	145 Million	26
Best school	4490 Million	60.6 Million	799 Million	62.3 Million	25
Pressure, Strain, tension	14.3 Million	1.79 Million	0.305 Million	0.598 Million	102
Good result	1670 Million	11.9 Million	180 Million	13 Million	102

Table 2: Pages returned by existing system

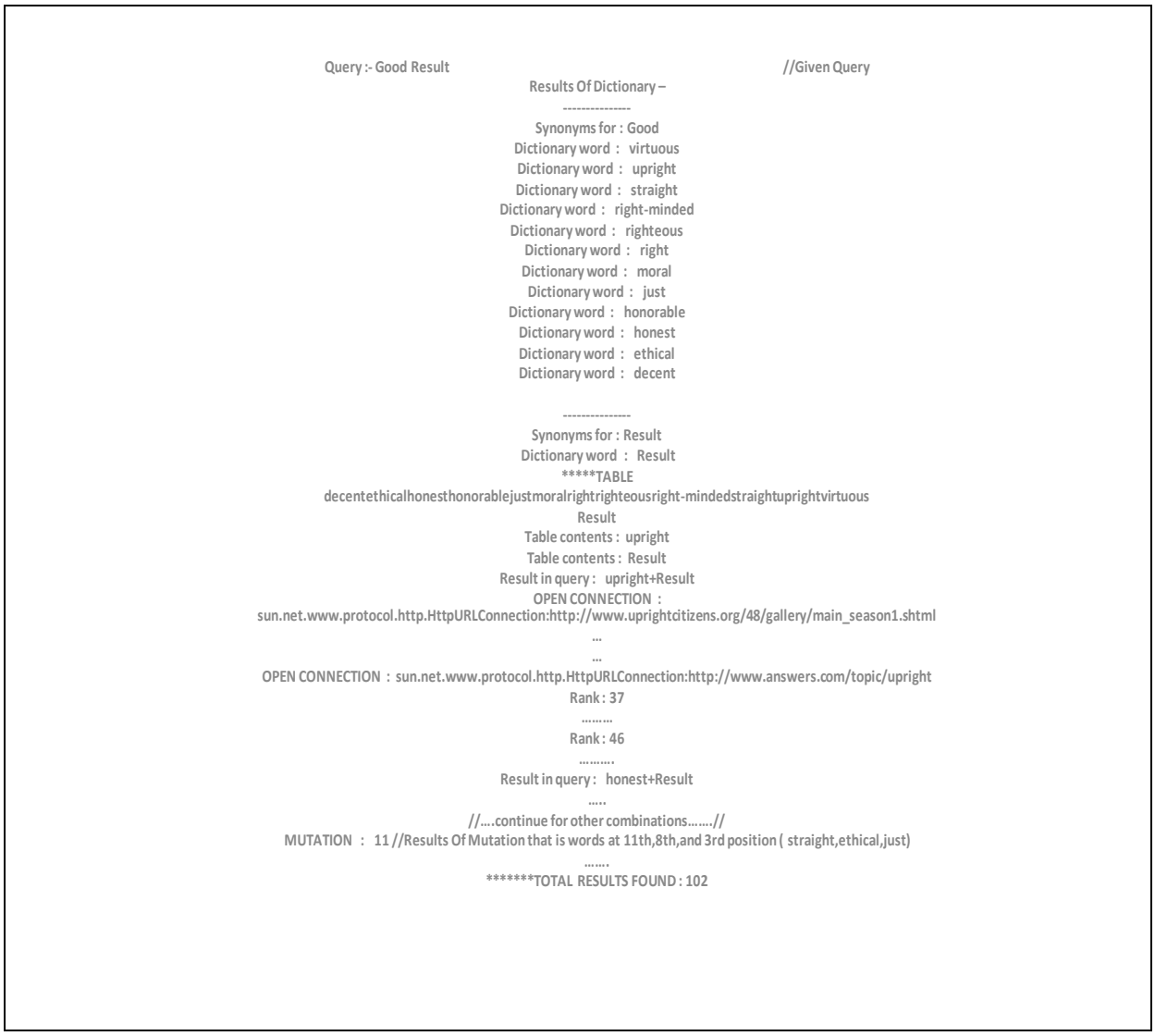

Figure 3: Results of Best pages for query Good Result

5. CONCLUSION

The results achieved after implementing the synonym based search system adopting the directory search mechanism with crossover of synonyms of terms used in the query, has saved time in visiting previously visited URL's for their information. The number of pages returned is enormously reduced. The adaptive mechanism gives results as desired by the user with a high ranking and more content. It searches all the possibilities with various combinations of terms and hence definitely gets the required results the user is searching for.

The future development in this system calls for use of meta-search engines so that larger information database will be available to search for more content and relevance.

6. REFERENCES

[1] Surgey Brin, Lawrence Page, "The Anatomy of a Large-Scale Hypertextual Web search Engine." 2000.

[2] Maria J. MartinBAutista, Maria Amparo Villa, "A Survey of Genetic feature Selection in Mining Issues", IEEE 1999 pp 1314-1321.

[3] Maria J. MartinBAutista, Maria Amparo Villa & Henrik L.Larsen,"Building Adaptive user profiles by a Genetic Fuzzy classifier with feature selection",IEEE 2000 O-7803-5877-5/00 pp 308-312.

[4] Chen, H., Chung, Y., Ramsey, M., Yang, C., Ma, P., Yen, J., "Intelligent Spider for Internet Searching," *Proceedings of the 30th Annual Hawaii International Conference on System Sciences - Volume IV*, Kailua-Kona, Hawaii, USA, January 1997.

[5] Tomca, N., *A Flexible Tool for Jaccard Score Evaluation* , B. Sc. Thesis, University of Belgrade, Belgrade, Serbia, Yugoslavia, December 1997.

[6] Mahbub, "Genetic Algorithm in Adaptive Web Search" Filed under Research, April 2007

[7] Mathew Richardson, Amit Prakash, Eric Brill, "Beyond Page Rank: Machine Learning for Static Ranking", WWW May 2006, ACM I-59593-323-9/06/0005.

[8] Mokoto Tsukada, Takashi Washio, Hiroshi Motoda, "Web-Page Classification by Using Machine Learning Methods", 1998.

[9] Fabio Ciravegna, Daniela Petrelli,"User involvement in adaptive information extraction", in Proceedings of IJCAI2001.

[10] Sadaqat Jan, Maozhen Li, Ghaidaa Al-Sultany and Hamed Al-Raweshidy "File Annotation and Sharing on Low-End Mobile Devices", Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2010.

[11] Windson Viana1, José Bringel Filho2, Jérôme Gensel, Marlène Villanova-Oliver, Hervé Martin, "A Semantic Approach and a Web Tool for Contextual Annotation of Photos Using Camera Phones",9th Workshop on Hot Topics in Operating Systems (HotOS IX).May 18-21, 2003.

[12] Soules CAN, Ganger GR "Why can't I find my files? New methods for automating attribute assignment", In: Proceedings of HotOS IX: the 9th workshop on hot topics in operating systems. USENIX Association, May 2003.

[13] M. J. Carman, M. Baillie, and F. Crestani, "Tag data and personalized information retrieval," in *SSM*, 2008, pp. 27–34.

[14] R. J¨aschke, L. B. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme, "Tag recommendations in folksonomies," in *PKDD*, 2007, pp. 506–514.

[15] 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

[16] Pooja Choudhary. "A Comparative Analysis of Various Web Search Engines". In International Journal of Computing and Business Research (IJCBR) ISSN (Online): 2229-6166 Volume 3 Issue 2 May 2012.

[17] G.Madhu, Dr.A.Govardhan and Dr.T.V.Rajinikanth, "Intelligent Semantic Web Search Engines: A Brief Survey", in International journal of Web & Semantic Technology (IJWesT) Vol.2, No.1, January 2011 DOI:10.5121/ijwest.2011.2103 34

[18] Ahmed Sameh and Amar Kadray, "Semantic Web Search Results Clustering Using Lingo and WordNet",in International Journal of Research and Reviews in Computer Science (IJRRCS) Vol. 1, No. 2, June 2010

[19] Joeran Beel, Bela Gipp, and Erik Wilde. " Academic Search Engine Optimization (ASEO): Optimizing Scholarly Literature for Google Scholar and Co.", in Journal of Scholarly Publishing, 41 (2): 176–190, January 2010

[20] Nandkishor Vasnik, Shriya Sahu and Devshri Roy, "Talash: A Semantic and Context Based Optimized Hindi Search Engine", in International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 201210.

[21] Jöran Beel and Bela Gipp. "Google Scholar's Ranking Algorithm: An Introductory Overview. In Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09), volume 1, pages 230–241, July 2009.

[22] Xing Wei, Fuchun Peng, Huishin Tseng, Yumao Lu, Xuerui Wang and Benoit Dumoulin," Search with Synonyms: Problems and Solutions", in coling 2010: Poster Volume, pages 1318–1326,Beijing, August 2010

[23] Angelos Hliaoutakis, Giannis Varelas, Epimenidis Voutsakis, Euripides G. M. Petrakis and Evangelos Milios,"Information Retrieval by Semantic Similarity", in. International Journal on Semantic Web & Information Systems, 2(3), 55-73, July-September 2006

[24] Yanhong Li, "Toward a Qualitative Search Engine". 1089-7801/98, IEEE INTERNET COMPUTING JULY • AUGUST 1998

[25] P.Sudhakar, G.Poonkuzhali and R.Kishore Kumar, "Content Based Ranking for Search Engines", in

proceedings of International MoltiConference of Engineers and Computer Scientists 2012 Vol I. IMECS Mar 2012

[26] Hang Cui, Ji-Rong Wen, Jian-Yun Nie, and Wei-Ying Ma, Member, "Query Expansion by Mining User Logs",In IEEE Transactions On Knowledge And Data Engineering, Vol. 15, No. 4, July/August 2003

[27] Kaushik Chakrabarti, Michael Ortega, Kriengkrai Porkaew and Sharad Mehrotra. " Query Refinement in Similarity Retrieval Systems" in Bulletin of the Technical Committee on Data Engineering Vol. 24 No. 3 IEEE Computer Society September 2001

[28] Xiaoou Tang, Ke Liu, Jingyu Cui,Fang Wen and Xiaogang Wang, "IntentSearch: Capturing User Intention for One-Click Internet Image Search" in IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 34, No. 7, July 2012