Google Search Filter Using Cosine Similarity Measure to Find All Relevant Documents of a Specific Research Topic

Mohammed Abdullah Hassan Al-Hagery[1]

***Abstract*—** A large number of scientific papers are retrieved using Search Engines from the electronic databases. Some of these Engines are limited and others have designed for a general purpose. A number of researchers wish to prepare a survey of a particular topic. They are facing a problem to find the most related topics to a particular research title. The other problem is rising as a result of a search in an electronic database, where some Search Engines displays Dozens of pages and hundreds of results, it needs also more effort to be scanned manually and decide which results are relevant and which should be excluded. During the search process and matching the contents, the Search Engine maybe ignore some important documents. Some of these documents are excluded although, it is relevant to the subject and some results are included but not important. This research concentrates on a development of a Multi Scanning Filter (MSF) algorithm, that works on research documents found in various scientific databases, such as ISI, SCOPUS or EBSCO, etc. The idea of this research depends on the Google Search Engine, where the proposed algorithm consists of three parts. It maximizes the search space and works as a filter to Google results based on the similarity measure. This algorithm reduces the final search result list, make it more accurate, eliminate the problem of results' dispersion in traditional Search Engines, and helps developers improve current Search Engines, such as Google, this in turn will assist researchers everywhere gather the most related topics to a particular research title in a short time.

***Keywords*—** Filtration Algorithm, Scientific Search Engines, Search Accuracy, Search Results Optimization, Search Results Quality, Search Using Synonyms, Similarity Measures.

I. Introduction

Search Engine optimization is a strategical technique to place a web document in top search results of a Search Engine. The main function of Search Engines and search processes is to deliver most related documents to users in minimum time. So granting fast time and effective accesses to the index is a major issue for performances of web Search Engines. The indexing processes are applied to the web pages after they have been collected by the crawler and placed into a data repository [1]. The increasing growth of the resources on the internet, the non-homogeneous subject information, the high volume of these resources, the variety of the users, and the informational needs of them are about the challenges, which make the Information Retrieval (IR) more difficult at any time on the web environment [2].

One of the problems that are facing all those who wish to review the historical background on a particular topic is the large number of documents that are archived in the electronic databases, but, the Search Engines during the search process bring a large number of documents, a high percentage of these documents are irrelevant to the required topic as a result of weakness of the matching process. This research aims to develop an algorithm to help researchers to collect related materials of the historical background of a research under consideration in a specific field. The proposed algorithm will work as a filter on the Google Search results, and then provides a short and an accurate list of results.

II. Literature Review

Search Engines on the Internet websites represent the most important components in the IR, so researchers and developers seeking to create new algorithms and tools to increase the quality of the search process and efficiency of Search Engines. Online Search Engines are undoubtedly the main fulcrum of web user activity and they are used for satisfying a great variety of search purposes. Among all the search intents, a considerable portion of traffic on traditional query-based Search Engines are characterized by complex search tasks that can be satisfied only through the aggregation of information from multiple sources [3].

Golovchinsky and Pickens [4] described a special framework for unifying transitions between various phases of exploratory search and presented how context from one step can be applied to the next. Another research work performed in [5]. They focused on the social search and proposed a model to realize the unexplored potential of folksonomies in satisfying complex search intents. They presented the relational folksonomy, an extension of the classical folksonomy that allows composing complex user-defined relations between objects and tagging them.

A number of approaches have discussed, which are relevant to web search, those that adopt a traditional, document-centric, information retrieval perspective that are limited by their refusal to consider the past search behavior of users during future search sessions [6]. In particular, they

[1] The author is in the Computer Science Department and currently is the head of Research Center in the College of Computer, (Qassim University), (KSA) (e-mail: hajry@qu.edu.sa, dr_alhagery@yahoo.com).

This work was supported by Deanship of Scientific Research, Qassim University, according to the agreement of the funded project No. SRD-014-2718, started in 2013, the author thanks, the sponsor of this work for their support.

argued that in many circumstances the users' search behavior is repetitive and regular. The same sorts of queries tend to recur and the same type of results that often selected. They described how this observation can lead to a novel approach to a more adaptive form of search, one that leverages past search behaviors as a means to re-rank future search results in a way that recognizes the implicit preferences of communities of searchers.

Several efforts have also been spent on improving the tagging systems quality of service by structuring and properly ranking search results in tag-based search[7] or proposing services or functions based on the information extracted from folksonomies, like personalized recommendation services [8], [9] or social link suggestion [10]. A research project accomplished in [11], the overall aim of this project was to provide an authoritative point of view with regards to the user effort required to obtain hits using a web-based Semantic Search Engine. The project seeks to compare Hakia as a Semantic Search Engine with Google. They collected queries from 30 university students and entered these queries into two Search Engines; Google, the most widely used Search Engine and Hakia as an upcoming Semantic Search Engine. Precision was thereafter calculated using a pre-determined formula. Their calculation revealed that Google outperforms. Hakia as it has a higher mean precision at 0.64 as compared to Hakia at 0.54. Google also has a lower standard deviation of 0.14 as compared to Hakia at 0.25.

A research group tried to make a search tool with some characteristics such as cost-effective, efficient, fast and user-friendly. The tool proposed to retrieve the most relevant documents, which have been stored into the database. The main goal was to make a web Search Engine that will retrieve the most matched web pages in the shortest possible time. They proposed an algorithm then designed it based on the basic principles of a tree structure and the crawler indexing method [12].

A description of an original specialized Search Engine introduced in [13], which uses advanced cross-lingual IR technologies to check information quality by synthesizing medical concepts, conclusions and references contained in the health literature, to identify accurate, relevant sources. The conducted results illustrated that the suggested Search Engine is perceived as informative in a high degree. This type is restricted to a specific domain and it is an alternative to general-purpose Search Engines. However, it left room for improvement. Arora and Bhalla [14], proposed a search based on synonyms. They used different synonyms for several keywords to find easily all relevant documents on the web based on these keywords and its synonyms, which organized based on a database (mapping table). This type of Search Engines planned to increase the ranking of a website collected on the Search Engine and to provide users with more accurate and relevant results, nevertheless this approach needs further improvement by implementing synonym table in a more effective way, which should include less space consumption and minimum access time.

Swaraj and Gunasekaran accomplished a comparative study on four different Search Engines, which tries to optimize itself by individual unique algorithms. They also shows the drawbacks and advantages of the previous algorithm and refinement in the latest algorithm. A number of vivid Search Engines are discussed such as Google, Bing, Yahoo, and ASK. They concluded that Google Search Engine is the most optimized and deals with better algorithm to throw out nearing results expected by users [15].

Singh and Gupta in [16] presented a distributed approach for web Crawlers including Data Mining. It works as a distributed system with a central Control Unit fixing or providing the jobs to a different computer, which are connected with a Network. Their approach is not exactly new as the largest Search Engine that can also distribute processing power by using a different number of computer systems linked with each other.

The retrieving process of meaningful information is very difficult. However to overcome this problem in Search Engines to retrieve meaningful information intelligently, semantic web technologies are playing a major role.

A preliminary survey presented over the existing literature regarding intelligent Search Engines and semantic search on the web. They concluded that different Search Engines return different search results due to the variation in indexing and search strategies [17].

Web-based Search Engines assist tenth and hundreds of millions of people to search for significant information, which is relevant to any subject, education, news, history, sport, business, to multimedia and many more. It has been of great importance to all people that have access to the Internet websites and its services. Search Engines have many influences on people in their dissimilar levels of lifestyle and social responsibility [18]. Minnie and Srinivasan analyzed the features of Vertical Search Engines and Meta Search Engines. A Vertical Search process provides the user with the results for queries on that field. Meta Search Engines send the user's search queries to several Search Engines and combine the final results together. They proposed a Meta Search Engine for searching and retrieving electronic documents on several fields in the internet networks and provide an interface to select the type of Search Engine[19]. Lewandowski analyzed the update strategies of the major web Search Engines for Google, MSN/Live.com, and Yahoo. He conducted a test of the updates of 40 daily updated pages and 30 irregularly updated pages. He used a set of data from a time span of six weeks through three years and identified an important problem, which is the delay in making crawled pages available for searching, which differs from one Search Engine to another [20]. Aravindhan and Shanmugalakshmi accomplished a survey about the semantic Search Engines to reveal the promising features of the semantic Search Engines and the object behind the reluctance of the users in adopting these sophisticated Search Engines [21]. A number of Semantic Web Search Engines have developed recently, which are based on different design principles and provides different levels of support for users and/or applications [22]. The IR performance of common Search Engines Investigated

to find Turkish documents, by using five Search Engines and a list of Turkish queries. Each query is run for each type one by one and the first twenty documents on each retrieval output are evaluated as being "relevant" and "non-relevant". The results used to compare all these Search Engines based on the precision, Google appears to be the best Search Engine in terms of average precision = 73% and normalized recall ratios=66%, on finding these type of documents [23]. Mostly, the problem of finding the Search Engine that performs best for a query is how to select the Search Engine carefully.

A research team analyzed problems and precision in the IR, in order to study the retrieval precision of website information and solve the problems existed in Search Engines based on experimental data and puts forward a new construction of a Search Engine. Web environment is important from two points of view; quality and quantity of information hidden and visible [24], [25]. Searching techniques in existing web are focused on discovering the documents via keywords in contrast to software and the semantic relations among the resources are ignored. De Silva introduced an algorithm to classify documents into various classes, and then he evaluated the algorithm through the use case of analyzing a set of reviews from Rotten Tomatoes. The results got with an accuracy of 53.6% [26].

This research focuses on the process of developing the quality of Google Search Engine, on the scientific research domain, using a filtration process. This type of filtration is used for separating the subjects based on the similarity measure and ranks the results as relevant or irrelevant. The irrelevant results/documents will be excluded and the relevant documents will be sorted according to their importance in a short list, instead of multiple lists.

III. Common Scientific Search Engines

The most popular Search Engines are covering wide areas of search and other functions, they provided additional services such as the possibility of creating a personal home page, provide free email service, Service IM Chat, and knowledge of the weather. One of the important search types is the exploratory search, and it is a difficult activity that requires iterative interaction. This iterative process helps the searcher to understand and to refine the information need. It also generates a rich set of data that can be used effectively to reflect on what has been found [4]. Search Engine quality depends on several factors; Index Quality, Quality of search features, Quality of the results, and Search Engine usability [20]. The most common Search Engines used by researchers, teachers, students in the education fields include the following types:

1) *Google Search Engine:* this type of search engines is the most improved, although, it has some drawbacks, it deals with better algorithm to throw out nearing results expected by users, for this reason Google was selected in this research.
2) *Go To*: This site is like an open market, where the researcher can find whatever he wants; information, services and goods easily and accurately.
3) *Yahoo:* This offers free service e-mail and displays for many of the products and goods through the channels.
4) *Alta Vista:* That is one of the largest sources of search, where the search deals with the component index of more than 140 million page views and information, which are upgraded every 28 days. The service also offers an interpretation of the various languages, an image search service, and email.
5) *Excite*: This Search Engine offers many topics that are grouped into one page, such as automotive, finance, family, computer, education, sports, and travel.
6) *Online JOurnal Search Engine (OJOSE):* is a Search Engine directed for Scientific Research and in education [2].
7) *Scirus:* is used for scientific information only, it is the most comprehensive scientific research tool on the web.
8) *Google Scholar (GS):* It is a freely accessible Search Engine on the web that indexes the metadata or full text of scholarly literature through an array of publishing formats. It is one of the most search tools in the academic field. GS is a combination of many methods organized in a single algorithm. There are a number of research works relevant to GS had been done. It includes an overlapping of data, with other Search Engines that have similar functionalities, such as web of Science and Scopus [27], GS is covering the literature in general and in particular research domains [28], [29], the appropriateness to use citation counts of GS for computing various indicators, such as the h-index [30] and the trustworthiness of GS as a significant source of information [31], [32].

IV. Research Problem

During the process of paper analysis, various attributes of a document should be examined, such as document title, abstracts, keywords, other details. Search Engines usually ignore a large number of papers. Some papers are excluded, although they are relevant to the subject, they do not provide relevant data, because they don't contain an empirical study or a historical background. In traditional methods, the researcher is willing to formulate a suitable historical background about a relevant subject. A lot of time is spent, maybe up to several days to weeks, this time is required to gather the most relevant references, and this is the first problem. The other problem becomes clear as a result of the search in the electronic Database, where some Search Engines display more results (maybe hundreds of results), which need additional effort to be filtered manually because a few numbers of these results are relevant and the most are irrelevant. As a result of these problems and to avoid the manual work, it is very important to develop additional filtration algorithms to reduce the search results and to improve the search quality on the web, precisely in the field of scientific research.

V. Similarity-Based Retrieval

A document is represented by a string or series of words, which can be identified by a set of keywords, user queries may use expressions of keywords, for instance, car and repair shop, tea or coffee, DBMS but not Oracle. Queries and retrieval should consider Polysemy and Synonyms, for example, repair and maintenance. The major complications of the model include; Synonymy and Polysemy. Synonym means, a keyword K does not appear anywhere in the document, even though the document is closely related to K, e.g., Data Mining. Polysemy: The same keyword may have different meanings in different contexts, e.g., Rock mining and Data Mining or Gold mining, etc.

To collect similar documents according to a set of shared keywords, results should have a high degree of relevance according to the closeness of the keywords, Keywords Relative Frequency (KRF), etc. There are basic methods for this task, such as stop list, word stem. Stop list is a set of words, which are considered "unrelated", although, they may appear frequently, for instance, of, the a, for, in, on are, with, to, when, where, etc. Word stem is a characteristic of a word, where several words are small syntactic variants of each other since they share a shared word stem, for example, drug, drugged, drugs. These words can be saved in a Frequency Table (FT), where each entry in the FT (i, j) is equal to the number of occurrences of the word wi in document di. Commonly, the ratio instead of the absolute number of occurrences is used. In this case, the similarity metrics used to measure the nearness of a document to a specific query depend on a set of keywords.

$$Sim(Key, Doc) = \frac{x_1.y_1 + x_2.y_2 + x_3.y_3 + \cdots x_n.y_n}{\left|\sqrt{{x_1}^2 + {x_2}^2 + {x_3}^2 + \cdots + {x_p}^2}\right| . \left|\sqrt{{y_1}^2 + {y_2}^2 + {y_3}^2 + \cdots + {y_p}^2}\right|} \quad (3)$$

Theoretically, it is the length of the vector. Similarly, ||y|| is the Euclidean norm of the vector y. The measure calculates the cosine of the angle between the two vectors x and y. The 0 values of cosine mean that the two vectors are at 90 degrees to each other, they are orthogonal and have no match. When, the value of cosine is equal to 1, the degree of similarity in this case is represented as a match between the two vectors, where the angle between them is 0. [26].

Any document can be appeared by hundreds of attributes, each showing the frequency of a specific word (phrase or keyword) in the document. So, each document contents can be represented by a Term Frequency Vector (TFV). Table I shows an example of four documents, various TFV. Doc1 contains six instances of the word "software", whereas the word "system" occurs three times. The word "improve" is absent from the entire document, as indicated by the count value that is equal to 0. Such data can be highly asymmetric. In this table the comparisons were done among each two documents.

Assume that X and Y are the first two TFV in Table I. That is, X = (6, 0, 3, 0, 2, 0, 0, 2, 0) and Y =

(4, 0, 2, 0, 1, 1, 0, 1, 1). How similar are X and Y?

To calculate the similarity of the two vectors, the cosine formula shown in "(1)," is applied and the result as follows:

X.Y=6×4+0×0+3×2+0×0+2×1+0×1+0×0+2×1+0×1=34

$$||X|| = \sqrt{6^2 + 0^2 + 3^2 + 0^2 + 2^2 + 0^2 + 0^2 + 2^2 + 0^2} = 7.28$$

$$||Y|| = \sqrt{4^2 + 0^2 + 2^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 1^2} = 4.9$$

A. Cosine Similarity Measure Application

The Cosine Similarity Measure can be used to compare documents or, say, give a ranking of documents, with respect to a given term of query words. Let X and Y be two terms for comparison. Using the cosine measure as a similarity function, cosine distance is shown in formula "(1)," and (3),". This measure is superior when compared to the other measures such as Jaccard measure, Euclidean, Pearson Correlation distance. The Cosine Similarity measure is particularly better for text documents [33].

$$Sim(X, Y) = \frac{X.Y}{|X|.|Y|} \quad (1)$$

Where, $||X||$ is the Euclidean norm of term x= x_1, x_2, x_3,… x_p, defined according to formula "(2)," as in [26].

$$||X|| = \sqrt{{x_1}^2 + {x_2}^2 + {x_3}^2 + \cdots + {x_p}^2} \quad (2)$$

$$\text{Sim}(X, Y) = \frac{X.Y}{|X|.|Y|} = \frac{34}{7.28 * 4.9} = 0.95$$

Based on formula "(1)," Sim (X, Y) = 0.953. Therefore, if the cosine similarity measure is used to compare these documents, based on the results of the given example, they would be considered quite similar. To apply this formula to find the similarity among a two documents, the first document will represent the first vector Xi, and the second document will represent the vector Yi.

VI. Research Methodology

The proposed filtration algorithm was designed as different components and applied using the Google Search Engine. It consists of three parts; the first part is searching based on a set of keywords. The second part is searching based on a set of Synonyms. The third part is using the results of the first and the second parts and applies the similarity measure, runs a filtration process, and generates the final results. This idea will give potential and features much better than those features of current Search Engines, in particular in Google.

A. Research Data

Three search examples were applied by Google Search Engine based on sets of Keywords and Synonyms. This process generates three samples of document files, which gathered from the electronic databases. The three samples that were applied in the filtration step (the third scan), these files include research papers retrieved as a PDF format. Some of these research papers were indexed in famous scientific

databases, such as in Scopus-Elsevier. Each sample contains 34 research files and the total number is 102 research papers in all samples.

Table I: Document TFV example

Docs	Keywords									‖X‖	‖Y‖	X×Y	Sim (X, Y)
	Software	Improve	System	Mining	Cell	Degree	Rank	Series	Maintain				
Doc1	6	0	3	0	2	0	0	2	0	7.28	4.9	34	0.95
Doc 2	4	0	2	0	1	1	0	1	1				
Doc1	6	0	3	0	2	0	0	2	0	7.28	9.85	19	0.26
Doc 3	1	4	1	3	1	2	5	4	0				
Doc1	6	0	3	0	2	0	0	2	0	7.28	6.32	2	0.04
Doc 4	0	2	0	5	1	1	0	0	3				
Doc 2	4	0	2	0	1	1	0	1	1	4.9	8.54	13	0.31
Doc 3	1	4	1	3	1	2	5	4	0				
Doc 3	1	4	1	3	1	2	5	4	0	9.85	6.32	26	0.42
Doc 4	0	2	0	5	1	1	0	0	3				

B. MSF *Description*

In general, search by Keyword enables researchers to accomplish explorations of more contents, such as flat files, text data fields. Commonly, there are two types of search mechanisms used in this research. Firstly, standard mechanism depends on a set of keywords, which uses direct queries on the contents to acquire the relevant results. The drawback is that queries are relatively slow, especially in the very large databases. On the other hand, there is no special format required by the system manager to search. Secondly, depends on a set of Synonyms to increase the search domain. Fig. 1 displays the general layouts of the proposed Search Engine processes. The research methodology describes all stages of the MSF.

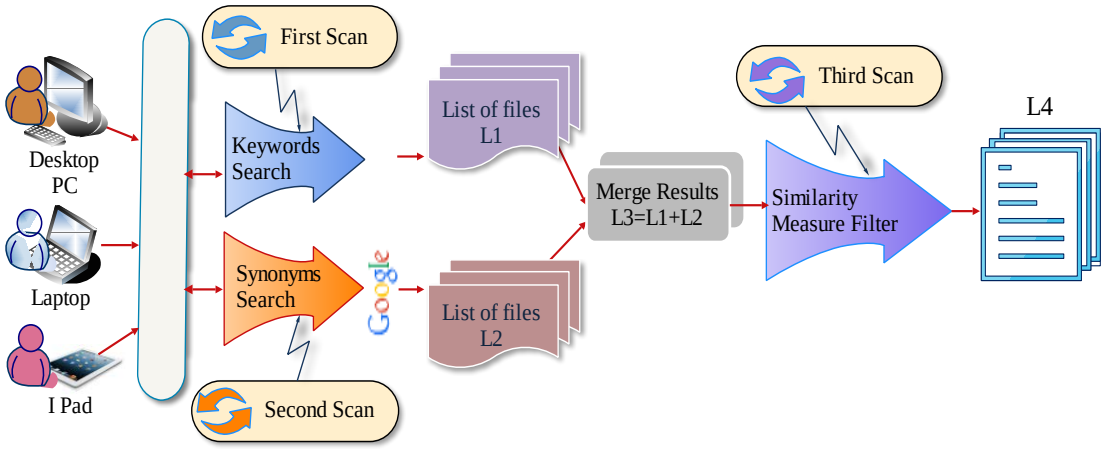

Fig. 1: MSF Processes Chart

1) The First Scan

The following points display main steps of the first scan's algorithm:

1. Start
2. Input user query (Keyword list) to the MSF interface.
3. Pass the query and process it using Google Search Engine.
4. Search and match user keywords with the contents of the Internet databases (DB1, DB2, DB3, DBn)
5. Collect all results = {doc1, doc2, doc3,…, docn}, from several pages P1, P2,…, Pn.
6. Create a list of the search results L1= {doc1, doc2, doc3,…, docn}, as illustrated in Fig. 2.
7. End

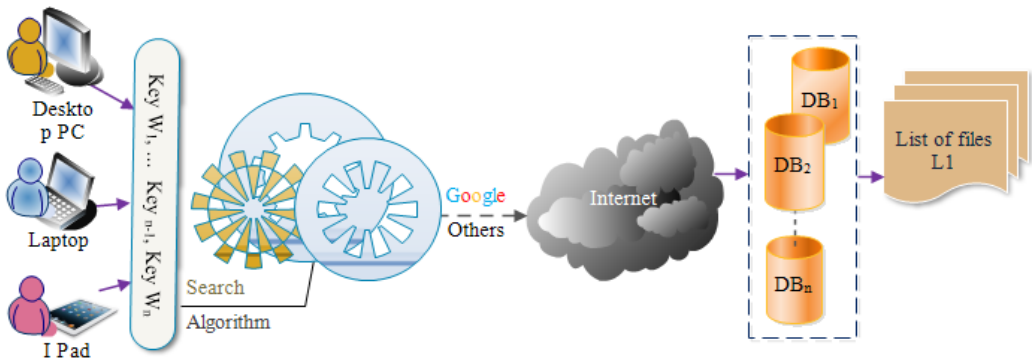

Fig. 2: First Filtration Scan

2) The Second Scan

The following algorithm describes the second scan steps:

1. Start
2. Input user query (Synonym list) to the MSF interface.
3. Repeat the steps of the first scan algorithm from 3 to 5.
4. Create an empty list (L2)
5. If there is a match add the file to L2, where L2 = L2 + (Fi), the details are shown in Fig. 3.
6. End

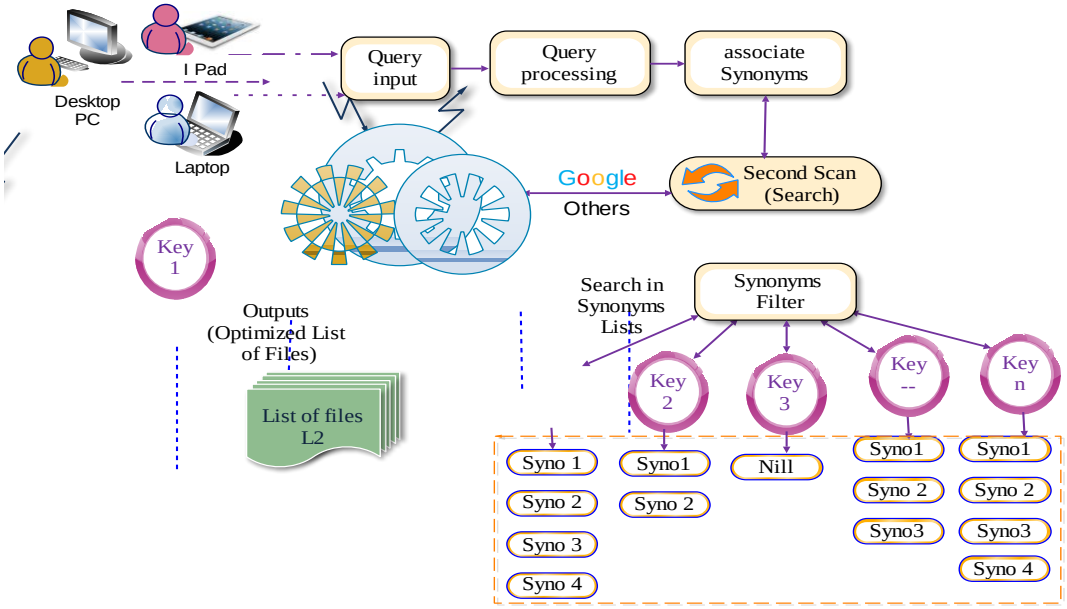

Fig. 3: Second Filtration Scan

3) The Third Scan

The objective of this type of scan is to identify an optimized list of relevant documents. The details are shown in Fig. 4. The following steps display the algorithm of this scan.

1. Create a Similarity Measure Table.
2. Use the keywords and their synonyms to fill in the table.
3. Create new list (L3), where L3=Merge of (L1+L2).
4. For each document file (Fi) in L3, search entire the contents.
5. Match, and count all recurrences of keywords and Synonyms.
6. Fill in the Similarity table.
7. Calculate the values of similarity using the formula "(1),".
8. Apply the similarity measure Sim (keyword, Doci) and Sim (Synonyms, Doci) shown in formula "(3),".
9. Fill in the Similarity table.
10. Enter the maximum length of references list (N).
11. Split the results of the Similarity table, which have similarity values rate > 0 in a new list L4.
12. Sort and display the contents of L4 in descending order, as an optimized list, where ($0 \leq$ L4 Length $\leq$ N).

So, it can be approved as the best list of references relevant to the suggested title, Fig. 4 illustrates headlines of filtration process (Scan 3).

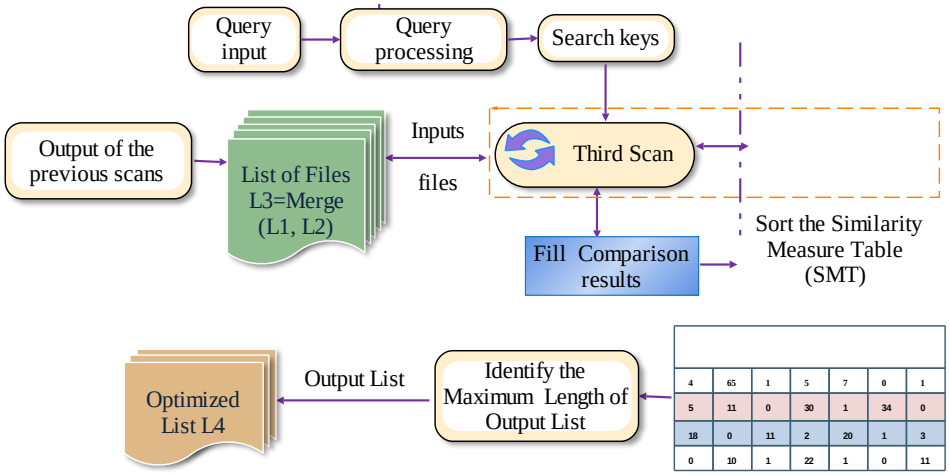

Fig. 4: The Third Filtration Scan

VII. Experimental Results

Three different experiments were applied on the proposed algorithm. The outputs are shown in the similarity tables, Appendix 1, Appendix 2, and Appendix 3. Each experiment includes 34 cases/document files. The contents of the first row in each table indicate the Keywords-Synonyms-Set (KSS), which were applied in Google searches and their values are shown in the second row, they were initialized to "1" for the comparison purpose. These values used to compute the Cosine Similarity between the KISS and their occurrences in all documents collected by the Google Search Engine, in the three experiments. Five keywords and three Synonyms in the first sample (Appendix 1), Five keywords and four Synonyms in the second sample (Appendix 2), and Five keywords and five Synonyms in the third sample (Appendix 3). The documents included in the three samples belong to various international databases, as a result of a Google search. The search results were organized in two separate lists L1 and L2. The two lists merged later in one single list to be used as an input to the third scan using the similarity measure, see the details in Fig. (2 and 3). The similarity measure was applied in the third scan and the results are shown in Appendix 1, 2, and 3, last column. The number of pages got from the Google search in the three experiments are 38, 35, and 41 pages respectively. The total number of items in each experiment is 369, 337, and 402 items respectively. These experiments accomplished to compare the results of the traditional Search Engines against the results of the proposed algorithm. The Cosine Similarity measure of formula "(1)," was applied to find the similarity amongst KSS and documents' contents. The results of Appendix 1, 2, and 3 were sorted in descending order based on the highest similarity values (last column). Each experiment covered various search results from many pages, only sub-sets of search results took randomly from various pages of the Google search results. For instance, in Appendix 1, the sub-sets selected as follows:

4 results (documents) took from page 1 and 4 from page 2. Also, 3 results took from page 3 and 2 results took from page 6. In addition, 2 results took from page 7, 1 from page 8, 3 results from page 10, and 2 results taken from page 38. The same idea followed by the contents of Appendix 2 and 3.

VIII. Results Evaluation and Discussion

When using the leading search engines, such as Google to find a specific subject, the expected results, supposed to be listed based on its importance and correlation degree with the required title. In this case, the search results are organized as a sequential list of the first item to last item, this list should be started with the results that have high correlation or high priority with the researcher's subject or title. This is a general perception for any user/researcher about the common search engines. This picture is that we had before. The results of this research have proven another vision based on the three experiments done using Google Search Engine and the proposed filter with Google.

The results of the three experiments were analyzed, it was found that the traditional Google search results produce dozens of pages and hundreds of results of documents. Mostly, these results appear to be scattered within these pages randomly without taking into account the priority criterion.

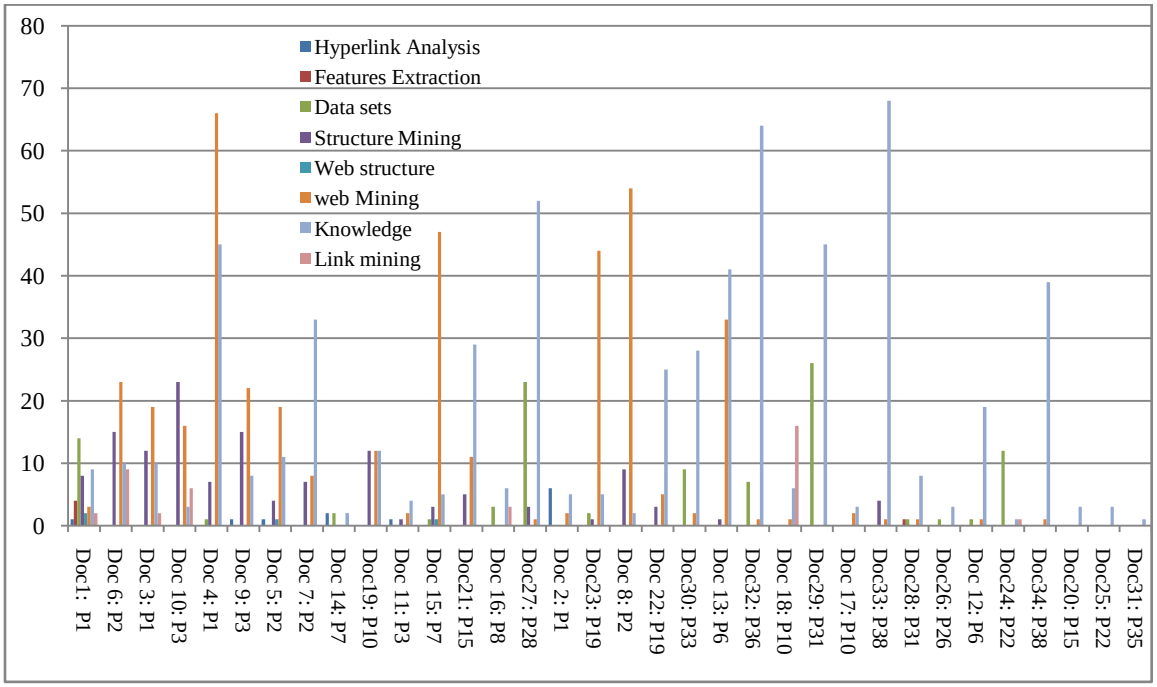

Fig. 5: Occurrences of the KSS in the documents of the first sample

Fig. 5, 6, and 7 show occurrences of KSS in the files found in the search results, some KSS appears once or more in a document. A number of KSS appears many times, such as "web mining as in Appendix 1, it appears 46 times in document 4, in page1, "knowledge" appears 64 times in document 32, in page 36, and "knowledge" appears in document 34, in page 38, 68 times. In addition, there is a number of KSS items that are completely irrelevant and absent in the results, its occurrence is 0, such as "Healthcare Data" in Appendix 2.

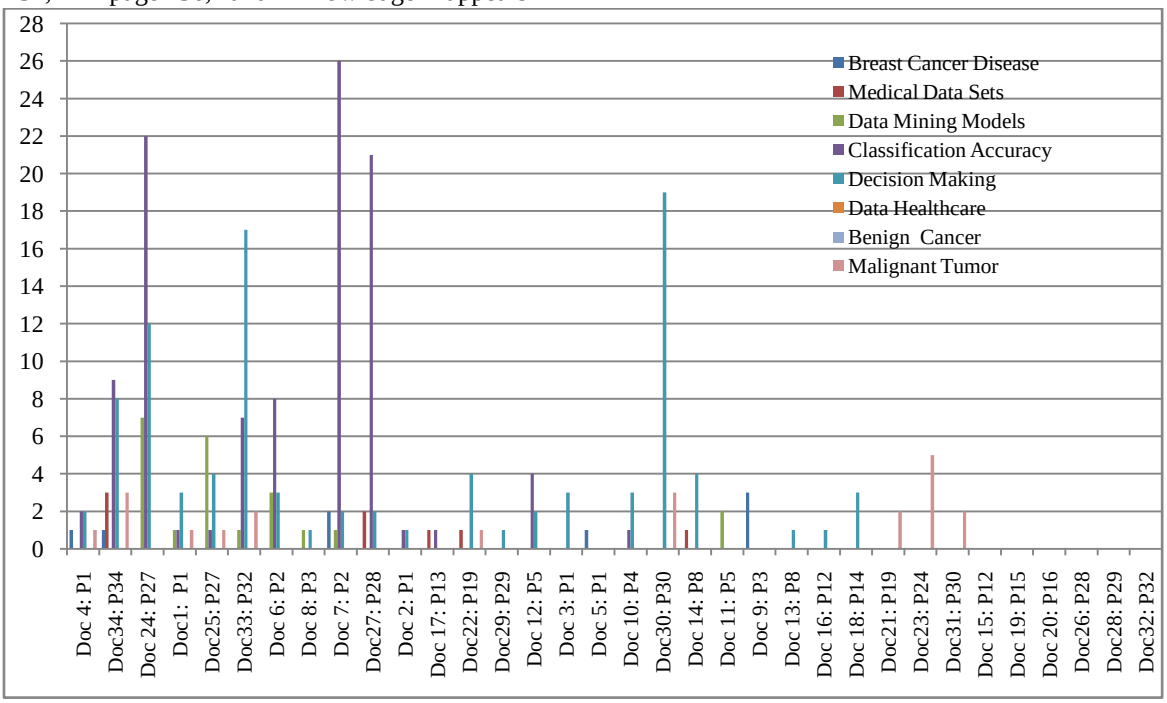

Fig. 6: Occurrences of the KSS in the documents of the second sample

The search results are distributed at different levels. Fig. 6 and 7 demonstrate these levels, according to the experiment 2 and 3. Some of these results seem to be inexact, based on the results classification in different cases:

1. Some of these documents (search results) displayed on the first page and the second page of the search, although, the degree of similarity is not high with the required topic (these are abnormal cases), some examples in Appendix 1

are supporting this case, such as doc 2 in page 1 and doc 8 in page 2. Also, in Appendix 2, docs 2, 3, 5 in page 1, and in Appendix 3, docs 2, 3 on page 1.

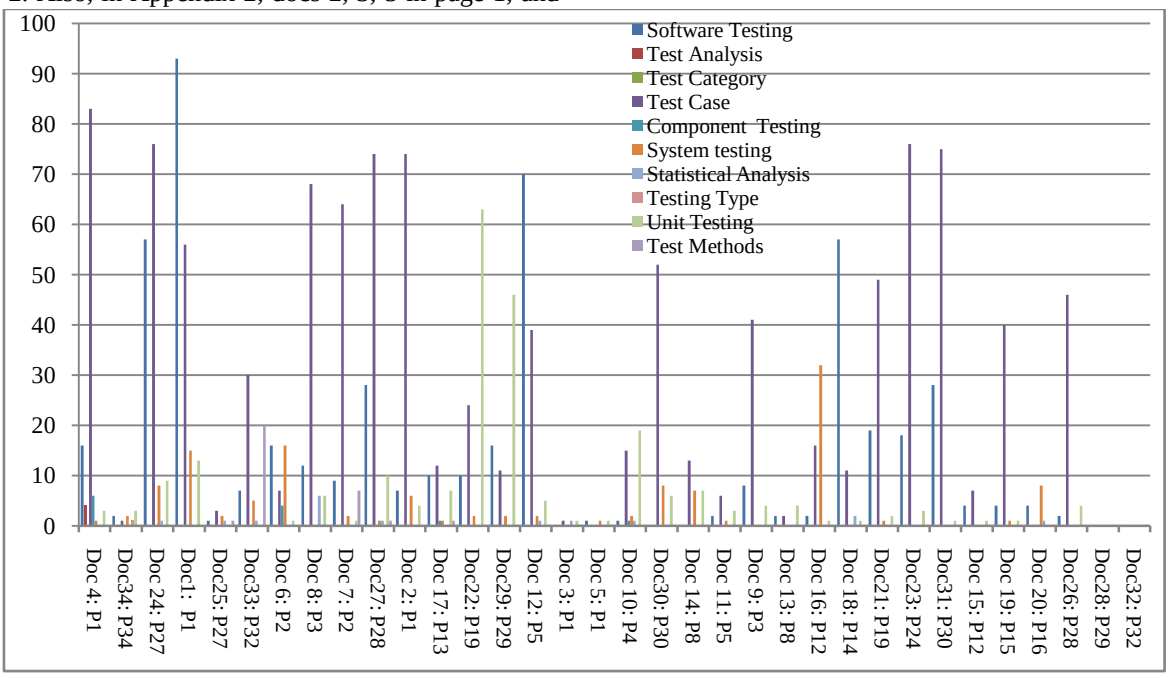

Fig. 7: Occurrences of the KSS in the third sample documents

2. Some of these documents came in the first page and the degree of similarity is high with the required topic (these are normal cases), for instance, in Appendix 1, docs 1, 3, and 4 on page 1. Also, docs 1, 4 in Appendix 2, and docs 1, 4 on page 1, Appendix 3).
3. Some of these documents placed at the end of search lists, although, it has a high similarity (from a logical point of view, these are abnormal cases), it represents a drawback of Google Search Engine, as observed in different cases, for example, in Appendix 1: doc 19 in page 10, doc 21 in page 15, doc 23 in page 19, doc 27 in page 28, in Appendix 2: doc 34 in page 34, doc 24 in page 27, doc 25 in page 27, doc 33 in page 32, and in Appendix 3: doc 29 page 30, doc 34 page 41, and doc 33 page 40. All these documents have a high rate of priority, but came late in the Google search results list.
4. Some of these documents came at the end of search lists with a very low similarity or irrelevant (normal case), many examples were shown in the three tables.

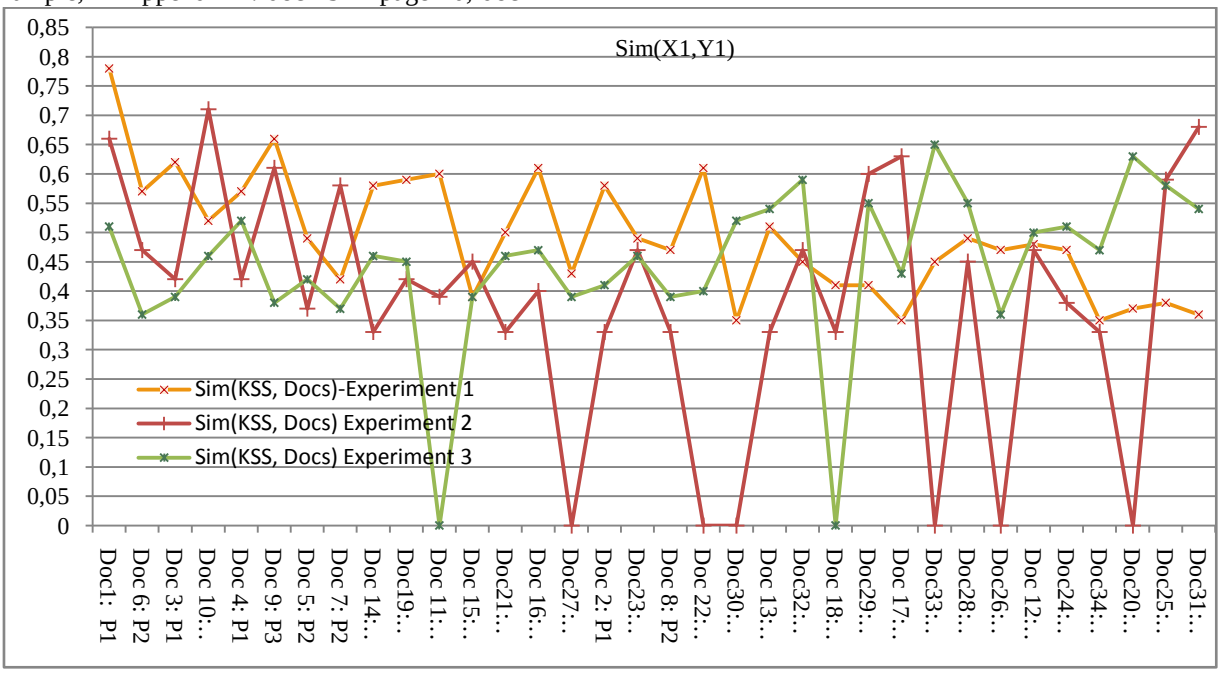

Fig. 8: Results randomization in google search

5. A set of documents is not related to researcher's subject, the similarity value is 0, these documents should be removed from the results, anyone can say, why Google puts them in the results? 6 cases in Appendix 1 and 2 cases in Appendix 3.
6. Other documents are spread between the cases discussed above (from 1 to 5), some of them are familiar cases and the other are unfamiliar cases, many examples are included in these tables. A clear distribution of relevant and irrelevant results is shown in Fig. 8.

Based on the proposed filtration algorithm, it was found that the research results are showing more accurate and optimal results. This algorithm displays the search results in the correct order based on the real similarity degree as in Fig. 9. The final list will start with all items appear in the top left of this Fig., the most relevant topics.

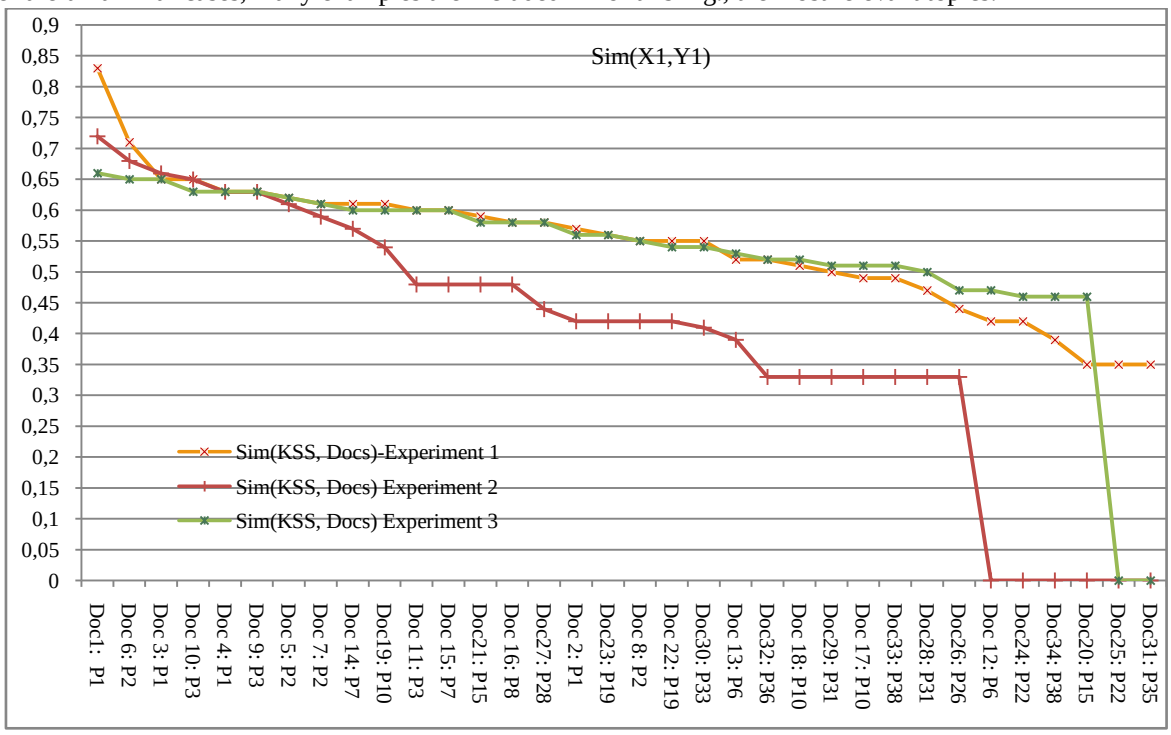

Fig. 9: Results of filtration using the proposed algorithm based on google

Where the curve of experiment 1 represents a sample of 34 random cases out of 369, the curve of experiment 2 represents a sample of 34 random cases out of 337, and the curve of experiment 3 represents a sample of 34 random cases out of 402.

So, in the results of Google, the researcher must check all results of each page from R_1 to R_N and must check all pages from $page_1$ to $page_N$, this is a time consuming when using manual filtration. This kind of results is shown in Fig. 8, it illustrates the results randomization in Google search. The important search results in Google are scattered in different pages, for example, in the first, the second and the third experiments there are 38 pages, 35 pages, and 41 pages, respectively. In this case, the user will search manually on these results, which consists of 369, 337, and 402 items respectively in the three experiments, but in the proposed method, it is enough for a researcher to check only the first part of results, as demonstrated in Fig. 9, which displays the experiments' results of the proposed MSF. Also, the researcher can determine the final list size. With the proposed method the results will be listed in a single list.

IX. Conclusion

The results got according to the proposed filter were much better than those produced according to the traditional Google Search Engine. The results' list was optimized from hundreds of topics organized in tens of pages to a small list controlled by a single parameter (List length), which can be used to determine the number of required references or similar subject. The search by KSS expands the search space and increases the related results. This in turn brings all the documents relating to the subject under study, the filtration process of similarity measure reduces a large number of topics, which have lower similarity or irrelevant.

This gives search results with more focus and more accuracy than those obtained from Google Search Engines. These are counted as an advantage and strength of the proposed filter (Google-Based MSF). This filter optimizes the final results of a Google Search, and increase its accuracy based on the multi scanning and filtering process.

Among the advantages offered by this research is saving researchers' time required for browsing a huge number of files, where the obtained results are focused mostly on the relevant topic. The disadvantages include the filtering time

required to measure the similarity, but when compared with reader time spend for select a list of the appropriate documents, this time does not affect at all, this means this disadvantage has not effect. In general, the Google-Based MSF will improve the search quality. The approach allows researchers, students collect and find a short list of references contains important materials that will be more helpful for them in their research subjects, especially in the historical background or literature review. In addition, this research will also help Search Engines' developers, for instance, Google developers' team to improve the precision and efficiency of their leading Search Engines. So, this paper provides a valuable contribution for improving the search results, which is very important generally for global society.

X. FUTURE WORKS

The research results represent an important and necessary work for a large segment of society, whether researchers, students, experts and other people. As a result of this importance, the research opens the door for additional research works for all those who have an interest to improve search methods, techniques, and tools of data and IR from the global network environment. A list of recommendations regarding to this aspect can be given as follows:

- Development of each part of the proposed algorithm separately, and merge all these parts as a single unit with the Google Search Engine through a flexible user interface that constitutes an integrated Search System with Google.
- The research function can be extended by adding more features to the current idea, based on the title and subject summary, using the text slices overlapping.
- Enhance the research idea to cover general topics, contents out of the research purpose after connecting the system with an extended database contains all famous Synonyms.
- Improve the filtration process, using other statistical measures, based on the Data Mining Tools and Techniques.

Appendices:

Appendix 1: Experimental results of the first sample

Documents	Keywords					Synonyms			Metrics			Measure
	Hyperlink Analysis	Features Extraction	Data sets	Structure Mining	Web analysis	Web Mining	Knowledge	Link Mining	‖X‖	‖Y‖	X×Y	Sim(X,Y)
RS Values	1	1	1	1	1	1	1	1	-	-	-	-
Doc1: page1	1	4	14	8	2	3	9	2	2.83	1.67	3.9	0.83
Doc 6: page2	0	0	0	15	0	23	10	9	2.83	1.95	3.9	0.71
Doc 3: page1	0	0	0	12	0	19	10	2	2.83	1.74	3.2	0.65
Doc 10: page3	0	0	0	23	0	16	3	6	2.83	1.57	2.9	0.65
Doc 4: page1	0	0	1	7	0	66	45	0	2.83	1.58	2.8	0.63
Doc 9: page3	1	0	0	15	0	22	8	0	2.83	1.63	2.9	0.63
Doc 5: page2	1	0	0	4	1	19	11	0	2.83	1.48	2.6	0.62
Doc 7: page2	0	0	0	7	0	8	33	0	2.83	1.46	2.5	0.61
Doc 14: page7	2	0	2	0	0	0	2	0	2.83	0.35	0.6	0.61
Doc19: page10	0	0	0	12	0	12	12	0	2.83	1.73	3	0.61
Doc 11: page3	1	0	0	1	0	2	4	0	2.83	0.47	0.8	0.6
Doc 15: page7	0	0	1	3	1	47	5	0	2.83	1.17	2	0.6
Doc21: page15	0	0	0	5	0	11	29	0	2.83	1.5	2.5	0.59
Doc 16: page8	0	0	3	0	0	0	6	3	2.83	0.73	1.2	0.58
Doc27: page28	0	0	23	3	0	1	52	0	2.83	1.45	2.4	0.58
Doc 2: page1	6	0	0	0	0	2	5	0	2.83	0.81	1.3	0.57
Doc23: page19	0	0	2	1	0	44	5	0	2.83	1.14	1.8	0.56
Doc 8: page2	0	0	0	9	0	54	2	0	2.83	1.36	2.1	0.55
Doc 22: page19	0	0	0	3	0	5	25	0	2.83	1.16	1.8	0.55
Doc30: page33	0	0	9	0	0	2	28	0	2.83	1.36	2.1	0.55
Doc 13: page6	0	0	0	1	0	33	41	0	2.83	1.42	2.1	0.52
Doc32: page36	0	0	7	0	0	1	64	0	2.83	1.22	1.8	0.52
Doc 18: page10	0	0	0	0	0	1	6	16	2.83	1.17	1.7	0.51
Doc29: page31	0	0	26	0	0	0	45	0	2.83	1.41	2	0.5
Doc 17: page10	0	0	0	0	0	2	3	0	2.83	0.36	0.5	0.49
Doc33: page38	0	0	0	4	0	1	67	0	2.83	1.08	1.5	0.49
Doc28: page31	0	1	1	0	0	1	8	0	2.83	0.82	1.1	0.47

Doc26: page26	0	0	1	0	0	0	3	0	2.83	0.32	0.4	0.44
Doc 12: page6	0	0	1	0	0	1	19	0	2.83	1.01	1.2	0.42
Doc24: page22	0	0	12	0	0	0	1	1	2.83	1.01	1.2	0.42
Doc34: page38	0	0	0	0	0	1	68	0	2.83	1	1.1	0.39
Doc20: page15	0	0	0	0	0	0	3	0	2.83	0.3	0.3	0.35
Doc25: page22	0	0	0	0	0	0	3	0	2.83	0.3	0.3	0.35
Doc31: page35	0	0	0	0	0	0	1	0	2.83	0.1	0.1	0.35

Appendix 2**:** Experimental results of the second sample

Documents	Keywords						Synonyms				Metrics			Measure
	Breast ancer Disease	Data Sets Medical	Data Mining Models	Accuracy Classification	Decision Making		Healthcare Data	Benign Cancer	Malignant Tumor	Mining Methods	‖X‖	‖Y‖	X×Y	Sim(X,Y)
RS Values	1	1	1	1	1		1	1	1	1	-	-	-	-
Doc 4: page1	1	0	0	2	2		0	0	1	2	3	0.37	0.8	0.72
Doc34: page34	1	3	0	9	8		0	0	3	3	3	1.32	2.7	0.68
Doc 24: page27	0	0	7	22	12		0	0	0	8	3	1.77	3.5	0.66
Doc1: page1	0	0	1	1	3		0	0	1	3	3	0.46	0.9	0.65
Doc25: page27	0	0	6	1	4		0	0	1	3	3	0.79	1.5	0.63
Doc33: page32	0	0	1	7	17		0	0	2	12	3	1.59	3	0.63
Doc 6: page2	0	0	3	8	3		0	0	0	4	3	0.99	1.8	0.61
Doc 8: page3	0	0	1	0	1		0	0	0	1	3	0.17	0.3	0.59
Doc 7: page2	2	0	1	86	2		0	0	0	5	3	1.16	2	0.57
Doc27: page28	0	2	0	21	2		0	0	0	4	3	1.11	1.8	0.54
Doc 17: page13	0	1	0	1	0		0	0	0	0	3	0.14	0.2	0.48
Doc 2: page1	0	0	0	1	1		0	0	0	0	3	0.14	0.2	0.48
Doc22: page19	0	1	0	0	4		0	0	1	0	3	0.42	0.6	0.48
Doc29: page29	0	0	0	0	1		0	0	0	1	3	0.14	0.2	0.48
Doc 12: page5	0	0	0	4	2		0	0	0	0	3	0.45	0.6	0.44
Doc 10: page4	0	0	0	1	3		0	0	0	0	3	0.32	0.4	0.42
Doc 3: page1	0	0	0	0	3		0	0	0	1	3	0.32	0.4	0.42
Doc 5: page1	1	0	0	0	0		0	0	0	3	3	0.32	0.4	0.42
Doc30: page30	0	0	0	0	19		0	0	3	0	3	1.04	1.3	0.42
Doc 14: page8	0	1	0	0	4		0	0	0	0	3	0.41	0.5	0.41
Doc 11: page5	0	0	2	0	0		0	0	0	11	3	1.02	1.2	0.39
Doc 13: page8	0	0	0	0	1		0	0	0	0	3	0.1	0.1	0.33
Doc 16: page12	0	0	0	0	1		0	0	0	0	3	0.1	0.1	0.33
Doc 18: page14	0	0	0	0	3		0	0	0	0	3	0.3	0.3	0.33
Doc 9: page3	3	0	0	0	0		0	0	0	0	3	0.3	0.3	0.33
Doc21: page19	0	0	0	0	0		0	0	2	0	3	0.2	0.2	0.33
Doc23: page24	0	0	0	0	0		0	0	5	0	3	0.5	0.5	0.33
Doc31: page30	0	0	0	0	0		0	0	2	0	3	0.2	0.2	0.33
Doc 15: page12	0	0	0	0	0		0	0	0	0	3	0	0	0
Doc 19: page15	0	0	0	0	0		0	0	0	0	3	0	0	0
Doc 20: page16	0	0	0	0	0		0	0	0	0	3	0	0	0
Doc26: page28	0	0	0	0	0		0	0	0	0	3	0	0	0
Doc28: page29	0	0	0	0	0		0	0	0	0	3	0	0	0
Doc32: page32	0	0	0	0	0		0	0	0	0	3	0	0	0

Appendix 3**:** Experimental results of the third sample

Documents	Keywords					Synonyms					Metrics			Measure
	Software Testing	Test Analysis	Test Category	Test Case	Component Testing	System testing	Statistical Analysis	Testing Type	Unit Testing	Test Methods	‖X‖	‖Y‖	X×Y	Sim(X,Y)
KS Values	1	1	1	1	1	1	1	1	1	1	-	-	-	-
Doc1: page1	93	0	0	83	0	15	0	0	13	0	3.16	2	4	0.63
Doc29: page 30	57	0	0	1	0	8	1	0	9	0	3.16	1.86	3.8	0.65
Doc 7: page 2	16	4	0	76	6	1	0	0	3	0	3.16	1.62	3.4	0.66
Doc 9: page 3	28	0	0	56	0	1	1	0	10	1	3.16	1.74	3.3	0.6
Doc34: page 41	7	0	0	3	0	5	1	0	0	20	3.16	1.66	3.3	0.63
Doc 4: page 1	10	0	0	30	0	2	0	0	63	0	3.16	1.74	3.2	0.58
Doc 8: page 3	12	0	0	7	0	0	6	0	6	0	3.16	1.65	3.2	0.61
Doc14: page 9	16	0	0	68	0	2	0	0	46	0	3.16	1.74	3.2	0.58
Doc33: page 40	16	0	0	64	4	16	0	0	1	0	3.16	1.63	3.2	0.62
Doc22: page 17	10	0	0	74	1	1	0	0	7	1	3.16	1.59	3	0.6
Doc 6: page 2	9	0	0	74	0	2	0	0	1	7	3.16	1.53	2.9	0.6
Doc 17: page 11	70	0	0	12	0	2	1	0	5	0	3.16	1.52	2.8	0.58
Doc 12: page 6	7	0	0	24	0	6	0	0	4	0	3.16	1.42	2.7	0.6
Doc30: page 32	1	0	0	11	1	2	1	0	19	0	3.16	1.44	2.5	0.55
Doc 3: page 1	0	0	0	39	0	8	0	0	6	0	3.16	1.41	2.4	0.54
Doc 5: page 2	0	0	0	1	0	7	0	0	7	0	3.16	1.41	2.4	0.54
Doc 10: page 3	2	0	0	0	0	32	0	0	1	0	3.16	1.43	2.3	0.51
Doc15: page 9	18	0	0	15	0	0	0	0	3	0	3.16	1.45	2.3	0.5
Doc18: page 11	57	0	0	52	0	0	2	0	1	0	3.16	1.43	2.3	0.51
Doc25: page 22	19	0	0	13	0	1	0	0	2	0	3.16	1.43	2.3	0.51
Doc19: page 14	8	0	0	6	0	0	0	0	4	0	3.16	1.34	2.2	0.52
Doc16: page 9	28	0	0	41	0	0	0	0	1	0	3.16	1.42	2.1	0.47
Doc 2: page 1	4	0	0	2	0	1	0	0	1	0	3.16	1.09	1.6	0.46
Doc28: page 27	2	0	0	16	0	0	0	0	4	0	3.16	1.1	1.6	0.46
Doc 13: page 6	4	0	0	11	0	8	1	0	0	0	3.16	0.9	1.3	0.46
Doc21: page 17	2	0	0	49	0	1	0	0	3	0	3.16	0.71	1.2	0.53
Doc31: page 37	4	0	0	76	0	0	0	0	1	0	3.16	0.81	1.2	0.47
Doc26: page 23	2	0	0	75	0	2	0	1	3	0	3.16	0.44	0.9	0.65
Doc20: page 14	2	0	0	7	0	0	0	0	4	0	3.16	0.49	0.8	0.52
Doc32: page 40	1	0	0	40	0	2	1	0	0	1	3.16	0.4	0.8	0.63
Doc24: page 20	0	0	0	0	0	0	1	0	1	0	3.16	0.17	0.3	0.56
Doc27: page 24	1	0	0	46	0	1	0	0	1	0	3.16	0.17	0.3	0.56
Doc 11: page 6	0	0	0	0	0	0	0	0	0	0	3.16	0	0	0
Doc23: page 20	0	0	0	0	0	0	0	0	0	0	3.16	0	0	0

REFERENCES

[1] P. Gupta and D. a. K. Sharma, "Context based indexing in search engines using ontology," *Int. J. Comput. Appl.*, vol. 1, no. 14, pp. 53–56, 2010.

[2] K. Selvakumar and S. Sendhilkumar, "Challenges and recent trends in personalized web search: a survey," *3rd Int. Conf. Adv. Comput. ICoAC 2011*, pp. 333–339, 2011.

[3] D. Donato, F. Bonchi, T. Chi, and Y. Maarek, "Do you want to take notes?: identifying research missions in Yahoo! search pad," *Proc. 19th Int. Conf. World wide web SE - WWW '10*, pp. 321–330, 2010.

[4] G. Golovchinsky and J. Pickens, "Interactive information seeking via selective application of contextual knowledge," *IIiX '10 Proceeding third Symp. Inf. Interact. Context*, no. C, pp. 145–154, 2010.

[5] R. Schifanella, L. M. Aiello, and G. Ruffo, "Tagging relations to achieve complex search goals," *Proc. 2011 7th Int. Conf. Next Gener. Web Serv. Pract. NWeSP 2011*, pp. 308–313, 2011.

[6] B. Smyth, E. Balfe, J. Freyne, P. Briggs, M. Coyle, and O. Boydell, "Exploiting query repetition and regularity in an adaptive community-based web search engine," *User Model. User-Adapted Interact.*, vol. 14, no. 5, pp. 383–423, 2004.

[7] A. Hotho, R. J??schke, C. Schmilz, and G. Stumme, "Information retrieval in folksonomies: search and ranking," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2006, vol. 4011 LNCS, pp. 411–426.

[8] L. Stoilova, T. Holloway, B. Markines, A. G. Maguitman, and F. Menczer, "GiveALink," *Proc. 3rd Int. Work. Link Discov. - LinkKDD '05*, pp. 66–73, 2005.

[9] B. Markines, B. Markines, C. Cattuto, C. Cattuto, F. Menczer, F. Menczer, D. Benz, D. Benz, A. Hotho, A. Hotho, G. Stumme, and G. Stumme, "Evaluating similarity measures for emergent semantics of social tagging," in *Proceedings of the 18th international conference on World Wide Web, WWW' 09, ACM*, 2009, pp. 641–650.

[10] R. Schifanella, A. Barrat, C. Cattuto, B. Markines, and F. Menczer, "Folks in folksonomies," *Proc. third ACM Int. Conf. Web search data Min. - WSDM '10*, p. 271, 2010.

[11] M. Andago, T. P. L. Phoebe, and B. A. M. Thanoun, "Evaluation of a semantic search engine against a keyword search engine using first 20 precision," no. 2005, pp. 55–63, 2010.

[12] D. Sharma and A. K. Sharma, "Crawler indexing using tree structure and its implementation," vol. 31, no. 6, pp. 34–39, 2011.

[13] A. Gaudinat, P. Ruch, M. Joubert, P. Uziel, A. Strauss, M. Thonnet, R. Baud, S. Spahni, P. Weber, J. Bonal, C. Boyer, M. Fieschi, and A. Geissbuhler, "Health search engine with e-document analysis for reliable search results," *Int. J. Med. Inform.*, vol. 75, no. 1, pp. 73–85, 2006.

[14] P. Arora and T. Bhalla, "A synonym based approach of data mining in search engine optimization," vol. 12, no. 4, p. 5, 2014.

[15] S. Decker and D. T. E. R. Eport, "Searching and browsing linked data with swse : the s emantic web search engine," *Science (80-.).*, vol. 1, no. 5, pp. 162–167, 2010.

[16] R. Singh and S. K. Gupta, "Search engine optimization - using data mining approach introduction :," vol. 2, no. 9, pp. 28–32, 2013.

[17] G. Madhu, A. Govardhan, and T. K. V. Rajinikanth, "Intelligent semantic web search engines: a brief survey," *Int. J. Web Semant. Technol.*, vol. 2, no. 1, pp. 34–42, 2011.

[18] O. K. Shade, O. O. Samuel, U. K. Richmond, and A. Oludele, "Trends in web-based search engine," *J. Emerg. Trends Comput. Inf. Sci.*, vol. 3, no. 6, pp. 942–948, 2012.

[19] E. Goldman, "Search engine bias and the demise of search engine utopianism," *Web Search. Inf. Sci. Knowl. Manag. vol. 14*, vol. 188, pp. 121–133, 2008.

[20] D. Lewandowski, "A three-year study on the freshness of web search engine databases," *J. Inf. Sci.*, vol. 34, no. 6, pp. 817–831, 2008.

[21] R. Aravindhan and R. Shanmugalakshmi, "Comparative Analysis of Web 3 . 0 Search Engines : A survey report," *Int. Conf. Comput. Commun. Informatics (ICCCI -2014)*, pp. 1–6, 2014.

[22] F. Zeshan and R. Mohamad, "Semantic Web service composition approaches : overview and limitations," *Int. J. New Comput. Archit. Their Appl.*, vol. 1, no. 3, pp. 640–651, 2011.

[23] R. G. Demirci, V. Kismir, and Y. Bitirim, "An evaluation of popular search engines on finding turkish documents," *Second Int. Conf. Internet Web Appl. Serv. ICIW'07*, pp. 7–11, 2007.

[24] W. Liangshen, H. Jie, X. Zaiyu, W. Xiaochen, Q. Caiyue, and L. Hui, "Problems and solutions of web search engines," *2011 Int. Conf. Consum. Electron. Commun. Networks*, pp. 5134–5137, 2011.

[25] J. A. Josephine and S. Sathiyadevi, "Ontology based relevance criteria for semantic web search engine," *2011 3rd Int. Conf. Electron. Comput. Technol.*, vol. 3, pp. 60–64, 2011.

[26] J. Han, M. Kamber, and J. Pei, *Data Mining : concepts and techniques Third Edition*. 2011.

[27] K. Yang and L. I. Meho, "Citation analysis: a comparison of google scholar, scopus, and web of science," vol. 43, 2006.

[28] W. H. Walters, "Google Scholar coverage of a multidisciplinary field," *Inf. Process. Manag.*, vol. 43, no. 4, pp. 1121–1132, 2007.

[29] J. J. Meier and T. W. Conkling, "Google scholar's coverage of the engineering literature: an empirical study," *J. Acad. Librariansh.*, vol. 34, no. 3, pp. 196–201, 2008.

[30] J. Bar-Ilan, "Which h-index? - A comparison of WoS, scopus and google scholar," *Scientometrics*, vol. 74, no. 2, pp. 257–271, 2008.

[31] P. Jacsó, "Google scholar: the pros and the cons," *Online Inf. Rev.*, vol. 29, no. 2, pp. 208–214, 2005.

[32] B. White, "Examining the claims of google scholar as a serious information source," n*ew zeal. library information management journal*, vol. 50, no. 1, pp. 11–24, 2006.

[33] R. Subhashini and V. J. S. Kumar, "Evaluating the performance of similarity measures used in document clustering and information retrieval," *2010 First Int. Conf. Integr. Intell. Comput.*, pp. 27–31, 2010.