Picture This: Preferences for Image Search

[Extended Abstract]

Paul N. Bennett
Microsoft Research
One Microsoft Way
Redmond, WA 98052
pauben@microsoft.com

David Maxwell Chickering
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
dmax@microsoft.com

Anton Mityagin
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
mityagin@microsoft.com

ABSTRACT

We demonstrate a system designed to elicit relative relevance judgments from users to rank images with respect to an image query. The system has been deployed and in use publicly for approximately one year. Furthermore, preference data collected from the users has been made available for research purposes.[1] Further details regarding research on this system is available from Bennett *et al.* [1].

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*selection process*

General Terms

Algorithms, Performance

Keywords

preference judgments, learning preferences

1. PROBLEM OVERVIEW

In many information retrieval applications, we desire to rank items in relation to an information need or task. The purpose may be to display directly the ranked items to a user of the system or to use the ranking as an intermediate step in another algorithm.

While the standard approach to ranking has been to rank items according to their relevance [5], and in particular *topical* relevance [2], an item may be preferred by a user based on a variety of other characteristics including quality, authoritativeness, and readability; these characteristics may be seen as defining a broader notion of relevance that incorporates the context of a task. For image search—which is the primary ranking task targeted by this system—a user may prefer an image because of its focus, composition, artistry, or a variety of other dimensions.

Ranking systems are typically constructed with data consisting of explicit relevance judgments for a set of training items. One approach to getting this training data is to elicit relevance judgments from a small number of "expert" judges or editors in a controlled setting. A difficulty of this approach is that, due to query ambiguity and personal preferences, it may be difficult for any single person except the query issuer to accurately judge the relevance of results.

An alternative approach to data collection is to sample judgments from a large population resembling the user base. This has motivated much of the research on mining click data. While click data has undeniable value, most notably because it comes from the issuer of the query, it also has potential weaknesses. In particular, items that have not been displayed cannot be clicked, and the lack of a click is often not informative because the search page itself may satisfy the information need. This latter problem is even more of a concern in image search because the search result page typically consists of actual images (potentially scaled down). Furthermore, if the designers of a ranking system experiment with the live system by (*e.g.*) swapping items or placing potentially non-relevant results in the top of the list [3, 4], there is a risk of frustrating the user and prompting him to switch search engines.

In this work, we attempt to solve the data-acquisition problem in the domain of image search with a social labeling game [6]. The game pairs two participants on the internet who are shown a sequence of queries and corresponding sets of image results; they are both asked to choose the best image for each query, and every time they agree they are awarded credits. After collecting enough credits, they may turn the credits in for prizes. Assuming participants are primarily seeking credits or prizes, the incentive mechanism encourages users to give their actual opinion of the best image because it is a good way to achieve agreement with high likelihood.

Bennett *et al.* [1] studied several probabilistic models that can be used to convert the preference data that results from *Picture This* play into a set of relevance scores for the items. In that work, they further described how the data resulting from game play is similar to click data but without issues of position bias, and is suitable for learning consensus rankings. Furthermore, they examined a number of preference learning models and demonstrated that two of these models

[1] A version of the data suitable for preference learning is available at `http://go.microsoft.com/?linkid=9648573` for use in research experiments.

KDD-HCOMP '09, June 28, 2009, Paris, France
Copyright 2009 ACM 978-1-60558-672-4 ...$5.00.

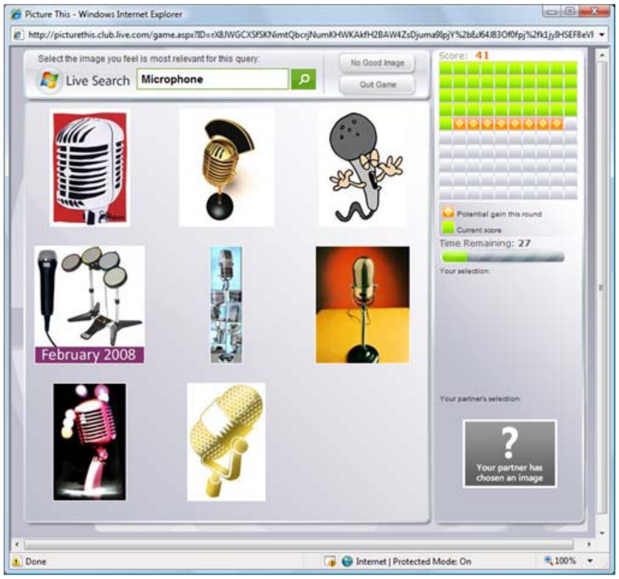
Picture This - Windows Internet Explorer
http://picturethis.club.live.com/game.aspx?ID=rX8JWGCXSfSKNimtQbcrjNumKHWKAkfH2BAW4ZsDjuma9lpjY%2b&J64J83Of0fpj%2fk1jy8HSEF8eVI
Select the image you feel is most relevant for this query:
Live Search
Microphone
No Good Image
Quit Game
Score: 41
February 2008
Potential gain this round
Current score
Time Remaining: 27
Your selection:
Your partner's selection:
?
Your partner has chosen an image
Done
Internet | Protected Mode: On
100%

Figure 1: The *Picture This* game interface. The query shown at the top is "Microphone" with eight candidate images. The green boxes in the grid at the top right show how many of the 100 possible points the player has earned; the orange boxes with diamonds show how many points will be earned on agreement (*i.e.*, k). In this case, the partner has made his selection (indicated in lower right corner) and is waiting for this player to choose.

– one based on logistic regression and the other a conditional online Bayesian model – not only perform well when given a large amount of data, but also learn rapidly when data is scarce.

2. GAME OVERVIEW

Picture This is a collaborative game where pairs of players are rewarded for agreeing on the best result for image-search queries. Currently the game can be accessed online at `http://picturethis.club.live.com`.

When a user starts the *Picture This* game he is first paired with a random partner. If no human player is available within a few seconds of the player starting the game, the player is assigned a robot as a partner. The two players then proceed through a set of rounds, working together for either two minutes or until they reach 100 points, whichever comes first. The players are synchronized such that the next round starts only when both players have completed the current round.

In each round, the two players are shown a query and two different random permutations of the same set of k images. The images are permuted both to eliminate positional bias in the preference data and to mitigate against fraud. k is initially set to three and changes throughout the game. Each player selects the image that, in his opinion, best matches the query. If the two players agree on the best image, they are both awarded k points and k is incremented by one if $k < 9$. If the players disagree on the best image, they are awarded zero points and k is decremented by one if $k > 2$. Thus the number of images displayed varies from 2 to 9 depending on player actions: when users agree, the number of images displayed (and the difficulty of the game) increases, and when users disagree, the number of images displayed decreases. Adapting the game difficulty to a player's performance allows us to both (1) take advantage of discerning players by effectively getting more preference judgments per click and (2) make the game more entertaining.

Players have the option to choose "no good image" to indicate that none of the images are a good result for the query. If either player chooses this option, the number of images k is incremented or decremented as usual depending on agreement, but no points are awarded for agreement. Players can also flag individual images as being "bad". If two players flag the same image as "bad", then they are awarded a time bonus of five seconds if they also agree on the best image. Flag matches are not rewarded if the players agree on "no good image".

In Figure 1 we show what the game interface looks like when a player is choosing an image. Note that the interface indicates in the lower right corner that the partner has already chosen. After the player makes his own selection, his partner's choice will be revealed.

After two minutes has passed or the pair of players has attained 100 points, the game ends. For players signed into the game site, these points are converted into a currency that can be spent on various items including t-shirts, computer hardware, computer software, and music. Players then have three options: they can choose to play again with the same partner, they can choose to play again with a new partner, or they can quit the game. If one player requests to play again with the same partner, that partner is given a choice to either accept or reject the invitation.

More details regarding game structure, query and image selection (which uses a rudimentary form of active learning), the design of partner robots, and the use of incentive structure and other mechanisms to mitigate fraud are discussed further by Bennett *et al.* [1].

3. REFERENCES

[1] P. N. Bennett, D. M. Chickering, and A. Mityagin. Learning consensus opinion: Mining data from a labeling game. In *WWW 2009*, 2009.

[2] C. Cleverdon. The significance of the cranfield tests on index languages. In *SIGIR '91*, pages 3–12, 1991.

[3] F. Radlinski and T. Joachims. Minimally invasive randomization for collecting unbiased preferences from clickthrough logs. In *AAAI 2005*, 2005.

[4] F. Radlinski, R. Kleinberg, and T. Joachims. Learning diverse rankings with multi-armed bandits. In *ICML 2008*, 2008.

[5] S. E. Robertson. The probability ranking principle in ir. *Journal of Documentation*, 33:294–304, 1977. Reprinted in: K. Sparck Jones and P. Willett (eds), Readings in Information Retrieval. Morgan Kaufmann, 1997. (pp 281-286).

[6] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *ACM CHI*, 2004.