Thumbs-Up: A Game for Playing to Rank Search Results

Ali Dasdan
Yahoo! Inc.
Sunnyvale, CA 94089, USA
dasdan@yahoo-inc.com

Chris Drome
Yahoo! Inc.
Sunnyvale, CA 94089, USA
cdrome@yahoo-inc.com

Santanu Kolay
Yahoo! Inc.
Sunnyvale, CA 94089, USA
santanuk@yahoo-inc.com

ABSTRACT

Human computation is an effective way to channel human effort spent playing games to solving computational problems that are easy for humans but difficult for computers to automate. We propose Thumbs-Up, a new game for human computation with the purpose of playing to rank search result. Our experience from users shows that Thumbs-Up is not only fun to play, but produces more relevant rankings than both a major search engine and optimal rank aggregation using the Kemeny rule.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search Process; I.2.6 [**Learning**]: Knowledge Acquisition

General Terms

Design, Experimentation, Human Factors, Measurement

Keywords

Games with a purpose, human computation, online games, rank aggregation, relevance, search engine

1. INTRODUCTION

Almost two-thirds of American households, irrespective of age, background, and gender (40% women) play computer or video games [3]. The insight behind "human computation" is to use "games with a purpose" (GWAPs) to channel this time and energy toward solving computational problems that are easy for humans but significantly hard for computers (i.e., automation) [5]. A very successful example of such games is the ESP game [5], also released as the Google Image Labeler, where players provide accurate labels for images while playing the game.

We propose Thumbs-Up as another example of a GWAP to address the (document) ranking problem for search engines. In Thumbs-Up, players are shown the same input (a query and images of two relevant search results) and must agree which search result is more relevant to the query.

Thumbs-Up is based on two fundamental hypotheses. The first states that given two documents, humans are better at ranking their (perceived) relevance than a computer. This is especially true if ranking involves personalization. Moreover, although computers (and search engines) employ sophisticated machine learning methods to rank search results, humans are still needed. They are directly employed to provide absolute or preference judgments, or are indirectly utilized as users to provide preference judgments through their clicks. The second hypothesis states that preference judgments are easier for humans to make than absolute judgments, which is well supported by some recent work, e.g., see [1] and the references therein.

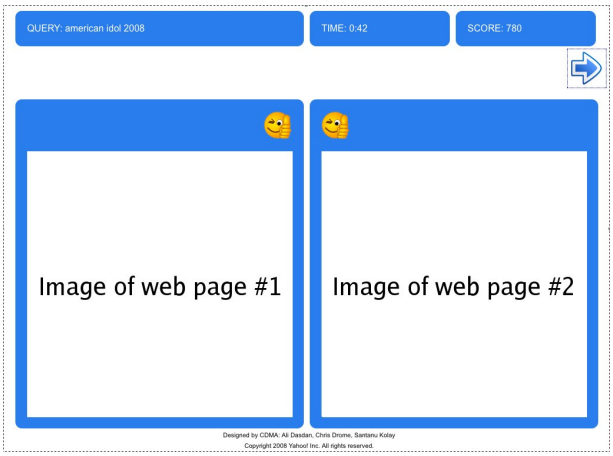

Figure 1: Screenshot of Thumbs-Up.

As for prior work, we are unaware of any application of GWAPs to search engine ranking [5]. In [1] and the references therein, methods are proposed to extract preference judgments from users but without using a game or GWAP framework. The analysis of the collected results is also performed differently.

2. METHODOLOGY

Game rules. A player logs in and is randomly matched with another player. Both players are shown the same input query and images of two web pages deemed relevant to the query. To increase their scores, the players must agree on the same page as more relevant.

Game features. We incorporated several well-known game features to make Thumbs-Up challenging and fun, which include: a time limit (60 s), score keeping (60 per successful match), daily and all time high score lists, and randomness in selecting partners, queries, and images. Features unique to Thumbs-Up include: image magnification on mouse-over and single-click selection.

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

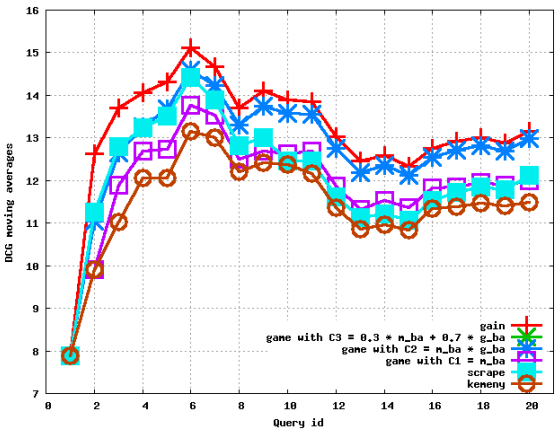

Figure 2: DCG moving averages.

Discounted cumulative gain (DCG). This metric is widely used by major search engines to measure the perceived relevance of ranked search results. For n queries q, each with k ranked results, DCG [4] is defined as

$$DCG = \frac{1}{n}\sum_{q} DCG(q) \text{ and } DCG(q) = \sum_{r=1}^{k} \frac{g(r)}{lg(1+r)} \quad (1)$$

where $g(r)$ is the gain (or judgment) for the document at rank r.

Rank aggregation. During the course of play, multiple rankings are produced which can be merged into a single ranking using rank aggregation methods. Many different methods to determine an aggregate ranking exist [2], where the optimal method depends on the application area. We use the Kemeny rule, which is recognized as one of the best overall [2].

The Kemeny rule minimizes the total number of disagreements between the aggregate ranking and the input rankings. Although determining the Kemeny ranking is NP-hard [2], it can be solved optimally for a small number of elements by using the following integer linear program (ILP):

$$\begin{aligned} &\textbf{minimize } C(x) = \textstyle\sum_{a \neq b} n_{ba} x_{ab} \\ &\textbf{subject to} \\ &x_{ab} + x_{ba} = 1 \ (\forall a, b : a \neq b) \\ &x_{ab} + x_{bc} + x_{ca} \leq 2 \ (\forall a, b, c : a \neq b, b \neq c, c \neq a) \\ &x_{ab} \in \{0, 1\} \ (\forall a, b : a \neq b) \end{aligned} \quad (2)$$

where for two elements a and b, $x_{ab} = 1$ if a is ranked ahead of b in the aggregate ranking and 0 otherwise, and n_{ba} is the number of input rankings that rank b ahead of a.

Game cost functions. We apply the ILP in Eq. 2 to Thumbs-Up results using three different cost functions: (1) $C_1(x) = \sum_{a \neq b} m_{ba} x_{ab}$, (2) $C_2(x) = \sum_{a \neq b} m_{ba} g_{ba} x_{ab}$, and (3) $C_3(x) = \sum_{a \neq b} (\alpha m_{ba} + (1-\alpha) m_{ba}) x_{ab}$ (where $\alpha \leq 0.3$ produced the same DCG as one by C_2). Here g_{ba} is the change in the gain for ranking b ahead of a (from Eq. 1), and m_{ba} is computed as follows.

For the ranking of a and b, let A_{ba} and D_{ba} denote the number of pairwise agreements and pairwise disagreements if b is ranked ahead of a. We then set m_{ba} and m_{ab} to $A_{ba}(1+f)$ and $A_{ab}(1+f)$ where $f = (D_{ab} + D_{ba})/(A_{ab} + A_{ba})$, which helps distribute disagreements to each side of agreements proportional to their magnitudes. If there are no agreements, we set $f = 0.5$, i.e., equal distribution.

3. RESULTS AND DISCUSSION

We selected queries randomly from web search user query logs. The top five search result (URLs) for each query were then scraped from major search engines. Each URL was judged by professional judges, and a digital image of the page was generated.

We implemented Thumbs-Up using a Java/Tomcat front-end with MySQL on the back-end, and released it internally to engineers at Yahoo! Web Search. Within one week we had collected the following statistics: 52 users, 20 queries, 1223 games, 1349 games (including skipped games), and 102 unique URLs viewed. Moreover, about 1/3 of the users played more than 20 games and four of them played more than 100. More than half of the games ended up with a score (agreement) for almost all users. The time spent per game varied from 5 s to 40 s, with an average of slightly less than 15 s. We require 10 games to rank five URLs for a given query, but collected 6 times more games on the average. These statistics seem to indicate that the game was fun and challenging to play.

We ranked the Thumbs-Up results using the following methods: (1) gain: ranking by judgments from professional judges, resulting in an upper bound on DCG; (2) C1, C2, and C3: rankings using the three cost functions C_1, C_2, and C_3, respectively; (3) kemeny: ranking using the Kemeny rule (C in Eq. 2); and (4) scrape: ranking of a major search engine.

Fig. 2 shows the DCG moving averages as each query is added. The DCG values for the last query give the final DCG values per method. From this figure, we see that "gain" performs the best, as expected, and "kemeny" performs the worst, which is surprising as it optimally minimizes the number of disagreements. We also see that our cost functions perform remarkably well: C_1 almost matches the major search engine (within 1%), and the other two are identical and close to "gain" (within 1%), beating the major search engine by up to 8% and the Kemeny rank aggregation by up to 13%. We believe this result is significant considering the magnitude of the investment in search engine ranking technology.

4. CONCLUSIONS

We developed a GWAP called Thumbs-Up for humans to play to rank search results. Our limited experience suggests that Thumbs-Up is not only fun and challenging to play, but also performs surprisingly well.

5. REFERENCES

[1] B. Carterette, P. Bennett, D. M. Chichering, and S. Dumais. Here or there: Preference judgments for relevance. In *Proc. Int. Euro. Conf. on IR (ECIR)*, pp. 16–27. ACM, 2008.

[2] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proc. Int. Conf. on WWW (WWW)*, pp. 613–622. ACM, 2001.

[3] The Entertainment Software Association: Industry facts. http://www.theesa.com/facts/, Dec 2008.

[4] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

[5] L. von Ahn and L. Dabbish. Designing games with a purpose. *Commun. ACM*, 51(8):58–67, 2008.