A Competitive Environment for Exploratory Query Expansion

David N. Milne, David M. Nichols and Ian H. Witten
Department of Computer Science
University of Waikato
Hamilton, New Zealand
+64 7 838-4246

{dnk2, dmn, ihw}@cs.waikato.ac.nz

ABSTRACT

Most information workers query digital libraries many times a day. Yet people have little opportunity to hone their skills in a controlled environment, or compare their performance with others in an objective way. Conversely, although search engine logs record how users evolve queries, they lack crucial information about the user's intent. This paper describes an environment for exploratory query expansion that pits users against each other and lets them compete, and practice, in their own time and on their own workstation. The system captures query evolution behavior on predetermined information-seeking tasks. It is publicly available, and the code is open source so that others can set up their own competitive environments.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval - *query formulation, search process.*

General Terms

Design, Human Factors.

Keywords

Query Expansion, game

1. INTRODUCTION

Formulating queries to digital libraries and web search engines is a major component of the daily work of most information workers. Although introductory material on "How to use the library" generally includes tips on information-seeking behavior [6], the vast majority of users lack any guidance on how to formulate search queries. They cannot measure their success, nor compare it with others (except anecdotally), and they have little opportunity to hone their skills in a controlled environment. Considering the everyday importance of search skills, it is astonishing that no support is available for practice, or even for detecting improvement.

JCDL '08, June 16–20, 2008, Pittsburgh, Pennsylvania, USA.

We have built a simple web-based system[1] that allows users to assess

- how good their searching skills are
- how they stack up against other searchers
- what can they do to improve their searching.

It provides a pre-determined set of tasks which users can attempt, in their own time and on their own workstation, and compare their success with others. Inspired by the so-called "ESP game" in which people help determine the contents of images by providing meaningful labels for them [1], it is intended to be both fun and to create valuable output—in this case, records of query refinement behavior by skilled information seekers on a known information task. Unlike the ESP game, it is based on a pre-existing database of items—in this case, information-seeking tasks and a corpus of documents.

2. THE KORU GAME

To play the game, a user logs in and the system responds with the screen shown in Figure 1. The user can proceed straight to a new task using the link supplied, or get additional information using the menu items along the top of the screen. For example, users can view a scoreboard in order to compare their performance with other players. Alternatively, they can examine the list of tasks (shown in Figure 2) and assess their performance on individual ones. They may attempt tasks in any order, and re-do them as often as they like in order to encourage performance improvement. Only the best query for each task is considered in the overall score.

Figure 3 shows a user in the process of performing a task called *wrongful convictions*. The task itself is shown in a box at the top, and can be expanded (using the *More* button) into a fuller description, in this case:

> *Find documents that discuss freed prisoners who have been wrongfully convicted based on faulty forensic evidence, poor police work, or false testimony.*
>
> *Documents about political prisoners who were freed because of incompetent prosecutions are relevant. However, documents that discuss prisoners who are pardoned or released on bond when their convictions are overturned are not relevant, nor are documents about prisoners freed to make a political statement or prisoners freed for an exchange.*

[1] http://www.nzdl.org/koru/game

Figure 1. The home page

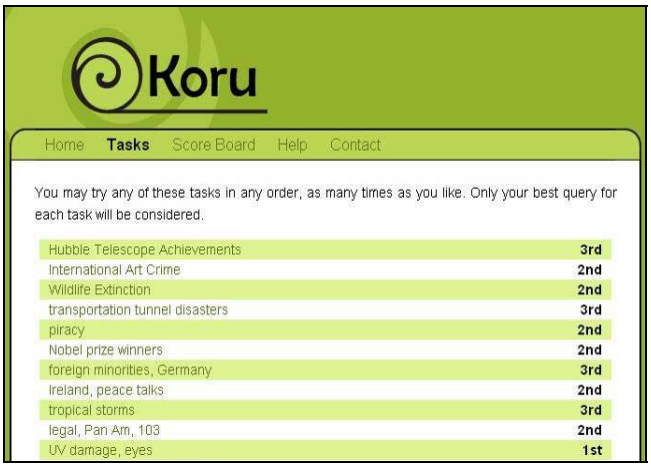

Figure 2. A user's placement on individual tasks

(The description is not expanded in the Figure to avoid obscuring part of the screen.) Near the top is a conventional query box, which is initialized with the task name ("wrongful convictions"); the user has expanded this with "death row" and clicked the *Search* button in order to see the results below. On the left is the *Query Results* panel, showing in this case documents 1–10 of over 100. The list is paginated into groups of 10 results, and the other pages are accessed using the small numbered tabs above.

On the right is the *Documents Tray* panel, which shows the contents of individual documents. In this case the second entry in *Query Results* has been clicked to reveal its corresponding document.

Individual query terms are highlighted in both the *Query Results* and *Document Tray*. More importantly, certain items under *Query Results* are also highlighted (in a different color). These correspond to documents that have been judged to be relevant to the query. There are two reasons for highlighting them: to save users from the very considerable amount of time required to make relevance judgments, and to ensure that users plainly understand the basis of the success score.

The success score for the query is displayed at the top right corner of the query results panel. The user's best query (shown beside the *Search* button) determines their final score for the task—in this case enough to place them 4th on the leader-board. As shown in the figure, scores can be moused over for a description of how they are calculated as the balanced f-measure of the results: the harmonic mean of relevance (proportion of all available relevant documents that are found) and precision (proportion of results returned that are relevant). Thus to obtain a high score one's query must return as many relevant documents as possible, while avoiding irrelevant ones.

3. IMPLEMENTATION

The game interface is built using the AJAX framework [4], and consequently requires nothing more than a standard web browser to play. This communicates with a Greenstone[2] digital library in which documents are indexed using Lucene.[3] Player statistics, tasks and query behavior (discussed in Section 4) are stored in a separate database.

The game mechanism depends on having interesting, challenging tasks for which all the relevant documents are known in advance. The tasks, documents and relevance judgments were obtained from the 2005 TREC HARD track [2], which pits retrieval techniques against each other on the task of high-performance retrieval. This data pairs 50 tasks, each comprising a 2- or 3-word title and a two-paragraph description and narrative in the same style as quoted above. The TREC HARD track's purpose was to provide tasks that are difficult enough to benefit from additional information about users and their intent. They are attractive for this work because they require users to think carefully about their query terms, and are unlikely to be satisfied by a single query or document. They provide players with a challenge, and significant interaction is required to obtain high scores.

The tasks are matched to the AQUAINT corpus of around a million newswire stories from Xinhua News, Associated Press, and the New York Times. For each task, a sample of approximately 750 relevance judgments are available from TREC; in which a document is identified as strongly relevant, weakly relevant, or irrelevant. Ideally, the game should have a judgment of relevance between every task and every document, but this is clearly unrealistic. Instead we restrict the game to documents with judgments for at least one task, and assume that any document that is not judged as relevant to a task is irrelevant. We also restrict the game to 20 tasks—to avoid becoming onerous—and to one of the sources (New York Times)—to avoid having the player read about the same events several times. The result is a collection of approximately 4,700 documents concerning a wide range of topics. The news stories are relatively short and concise—an average of 1200 words each—which helps to maintain player motivation.

4. QUERY EVOLUTION BEHAVIOR

The Koru game was originally conceived not just as an educational tool, but primarily as a way of collecting realistic data on how skilled users evolve their queries to satisfy an information need. We were interested in the decisions and strategies these users employ, and how this can be used to guide the development of an intelligent search engine [8].

[2] http://www.greenstone.org; we use Greenstone version 3.03
[3] http://lucene.apache.org/

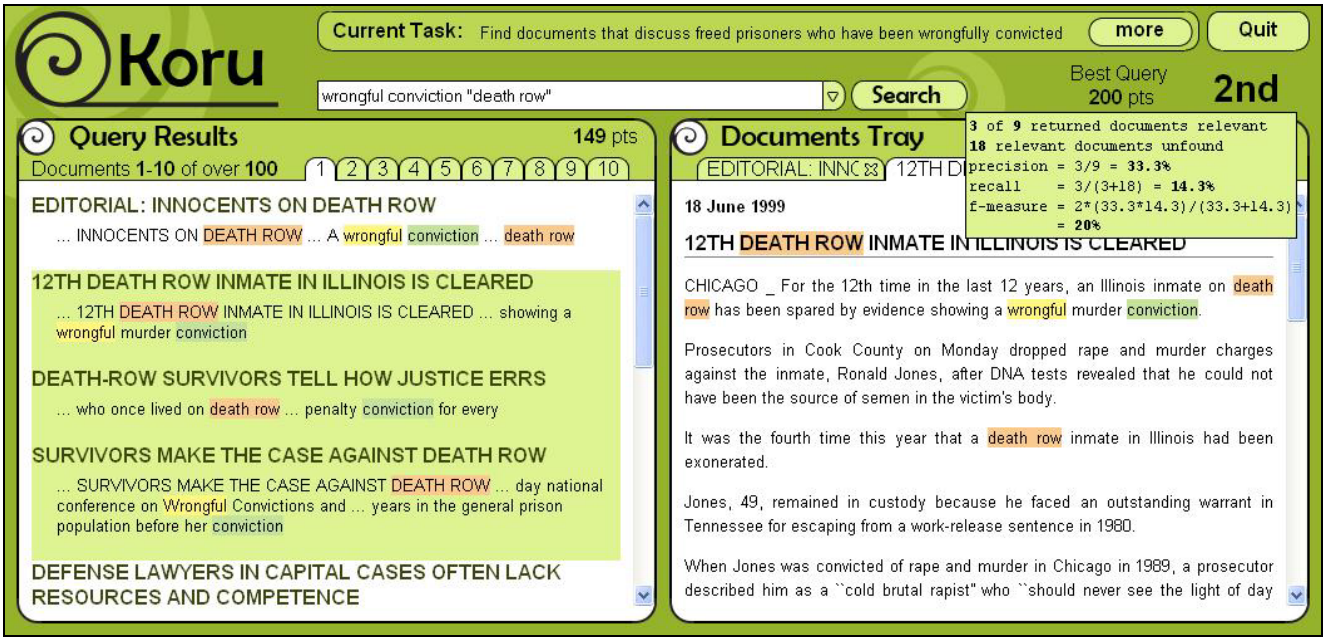

Figure 3. Performing a task (three relevant documents are highlighted in the *Query Results* pane)

Take the example shown in Table 1, which records one player's attempt to produce an ideal query for

What disasters have occurred in tunnels used for transportation?

This player has consistently performed well, and can thus be identified as someone whose behavior should be emulated. The individual decisions that should be emulated can similarly be identified as those that increase the player's score. The information available to them when making these decisions can be identified through logs of document surrogates and content that was shown to them at the time.

Table 1 reveals three key decisions, all of which are fairly mechanical and could be emulated automatically:

- The player insisted that each document must contain both *tunnel* and *die*, which dramatically increased precision. This could be emulated automatically by insisting that semantically unrelated terms be joined by an AND operator.
- They supplemented *die* with *hurt*, *death* and *damage* to increase recall. This could be emulated by identifying synonyms and closely related terms—either from retrieved documents or an external thesaurus—and grouping these with OR operators.
- They increased precision further by specifying that returned documents must not contain *power, wind* or *carpal tunnel.* This could be emulated by a form of relevance feedback in which significant terms are identified in unhelpful documents and excluded from the query.

It seems that this player owes most of her success to mechanical application of Boolean operators. Most users have little or no understanding of these, which reduces their ability to search effectively [5]. This disadvantage could well be avoidable: their use might be mechanical enough to be automated. This is one example of the end goal of this work: to identify the strategies that good players employ and investigate ways in which search engines themselves can take responsibility for issuing effective queries.

5. DISCUSSION

There are several possible uses for the Koru game. First, it can be used by those wishing to learn about or improve their search skills, because it is known that searchers have difficulty identifying useful terms for effective query expansion [9]. Second, it can be used in library and information science education as a tool to explore relevance and its influence on user behavior and interface design. Third, it can be used as a source of research data on query evolution.

The Koru game differs from a normal search interface interaction in three ways: it is social (the users are aware of other users through their rankings), it shows relevance judgments in result lists (only available with specific collections), and it uses pre-specified topics (rather than the users' interests).

In addition to its social element, the system is explicitly *competitive*: users are scored and ranked by their success in generating effective queries. The ESP game [1] is based around simple agreement on image labeling. Building on this approach the *1001 Paraphrases* game uses partial obfuscations of previous (correct) user contributions [3]. Its designer suggests that his approach may be applicable in other contexts, and the Koru game is well-suited to exploring this method of incentivizing users. High (or higher) scoring queries can be obfuscated, or degraded, and used as new starting points—or hints—for users who are scoring poorly or making little progress.

One method of addressing the observed issue that subjects tend to concentrate on terms for *new* queries rather than modified or refined queries [9] is to enhance the presentation and value of query modifications. For example, the scoring system could be adjusted to favour modifications over new queries. Game-based

Query	Recall	Precision	Score
transportation tunnel disasters	69%	9%	159
transportation tunnel disasters die	85%	11%	195
(tunnel AND die)	23%	30%	261
(tunnel AND die) OR (train AND die)	23%	13%	162
(tunnel AND die)	23%	30%	261
(tunnel AND die) OR (underground AND die)	23%	15%	182
(tunnel AND (die OR hurt OR death OR damage))	62%	26%	364
(tunnel AND (die OR hurt OR death OR damage)) NOT (power OR wind)	62%	31%	410
(tunnel AND (die OR hurt OR death OR damage)) NOT (power OR wind OR "carpel tunnel")	62%	31%	410
(tunnel AND (die OR hurt OR death OR damage)) NOT (power OR wind OR "carpal tunnel")	62%	33%	432

Table 1. Evolution of a particular query with recall, precision, and the game score (shown to the user)

approaches to query expansion may be one way in which we can address the suggestion that "we should work to design clever and creative techniques for encouraging users to be loquacious rather than reticent, both during their initial querying and during follow-up interactions" [7].

As with other systems, the recorded queries are a potential source of research data. As Koru is used it collects query paths and the associated scores at each point on the path. Over time we can identify tactics that produce large increases in query scores (e.g. new query terms, new logical expressions) and analyze user's behavior in conjunction with their scores.

A key question is whether the display of the known relevant documents creates a situation that is insufficiently authentic for the data gathered to be useful for future analysis or system design. In real usage searchers modify both queries and their information needs as they interact with surrogates and documents. The game environment uses a fixed (but non-personal) task with explicit feedback based on known answers.

Although in general systems will not explicitly highlight relevant documents in search results as the Koru game does (e.g., see Figure 3) there are certain situations where similar interactions already occur. Systems that have access to user history data can infer relevance from past user actions; customers at Amazon.com receive 'you purchased this item on …' messages added to their result lists.

6. CONCLUSION

In this paper we have introduced the Koru Game, a novel system that differs from traditional search engines in three key ways.

First, it is educational. Despite the widespread use of search engines, little information is available about how to put them to use. Consequently, the majority of users have scant understanding of how to issue effective queries [5]. The Koru game aims to address this by providing a controlled environment in which players can practice effective information seeking.

Second, it produces valuable information for developing and improving search engines, in the form of query logs. While the analysis and application of query logs is nothing new, it is hampered by missing information. By themselves, queries can provide only a vague impression of the underlying information needs and the extent to which they have been fulfilled. It is difficult to separate good queries from bad, or the users who should be emulated from those who need help. The Koru game makes it trivial to make such distinctions, because every query is scored against a predefined task, and every user is scored against the other players.

Third, the game is social: it allows users to see how their search skills measure up against others. This fosters a competitive environment in which users are motivated to continue playing. Doing so will provide immediate benefit in information seeking proficiency. It may yield much more in the long term by guiding the development of new, intelligent search engines that close the gap between those who play the game well and those who don't.

ACKNOWLEGEMENTS

We would like to acknowledge the entire New Zealand Digital Library Project team for their unstinting work in providing an environment that makes this kind of research meaningful—and enjoyable.

REFERENCES

[1] von Ahn, L. and Dabbish, L. (2004) Labeling images with a computer game. *Proc. CHI '04*. 319-326.

[2] Allan, J. (2005) HARD Track overview in TREC 2005 high accuracy retrieval from documents. *Proc of TREC-2005.*

[3] Chklovski, T. (2005) Collecting paraphrase corpora from volunteer contributors. *Proc. of K-CAP '05*. 115-120.

[4] Crane, D., Pascarello E. and James, D. (2005) *Ajax in Action*. Greenwich, CT: Manning Publications Co.

[5] Jansen, B., 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 & Management 36*(2).

[6] Kapoun, J.M. (1995) Re-thinking the Library Pathfinder, *College & Undergraduate Libraries* 2(1) 93-105.

[7] Kelly, D. and Fu, X. (2007) Eliciting better information need descriptions from users of information search systems. *Information Processing & Management* 43(1) 30-46.

[8] Milne, D., Witten, I.H. and Nichols, D.M. (2007) A knowledge-based search engine powered by Wikipedia. *Proc. of CIKM '07*. 445-454.

[9] Ruthven, I. (2003). Re-examining the potential effectiveness of interactive query expansion. *Proc. of SIGIR'03*. 213-220.