Evaluating the effectiveness of Google, Parsijoo, Rismoon, and Yooz to retrieve Persian documents

Mahdi Zeynali Tazehkandi and Mohsen Nowkarizi
Department of Knowledge and Information Science, Ferdowsi University of Mashhad, Mashhad, Iran

Received 12 November 2019
Revised 7 February 2020
Accepted 8 February 2020

Abstract

Purpose – The purpose was to evaluate the effectiveness of Google (as an international search engine) as well as of Parsijoo, Rismoon, and Yooz (as Persian search engines).

Design/methodology/approach – In this research, Google search engine as an international search engine, and three local ones, Parsijoo, Rismoon, and Yooz, were selected for evaluation. Likewise, 32 subject headings were selected from the Persian Subject Headings List, and then simulated work tasks were assigned based on them. A total of 192 students from Ferdowsi University of Mashhad were asked to search for the information needed for simulated work tasks in the selected search engines, and then to copy the relevant website URLs in the search form.

Findings – The findings indicated that Google, Parsijoo, Rismoon, and Yooz had a significant difference in the precision, recall, and normalized discounted cumulative gain. There was also a significant difference in the effectiveness (average of precision, recall, and NDCG) of these four search engines in the retrieval of the Persian resources.

Practical implications – Users using an efficient search engine will attain more relevant documents, and Google search engine was more efficient in retrieving the Persian resources. It is recommended to use Google as it has a more efficient search.

Originality/value – In this research, for the first time, Google has been compared with local Persian search engines considering the new approach (simulated work tasks).

Keywords Evaluation, Information retrieval, Effectiveness, Search engines, Google, Online retrieval

Paper type Research paper

Introduction

Traditionally, libraries and information centers were considered to be the only places where information resources were available to meet the information needs of the users. With the advent of modern information and communication technologies (ICTs), especially the World Wide Web, significant changes occurred in the production, distribution, dissemination, and access to the information resources, and the Web became one of the most important sources of information. The number of users and the amount of information presented through the Web are tremendously increasing; according to Lawrence and Giles (1999), in December 1997, there were 800 million pages, while the indexable Web in 2019 is estimated at more than six billion pages (Kunder, 2019). This indicates that, from these Web pages, not anybody can get one's information needs; however, everybody requires tools to help him/her attain the most relevant Web pages and fulfill one's information needs. That is why, shortly after the invention of the Web, search tools were designed (Anderson, 2006; Poulter, 1997).

The authors thank Dr Dickson K.W. Chiu, Faculty of Education, University of Hong Kong; and Dr Asefeh Asemi, Business Informatics, Corvinus University of Budapest; and Mohammad Dolati, PhD student of Plant Biotechnology, Ferdowsi University of Mashhad for comments that greatly improved manuscript.

Library Hi Tech

0737-8831
DOI 10.1108/LHT-11-2019-0229

Search tools are categorized into three groups: directories, search engines, and metasearch engines (Poulter, 1997; Green, 2000; Oppenheim *et al.*, 2000), the most notable of which are search engines, as more than 80 percent of Internet users benefit from their use to find the information they need (Kumar and Sampath Pavithra, 2010; Kerkmann and Lewandowski, 2012). More than 1,000 search engines are listed on the Website of the "The Search Engines List" (2010), and the number is increasing daily. Nonetheless, search engines for specific languages are also designed. Nowadays, more than ten Persian search engines are operating and rendering services to the users. Each of them has its own strategy and policy, search features, and facilities. They are different in terms of the database size, retrieval algorithm, the user-interface, and so forth.

Despite these differences in search engines, on the one hand, their designers claim their own engine to be the best and most suitable platform for searching. On the other hand, the number of the search engines is constantly increasing. It leads to the selection of the most practical tools. The differences include coverage, content, search strategy, resource provision, and ranking, and the set of tools they provide to help users to cause each of the search engines to find different results for searching the same information needs (Clarke, 2000). A look at the studies on the search engines reveals the number of studies comparing the international and the Persian search engines is very low.

According to the above-mentioned issues and as long as the function of the search engines is to retrieve documents relevant to the needs of the users, their evaluation will lead to greater knowledge of the search engines' abilities by the users (Croft *et al.*, 2010); Lewandowski (2008a) stated that international search engines faced difficulties when the results were restricted to a local language, the basic problem being comparing the effectiveness of Persian search engines with international ones and to determine which search engine can find more relevant results. In order to achieve the research objectives, the following hypotheses have been adopted:

Main hypothesis

(1) There is a significant difference between the effectiveness of Google, Parsijoo, Rismoon, and Yooz search engines.

Sub-hypotheses

(1) There is a significant difference between the precision ratio of Google, Parsijoo, Rismoon, and Yooz.

(2) There is a significant difference between the recall ratio of Google, Parsijoo, Rismoon, and Yooz.

(3) There is a significant difference between the NDCG of Google, Parsijoo, Rismoon, and Yooz.

Literature review

Many researchers in information science, such as Resnick (1961), Saracevic (1995), Tague-Sutcliffe (1996), Chu and Rosenthal (1996), Hawking *et al.* (1999), Oppenheim *et al.* (2000), Voorhees (2001), MacFarlane (2007), Bilal (2012), Lewandowski (2015), and Damessie *et al.* (2016), discuss about "evaluation." Evaluation means assessing the performance or value of a system, process (technique, procedure. . .), product, or policy (Saracevic, 1995). The value of the information retrieval system, especially the search engines, is to retrieve documents relevant to users' needs (Cooper, 1971; Voorhees, 2001; Saracevic, 2007). In other words, the value of information retrieval systems refers to effectiveness and efficiency. It should be noted that effectiveness differs from efficiency (Goel and Yadav, 2012). Effectiveness, loosely

speaking, measures the ability of the search engine to find the relevant information, and efficiency measures how quickly this is done. Evaluation is the key to making progress in building better search engines. It is also essential to understand if a search engine is being used effectively in a special application (Croft *et al.*, 2010). The approaches of evaluation in some studies such as Borland (2003), Saracevic (2007), and Hjørland (2010), which are cited in the following, have been discussed.

System-based evaluation

The main studies on information retrieval evaluation (from the Cranfield studies in the 1950s and the 1960s, to the evaluations of the Text Retrieval Conference (TREC) in the 1990s), such as Kent *et al.* (1955) and Cleverdon (1967), are based on the system-oriented approach. It is based on a model of IR, called the traditional or laboratory IR model, in which the emphasis is on systems processing information objects and matching them with queries (Saracevic, 1995). Document and queries represent by representation method (algorithm), and matching them is a base for relevance judgment (Järvelin, 2007). To evaluate search engines, some researchers have used this approach, and several researchers (Wu and Crestani, 2003; Ali and Beg, 2011; Bar-Ilan *et al.*, 2006; Can *et al.*, 2004; Cen *et al.*, 2009; Chowdhury and Soboroff, 2002; Hou, 2009; Isfandyari Moghaddam and Parirokh, 2006; MacFarlane, 2007; Nuray and Can, 2006; Sadeghi, 2011; Shi *et al.*, 2010; Zhang and Fei, 2010) have suggested various automatic methods to compare search engines in terms of retrieval effectiveness, which are discussed below.

Bar-Ilan (1998) investigated the retrieval effectiveness of six search engines (AltaVista, Excite, Infoseek, Lycos, Magellan, and Opentext) on a simple query (Erdos). He specified the degree of relevance of documents by considering whether the word "Erdos" was in the title, URL, and. . . or not.

Shang and Li (2002) use four relevance evaluation algorithms for calculating the precision of six search engines, namely, AltaVista, Fast, Google, Go, Won, and Northern Light. The algorithms were vector-space model, Okapi similarity measurement, cover density ranking, and three-level scoring methods.

Chowdhury and Soboroff (2002) evaluated five search engines, namely, Lycos, Netscape, Fast, Google, and HotBot, using the automatic evaluation method.

Nuray and Can (2006) introduced new methods for automatic ranking of retrieval systems. Their method includes two parts, namely, selecting systems for data fusion and selecting documents as pseudo-relevant documents in the fusion result.

Isfandyari Moghaddam and Parirokh (2006) introduced a new method using overlap between search engines and metasearch engines. In this method, you can select keywords and search in each search engine and metasearch engine. Two lists were prepared: one list was based on the first ten results recalled by the search engine, and the other was based on the first 40 results recalled by the metasearch engines. Then, based on the overlap of the results, the retrieval ability of the metasearch engines is calculated. In a similar study, Spoerri (2007) suggested a new method, using the structure of overlap between search results to rank retrieval systems.

Ali and Beg (2011) review some of the efforts made for the evaluation of Web search systems. They presented an automatic Web search evaluation system that combines the different evaluation techniques using a rough set based rank aggregation technique.

Kumar and Prakash (2009b) compared the retrieval effectiveness of Google and Yahoo based on matching words in queries and Web pages retrieved through search engines.

Sadeghi (2011) introduces two new automatic methods for evaluating the performance of search engines, namely, "tendency degree" and "coverage degree." In other words, he

employed metasearch engines as judges to evaluate the performance of search engines without the need for human relevance judgments. Then, he evaluated experimentally the performance of the search engines (Ask.com, Bing, and Google) based on the 50 topics of the 2002 TREC Web track.

Human-based evaluation

Vickery (1959) discussed the weaknesses of the system-oriented approaches. Like Vickery (1959), Dervin and Nilan (1986) also contrasted the system-oriented evaluation with the human-oriented evaluation, and issued an impassioned call for a paradigm change or shift in IR evaluations from system-oriented to human-oriented evaluations. Hawking *et al.* (1999) discussed the challenges in Text Retrieval Conference evaluation. From this time onward, the judgment of relevance by humans has emerged as the dominant approach. In this regard, Griesbaum and Spink (2004) and Bar-Ilan (2005) believe that for comparing rankings of different tools, one can compute similarity measures without the involvement of human judges, but judging the rankings produced by a specific search tool, the best method is human judgment. Since then, various researchers such as Griesbaum and Spink (2004), Demirci *et al.* (2007), Luyt *et al.* (2009), Teixeira Lopes and Ribeiro (2011), and Golzardi *et al.* (2013) have evaluated the search engines using human judgment, some of which are mentioned below.

Modifying their own pilot study of 1998, Su and Chen (1999) evaluated four search engines: Altavista, Excite, Lycas, and Infosys. They asked 36 students to search for their own information needs in the engine, and then evaluate them in terms of five criteria, including relevance, utility, efficiency, user satisfaction, and connectivity (continuity).

Smith (2003) used New Zealand topics to search in three local New Zealand search engines, four major global search engines, and three metasearch engines. He calculated recall metrics to evaluate them.

Goh and Ang (2003) compared the retrieval effectiveness of Overture and Google. They submitted 45 queries to each of the mentioned search engines, and the first ten documents returned were analyzed using different relevancy criteria.

Xie (2004) asked 21 undergraduate students to search two kinds of topics on an online database system (Dialog for health-related topics and Factiva for business-related topics), a directory (Yahoo), a search engine (Google), a metasearch engine (MetaCrawler), and a specialized search engine of their own choices. Finally, he used open-ended questions to elicit information regarding what the participants liked the most and disliked the most about Web search tools.

Lewandowski (2008a) tests the ability of Google, Yahoo, MSN, and Ask to distinguish between German- and English-language documents. He used advanced search by the mentioned search engine to look for 50 words that were in both German and English.

Tawileh *et al.* (2010) compared the effectiveness of five different search engines, two of which were Arabic engines: Araby and Ayna. The others were international Arabic-enabled engines such as Google, MSN, and Yahoo. They used 50 randomly selected queries from the top searches on Araby. The relevance of the top ten results and their descriptions retrieved by each search engine for each query were evaluated by independent jurors.

Garoufallou (2012) asked 16 librarians to search for predetermined topics in four international search engines (Altavista, Exalead, Google, and Yahoo) four Greek search engines (Find, Google -google.gr-, In, Robby), and complete a questionnaire for each search engine.

Sampath Kumar and Kumar (2013) examined the use of various search engines and metasearch engines. They also intended to know whether the Indian academics used a search strategy of various search engines for information retrieval or not. They used a questionnaire to gather the data.

User-based evaluation

With a cursory glance at the above-mentioned research, it is clear that the human-oriented approach has become the dominant approach in evaluating information retrieval systems. The point to note in this regard is that human is a general concept. Human can be an expert in any field or a researcher himself, or can be someone who really needs the information. Hjørland (2010) emphasized that relevance judgment can be human judgment, but not user judgment. For example, it is better to tell the end-users when a researcher or an expert in a field determines the degree of relevance, as it is human judgment but not the end-user type. In this regard, Saracevic (2007) emphasized that in a user-oriented approach, only the end-users can be the judge. In this regard, Bookstein (1979) believes that anyone who disagrees with the end-user judgment is necessarily and wrongly mistaken. Harter (1996) considers this approach of relevance as psychological relevance. According to Zeynali Tazehkandi and Nowkarizi (2020), this approach is rooted in the philosophy of Heraclitus. As Fidel (1993) and Wilson (2000) point out, the discussions of the qualitative research method have influenced the development of this evaluation approach in information retrieval. In other words, in a user-oriented approach, just anyone can judge the relevance of a document to their own information needs. Thus, these researchers distinguished between user and non-user judgments. However, both are human judgment. Various researchers have carried out research with this idea, some of which are mentioned below.

Hariri (2011) evaluated the relevance of retrieved documents by Google. She asked 34 graduate students from various disciplines to search their information needs in Google, and then designate a degree of relevance (most relevant, partially relevant, and irrelevant) to each document retrieved by Google.

In a similar study, Liu (2011) evaluated Bing, Blekko, and Google. He asked 35 computer science students to search their own information needs using them; consequently, they specified their satisfaction with the retrieved results with three degrees (fully satisfied, relatively satisfied, and dissatisfied).

Integration-based evaluation

With a superficial look at the above studies, it is clear that on the one hand, the human-based approach, and then the user-based approach, have become the dominant approaches in evaluating information retrieval systems. Nevertheless, some researchers such as Chowdhury and Soboroff (2002), Can *et al.* (2004), and Sadeghi (2011) have proposed automatic methods (software and algorithmic methods) to evaluate information retrieval systems. On the other hand, other researchers have paid attention to meta-evaluation (an evaluation of the evaluation). In this regard, Resnick and Savage (1964) and Hoffman (1965) examined the consistency of human judgment in evaluating information retrieval systems. Voorhees (2001) discussed the fundamental assumptions and appropriate uses of the Cranfield paradigm. Damessie *et al.* (2016) examined the effect of document order and topic difficulty on judges. Saracevic has written several articles (1995, 2007, 2012; 2015) that have accounted for a critical and historical analyses of evaluations of IR systems and processes. He discussed the strengths and weaknesses of the two approaches. Although the human approach also has some disadvantages, different researchers have used human judgment to evaluate the relevance of the document. Saracevic (2007) reported a total of six (6) separate levels, namely, the engineering (ENG) level, the input (IPT) level, the processing (Hripcsak and Rothschild, 2005) level, the output (OPT) level, the user and use (UAU) level, and the social level. The system-oriented approach covers the first three levels, and the user-oriented approach consists of the last three levels (Saracevic, 2007; Akhigbe *et al.*, 2011). Therefore, it is clear that by applying any of these approaches in the evaluation of information retrieval systems, each of the six levels is not measured. Fidel (1993) believed that the system-oriented

approach was influenced by the quantitative method, and the user-oriented approach by the qualitative method. Like Fidel (1993), Thornley (2005) stated that information retrieval (IR) had two main research traditions. These were the empirical or quantitative tradition, which dealt primarily with symbol manipulation and matching and did not explicitly concern itself with the problem of meaning, and the cognitive or qualitative tradition, which proposed that the problem of meaning was central to IR. In this relation, Saracevic (2007, p. 1925) stated, "As it turns out, both sides in the battle are wrong. Dervin and Nilan and followers were wrong in insisting on the primacy or the exclusivity of the user approach. Systems people were wrong in ignoring the user side and making the traditional IR model an exclusive foundation of their research for decades on end. Neither side got out of their box. Deep down the issue is not a system versus user approach. It is not system relevance against user relevance. The central issue and the problem is: How can we make the user and system side work together for the benefit of both? When IR systems fail that we are needed both the system-oriented and user-oriented approach to information retrieval evaluation." In this regard, Fidel (2008), in his review of research methods in the field of information science, states that a new approach has been proposed, called "mixed method," which claims to integrate the two quantitative and qualitative approaches of research method. Huang and Soergel (2013) also agreed that both approaches are complementary. Bates (2002) also believed that integrating the two approaches is essential. In this regard, Borlund and Schneider (2010) suggested the use of simulated work tasks to evaluate information retrieval systems to address both approaches. Using simulated work tasks draws attention to the context. Also, Bouramoul *et al.* (2011) believed that the exploitation of contextual elements could be a very good way to evaluate the search tools.

Zhang *et al.* (2013) compared the effectiveness of Google and Baidu search engines. They designed 20 search tasks focusing on the four subject domains, including medicine and health, culture and education, information and technology, and business and economy, with five search tasks assigned for each one. The results indicated that the effectiveness of the Google search engine was higher than Baidu's.

Kelly *et al.* (2015) used 48 participants in the context of a laboratory interactive information retrieval (IIR) experiment to investigate the understanding of the tasks. Borlund (2016) reviewed many articles in the evaluation of the information retrieval system field to investigate how various researchers were using simulated work tasks in their research.

Retrieval of Persian documents and search engines

Various studies such as Hayati and Alijani (2012), Mirgood *et al.* (2015), and Aslanian *et al.* (2016) have investigated the retrieval of Persian documents through international and local search engines, some of which are discussed. Erfanmanesh and Didegah (2012) evaluated 16 Persian search engines based on six criteria, including traffic, links, page views, time spent on site per user, and Iranian and foreign visitors. They collected their data from the Alexa Website and used the Statistical Analysis System (SAS) software to analyze them.

Hariri (2011) evaluated the relevance of retrieved documents through Google. She asked 34 graduate students from various disciplines to search their information needs in Google, then designated a degree of relevance (most relevant, partially relevant, and irrelevant) to each document retrieved by Google.

Mahmoudi *et al.* (2014) evaluated the performance of Google as a well-known search engine, and that of Parsijoo as a Persian search engine by investigating how the search engines behaved for the Persian queries. They posed 2000 queries to three search engines supporting the Persian language. Then mean reciprocal rank and success *N* measures were used for evaluation of the information retrieval effectiveness.

Morvarid *et al.* (2016) reported the average rank of Google, Yooz, Parsijoo, and Yahoo as 521.28, 507.88, 496.27, and 476.57, respectively. Comparing the results of this study with theirs indicates that in both, Google's effectiveness has been higher than that of Parsijoo and Yooz. In another study, Nowkarizi and Zeynali Tazehkandi (2017) evaluated four Persian search engines including Parsijoo, Yooz, Parseek, and Rismoon by using 32 queries. Based on Persian search engine evaluation studies, it can be said that although there are more than 20 local search engines to retrieve Persian documents, Parsijoo and Yooz are more well-known than other local search engines.

Evaluation matrix or evaluation criteria

Many information scientists advocate that evaluation of information retrieval systems should pay more attention to those factors that can provide improved services to the users. In this regard, Cleverdon *et al.* (1966) proposed six criteria for the evaluation of an information retrieval system, including the ability of the system to present all the relevant items (recall), the ability of the system to present only those items that are relevant (precision), the average interval between the time the search request is made and the time an answer is provided (time lag), the effort – intellectual as well as physical – required from the user in obtaining answers to the search requests, the form of presentation of the search output, which affects the user's ability to make use of the retrieved items, and the coverage of the collection, that is, the extent to which the system includes relevant matter. In a similar study, Vickery (1970) identified six criteria for the evaluation of information retrieval systems. He grouped them into two sets as follows: set 1 (1. coverage – the proportion of the total potentially useful literature that has been analyzed, 2. recall – the proportion of such references that are retrieved in a search, and 3. response time – the average time needed to obtain a response from the system); and set 2 (1. precision – the ability of the system to screen out irrelevant references; 2. usability – the value of the references retrieved, in terms of such factors as their reliability, comprehensibility, currency; and 3. presentation – the form in which search results are presented to the user).

In another study, Lancaster (1971) identified five evaluation criteria, including the coverage of the system, the ability of the system to retrieve relevant documents or recall, the ability of the system to avoid retrieval of irrelevant documents or precision, the response time of the system, and the amount of effort required by the users. Lancaster (1971) identified five evaluation criteria including the coverage of the system, the ability of the system to retrieve relevant documents or recall, the ability of the system to avoid retrieval of irrelevant documents or precision, the response time of the system, and the amount of effort required by the users. Salton and McGill (1983) identified the various parameters of an information retrieval system as related to each of five evaluation criteria, including (1) recall and precision (indexing exhaustivity – recall tends to increase the exhaustivity of indexing terms; term specificity – precision increases with the specificity of the index terms; indexing language – availability of measures of recognition of synonyms, terms relations, etc., which improve recall; query formulation – ability to formulate an accurate search request; search strategy – ability of the user or intermediary to formulate an adequate search strategy); (2) response time (organization of stored documents; type of query; location of information center; frequency of receiving user's queries; size of the collection); (3) user effort (accessibility of the system; availability of guidance by system personnel; volume of retrieved items; facilities for interaction with the system); (4) form of presentation (type of display device; nature of output – bibliographic reference, abstract, or full text); (5) collection coverage (type of input device and type and size of storage device; depth of subject analysis; nature of users' demand; and physical forms of documents).

A review of the background reveals since the emergence of Web search tools; their evaluation has also been studied. One of these studies is to evaluate international and local

search engines in retrieving documents in local languages. In this regard, various researchers have focused on the retrieval of Persian language documents. Generally, early studies in evaluating information retrieval systems were based on a system-oriented approach (Kent *et al.*, 1955; Cleverdon, 1967). Notwithstanding, some researchers have explicated the weaknesses of this approach. Consequently, the human-oriented approach becomes the dominant approach in evaluating information retrieval systems such as (Su and Chen, 1999). After the human-oriented approach was the dominant one, and although the proponents of this approach agreed in theory, they were acting differently. Therefore, some researchers such as Saracevic (2007) and Hjørland (2010) revealed the distinction between the user (human) and non-user (human) judgment. Accordingly, words such as end-user, context, and situation are introduced.

In recent years, researchers such as Bates (2002), Saracevic (2007), and Huang and Soergel (2013) have been calling to integrate systems-oriented and user-oriented approaches. In this regard, some researchers have referred to dualist models (Saracevic, 2007), approaches (Thornley, 2005; Thornley and Gibb, 2007), and methods (Fidel, 2008; Ma, 2012) for evaluating the information retrieval systems. Also, Wilson (2002), Budd (2004), and Hjørland (2010) have tied search engine evaluation research in philosophy. In this regard, Thornley (2005), Thornley and Gibb (2007), and Zeynali Tazehkandi and Nowkarizi (2020) believe that the dialectical philosophical approach is appropriate to this subject. In addition, Zeynali Tazehkandi and Nowkarizi (2020) emphasize that the use of simulated work tasks is necessary but not sufficient; rather, all the steps used to evaluate search engines must be rooted in a philosophical approach. In other words, they point out that all stages of data collection must come from a composite philosophical approach that composed both system-oriented and user-oriented approaches. It can also be said that, according to them, composing is a higher level of integration. They believe that the dialectical approach composed these two approaches. Dialectics is a method of philosophical reasoning that involves some sort of contradictory process between opposing sides (Thornley and Gibb, 2007) that includes two stages: collecting and dividing. However, researchers such as Bates (2002), Saracevic (2007), and Huang and Soergel (2013) have suggested using an approach that integrates both system-oriented and user-oriented approaches. Also, researchers such as Thornley and Gibb (2007) and Zeynali Tazehkandi and Nowkarizi (2020) have introduced a dialectical approach that composed both system-oriented and user-oriented approaches. Therefore, in this study, a dialectical approach is used to evaluate search engines.

Design

This was an evaluative study, which was conducted according to a composite approach (quantitative and qualitative). Evaluative research is a particular type of applied research dependent variable which is a value, a goal, or an effect that is implemented in the real environment. Performance measurement is a more specific type of evaluative research that deals with output and efficiency indicators rather than merely considering the inputs (Connaway and Powell, 2010).

The population includes three sub-populations: local search engines, simulated work tasks (subject headings), and participants (students at FUM), each of which is described below and how they were chosen.

Selection of the search engines

After identifying the local search engines and providing a list of them, the researchers visited their Web pages. Their various characteristics were examined, some of which were their emergence history, accessibility, information retrieval facilities, number of indexed Web pages, non-promotional activities, and retrieving the relevant results.

Another criterion for selecting local search engines was to choose the search engines introduced and investigated by other researchers. Additionally, each search engine rank on Alexa website was also considered as a criterion to select it. After considering the rank of international search engines on the Alexa website, as well as focusing on the effectiveness of search engines reported in the previous studies on the topic, Google was also chosen as an international search engine.

Preparation of simulated work tasks (selection of subject headings)

As background research has shown, researchers such as Borlund and Schneider (2010), Bouramoul *et al.* (2011), and Borlund (2016) have suggested the use of simulated work tasks to evaluate the information retrieval system. So, we used it to consider contextual elements. In this regard, it should be clear how many simulated work tasks are involved and what their topics are. Based on a review of the literature, it can be seen that different researchers have used different numbers of queries or simulated work tasks to evaluate search engines, as mentioned in Table I.

According to Table I, it seems that using 20 to 40 queries or simulated work tasks is reasonable. So, to avoid personalized and arbitrary selection, 32 subject headings were selected through a stratified sampling method from the Persian Subject Headings (PSH) as a basis to choose the simulated work tasks. To do this, the PSH list and its appendices were divided in terms of the number of pages, into 32 sections, from which a page was selected as a sampling unit using the random number table (Connaway and Powell, 2010). Then each target page in the PSH list was opened, and a subject heading was selected by closing the eyes and putting the fingertip on one of them. The selected subject headings are presented in Table AI.

According to the selected subject headings for the sample, some simulated work tasks were formulated by surveying three experts in the field of Library and Information Science (Van Rijsbergen). One of the subject headings was "Writing a Resume," a simulated work task that is presented below.

Assume you have been graduated. You would like to write and submit your CV for a recruitment ad or an employment agency, but you do not know how to write it. Thus, you should consult the appropriate resources for knowing how to write the resume and do it.

Table I. Number of queries in the studies

Author	Year	Number of queries	Author	Year	Number of queries
Bar-Ilan	1998	1	Luyt *et al.*	2009	14
Su and Chen	1999	36	Tawileh *et al.*	2010	50
Leighton and Srivastava	1999	15	Deka and Lahkar	2010	50
Yaltaghian and Chignell	2002	7	Kumar and Sampath Pavithra	2010	15
Goh and Ang	2003	45	Lewandowski	2011	100
Smith	2003	10	Sadeghi	2011	50
Wu and Li	2004	24	Hariri	2011	34
Can *et al.*	2004	25	Liu	2011	35
Jansen and Molina	2006	100	Bilal	2012	30
Moghaddam and Parirokh	2006	5	Zhang *et al.*	2013	20
Demirci *et al.*	2007	12	Mahmoudi *et al.*	2014	2000
MacFarlane	2007	50	Ajayi and Elegbeleye	2014	5
Lewandowski	2008a	50	Morvarid *et al.*	2016	5
Lewandowski	2008b	40	Nowkarizi and Zeynali Tazehkandi	2017	32
Sampath Kumar and Prakash	2009	15	Wu *et al.*	2019	200

Guidance: Copy the URLs that may help you write the resume in a word file.

Selection of the participants (judgments)

The participants were FUM students. Different researchers have used different numbers of participants; for example, Lewandowski (2008a), Hariri (2011), Garoufallou (2012), and Sampath Kumar and Kumar (2013) used 40, 34, 16, and 450 participants, respectively. So, regarding counseling faculty members of the LIS of FUM, 192 students were selected, by stratified random sampling, from various educational degrees and studying in different FUM disciplines. Some variables such as gender and age (Vakkari and Järvelin, 2005), educational degree (Saracevic and Kantor, 1988), and educational area (Davidson, 1977) are considered as the factors influencing the process of information retrieval.

The process of the research implementation

After preparing simulated work tasks, in each search form, we designed two of them and submitted them to the participants along with the search execution instructions. The participants read each simulated work task, and then some information needs were formed in them. They searched these information needs originated from the work tasks through the determinate search engines. Then they read the retrieved Websites, copied the website URLs relevant to the simulated work task, and recorded them in an electronic search form. Finally, they sent the filled form to the researchers' email address or a designated telegram group introduced to them.

Determining the relevance score of the URLs

After receiving the search forms from the participants, we recorded the participant-selected URLs in the Excel file; they were arranged in alphabetical order to calculate, easily, the frequency of each URL by using the sort order. As a result, the frequency of each was calculated.

Up to this stage, the rate of the URLs related to the subject introduced by the participants was designated. In other words, the rate of the URLs selected by the participants as a relevant URL determined the relevance score of them. To normalize the URLs' relevance score and measure the precision, recall, and normalized discounted cumulative gain (NDCG), the relevancy of each URL should be calculated ranging from 0 to 1. In this regard, according to Jain *et al.* (2005) and Jain and Bhandare (2011), if we want to normalize our data, we can simply calculate:

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

where $x = (x_1, \ldots, xn)$ and z_i is now our ith normalized data. The relevance score of a URL can be zero. So, we have used the following formula:

$$\text{Normalized relevance score of } n \text{ URL in A SWT} = \frac{\text{frequency of } n \text{ URL in A SWT}}{\text{The highest frequency in A SWT}}$$

For example, suppose that the URL n for the simulated work task A had been selected three times by participants, and the URL m with ten times had the maximum frequency of selection. Then we would have:

$$\text{Normalized relevance score of } n \text{ RL in A SWT} = \frac{3}{10}$$

Up to this point, how to determine the degree of relevance of the documents has been determined, but we need to use evaluation metrics to compare search engines. This is discussed below.

Metrics

Any evaluation metric considers different characteristics of information retrieval systems. So According to Baccini *et al.* (2012), it is better to use different metrics to evaluate search engines. If different metrics are used, most of the features of information retrieval systems will be measured. So, according to Croft *et al.* (2010) and Bama *et al.* (2015), three metrics, including precision, recall, and NDCG, were used to evaluate search engines, which are explained as follows:

Precision

Precision is one of the most commonly used metrics (Bar-Ilan, 1998; Leighton and Srivastava, 1999; Gordon and Pathak, 1999; Kumar and Sampath Pavithra, 2010; Hariri, 2011). According to Harter (1996), Powers (2011), and Buckland (2017), precision refers to the proportion of retrieved documents that are relevant to the query. According to Kumar and Prakash (2009a) and Saracevic (2012), we modified the binary formula of precision and recall. Therefore, the following formula has been used to calculate these metrics:

$$\text{precision of B search engine for A SWT in D e-form} = \frac{\text{sum of the normalized relevance score of URLs retrieved by B search engine for A SWT in D e-form}}{\text{total number of URLs retrieved by B search engine for A SWT in all e-form}}$$

Recall

Recall is another measure that has been used in most studies to evaluate information retrieval systems (such as (Bar-Ilan, 1998; Bitirim *et al.*, 2002; Robinson and Wusteman, 2007). Intuitively, recall measures how well the search engine is doing at finding all the relevant documents fitted to a query (Harter, 1996; Buckland, 2017; Croft *et al.*, 2010). Since the relevance score of the document is continuous and comparative, here the following formula has been used to calculate the measures:

$$\text{recall of B search engine for A SWT in D e-form} = \frac{\text{sum of the normalized relevance score of URLs retrieved by B search engine for A SWT in De-form}}{\text{sum of the normalized relevance score of URLs retrieved by 4 search engine for A SWT in all e-form}}$$

NDCG

According to Croft *et al.* (2010), the focus of an effective measure should be on how well the search engine performs at retrieving relevant documents at very high ranks. It seems the precision (above-mentioned formula) and recall are not convenient measures. In this regard, Järvelin and Kekäläinen (2002) believe that the focus of effective measures should be to determine how the search engines can retrieve more relevant documents before relevant documents. This is calculated using the following formula:

$$\text{NDGC} = \frac{\text{DCG}}{\text{ideal DCG}}$$

Effectiveness

Effectiveness is a term used in many information retrieval studies such as (Lewandowski, 2008a; Goel; Yadav, 2012). Effectiveness refers to the ability of the search engine to find the relevant information which measures, using precision, fallout, recall, BPREF, NDCG, and so on (Van Rijsbergen, 1974; Croft *et al.*, 2010). To calculate the effectiveness of B search engine, the following formula is used:

Effectiveness of B search engine for A SWT in D e-form = sum of precision, recall, and NDCG scores of B search engine for A SWT in D e-form divided by 3.

After measuring, the data obtained from the mentioned calculations were entered in SPSS20. Then the convenient statistical tests were used, according to the existing conditions, which are described in detail below.

Validity and reliability of research tools

The validity was confirmed through the researchers' studies and the views of the faculty members and related texts, particularly (Borlund, 2003; Saracevic, 2012; Borlund, 2016). Additionally, during the phases of implementation and search forms, tasks and other related issues were reviewed and revised by some experts in the field of LIS. Then, the search forms and simulated work tasks were submitted to the LIS faculty members of FUM (six people). Finally, according to the received comments, the necessary items were corrected and finalized. To measure reliability, six SWT were given to two groups of participants (each group includes 32 persons), and they were asked to search their information needs created from the study of SWTS in the determined search engines and copy the relevant URLs in a word file (2 SWTs given to each participant, so that they totally conducted 128 searches). Then the precision, recall, and NDCG scores were calculated for the search engine at two different groups of participants. Finally, the correlation of the test and retest phases was measured, and this was 0.739, which confirmed the tool reliability.

Findings

After being collected, the data were entered into SPSS (version 20). Since the scale of data was quantitative in all the hypotheses, the first condition of the parametric tests was true. Then, in each hypothesis, the normal distribution of variables was estimated by using appropriate tests.

Before testing the main, overarching hypothesis, it was branched into more specific and better testable sub-hypotheses, which were tested. Finally, the main hypothesis were concluded from the three sub-hypotheses.

First sub-hypothesis

As mentioned, since the normality test showed the distribution was normal, the repeated measures test was used. Mauchly's test of sphericity was applied to verify the uniformity of covariance. Its significant level (0.13) indicated that the covariance uniformity assumption is confirmed. Therefore, the sphericity assumed test was used to examine the precision of the search engines. The significance of the sphericity assumed (0.001) test was less than alpha (0.05). Therefore, there was a significant difference between the precision of Google, Parsijoo, Rismoon, and Yooz. Then, a pair-sampled *t*-test had been used to indicate which two means were significantly different. The results were shown in Table II.

As it is shown in Table II, there was no significant difference between the precision of Google and Parsijoo, Google and Yooz, and Parsijoo and Yooz, while there were significant differences between the precision of Google and Rismoon (0.00), Parsijoo and Rismoon (0.00), and Yooz and Rismoon (0.00). For more specific information on the precision of the mentioned

search engines, the mean confidence interval (95 percent) of the search engines' precision was plotted in Figure 1.

According to Figure 1, it is clear that the precision of the three search engines, Google, Parsijoo, and Yooz was very close together and did not differ significantly, while Rismoon had less precision than the other three ones. More generally, precision is relevant to measure the ability of systems to answer queries for which a small set of documents is expected by the users. This is the case for question/answering searches for which a single relevant answer is enough and that needs a few answers (Baccini *et al.*, 2012). So there is not much difference between Google, Parsijoo, and Yooz to retrieve related simple answers.

Second sub-hypothesis

Since the recall of the search engines had a normal distribution, the repeated measure test was used. The uniformity of covariance was tested by the Mauchly's test of sphericity, which indicated that the assumption of uniformity of covariance was not confirmed (p-value = 0.001). Hence, Greenhouse–Geisser was used to test the difference between the search engines' recall. Accordingly, the test p-value (0.001) indicated that there was a significant difference between the search engines' recall. Therefore, a pair-sampled t-test had been used to indicate which two means were significantly different. The results are shown in Table III.

As it is shown in Table III, the significance of all six pairs of search engines was less than (0.05), which means that there was a significant difference between the recall of each of the six pairs of search engines. To better represent the search engines' recall, the confidence interval (95 percent) of them was plotted in Figure 2.

As shown in Figure 2, the recall of the Google search engine was more than that of the local search engines and was estimated to be around 21 to 33 percent. In addition, the recall of Parsijoo, Yooz, and Rismoon was estimated, respectively, 15–25 percent, 12–24 percent, and

	Pairs	Mean	Test statistic	Df	p-value
Pair 1	Google and Parsijoo	0.05	1.37	31	0.20
Pair 2	Google and Yooz	0.07	1.78	31	0.08
Pair 3	Google and Rismoon	0.29	5.24	31	0.001
Pair 4	Parsijoo and Yooz	0.01	0.46	31	0.64
Pair 5	Parsijoo and Rismoon	0.24	4.49	31	0.001
Pair 6	Yooz and Rismoon	0.22	4.06	31	0.001

Table II. Pair-sampled t-test to determine the difference of pairs of search engines precision

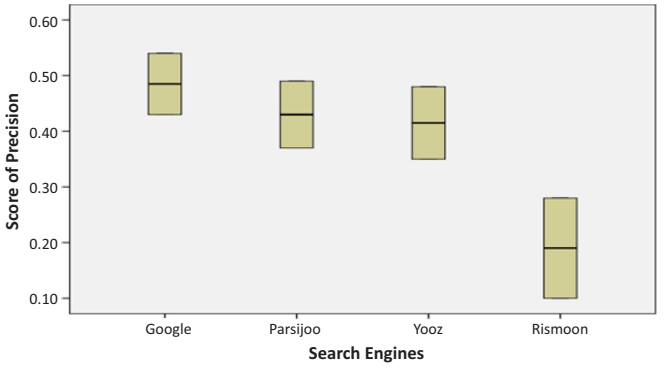

Figure 1. Estimation of the mean confidence interval of the search engines' precision

0.01–0.04 percent at a 95 percent confidence level. Considering a user's point of view, the recall, which is related to the capacity of a system to retrieve most of the relevant documents, is very important (Su, 1994). Because sometimes users intend to dominate the topic, so access to all or most of the resources matters. For example, this is especially interesting in tasks like science monitoring when sets of documents have to be gathered for further analysis or text mining, as well as to access to various aspects of the subject. Thus, in such a situation, the use of Google is recommended.

Third sub-hypothesis

For this variable, a normality test showed that the distribution of data was not normal. Therefore, for measuring the significant difference in NDCG of the surveyed search engines, Friedman's nonparametric test was used. The result showed that there was a significant difference between the search engines' NDCG (p-value = 0.001). But this test alone does not indicate which meanings are significant. Then to identify the differences between each pair of search engines, the sign test was used. The results are shown in Table IV.

As it is shown in Table IV, there was a significant difference between the NDCG of Google and Rismoon, Yooz and Rismoon, and Parsijoo and Rismoon, while the differences between the NDCG of Google and Parsijoo, Google and Yooz, and Parsijoo and Yooz were not significant. To better represent these differences of similarities, Figure 3 was drawn.

According to Figure 3, it is found that in the NDCG, the Rismoon got fewer scores than the other three search engines. Although Google, Yooz, and Parsijoo had better performance in this regard, there were no differences between them. Since NDCG score indicates the quality of the documents ranking of the search engines, as shown in Figure 3, the quality of Rismoon

Table III.
Pair-sampled t-test to determine the difference of pairs of search engines recall

	Pairs	Mean	Test statistic	Df	p-value
Pair 1	Google and Parsijoo	0.06	2.78	31	0.009
Pair 2	Google and Yooz	2.26	17.38	31	0.001
Pair 3	Google and Rismoon	0.24	7.57	31	0.001
Pair 4	Parsijoo and Yooz	2.19	16.52	31	0.001
Pair 5	Parsijoo and Rismoon	0.17	6.67	31	0.001
Pair 6	Yooz and Rismoon	2.01	14.04	31	0.001

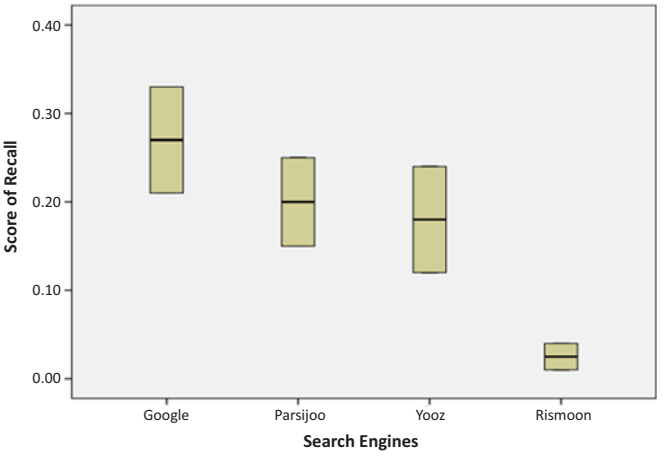

Figure 2.
Estimation of the mean confidence interval of the search engines' recall

was very low. Therefore, if users intend to use a search engine to rank these documents based on their relevance, it is suggested not to use Rismoon.

Finally, the main, overarching hypothesis was tested to find the total effectiveness differences of the search engines.

Main and overarching hypothesis

For more information on the status of Google, Parsijoo, Rismoon, and Yooz, their effective mean of precision, recall, and NDCG were drawn in terms of the SWTs shown in Figure 4.

As shown in Figure 4, the effectiveness dispersion of Rismoon was very high. At the same time, its effectiveness was higher than other search engines, while its performance in other search tasks was weak. It also indicated that Google dispersion was very low, while those of Parsijoo and Yooz were close to each other. Hence, it seems that Google gives the same importance to various subjects (search task). It also covers the different subjects similarly. Although Rismoon retrieved a few documents for some of the subjects, in some other areas, it managed to retrieve the most relevant documents.

The effectiveness of four search engines at the sample level was shown in Figure 4, but in order to know their effectiveness at the population level, it was necessary to use statistical tests. Thus, first, the normality test showed that the total data distribution (mean of precision, recall, and NDCG) was normal. Therefore, to test the hypothesis, repeated measures test was used.

The uniformity of covariance was tested by Mauchly's test of sphericity. It indicated that the assumption of uniformity of covariance was not confirmed (0.001). Hence, the Greenhouse–Geisser test was used to examine the difference between the search engines'

	Search engines	*p*-value
Pair 1	Google and Parsijoo	0.201
Pair 2	Google and Yooz	1
Pair 3	Google and Rismoon	0.001
Pair 4	Parsijoo and Yooz	1
Pair 5	Parsijoo and Rismoon	0.001
Pair 6	Yooz and Rismoon	0.001

Table IV.
Sign test to determine the difference of the pairs of search engines NDCG

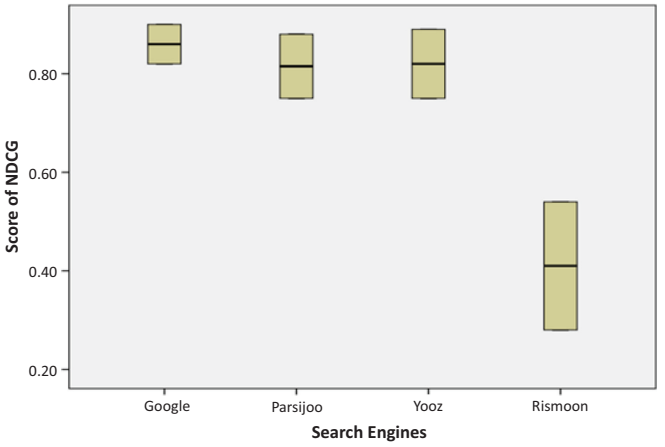

Figure 3.
Estimation of the mean confidence interval of the search engines' NDCG

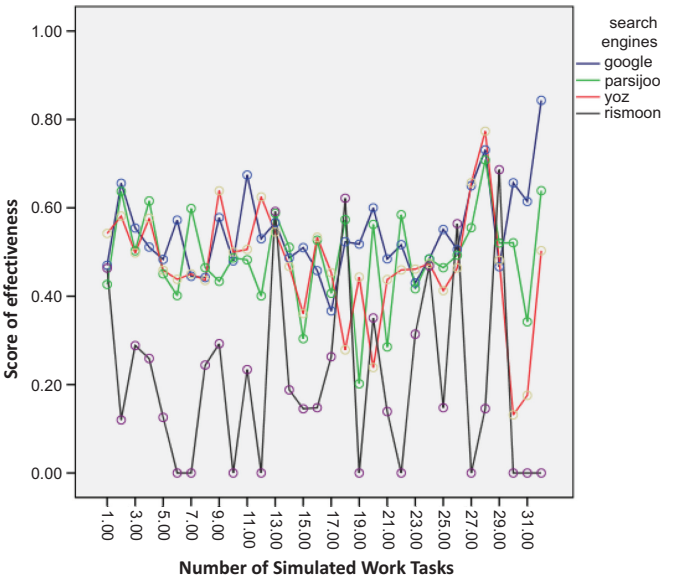

Figure 4.
Effectiveness of four search engines in SWT separately

effectiveness. The results indicated that there was a significant difference between the effectiveness of the search engines ($p = 0.001$). To show the detailed differences between pairs of search engines, paired-sample *t*-test has been used, and the results are shown in Table V.

As shown in Table IV, there was a significant difference between the effectiveness of Google and Parsijoo, Google and Yooz, Google and Rismoon, Parsijoo and Rismoon, Yooz and Rismoon, while any significant difference was not observed between the effectiveness of Parsijoo and Yooz. Figure 4 is illustrated for a more precise and objective knowledge of the effectiveness of each of the surveyed search engines in relation to each other.

As shown in Figure 5, the effectiveness of one of the local search engines, namely, Rismoon, was significantly less than the three others, as well as the effectiveness of the other two ones, namely, Parsijoo and Yooz, was very close and similar. Finally, it seemed Google had a better performance than the local search engines.

Discussion

The World Wide Web, with its short history, has experienced significant changes. Also, the earlier search engines were established based on the traditional database and information retrieval methods, and many other algorithms and methods have since been added to them to improve their results. The dynamic nature of the Web and the shifting of search engines over

Table V.
Paid-Samples *t*-test to determine the difference of effectiveness of pairs of search engines

Pairs of search engine		Mean	Test statistic	Df	*p*-value
Pair 1	Google and Parsijoo	0.056	2.75	31	0.01
Pair 2	Google and Yooz	0.07	2.67	31	0.012
Pair 3	Google and Rismoon	0.32	7.37	31	0.001
Pair 4	Parsijoo and Yooz	0.01	0.721	31	0.476
Pair 5	Parsijoo and Rismoon	0.27	7.08	31	0.001
Pair 6	Yooz and Rismoon	0.25	5.93	31	0.001

time require continuous evaluation of search engines. So although in recent studies Persian local search engines have been compared to Google, their evaluation is still important because both search engines are changing and evaluation approaches have developed. So, in this study, we compared the Google, Parsijoo, Yooz, and Rismoon search engines using three metrics of precision, recall, and NDCG. As seen in Table II, the precision of all search engines was significantly different, but there was no significant difference between Google and Parsijoo and Google and Yooz. Regarding the recall metric, it can be seen in Table III that except for Google and Parsijoo, the recall rates of all the search engines were significantly different from each other, but in relation to the NDCG metric, it is observed (Table IV) that the Rismoon search engine had significantly poorer performance, but no significant difference was observed between the other search engines. Finally, the effectiveness of the search engines was tested according to these three metrics. As can be seen in Table V, the effectiveness of all search engines was significantly different except for the Parsijoo and Yooz pairs. Also, Google's search engine performed significantly better than the local search engines.

As mentioned earlier, some researchers have compared the effectiveness of international search engines to local ones. In this regard, Morvarid *et al.* (2016) reported the average rank of Google, Yooz, Parsijoo, and Yahoo to be 521.28, 507.88, 496.27, and 476.57, respectively. Comparing the results of this study with these indicates that in both, Google's effectiveness has been more than Persijoo and Yooz. But unlike their results, there was no significant difference between the effectiveness of Parsijoo and Yooz in this study, which may be due to the difference in the queries done in the search engines.

In their study, Mahmoudi *et al.* (2014) reported that Parsijoo's effectiveness was more than Google's, which contradicted the results of this study. Most likely, this result seems to be due to the fact that in their study, only the navigational queries were searched for, while in the present study, the queries were not limited by the topic or type. In another study, Hariri (2011) indicated that Google's effectiveness was 50 percent. In this research, Google's effectiveness was estimated to be between 50 and 57 percent, based on which we can consider the two results similar. The partial difference observed may be due to the time passed. As time goes on, Google's retrieval policies and algorithms are revised. It may strengthen the search engine effectiveness.

According to the results reported by Tawileh *et al.* (2010), Google's effectiveness was higher than Arabic search engines. In addition, the results of Zhang *et al.* (2013) (2013) indicated that Google's effectiveness was higher than Baidu's. In the light of the present research's findings and some previous studies such as Griesbaum and Spink (2004), Demirci *et al.* (2007), Luyt *et al.* (2009), Deka and Lahkar (2010), and Garoufallou (2012), it can be said that Google performs better.

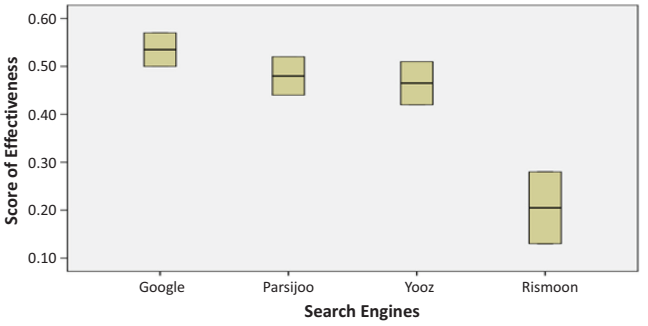

Figure 5.
Estimation of the mean confidence interval of the search engines' effectiveness

Conclusion

Generally, a search engine consists of several components that can include crawler, indexer, retrieval algorithm, query processor, interface, and ranker. In this study, three local Persian search engines (Parsijoo, Yooz, and Rismoon) and Google were evaluated for all of these components using the dialectical approach. In detail, each participant searched two tasks in the determined search engines (384 searches). They recorded 1,243 URLs as relevant. Finally, Google's effectiveness was estimated from 50 to 57 percent; Parsijoo, 44–52 percent; Yooz, 42–51 percent; and Rismoon, 13–28 percent. This indicated that Google's effectiveness was more than that of Persian local search engines. If users use efficient search engines to get the information they need, they will gain access to relevant information in less time, and the present study shows despite the focus of local search engines on specific language resources, Google is still better than the local ones even in the local languages. Hence, it is better for users of search engines, especially Persian language users, to search the information needs in Google for retrieving the Persian information resource.

References

Ajayi, O.O. and Elegbeleye, D.M. (2014), "Performance evaluation of selected search engines", *Performance Evaluation*, Vol. 4, pp. 4-15.

Akhigbe, B.I., Afolabi, B.S. and Adagunodo, E.R. (2011), "Assessment of measures for information retrieval system evaluation: a user-centered approach", *International Journal of Computers and Applications*, Vol. 25, pp. 6-12.

Ali, R. and Beg, M.S. (2011), "An overview of Web search evaluation methods", *Computers and Electrical Engineering*, Vol. 37, pp. 835-848.

Anderson, B. (2006), "Indexing the internet", *Behavioral and Social Sciences Librarian*, Vol. 25, pp. 135-139.

Aslanian, H., Saeed, G.M. and Javad, S. (2016), "Comparative study on selected search engines in retrieving information cleft Lip & Palate in 2013-2015", *Navid No*, Vol. 18, pp. 8-15.

Baccini, A., Dejean, S., Lafage, L. and Mothe, J. (2012), "How many performance measures to evaluate information retrieval systems?", *Knowledge and Information Systems*, Vol. 30, pp. 693-713.

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

Bar-Ilan, J. (1998), "On the overlap, the precision and estimated recall of search engines. A case study of the query 'Erdos'", *Scientometrics*, Vol. 42, pp. 207-228.

Bar-Ilan, J. (2005), "Comparing rankings of search results on the web", *Information Processing and Management*, Vol. 41, pp. 1511-1519.

Bar-Ilan, J., Mat-Hassan, M. and Levene, M. (2006), "Methods for comparing rankings of search engine results", *Computer Networks*, Vol. 50, pp. 1448-1463.

Bates, M.J. (2002), "Toward an integrated model of information seeking and searching", *New Review of Information Behaviour Research*, Vol. 3, pp. 1-15.

Bilal, D. (2012), "Ranking, relevance judgment, and precision of information retrieval on children's queries evaluation of Google, Yahoo!, Bing, Yahoo! Kids, and ask Kids", *Journal of the American Society for Information Science and Technology*, Vol. 63, pp. 1879-1896.

Bitirim, Y., Tonta, Y. and Sever, H. (2002), "Information retrieval effectiveness of Turkish search engines", *International Conference on Advances in Information Systems*, Springer, pp. 93-103.

Bookstein, A. (1979), "Relevance", *Journal of the American Society for Information Science*, Vol. 30, pp. 269-273.

Borlund, P. (2003), "The IIR evaluation model a framework for evaluation of interactive information retrieval systems", *Information Research*, Vol. 8 No. 3, pp. 3-8.

Borlund, P. (2016), "A study of the use of simulated work task situations in interactive information retrieval evaluation a meta-evaluation", *Journal of Documentation*, Vol. 72, pp. 394-413.

Borlund, P. and Schneider, J.W. (2010), "Reconsideration of the simulated work task situation: a context instrument for evaluation of information retrieval interaction", *Proceedings of the Third Symposium on Information Interaction in Context*, ACM, pp. 155-164.

Bouramoul, A., Kholladi, M.-K. and Doan, B.-L. (2011), "Using context to improve the evaluation of information retrieval systems", arXiv preprint arXiv:1105.6213.

Buckland, M. (2017), *Information and Society*, Massachusetts Institute of Technology Press, Cambridge.

Budd, J.M. (2004), "Relevance: language, semantics, philosophy", *Library Trends*, Vol. 52, pp. 447-462.

Can, F., Nuray, R. and Sevdik, A.B. (2004), "Automatic performance evaluation of Web search engines", *Information Processing and Management*, Vol. 40, pp. 495-514.

Cen, R., Liu, Y., Zhang, M., Ru, L. and Ma, S. (2009), "Automatic search engine performance evaluation with the wisdom of crowds", *Asia Information Retrieval Symposium*, Springer, pp. 351-362.

Chowdhury, A. and Soboroff, I. (2002), "Automatic evaluation of world wide web search services", *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 421-422.

Chu, H. and Rosenthal, M. (1996), "Search engines for the World Wide Web: a comparative study and evaluation methodology", *Proceedings of the Annual Meeting-American Society for Information Science*, pp. 127-135.

Clarke, S.J. (2000), "Search engines for the world wide web", *Journal of Internet Cataloging*, Vol. 2, pp. 81-93.

Cleverdon, C.W. (1967), "The Cranfield tests on index language devices", *Aslib Proceedings*, MCB UP, pp. 173-194.

Cleverdon, C.W., Mills, J. and Keen, E.M. (1966), "Factors determining the performance of indexing systems, (Volume 1: design)", Cranfield: College of Aeronautics, Vol. 28.

Connaway, L.S. and Powell, R.R. (2010), *Basic Research Methods for Librarians*, ABC-CLIO.

Cooper, W.S. (1971), "A definition of relevance for information retrieval", *Information Storage and Retrieval*, Vol. 7, pp. 19-37.

Croft, W.B., Metzler, D. and Strohman, T. (2010), *Search Engines: Information Retrieval in Practice*, Addison-Wesley, Reading.

Damessie, T.T., Scholer, F., Järvelin, K. and Culpepper, J.S. (2016), "The effect of document order and topic difficulty on assessor agreement", *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ACM, pp. 73-76.

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

Deka, S.K. and Lahkar, N. (2010), "Performance evaluation and comparison of the five most used search engines in retrieving web resources", *Online Information Review*, Vol. 34, pp. 757-771.

Demirci, R.G., Kismir, V. and Bitirim, Y. (2007), "An evaluation of popular search engines on finding Turkish documents", *Second International Conference on Internet and Web Applications and Services (ICIW'07)*, IEEE, 61-61.

Dervin, B. and Nilan, M.S. (1986), "Information needs and use", *Annual Review of Information Science and Technology*, Vol. 21, pp. 3-33.

Erfanmanesh, M. and Didegah, F. (2012), "Evaluating function of Persian search engines on the web using correspondence analysis", *International Journal of Integrated Supply Management*, Vol. 8, pp. 77-87.

Fidel, R. (1993), "Qualitative methods in information retrieval research", *Library and Information Science Research*, Vol. 15, pp. 219-247.

Fidel, R. (2008), "Are we there yet?: mixed methods research in library and information science", *Library and Information Science Research*, Vol. 30, pp. 265-272.

Garoufallou, E. (2012), "Evaluating search engines: a comparative study between international and Greek SE by Greek librarians", *Program*, Vol. 46, pp. 182-198.

Goel, S. and Yadav, S. (2012), "An overview of search engine evaluation strategies", *International Journal of Applied Information Systems*, Vol. 1, pp. 7-10.

Goh, D.H. and Ang, R.P. (2003), "Relevancy rankings pay for performance search engines in the hot seat", *Online Information Review*, Vol. 27, pp. 87-93.

Golzardi, E., Meghdadi, M. and Ghaderzadeh, A. (2013), "Comparison of search engine to retrieve Persian web pages", *National Conference on Computer Engineering and Sustainable Development Focusing on Computer Networks, Modeling and systems Security*.

Gordon, M. and Pathak, P. (1999), "Finding information on the World Wide Web: the retrieval effectiveness of search engines", *Information Processing and Management*, Vol. 35, pp. 141-180.

Green, D. (2000), "The evolution of Web searching", *Online Information Review*, Vol. 24, pp. 124-137.

Griesbaum, J. and Spink, A. (2004), "Evaluation of three German search engines: Altavista. de, Google. de and Lycos. de", *Information Research*, Vol. 9.

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

Harter, S.P. (1996), "Variations in relevance assessments and the measurement of retrieval effectiveness", *Journal of the American Society for Information Science*, Vol. 47, pp. 37-49.

Hawking, D., Craswell, N., Thistlewaite, P. and Harman, D. (1999), "Results and challenges in web search evaluation", *Computer Networks*, Vol. 31, pp. 1321-1330.

Hayati, Z. and Alijani, R. (2012), "The web search engines and general reference questions", *International Journal of Integrated Supply Management*, Vol. 3, pp. 18-32.

Hjørland, B. (2010), "The foundation of the concept of relevance", *Journal of the American Society for Information Science and Technology*, Vol. 61, pp. 217-237.

Hoffman, J.M. (1965), *Experimental Design for Measuring the Intra-and Inter-group Consistency of Human Judgment of Relevance*, Georgia Institute of Technology.

Hou, J. (2009), "Research on design of an automatic evaluation system of search engine", *ETP International Conference on Future Computer and Communication*, 2009, IEEE, pp. 16-18.

Hripcsak, G. and Rothschild, A.S. (2005), "Agreement, the f-measure, and reliability in information retrieval", *Journal of the American Medical Informatics Association*, Vol. 12, pp. 296-298.

Huang, X. and Soergel, D. (2013), "Relevance an improved framework for explicating the notion", *Journal of the American Society for Information Science and Technology*, Vol. 64, pp. 18-35.

Isfandyari Moghaddam, A. and Parirokh, M. (2006), "A comparative study on overlapping of search results in metasearch engines and their common underlying search engines", *Library Review*, Vol. 55, pp. 301-306.

Jain, A., Nandakumar, K. and Ross, A. (2005), "Score normalization in multimodal biometric systems", *Pattern Recognition*, Vol. 38, pp. 2270-2285.

Jain, Y.K. and Bhandare, S.K. (2011), "Min max normalization based data perturbation method for privacy protection", *International Journal of Computer and Communication Technology*, Vol. 2, pp. 45-50.

Jansen, B.J. and Molina, P.R. (2006), "The effectiveness of Web search engines for retrieving relevant ecommerce links", *Information Processing and Management*, Vol. 42, pp. 1075-1098.

Järvelin, K. (2007), "An analysis of two approaches in information retrieval: from frameworks to study designs", *Journal of the American Society for Information Science and Technology*, Vol. 58, pp. 971-986.

Järvelin, K. and Kekäläinen, J. (2002), "Cumulated gain-based evaluation of IR techniques", *ACM Transactions on Information Systems*, Vol. 20, pp. 422-446.

Kelly, D., Arguello, J., Edwards, A. and Wu, W.-C. (2015), "Development and evaluation of search tasks for IIR experiments using a cognitive complexity framework", *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*, ACM, pp. 101-110.

Kent, A., Berry, M.M., Luehrs, F.U. Jr and Perry, J.W. (1955), "Operational criteria for designing information retrieval systems", *American Documentation*, Vol. 6, pp. 93-101.

Kerkmann, F. and Lewandowski, D. (2012), "Accessibility of web search engines", *Library Review*, Vol. 61, pp. 608-621.

Kumar, B. and Prakash, J. (2009a), "Precision and relative recall of search engines: a comparative study of Google and Yahoo", *Singapore Journal of Library and Information Management*, Vol. 38, pp. 124-137.

Kumar, B.S. and Prakash, J. (2009b), "Precision and relative recall of search engines: a comparative study of Google and Yahoo", *Singapore Journal of Library and Information Management*, Vol. 38, pp. 124-137.

Kumar, B.T. and Sampath Pavithra, S.M. (2010), "Evaluating the searching capabilities of search engines and metasearch engines: a comparative study", *Annals of Library and Information Studies*, Vol. 57, pp. 87-97.

Kunder, D.M. (2019), "The size of the world wide web", available at: https://www.worldwidewebsize.com/.

Lancaster, F. (1971), "An evaluation of ears (epilepsy abstracts retrieval system) and factors governing its effectiveness", Report to N1NDS, University of Illinois.

Lawrence, S. and Giles, C.L. (1999), "Accessibility of information on the web", *Nature*, Vol. 400, p. 107.

Leighton, H.V. and Srivastava, J. (1999), "First 20 precision among World Wide Web search services (search engines)", *Journal of the American Society for Information Science*, Vol. 50, pp. 870-881.

Lewandowski, D. (2008a), "Problems with the use of web search engines to find results in foreign languages", *Online Information Review*, Vol. 32, pp. 668-672.

Lewandowski, D. (2008b), "The retrieval effectiveness of web search engines: considering results descriptions", *Journal of Documentation*.

Lewandowski, D. (2011), "The retrieval effectiveness of search engines on navigational queries", *Aslib Proceedings: New Information Perspectives*, Emerald Group Publishing, pp. 354-363.

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*, Vol. 66, pp. 1763-1775.

Liu, B. (2011), *User Personal Evaluation of Search Engines-Google, Bing and Blekko*, University of Illinois at Chicago.

Luyt, B., Goh, D. and Sian Lee, C. (2009), "Searching locally: a comparison of Yehey! and google", *Online Information Review*, Vol. 33, pp. 499-510.

Ma, L. (2012), "Some philosophical considerations in using mixed methods in library and information science research", *Journal of the American Society for Information Science and Technology*, Vol. 63, pp. 1859-1867.

MacFarlane, A. (2007), "Evaluation of web search for the information practitioner", *Aslib Proceedings*, Emerald Group Publishing, pp. 352-366.

Mahmoudi, M., Badie, R., Zahedi, M.S. and Azimzadeh, M. (2014), "Evaluating the retrieval effectiveness of search engines using Persian navigational queries", *7'th International Symposium on Telecommunications (IST'2014)*, IEEE, pp. 563-568.

Mirgood, S.H., Ghiasi, M., Daliri, S., Kouchakinejad, E.I.and Abbasian, J.A. (2015), "A comparison of accuracy in specialized medical search and general search engines for retrieving medical image", *Educational Development of Jundishapur*, Vol. 6 No. 2, pp. 131-138.

Moghaddam, A.I. and Parirokh, M. (2006), *A Comparative Study on Overlapping of Search Results in Metasearch Engines and Their Common Underlying Search Engines*, Library Review.

Morvarid, N., Behzadi, H. and Raddad, I. (2016), "Qualitative ranking of Persian and non-Persian search engines in information retrieval of Islamic subjects", *Library and Information Science*, Vol. 19, pp. 44-77.

Nowkarizi, M. and Zeynali Tazehkandi, M. (2017), "The overlap and coverage of 4 local search engines: Parsijoo, Yooz, Parseek and Rismoun", *Human Information Interaction*, Vol. 4, pp. 48-59.

Nuray, R. and Can, F. (2006), "Automatic ranking of information retrieval systems using data fusion", *Information Processing and Management*, Vol. 42, pp. 595-614.

Oppenheim, C., Morris, A., McKnight, C. and Lowley, S. (2000), "The evaluation of WWW search engines", *Journal of Documentation*, Vol. 56, pp. 190-211.

Poulter, A. (1997), "The design of World Wide Web search engines: a critical review", *Program*, Vol. 31, pp. 131-145.

Powers, D.M. (2011), "Evaluation: From precision, recall and f-measure to ROC, informedness, markedness and correlation", *Journal of Machine Learning Technologies*, Vol. 2 No. 1, pp. 37-63.

Resnick, A. (1961), "Relative effectiveness of document titles and abstracts for determining relevance of documents", *Science*, Vol. 134, pp. 1004-1006.

Resnick, A. and Savage, T. (1964), "The consistency of human judgments of relevance", *American Documentation*, Vol. 15, pp. 93-95.

Robinson, M.L. and Wusteman, J. (2007), "Putting Google Scholar to the test: a preliminary study", *Program*, Vol. 41, pp. 71-80.

Sadeghi, H. (2011), "Automatic performance evaluation of web search engines using judgments of metasearch engines", *Online Information Review*, Vol. 35, pp. 957-971.

Salton, G. and McGill, M. (1983), *Introduction to Modern Information Retrieval.*

Sampath Kumar, B. and Kumar, G. (2013), "Search engines and their search strategies: the effective use by Indian academics", *Program*, Vol. 47, pp. 437-449.

Sampath Kumar, B. and Prakash, J. (2009), "Precision and relative recall of search engines: a comparative study of Google and Yahoo", *Singapore Journal of Library and Information Management*, Vol. 38, pp. 124-137.

Saracevic, T. (1995), "Evaluation of evaluation in information retrieval", *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, Citeseer, pp. 138-146.

Saracevic, T. (2007), "Relevance: a review of the literature and a framework for thinking on the notion in information science. Part III: behavior and effects of relevance", *Journal of the American Society for Information Science and Technology*, Vol. 58, pp. 2126-2144.

Saracevic, T. (2012), "Research on relevance in information science: a historical perspective. International perspectives on the history of information science and technology", *Proceedings of the ASIS&T*, pp. 49-60.

Saracevic, T. (2015), "Why is relevance still the basic notion in information science. Re: inventing information science in the networked society", *Proceedings of the 14th International Symposium on Information Science (ISI 2015)*, pp. 26-35.

Saracevic, T. and Kantor, P. (1988), "A study of information seeking and retrieving. III. Searchers, searches, and overlap", *Journal of the American Society for Information Science*, Vol. 39, pp. 197-216.

Shang, Y. and Li, L. (2002), "Precision evaluation of search engines", *World Wide Web*, Vol. 5, pp. 159-173.

Shi, Z., Wang, B., Li, P. and Shi, Z. (2010), "Using global statistics to rank retrieval systems without relevance judgments", *International Conference on Intelligent Information Processing*, Springer, pp. 183-192.

Smith, A.G. (2003), "Think local, search global? Comparing search engines for searching geographically specific information", *Online Information Review*, Vol. 27, pp. 102-109.

Spoerri, A. (2007), "Using the structure of overlap between search results to rank retrieval systems without relevance judgments", *Information Processing and Management*, Vol. 43, pp. 1059-1070.

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

Su, L.T. and Chen, H.-l. (1999), "Evaluation of web search engines by undergraduate students", *Proceedings of the ASIS Annual Meeting*, ERIC, pp. 98-114.

Tague-Sutcliffe, J.M. (1996), "Some perspectives on the evaluation of information retrieval systems", *Journal of the American Society for Information Science*, Vol. 47, pp. 1-3.

Tawileh, W., Mandl, T., Griesbaum, J., Atzmueller, M., Benz, D., Hotho, A. and Stumme, G. (2010), "Evaluation of five web search engines in Arabic language", *LWA*, pp. 221-228.

Teixeira Lopes, C. and Ribeiro, C. (2011), "Comparative evaluation of web search engines in health information retrieval", *Online Information Review*, Vol. 35, pp. 869-892.

The Search Engines List(2010), "The Search Engine List: the comprehensive list of search engines", available at: http://www.thesearchenginelist.com/.

Thornley, C. and Gibb, F. (2007), "A dialectical approach to information retrieval", *Journal of Documentation*, Vol. 63, pp. 755-764.

Thornley, C.V. (2005), *A Dialectical Approach to Information Retrieval: Exploring a Contradiction in Terms*.

Vakkari, P. and Järvelin, K. (2005), *Explanation in Information Seeking and Retrieval. New Directions in Cognitive Information Retrieval*, Springer, pp. 113-138.

Van Rijsbergen, C.J. (1974), "Foundation of evaluation", *Journal of Documentation*, Vol. 30, pp. 365-373.

Vickery, B.C. (1959), "The structure of information retrieval systems", *Proceedings of the International Conference on Scientific Information*, pp. 1275-1290.

Vickery, B.C. (1970), *Techniques of Information Retrieval.*

Voorhees, E.M. (2001), "The philosophy of information retrieval evaluation", *Workshop of the Cross-Language Evaluation Forum for European Languages*, Springer, pp. 355-370.

Wilson, T.D. (2000), "Recent trends in user studies: action research and qualitative methods", *Information Research*, Vol. 5, p. 76.

Wilson, T.D. (2002), "Philosophical foundations and research relevance: issues for information research (Keynote address)", *proceedings of Fourth International Conference on Conceptions of Library and Information Science: Emerging Frameworks and Method*, University of Washington, Seattle, USA.

Wu, S. and Crestani, F. (2003), "Methods for ranking information retrieval systems without relevance judgments", *Proceedings of the 2003 ACM Symposium on Applied Computing*, pp. 811-816.

Wu, S. and Li, J. (2004), "Effectiveness evaluation and comparison of Web search engines and meta-search engines", *International Conference on Web-Age Information Management*, Springer, pp. 303-314.

Wu, S., Zhang, Z. and Xu, C. (2019), "Evaluating the effectiveness of Web search engines on results diversification", *Information Research: An International Electronic Journal*, Vol. 24 No. 1.

Xie, H. (2004), "Online IR system evaluation: online databases versus Web search engines", *Online Information Review*, Vol. 28, pp. 211-219.

Yaltaghian, B. and Chignell, M. (2002), "How good is search engine ranking? a validation study with human judges", *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, SAGE Publications Sage CA, Los Angeles, CA, pp. 1276-1280.

Zeynali Tazehkandi, M. and Nowkarizi, M. (2020), "A dialectical approach to search engine evaluation", *Libri - International Journal of Libraries and Information Services*, Vol. 70.

Zhang, J. and Fei, W. (2010), "Search engines' responses to several search feature selections", *The International Information and Library Review*, Vol. 42, pp. 212-225.

Zhang, J., Fei, W. and Le, T. (2013), "A comparative analysis of the search feature effectiveness of the major English and Chinese search engines", *Online Information Review*, Vol. 37, pp. 217-230.

Further reading

Budd, J.M. (2001), *Knowledge and Knowing in Library and Information Science: A Philosophical Framework*, Scarecrow Press, Boson Way, Lanham, MaryLand.

Appendix

Row	Subject headings	Row	Subject headings
1	Symptoms of pregnancy	17	Introduction to the geography of Saudi Arabia
2	Introducing Iran Khodro Company	18	Ways to prevent ozone layer destruction
3	Cross-cultural communication	19	Establish a close relationship with the child
4	Impact of climatic conditions on physical properties	20	Employment of people with disabilities
5	Introducing Iranshahr	21	Postmodern architecture
6	Types of xylophone	22	Introducing the book "Fei Ma Fei"
7	Causes of seizures in children	23	The pioneers of spirituality
8	How to write a resume	24	Iranian social unrest
9	Pro-peace movements	25	Purpose, types, and time of Retreat
10	Side-effects of cannabis use	26	Accidents, casualties, and problems arising from the air bombardment of the Iran–Iraq war
11	Symptoms of circulatory disorder	27	Organize file on computer
12	The importance and benefits of tree planting	28	Introducing Shiraz tourist places
13	Learn Turkish	29	The concept, types, ways, and obstacles of theology
14	Types of pottery and pottery styles in Iran	30	The materials needed for better spinach growth
15	The principle of harm and loss	31	Children cartoon summary
16	Introducing children filmmakers	32	Introducing Organization of Islamic Countries Summit

Table AI.
Subject heading

Corresponding author
Mahdi Zeynali Tazehkandi can be contacted at: ma.zeynali@mail.um.ac.ir; Mohsen Nowkarizi can be contacted at: mnowkarizi@um.ac.ir