Fu-Finder: A Game for Studying Querying Behaviours

Test your Search-Fu

Carly O'Neil
School of Computing Science
University of Glasgow
United Kingdom
oneilcl@dcs.gla.ac.uk

James Purvis
School of Computing Science
University of Glasgow
United Kingdom
0801303P@student.gla.ac.uk

Leif Azzopardi
School of Computing Science
University of Glasgow
United Kingdom
leif@dcs.gla.ac.uk

ABSTRACT

Usually the focus of evaluation within Information Retrieval has been placed largely upon the system. However, the individual user and their submitted queries are typically the greatest source of variation in the search process. This demonstration paper presents Fu-Finder, a fun and enjoyable game that measures the user's querying abilities (or search-fu). This game provides useful data for the study of user querying behaviour and assesses how well users can find specific web pages using different search engines.

Categories and Subject Descriptors: H.3.3 Information Storage and Retrieval: Information Search and Retrieval

General Terms: Performance, Experimentation

Keywords: Search-Fu, Querying Behaviour, Findability

1. INTRODUCTION

Searching and finding particular webpages can sometimes feel like a game of skill and chance. Choosing the "right" query terms to illicit the desired response is largely dependant upon the user. The user needs to be able to understand their information need, the system that they are using, how it works, what it is in the document (imagined, ideal, actual and/or known) and what terms would help to distinguish it from other documents. Not a particularly easy task. This has led to the term, *search-fu* being coined which denotes the skill of user's search abilities i.e., someone with weak search-fu will be unlikely to find relevant documents, while someone with strong search-fu will be able to find almost anything. So the question is, how much search-fu does a user possess? Or stated more formally, how well can a user satisfy their information needs when using a search engine?

The prototype system we have developed to help answer this question is a human computation game (HCG). HCGs have been developed to obtain data for various purposes (such as image annotation, character recognition, etc) [8]. Recently, the game Page Hunt was developed to help optimise the Bing search engine as a way to help annotate pages that are difficult to find [3]. However, rather than trying to optimise a particular search engine, we have developed Fu-Finder to evaluate the user's ability at querying for pages across a number of search engines. The game is currently hosted and available to play[1], while the code is open source and freely available to download as part of the PuppyIR project housed on SourceForge[2]. In this paper, we describe the game developed and some of the initial results.

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.

2. EVALUATIONS USING GAMES

Human computation games provide a novel tool for solving computationally challenging problems (which are easy for humans, but hard for computers). The success of such games has been their ability to generate large quantities of reliable and useful data by providing users with enjoyment and points as payment rather than money or course credit. The added advantage is that these games can be undertaken in a controlled environment (i.e. within the remit of the game rules/setup) much like a standard or traditional IR experiment, but without the problems of using paid-workers through crowd sourcing. While HCGs are often developed to solve a difficult problem, such as von Ahn's ESP game, where players independently annotate an images [7], in IR human computation games have been used for evaluation purposes.

Two notable games developed in IR are: Book Explorer [2] and PageHunt [3]. In [2], they developed a human computation game to assess relevancy of electronic books to topics. The main motivation for incorporating this task in a game was to provide a means of obtaining relevancy judgements for large document collections without needing to provide fiscal payment to study participants. This has been outlined as a prime motivation for including a computationally-hard task within a game with a purpose [7]. The Book Explorer game, whose model is a variant of the inversion-problem model [8], has two player roles, which can be adopted by individuals or teams. The *explorer* locates and highlights relevant sections of a document; while the *reviewer* provides relevancy feedback on documents [2]. Reviewers can only enter relevancy judgements for documents that have been evaluated by at least one explorer player. Furthermore, reviewers also serve as a cheating-detection mechanism as they examine explorer input, thus creating a multi-level validation system.

While the results from the study appear to show some promise for obtaining reasonably good relevance judgements,

[1] http://www.dcs.gla.ac.uk/access/fufinder
[2] http://sourceforge.net/projects/puppyir/

whether it can be truly considered a game is matter of debate. The Books Explorer game places a high cognitive load on the reviewer, and requires them to read a large amount of text. Such a burden violates the general principles of designing a fun and enjoyable game (unless of course the gamers are also avid book fans).

On the other hand, Page Hunt [3, 4] was a game that was developed to optimise of the Bing search engine for pages that were difficult to find using Bing. Pages included within the game were selected because the labels for these pages used by Bing were ineffective, and additional metadata was required [4]. The rules of Page Hunt were simple: given a screen shot of a web page, enter a query to find the page, if successful they are presented with a new page to look for, otherwise they can either query again or pass and look for a new page. The goal was to find as many of the pages shown as possible in the 3 minutes (although the timer itself was not visible during the game). Queries were sent to the Bing search engine, and a player awarded a point if the web page was found within the top five results. These results were viewable in a pane presented over the top of the screenshot view. The interaction metaphor present within the game is based on the traditional user interaction with search engines. The inclusion of such metaphors generally improves the usability of game interfaces [6].

Using the Page Hunt game, they performed an experiment that lasted ten days and attracted 10,000 registered users from the Web community. User participation was strictly voluntary, and users were not contacted directly by the researchers to be encouraged to play. This voluntary participation suggests that the results obtained were not affected by any payment-related bias. The sole motivation for users continuing gameplay was enjoyment. The game was seeded with 698 URLs for which better labels were required for Microsoft's Bing search engine [3]. The usage statistics obtained for this short study were very promising. The average user contributed 12 labels, and each URL obtained over 170 distinct labels [3]. The collected data indicated that some pages present were easily found in the top N results, while others were more difficult to obtain. Percentage findability of a page was defined as the percentage of users who found the page in the top N search results, where $N = 5$ [3]. It was found that 27% of pages exhibited 100% findability, while 26% has 0% findability. From this they concluded that the number of pages with 100% findability is approximately equal to the number of pages with 0% findability [4]. The length of the page URL was also found to be a contributing factor in page findability within Bing. As the URL length increased, the probability of finding said page within the Bing search engine rankings was found to decrease [3].

3. GAME OVERVIEW

Fu-Finder is a variant of the Page Hunt [3] game, but provides a couple of different features that changes the game dynamics and the motivations of users. Before we describe these changes, we shall outline the game and the gameplay. The premise of the game is based on testing users ability to search (i.e. their search-fu) in a competitive environment. The game starts by presenting a user with a screen shot of a page. This page is what they need to find by querying the search engines. The game returns the results from three search engines and denotes whether the target page was returned in the top five results or not. The gamer has the option of (i) searching for the same target page again, or (ii) moving on to the next target page. The game lasts for three minutes and gamers are awarded points according to the number and rank positions of target pages that they find. They obtain more points for retrieving the page higher up in the ranking, and for retrieving the page across multiple search engines. In contrast to the gameplay within PageHunt, the game contains multiple search engines, and aims to motivate users to pose better queries to obtain a substantially larger amounts of points.

The interface of the main game page is shown in Fig 1. The target page is shown on the right, gamers enter queries in the middle box, and the scores and results of the query entered is shown on the right. Below the title and url of the results retrieved from each engine is shown in two colours (pink to indicate the result not-matching, and green to indicate a match with the target page). In the Figure, it shows the results from a very successful query which returns the target page at rank one across all the search engines. At the top of the screen, the game displays the user currently logged in to the system, and the time remaining for the current game. The Fu-Finder game required users to read an information sheet, consent to participating in the study, and register as a game participant. By including a user registration component we anticipated that this would reduce the number of participants, but it was necessary to provide informed consent to users, and also enabled us to collect data across particular users (as opposed to sessions).Following completion of the game, users are shown the status, any rewards that they achieved, the leaderboard of results, and asked whether they would like to play again.

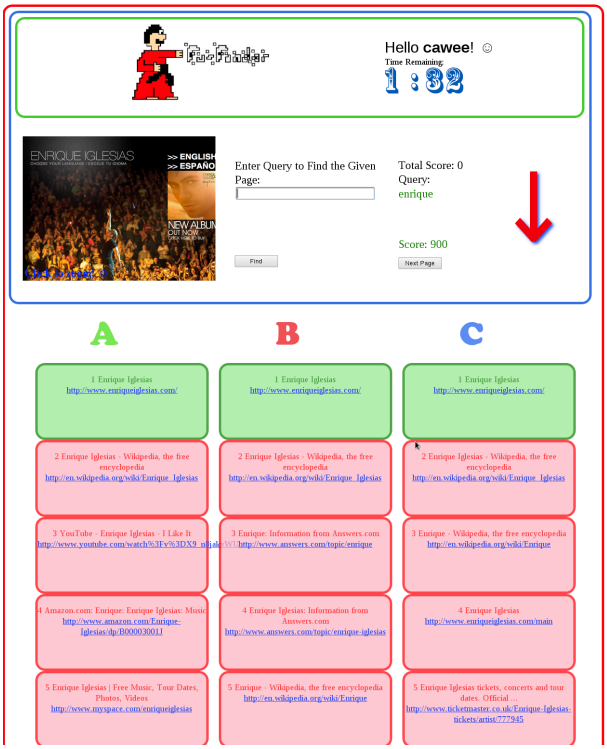

Figure 1: A screenshot of the main game screen, containing: The logo, username and timer (top), the target page (left), the query (middle), the score for the query and actions (right), & the results (below).

3.1 Game Properties

Time Limit The primary purpose of the time limit is to focus the user's attention[3]. The visible time pressure imposed on the gamers was to encourage them to issue queries quickly. This is because we wanted to try and illicit queries that were as realistic as possible, i.e. relatively short, containing spelling mistakes, etc. We believed that this would lead to more reformulations by the user, and help to build up a picture of how queries are modified. If, instead of using a time limit, we asked users to try and find a set number of pages without the pressure of a time limit, then we felt this would have led to less natural and artificially long queries.

Points Allocation To encourage gamers to enter high quality queries a points system was introduced. As previously mentioned, the points system rewarded queries which retrieved the page at higher ranks and across more search engines. Obtaining points for queries provides *positive reinforcement* [5] to gamers, which should encourage them to improve their queries and continue playing. The number of points allocated *po* for a given attempt a over the set of search engines E is formally defined as:

$$po(a) = \sum_{e \epsilon E} f(a, e) \quad (1)$$

where e is a given engine belonging to the set of nominated search engines E. $f(a, e)$ is a function defining the points allocated for attempt a to engine e, and is defined as: $\frac{c}{r(p,e)}$ when $r(p, e) \leq 5$, else zero. $r(p, e)$ is the rank of a page p for engine e and c is a constant, which we set to 100 i.e. a page returned a rank one, two and three obtained 100, 50 and 33, points respectively. If all engines returned the target page at rank one a bonus multiplier of 3 was applied to the points allocated.

Reward and Status Initiatives The first mechanism adopted within the game is the use of player statuses. These statuses are based on the points obtained and are shown in Table 1. The second mechanism was the inclusion of leaderboards of high scores to introduce a competitive element to the game (i.e. to develop rivalry among gamers) and to encourage gamers to play multiple times and improve their search-fu and status. In terms of behaviour, these rewards constitute as *positive partial reinforcement* [5].

3.2 Initial Evaluation of Fu-Finder Game

A pilot study of the Fu-Finder game was undertaken from the 28th of February until the 4th of April in 2011[4]. The game was advertised locally, and online via Reddit. This resulted in 1026 participants, who played a total of 9075 games and issued 16064 queries in total. We used a subset of the pages used in PageHunt, and the 540 pages used seeded the game[5]. The set of pages used were from a number of different domains, such as health, movies, sports, etc and provided spectrum of retrieval changes (i.e from navigational to informational). The distribution of gamers vs. games was heavily skewed with most players only playing one or two games, while 28 participants played more than 50 games, 7 gamers played more than 100 games and one participant played 301 games. The most avid participant spent a total of 903 minutes using Fu-Finder. The highest score was 12,700 points for which the participant achieved "Search-Fu Grandmaster" status. However, only 304 out of the 1026 participants successfully retrieved at least one page, the unsuccessful participants only played one game. This suggests that either the game was too hard (i.e. presenting too many *hard-to-find* pages discouraging users), the game was not very intuitive, or worse was not particularly entertaining. In future versions, we shall provide a few easy pages first to ease the player into the game and to encourage them to continue playing. In the next section, we shall describe some of the initial findings regarding querying behaviour, the findability of pages, and the search-fu of gamers.

Status	Points Required
Search-Fu Grandmaster	10,000
Web Crawler	7000
Net Navigator	4400
Lycos Retriever	2000
Lynx Net Noob	0

Table 1: Points Required for Status, where the titles were inspired by iconic web browsers.

4. EMPIRICAL FINDINGS

From this initial study we observed that the average length of a query was 3.31 terms in length (which is within the usual range for web queries [1]), and the queries were approximately 20 characters in length on average. Out of the 16064 queries, 8966 queries were unique. This suggests that gamers were generally consistent in formulating similar queries for the same target page. We also observed that 78% of queries contained no URLs within the query, while 14% contained a mixture of keywords and URLs, and 8% of queries contained URLs only. These observations show some of the different tactics employed to precisely identify the target page (i.e. either, keywords, URL + Keywords, or URL only). Since many of the target pages contained visual features such as the URL it is not surprisingly that about a quarter of the queries contained such a feature.

Page and User Percentage Findabiltiy: In [3], they proposed the percentage findability measure for pages. Page %Findability is the percentage of users that found a page within the top n results. This measure provides an indication of how easily users could find a particular page. A similar measure can be defined for users, which provides an indication of how good a particular user is at finding pages (and thus a measure of their search-fu.) User %Findability is the percentage of pages that the user found within the top n results. Given these measures, we calculated the page and user findability and plotted histograms to show the distribution across the measurements obtained in Figure 2. From the top plot, we can see that for Page %Findability the histogram consists of a large spike at 0% Findability, and a skewed, flatten normal distribution peaking at 60-70% Findability for pages. This suggests that while most of the pages in this sample were findable to some degree by users, a substantial number of pages were essentially un-findable. This further suggests that there is a divide among pages, such that a page is either findable or not. For the most findable pages (i.e. 70%+, or found 7 or more out of ten

[3]Note, that the use of a time limit to provide an explicit user goal complies with design guidelines defined by von Ahn and Dabbish [8].

[4]Ethics approval was obtained from College of Science and Engineering (ETHICS-CSE00795).

[5]Many thanks to Raman Chandrasekar for suppling the list of URLs used in [3]. Of the 698 pages used in [3] only 540 still existed and formed the set of pages used in this study.

attempts), these pages tended to be either very distinctive site pages, or pages from well known or respected websites such as `imdb.com` or `medlineplus.com`. While the least findable tended to be pages that were deeper within sites or very obscure site pages.

The bottom plot shows the histogram of User %Findability for all participants that successfully retrieved at least one page (note: 772 participants obtained 0%, and only played one game). From the histogram, we can see that most participants obtained a User %Findability score of between 30-50%, suggesting that they were successful 3-5 times out of every ten queries that they submitted. However, there was a small portion of users with User %Findability above 80%, suggesting that there is perhaps a class of power searchers (with very strong search-fu). But, what is quite clear from the plot is that there large variation across the abilities of users to retrieve pages (and this is despite the fact that users could see the target page).

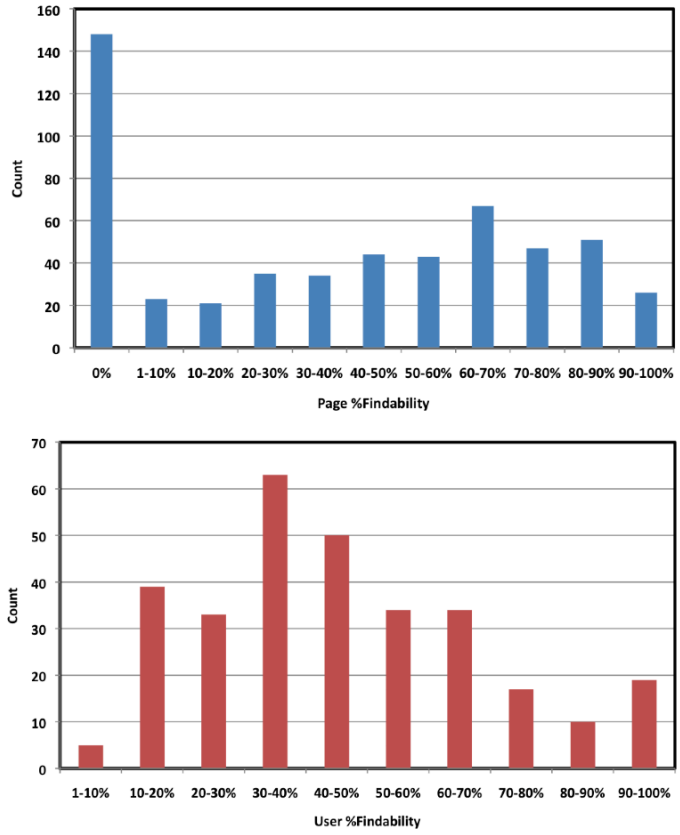

Figure 2: Percentage Findability Histograms: Findability across pages (top), Findability across users (bottom).

5. SUMMARY & FUTURE WORK

We have provided an overview of the Fu-Finder system that aims to study user querying behavior by employing a Human Computation Game to obtain data. The initial results show that users exhibit a wide range of abilities when it comes to searching for known-item pages, and that across pages there is also large amount of variation in terms of findability. In future work, it would be interesting to explore the different relationships between pages, and the difficulty of finding them, in more detail. For example, to consider the depth of the URL/page within a given site, and determine more precisely the relationship between depth and findability. Also, of interest would be to consider how search experience and other user factors affect the user's search-fu, including topicality. Here we would like to extend Fu-Finder to support different slide decks (i.e. use different sets of web pages as the seed for the game). This would support testing across different topical domains to determine whether users find it easier to retrieve pages which they are interested or knowledgeable about. And conversely, whether they find it harder to retrieve pages from new or unknown domains/topics. In particular, and inline with the remit of the PuppyIR project we are especially interested in examining the differences in querying abilities of adults and children. So having different slide decks to represent pages that would be of interest to children and adults (specifically, their parents), would enable further testing in a relatively unexplored area of research. This is where the use of a game, as a scientific methodology, is particularly applicable, because studying querying behavior of children would be quite difficult under standard experimental conditions. Having a game for children to interact with is likely to result in more reliable data from participants. In future work, we will be extending the Fu-Finder game to explore research in this direction.

Acknowledgments This research is supported the PuppyIR project and is funded by the EC's FP7 2007-2013 under grant agreement no. 231507.

6. REFERENCES

[1] A. Arampatzis and J. Kamps. A study of query length. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 811–812, 2008.

[2] Kazai, G., Milic-Frayling, N. and Costello, J. Towards Methods for the Collective Gathering and Quality Control of Relevance Assessments. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'09, pages 452–459, Boston, MA, USA.

[3] Ma, H., Chandrasekar, R. Quirk, C. and Gupta, A. Improving Search Engines Using Human Computation Games. In *Proceedings of 18th ACM International Conference on Information and Knowledge Management*, pages 275–284, Hong Kong, China, 2009.

[4] Ma, H., Chandrasekar, R., Quirk, C. and Gupta, A. PageHunt: Improving Search Engines Using Human Computation Games. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'09, pages 746–747, Boston, MA, USA, July 2009. ACM.

[5] Malim, T. and Birch, A. *Introductory Psychology*, chapter 6: Learning and Behaviour, pages 124–137. Palgrave Macmillan, 1998.

[6] Malone, T.W. Heuristics for Designing Enjoyable User Interfaces: Lessons From Computer Games. In *Proceedings of the 1982 Conference on Human Factors in Computing Science*, CHI'82, pages 63–68, Gaithersburg, Maryland, United States, 1982. ACM.

[7] von Ahn, L. and Dabbish, L. Labelling Images With a Computer Game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI'04, pages 319–326, Vienna, Austria, April 2004.

[8] von Ahn, L. and Dabbish, L. Designing Games With a Purpose. *Communications of the ACM*, 51:58–67, 2008.