Three approaches to measuring recall on the web: A systematic review

Article *in* The Electronic Library · May 2020

DOI: 10.1108/EL-12-2019-0287

2 authors:

Mahdi Zeynali Tazehkandi
Ferdowsi University Of Mashhad
20 PUBLICATIONS **8** CITATIONS
SEE PROFILE

Mohsen Nowkarizi
Ferdowsi University Of Mashhad
74 PUBLICATIONS **43** CITATIONS
SEE PROFILE

Three approaches to measuring recall on the Web: a systematic review

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

Received 20 December 2019
Revised 2 March 2020
29 March 2020
7 May 2020
Accepted 14 May 2020

Abstract

Purpose – The purpose of this paper is to present a review on the use of the recall metric for evaluating information retrieval systems, especially search engines.

Design/methodology/approach – This paper investigates different researchers' views about recall metrics.

Findings – Five different definitions for recall were identified. For the first group, recall refers to completeness, but it does not specify where all the relevant documents are located. For the second group, recall refers to retrieving all the relevant documents from the collection. However, it seems that the term "collection" is ambiguous. For the third group (first approach), collection means the index of search engines and, for the fourth group (second approach), collection refers to the Web. For the fifth group (third approach), ranking of the retrieved documents should also be accounted for in calculating recall.

Practical implications – It can be said that in the first, second and third approaches, the components of the retrieval algorithm, the retrieval algorithm and crawler, and the retrieval algorithm and crawler and ranker, respectively, are evaluated. To determine the effectiveness of search engines for the use of users, it is better to use the third approach in recall measurement.

Originality/value – The value of this paper is to collect, identify and analyse literature that is used in recall. In addition, different views of researchers about recall are identified.

Keywords Information retrieval, Webometrics, Search engine effectiveness, Relevance judgement, Evaluation metrics

Paper type Literature review

1. Introduction

Evaluating an information retrieval system, especially a search engine, is one of the fundamental topics in the field of library and information science. In this regard, various researchers, such as Bar-Ilan (1998), Bilal (2012), Deka and Lahkar (2010), Demirci *et al.* (2007), Fattahi *et al.* (2016), Gordon and Pathak (1999), Hussain *et al.* (2019), Lewandowski (2015), Wani Zahid and Ahmad Sofi (2016) and Zeynali Tazehkandi and Nowkarizi (2020), evaluated search engines.

The authors thank Professor Jeonghyun (Annie) Kim, Department of Information Science, University of North Texas; Professor John M. Budd, Professor Emeritus, School of Information Science and Learning Technologies, College of Education, University of Missouri; Mr Iman Tahamtan, Graduate Research Assistant, School of Information Sciences, College of Communication and Information, University of Tennessee; Mr Mohammad Dolati, PhD student of Plant Biotechnology, Ferdowsi University of Mashhad; and Anonymous Reviewers for valuable comments that greatly improved the manuscript.

The Electronic Library
Vol. 38 No. 3, 2020
pp. 477-492

0264-0473
DOI 10.1108/EL-12-2019-0287

Generally, there are two steps to evaluate search engines: selecting an evaluation approach and a metric. Some researchers, such as Ali and Beg (2011) and Clough and Sanderson (2013), have addressed the two steps. Others, such as Budd (2004), Cooper (1971) and Hjørland (2010), have studied different approaches to evaluation. While others, among them Baccini *et al.* (2012), Cleverdon (1970), Magdy and Jones (2010) and Sirotkin (2013), have investigated various evaluation metrics.

The results of the evaluation of information retrieval studies are influenced by the two steps mentioned. Both groups of these studies are significant. However, the purpose of the present study is to present a systematic review of the literature concerning the different views of researchers on recall because Saracevic (2015) stated that recall is one of the standard evaluation metrics, the title of which, however, is ambiguous. Also, Budd (2001) and Hjørland (2005) emphasized that the methods and techniques used in every type of research are not independent of the theoretical foundations which determine the way researchers operate. Thus, there is a research gap in both the descriptions and different formulas of recall and their relationship. In addition, Budd (2001) stated that progress will occur, provided that the workings of every field rely on knowing where it came from, how it got here and where it is going to, and are deeply and critically investigated. Therefore, the following questions are designed to conduct such a study on the recall metric:

Q1. What are the different definitions of recall?

Q2. What are the different formulas of recall?

Q3. What is the corresponding definition of each of the formulas?

Q4. Which of the definitions and formulas is consistent with the origin (etymology and the designer's viewpoint) of recall?

Q5. In each of the existing formulas, what components of the search engines are measured?

2. Methodology

This review article has followed the rules of systematic review (Khan *et al.*, 2003). Search terms, such as "recall metric", "recall measure", "evaluation metrics", "evaluation measure", "search engine evaluation", "information retrieval system evaluation", "relevance evaluation measure" and "relevance evaluation metrics", were used to identify studies conducted in this subject area. Electronic databases – Emerald Publishing, Google Scholar, Sage Publishing and Springer – were used in the search for documents.

The searches were carried out between February and March 2019 and 3,200 documents (8 queries * 4 databases * 100 first results) were identified. The relevance of the articles was judged based on the title and the abstract. To determine the relevance of the books, their tables of contents were considered. Overall, about 85 documents were used for this review. It should be noted that the documents selected were only in the English language, regardless of when they were published. Figure 1 provides more information about each of these documents.

As can be seen in Figure 1, the relevant documents were grouped based on their years of publication. For each year, the author's (or authors') name as a citation was included in alphabetical order based on the first author's name. Finally, a total number of documents assigned under each year was included. In addition, an analysis of the literature showed that they fell into different types of publications, such as journal articles, conference proceeding articles, book chapters and so on (Figure 2).

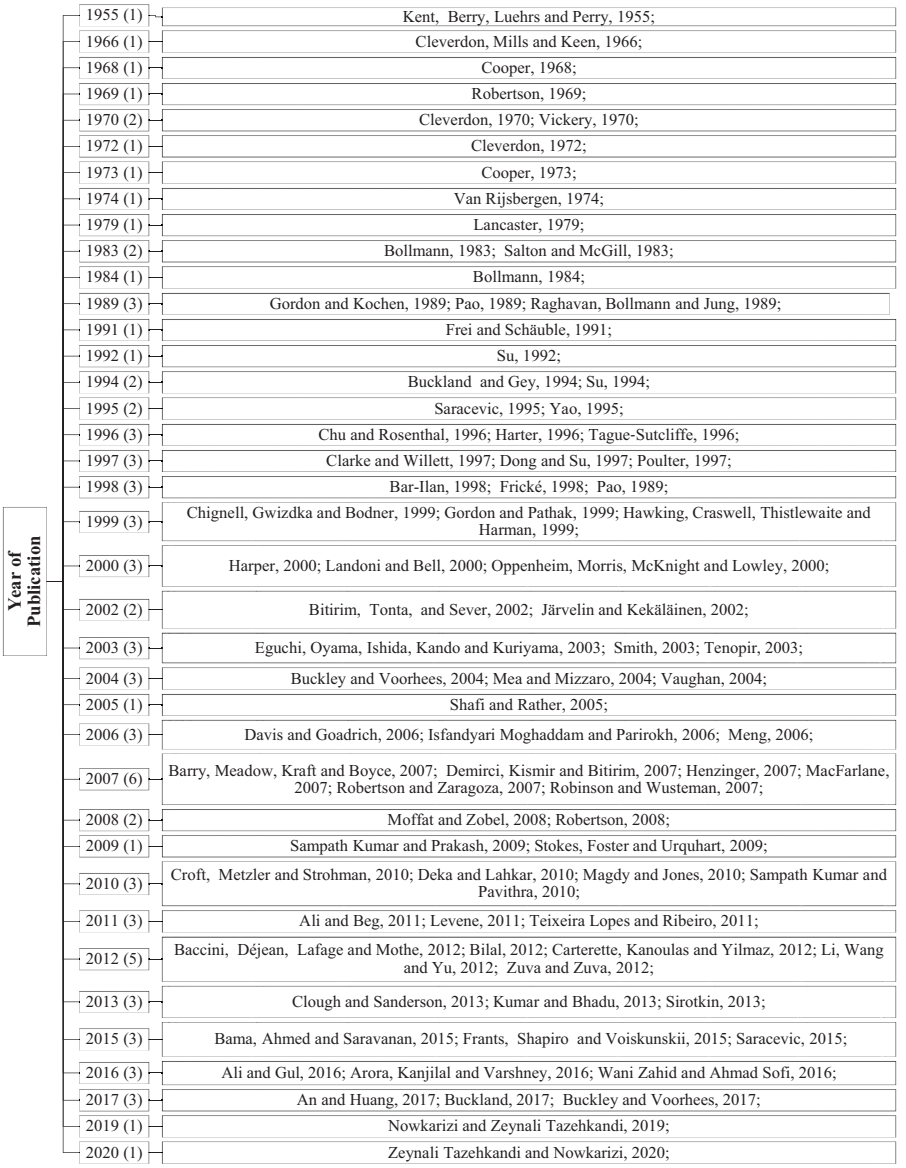

Figure 1.
Number of identified relevant documents per year

As can be seen in Figure 2, the identified documents were grouped based on their type of publication. For each journal article, the journal's name and, for each conference paper, the name of the conference and so on, their publishers were presented in alphabetical order. Different types of publications were identified, but the majority of the relevant literature fell into the journal article type (55 items). Finally, it should be noted that to answer the research

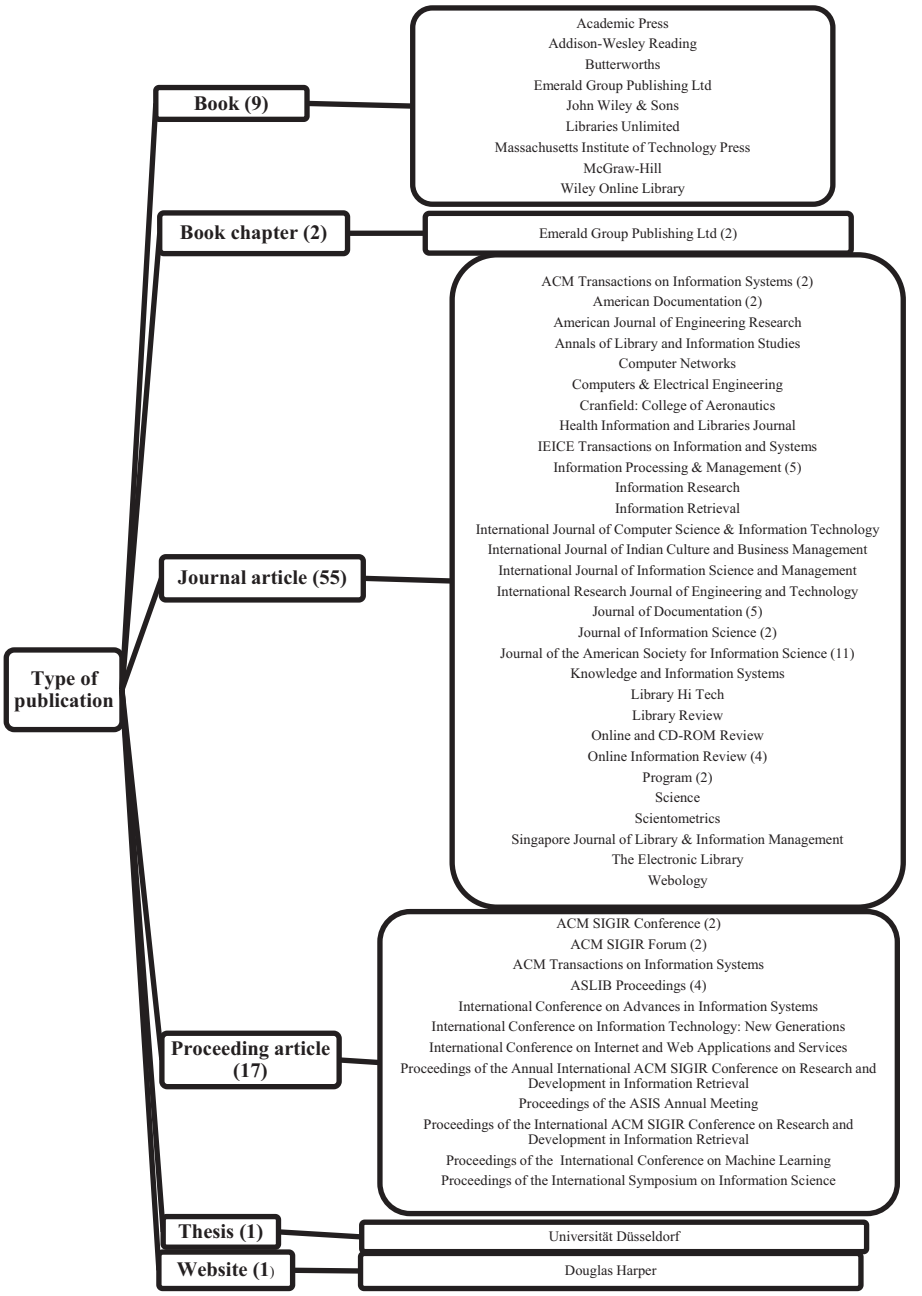

Figure 2.
Type of publication of the identified documents

questions, each literature was then assigned to each question; wherever a document was related to more than one, it was assigned to more than one question (Figure 3).

As can be seen in Figure 3, the documents were categorized based on their relationship to the questions mentioned. For each group, the documents were arranged according to their date of publication.

3. Findings

3.1 What are the different definitions of recall?

As to the description of recall, the first group, such as Buckland (2017), Buckland and Gey (1994) and Croft *et al.* (2010), refer to the completeness or all the relevant documents, but it is not specified where all the relevant documents are located. For the second group, such as Barry *et al.* (2007), Lewandowski (2015), Shafi and Rather (2005) and Zuva and Zuva (2012), the word "collection" and "file" are added to the first description, but remain ambiguous. Hence, this question arises as to what the "collection" and "file" refer? This question derived from the descriptions of the second group is answered in the descriptions of the third and the

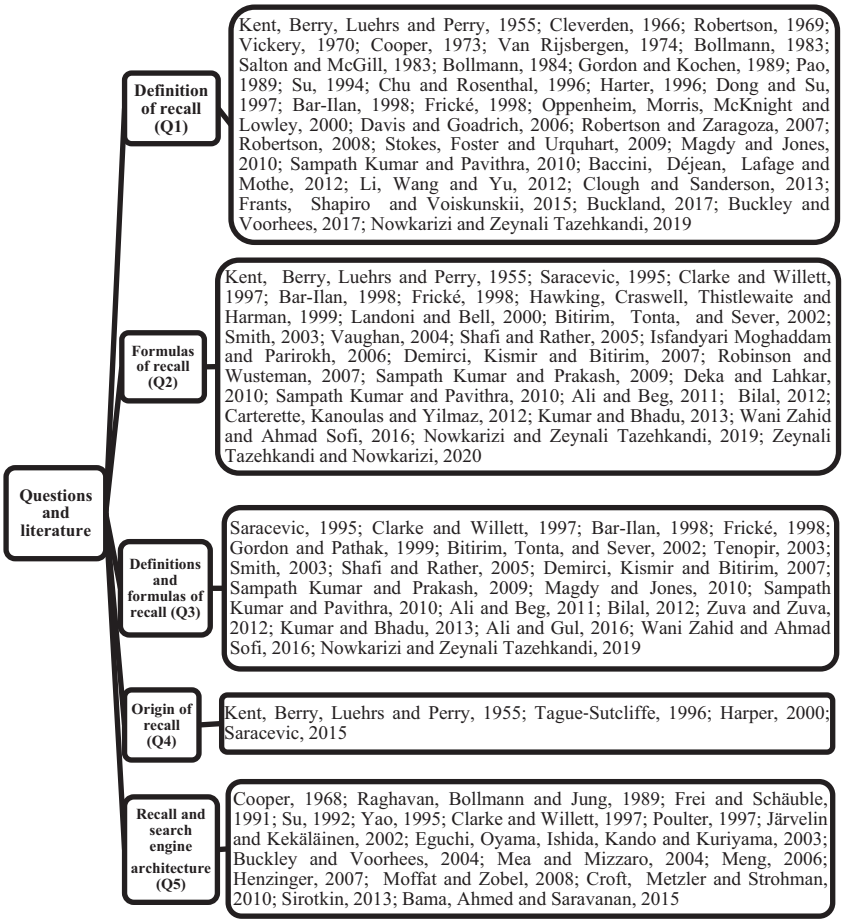

Figure 3.
Documents assigned to answer the questions

fourth groups. In the description of the third group, such as Bilal (2012), Clarke and Willett (1997), Lancaster (1979) and Nowkarizi and Zeynali Tazehkandi (2019), all the relevant documents are in the index of the search engine, while in the fourth group, such as Bar-Ilan (1998), Chu and Rosenthal (1996), Frické (1998) and Robinson and Wusteman (2007), the Web is the collection where all the relevant documents are located. In the fifth group, such as Bitirim *et al.* (2002), Bollmann (1983) and Yao (1995), recall considers that relevant documents are retrieved before irrelevant ones.

3.2 What are the different formulas of recall?

Different researchers have used different formulas to calculate the recall ratio of search engines, which are stated as follows. For the first group of researchers, such as Clarke and Willett (1997) and Nowkarizi and Zeynali Tazehkandi (2019), recall is the ratio of the number of relevant documents retrieved to the total number of relevant documents that exists in the index or document data store of a search engine which is calculated by the following formula:

$$Recall = \frac{\text{number of documents retrieved by a search engine}}{\text{total number of relevant documents in the index's search engine}} \quad \text{(Formula 1)}$$

A number of researchers, such as Bar-Ilan (1998), Kumar and Bhadu (2013), Kumar and Pavithra (2010), Sampath Kumar and Prakash (2009), Shafi and Rather (2005) and Usmani *et al.* (2012), have used other formulas. For these researchers, in the denominator of the recall formula, the relevant documents available on the Web are considered (Bar-Ilan, 1998), while in Formula 1, only the relevant documents available in the index or document data store of the search engine appear in the denominator. In other words, these researchers calculate the recall metric through the following formula:

$$Recall = \frac{\text{number of documents retrieved by a search engine}}{\text{total number of relevant documents available on the web}} \quad \text{(Formula 2)}$$

It should be noted that in this formula two or more search engines are examined to determine how many documents are relevant to a particular topic on the Web in search engine evaluation studies. Hence, the number of relevant documents retrieved by all search engines related to a particular topic assumes all the relevant documents available on the web. In this respect, Bar-Ilan (1998) and Kumar and Pavithra (2010) have used this method to calculate the total number of relevant documents available on the web, but it is clear that a researcher in his/her research cannot search a subject or query in all search engines to obtain all the documents available on the web by aggregating the relevant documents found in the search engines. In this regard, Frické (1998) believed that recall is a relative metric and cannot be accurately calculated. However, Su (1994) stated that sometimes users seek to dominate the subject and tend to access the maximum number of relevant documents available on the Web; this may be considered as comprehensiveness as denominated by Nowkarizi and Zeynali Tazehkandi (2019).

Another important point is that, in mathematics, the total of the elements of two sets A and B is calculated by the union formula (total of two sets A and B = $|A| + |B| - |A \cap B|$), not a plus formula (total of two sets A and B $\neq |A| + |B|$) (Hazewinkel, 2001). So, there is no need to use the plus formula to calculate all the relevant documents, but rather, if one

document is retrieved by several search engines, it should be calculated only once and use the union formula.

In the third group, both the denominator and numerator are different from the other two groups mentioned so it is calculated by the following formula:

$$R_{norm} = \frac{1}{2}\left(1 + \frac{R^{+} - R^{-}}{D}\right) \quad \text{(Formula 3)}$$

This formula is called the *normalized recall* that was defined by Rocchio in 1964 (Bitirim *et al.*, 2002; Bollmann, 1983; Yao, 1995) where R^{+} stands for the number of document pairs in which a relevant document is ranked higher than a non-relevant document, R^{-} stands for the number of document pairs in which a non-relevant document is ranked higher than a relevant one, and D stands for the maximal number of $R+$ or the number of documents (Bitirim *et al.*, 2002).

3.3 What is the corresponding definition of each of the formulas?

Five groups of descriptions were identified for recall. Budd (2001) and Hjørland (2005) emphasized the theoretical and philosophical foundations which determine the type of research methods and techniques. Thus, it could be said that the different descriptions proposed for each metric such as recall have resulted in using different formulas for its measurement; however, unclear theoretical descriptions are assumed not to result in corresponding formulas. It can be concluded that the descriptions of the first and the second groups are ambiguous to the readers, but their suppliers eventually agree with one of the descriptions of the third, fourth and fifth groups which have their own corresponding formulas; therefore, there will be three different formulas (at a practical level) for the recall. In other words, any formula used by the researchers to evaluate the search engines was rooted in the descriptions already provided. Hence, the formulas presented above have their corresponding descriptions. Of course, the descriptions of the first and second groups may be eliminated, as the first and second descriptions are not clear and may be incorporated into other descriptions. It can be said that the first, second and third formulas are derived from the third, fourth and fifth descriptions, respectively. Also, from now on, the third definition and the first formula, the fourth definition and the second formula, the fifth definition and the third formula are called the first, second and third approaches, respectively.

3.4 Which of the definitions and formulas is consistent with the origin (etymology and the designer's viewpoint) of recall?

When humans use language and terms, they perform a type of act, with the intention of accomplishing something (Capurro and Hjørland, 2003). In this regard, Nowkarizi and Zeynali Tazehkandi (2019) paid attention to the etymology of the word "recall". Recall consists of two parts: *re* and *call* where the prefix "re" means "again" and refers to repeating an action, and the verb "call" refers to the act of calling or finding (Online Etymology Dictionary, 2018). Therefore, recall means re-finding. As for the structure of the search engine, it has been determined that a crawler initially found the websites based on the search engine indexing policy (find, search, or call action). Finally, the indexing words and the corresponding links to the documents are stored in the search engine database. When a user enters a query into the search field, the documents retrieved by the crawler (the documents available in the search engine index or database) – not those available on the Web – are searched and the documents relevant to the user's query are displayed (re-find or re-call action). Likewise, the recall metric does not focus on the web; but rather it focuses on

the search engine database; that is, the document available in the search engine database or index. As such, Clarke and Willett (1997) and Nowkarizi and Zeynali Tazehkandi (2019) considered the first approach as an appropriate one for recall due to the nature of the word, based on the etymology of recall (Online Etymology Dictionary, 2018). Similarly, Ali and Beg (2011) stated that the study of Clarke and Willett (1997) is a good example of evaluation studies on recall in which the method of calculating recall is realistic. Furthermore, Budd (2001) emphasized that it is important to know where everything came from. This refers to the origin of a phenomenon. Recall was initially introduced by Kent *et al.* (Saracevic, 2015; Tague-Sutcliffe, 1996). According to Kent *et al.* (1955) information retrieval should be defined as the process of identifying what documents or records contain information of pertinent interest in a given information query. Thus, they provide a metric as the fraction of the pertinent documents to which the system directed attention:

$$\frac{w}{x} = \text{fraction of the pertinent documents to which the system directed attention}$$

In this formula, w stands for the number of documents of actual pertinent interest by personal inspection of the documents selected by a particular system and x represents the number of documents of pertinent interest from among all the documents embraced by a particular system. Hence, the first approach to recall is consistent with the idea of Kent *et al.* (1955) as well as to the origin of the recall metric.

3.5 In each of the existing formulas, what components of the search engines are measured?
If the different formulas are used to evaluate search engines, then different components of search engines will be evaluated (Buckley and Voorhees, 2004). To reveal this, suppose that the following documents are available about a subject on the Web (Figure 4).

Also, suppose that the crawler of search engine A collects the documents 1, 0.95, 0.9, 0.75, 0.6, 0.7 and 0.8 and stores them into its document data store. A user submits a query and a ranked list of documents is displayed: for example, 0.6, 0.95 and 0.9. Also, suppose that the crawler of search engine B collects the documents 0.35, 0.75, 0 and 0.65 and stores them into its database. A user submits a query into search engine B and a ranked list of documents are displayed to the user: for example, 0 and 0.75. Now, if Formula 1 is used to calculate the recall ratio of the two search engines, it will be as follows (Clarke and Willett, 1997):

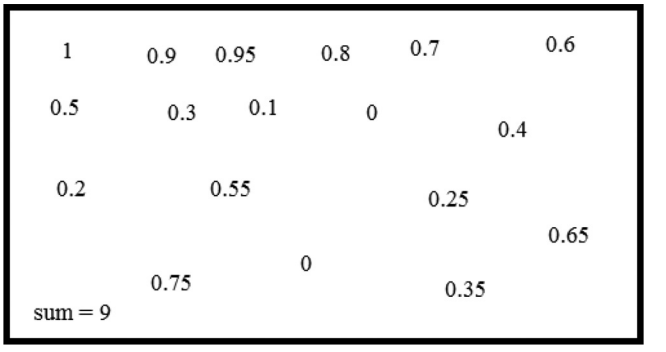

Figure 4.
The Web

$$Recall = \frac{relevance\ score\ of\ the\ retrieved\ documents}{relevance\ score\ of\ the\ documents\ in\ the\ index\ or\ document\ data\ store}$$

$$Recall\ of\ search\ engine\ A = \frac{0.6 + 0.95 + 0.9}{1 + 0.95 + 0.9 + 0.75 + 0.6 + 0.7 + 0.8} = \frac{2.45}{5.7} = 0.42$$

$$Recall\ of\ search\ engine\ B = \frac{0.75}{0.35 + 0.75 + 0.65 + 0} = \frac{0.75}{1.75} = 0.42$$

The above example clearly shows that although the two search engine crawlers have retrieved different documents, they have produced equal recall scores. It is because this approach does not evaluate the crawler performance in calculating recall. As a result, this approach can be called a system-oriented approach, because what matters to the users is the relevant documents on the Web – whether indexed by a search engine or not (Frické, 1998). Clarke and Willett (1997) assumed that a query is searched using the two search engines A and B, and that these searches retrieve the relevant documents of a and b, respectively. They further assumed that the two sets of documents do not overlap so that the total pool containing a + b is relevant. They explained that it cannot immediately conclude that the recall of A and B are $a/(a+b)$ and $b/(a+b)$, respectively, because some of the b relevant documents retrieved by search engine B may not have been available to search engine A, and vice versa. Accordingly, it may obtain a figure for recall performance of A only after checking how many of this b relevant could have been retrieved by A. In other words, how many of b relevant documents were processed by the crawler of search engine A. If these documents are named c ($c < b$), then the true recall for A is $a/(a+c)$.

If Formula 2 is used to calculate the recall ratio of the two search engines noted, it will be as follows (Bar-Ilan, 1998):

$$Recall = \frac{\text{relevance score of the retrieved documents}}{\text{relevance score of documents on the web}}$$

$$Recall\ of\ search\ engine\ A = \frac{0.6 + 0.95 + 0.9}{1 + 0.9 + 0.95 + \cdots + 0.35} = \frac{2.45}{9} = 0.27$$

$$Recall\ of\ search\ engine\ B = \frac{0.75}{1 + 0.9 + 0.95 + \cdots + 0.35} = \frac{0.75}{9} = 0.08$$

This approach accounts for the number of relevant documents available on the Web in the denominator of the recall formula. Nowkarizi and Zeynali Tazehkandi (2019) stated that the two concepts of "recall" and "comprehensiveness" were assumed synonymous in the literature of information retrieval, but comprehensiveness is different from recall as explained by Clarke and Willett (1997). Literally, comprehensiveness means to cover the full and wide [1], to include and attend everything or nearly all elements and aspects of something [2] to include necessary components, to include most parts and aspects of something [3].

Three words are notable in the above descriptions: what do "components", "all" and "necessary" mean? Obviously, in information retrieval, the focus is on the relevant

documents to which the concept of "components" also refers. Thus, "all the components" means "all the relevant documents" available on the Web and not in the search engine's document data store or index. In this approach, the performance of the crawler is also evaluated. However, the weakness of Formula 2 is that it does not take into account the ranker performance. As the documents are displayed to the user in any order, the recall score does not change. If Formula 3 is used to calculate the recall ratio, the order of the documents will be considered as well. Now, assume that the two search engines A and B display the same documents in a different order; then it will be as follows:

Search engine A: 0.95, 0.6, 0.9
Search engine B: 0.95, 0.9, 0.6

$$R_{norm} = \frac{1}{2}\left(1 + \frac{R^{+} - R^{-}}{D}\right)$$

(Bitirim *et al.*, 2002)

$$\textit{Recall of search engine } A = \frac{1}{2}\left(1 + \frac{1-1}{3}\right) = \frac{1}{2}\left(1 + \frac{0}{3}\right) = 0.5$$

$$\textit{Recall of search engine } B = \frac{1}{2}\left(1 + \frac{2-0}{3}\right) = \frac{1}{2}\left(1 + \frac{2}{3}\right) = 0.83$$

As seen in Formula 3, the order of the documents influences the recall score. This formula considers the ranker's performance. It should be noted that the performance of the retrieval algorithm component is evaluated in all of the three formulas, but none of them take the query processor performance into account.

4. Conclusion

Searching in search engines has become an essential part of a user's daily life. Users prefer to access the relevant documents with the least effort – a major goal (Spink and Jansen, 2006; Wolfram *et al.*, 2001). To achieve this major goal, the following two minor goals should be achieved.

The first minor goal is that the search engine should initially provide all the relevant documents to the users. To achieve this minor goal, firstly, the crawler should report the relevant documents to the search engine (crawler function) and, secondly, the retrieval algorithm retrieves all relevant documents available in the search engine's document data store while searching for the users (retrieval algorithm function). In this regard, Lancaster (1979) emphasized that a document should first be covered by the system to retrieve the same document while searching for users.

The next minor goal is that the search engine should display the relevant documents before the irrelevant ones. To reach this minor goal, the ranker should also rank the relevant documents before the irrelevant ones so that users could see the relevant document first. It can be concluded that from the user's point of view, an appropriate formula for recall should be able to measure the performance of all three of the components of retrieval algorithm, crawler and ranker and not merely one of them. For this reason, the third approach may be labelled as a user-oriented approach because users prefer to access the relevant documents and display the relevant documents before the irrelevant ones (Jansen *et al.*, 2000; Spink and Jansen, 2006; Wolfram *et al.*, 2001). It should

also be noted that new metrics in information retrieval, such as Bpref by Buckley and Voorhees (2004), Normalized Discounted Cumulative Gain by Järvelin and Kekäläinen (2002) and average distance measure by Mea and Mizzaro (2004), have been designed to measure the three components of search engines which were mentioned earlier. Therefore, "where everything is going to", as emphasized by Budd (2001) in his book, can be understood. So, it is recommended that researchers use the third approach of recall to evaluate search engines.

Notes

1. Comprehensive (2018), in the Online Cambridge Dictionary, available at: https://dictionary.cambridge.org/dictionary/english/comprehensive (accessed 12 November 2018).
2. Comprehensiveness (2018), in the Online Oxford Dictionary, available at: www.oxfordlearnersdictionaries.com/definition/english/comprehensiveness (accessed 12 November 2018).
3. Comprehensive (2018), in the Online Macmillan Dictionary, available at: www.macmillandictionary.com/dictionary/british/comprehensive (accessed 12 November 2018).

References

Ali, R. and Beg, M.S. (2011), "An overview of web search evaluation methods", *Computers and Electrical Engineering*, Vol. 37 No. 6, pp. 835-848.

Baccini, A., Déjean, S., Lafage, L. and Mothe, J. (2012), "How many performance measures to evaluate information retrieval systems?", *Knowledge and Information Systems*, Vol. 30 No. 3, pp. 693-713.

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 No. 2, pp. 207-228.

Barry, C., Meadow, C.T., Kraft, D.H. and Boyce, B.R. (2007), *Text Information Retrieval Systems*, Academic Press, San Diego, CA.

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 No. 9, 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, Berlin, Heidelberg, pp. 93-103.

Bollmann, P. (1983), "The normalized recall and related measures", *ACM SIGIR Forum*, Association for Computing Machinery, New York, NY, pp. 122-128.

Buckland, M. (2017), *Information and Society*, MIT Press, Cambridge, MA.

Buckland, M. and Gey, F. (1994), "The relationship between recall and precision", *Journal of the American Society for Information Science*, Vol. 45 No. 1, pp. 12-19.

Buckley, C. and Voorhees, E.M. (2004), "Retrieval evaluation with incomplete information", *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, New York, NY*, pp. 25-32.

Budd, J.M. (2001), *Knowledge and Knowing in Library and Information Science: A Philosophical Framework*, Academic Press, New York, NY.

Budd, J.M. (2004), "Relevance: language, semantics, philosophy", *Library Trends*, Vol. 52 No. 3, pp. 447-462.

Capurro, R. and Hjørland, B. (2003), "The concept of information", *Annual Review of Information Science and Technology*, Vol. 37 No. 1, pp. 343-411.

Chu, H. and Rosenthal, M. (1996), "Search engines for the world wide web: a comparative study and evaluation methodology", *Proceedings of the Annual Meeting of the American Society for Information Science*, Vol. 33, pp. 127-135.

Clarke, S.J. and Willett, P. (1997), "Estimating the recall performance of web search engines", *ASLIB Proceedings*, Vol. 49 No. 7, pp. 184-189.

Cleverdon, C.W. (1970), "Evaluation tests of information retrieval systems", *Journal of Documentation*, Vol. 26 No. 1, pp. 55-67.

Clough, P. and Sanderson, M. (2013), "Evaluating the performance of information retrieval systems using test collections", *Information Research*, Vol. 18 No. 2.

Cooper, W.S. (1971), "A definition of relevance for information retrieval", *Information Storage and Retrieval*, Vol. 7 No. 1, pp. 19-37.

Croft, W.B., Metzler, D. and Strohman, T. (2010), *Search Engines: Information Retrieval in Practice*, Pearson, New York, NY.

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 No. 5, pp. 757-771.

Demirci, R.G., Kismir, V. and Bitirim, Y. (2007), "An evaluation of popular search engines on finding turkish document", *Second International Conference on Internet and Web Applications and Services (ICIW '07)*, IEEE, Turkey, pp. 1-5.

Fattahi, R., Parirokh, M., Dayyani, M.H., Khosravi, A. and Zareivenovel, M. (2016), "Effectiveness of Google keyword suggestion on users' relevance judgment", *The Electronic Library*, Vol. 34 No. 2, pp. 302-314.

Frické, M. (1998), "Measuring recall", *Journal of Information Science*, Vol. 24 No. 6, pp. 409-417.

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 No. 2, pp. 141-180.

Hazewinkel, M. (2001), "Union of sets", in Rehmann, U. (Ed.), *Encyclopedia of Mathematics*, 2nd ed., Springer, Berlin.

Hjørland, B. (2005), "Library and information science and the philosophy of science", *Journal of Documentation*, Vol. 61 No. 1, pp. 5-10.

Hjørland, B. (2010), "The foundation of the concept of relevance", *Journal of the American Society for Information Science and Technology*, Vol. 61 No. 5, pp. 217-237.

Hussain, A., Gul, S., Shah Tariq, A. and Shueb, S. (2019), "Retrieval effectiveness of image search engines", *The Electronic Library*, Vol. 37 No. 1, pp. 173-184.

Jansen, B.J., Spink, A. and Saracevic, T. (2000), "Real life, real users, and real needs: a study and analysis of user queries on the web", *Information Processing and Management*, Vol. 36 No. 2, pp. 207-227.

Järvelin, K. and Kekäläinen, J. (2002), "Cumulated gain-based evaluation of IR techniques", *ACM Transactions on Information Systems (Tois)*, Vol. 20 No. 4, pp. 422-446.

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 No. 2, pp. 93-101.

Khan, K.S., Kunz, R., Kleijnen, J. and Antes, G. (2003), "Five steps to conducting a systematic review", *Journal of the Royal Society of Medicine*, Vol. 96 No. 3, pp. 118-121.

Kumar, K. and Bhadu, V. (2013), "A comparative study of BYG search engines", *American Journal of Engineering Research*, Vol. 2 No. 4, pp. 39-43.

Kumar, S.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 No. 1, pp. 124-137.

Kumar, B.T. and 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 No. 2, pp. 87-97.

Lancaster, F.W. (1979), "Information retrieval systems: Characteristics", *Testing, and Evaluation*, 2nd ed., Wiley, New York, NY.

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 No. 9, pp. 1763-1775.

Magdy, W. and Jones, G.J. (2010), "PRES: a score metric for evaluating recall-oriented information retrieval application", *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, Geneva*, pp. 611-618.

Mea, V.D. and Mizzaro, S. (2004), "Measuring retrieval effectiveness: a new proposal and a first experimental validation", *Journal of the American Society for Information Science and Technology*, Vol. 55 No. 6, pp. 530-543.

Nowkarizi, M. and Zeynali Tazehkandi, M. (2019), "Rethinking the recall measure in appraising information retrieval systems and providing a new measure by using Persian search engines", *International Journal of Information Science and Management*, Vol. 17 No. 1, pp. 1-19.

Robinson, M.L. and Wusteman, J. (2007), "Putting Google scholar to the test: a preliminary study", *Program*, Vol. 41 No. 1, pp. 71-80.

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

Shafi, S. and Rather, R.A. (2005), "Precision and recall of five search engines for retrieval of scholarly information in the field of biotechnology", *Webology*, Vol. 2 No. 2, pp. 42-47.

Sirotkin, P. (2013), "On search engine evaluation metrics", Thesis, available at: https://docserv.uni-duesseldorf.de/servlets/DerivateServlet/Derivate-25066/On%20Search%20Engine%20Evaluation%20Metrics%20-%20final.pdf (accessed 12 November 2018).

Spink, A. and Jansen, B.J. (2006), *Web Search: Public Searching of the Web*, Kluwer Academic Publisher, New York, NY.

Su, L.T. (1994), "The relevance of recall and precision in user evaluation", *Journal of the American Society for Information Science*, Vol. 45 No. 3, pp. 207-217.

Tague-Sutcliffe, J.M. (1996), "Some perspectives on the evaluation of information retrieval systems", *Journal of the American Society for Information Science*, Vol. 47 No. 1, pp. 1-3.

Usmani, T.A., Pant, D. and Bhatt, A.K. (2012), "A comparative study of Google and Bing search engines in context of precision and relative recall parameter", *International Journal on Computer Science and Engineering*, Vol. 4 No. 1, p. 21.

Wani, Z.,A. and Ahmad Sofi, A. (2016), "Retrieval efficiency of select search engines vis-à-vis diverse open courseware formats", *The Electronic Library*, Vol. 34 No. 3, pp. 457-470.

Wolfram, D., Spink, A., Jansen, B.J. and Saracevic, T. (2001), "Vox populi: the public searching of the web", *Journal of the American Society for Information Science and Technology*, Vol. 52 No. 12, pp. 1073-1074.

Yao, Y. (1995), "Measuring retrieval effectiveness based on user preference of documents", *Journal of the American Society for Information Science*, Vol. 46 No. 2, pp. 133-145.

Zeynali Tazehkandi, M. and Nowkarizi, M. (2020), "Evaluating the effectiveness of Google, Parsijoo, Rismoon, and Yooz to retrieve Persian documents", *Library Hi Tech*, doi: 10.1108/LHT-11-2019-0229.

Zuva, K. and Zuva, T. (2012), "Evaluation of information retrieval systems", *International Journal of Computer Science and Information Technology*, Vol. 4 No. 3, pp. 35-43.

Further readings

Ali, S. and Gul, S. (2016), "Search engine effectiveness using query classification: a study", *Online Information Review*, Vol. 40 No. 4, pp. 515-528.

An, X. and Huang, J.X. (2017), "geNov: a new metric for measuring novelty and relevancy in biomedical information retrieval", *Journal of the Association for Information Science and Technology*, Vol. 68 No. 11, pp. 2620-2635.

Arora, M., Kanjilal, U. and Varshney, D. (2016), "Evaluation of information retrieval: precision and recall", *International Journal of Indian Culture and Business Management*, Vol. 12 No. 2, pp. 224-236.

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 No. 2, pp. 1015-1020.

Bollmann, P. (1984), "Two axioms for evaluation measures in information retrieval", *Proceedings of the 7th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, BCS Learning and Development, Swindon*, pp. 233-245.

Buckley, C. and Voorhees, E.M. (2017), "Evaluating evaluation measure stability", *ACM SIGIR Forum*, Association for Computing Machinery, New York, NY, pp. 235-242.

Chignell, M.H., Gwizdka, J. and Bodner, R.C. (1999), "Discriminating meta-search: a framework for evaluation", *Information Processing and Management*, Vol. 35 No. 3, pp. 337-362.

Cleverdon, C.W. (1972), "On the inverse relationship of recall and precision", *Journal of Documentation*, Vol. 28 No. 3, pp. 195-201.

Cleverdon, C.W., Mills, J. and Keen, E.M. (1966), *Factors Determining the Performance of Indexing Systems*, Volume 1, Design, College of Aeronautics, Cranfield, CT.

Cooper, W.S. (1968), "Expected search length: a single measure of retrieval effectiveness based on the weak ordering action of retrieval systems", *American Documentation*, Vol. 19 No. 1, pp. 30-41.

Cooper, W.S. (1973), "On selecting a measure of retrieval effectiveness", *Journal of the American Society for Information Science*, Vol. 24 No. 2, pp. 87-100.

Davis, J. and Goadrich, M. (2006), "The relationship between precision-recall and ROC curves", *Proceedings of the 23rd International Conference on Machine Learning, ACM*, pp. 233-240.

Dong, X. and Su, L.T. (1997), "Search engines on the world wide web and information retrieval from the internet: a review and evaluation", *Online and CD-ROM Review*, Vol. 21 No. 2, pp. 67-82.

Eguchi, K., Oyama, K., Ishida, E., Kando, N. and Kuriyama, K. (2003), "Evaluation methods for web retrieval tasks considering hyperlink structure", *IEICE Transactions on Information and Systems*, Vol. 86 No. 9, pp. 1804-1813.

Frei, H.P. and Schäuble, P. (1991), "Determining the effectiveness of retrieval algorithms", *Information Processing and Management*, Vol. 27 Nos No. 2-3, pp. 153-164.

Gordon, M. and Kochen, M. (1989), "Recall-precision trade-off: a derivation", *Journal of the American Society for Information Science*, Vol. 40 No. 3, pp. 145-151.

Harper, D. (2000), "Online etymology dictionary", available at: www.etymonline.com (accessed 12 November 2018).

Harter, S.P. (1996), "Variations in relevance assessments and the measurement of retrieval effectiveness", *Journal of the American Society for Information Science*, Vol. 47 No. 1, pp. 37-49.

Hawking, D., Craswell, N., Thistlewaite, P. and Harman, D. (1999), "Results and challenges in web search evaluation", *Computer Networks*, Vol. 31 Nos No. 11-16, pp. 1321-1330.

Henzinger, M. (2007), "Search technologies for the internet", *Science*, Vol. 317 No. 5837, pp. 468-471.

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 No. 5, pp. 301-306.

Landoni, M. and Bell, S. (2000), "Information retrieval techniques for evaluating search engines: a critical overview", *ASLIB Proceedings*, Vol. 52 No. 3, pp. 124-129.

Levene, M. (2011), *An Introduction to Search Engines and Web Navigation*, Wiley, Canada.

Li, K., Wang, Y. and Yu, W. (2012), "Chapter 7: personalised search engine evaluation: methodologies and metrics", *Web Search Engine Research*, Emerald Group Publishing, pp. 163-202.

Meng, X. (2006), "A comparative study of performance measures for information retrieval systems", *Third International Conference on Information Technology: New Generations*, *IEEE*, pp. 578-579.

Moffat, A. and Zobel, J. (2008), "Rank-biased precision for measurement of retrieval effectiveness", *ACM Transactions on Information Systems*, Vol. 27 No. 1, p. 2.

Oppenheim, C., Morris, A., McKnight, C. and Lowley, S. (2000), "The evaluation of WWW search engines", *Journal of Documentation*, Vol. 56 No. 2, pp. 190-211.

Pao, M.L. (1989), *Concepts of Information Retrieval*, Libraries Unlimited, Englewood, CO.

Poulter, A. (1997), "The design of world wide web search engines: a critical review", *Program*, Vol. 31 No. 2, pp. 131-145.

Raghavan, V., Bollmann, P. and Jung, G.S. (1989), "A critical investigation of recall and precision as measures of retrieval system performance", *ACM Transactions on Information Systems (Tois)*, Vol. 7 No. 3, pp. 205-229.

Robertson, S. (1969), "The parametric description of retrieval tests", *Journal of Documentation*, Vol. 25 No. 2, pp. 93-107.

Robertson, S. (2008), "On the history of evaluation in IR", *Journal of Information Science*, Vol. 34 No. 4, pp. 439-456.

Robertson, S. and Zaragoza, H. (2007), "On rank-based effectiveness measures and optimization", *Information Retrieval*, Vol. 10 No. 3, pp. 321-339.

Salton, G. and McGill, M. (1983), *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, NY.

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.

Smith, A.G. (2003), "Think local, search global? Comparing search engines for searching geographically specific information", *Online Information Review*, Vol. 27 No. 2, pp. 102-109.

Stokes, P., Foster, A. and Urquhart, C. (2009), "Beyond relevance and recall: testing new user-centred measures of database performance", *Health Information and Libraries Journal*, Vol. 26 No. 3, pp. 220-231.

Su, L.T. (1992), "Evaluation measures for interactive information retrieval", *Information Processing and Management*, Vol. 28 No. 4, pp. 503-516.

Tenopir, C. (2003), "Information metrics and user studies", *ASLIB Proceedings*, Vol. 55 Nos 1/2, pp. 13-17.

Valery, J.F., Jacob, S. and Vladimir, G.V. (1997), *Automated Information Retrieval: Theory and Methods*, Emerald Group Publishing, Bingley.

Van Rijsbergen, C.J. (1974), "Foundation of evaluation", *Journal of Documentation*, Vol. 30 No. 4, pp. 365-373.

Vaughan, L. (2004), "New measurements for search engine evaluation proposed and tested", *Information Processing and Management*, Vol. 40 No. 4, pp. 677-691.

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

Yao, Y. (1995), "Measuring retrieval effectiveness based on user preference of documents", *Journal of the American Society for Information Science*, Vol. 46 No. 2, pp. 133-145.

About the authors

Mahdi Zeynali Tazehkandi is a PhD candidate in the Department of knowledge and Information Science, Ferdowsi University of Mashhad, Mashhad, Iran. His main research interests are Philosophy of Science, Evaluation of Information Retrieval Systems and Search Engines. He has authored ten articles in prestigious Persian journal and three articles in international journal such as *Library Hi Tech*, *Libri* and *International Journal of Information Science and Management*. Mahdi Zeynali Tazehkandi is the corresponding author and can be contacted at: ma.zeynali@mail.um.ac.ir

Mohsen Nowkarizi is a Professor in the Department of knowledge and Information Science, Ferdowsi University of Mashhad, Mashhad, Iran. His main research interests are user interface, knowledge management, entrepreneurship in libraries, user behaviors and information retrieval. He has some experiences in supervising MS and PhD students in these areas. He has authored several articles in prestigious Persian journal and some articles in international journals such as *Journal of Librarianship and Information Science*, *Libri*, *International Journal of Information Science and Management* and so on.