International Journal of Information Science and Management
Vol. 17, No. 1, 2019, 1-16

Rethinking the Recall Measure in Appraising Information Retrieval Systems and Providing a New Measure by Using Persian Search Engines

Mohsen Nowkarizi
Associate Prof. Department of Knowledge & Information Sciences, Faculty of Education Sciences & Psychology, Ferdowsi University of Mashhad, Mashhad, Iran
Corresponding Author: mnowkarizi@um.ac.ir

Mahdi Zeynali Tazehkandi
M.A. Department of Knowledge & Information Sciences, Faculty of Education Sciences & Psychology, Ferdowsi University of Mashhad, Mashhad, Iran

Abstract

The aim of the study was to improve Persian search engines' retrieval performance by using the new measure. In this regard, consulting three experts from the Department of Knowledge and Information Science (KIS) at Ferdowsi University of Mashhad, 192 FUM students of different degrees from different fields of study, both male and female, were asked to conduct the search based on 32 simulated work tasks (SWT) on the selected search engines and report the results by citing the related URLs. The Findings indicated that to measure recall, one does not focus on how documents are selecting, but the retrieval of related documents that are indexed in the information retrieval system database is considered While to measure comprehensiveness, in addition to considering the related documents' retrieval in the system's database, the performance of the documents selecting on the web (performance of crawler) was also measured. At the practical level, there was no strong correlation between the two measures (recall and comprehensiveness) however, these two measure different features. Also, the test of repeated measures design showed that with the change of the measure from recall to comprehensiveness, the search engine's performance score is varied. Finally, it can be said, if the study purpose of the search engines evaluation is to assess the indexer program performance, the recall use will be suggested while, if its purpose is to appraise the search engines to determine which one retrieves the most relevant documents, the comprehensiveness use will be proposed.

Keywords: Recall, Comprehensiveness, Evaluation of Information Retrieval Systems, Search Engines.

Introduction

For the past years, the evaluation of the information retrieval system performance has been an important issue in information retrieval studies. Hence, in the context of information retrieval literature, more attention has been paid to determine the evaluation measures. More than 130 measures are used to evaluate information retrieval in the studies. According to Buckley and Voorhees (2005), Bama, Ahmed and Saravanan (2015) each of these measures

assesses different characteristics of information retrieval systems such as search engines. In this regard, Soleimani (2009) and Biranvand (2012), consider the search engine structure include spider, crawler, indexer, database, and Ranker. On the other hand, Davarpanah (2008) and Kousha (2003) believe that it has three main components including spider, database, or index repository, and search program or interface. While, according to Henzinger (2007), a search engine has two components: an offline component that gathers web pages and creates an in-house representation of them which is called inverted file, and the online component that meets user requests and is responsible for finding relevant documents and sorting them. According to Croft, Metzler and Strohman (2015) a search engine consists of crawler, indexer, database, user interface, ranker and evaluator. Although there is no consensus on the architecture of the search engines, the basic search engine components that are also represented in most resources, generally include three main parts that may be expanded to eight components (Aqdasi almdari, Pormanaf, AbdulJabarpourniyavar, 2015). Finally, it can be said that these three main components include a robot or crawler, an indexing program or database, and a search program. A search engine creates a database based on the data that its crawler has reported. In this section, the information retrieved by crawlers is first indexed using indexing program in the form of various criteria and parts, then stored in a repository or database (Davarpanah, 2008). One of the major differences between search engines is the different methods of indexing process in their databases (Montazer, 2005), because these indexes and databases are the basis of the search engines practice in ranking results and combining logically the words to retrieve information on the Internet. When the users submitted the queries to the search engine, the database is searched and all of the pages that are relevant to the user's request are identified. Then, based on the ranking rules and algorithms, the retrieved results are ranked according to the user's request, and the documents are displayed to the user based on their relevance (Davarpanah, 2008).

As mentioned, more than 130 measures are used to evaluate information retrieval in the studies. According to Buckley and Voorhees (2005) in different evaluation measures, various features are given about what are the user criteria, how they are interpreted, their values and their strengths in evaluating the results. In other words, the ontological assumptions determine the epistemological assumptions; these ones create in turn the methodological requirements, and finally, a technique is offered which matches these methodological requirements. Recall measure is one of the most important measures of information retrieval system. Recall is the number of relevant documents that are actually retrieved from the whole collection of relevant documents in the file (Pao, 2008; Clark & Willett, 1997). In other words, the ability of information retrieval system to find relevant documents that exists in the database (Yilmaz, Carterette & Kanoulas, 2012). According to Saracevic (2015) in most studies on information retrieval assessment, it can be seen that recall is considered as a measure to evaluate information retrieval systems. On the other hand, recall can't be ignored. To measure it, one do not usually take into account the crawling performance of the search engines. In this regard Clarke and Willett (1997) say that assume that a query is searched using two engines, A and B, and that these searches retrieve a and b relevant documents, respectively. Assume further

that there is no overlap in these two sets of documents so that the total pool contains a+b relevants. We do not conclude that the recall of A and B are a/ (a+b) and b/ (a+b), respectively since it may be that some of the b relevant documents retrieved by B were not available to A, and vice versa. Accordingly, we can obtain a figure for the recall performance of A only after checking how many of these b relevant documents have been retrieved by A. If there are c such relevants (c < b) then the recall for A is a/ (a+c). In other words, to measure it, researchers only pay attention to document indexing quality. However, users do not notice the quality of document indexing (e.g the pool contains a+c), what is imortant to them is to find sets of documents so that the total pool contains a+b relevant. In addition, Saracevic (2015) believes that recall was first called as relevance, and later it was assigned this term which is ambiguous. Therefore, the necessity to rethink the concept and, consequently, revise it is a vital issue that has been considered in this study. Based on this, a new measure is propose as *comprehensiveness*, and then the two measures of *comprehensiveness* and *recall* are compared using two hypotheses.

Hypothesis 1: There is a significant correlation between the two measures, recall and comprehensiveness of Parsijoo, Rismoon and Yooz search engines.

Hypothesis 2: Shifting the measure from recall to comprehensiveness causes to change a performance score of the Parsijoo, Rismoon and Yooz search engines.

In order to examine the second hypothesis, two sub-hypotheses are designed and the results of these two sub-hypotheses are compared together as follow:

First sub-hypothesis: There is a significant difference between the recall of Parsijoo, Rismoon and Yooz.

Second sub-hypothesis: There is a significant difference between the comprehensiveness of Parsijoo, Rismoon and Yooz.

Literature review

Evaluation of information retrieval systems is fundamental topics in Library and Information Science. In this regard, various researches such as Ahlgren and Grönqvist (2008), Sakai and Kando (2008) Bama, Ahmed and Saravanan (2015) have been carried out on the measurement of information retrieval systems. Recall is one of the most important measures of information retrieval system. It computes the ability of information retrieval system to find relevant documents that exists in the database (Yilmaz, Carterette & Kanoulas, 2012) which is calculated through the following formula:

$$Recall = \frac{number\ of\ relevant\ retreived\ documents}{number\ of\ relevant\ documents\ in\ database}$$

The above formula was designed for recall based on a binary classification, but one may define it for a continuum and comparatively scale as follows:

$$Recall = \frac{relevance\ score\ of\ retreived\ documents}{relevance\ score\ of\ documents\ in\ database}$$

Regarding recall, apparently a question is posed: what are the relevant documents that users are looking for: The relevant documents indexed in the search engine database or the relevant documents available on the web?

As stated, in system-oriented approach, one addresses the documents organization or organizing the documents selected by the search engine crawler. Recall also depends on the quality of organizing the documents. In other words, in the calculation of recall, the relevant documents available on the search engine database - not the relevant documents available on the web- are placed in the denominator, so the crawler function is not considered because the crawler chooses the documents on the web. In recall, however, one pays attention to the next phase of selecting the documents as only the relevant documents available in the database are addressed. In the same way, Lancaster (2003) points out that database coverage may also affect its search success or failure. Concerning some databases such as Emerald, Science Direct, etc, he adds that the database producer first chooses a set of newly published titles (journals or articles) that match the selection criteria (indexing policy). So, the coverage of the databases is different. This may clarify the impact of the theoritical foundations on the techniques .In addition, the literary structure of recall term in English and Persian supports the recent issue.

The prefix “Baz” in Persian and “Re” in English refers to the “again”. In other words, it shows doubling and repeating an action. The verb “Yaft” and “Yabesh” refers to the act of finding. In English, "call" means address, quest, want. So, in the Persian language, it means re-find and in English it means request that it has the same meaning.

In computing, recall refers to the fraction of the relevant documents retrieved from a database to meet a question (English Oxford Living Dictionary, 2018).

Concerning the search engine structure, it has been determined that a crawler first finds websites based on the search engine indexing policy (find or search or call action). Finally, indexing words and links of these documents are stored in the search engine database.

When a user enter the query into the search box, the documents retrieved by the crawler (the documents available in the search engine index or database) not the documents available on the web- are searched and the relevant documents to the user's query are presented to him (re-find or re-call action). Hence, the literary structure also confirms that in recall the crawler function is not considered. On one hand, according to Saracevic (2015) in the mid-1950s, Allen Kent and James Perry, two chemists, wrote some articles on information retrieval techniques. In one of them, they provided two measures (precision and recall) to evaluate the relevance of information retrieval systems. Since the initial approach to assess information retrieval has been a quantitative and system-oriented approach, recall has taken a rise out of system-oriented approach While system-oriented approach in the information retrieval evaluation has already been severely criticized and user-oriented approach has been suggested as an appropriate approach. Accordingly, in recent decades, various researchers have tried to propose new measures. In this regard, varias studies have been conducted on the presentation of new measures, which are referred to some of the most important ones.In one of these studies, Mea and Mizzaro (2004) have proposed a new retrieval effectiveness measure,

named *Average Distance Measure* (ADM), which simply measures the average distance—or difference—between UREs (The relevance score of the documents assigned by the users) and SREs (The relevance score of the documents assigned by the Information Retrieval System). In a more formal way, for a given query, we can define two relevance weights for each document in the database D_ADM is a score between zero and one, which zero indicates weak system performance and one shows its best performance.

One of the other studies about information retrieval measures is Järvelin & Kekäläinen research. According to Järvelin & Kekäläinen (2002), modern large retrieval environments tend to overwhelm their users by their large output. Since all documents are not of equal relevance to their users, highly relevant documents, or document components, should be identified and ranked first for presentation. This is often desirable from the user point of view. In other words, they have argued that the focus of effective measures should be on how far the search engines can retrieve more relevant documents ahead of relevant documents. So, precision and recall measures are not appropriate measures. Hence, they have introduced a new measure called *Normalized Discounted Cumulative Gain* (NDCG). For each query, the NDCG is computed as DCG /ideal DCG.

Buckley & Voorhees (2004) have defined a new measure called *Bpref*. It computes a preference of whether judged relevant documents are retrieved ahead of judged non-relevant documents. As we can see, Bpref allows us to simply look at how known relevant and non-relevant documents are ranked rather than expecting to know all the relevant documents in the collection. One problem with Bpref definition is that only the same irrelevant documents as the number of relevant document are used for the calculation. This is a problem especially where the number of known relevant document is low. Hence, Grönqvist (2005) has introduced a new measure called *Rank eff*. This measure is similar to Bpref but it does not suffer from the same weakness as Bpref: it uses all the relevance judgments, it can handle data set with any number of relevant and irrelevant documents, it handles small number of document better than Bpref. Finally, it can be said that this measure does not affect the number of documents. Some researchers have tried to analyze assessment measures, which are referred to some of the most important ones. Cooper (1968) is defined and compared measure of document retrieval system performance called the "expected search length reduction factor (ESL)" with measures of precision and recall. He concludes that this measure (ESL) provides a single figure of merit; allows for gradations of retrieval status through the concept of a weak ordering; evaluates retrieval performance relative to random searching; and takes into account the needed amount of relevant material.

Clarke and Willett (1997) have attempted to explain the conceptual and theoretical foundations of recall measure. They first mentioned the structure of the search engine, and then point out that crawler performance is not considered in recall measures. They provide a method for accurately calculating the recall measure. Then, using the described method, compare the retrieval effectiveness of the Alta Vista, Excite and Lycos Web search engines.

Bar-Ilan, Mat-Hassan, and Levene (2006) computed five measures: the overlap, Spearman's footrule, F, Fagin's G measure, and the new M measure. Reason of them for

introducing this new measure was to minimize the problems related to the other measures. They conclude the overlap ignores rankings, Spearman's footrule is based only on the relative rankings and ignores the non-overlapping elements completely, and, finally, Fagin's measure gives far too much weight to the size of the overlap. The new measure attempts to take into account both the overlapping and the non-overlapping elements, and gives higher weight to the overlapping URLs among the top-ranking results. It seems that the M measure better captures our intuition regarding the quality of rankings.

Sakai and Kando (2008) compares the robustness of IR metrics to incomplete relevance assessments, using four different sets of graded-relevance test collections with submitted runs—the TREC 2003 and 2004 robust track data and the NTCIR-6 Japanese and Chinese IR data from the crosslingual task. According to these experiments, Q0, nDCG0 and AP0 proposed by Sakai are superior to bpref proposed by Buckley and Voorhees and to Rank-Biased Precision proposed by Moffat and Zobel. They point out some weaknesses of bpref and Rank-Biased Precision by examining their formal definitions.

As noted above, some researchers have compared different measurements, and ultimately identified their weaknesses and strengths and some researchers have developed to evaluate information retrieval systems especially search engines. Nevertheless in most studies on information retrieval assessment, it can be seen that recall is considered as a standard measure to evaluate information retrieval systems While, the recall measure calculates only the performance of the indexer program. Based on this, a new measure is designed as Comprehensiveness, to pay consider the spider's performance in addition to the indexer program's one which is discussed below.

Sometimes users seek to dominate the subject and tend to access the maximum number of relevant documents available on the web (Su, 1994), this may be considered as comprehensiveness. Comprehensiveness is different from recall. However, theses two concepts are considered synonym in the information retrieval literature. The meaning of recall was explained. But the comprehensiveness meaning in dictionaries is: to cover the full and wide (Cambridge English Dictionary, 2018), to include and attend everything or nearly all elements and aspects of something (Englisch Oxford Living Dictionary, 2018), to include necessary components (Colin Dictionary, 2018) , to include most parts and aspects of something (MacMilan Dictionary, 2018). According to the definitions of dictionaries, one may conclude three points:

1. All definitions refer to the coverage of inclusion of something.
2. Some definitions refer to inclusion of all or most components.
3. Some definitions refer to the inclusion of the necessary components

Regarding the points, it is important to consider three words in the definitions: what do components, all and necessary mean?

Obviously, in the information retrieval, the focus is on relevant documents, the concept of "components" also refers to them.

So the meaning of all the components refers to all the relevant documents which are available on the web not in the search engine database. Only some of the relevant documents

are indexed in a search engine database. In the case of "necessity", it should be pointed out that no one may consider necessary some of the relevant documents and unnecessary some others. Finally, it may include all the relevant documents that are available on the web and provide the following definition or formula for it.

$$Comperhensiveness = \frac{number\ of\ relevant\ retreived\ documents}{number\ of\ relevant\ documents\ in\ web}$$

The above formula designed for comprehensiveness based on a binary classification which can be defined as follows for a continuum and comparatively scale:

$$Comperhensiveness = \frac{relevance\ score\ of\ retreived\ documents}{relevance\ score\ of\ documents\ in\ web}$$

Obviously, it is not possible to identify all the relevant documents on the web; therefore, in the comprehensiveness formula, one had to, for example, place all the relevant documents retrieved by several search engines. So it will be relative. In this way, the difference between the two measures, recall and comprehensiveness, is identified. In calculating a search engine recall ratio, the relevant documents available in a search engine database are placed in the denominator while to calculate a search engine comprehensiveness, the sum of related documents retrieved by some search engines is placed in the denominator.

Traditional test-collection experiment evaluates several different retrieval strategies by applying them to a common set of documents, and such a common database does not exist in the context of the Web. The reason for this is that the spider programs associated with different search engines adopt different criteria both for exploring the Web and for selecting the pages that should be indexed. There may well be a substantial degree of overlap between the pages indexed by two engines but the resulting databases that are searched will not be the same unless identical spider and indexing programs have been used. There is no reason to believe that this is the case, as even a cursory glance at the promotional material at the engines' home pages will demonstrate, and we must accordingly take this differential coverage into account when evaluating the recall of the searching component of a search engine (Clarke and Willett, 1997). Formally, assume that a query is searched using two engines, A and B, and that these searches retrieve a and b relevant documents, respectively. And overlap of these two sets of documents is zero. Hence the total pool contains a+ b relevant document. Since it may be that some of the b relevant documents retrieved by B were not available to A, and vice versa. Accordingly, the recall for A is a/(a+c) which c is a subset of b, While users are willing to access all the relevant documentation available in the total pool contains a+ b relevant document. Based on theoretical foundations, one pays attention to a crawler performance in the calculation of comprehensiveness while in recall, a crawler function is ignored. As explained, theoretically, recall and comprehensiveness have different meanings. The question now arises whether these two measures are actually different in practice.

Research Methodology

The fundamental of a developmental research is to implement organized studies to improve and innovate some tools for developing or promoting the quality of products, services and techniques. Since in this research, recall has been revised and a new measure entitled comprehensiveness is presented both at the theoretical (conceptual) and practical level (the formulation), this is a developmental one choosing the approach to judge the relevance of documents is important in information retrieval research. In this regard, various researchers such as Saracevic (2007), Thornley (2012), Huang and Soergel (2013) have addressed this issue. By reviewing related literatures it was found that evaluation of information retrieval systems is successful if it benefits from composite approach. Therefore, in this research, a composite (dialectical) approach was used to determine the relevance of the documents. The research design is fully explained in the follow.

In order to test the second hypothesis, Parsijoo, Rismoon and Yooz (3 Persian search engines) were selected, but to calculate comprehensiveness, an extensive list of the relevant web documents was necessary. In the way, it was necessary to use a public and exhaustive search engine to retrieve the documents not indexed in those search engines, and place them in the relevant documents.

As mentioned earlier, what is important in the comprehensiveness is the maximum amount of relevant documents not the relevant documents indexed in a search engine database. Therefore, since it is intended to compare the comprehensiveness of three Persian search engines to each other, it is necessary to search one another search engine other than the three mentioned search engines to reach the maximum number of related documents. Thus, according to Alexa's website (2016), Google is the most exhaustive search engine used in Iran. Based on various studies such as Riahinia et al (2016) and Lewandowski (2015), Google has a better performance than other public search engines. Thus, Google was used as the best search engine to access the maximum relevant documents.

Since in most of information retrieval studies (such as Wu & Li, 1999; Llic. Bessell, Silagy, & Green., 2003; Tang, Craswell, Hawking, Griffiths & Christensen, 2006; Knight, Holt and Warren, 2009; Lewandowski, 2008; Hariri, 2011) the number of judges have been less than 50 people but recently the researchers tend to use more people to judge in these situations, then in consultation with three experts from the Department of Knowledge and Information Science (KIS) at Ferdowsi University of Mashhad (FUM), a sample of 192 FUM students of different degrees engaging in different fields of study both male and female were selected through stratified random sampling. Because in some studies (Davidson, 1977; Huang, and Wang, 2004; Vakkari & Järvelin, 2005; Saracevic, 2007) these factors are shown effective in information retrieval. Two simulated work tasks (SWTs) were considered for each student (participant), and they received the SWTs along with the search instructions. The participants read each SWT and then formed an information need in their minds. In the next step, they entered the information need in the form of a query in the search box of each engine, browsed the retrieved websites and, then, recorded the URL of each website related to the SWTs in an electronic search form. Finally, they sent the completed form to the

researchers' email address. After receiving the search forms, the researchers created the relevant documents' pool. In other words, all the URLs selected by the participants were placed in the pool, and their relevancy was determined by the number of times that the same URL was chosen by them. For example, suppose that the URL N for the subject A has been selected 4 times and the URL M 16 times (maximum selection times) by them, so the relevancy of the URL N was 0.25.

In this way, the relevancy of each URL has been determined. Below, recall and comprehensiveness ratios were calculated for Parsijoo, Rismoon and Yooz:

Calculate the recall for the Parsijoo search engine

$$R = \frac{\text{relevance score of document retrieved by Parsijoo search engine in one search in subject A}}{\text{relevance score of document retrieved by Parsijoo search engine in all search in subject A}}$$

Calculate the comprehensiveness for the Parsijoo search engine:

$$C = \frac{\text{relevance score of document retrieved by Parsijoo search engine in one search in subject A}}{\text{relevance score of document retrieved by Parsijoo and Rismoon and Yooz search engine in all search in subject A}}$$

Finally, the findings yielded of the calculations in the Excel file were entered into the SPSS20 and, according to the conditions, the convenient statistical tests were used which are described in detail in the next sections. The validity of the instrument to calculate the relevance was approved through consulting the previous studies, seeking the faculty members' views, and referring related texts (in particular Saracevic, 2007, Huang and Soergel, 2013) .During the implementation phase, the search forms, SWTs, and so on were also reviewed and revised by several experts in the field of KIS.

Then, the search forms and SWTs were submitted to the KIS faculty members at the FUM. Finally, according to the received points, the necessary items were modified and finalized. To measure the reliability, six SWTs were given twice to some users over a two-week period, and they were asked to search the information need from SWTs in the search engines and record the related URLs. In the end, the two tests' correlation coefficient was calculated. Since it was 0.739, the reliability of the research tool was confirmed. One of the limitations was the lack of access to computers at all campus sites. To resolve this, the researchers had to carry out their own laptop. Also, to gather the huge collection of data required a great deal of time, thus 3 information science experts were also contributed in data gathering

Since the data was quantitative in two hypotheses, so Pearson test was used in order to measure the correlation between recall and comprehensiveness and Greenhouse-Geisser Test was used in order to measure recall of search engine and repeated measurement test to measure the differences of search engine comprehensiveness.

Findings

First hypothesis

There is a significant correlation between recall and comprehensiveness. As the data were

quantitative and Kolmogorov-Smirnov (K-S) test value was 0.42, Pearson's test was used to examine the correlation between recall and comprehensiveness. The results are presented in Table 1.

Table1

Pearson test in order to measure correlation recall and comprehensiveness

Variable	Number	Test statistic	p-value
Recall-comprehensiveness	96	0.64	0.001

Since the significance level was 0.64 and the p-value was less than 0.05(P <0.001), there was a significant correlation between them. When the correlation between two variables is between 0.4 and 0.7, there is a moderate correlation between them. On the other hand, when two measures of common factors of something are affected, there will be a strong correlation (0.7 to 1) between them and they will be categorized in one group so that they will measure some common characteristics (Baccini, Déjean,, Lafage, and Mothe, 2012). In this case, it is not necessary to calculate more than one measure in evaluating information retrieval systems. Since there was a moderate correlation between recall and comprehensiveness, although the first hypothesis was confirmed, it can be concluded that recall and comprehensiveness were different characteristics because there was not a strong correlation between them. This result is acceptable because the performance of text transformation component or indexer is considered in the calculation of the recall, while in the calculation of the comprehensive measure, the text acquisition or crawler performance is also considered (Clarke and Willett, 1997).

Second hypothesis: Shifting the measure from recall to comprehensiveness causes to change a performance score of the search engines. To answer this hypothesis, first two sub-hypotheses have been considered, which are:

First sub-hypothesis: There is a significant difference between the recall of Parsijoo, Rismoon, and Yooz.

Since the data were quantitative and their normality was approved, repeated measures test was used to determine the difference.

To determine the appropriate test, the assumption of uniformity of covariance is mandatory. Mauchly's Test of Sphericity was applied to test it. In table 2, the results are drawn.

Table2

Mauchly's Test of Sphericity in order to identify uniformity of covariance

Variable	Test statistics	df	p-value
Recall	0.29	2	0.001

As shown in Table 2, the significance level of Mauchly's Test was less than 0.05 (0.001), so the zero assumption was rejected. In other words, the data spherity were not confirmed. Hence, to identify the difference between recall of the engines, Greenhouse-Geisser Test was

used whose results are drawn in Table 3.

Table3

Greenhouse-Geisser Test in order to measure recall of search engine

Variable	mean square	Test statistics	df	p-value
Recall	0.26	2.31	1.17	0.13

As indicated in Table 3, the significance level of Greenhouse-Geisser test (0.13) was greater than 0.05, so the zero assumption was confirmed. In other words, there was no significant difference between the recall of Parsijoo, Rismoon, and Yooz. Thus, the first sub-hypothesis was not confirmed.

Second sub hypothesis: There is a significant difference between the comprehensiveness of Parsijoo, Rismoon, and Yooz. Since the data were quantitative, and the normality was confirmed (K-S sig= 0.23), Repeated Measures design was used.

To test, the uniformity of covariance, Mauchly's Test of Sphericity has been applied whose results were drawn in table 4.

Table 4

Mauchly's Test of Sphericity in order to identify the uniformity of covariance

Variable	Test statistics	df	p-value
comprehensiveness	0.941	2	0.403

As shown in Table 4, its significance level (Sig= 0.403) was greater than 0.05, so the zero assumption was confirmed.

In other words, the data sphericity has been confirmed. Therefore, to identify the comprehensiveness of the search engines, Repeated Measures Test was used whose results were presented in Table 5.

Table 5

Repeated measurement test to measure the differences of search engine comprehensiveness

Variable	mean square	Test statistic	df	p-value
Comprehensiveness	0.3	22.67	2	0.001

As indicated in Table 5, its significance level (Sig= 0.001) was less than 0.05, so the zero assumption was not confirmed. In other words, there was a significant difference between the comprehensiveness of Parsijoo, Rismoon, and Yooz. In this way, the second sub-hypothesis was confirmed. Finally, to test the second hypothesis, it can be concluded that there was no significant difference between the recall of Parsijoo, Rismoon, and Yooz. While there was a significant difference between the comprehensiveness of the mentioned search engines, then, shifting the measure from recall to comprehensiveness caused a significant difference between the search engines' performance, so the second hypothesis has been confirmed.

Discussion and Conclusion

Research methods and techniques are related to philosophical and theoretical foundations, and they follow a series of epistemological and theoretical foundations about which most researchers are unknown in many cases. However, if some techniques are applied regardless of their related theoretical foundations, it won't certainly lead to useful results, and perhaps these results will be questionable. Today, the studies on the evaluation of information retrieval systems, especially search engines, have a major contribution to information science studies and their findings are of particular importance. These findings depend on the measures they use, since each measure takes into account some specific features of the information retrieval system.

In this research, recall was re-defined and a new measure was considered as comprehensiveness, then these two measures were compared. Recall is one of the most widely used measures of information retrieval assessment that has been focused on over time in information retrieval evaluation studies. However, since it is rooted in a system-oriented approach, only the algorithms of information retrieval systems are measured. In other words, the measure focuses on how the documents are indexed and organized by the system. Hence, the recall ratio calculation does not provide users with helpful information. In the case of search engines, one may say that it is not important to users how those engines are indexing the documents. Rather, they tend to access the relevant documents which meet their information needs through the Web. Recall take into account only the relevant documents available on and indexed in the search engine database While, what does matter to the users are the relevant documents on the Web, whether indexed or not indexed by a search engine. It was necessary to rethink recall. Thus, recall was redefined regarding new approaches to the evaluation of information retrieval system, and a new measure was introduced as comprehensiveness. In addition to include the features of recall, comprehensiveness also focuses on how the documents are selected by the system, and how the crawler perform or play its role.To explain more about the difference between comprehensiveness and recall, we point out an example in the research. There were 30 relevant documents on the Web on “Etiquaf” that Rismoon indexed 2 of them at its database and provided users with 1 document while being searched on the subject. Its recall and comprehensiveness ratio were respectively 0.5 and 0.06. Parsijoo stored 10 ones on that subject, among of the 30 relevant documents, in its database and when submitting, provided the user with 4 documents, its recall ratio and comprehensiveness were respectively 0.4 and 0.13.

Now, if you intend to use recall to compare these two search engines (Rismoon, Parsijoo), the "Rismoon" will be introduced as an efficient one, and if the comparison is focused on comprehensiveness to evaluate them, “Parsijoo” will be more efficient. As “Rismoon” has retrieved 1 and “Parsijoo” 4 documents while being searched, it can be said that in fact “Parsijoo” is more efficient than " Rismoon" . In this way, it will be clear that a search engine may perform better using recall yet it is less comprehensive. Altogether, it is more helpful for users to use the search engine which will be more comprehensive. So, in comparison with recall, comprehensiveness illustrates more efficiently the performance difference of search

engines. At the practical level, in the first hypothesis, the findings indicated that there was a moderate correlation between recall and comprehensiveness. So, this conclusion is logical because both in recall and in comprehensiveness calculation, the numerator is the same and only the denominator differs.

If there is a strong correlation between the two measures, they show the same characteristic. But, in this study there was not a strong correlation between them.

Hence, each of these takes into account to some extent the specific features of the information retrieval system. By examining the second hypothesis, it was also observed that there was not a significant difference between recall ratios of each one, but their comprehensiveness ratio differed significantly. In this way, it can be said obviously that at the practical level comprehensiveness takes into account different characteristics in comparison with recall. The latter only measures the function of indexer program while the former in addition to it, considers robot or crawler. Thus the latter addresses the extent to which a search engine can provide the relevant documents available in its database to user during searching while the former considers to what extent a search engine may provide users with the relevant documents available on the indexable web through searching. Finally, it can be said that users tend to access relevant documents on the web, whether a document is indexed in the search engine database or not. Since in recall calculation of a search engine, in a fraction denominator, only the relevant documents available in its own database are considered, while in the denominator of comprehensiveness, the relevant documents available on the indexable web are taken into account, recall is of interest to system, while comprehensiveness is important to users.

The findings showed that P-value of Greenhouse-Geisser test was 0.13 in the comparison of the three native Persian (Parsijoo, Yooz and Rismon). Therefore, if in the evaluation of these engines, one use recall, there will be no difference in the performance of them, and their effectiveness will be the same. While the Greenhouse-Geisser P-value was 0.001 in evaluating these search engines using comprehensiveness. Thus, the evaluation of them using comprehensive showed that their effectiveness is different. Finally, it can be said that comprehensiveness, in comparison with recall, shows their performance differences more accurately and precisely. Hence, researchers in the field of information retrieval assessment, which explores search engines to determine the most efficient, are suggested to use comprehensiveness measure rather than recall in their studies.

Reference

Ahlgren, P., & Grönqvist, L. (2008). Evaluation of retrieval effectiveness with incomplete relevance data: Theoretical and experimental comparison of three measures. *Information Processing & Management*, 44(1), 212-225.

Alexa (2017). Google.com Traffic Statistics. Retrieved from https://www.alexa.com/siteinfo/google.com

Aqdasi Alamdari, P., Poormanaf, V. & Abdul-Jabar Pourniyavar, F. (2015). Check the performance of search engines. National Conference on Computer Engineering and

Information Technology Management. Retrieved from http://www.civilica.com/ Paper-CSITM02-CSITM02_103.html.

Baccini, A., Déjean, S., Lafage, L., & Mothe, J. (2012). How many performance measures to evaluate Information Retrieval Systems? *Knowledge and Information Systems*, *30*(3), 693-713.

Bama, S. S., Ahmed, M. I., & Saravanan, A. (2015). A survey on performance evaluation measures for information retrieval system. International Research Journal of Engineering and Technology, 2(2), 1015-1020.

Bar-Ilan, J., Mat-Hassan, M., & Levene, M. (2006). Methods for comparing rankings of search engine results. *Computer networks*, 50(10), 1448-1463.

Biranvand, A. (2012). Computer and Internet Basics. Tehran: Chapar.

Buckley, C., & Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. In Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval (pp. 25–32).

Cambridge English Dictionary (2018). Comprehensiveness. Retrieved from https://dictionary.cambridge.org/dictionary/english/comprehensive

Clarke, S. J., & Willett, P. (1997). Estimating the recall performance of Web search engines. *Aslib Proceedings*, 49 (7), 184-189.

Collin Dictionary (2018). Comprehensiveness. Retrieved from https://www.collinsdictionary .com/dictionary/english/comprehensive

Cooper, W. S. (1968). Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *American documentation*, 19(1), 30-41.

Croft, W. B., Metzler, D., & Strohman, T. (2015). *Search engines: Information retrieval in Practice*. London: Pearson Education.

Davarpanah, M. (2008). Search for scientific and research information in print and electronic resources. Tehran: Dabizesh.

Davidson, D. (1977). The effect of individual differences of cognitive style on judgments of document relevance. *Journal of the American Society for Information Science*, 28(5), 273–284.

English Oxford Living Dictionary (2018). Comprehensiveness. Retrieved from https://en.oxforddictionaries.com/definition/comprehensiveness

English Oxford Living Dictionary (2018). Recall. Retrieved from https://en.oxforddictionaries.com/definition/recall

Grönqvist, L. (2005). Evaluating latent semantic vector models with synonym tests and document retrieval. In *ELECTRA Workshop on Methodologies and Evaluation of Lexical Cohesion Techniques in Real-world Applications (Beyond Bag of Words)* (Vol. 5).

Hariri, N. (2011). Relevance ranking on Google: Are top ranked results really considered more relevant by the users? *Online Information Review*, 35(4), 598-610.

Henzinger, M. (2007). Search technologies for the Internet. *Science*, 317(5837), 468-471.

Huang, M., & Wang, H. (2004). The influence of document presentation order and number of documents judged on users' judgments of relevance. *Journal of American Society for Information Science and Technology*, 55(11), 970–979.

Huang, X., & Soergel, D. (2013). Relevance: An improved framework for explicating the notion. *Journal of the American Society for Information Science and Technology*, 64(1), 18-35.

Ilic, D., Bessell, T. L., Silagy, C. A., & Green, S. (2003). Specialized medical search-engines are no better than general search-engines in sourcing consumer information about androgen deficiency. *Human Reproduction*, 18(3), 557-561.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4), 422-446.

Knight, D., Holt, A., & Warren, J. (2009). Search engines: a study of nine search engines in four categories. *Journal of Health Informatics in Developing Countries*, 3(1), 1-8.

Kousha, K. (2003). Internet search Tools: Principles, Skills and Web Search Options. Tehran: Ketabdar.

Lancaster, F. W. (2003). Indexing and abstracting in theory and practice. Translated by Abbas Gilvari. Tehran: Chapar.

Lewandowski, D. (2008). The retrieval effectiveness of web search engines: considering results descriptions. *Journal of Documentation*, 64(6), 915-937.

Lewandowski, D. (2015). Evaluating the retrieval effectiveness of Web search engines using a representative query sample. *Journal of the Association for Information Science and Technology*, 66(9), 1763-1775.

Macmillan Dictionary (2018). Comprehensiveness. Retrieved from https://www.macmillandictionary.com/dictionary/british/comprehensive

Mea, V. D., & Mizzaro, S. (2004). Measuring retrieval effectiveness: A new proposal and a first experimental validation. *Journal of the American Society for Information Science and Technology*, *55*(6), 530-543.

Meriam webester dictionary (2018). Comprehensiveness. Retrieved from https://www.merriam-webster.com/dictionary/ Comprehensiveness.

Montazer, G. A. (2005). *Internet search engines: Income on optimal information retrieval*. Tehran: Kavir [In Persian].

Pao, M. L (2000). Concepts of information retrieval. Translated by Asdollah Azad and Ramatollah Fattahi. Mashhad: Ferdowsi University of Mashhad.

Riahinia, N, Bakshyan, L, Latifi, M., & Rahimi, F. (2016). Evaluation the accuracy and recall in general search engines, based on the system relevance and search logic. *Journal of Academic Librarianship and Information Research*, 50(1), 3-24. [In Persian]

Sakai, T., & Kando, N. (2008). On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Information Retrieval*, 11(5), 447-470.

Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in Information Science. Part II: Nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology*, 58(13):1915–

1933.

Saracevic, T. (2015). Why is relevance still the basic notion in Information Science? In: F. Pehar, C. Schlögl, C. Wolff (Eds.) *Reinventing Information Science in the Networked Society* (pp. 26-36), *Proceedings of the 14th International Symposium on Information Science* (ISI 2015), ZadarCroatia, 19th– 21st May 2015. Retrieved Dec. 1, 2017 from https://zenodo.org/record/17964/files/keynote2.pdf

Soleimani, H. (2009). Learning to search the web for databases. Tehran: Soleimani. [In Persian]

Su, L. T. (1994). The relevance of recall and precision in user evaluation. *Journal of the American Society for Information Science*, *45*(3), 207-217.

Tang, T. T., Craswell, N., Hawking, D., Griffiths, K., & Christensen, H. (2006). Quality and relevance of domain-specific search: A case study in mental health. *Information Retrieval*, 9(2), 207-225.

Thornley, C. (2012). Information retrieval (IR) and the paradox of change: An analysis using the philosophy of Parmenides. *Journal of Documentation*, 68(3), 402-422.

Vakkari, P., & Järvelin, K. (2005). Explanation in information seeking and retrieval. *New directions in cognitive information retrieval.* Dordrecht, Netherland: Springer

Wu, G., & Li, J. (1999). Comparing Web search engine performance in searching consumer health information: evaluation and recommendations. Bulletin of the Medical Library Association, 87(4), 456-461.

Yilmaz, E., Carterette, B., & Kanoulas, E. (2012). Evaluating Web Retrieval Effectiveness. In Dirk lewandowski , *Web search engine research*. Bingley, west Yorkshire: Emerald Group Publishing.