Training students to evaluate search engines

Mark Sanderson

School of CS & IT, RMIT University, Melbourne, Australia

Amy Warner

The National Archives, London, UK

Abstract In this chapter, two exercises that were part of an information retrieval course are described. Both exercises were set up to assess the qualities of a search engine and to consider how the engine could be improved. Students were engaged in problem-based learning. The exercises were run for several years and found to be successful both in engaging students in the course, but also in establishing links with organizations interested in evaluation of search.

Other things to teach

It would appear that there are two classic expectations about the sorts of students that will result from information retrieval courses. In Computer Science (CS), students who go on to use the skills they learn will either implement IR systems or they will research novel techniques which will eventually be of interest to developers of IR systems. Within courses taught in Information Schools (IS), there is an expectation that students will learn how search engines work and so in their eventual place of employment they will be effective managers of information and/or know how to query for data expertly. However, it is increasingly clear that there is an emerging set of skills that need to be taught in both types of department.

Ever increasing numbers of organizations offer search of their information for either internal or external use, using common commercial search solutions set in their default configuration. There is a growing realization amongst the organizations that these standard settings are not producing the ideal searching system and that adjustment and customization could provide a substantially improved service. However, exactly how such customization can be done is poorly understood in most organizations.

White (2006) suggested that such improvements will be brought about by a multi-disciplinary search team composed of both CS and IS trained people monitoring and manipulating the performance of the engine for an organization. We have started exercises within our IR course to address the training needs for the IS people in White's suggested teams. To this end, we created a course that teaches both how information retrieval systems work and the range of possible ways (from IR research) that the systems could be improved. As the only way that improvements and customizations can be assessed is through evaluation, this topic was a key component of the course delivered both through lectures and coursework. The coursework on evaluation is what is described in this chapter.

It was decided that a problem led teaching approach would be taken with the coursework with students set two different evaluations:

1. The students were instructed to find a web-based search engine, which they had to assess, reporting on its qualities as well as detailing ways in which it could be improved.
2. The students collectively conducted a test collection-like measurement of the effectiveness of two search engines, which were indexing the same information. Here, the work was conducted in collaboration with an information provider.

These two evaluations are now described in more detail. The work described here was carried out while the author was working at the Information School at the University of Sheffield, UK.

First evaluation – assessing a search engine

The inspiration for this style of coursework was taken from the work of (Heuwing et al., 2009) who assessed a wide range of publicly accessible company based extranet search engines across a range of evaluation criteria. The broad finding was that most of the engines were poorer than a Google search restricted to the company's public web pages (commonly known as a *site search*). However, it was clear from the study that it was possible to optimize a locally run search engine to out-perform Google. This work showed the generally poor state of locally run search services on the web, while at the same time demonstrating the improvement to search quality that local knowledge of information and users can bring about. It was decided, therefore to set coursework where the students were given the task to seek out and assess a poorly performing search engine and recommend ways in which it could be improved.

Education goals and objectives

The Educational goals of the exercise were to get students to assess and suggest improvements for a search engine that they had found. The educational objectives of the exercise were that at the end of the coursework students would be able to

- Find a suitable search engine to study in their coursework.

- Identify the general characteristics/qualities of the users of the selected search engine
- Determine the needs of those users
- Assess the ability of the search engine to meet those needs
- Apply a relevant sub-set of the course's lecture material so as to meet the identified needs.

An additional objective of the work was to demonstrate to the students that there was great potential for improving established search services and that with the knowledge gained during the course that they were potentially in a position to provide those improvements.

Practicalities of the exercise

A structure was provided to the students to help them complete the work successfully. In order to maximise the students' chances of finding a search engine in need of improvement, they were told to focus on candidate engines from relatively small organizations. They were told to avoid organizations where search was their core business (e.g. Google, Bing, YouTube, etc) as it was unlikely that students would be able to come up with valuable suggestions for improvements. The students were asked to demonstrate their understanding of the search engine audience by constructing a set of queries, which they considered might typically be submitted to their chosen system.

Using Broder's two main categories of queries (Broder, 2002), they were asked to detail which were navigational (queries aimed at locating a home page within the site) and which were informational (queries seeking documents containing relevant information). The students were asked to demonstrate the workings of the search site, the content of search results and how they were laid out. Finally, the students were asked to describe which ideas from course materials or even independently found research papers could be used to potentially improve the search engine.

The running of the exercise

This exercise was run at the University of Sheffield for three years in a course taken each year by both undergraduate and Masters level students. Such is the range of search sites available on the net, across the 90 or so students who have taken this exercise, only a small handful of sites were assessed more than once. Highlights from the exercises run so far include

- A detailed critique of the UK Government's DirectGov search engine[1] using both an extensive literature review and multiple examples of queries that fail to find any relevant item. Along with suggestions on how the search engine could be improved.
- An examination of the search engine of a UK regional newspaper, where the students conducting the evaluation, upon finding an extensive audio interview archive on the newspaper web site, suggested the addition of speech recognition technology to provide transcripts which could be searched.
- A student while trying to locate the sort of searches conducted on a UK University web site used the Alexa service[2] to research typical queries used.
- Finally, an examination of the search engines behind a Pakistani newspaper, which published its articles simultaneously in Urdu and English. Different search engines were used for the two languages and the student used this situation to contrast the features of the two engines by issuing the same query (translated to the correct language) to both engines. The value of spell correction or stemming present in one search engine but not the other could be shown in this situation.

Students reported their work back through a ten minute presentation to the class and in a written report.

Second evaluation – test collection evaluation

Test collections have long been a standard way in which a searching system can be evaluated. The three classic components of a document collection, a set of queries (also known as topics), accompanied with a set of relevance judgements (also known as qrels) have their origins in work dating back to the early 1950s (Thorne, 1955); (Gull, 1956). Used in conjunction with an evaluation measure, the test collection has been a classic form of evaluation since that early work. Even in the past ten years where the rise of the query log analysis has taken a greater role in assessing search effectiveness, it is clear that test collections still form a core part of the evaluation processes of major search companies; see (R. W. White & Morris, 2007), (Chapelle, Metlzer, Y. Zhang, & Grinspan, 2009) for brief descriptions of test collections used in Microsoft and Yahoo! respectively.

It was decided to engage the students in building a component of a test collection for a search engine as part of their coursework. At this time, the UK Government's archive, TNA (The National Archives) approached the author asking if it would be possible to run a student project to assess the effectiveness of their pub-

[1] In more recent years, the quality of this search engine has been much improved.

[2] http://www.alexa.com/

lically facing search engine[3] and, as (Heuwing et al., 2009) had done, contrast it with Google site search. In the past, the department had run Masters student projects on measuring search output, the most notable of which had resulted in the creation of a well used test collection (Davies, 1983). However, the outcomes of such projects were entirely dependent on the skills of a single student. Therefore, a more fault tolerant less risky approach was considered.

Analysing search logs from TNA, the most popular queries to the search engine were identified for use in the test collection; more detail is provided in the "Queries" section. As building relevance assessments for such a collection can be a time consuming process it was decided to partition the gathering of assessments across the students taking the course. The students were each given a set of queries against which they were asked to assess the relevance of the top *n* documents returned by the TNA search engine and Google site search. The students were given different sets of queries and by working in a "crowdsourcing fashion" they were hopefully able to assess a sufficient number of queries to produce a statistically meaningful evaluation.

Educational goals and objectives

The educational goals of the exercise were to engage students in an actual comparison of two search engines. Students would conduct the evaluation and write a report on their experiences of the two search engines as well as their view on which was the better engine. Thus the educational objectives were that at the end of the exercise the students would be able to

- Conduct a test collection-like evaluation comparing two search engines
- Compute an evaluation measure over a set of results
- Conduct a broader evaluation of the two systems taking into account a wide range of qualities of the search engine
- Describe how hard it can be to judge the relevance of documents

As with the earlier exercise, an additional objective of this exercise was to show students the potential for improving the quality of search systems, and a real world example of the importance of a search engine.

Running the exercise

As with any coursework, it was important to ensure that students were clear on what they had to do. As the plan was to use outputs from each student in a formal comparison of search engines, the importance of the clarity of instructions to stu-

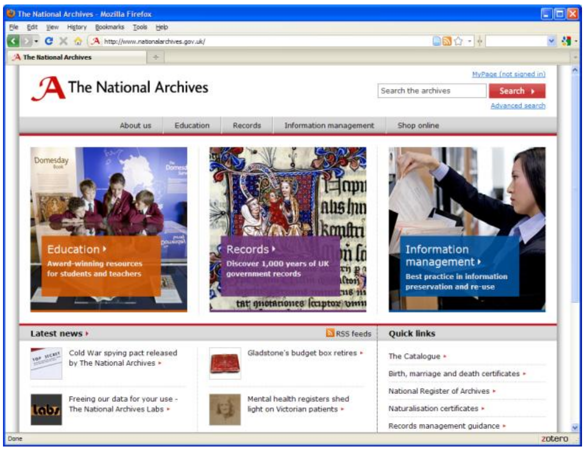
[3] http://www.nationalarchives.gov.uk/

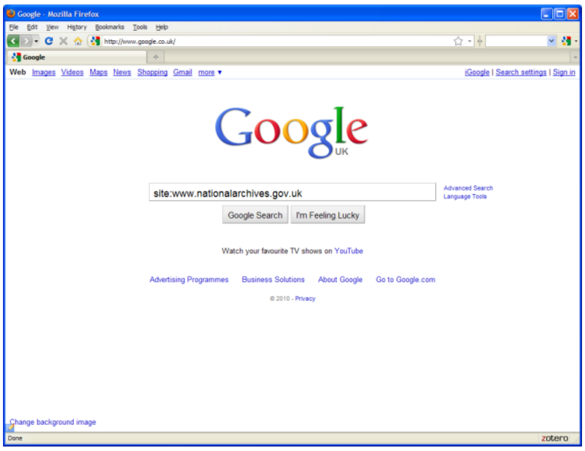
dents was even stronger. It was important to gather data that enabled the course lecturer to check the information that students returned. Several iterations of the exercise were run, each introducing a new element that reduced the chance of error and speeded up the use of the student generated evaluation data.

Here we describe the four main components of the exercise, how students worked with or generated those components and how the exercise has evolved over its multiple running.

Search engines

The exercise involved comparing two search engines which indexed the same collection. The first engine was TNA's search tool available in the top right hand corner of all the pages on its website, the second was Google searching over just TNA web pages, invoked with the command "site:www.nationalarchives.gov.uk" typed into the search box in addition to any entered query, see Figure 1.

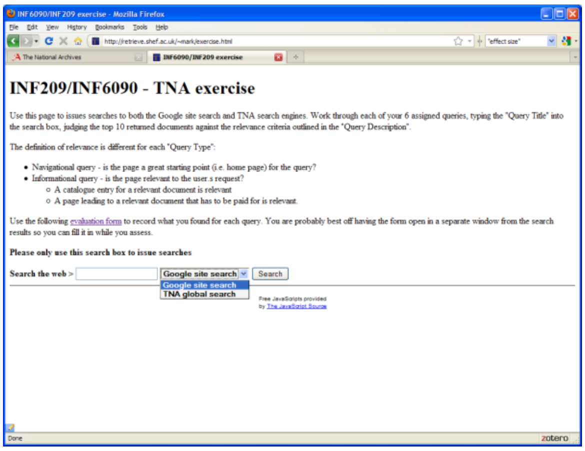

The National Archives - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.nationalarchives.gov.uk/
The National Archives
MyPage (not signed in)
The National Archives
Search the archives
Search
Advanced search
About us
Education
Records
Information management
Shop online
Education
Award-winning resources for students and teachers
Records
Discover 1,000 years of UK government records
Information management
Best practice in information preservation and re-use
Latest news
RSS feeds
Cold War spying pact released by The National Archives
Gladstone's budget box retires
Freeing our data for your use - The National Archives Labs
Mental health registers shed light on Victorian patients
Quick links
The Catalogue
Birth, marriage and death certificates
National Register of Archives
Naturalisation certificates
Records management guidance
Done
zotero



site:www.nationalarchives.gov.uk

Figure 1: screenshots of the two search engines to be compared

Monitoring of outputs from the students indicated that most successfully followed the instructions to search on these two web sites. However a number were found to have issued different queries or they did not use the site restriction on Google correctly. In an attempt to reduce these errors, a more managed system was created with a standard web based form into which the students entered their query and choose from a popup menu which search engine the query was to be sent to, see Figure 2.


INF6090/INF209 exercise - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://retrieve.shef.ac.uk/~mark/exercise.html
"effect size"
The National Archives
INF6090/INF209 exercise
Done
zotero

INF209/INF6090 - TNA exercise

Use this page to issues searches to both the Google site search and TNA search engines. Work through each of your 6 assigned queries, typing the "Query Title" into the search box, judging the top 10 returned documents against the relevance criteria outlined in the "Query Description".

The definition of relevance is different for each "Query Type":

- Navigational query - is the page a great starting point (i.e. home page) for the query?
- Informational query - is the page relevant to the user.s request?
  - A catalogue entry for a relevant document is relevant
  - A page leading to a relevant document that has to be paid for is relevant.

Use the following evaluation form to record what you found for each query. You are probably best off having the form open in a separate window from the search results so you can fill it in while you assess.

Please only use this search box to issue searches

Search the web > Google site search

Figure 2: Customized query entry form

Use of this form in the most recent running of the exercise reduced the number of errors in the data provided by students.

Queries

The queries that the students ran against the search engines were manually selected from query logs drawn from the TNA search engine. The queries chosen were the most popular queries submitted to the engine. Only the text of each query was preserved in the log, therefore a TNA staff member familiar with the content of the collections selected, studied and augmented a subset of popular queries with a written description detailing the searcher's probable information need. After an initial analysis of the logs, it was discovered that there were both informational and navigational queries being submitted to the TNA system, from the categories of (Broder, 2002), therefore queries were also tagged as either navigational or informational. Determination of these two types were made by the TNA staff member based on their knowledge of the internal structure of the TNA web site, and the data TNA held. For navigational queries, there were certain home pages known to be good starting points for types of search, such as family history or census queries.

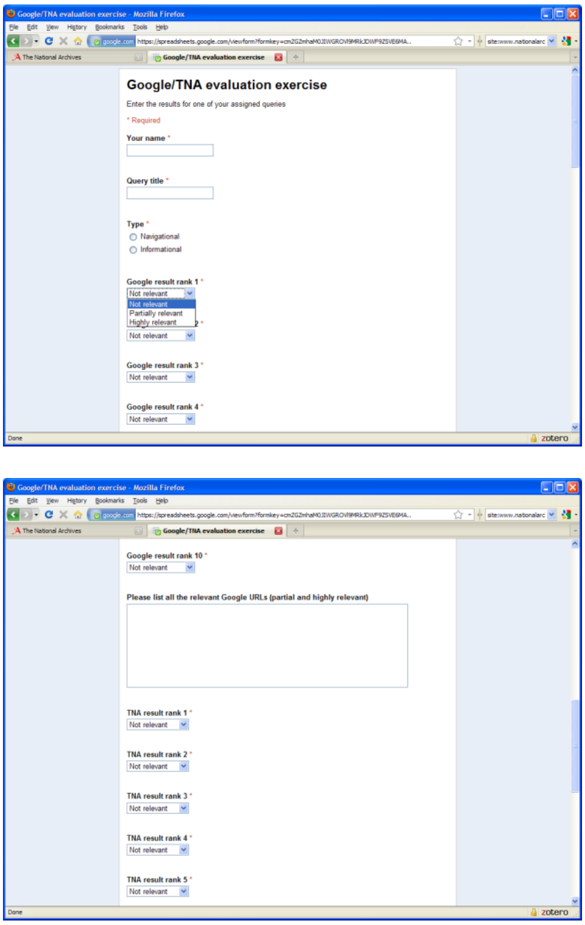
In total, 48 queries – 24 informational, 24 navigational – were selected; each student was given 5-6 to run on both search engines. The queries were given to at least two students so that if one student was found to have made a mistake, hopefully the topic would still be assessed correctly by the other.

Relevance judgments

Students were told to run each query on both search engines and to note down the relevance of the documents returned. The difference between informational and navigational queries was described to students. Part of the aim of the exercise was to show students how hard it can be to judge the relevance of items returned by a search engine. Therefore the level of guidance to students on exactly how a document should be judged as relevant or not was limited to simply explaining the difference between navigational and informational queries and the types of documents expected to be relevant to each query type. No training was provided. Students were told to be consistent across their judgments, but given no guidance on the degree of relevance or topic coverage in a document necessary to cause a document to be judged relevant. Although this would result in variations in the way that judgements were made in the test collection, there is good evidence from the literature (Voorhees, 1998) that such variation does not adversely affect decisions on which search engine is the better.

The students conducted the exercise in their own time. They were asked to record the URLs of documents they judged to be relevant. Getting the students to do this was found to help the course lecturer monitor for any mistakes made by students. In the initial running of the exercise students were told to email results to the course coordinator, in subsequent running of the exercise, an online form was created to ensure that data returned was structured and complete. The level of detail gathered in the form was increased over time.

In the current version (see Figure 3), the students enter their name, the title of the query, indicate the query type, and detail the degree of relevance for each of the top 10 results for both search engines on that query. A large text entry box is provided where the students copy the URL of any relevant or partially relevant documents found.


Google/TNA evaluation exercise - Mozilla Firefox
File Edit View History Bookmarks Tools Help
google.com https://spreadsheets.google.com/viewform?formkey=cmZGZmhaM0JIWGROVl9MRkJDWF9ZSVE6MA..
site:www.nationalarc
The National Archives
Google/TNA evaluation exercise
Done
zotero

Google/TNA evaluation exercise

Enter the results for one of your assigned queries

* Required

Your name *

Query title *

Type *

○ Navigational

○ Informational

Google result rank 1 *

Not relevant

Not relevant
Partially relevant
Highly relevant

2 *

Not relevant

Google result rank 3 *

Not relevant

Google result rank 4 *

Not relevant

Google result rank 10 *

Not relevant

Please list all the relevant Google URLs (partial and highly relevant)

TNA result rank 1 *

Not relevant

TNA result rank 2 *

Not relevant

TNA result rank 3 *

Not relevant

TNA result rank 4 *

Not relevant

TNA result rank 5 *

Not relevant

Figure 3: screenshots of the online form used by students to record the relevance of search results.

Evaluation measure

The evaluation measure, precision at rank 10, was initially used chosen because of its long standing popularity. As degrees of relevance were being gathered, Discounted Cumulative Gain (DCG) measured at rank 10 (Järvelin & Kekäläinen, 2000) was also calculated. Using the most recent relevance judgement form, the measures could be computed automatically. In addition, statistical significance and power analysis tests were conducted to determine the importance of comparisons made between the two search engines' outputs.

Feeding results back to students

The first two times the exercise was run, the results from the students were collated by the course instructor and a brief summary of the comparisons between the systems was fed back to students in the next lecture. However, it was realised that more could be done with this part of the coursework. In the last two runs of the exercise, a longer feedback session was held. Here a spreadsheet of all the data from the students was displayed in a lecture and the details of the way in which the data was manipulated to produce a scientific result were demonstrated live to the students.

This included showing how the results from the students were checked for errors with mistaken input removed from the calculations. Then the scores for the same query from multiple students were merged before an overall score for the two search engines was calculated and compared. Students were shown the comparisons as well as details of statistical significance tests that were run to determine the probability of the differences observed being due to chance. Splitting the queries by type (i.e. navigational/informational) and re-computing the comparisons was also done as were comparisons with results from previous runs of the experiment.

The chief priority of the work was to construct an experimental design that students would be able to conduct correctly and complete, without being over burdened with unnecessary procedure. However, taking a simplistic approach meant there were flaws in the experimental methodology:

- There was no control on the order in which students assessed the two search engines. Order effects are a well known problem in experimentation involving people (Hogarth & Einhorn, 1992).
- Students can see which search engine produced which result. Their likely familiarity with Google might influence their relevance judgements on documents returned by that search engine; there is experimental evidence of such a bias towards the Google brand (Jansen, M. Zhang, & Schultz, 2009).

However, the flaws were made a feature of the exercise; in the feedback session, students were asked to critique the design of the experiment and suggest improvements. In the first running of this feedback session, students pointed out the potential problems of familiarity with the Google brand and their judgements as well as issues with order effects. They also suggested potential problems with their own consistency of assessment and their lack of familiarity with some of the topics behind the selected queries. This facilitated discussion of relevant research on topics such as assessor consistency (Voorhees, 1998) and assessor expertise (Bailey et al., 2008).

Execution of the exercise

The exercise was run four times in total across two modes of delivery. Twice as an exercise given to students to conduct in their own time returning results by a deadline; twice as a 1 hour exercise run live in class, while instructors were present to answer any questions. The exercise was given to 2nd year undergraduate students and taught Masters students as well as to PhD students attending an information retrieval summer school. All four times the exercise was run successfully to conclusion producing statistically significant meaningful results, which confirmed the validity of this crowd sourced approach to evaluation. As it is unlikely that students will have encountered such an exercise like this elsewhere, the instructions and procedures need to be explained clearly in order to ensure that it runs smoothly. It was also clear that the job of judging relevance is a repetitive somewhat tedious task, so ensuring that the student do not have to undertake a large number of judgements is also important.

It was found that across each of the four runs of the exercise, a consistent picture emerged of one of the compared engines outperforming the other. Differences between the two were statistically significant and different results were found for the two query types: informational and navigational. One of the search engines was found to be particularly good at navigational queries, while the difference between the engines for informational queries was smaller. Running the exercise multiple times with different groups of students provided a form of internal validation. More recently the results derived from one of the student based exercises (same queries, same collection) were repeated using a group of paid assessors. The results of the search engine comparisons from this experiment were again consistent with the results from the student-based exercise, which leads us to be more confident that the student based exercise not only provided a learning experience for the students but also generated valuable research outputs for search engine owners.

Impact of the exercise

A clear peripheral benefit of running the exercise was the ability to develop a relationship with a large organization, in our case TNA, for which search is an important service. The organization was able to get, for free, a series of sufficiently rigorous experiments comparing their search engine with another. In the case of TNA, the findings of the first two running of the exercises were found to agree with internal tests run by the organisation. Consequently, the results fed into decisions made by the search engine team on how to develop their services further.

The impact for the students was seeing that they were active participants in research. An introductory lecture from TNA was given to the students about the practical running of a search engine. Students saw the importance of their work when TNA staff attended the meeting showing the results of the exercise. This relationship between TNA and the department led to a number of closer ties between the two, including a small funded research project to explore evaluation further.

A formal request for feedback from students revealed that both kinds of search engine evaluations were viewed as one of the best parts of the information retrieval course.

Conclusions and potential extensions

The motivation for both forms of evaluation exercise described here were derived from a realisation that there is a large relatively well established group of search engine users who are neither researchers nor developers, but rather information providers. These are people who have bought in a search system, are provided with many options to customise, change or improve their search services, but do not necessarily know how to go about such alterations. While the course content is relatively standard for an information studies department, the exercises are specifically targeted at enabling students to develop skills to help them assess the quality of a search engine and at the same time conduct different forms of evaluation including a conventional test collection like assessment. The evaluations have proved to be the most popular part of the course and have enabled the development of links with an organization that is keen to get access to such skills.

In planning for future versions of the exercises we examined the pedagogical literature on project based learning. The idea of engaging students in learning centred on projects is long standing (Kilpatrick, 1918). However, as (Barron et al., 1998) point out, the level of adoption of such an approach is relatively low; they suggest four principles that should underlie successful project-based learning. We list them here; detail what is currently done to achieve these principles in our exercise and what is planned to be done in the future.

1. *Learning-appropriate goals* – Barron et al state that it is not sufficient to set goals to ensure the student achieves the project outcomes, it is also necessary to set goals for learning also. In both exercises, so-called "additional objectives" were set, which could be viewed as learning-appropriate goals. However, no form of summative assessment of these objectives was carried out. Future versions of the exercises will conduct such evaluation.
2. *Scaffolds that support both student and teacher learning* – Barron et al state it is necessary to structure exercises to ensure that students address all stated objectives. In early versions of the exercise there was little provision of such scaffolds. Over time, the definition of the two exercises became more structured providing the structure Barron et al require. However, in the first exercise the evaluations conducted by the students are relatively ad hoc, what is needed is a more methodical way of considering how well an engine is successfully providing services to users.
3. *Frequent opportunities for formative self-assessment and revision* – while conducting exercises, Barron et al state students should be able to assess their progress and revise their project if necessary. Particularly with the second exercise, it was found that students often requested help from the course tutor to check that they were "doing things right". Explicit provision of formative self-assessment – such as example queries and relevance judgements – will be provided in revised versions of the exercise.
4. *Social organizations that promote participation and result in a sense of agency* – the first exercise arranged students in groups to jointly conduct the work, which it was hoped would improve participation; in the second exercise visits by TNA staff to the class was hoped to encourage a sense of agency in the students. In future versions of the exercises, however the extent to which these arrangements succeeded will be examined through student feedback.

There are also a number of IR specific improvements that could be made. The current requirement for students in the first exercise to examine only the search engines of potentially weaker searching systems is perhaps rather limiting. Even state of the art search engines have plenty of opportunities to be improved; therefore, allowing students the possibility to assess any type of search engine will be examined in future.

With the second exercise, the primary aim for improvement is to consider how to refine the experimental procedure to remove biases in the measurement of the two search engines. For example, were students likely to be biased in judging the quality of Google search results? While the IR community has a well established method for assessing the effectiveness of a search engine using test collections, there is less of a commonly agreed standard for evaluating the other aspects of the engines, which is an aspect to be better understood for this exercise. In addition, given the success of involving search providers in the second exercise, seeking information providers interested in being involved in having their system assessed in the first exercise is another hopeful extension the course.

Acknowledgements

The author is grateful to TNA for allowing their search engine be used in the evaluation exercise. He also wishes to thank the reviewers of this chapter for their invaluable comments. Finally, many thanks to the students who took part in the exercises and whose innovation, feedback, and comments made running the IR course such a joy.

References

Bailey, P., Craswell, N., Soboroff, I., Thomas, P., de Vries, A. P., & Yilmaz, E. (2008). Relevance assessment: are judges exchangeable and does it matter. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 667-674). ACM New York, NY, USA.

Barron, B. J. S., Schwartz, D. L., Vye, N. J., Moore, A., Petrosino, A., Zech, L., & Bransford, J. D. (1998). Doing with understanding: Lessons from research on problem-and project-based learning. *Journal of the Learning Sciences*, *7*(3), 271-311.

Broder, A. (2002). A taxonomy of web search. *SIGIR Forum*, *36*(2), 3-10. doi:10.1145/792550.792552

Chapelle, O., Metlzer, D., Zhang, Y., & Grinspan, P. (2009). Expected reciprocal rank for graded relevance. In *Proceeding of the 18th ACM conference on Information and knowledge management* (pp. 621-630). ACM Press New York, NY, USA.

Davies, A. (1983). *A document test collection for use in information retrieval research* (Dissertation). Department of Information Studies, University of Sheffield.

Gull, C. D. (1956). Seven years of work on the organization of materials in the special library. *American Documentation*, *7*(4), 320-329. doi:10.1002/asi.5090070408

Heuwing, B., Mandl, T., Womser-Hacker, C., Braschler, M., Herget, J., Schäuble, P., & Stucker, J. (2009). Evaluation der Suchfunktion deutscher Unternehmenswebsites. In *Wissensorganisation 09: "Wissen - Wissenschaft - Organisation" 12*. Tagung der Deutschen ISKO (International Society for Knowledge Organization).

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model, *Cognitive Psychology*, *24*(1), 1-55. doi:10.1016/0010-0285(92)90002-J

Jansen, B. J., Zhang, M., & Schultz, C. D. (2009). Brand and its effect on user perception of search engine performance. *Journal of the American Society for Information Science and Technology*, *60*(8), 1572-1595.

Järvelin, K., & Kekäläinen, J. (2000). IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 41-48). ACM New York, NY, USA.

Kilpatrick, W. (1918). The project method. *The Teachers College Record*, *19*(4), 319-335.

Thorne, R. (1955). The efficiency of subject catalogues and the cost of information searches. *Journal of documentation*, *11*, 130-148.

Voorhees, E. M. (1998). Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 315-323). ACM Press New York, NY, USA.

White, M. (2006). *Making Search Work: Implementing Web, Intranet and Enterprise Search*. Facet Publishing.

White, R. W., & Morris, D. (2007). Investigating the querying and browsing behavior of advanced search engine users. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 255-262). ACM Press New York, NY, USA.