Special requirements for comparative evaluation of web search engines

Wissam Tawileh, Thomas Mandl, Joachim Griesbaum
Institute of Information Science
University of Hildesheim
Hildesheim, Germany
tawilehw@uni-hildesheim.de

ABSTRACT: *Performance evaluation of classical information retrieval systems usually aims to assess the ability of these systems to find documents considered relevant to a certain search query based on a specific evaluation criteria. This approach, however, is not suitable to adequately evaluate some information retrieval applications such as web search engines. The web special characteristics make information retrieval tasks and the evaluation of search engines on the web face multiple challenges. Different web-specific, user-specific and language-specific requirements should be considered when designing and performing evaluation tests on operational web search engines. This paper discusses the special requirements for comprehensive comparative evaluation of different web search engines and highlights some language-specific considerations for evaluation in Arabic language.*

Keywords: Evaluation Requirements; Web Search Engines; Comparative Evaluation; Information Retrieval; Arabic Language

Received: 2 February 2011, Revised 8 March 2011, Accepted 15 March 2011

1. Introduction

Users of web search engines try to satisfy their personal and professional needs for information by searching the web and accessing different types of documents available online presented by the used search engine in a search results list as links to relevant documents.

The ability of a web search engine to find the most relevant documents to a search query entered by a user is evaluated as a key performance indicator of this search engine. This evaluation is mainly dependent on relevance definition which varies from one user to another and can affect users' preferences when choosing their favorite web search engine.

Finding the most relevant documents on the web is a challenging task for the web search engines. Multiple factors affect the work of these engines as information retrieval systems dealing with the web as a document collection. In addition to its' huge size, other characteristics of the web are influencing the performance of web search engines when trying to meet users expectations in a search session.

Users who look for information in a specific language may not be able to read documents in other languages. Arabic language for example is the world's fifth spoken language in term of first-language speakers' count [1] but most of its internet users are not comfortable with English language and only 35% of them speak it [2]. Web search engines have to take language-specific requirements into account when trying to help this users' group to finde "useful" information on the web.

Users' behavior is also an important issue to deal with while evaluating the performance of web search engines from the users' perspective. How do web users deal with web search engines is a subject of multiple studies in different languages and users groups.

The comparative evaluation test performed on five web search engines in Arabic language presented in [3] and [4] followed the design recommendations for classical information retrieval evaluation tests and considered multiple web-specific test requirements to compare the performance of the tested search engines dealing with Arabic language from the user's point of view.

The following section goes through test types and recommendations suggested in the literature for proper evaluation tests of web search engines as a special information retrieval application. The main web-specific characteristics which can influence the design and implementation of these tests are then reviewed.

Further requirements for comparative evaluation of web search engines are presented in the third section and the fourth section discusses the language-specific considerations for evaluation tests on web search engines in Arabic language. The last section is a conclusion of this paper.

2. Evaluation of Web Search Engines

The following two types of web search engines evaluation were identified in [5]:

- *Testimonials:* which are test-drives conducted by press or computer organizations to assess the search engine features directly visible to users like ease of use, interface design and response times,

- *Shoot outs:* which are performance evaluations performed by processing different real searches on tested search engines to assess their ability to find relevant results. Shoot outs are the tests similar to information retrieval evaluation experiments conducted in a laboratory or operational environment.

A methodology for well designed and conducted classical information retrieval evaluation tests was presented in [6] and recommended the researcher to address the following ten issues while preparing for the evaluation test:

1. Need for testing
2. Type of test
3. Variables definition
4. Search engines selection
5. Finding queries
6. Processing queries
7. Experimental design
8. Data collection
9. Data analysis
10. Presenting results

This methodology aims to ensure the validity of the experiment, the reliability of the results and the efficiency of the test procedures. In addition to these general information retrieval test requirements, seven features were introduced in [5] to assure the quality of evaluation tests designed for web search engines as special information retrieval systems. These features were discussed and reduced in [7] to the following five essential test requirements for web search engines:

1. Searches should be motivated by genuine user need.
2. If a search intermediary is employed, the primary searcher's information need should be as fully captured as possible and transmitted in full to the intermediary.
3. A large number of search topics must be used.

4. Most major search engines should be included.

5. Experiments should be well designed and conducted.

From an information retrieval perspective, evaluation has multiple levels and can be classified in the two broad categories: *system-centered evaluation level* and *user-centered evaluation level* [8]. Taking both aspects into consideration, a quality measurement approach of web search engines proposed in [9] was expanded in [10] to include the following four dimensions for an overall evaluation test:

- Index quality
- Quality of the results
- Quality of search features
- Search engine usability

The comparison study in [10] concluded that the lack in research regarding user-centered overall quality evaluation of web search engines can be attributed to the limited resources usually available for evaluation experiments and the need for comprehensive studies to cover all aspects of search engines quality evaluation from the user's perspective.

A modern approach for evaluation of web search engines based on a *crowdsourcing* methodology was recently presented in [11]. The evaluation paradigm called *Technique for Evaluating Relevance by Crowdsourcing (TERC)* allows more jurors to involve in the relevance evaluation process and get paid for their participation through an online service offered by Amazon Internet Services Company. Amazon Mechanical Turk (MTurk 1) enables developers to include human intelligence in performing tasks for their applications [12]. TERC provides a fast, low cost, high quality and flexible relevance evaluation process through configurable features of the (MTurk) evaluation tasks and participation requirements.

As participated users are evaluating given search results for given test queries and do not have their own original information needs involved, this paradigm is limited by the artificiality of the task. The same tool, however, can be used to accomplish information needs reconstruction tasks as implemented in this test but on a larger scale, which may provide more adequate simulation of the original information needs which motivated the use of the selected search queries.

Participated users are, moreover, unknown to the researcher, which makes it unclear to which extent they possess certain prerequisites to participate in the evaluation process. To overcome such issues, MTurk offers "qualification tests 2 " which obligate users to self-assess their suitability for a particular task available on the website. A user can not start processing a task (thus get paid) before passing a qualification test defined and configured by the task requestor (the evaluation researcher).

Web search engines were introduced to help users find documents relevant to their information needs on the web as a huge document collection. Specific properties of this document collection are characterized in [13] as challenges for web search engines to work effectively and efficiently. Some of these properties are listed here.

2.1 Size of the Web

A detailed study published in [14] at the University of California estimated the total size of the internet in the year 2002 with 532,897 Terabytes. This includes all types of data and both the surface web and the deep web (also called invisible web) which is the part of the web that search engines do not add to their indices [10]. Estimating the size of the deep web is even more difficult than measuring the surface web. In [15] the deep web is referred to as the dynamic web pages generated on demand (from a database for example) and is estimated to be up to 550 times larger than the surface web [15]. However, this estimation was found in [16] to be exaggerated.

A method is proposed in [17] to estimate the total count of web pages indexed by different search engines and developed a website to estimate the size of the world wide web on a daily bases. In mid May 2009, at least 25 billion web pages were estimated. All these estimates show how big the web as a document collection is, which translates into a real challenge for web search engines to explore, analyze and index the largest possible amount of web pages.

[1] https://www.mturk.com/mturk/welcome

[2] https://www.mturk.com/mturk/findquals

The other important factor facing web search engines is the rapid expansion of the web, which enforce search engines to optimize their crawling and indexing strategies if they want to keep their indexes up to date and to be able to find latest relevant and available information. A study predicts that the internet doubles it size every 5.32 Years [18].

2.2 Heterogeneity

Heterogeneity on the web has multiple dimensions of which the two are important for search engines to deal with:

- Data types: In addition to different formats of images and audio and video streams, the web contains multiple formats of text documents which make the identification of the document type and the detection of its content more complicated for search engines.

- Languages: The detection of content language in a text document can be very helpful for search engines to classify the documents by language and to avoid the wrong retrieval of documents containing a keyword which has different usage in multiple languages. This issue is particularly relevant for language-specific evaluation tests like in Arabic language.

2.3 Granularity and Hyperlinks

Most documents on the web use hypertext structures to embed images and text paragraphs from external sources inside the document text. Web pages containing frame sets reference to the original pages using hyperlinks as well. This makes exploring web pages and relations between them in need for extra effort from search engines to add web pages to their indices and retrieve useful information even when they are located on further links contained in a web page.

2.4 Web Dynamics

Other than conventional information retrieval systems, web search engines must update their indices very frequently as the internet is a very dynamic medium. Thousands of web pages are being added, modified and deleted every day. The actuality of the search engine index is very important to retrieve currently available information. It is considered in [19] as a key success factor for web search engines. Users information needs will not be met if most of the retrieved pages were deleted or replaced with different content by the time of their retrieval.

2.5 Authors and users

The historical development of the internet to be a public information medium and later a commercial medium changed the structure of documents available online to meet the targets of websites owners and professional websites design and development service providers. No standards have been developed to help search engines process and index web pages.

The visual design was for professional website developers more important to offer competitive, attractive websites than technical needs for automatic search processing. Many developers hide the source codes of web pages they generate, making automatic processing more complicated.

2.6 Content and Spamming

Another type of challenges for web search engines is the efforts of website owners to get the highest possible traffic rates to their sites. They try to make their websites appear at the top of search results lists. Multiple tricks could give search engines false information about a site content to place it among the first search results. This misuse of web development techniques is called (like other web misuses) spamming and can negatively affect the indexing and results presentation performance of web search engines.

In addition to these and other web-specific challenges facing web search engines, language-specific and users-related issues can also affect their performance and influence the design and results of evaluation tests assessing their functionality in specific language or region.

3. Comparative Evaluation of Web Search Engiens

All evaluated operational commercial web search engines proofed, in so many studies, to be far from the perfect case and to need improvement in multiple aspects. Comparing the performance of different web search engines helps, however, to understand users' preferences in web search and considering alternative web search engines for special requirements or users groups.

This comparative evaluation needs attention to be paid for the selection of tested web search engines and their special

features. Moreover, the design of a comparative evaluation test is also affected by some factors. Based on previous studies (like [20] and [21]) and remarks noted during the evaluation test presented in [3], issues that may affect the test results and should be considered in a comparative evaluation test include:

• Special search features: Does a specific search engine offer search features for the comfort of the users' group targeted in the test, such as restricting the search to documents in a specific language or originated from a specific country or region? This may be important for users to find information more relevant to their own context.

Another useful feature for users in a specific location is the ability to save their search preferences and localize their search experience to benefit from these settings each time they use the same search engine again. Some search engines integrate new special features in their default search interface to offer a comfortable search experience without saving special preferences or using advanced search options. "Google Instant1" for example displays results as the user is typing the search query.

• Search engine interface: For users with little knowledge of forgiven languages (mostly English), a local interface of the web search engine would be of a great help and meaning. Users want to find satisfy their information needs with a minimum effort and would probably prefer a web search engine with which they can deal in their mother tongue. International web search engines interested in local markets usually offer multi-lingual search interfaces, or even extra websites in the local top level domains1. These country specific websites should present search results more relevant to local users needs.

• Queries processing: Test queries should be processed simultaneously on all tested search engines when evaluating multiple web search engines. All test engines should have an equal chance with the current index at a given point of the test.

• Users' preferences: Search results should be anonymized and presented in a randomized list to the jurors to avoid giving a better chance for a certain search engine with more users' preferences. This can also avoid the learning effect when multiple engines deliver results in the same ranking.

The effect of these issues may not be obvious in the initial comparative test design, but can affect the overall performance judgment of a web search engine from user's perspective.

Comparative evaluation tests can also be designed for different types of web search engines to compare the performance of different search methods, or the ability of different types of search engines to deal with a certain searth method. An example of these multi-type comparative evaluation tests is presented in [22] where the semantic search performance of three keyword-based search engines (Google, Yahoo and MSN) and one semantic search engine (Hakia) is compared. The study found that the performance of the tested search engines varied depending on the evaluation measure. On the other hand, both types of search engines, the keywordbased and the semantic search engines, showed low semantic search performance. This can help understanding how effective a search method is and which improvements are required.

4. Language- Specific Requirements

Following the test methodology introduced in [6] for information retrieval evaluation and considering the five requirements for web search engines evaluation tests in [7], the evaluation test designed and conducted on the five web search engines (Araby, Ayna, Google, MSN and Yahoo) in Arabic language [3] faced multiple language-specific issues which influenced its' design, performance and results. These issues were discussed in [23] and are summarized here.

4.1 Finding queries

Fifty queries were selected from the top searches reported by the Arabic web search engine (Araby.com [5]) on its' homepage. Most of these queries were general and consisted of one or two terms only3. Search queries should be randomly selected for the evaluation test. However, the following four language- and location-dependent factors affected the query selection process for evaluation in Arabic language:

4.1.1 Arabic language vowels

Some Arabic words can have different meanings when using language-specific diacritics [24]. The simple term "شعر" for example can point to "شَعر" which means "Poetry" or "شِعر" which means "Hair". The difference in the meaning of a singleterm polysematic query affects relevance judgment, as different users' information needs can not be guessed and relevant documents can not be clearly defined.

[3] http://www.google.com/instant/

[4] Google offers multiple websites for different Arab countries (googl.com.sa, google.com.qa, google.com.eg …etc)

The evaluation measure Importance of completeness of search results may overcome similar issues as supposed in [25]. As there is still no adequate solution for this problem (especially for Arabic language), polysemantic Arabic search queries had to be eliminated in this evaluation test.

4.1.2 Backgrounds and origins of users

Country-specific and religious topics should be also avoided, even when they appear in the top searches on the search engine Araby.

Arab users with limited knowledge of local and national subjects in different Arab countries should be able to participate in a general evaluation test in Arabic language on web search engines. Relevance judgment of search results related to such queries can be problematic and influenced by users' culture and background.

Although defined relevance judgment criteria were provided to jurors to decide on the documents relevance, avoiding obviously religious and location-specific queries could secure more valid relevance judgment.

4.1.3 Censorship

Systematic content filtering in most Arab countries limits accessibly for internet users [26]. Religiously and culturally objectionable content is filtered by political regimes. Access to specific political and conflict related content is forbidden in these countries. Some countries restrict access to an entire top level domain for political reasons.

A juror may not be able to access some or all of the search results presented for a certain search query from his/her location when they point to a blocked content.

Test search queries should be selected in a way that ensures a high level of accessibility to search results from most all countries in the Arab region. This can be particularly difficult and requires knowledge of local political situations as some topics can be temporarily objectionable in a certain country.

4.1.4 Content types

According to "Google Insights for Search1", images, audio and video files are usually in the top ten searches initiated by users from the Arab world. Very general single-term queries like "songs", "videos" and "photos" are used by almost all Arab users to search for specific content types.

The complication in defining information needs and relevance judgment for these content types, in addition to limited access to specific providers on the web (like youtube.com) in some Arab countries make it harder to include search queries from similar topics in the test.

4.2 Original information needs

Exploring users' information needs originally motivated their use of a particular search query is very difficult when general short queries are used. A simulation of original information needs can be constructed from multiple descriptions of the information a user might have needed when he/she searched for a query. These descriptions can be collected from volunteers participating in a reconstruction process.

4.3 Finding motivated participants

Like in many other information retrieval tests, enough volunteers were required for the information needs description task and for the relevance evaluation tasks. Fifty evaluation jurors are ideally required for fifty test queries. A juror has to evaluate a single query on all tested search engines. The juror has to be a native Arabic speaker and to have basic knowledge in internet browsing and dealing with web search engines.

An evaluation test in a laboratory environment was difficult because of the following reasons:

- Organization and logistics: Timing was extremely difficult to plan. Logistics issues would rise with a number of fifty jurors participating in a limited lab capacity within a short time.

[5] Araby.com suspended its operations in May 2010 http://www.souq.com/Announcement.html

[6] A translated list of selected queries is available online: http://uni-hildesheim.de/~tawilehw/se/queries.html

• Test acceptance: Test tasks can be complicated and confusing. A total of six tasks should be performed including the evaluation of fifty descriptions and fifty results and the documentation of multiple steps.

The TERC paradigm proposed in [11] helps overcoming some of these issues. Users in this online croudsourcing methodology are financially motivated and get paid for every completed evaluation task. The task requester (the researcher) can reject a particular completed task for poor quality and deny the payment.

Participants through MTurk can be enforced to pass a qualification test for Arabic language proficiency before starting the evaluation process. A certain knowledge level of local culture may also be required and proofed through such tests. A user with excellent Arabic language skills may not be able to evaluate the relevance of some web documents within a specific context.

Performing the evaluation tasks online proved to be suitable for this and other evaluation tests when a large number of jurors should be involved in the process. The long latency in the evaluation completion faced in this test can be solved be defining a due date and time for the MTurk tasks. Interested participant will be willing to meet the evaluation deadline and get paid for their work.

4.4 Search engines selection

Arab users should be able to use the search engine with minimal effort. The most popular five search engines according to Alexa were selected. Araby (www.araby.com) and Ayna (www.ayna.com) were the only ranked native Arabic engines by the time of the test. The three search engines Google (www.google.com.sa), MSN (www.live.com 2) and Yahoo (www.yahoo.com) are the top used international Arabicenabled search engines in the region.

To use all tested search engines as Arabic specialized search engines working on an Arabic document collection, the tested search engines should be accessed through their Arabic interface when available, and the engines should be configured to limit their search to Arabic web pages if possible.

4.5 Arab users' behavior

As there are no published studies about Arab users' behavior, international users' behavior regarding short search sessions and the seldom usage of search operators in the queries could be generalized from international studies (like [28]) to Arab users as well. Arab users seemed also to use very short and rather unspecific search queries

5. Conclusions

The evaluation of operational web search engines has to consider special characteristics of the web as a document collection, its' different content types and its' users.

Comparative evaluation of different web search engines may need design adjustments to give all tested search engines an equal chance in performance evaluation.

In addition to web-specific requirements, language-specific issues should also be considered when designing and performing comparative evaluation tests. This depends on the particular language targeted by the test.

More research is still needed to explore the ability of web search engines to deal with specific languages, and Arabic in particular. An overall performance evaluation is still a subject of multiple international studies and need to be further discussed and developed.

The recently proposed relevance evaluation approach TERC can help in designing and performing comparative evaluation tests on web search engines effectively and efficiently using a configurable, low cost, location-independent and flexible tool (MTurk) available online for participants from a wide geographic area.

[7] http://www.google.com/insights/search/

[8] Officially replaced on 03.06.2009 with a new search service from Microsoft (www.bing.com) [27]

References

[1] Graddol, D. (2009). The Future of English? A guide to forecasting the popularity of the English language in the 21st century. London, British Council. <http://www.britishcouncil.org/learning-elt-future.pdf> (Accessed 10.08.09)

[2] Hancock, M.(2006). Arab Internet use up by nine million." ITP Digital, 24. September<http://www.itp.net/index.php?view=article&id=487633&Itemid=1&opt ion=com_content> (Accessed 10.08.09)

[3] Tawileh,W., Mandl, T., Griesbaum, J.(2010). Evaluation of Five Web Search Engines in Arabic Language, *In*: M. Atzmueller, Benz, D., Hotho, A., Stumme, G.,. (Eds.) "LWA 2010 – Lernen, Wissen & Adaptivität: Workshop Proceedings. 04.-06. October , Universität Kassel. Workshop Information Retrieval. p. 221-228

[4] Tawileh, W., Mandl, T., Griesbaum, J. (2010). *Search results presentation and interface design - A comparative evaluation study of five web search engines in Arabic language*, *In*: Proceedings of the 10th International Conference on Intelligent Systems Design and Application (ISDA2010), 29. November - 01. December, Cairo – Egypt. IEEE, p. 592-597.

[5] Gordon, M., Pathak, P. (1999). Finding Information on the World Wide Web: the Retrieval Effectiveness of Search Engines, *In*: Information Processing & Management 35, 141-180.

[6] Tague-Sutclife, J.(1992). The Pragamatics of Information Retrieval Experimentation, Revisited, *Information Processing & Management*, 28 (4) Elsevier, p. 467-490.

[7] Hawking, D., Craswell, N., Bailey, P., Griffiths, K., (2001). Measuring Search Engine Quality, *Information Retrieval*, 4 (1) Springer, Netherlands, 33-59.

[8] Saracevic,T. (1995). Evaluation of evaluation, *In*: Information Retrieval, Paper presented at the SIGIR2 95, Seattle, CA. ACM Press, p. 138-146.

[9] Lewandowski, D., (2006) Zur Bewertung der Qualität von Suchmaschinen. *In*: J. Eberspächer and S. Holtel (Eds.), „Suchen und Finden im Internet, Springer, Heidelberg, p. 195-199.

[10] Lewandowski, D., Höchstötter, N. (2008). Web Searching: A Quality Measurement Perspective, *In*: A. Spink and M. Zimmer (Eds.). Web Searching: Multidisciplinary Perspectives, Springer, Berlin. 2008, p. 309-340.

[11] Alonso,O., Rose, D. E., Stewart, B.(2008). Crowdsourcing for relevance evaluation, *SIGIR Forum* 42, 2 (November), 9-15.

[12] Barr, J., Cabrera, L. F. (2006). AI Gets a Brain. *Queue* 4, 4 (May) 24-29.

[13] Ferber, R. (2003). Information Retrieval, Suchmodelle und Data-Mining- Verfahren für Textsammlungen und das Web, dpunkt Verlag GmbH, Heidelberg.

[14] Lyman, P., Hal, R. V. (2003). How Much Information 2003?" <http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/internet.htm> (Accessed 10.08.09)

[15] Bergman, M. K. (2001). The deep web: surfacing hidden value, *Journal of Electronic Publishing* 7, University of Michigan.

[16] Lewandowski, D., Mayr, P. (2006).Exploring the academic invisible web, *Library Hi Tech* 24, p. 529–539.

[17] de Kunder, M., (2007) "Estimated size of the indexed World Wide Web." Unpublished masters thesis, Tilburg University,. <http://www.dekunder.nl/Media/Scriptie%20Maurice%20de%20Kunder%20-%20Grootte%20geindexeerde%20web.pdf> (Accessed 10.08.09)

[18] Zhang, G.Q., Zhang, G. Q., Yang, QF., Cheng, SQ., Zhou, T. (2008). Evolution of the Internet and its cores, *In*: New Journal of Physics, Vol. 10, December 2008, <http://www.iop.org/EJ/article/1367- 2630/10/12/123027/njp8_12_123027.html> (Accessed 10.08.09)

[19] Lewandowski, D., (2006) 2006 Aktualität als erfolgskritischer Faktor bei Suchmaschinen. *In*: IWP - Information: Wissenschaft und Praxis 57, 3, 141-148.

[20] Griesbaum, J., Rittberger, M., Bekavac, B. (2002). Deutsche Suchmaschinen im Vergleich: AltaVista.de, Fireball.de, Google.de und Lycos.de", *In*: Hammwöhner, R., Wolff, C., Womser-Hacker, C., (Ed.) Information und Mobilität, Optimierung und Vermeidung von Mobilität durch Information", Proceedings des 8. Internationalen Symposiums für Informationswissenschaft, Konstanz. p. 201-223.

[21] Lewandowski, D., (2008). Problems with the use of web search engines to find results in foreign languages, *Online Information Review* 32 (5) 668-672.

[22] Tümer, D., Shah, M. A., Bitirim, Y. (2009). An Empirical Evaluation on Semantic Search Performance of Keyword-Based and Semantic Search Engines: Google, Yahoo, Msn and Hakia, *In*: 4th International Conference on Internet Monitoring and Protection (ICIMP '09) 24-28. May , p.51-55.

[23] Tawileh, W., Mandl, T., Griesbaum, J. (2011). Evalutaion of Web Search Engines and Challenges for Arabic Language, *In*: Proceedings of the 2011 International Conference on Information and Computer Applications (ICICA2011), 18 - 20. March , Dubai – UAE, IEEE . p. 334-338.

[24] Hammo, B., (2009). Towards enhancing retrieval effectiveness of search engines for diacritisized Arabic documents, *Information Retrieval*, 12 (3) Springer, Nietherlands, p. 300-323.
[25] Lewandowski, D., (2007). Mit welchen Kennzahlen lässt sich die Qualität von Suchmaschinen messen? *In:* M. Machill and M. Beiler (Edt.), ie Macht der Suchmaschinen / The Power of Search Engines. Köln: Herbert von Halem Verlag. p. 243-258.
[26] Noman, H., Zarwan, E., (2007). Regional Overview: Middle East and North Africa. OpenNet Initiative, 15. May. <http://opennet.net/research/regions/mena> (Accessed 10.08.09)
[27] Microsoft Corporation. Online Press Release, 28, May (2009) <http://www.microsoft.com/presspass/press/2009/may09/05-28NewSearchPR.mspx> (Retrieved 10.08.09)
[28] Jansen, B. J., Spink, A., Saracevic, T., (2000). Real life, real users, and real needs: a study and analysis of user queries on the Web, *Information Processing & Management*, 36 (2) 207-227.