International Journal on Advanced Science Engineering Information Technology

Vol.10 (2020) No. 3
ISSN: 2088-5334

Improving Faceted Search Results for Web-based Information Exploration

Mohammed Najah Mahdi[a], Abdul Rahim Ahmad[b], Roslan Ismail[c]

[a] *Institute of Informatics and Computing in Energy, College of Computing & Informatics (CCI), Universiti Tenaga Nasional, Malaysia*
E-mail: najah.mahdi@uniten.edu.my

[b] *Department of System & Networking, College of Computing & Informatics (CCI), Universiti Tenaga Nasional, Malaysia*
E-mail: abdrahim@uniten.edu.my

[c] *Department of Software Engineering, College of Computing & Informatics (CCI), Universiti Tenaga Nasional, Malaysia*
E-mail: roslan@uniten.edu.my

***Abstract*— The World Wide Web (WWW), a fast-growing store, contains a significant portion of human knowledge. However, the sheer scale of the Web, along with the fact that it is decentralized, highly redundant, and largely inaccurate, causes the use of the knowledge is quite cumbersome. The present search engines (SEs) use the query, and the response lookup process is incapable of producing a precise result. Thus, researchers work beyond this paradigm to explore a new class of methods to seek information, which known as an exploratory search (ES). This ES is open-ended, and its faceted search (FS) improves the overall search process. The search engine presented in this study is running in the cloud computing platform environment. Its development is based on the idea of improving visual ES while exploring information on the Web. This notion reflects the process of seeking and combing the vast information by using the coordinated visualization method, apart from minimizing the effort spent in seeking information per query. Finally, we evaluate the proposed prototype against the Internet Movie Database (IMDb) search engine, an online database of information related to films, television programs, home videos, video games, and streaming content online including cast, production crew, and personal biographies, plot summaries, trivia, fan, and critical reviews, and ratings. The results show that the proposed search engine gives more relevant search results as compared with the others.**

***Keywords*— exploratory search; faceted search; visualization; search engine.**

I. INTRODUCTION

When considering the rise in the demand for information from the search engine, the quantity of data deposited on the WWW keeps growing. For example, in May 2016, the volume of the WWW was roughly 4.6 billion webpages [1]. The Web is continuously increasing in the amount of content, as well as their diversity. These include the number of multimedia content such as blogs and Wikipedia.

With the emergence of the WWW, several users using the internet as a medium for exploring, finding, and discovering credible information has been increasing quite tremendously. People are heavily relying on online materials to access the information to fulfill their needs [2]. This contributes to the development of a more complex and exploratory search engine to satisfy their tasks in finding the related documents. Nevertheless, the existing online SEs, have a limited capability to rank the retrieve information as well as the inefficient and ineffective process involved [3]. Examining the SE process of the user, an essential aspect beyond the scope of present SEs, is individually performing a documents SE task. Though the SEs smartly evolve to track the histories and preferences of users for personalizing the searched results and suggesting the queries, they often do not concentrate on the users' searched tasks [4]. Thus, they fail to intelligently support SE path suggestions, including the next query to be executed, the queries to be excluded, the Web offering helpful documents for the task, or the relevant documents to be considered to accomplish the researcher task goal[5].

With the advancement of the Internet, IT support systems is compelled to support enough storage space and faster computing facilities, for example, search engines for Internet applications. These problems can be effectively solved through the technique of cloud computing. Following this,

we presented a SE prototype system, in this paper, based on the cloud-computing platform [6]. The idea of a visual ES engine exploring information on the Web has further developed in this work. This concept aims at supporting techniques of finding information through coordinated visualizations by reducing the average searched effort required per each query. These mentioned approaches are habitually by commercial searched engines.

The structure of the paper is organized as follows:

- The overview of related works, interactive visuals, and guides in general on visualization based faceted search for SEs, are provided in section II.
- In section III, the overview of the SE framework outlined with some illustrations of the methodology, through which the data were collected and the results were implemented.
- The test and evaluation of the proposed prototype search system is carried out in section IV.
- Finally, the conclusion remark on how to improve the relevancy of Web search results has been presented.

II. Materials and Method

An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it. The relevant literature on information exploration and navigation runs across multiple research groups. This part elaborates several essential concepts within the scope of faceted search and information visualization. ES is any SE behavior identified by considerable uncertainty about the target of the search or lack of knowledge about the domain [7]. Thus, researchers attempt to seek information may require support that can assist them search through the unclear domain. Due to the level of uncertainty in the SE process and the required information needed by a researcher, it can be hard to provide help to researchers working on ES to find their search target. ES, searching technique, allows the information searchers to undergo complex cognitive tasks leading to learning and discovering [8]. When conducting searching tasks, ES learns and develop skills through the researchers' history and browsing behavior. Researchers who managing ES will compare information with other search engines using a similar command, which several properties can be understood (see Fig 1).

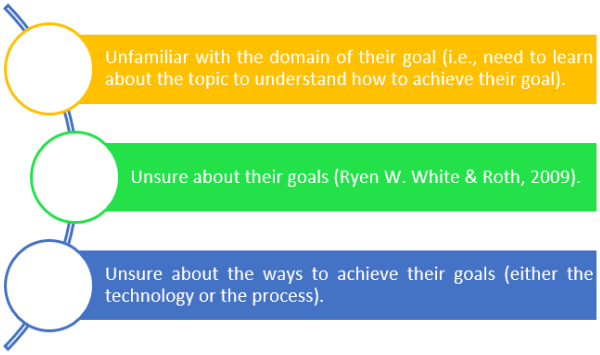

Fig. 1 Exploratory Search Classification Activities

FS, a search technique, has growing to be well-developed and popular for business and educational websites[10]. Because of the enormous information on the Web, researchers who look for information provider that can help them analyze such information and provide a valuable output. FS is a novel exploratory search engine process that helps a researcher discovers useful information within a large data repository. This enables researchers to filter or explore a set of documents by utilizing a number of discrete attributes or facets[11].

The target of this paper is to investigate the use of ES in information visualization for faceted search. The goal of any search engine is to simplify the comprehension of a complex structure and turn it into in a simple form [15-17]. Therefore, this study tries to answer two questions; first. Is it feasible to implement a cloud based search engine and if yes, what is its impact on the search engine's performance. Secondly, how to evaluate the proposed ES engine and estimate its capability.

We propose the SE prototype to overcome the limitations of the existing ones. The features in the proposed ES engine of information visualization are used to improve the comprehension of huge amounts of data or information through graphical representations. Such system can be suitable for the cloud computing platform environment. The prototype is focused on the idea of a visual ES for documents discovery on the Web. This concept is aimed at supporting an active means of finding and discovering information through coordinated visualizations and reducing the average quantity of search efforts required per query. The prototype is evaluated regarding its capability and is then compared with IMDb SE. In addition, the approach attempts to extend the current consensus around traditional search naturally. The system is built with the idea that no loss should be incurred in moving from one paradigm to the next. That is, whatever could be performed in the old setting will still hold in the new setting. The system eventually fulfils the list of what the researcher believes would be the main natural component of ES.

A. *Cloud Computing*

Cloud Computing (CC) can be considered as a model for computing infrastructure, hosting, and managing such infrastructure. Using CC helps in the management of resources, lowering the costs, and accessing the resources whose sizes fluctuate with demands [12]. CC has been widely applied to host web sites, process large batch of jobs, and many similar jobs. Despite the potentials and advantages of CC, the technique has not been extensively applied to enterprise applications including backup, shared file systems, as well as other internal systems [13]. To make a CC suitable for these applications, several challenges should be solved. These include the cost, security, performance, and interface mismatch. CC can be considered as a way of outsourcing computer hardware, which can make it easier to build many new applications. For example, hardware outsourcing has been used for some time since at least the rise of the Web. Web hosting supports have been offering servers for rent to controlled data centers [14].

B. *Advantages of Faceted Search*

Faceted search provides a more dynamic means to browse and search for resources than traditional "advanced" search form, where all the available search fields are provided at

once [18]. In the traditional manner, users must set up the search criteria at the beginning of the search. However, users may not be completely clear about the keywords in all the dimensions when they initiate the search, and thus, the traditional search is unsuitable for typical searches. Moreover, enterprises attempt to provide additional valuable information to the existing structured template [19]. Consequently, structured properties may increase to an extremely large number, which leads to a challenging search task and a loss of search focus. In addition, users may want to select a combination of values that does not even exist in the document data set, and thus, an ideal solution is a navigational system provided by FS that can guide users to their areas of interest [20].

C. Information Visualization for Search and Similarity Search

FS provides users the opportunity to explore within the collection of documents. However, only a fixed mode of interaction can be featured in most mainstream search systems [21]. For example, searched outputs are most often reproduced as a list of text with smaller interactions, likes sorting or paging. Understanding new information, which allows for various interaction modes, is required. As stated to White and Roth [22], ES system is supposed to focus on user control and responsibility. This feature should allow the researcher to choose how data is visualized depending on the goal of interest. Therefore, this study goes beyond the traditional FS to investigate the way information visualization could be employed to make the user experience more exploratory.

First, the query terms will be revisited, and then, several examples of visualizations that can be applied to SE results or to facets will be covered. A typical search situation indicates how a user can obtain a set of matching documents when inputting a set of query terms. The query terms usually exist in the search box. To rewrite such queries, the researcher can click in the input SE box and edit query terms, manually. A various approach allows users to interact with keyword terms more directly. These terms are usually reproduced in the form of tags with actions, likes removing, toggling, or clearing. The user can easily manipulate the query faster, thereby obtaining reduced or broader SE outputs[23].

Different visual play on the keyword is based on supporting related suggestions. Query suggestions are as a result of extensive research on IR in query expansion[24, 25]. The idea behind this query suggestion is to provide the researcher with additional keywords, which could guide the SE towards related information. In its most simple usage, the suggested query terms simply act as shortcuts to old typed keywords [22]. However, suggestions can assist the researcher to explore a group of query terms leading to new information of interest. Query suggestion can be commonly implemented within large commercial SE by applying substantial SE logs[26]. If a researcher was to type the query, for example “The Godfather”, the key in by other researchers may have led to suggest “movie”. The user may have been unaware of the movie, “The Godfather”. If the researcher had been replaced the looking for “Mario Puzo”, then the SE might more simply, have suggested the query term “book”. These suggestions most often presented in the form of a list. However, other SEs have tried more attractive visualizations with more accurate results since a SE can respond to a user query with enormous related web pages.

Quintura is a visual search engine (sometimes called find engine) that uses tag clouds and a rich interface to dynamically display and refine results. The interface includes a query box at the top of the page, a tag cloud on the left side of the monitor, and a rich results display on the other side of the page. (See Fig 2).

A flaw in query suggestion is how the method may influence the researchers to the most established approaches, thus reducing discovering. This issues is usually identified as “query drifting”[27]. In this scenario, most users present a much narrower group of the Web. The outputs retrieved by Google may be coming from its cache, as well[28]. One way of addressing this problem is by providing a clear response between the query and the related results. This feature would then lead to another topic of discussion, tight coupling between query terms and searched outputs in the form of dynamic queries [29].

FS considers another fundamental feature; namely, the representation of facets that are related to both the metadata and the search subject. This can be used to represent the data in a significant way [30].

Fig. 2 Quintura Represents Suggested Query Terms

Di Sciascio et al. [31] used the social search concept to develop an ES interface that can influence the data left by past users of such information models. This increases the power of the ES prototype. To improve the ranking of results, they also proposed a prototype that incorporates the functionality of social search into the ES prototype, which can be integrated with ES features. This prototype allows the user to (1) find the information required by using relevant keywords and to (2) combine social knowledge based on tag matching and user similarity.

Ahn and Brusilovsky [32] provide a particular way to combine personalized search and interactive visualization by presenting a base search engine adaptive visualization called Visual Information Browsing Environment (VIBE). This can perform the task. The authors suggested that both the search engine and interactive visualization approaches could be very useful to merge the powerful parts of a personalized

search. In addition, this can be taken when discovering the personalization value in a more interactive SE context, which is like the VIBE approach to search interactive visualization results. Furthermore, the precision and productivity can be enhanced using VIBE especially in personalized search.

Beecham, et al. [33] proposed the FS Views of Varying Emphasis (FaVVEs). This was designed at three levels of flexibly: space, time and description. When varying the abstraction level, specific perspectives can be brought into abstraction for each perspective, or out of focus. Beecham focused on geo-located event data. When using a dataset of crime reports in Chicago, the view combinations were built. To recap briefly, the contributions of FaVVE are given: (a) designing a framework for multi-perspective; (b) building and applying this framework to spatiotemporal-thematic event documents; and (c) evaluating the presented structure via a researcher.

D. Overview of The Search Engine Framework

In Fig. 3, we outlined the overview of the SE framework. Using specific starting keywords, users can visually and interactively discover and explore the relationship graphs with the aid of the client side of the system. Meanwhile, the CC server can construct and return relationship graphs through metasearch and knowledge integration. The framework contains three most important modules: inference and representation of visualization relationship graphs on the Web via Similarity Search, visual ES of relationship graphs through both browsing strategies and human-computer interactions, and querying via the multi- interface.

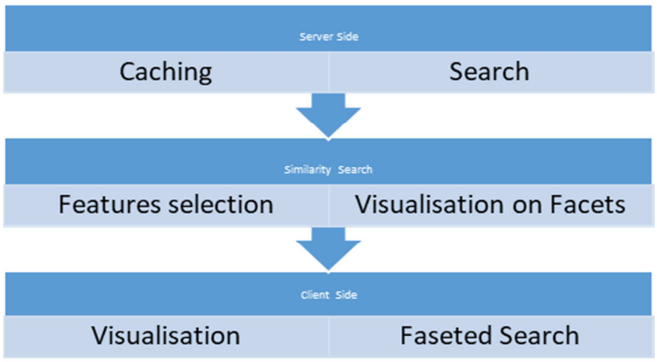

Fig. 3 Overview of the SE Framework

E. Similarity Search

We have discussed faceted search and gone beyond this paradigm with information visualization. By employing the appropriate visualization, either for search results or facets, the user can make greater sense of the data, making his experience more exploratory[34].

Presenting similar sets of results, as well as discovering new ones, is another function of an exploratory search system. Therefore, our system should be accomplished to retrieve the sets of results, not only possible using the full text search. This technique is the most important one since the data at present is multimedia in nature including images and videos. We, therefore, turn our attention to retrieval methods focusing on the whole content of documents[35].

F. Search in Metric Space

Upon devising the feature space, one may choose a proper distance measure to match the documents, an approach known as "NNS". The simple distance measures between vectors[35]. However, NNS can be more difficult when the dimension of such space increases. This problem is simply meaning the "curse of dimensionality." To overcome it, we proposed two techniques. While the first approach deals with increasing the efficiency of NNS, the second one collates the features into textual fingerprints[36].

G. Distances

Huge distances measure several feature spaces. Perhaps, the simple metric between real value feature vectors of a constant dimension n are those presented by the Minkowski norm L_p:

$$\text{d-p(v,w)=l-p(w-v)} \tag{1}$$

whereby v and w represent the two vectors in R^n, and l_p is simply the Minkowski norm which was defined as:

$$w \rightarrow L_p(w) = |w|_p = \left(\sum_{i=0}^{n} |w_i|^p\right)^{1/p} \tag{2}$$

L_2 is the Euclidean distance between the two points.L_1 is the "Manhattan norm," representing the distance through which a car must travel in a rectangular street grid to reach point b from point a. L_∞ stands for the maximum norm or Chebyshev norm corresponding to the maximum of the components. A norm must strictly present not all measures of similarity. As example is cosine similarity, measuring the angle between two vectors:

$$d_{cos}(v,w) = 1 - \frac{v \,.\, w}{L_2(v)\, L_2(w)} \tag{3}$$

Many measures of similarity exist. This depends on the nature of the feature space. For example, if the feature vectors are given as probability vectors (vectors with non-negative components that add up to one), the Kullback-Leibler divergence can be a good measure. The divergence of Kullback-Leibler can be used to measure the difference between any two probability distributions, say v and w. More precisely, the expected number of extra bits required to code samples from v when using a code based on w, rather than using a code based on V can be it measured.

$$d_{KL}(v,w) = \sum_i v_i \log \frac{v_i}{w_i} \tag{4}$$

However, d_KL is neither a metric nor symmetric. The operator d_KL is also not finite. It can tend to infinity only when one of the components of W tends to zero. This would be difficult if the feature vectors contain arbitrarily small components. Consequently, the Jensen-Shannon is usually given as follows:

$$d_{JS}(v,w)=(d_{KL}(v,m)+d_{KL}(w,m))/2 \tag{5}$$

whereby m=(v+w)/2. This Jensen-Shannon divergence can sometimes be considered as the metric and finite version of the Kullback-Leibler divergence. Upon devising a feature space with a distance measure, the retrieval can be minimized to match nearby objects.

H. User Interaction

The illustrations built using our prototype ~~all~~ share the same kind of user interaction regardless of the type of data being explored. The features of search, facets, visualization, and query by example are all included. To build the interface, a conventional FS interface is adopted and extended with the exploratory capabilities. The overall sought user interaction will be focused based on [37]. The details of the interface and its customization will be discussed in the next sections.

I. Experiment Setup

The experiments are carried out using ~~on~~ a HP OMEN - 15-cb511tx, Intel® Core™ i7-7700HQ (2.8 GHz base frequency, up to 3.8 GHz) laptop with 8 GB DDR4-2400 SDRAM memory. The prototype system is developed based on an open source search server - Sphinx 2.2.10 and programmed in C++ on the Linux platform (Ubuntu 16.04 LTS) supported by Python 3.5.2 programming language. On average, the system can fetch a result page in 0.02 seconds and the average time for presenting results of one page is less than 0.011 seconds. Even for the most complicated cases (deep searching with the same results more than ones), **we are of** the opinion that ten result pages can be enough for the testing[38, 39].

III. Results and Discussion

The test and evaluation of the proposed prototype are described to evaluate the searching capabilities of the search engines. The next section describes the details of the processes involved.

A. Evaluating the Searching Capabilities of SEs

Before we proceed with the details of the experiments, the settings and the processes involved are described. These include the outlines of the main characteristics of the dataset and queries, and the preparation steps. Finally, the discussion of the results related to the IMDb Datasets and Queries and the evaluation of results are presented.

B. Datasets and Queries

The test dataset is contained in the part of IMDb. The current framework for searching data is the most widely used for essential arguments in defining query reply as tuples. The searching framework for this dataset can be structured similarly to the IMDb. For example, a movie listing may contain cast documents such as, the actors and characters, through which each actor acts. The actor's page can display all films in which the identified actor has presented along with the corresponding characters. Meanwhile, when presenting a character's record, one can see all actors who have ever acted the character and the films in which the actor shows. In an effort of providing the most uniform evaluation with old work, the queries are often adapted from the evaluation of the SPARK system [40]. The database used by SPARK differs with that which contents film genres; queries specifying a particular genre were replaced with a query specifying a particular cast role (e.g., actor, actress, director). Furthermore, several more new queries attempt to replace the original queries from SPARK evaluation since the related outputs must not be identified; the original documents selected was unclear from the keyword or the database lacked tuples containing the selected queries.

For each keyword, the documents required (mostly generic, i.e., documents about a specific actor) was identified. The query outputs can be judged as related only if they can have addressed the documents required. In all, 8 of the 22 original keywords were formulated. The keywords are presented in Table I.

TABLE I
Luo, et al. [40] Queries

No.	Modified Queries
Q1	Bryan Cranston Breaking Bad
Q2	Keanu Reeves Toy Story
Q3	Wachowski Trinity
Q4	Lucy Truth or Dare
Q5	Adam Sandler
Q6	Pulp Fiction director
Q7	2019 Thompson
Q8	Fight Club
Q9	Hanks Forrest Gump
Q10	Castle in the Sky

Kumar and Pavithra [39] conducted an in-depth analysis when comparing the searching capabilities of two SEs (Yahoo and Google) and two Metasearch SEs (Metcrawler and Dogpile). This was done based on the precision value and return recall. Fifteen queries representing a broad range of library and documents science topics were chosen, through which every query was submitted to the SEs.

We conducted the search for our experiments based on the method used in Kumar and Pavithra [39] by comparing the searching capabilities of search engines. These comparisons were done on the precision value, as well as the relative recall. Twelve queries representing a broad range of IMDb queries (nine were adapted from SPARK and three new ones) were used. Each of these queries was submitted to the SEs. The first 100 outputs obtained in each scenario were evaluated. Following this, every query was executed in the two SEs at nearly the same time to avoid temporal variations. This can be used to retrieve related data from each SEs.

When conducting searches in response to their queries, the user cannot always retrieve the related documents. The quality of searching the right documents accurately is simply identified as the precision value of the SE[41]. In the current study, the search outputs retrieved by the SEs are grouped into 5 categories as depicted in Table II on the basis of the criteria by Kumar and Pavithra [39].

TABLE II
The Search Engine Results Criteria[39]

No	**Category**	**Criteria**
1	more relevant	If the content of the web page closely matched the subject matter of the search query, then it was categorized as “more relevant” and it was given a score of 2.
2	less relevant	If the content of the web page is not closely related to the subject matter, but consists of some relevant aspects to the subject matter of the search query, then it was categorized as “less relevant” and it was given a score of 1
3	irrelevant	If the content of the web page is not related

		to the subject matter of the search query, then it was categorized as "irrelevant" and it was given a score of 0
4	links	If the content of the web page consisted of a whole series of links, rather than the information required, then it was categorized as "links" and it was given a score of 0.5, if inspection of one or two of the links proved to be useful.
5	Site can't be accessed	If the site can't be accessed for a particular URL, then the page was checked later. If this message repeatedly occurred, then the page was categorized as "site can't be accessed" and it was given a score of 0.

The scores for each site retrieved by the search engine is given manually based on the categories.

Using these criteria can allow the calculation of the precision, as well as the relative recall of SEs for every queries using the following formula[39]:

$$Precision = \frac{Sum\ of\ the\ scores\ of\ sites\ retrieved\ by\ SE}{Total\ number\ of\ sites\ retrieved} \quad (6)$$

$$Relative\ Recall = \frac{Sum\ of\ sites\ retrieved\ by\ the\ two\ SE}{Total\ number\ of\ sites\ retrieved\ by\ SE} \quad (7)$$

C. Precision of IMDb Database Search Engine

When searching for the 12 queries, not all of them can give results. The total sites that returned the results were 429, out of which only a total of 160 sites were selected for comparisons with those reported by Kumar [38]. Majority of the selected query results were from query 12 because it returned the most. Table 4 illustrates the relevant statistics for the 12 queries and the selected sites. It is clear from the table that 31% of sites are more related (see category 1) and 7.5% of sites are less relevant (category 2). Summing these two categories, the total percentage of relevant sites is 38%, which is almost half of the queries. Meanwhile, only 18% are irrelevant. The mean precision of the search is found to be 0.1.

From the table 3, the results of query with respect to the search precision for the different queries (query Q1 to query Q12) can be analyzed and summarized. It is important to note that the aim of the searching was only on the exact query matching with relevant sites or irrelevant sites. For queries Q1, Q3, Q4, Q8, Q9 and Q11, no site was returned because the query keywords did not match with the IMDb database for the selected queries (exact keywords). This result in the Precision was calculated to be zero.

TABLE III
PRECISION OF IMDB DATABASE SEARCH

			Categories					
Search Queries	Total no. of sites	Selected Sites	1	2	3	4	5	Precision
Q1	0	0	0	0	0	0	0	0
Q2	4	4	1	0	1	0	2	1
Q3	10	10	0	0	8	0	2	0
Q4	2	2	0	0	2	0	0	0
Q5	10	10	3	0	7	0	0	0.6
Q6	9	9	3	0	6	0	0	0.67
Q7	10	10	1	0	9	0	0	0.2
Q8	0	0	0	0	0	0	0	0
Q9	0	0	0	0	0	0	0	0
Q10	15	15	0	0	3	7	5	0
Q11	0	0	0	0	0	0	0	0
Q12	369	100	43	12	15	23	7	1.44
Total	429	160	51	12	61	30	16	0.9
% Total			31%	7.5%	38%	18%	1%	

For query Q7, all 10 returned sites were selected, one of which was only from the more relevant category (Category 1), while the remaining nine were from irrelevant sites (Category 3). Consequently, the Precision has been calculated as 0.2.

For query Q10, similar to Query 9, all the returned sites were selected, three of which were from the irrelevant category (Category 3), while seven were from the Links (Category 4). These returned sites are web pages consisting a whole series of links, based on the scores retrieved for each site, no matching for the selected queries. This result in the Precision was calculated as zero. For query Q12, 100 sites were selected from the total of 369 sites. When combining the results of more relevant (Category 1) and less relevant (Category 2), we obtained 55. Meanwhile, the combined results of irrelevant (Category 3) and that of Links sites (Category 4) were 30. Query Q12 has more returned results and highest Precision score of 0.9 when compared with the other queries

D. Precision of the proposed Search Engine

As in the previous two experiments, 12 queries were done and not all the queries returned results. The total site returned the results was 525, through which only 133 sites were selected for comparison with the study by Kumar [38]. Majority of the selected query results were from query 12 as it returned the most. Table IV illustrates the relevant statistics for the 12 queries and the selected sites. It is clear from the table that almost 73% of the sites are relevant (categories 1 and 2), while less than 30% are irrelevant. This means that the mean precision is 1.93.

Table IV shows that the results of queries with respect to their search precisions (for query Q1 through query Q12) can be analyzed and summarized. The prototype has a better

result than IMDb since the former adopted additional features, as well as item based search. This helps to obtain more accurate matching and more similarity results. Some analyses are given below: For queries (Q1, Q2, Q3, Q4 and Q11), the results indicate a higher precision as shown in SERP since the order of queries match with facet and item based search. The matching results retrieved all relevant sites that are retrieved in IMDb sets. Therefore, the precision for those queries are 1.

TABLE IV
PRECISION OF IMDB DATABASE SEARCH

			Categories					
Search Queries	Total no. of sites	Selected Sites	1	2	3	4	5	Precision
Q1	2	2	1	0	1	0	0	1
Q2	6	2	1	4	1	0	0	1
Q3	3	1	1	0	0	0	0	1
Q4	2	2	1	0	1	0	0	1
Q5	316	10	3	3	4	0	0	0.9
Q6	39	9	9	0	0	0	0	2
Q7	1	1	1	0	0	0	0	2
Q8	1	1	1	0	0	0	0	2
Q9	1	1	1	0	0	0	0	2
Q10	3	3	3	0	0	0	0	2
Q11	1	1	1	0	0	0	0	1
Q12	150	100	65	5	30	0	0	1.35
Total	525	133	88	9	40	0	0	1.39
% Total			66.16%	6.7%	30%			

For queries (Q5 and Q12), the total selected site for Q5 and Q12 was 10 and 100, respectively. The matching shown on SERP revels that the results of queries with respect to the Precision were high. The most relevant sites were in Category 1, while the slightly less relevant ones were given in Category 2. Both categories retrieved more than 100 sites. The SERP showed the total of irrelevant sites in Category 3 as 30.

For queries (Q6, Q7, Q8, Q9 and Q10), the results of queries with respect to SERP showed a higher precision of matching and similarities with features facet and item based search. The sites selected the same SERP matching query results, and the results retrieved all the relevant sites in IMDb sets. This was more than the precision for those queries in category 2. These results are 66% accurate. This is a better result when compared to IMDb, which is 31%.

The comparative precision of IMDb and prototype are shown in Figure 4. These results were plotted into a Search Queries and precision graph. The graph was interpolated at 12 points to show the precision value of the retrieval performance capabilities at each Query point in the searched results. Figure 4 summarizes the effectiveness of all systems and the algorithms used per graph. Item Based search and IMDb outperformed the other algorithms. The system IMDb had precision values started below -0.1. Few precision points reach zeros, and then slightly increase. This shows that there were no relevant data for the selected query, or that they have different results, which were not related to the selected query. Results of prototype, when compared to IMDb, have precision values above 0.8 or even more at every Query point.

This shows that the covering of the prototype was higher than other systems. For example, at query No. 10, Figure 4 shows that the high precision reached was precision 2, which is less than that of prototype for the same query. Prototype has retrieved more relevant data when compared with others systems. The prototype, also, has higher precision improvement with selected query. The system shows an improvement the precision. Furthermore, prototype performance curves capabilities are shown in Figure 4.

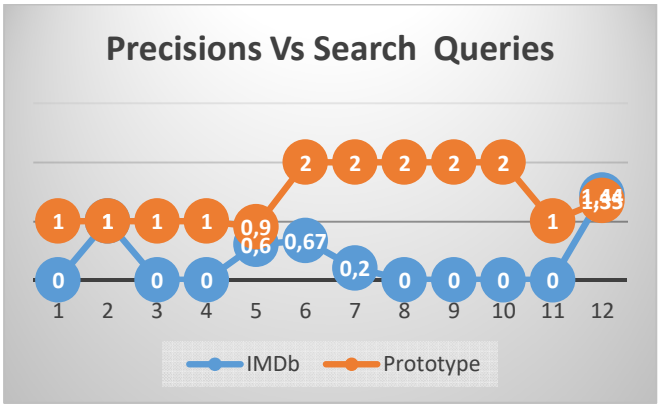

Fig. 4 Precision of IMDb and Prototype

E. Relative Recall of Search Engines

The term "recall" is a measure indicating how a specific item is retrieved, or the degree at which the returned of wanted items happens. Therefore, Recall is the ability of a related search to obtain all or most of the related information in the collection. The related recall can be obtained by applying the following formula equation 6.

F. Relative Recall of IMDb and Proposed Search Engines

Using the formula equation 7, the relative recall is calculated and presented in Table V.

It can be observed from the above table that the Total Relative recall of the IMDb is 2.83, and prototype is 9.12. Considering the prototype, search queries have highest recall value (1) followed by a query 12 (0.28), while the least recall is for query 2 (0.75). From the table above, the results of the queries with respect to the search precision for the different queries (query Q1 through query Q12) can be

analyzed and summarized. The prototype has a better result compared to IMDb since it adopted additional features as well as item based search to achieve more accurate matching and more similarity results. Further analysis is given as follows: For queries (Q1, Q3, Q4, Q8, Q9 and Q11), when compared to SERP, the results of queries showed a higher relative recall of queries. Meanwhile, the prototype showed the exact relative recall of other search engines and the matching results that retrieved all relevant sites that have in IMDb sets. For queries (Q2, Q7, Q10 and Q12), the results of the selected queries showed that the prototype retrieved the only sites that were saved in IMDb. For queries (Q5 and Q6), the results of both queries compared to SERP showed the highest recall from the total number of sites. The relative recall for Q5 is 0.9, taken from the total number of sites 316, compared to other SEs.

TABLE V
RELATIVE RECALL OF IMDB AND PROPOSED SE

	IMDb		Prototype	
Search Queries	Total no. of sites	Relative Recall	Total no. of sites	Relative Recall
Q1	0	0	2	1
Q2	2	0.25	6	0.75
Q3	0	0	3	1
Q4	0	0	2	1
Q5	10	0.03	316	0.96
Q6	9	0.9	39	0.81
Q7	10	0.18	1	0.09
Q8	0	0	1	1
Q9	0	0	1	1
Q10	10	0.76	3	0.23
Q11	0	0	1	1
Q12	369	0.71	150	0.28
Total	410	2.83	525	9.12

The Relative Recall of IMDb and prototype are shown in Figure 5. The results were plotted into a Search Queries and Relative Recall graph. This graph was interpolated at 12 points to show the precision value of the Relative Recall performances capabilities at each Query point in the searched results.

Figure 5, also, summarizes the effectiveness of all systems and the algorithms used per graph. Item Based search and IMDb outperformed the other algorithms. The Relative Recall of IMDb values started below 0.1 because this search engine gave more irrelevant results. The Relative Recall of prototype then slightly increased to cover all retrieval relevant data for the selected query. Query points in the search results are between the two curves. The prototype returned relevant results matching the selected query. Prototype has the highest Recall improvement with each query and returned a more relevant data.

A. *Conclusion the Searching Capabilities of Prototype*

Today, millions of users' access information ranging from keeping them updated with the latest news to searching on various topics of their interests. This makes the search engines the most effective searching tools [42]. Despite the fact that enormous volume of information can be retrieved using search engines at impressive speed, most results retrieved from these SEs may not be relevant[43]. Addressing this issue, results obtained in this study reported that the prototype engine can retrieve more relevant information on the WWW. Though the smaller number of sites was retrieved by the IMDb for all queries, the mean precision of prototype engines is relatively higher when compared to the other SEs. This clearly present how the prototype fails to achieve higher precision as the IMDb does.

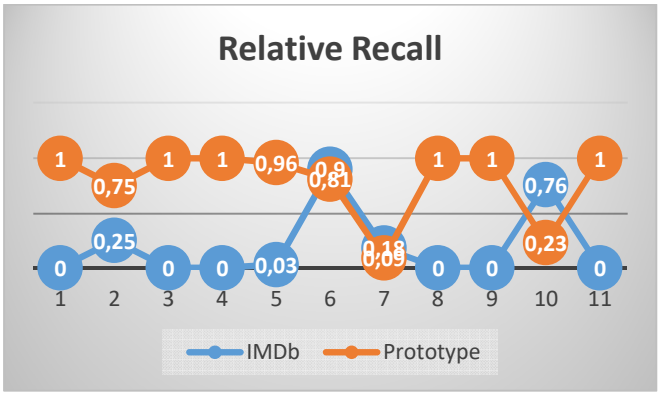

Fig. 5 Relative Recall of IMDb and Prototype

The results in above Tables show how the prototype SE have improved in term of recall in all queries when compared with the proposed approaches. However, our proposed approach in prototype gives better precision values, which subsequently result in high mean precision.

IV. CONCLUSION

In recent years, how to improve the relevancy of SE results is of increasing interest. This is associated with the heterogeneity, large volumes, and varied structures nature of the Web. Therefore, finding best results that can suit the needs of every individual searcher is challenging. Different approaches that can increase the capabilities of Web to handle large number of results have been proposed. Two of such approaches are interactive graphical and visualization methods. These not only increase the ability of the display in handling large results, but also present several attributes for

every page. Moreover, the SE can control the query reformulation and reconstruction itself.

This paper presents the testing and evaluation of the proposed prototype SE, through which the searching capabilities of search engines were reported. When confronting the challenges of handling large number of searches, the results obtained by precision and relative recall have present the ability to resolve the search engine problems. For the user to access more relevant searched documents, the results provided by the prototype SE system shows their ability to resolve the ambiguity of SE results. In order to evaluate the issues of further verification, the outputs were further compared with those obtained by other SEs. The outputs obtained by this study reported that the prototype search system is more efficient, reliable and accurate. This is because it has the capacity of storing all documents in a system when compared with the benchmark obtained by other approaches.

ACKNOWLEDGEMENT

This research was sponsored and supported under the Universiti Tenaga Nasional (UNITEN) internal grant no J510050783 (2018). Many thanks to the Innovation & Research Management Center (iRMC), UNITEN who provided their assistance and expertise during the research.

REFERENCES

[1] M. d. Kunder, "World Wide Web Size. ," https://www.worldwidewebsize.com/, 2016.
[2] J. Curran, N. Fenton, and D. Freedman, Misunderstanding the internet: Routledge, 2016.
[3] A. Selcuk, C. Ã–rencik, and E. Savas, "Private search over big data leveraging distributed file system and parallel processing," 2015.
[4] S. M. Y. Esbitan, "A Personalized Context-Dependent Web Search Engine Using Word Net (Sama Search Engine)," A Personalized Context-Dependent Web Search Engine Using Word Net (Sama Search Engine), 2012.
[5] H. Wachsmuth, M. Potthast, K. Al Khatib, Y. Ajjour, J. Puschmann, J. Qu, et al., "Building an argument search engine for the web," in Proceedings of the 4th Workshop on Argument Mining, 2017, pp. 49-59.
[6] M. N. Mahdi, A. R. Ahmad, and R. Ismail, "A Real Time Visual Exploratory Search Engine for Information Retrieval in a Cloud," International Journal of Future Computer and Communication, vol. 4, p. 216, 2015.
[7] K. Collins-Thompson, S. Y. Rieh, C. C. Haynes, and R. Syed, "Assessing learning outcomes in web search: A comparison of tasks and query strategies," in Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval, 2016, pp. 163-172.
[8] M. Dörk, "Visualization for Search: Exploring Complex and Dynamic Information Spaces," Doctoral dissertation, University of Calgary, 2012.
[9] G. Marchionini, "Exploratory search: from finding to understanding," Communications of the ACM, vol. 49, pp. 41-46, 2006.
[10] D. Sonntag and H.-J. Profitlich, "An architecture of open-source tools to combine textual information extraction, faceted search and information visualisation," Artificial intelligence in medicine, vol. 93, pp. 13-28, 2019.
[11] M. Breeding, "We need to go beyond Web 2.0," Computers in libraries, vol. 27, pp. 22-25, 2007.
[12] E. Gorelik, "Cloud computing models," Doctoral dissertation, Doctoral dissertation, Massachusetts Institute of Technology, 2013.
[13] B. R. Prasad and S. Agarwal, "Comparative Study of Big Data Computing and Storage Tools: A Review," International Journal of Database Theory and Application, vol. 9, pp. 45-66, 2016.
[14] H. Al-Aqrabi, L. Liu, R. Hill, L. Cui, and J. Li, "Faceted Search in Business Intelligence on the Cloud," in Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing, 2013, pp. 842-849.
[15] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: an open source software for exploring and manipulating networks," in Third international AAAI conference on weblogs and social media, 2009.
[16] M. Card, Readings in information visualization: using vision to think: Morgan Kaufmann, 1999.
[17] R. Spence, Information visualization vol. 1: Springer, 2001.
[18] S. Hadlak, H. Schumann, and H.-J. Schulz, "A survey of multi-faceted graph visualization," in Eurographics Conference on Visualization (EuroVis), 2015, pp. 1-20.
[19] W. Dakka, R. Dayal, and P. G. Ipeirotis, "Automatic discovery of useful facet terms," in SIGIR Faceted Search Workshop, 2006, pp. 18-22.
[20] W. Kong, "Extending Faceted Search to the Open-Domain Web," University of Massachusetts Libraries, 2016.
[21] M. N. Mahdi, A. R. Ahmad, and R. Ismail, "Paradigm Extension of Faceted Search Techniques A Review," Journal of Telecommunication, Electronic and Computer Engineering (JTEC), vol. 9, pp. 149-153, 2017.
[22] R. W. White and R. A. Roth, "Exploratory search: beyond the query-response paradigm (Synthesis lectures on information concepts, retrieval & services)," Morgan and Claypool Publishers, vol. 3, 2009.
[23] W.-C. Lin, S.-W. Ke, and C.-F. Tsai, "SAFQuery: a simple and flexible advanced Web search interface," The Electronic Library, vol. 34, pp. 155-168, 2016.
[24] C. Costa and M. Y. Santos, "Big Data: State-of-the-art concepts, techniques, technologies, modeling approaches and research challenges," IAENG International Journal of Computer Science, vol. 43, pp. 285-301, 2017.
[25] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma, "Query expansion for short queries by mining user logs," IEEE Trans. Knowl. Data Eng, vol. 15, pp. 829-839, 2002.
[26] J. Teevan, E. Adar, R. Jones, and M. A. Potts, "Information re-retrieval: repeat queries in Yahoo's logs," in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 2007, pp. 151-158.
[27] R. W. White and S. M. Drucker, "Investigating behavioral variability in web search," in Proceedings of the 16th international conference on World Wide Web, 2007, pp. 21-30.
[28] N. Dalum Hansen, K. Mølbak, I. J. Cox, and C. Lioma, "Seasonal Web Search Query Selection for Influenza-Like Illness (ILI) Estimation," in Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017, pp. 1197-1200.
[29] C. Ahlberg, C. Williamson, and B. Shneiderman, "Dynamic queries for information exploration: An implementation and evaluation," in Proceedings of the SIGCHI conference on Human factors in computing systems, 1992, pp. 619-626.
[30] D. R. Harris, "Modeling Integration and Reuse of Heterogeneous Terminologies in Faceted Browsing Systems," in Information Reuse and Integration (IRI), 2016 IEEE 17th International Conference on, 2016, pp. 58-66.
[31] C. Di Sciascio, P. Brusilovsky, and E. Veas, "A Study on User-Controllable Social Exploratory Search," in 23rd International Conference on Intelligent User Interfaces, 2018, pp. 353-364.
[32] J.-W. Ahn and P. Brusilovsky, "Adaptive visualization for exploratory information retrieval," Information Processing & Management, vol. 49, pp. 1139-1164, 2013.
[33] R. Beecham, C. Rooney, S. Meier, J. Dykes, A. Slingsby, C. Turkay, et al., "Faceted Views of Varying Emphasis (FaVVEs): a framework for visualising multi-perspective small multiples," in Computer Graphics Forum, 2016, pp. 241-249.
[34] V. T. Lee, A. Mazumdar, C. C. del Mundo, A. Alaghi, L. Ceze, and M. Oskin, "POSTER: Application-Driven Near-Data Processing for Similarity Search," in Parallel Architectures and Compilation Techniques (PACT), 2017 26th International Conference on, 2017, pp. 132-133.
[35] L. Chen, Y. Gao, X. Li, C. S. Jensen, and G. Chen, "Efficient Metric Indexing for Similarity Search and Similarity Joins," IEEE Transactions on Knowledge and Data Engineering, vol. 29, pp. 556-571, 2017.
[36] A. Jain, N. Lupfer, Y. Qu, R. Linder, A. Kerne, and S. M. Smith, "Evaluating TweetBubble with ideation metrics of exploratory browsing," in Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition, 2015, pp. 53-62.

[37] H. C. L. Hsieh and N. C. Cheng, "A Theoretical Model for the Design of Aesthetic Interaction," in International Conference on Human-Computer Interaction, 2016, pp. 178-187.
[38] G. Kumar, "Top 10 search Engines List Learn more about them," 2016.
[39] B. Kumar and S. Pavithra, "Evaluating the searching capabilities of search engines and metasearch engines: A comparative study," 2010.
[40] Y. Luo, W. Wang, X. Lin, X. Zhou, J. Wang, and K. Li, "Spark2: Top-k keyword query in relational databases," IEEE Transactions on Knowledge and Data Engineering, vol. 23, pp. 1763-1780, 2011.
[41] T. A. Usmani, D. Pant, and A. K. Bhatt, "A comparative study of google and bing search engines in context of precision and relative recall parameter," International Journal on Computer Science and Engineering, vol. 4, p. 21, 2012.
[42] J. Uddin, S. M. Ahmad, S. U. Jan, and A. Reba, "Precision and Relative Recall of Search Engines using Education Keywords: A Comparative study of Google, Yahoo and Refseek," PUTAJ-Humanities and Social Sciences, vol. 25, pp. 99-112, 2017.
[43] C. L. Smith, J. Gwizdka, and H. Feild, "Exploring the Use of Query Auto Completion: Search Behavior and Query Entry Profiles," in Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval, 2016, pp. 101-110.