# IRVILAB: Gamified Searching on Multilingual Wikipedia

Paavo Arvola
Faculty of Information Technology
and Communication Sciences
Tampere University
Tampere, Finland
paavo.arvola@tuni.fi

Tuulikki Alamettälä
Faculty of Information Technology
and Communication Sciences
Tampere University
Tampere, Finland
tuulikki.alamettala@tuni.fi

## ABSTRACT

Information retrieval (IR) evaluation can be considered as a form of competition in matching documents and queries. This paper introduces a learning environment based on gamification of query construction for document retrieval, called IRVILAB (Information Retrieval Virtual Lab). The lab has modules for creating standard evaluation settings, one for topic creation including relevance assessments and another for performance evaluation of user queries. In addition, multilingual Wikipedia online collection enables a module, where relevance assessments are translated to other languages. The underlying game utilizes IR performance metrics to measure and give feedback on participants' information retrieval performance. It aims to improve participants' search skills, subject knowledge and contributes to science education by introducing an experimental method. Distinctive features of the system include algorithmic relevance assessments and automatic recall base translation.

## CCS CONCEPTS

• Information systems → Information retrieval; Evaluation of retrieval results; Test collections • Human-centered computing → Collaborative and social computing; Social Tagging systems

## KEYWORDS

Retrieval evaluation, gamification, serious gaming, IIR, Integrated Learning of Content and Language (CLIL)

**ACM Reference format:**
Paavo Arvola and Tuulikki Alamettälä. 2022. IRVILAB: Gamified Searching on Multilingual Wikipedia. In *Proceedings of ACM SIGIR '22 Conference, July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3477495.3531662


*SIGIR '22, July 11–15, 2022, Madrid, Spain*

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531662

## 1 Introduction

Effective query formulation improves search results and information interaction [11, 4], which can be measured with IR metrics using controlled settings as described in this paper. Information search skills are an important part of *information literacy*, which has been acknowledged as an essential prerequisite for engaged citizenship and lifelong learning [2, 5]. Without relevant search results, it is difficult to proceed further in the process: into evaluating and using information. An ability to use the search engine and query formulation is the foundation of the search process.

### 1.1 Learning query construction

Query formulation involves constructing a query for a search engine to express an information need. Typically, a user expresses it in the form of keywords and phrases. The queries are matched against a document collection by a search engine. The quality of results is heavily dependent on the query expression.

Earlier research indicates that query formulation and reformulation can be one of the most problematic and challenging tasks for users [7, 12]. To overcome these challenges, we have developed a system that trains information search skills. It measures and gives feedback based on participants' information retrieval performance. The pedagogical goal is to improve participants' search skills in an experimental environment, a LAB. Thus, using IR measures as game components contributes to science education by introducing an experimental method. The system works in a multilingual environment fostering CLIL (Content and Language Integrated Learning), where topics are searched using a foreign language to be learned.

Although many user groups are quite possible, preferred players (searchers) are second degree students, while topic creators and assessors would be their teachers or university students. The latter group would foster much needed co-operation between the second-degree education and universities. This co-operation contributes to the science education and search skills, disseminating and promoting information retrieval related studies in the universities at the same time.

### 1.2 IR as a game

Generally, it is meaningful to gamify information retrieval for various settings (e.g., [6]). Particularly, the IR document retrieval evaluation scheme (laboratory model) is intrinsically a game,

where methods are compared against the others in a competitive setting. In traditional evaluations IR methods or systems are evaluated given a query, but alternatively queries can be evaluated with a given IR system.

Accordingly, our approach is based on an old idea of analyzing query performance (QPA) by using the standard IR metrics [10]. The QPA style game setting followed the IR laboratory model with controllable search settings including frozen standard test collections and IR metrics.

The IRVILAB Wikipedia edition[1] is tailored for the online Wikipedia, where the articles are contemporary and meaningful items for the participants to search. In IRVILAB, the topics can be created, and articles assessed by anyone, if necessary. In addition, IRVILAB supports rapid (semi) automatic recall-base construction, using a query as a seed. For multi-lingual purposes it offers a feature where the assessments can be translated virtually to any other language through *cross-language links*.

The presented demonstration uses online Wikipedia as a familiar environment for the participants to train searching and compete against themselves and one another based on their queries and material they find. The game can be tailored for virtually any topic and language, thus suitable for learning subject knowledge and foreign languages through query formulation, search and personal relevance assessments.

### 1.3 The online Wikipedia as a multilingual collection

Wikipedia is a free, multilingual online encyclopedia covering a plethora of topics. While being one of the most popular sites on the Internet having editions in 325 languages, it is known worldwide. Because of all this and more, Wikipedia has been used in various IR settings as a frozen edition for repeatable IR evaluation settings (e.g., [1]). Online Wikipedia, in turn, is not frozen, but more like an ever-growing organism, not directly suitable for robust IR evaluations over periods of time. However, apart from the Web, as an encyclopedia it is still a much less dynamic environment encapsulating subjects as individual articles. These articles often represent related concepts or entities with constantly and mostly improving quality and thus relevance in their content. In addition, many significant articles have their unique counterparts in Wikipediaes of other languages, forming a coarse-grained parallel corpus. Technically, these articles are yielded through cross-language links.

## 2 System description

In IRVILAB anyone can set up a game. Game setting is basically writing a topic description and making (graded) relevance assessments within the collection based on the description. The player (searcher) is given a standard "bag-of-words" query field, where the player inputs a query based on the topic description and gets results with immediate feedback in the form of result list and performance analysis as IR measures. After seeing the results, the player can suggest relevance values for documents with missing relevance judgements or disagree with the existing ones, contributing to the collaborative effort for relevance assessments.

### 2.1 System architecture

IRVILAB works as an online browser-based system offering a standard HTML client, with some Javascript. The user interface is simplistic and has been successfully tested with most commercial browsers.

The back-end is programmed using PHP with MariaDB database access. While it is intended for educational purposes, it offers views for both students and teachers. No standalone installation is required, instead the system is scalable and can be used from anywhere.

The system architecture is lightweight; the IR system and document maintenance are based on technology offered by Mediawiki. The search utilizes Mediawiki API[2] on Wikipediaes of different languages. In other words, using the CirrusSearch query language, the documents are retrieved directly from online Wikipediaes. The inner functions are described here as program modules for topic design, translation and query and information seeking evaluation module as the game (See Figure 1).

### 2.2 Topic creation module

Using the Topic creation module, topics are created for the participants to compete. This involves writing a topic description (information need) and related relevance assessments of the documents. Topics can be created by a teacher, if used on a course, or anyone for that matter. Alternatively, or in addition, they can be done in a multi-user setting collaboratively as well while playing. Apart from manual assessments following a topical relevance criterion, the system offers automatic relevance assessment process following algorithmic relevance criterion [9]. In the latter, option documents' relevance grading is based on the assessors one or more queries, where the hits are marked as relevant. Algorithmic relevance assessments may be sufficient for a basis of a game and learning instead. The assessments can be altered manually later on, making the assessment process semi-automatic. Graded relevance assessments are supported.

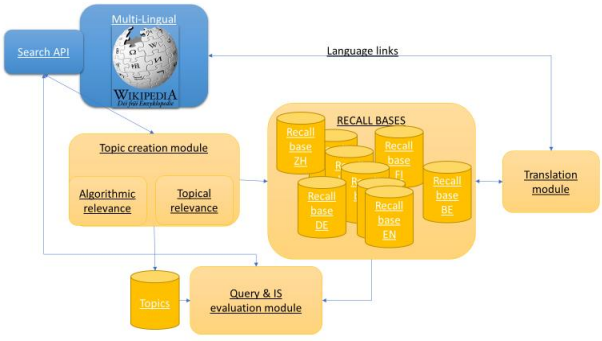


**Figure 1: System diagram with modules**

[1] https://webpages.tuni.fi/irvilab/lab/

[2] https://www.mediawiki.org/wiki/API:Search

**Table 1: Two sample Ad hoc topics from [1] and related recall base sizes (#docs) per relevance grade on Finnish Wikipedia.**

| | 0 | 1 | 2 | 3 | Tot. |
|---|---|---|---|---|---|
| GMO (topic number 663): Safety of genetically modified food | 504 | 62 | 12 | 4 | 582 |
| TT (topic number 557): Time travel theories | 532 | 80 | 41 | 13 | 666 |

## 2.3 Translation module

The relevance assessments can be used in Wikipediaes of other languages, because the same article can be present in other Wikipediaes, too. Namely, each Wikipedia article has a set of cross-language links to the corresponding articles in Wikipediaes of other languages. Assuming their similar relevance with the already assessed source document enables automatically "translated" recall base instances. In other words, after the translation, a topic is directly available for search using another language. An obvious use case for this is CLIL (Content and Language Integrated Learning), where topics are searched using a foreign language to be learned. Another use case is to force participants (searchers) to use a language completely alien to them to illustrate the computers' viewpoint in matching.

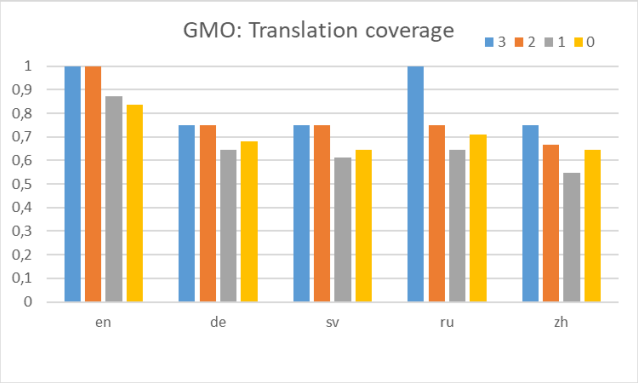


**Figure 2: Translation coverage from Finnish to five languages (English, German, Swedish, Russian and Chinese) of topic GMO. See Table 1 for initial recall base size and topic name.**

The differences between Wikipediaes may rise a concern of the completeness of the translated recall base. As a partial answer, Figures 2 and 3 present the recall base translation coverage of topics presented in Table 1 by their relevance grade (3 denoting very relevant to 0 not relevant). The relevance assessments were initially done in Finnish Wikipedia collaboratively by 50 participants and then translated automatically into recall bases of five languages (English, German, Swedish, Russian and Chinese) using the module. It seems that relevant documents are found slightly more often from other Wikipediaes than non-relevant.

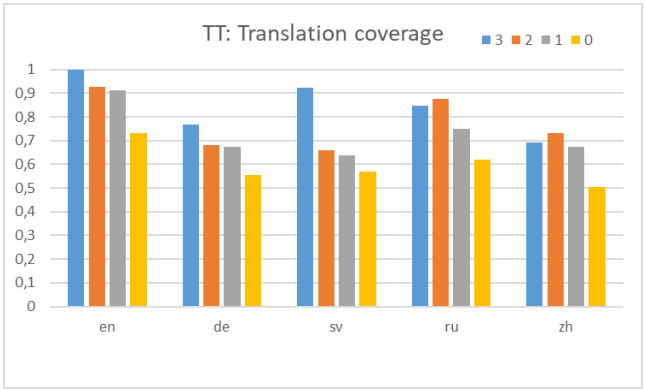


**Figure 3: Translation coverage from Finnish to five languages of topic TT (English, German, Swedish, Russian and Chinese).**

## 2.4 Query and information seeking evaluation module

The players construct queries for a given topic and given IR system and try to get as good results as they possibly can. The results are evaluated against relevance assessments (2.2 and 2.3). The gamification of query evaluation is based on IR metrics and related measures and diagrams. Related metrics include but are not limited to CG based metrics and precision-recall metrics. Euler diagram[3] is used to illustrate precision and recall for the result set and CG curves[4] for the rankings. Leaderboards are provided based on these measures for competition.

The ground truth is not hidden in the query evaluations, but provided for the players. In contrast, information seeking is another game alternative and is measured by comparing participants' relevance assessments to the ground truth hidden from the participants, or assessments done by other participants. The participants make several queries and read and assess the retrieved documents and get points whenever they find a suitable one.

Measuring IR performance in a dynamic collection raises question about the reliability of the results over time using the same relevance assessments. Even though the system enables collaborative maintenance of relevance assessments, it is not realistic to monitor them too often.

Even without any updates, we assume that for many topics the replicability remains satisfactory at least for less robust educational gaming purposes. To illustrate this, Figures 4 and 5 represent query performances in December 2020 and again in January 2022. In this experiment, 59 participants were given a task to figure out best possible query expression for two distinct topics without seeing the relevance assessments, just the result documents. The relevance assessments were done collaboratively during 2020. Right after the test the queries were evaluated against the assessments, and then again after 13 months. The

[3] https://github.com/benfred/venn.js
[4] https://developers.google.com/chart

performances rankings between these points are highly correlated and show plausible long-term sustainability at least for educational gaming purposes. Naturally, this warrants for more in-depth research.

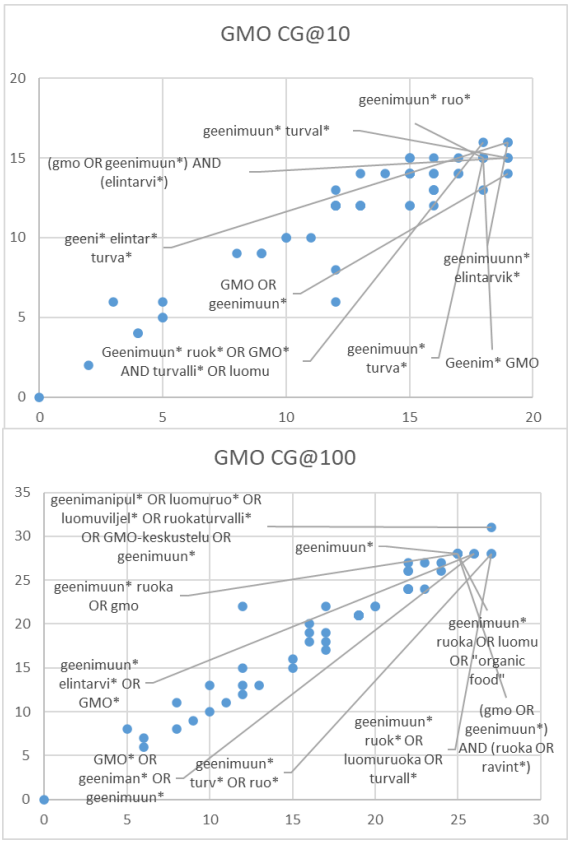


**Figure 4: Cumulated Gain at cut-offs 10 and 100 for topic GMO. x-axis Dec. 2020 score, y-axis Jan 2022 score. Top queries marked. Spearman correlations CG@10: 0.89 and CG@100: 0.97**

To improve reliability, for instance Cumulated Gain [3] metrics using condensed lists is proven to be quite accurate for some settings with incomplete assessments (e.g., [8]). In other words, for the calculation the unassessed documents are omitted from the results.

## 3 Discussion

In traditional IR system evaluations, it is necessary to find the best possible method with statistical significance. Thus, a highly controlled laboratory setting with multiple topics is preferred. The rationale behind this demonstration is that comparing queries of human participants can be performed in a less robust environment allowing more degrees of freedom in topic creation, rapid recall-base construction, and more realistic and up-to-date collections such as online Wikipedia. With human participants, a far more important factor than a perfect IR test setting and accurate results are meaningful topics, familiar and useful collection, learning and gaming. However, it seems that the system presented in this paper delivers results accurate enough to measure query performance for a learning game. To further improve robustness and reliability is one of the design goals in the future.

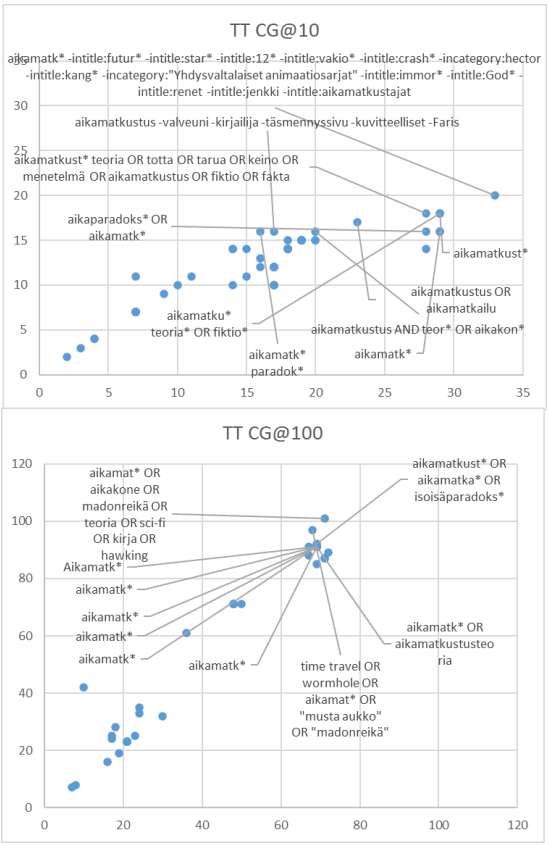


**Figure 5: Cumulated Gain at cut-offs 10 and 100 for topic TT x-axis Dec. 2020 score, y-axis Jan 2022 score. Top queries marked. Spearman correlations CG@10: 0.87 and CG@100: 0.88.**

## ACKNOWLEDGMENTS

This paper was partially funded by the European Union Erasmus+ programme with grant number 2021-1-FI01-KA220-SCH-000029713.

## REFERENCES

[1] Paavo Arvola, Shlomo Geva, Jaap Kamps, Ralf Schenkel, Andrew Trotman and Johanna Vainio. 2011. Overview of the INEX 2010 Ad Hoc Track. In S. Geva, J. Kamps, R. Schenkel, & A. Trotman (Eds.), *Comparative Evaluation of Focused Retrieval: 9th International Workshop of the Inititative for the Evaluation of XML Retrieval* (INEX 2010). Lecture Notes in Computer Science, vol 6932. Springer, Berlin, Heidelberg. 1-32. https://doi.org/10.1007/978-3-642-23577-1_1

[2] Alton Grizzle, Penny Moore, Michael Dezuanni, Sanjay Asthana, Carolyn Wilson, Fackson Banda, and Chido Onumah. 2014. *Media and information literacy: Policy and strategy guidelines.* UNESCO.

[3] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20, 4 (2002), 422–446. https://doi.org/10.1145/582415.582418

[4] Kalervo Järvelin, Pertti Vakkari, Paavo Arvola, Feza Baskaya, Anni Järvelin, Jaana Kekäläinen, Heikki Keskustalo, Sanna Kumpulainen, Miamaria Saastamoinen, Reijo Savolainen, and Eero Sormunen. 2015. Task-Based Information Interaction Evaluation: The Viewpoint of Program Theory. *ACM Trans. Inf. Syst.* 33, 1, 1-30. https://doi.org/10.1145/2699660

[5] Library and Information Association. 2018. *What is information literacy?* Retrieved from https://www.cilip.org.uk/page/informationliteracy

[6] Cristina Ioana Muntean and Franco Maria Nardini. 2015. Gamification in Information Retrieval: State of the Art, Challenges and Opportunities. In *Proceedings of the 6th Italian Information Retrieval Workshop* (IIR'2015).

[7] Soo Young Rieh and Hong Xie. 2006. Analysis of multiple query reformulations on the web: The interactive information retrieval context. *Information Processing & Management*, 42, 751-768. https://doi.org/10.1016/j.ipm.2005.05.005

[8] Tetsuya Sakai. 2007. Alternatives to Bpref. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '07). ACM, New York, NY, USA, 71–78. https://doi.org/10.1145/1277741.1277756

[9] Tefko Saracevic. 1975. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for information science*, *26*, 6, 321-343.

[10] Eero Somunen, Sakari Hokkanen, Petteri Kangaslampi, Petri Pyy, and Bemmu Sepponen. 2002. Query performance analyser -: a web-based tool for IR research and instruction. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '02). ACM, New York, NY, USA, 450. https://doi.org/10.1145/564376.564491

[11] Ryen W.White. 2016. *Interactions with search systems.* Cambridge University Press.

[12] Ryen W.White, Matthew Richardson, and Wen-tau Yih. 2015. Questions vs. Queries in Infrmational Search Tasks. In *Proceedings of the 24th International Conference on World Wide Web* (*WWW '15 Companion*). ACM, New York, NY, USA, 135–136. https://doi.org/10.1145/2740908.2742769