Classifying Images with Image and Text Search Clickthrough Data

Gavin Smith, Michael Antunovic, and Helen Ashman

WebTech and Security Lab, University of South Australia
{gavin.smith,helen.ashman}@unisa.edu.au

Abstract. Clickthrough data from search logs has been postulated as a form of relevance feedback, which can potentially be used for content classification. However there are doubts about the reliability of clickthrough data for this or other purposes. The experiment described in this paper gives further insights into the accuracy of clickthrough data as content judgement indicators for both HTML pages and images. Transitive clickthrough data based classification of images contained in HTML pages has been found to be inferior to direct classification of images via image search clickthrough data. This experiment aimed to determine to what extent this is due to the inferior accuracy of clickthrough-based classification accuracy in HTML. The better classifications resulting from clickthroughs on image searches is confirmed.

1 Introduction

1.1 Background

Clickthrough data is increasingly seen as a potential indicator of relevance feedback on search results, not just for search ranking purposes but also for other applications such as classification of content.

Clickthrough data is generated when a user selects (clicks on) results from a page returned from a search. The implication is that the user will primarily click on results of most relevance to their information requirement as expressed in the query term submitted to the search engine. Thus clickthrough is a form of relevance feedback that rates the relevance of the tendered results to the query.

While simple in principle there are a number of known issues (e.g. trust bias, see section 2) with the reliance on clickthrough data as implicit judgment. Such issues and the severity of their impact then vary depending on the type of search (document vs. image) providing the clickthrough data. They introduce noise, and the nature of click-through logs [13] introduces low coverage. However, the existence of such cheap, large and continuously generated logs inspires attempts to resolve these issues of noise and increase coverage. This paper evaluates the potential of web search clickthrough data to contribute to image classification via a transitive labeling method [1]. Such a method seeks to alleviate the sparsity problem by using an alternative resource

J. Liu et al. (Eds.): AMT 2009, LNCS 5820, pp. 385–396, 2009.

to the click-through data from the resource's own search (as in [4]) and one for which a potentially larger source exists[1].

1.2 Motivation – Transitive Classification of Images

It seems that reliable clickthrough data should be a primary source of relevance feedback on searches especially in terms of ranking of image results, as has been proposed for web document search results [3] [6]. So far this has been seldom reported in the open literature and this may be because of the difficulty in procuring web log data that includes image searches. However, clickthrough data can also be used to extend the applicability of established labels and classifications, such as in the *transitive* method for image labelling which inherits a classification or label for an HTML or text page onto potentially every image contained in that document [1]. This method additionally requires filtering out "non-content" images which were identified by inspection as being frequently generated by advertisements, formatting images (such as bullet point icons and lines), and banners.

One purpose of the experiment described in this paper is to determine whether this transitive labelling method can accurately supplement existing image classification technologies. We experimentally validated the image labels generated by this transitive method (along with the direct method and 4 other widely-used methods) with ground-truthing by over 100 human assessors. We found that these transitive labels were inferior to both the Google Image Search and the derivative direct labelling method (which consists of applying a search term as a label to an image if it has 2 or more clickthroughs from an image search on that term which confirm the image's relevance to the term) [1]. However it was not immediately clear whether the errors in the transitive method arose from a poor filtering algorithm or whether it was an artifact of "garbage in, garbage out", that is, whether the image classifications were poor because they inherited poor labels from their containing web pages. Given the doubt about text search clickthrough validity (see section 2.1), the page labels themselves could have created a significant proportion of the errors which were inherited onto the images they contained.

This paper thus reports on an experiment that establishes whether a set of Web pages, containing images which were labelled using the transitive method, was accurately labelled. This first result demonstrates that the accuracy of images labelled with clickthroughs using the direct labelling method is significantly higher than the accuracy of HTML pages labelled in the same way. This indicates how much "garbage" goes into the transitive labelling algorithm. We then go on to correlate the accuracy of transitively-generated image labels with the accuracy of HTML page labels, finding that the filtering algorithms that purportedly screened out "non-content" images and left in "content" images were at least partly responsible for the inaccuracies of the transitive labelling method.

We next go on to section 2, which considers related work by others evaluating the use of clickthrough data. Section 3 then describes the experiment that ground-truths

[1] In the three years of collected Web logs from the University of Teesside used in this experiment, we found that image searches amounted to only around 5% of the total amount of web image and text searches.

the accuracy of HTML search clickthrough data, while section 4 details the results of the experiment. Section 5 discusses the results, and section 6 concludes.

2 Related Work

In this section we look at other assessments of the relevance of clickthrough data. We find that most such assessments to date focus on text searches.

2.1 Clickthroughs to Web Pages

Clickthrough data from traditional web page search has been the subject of much recent work. Proposed for a wide range of uses, in 2005 the usefulness of this data began to be questioned. From the research it became clear that it was not correct to rely on the assumption that clickthrough data could be directly used as an absolute judgment of relevance [5] [8]. Others [4] [7] found that users of the search systems were biased in a number of ways when it came to clicking results, causing them to click on results that may not be the most relevant. Specifically quality-of-context, trust and position biases were identified. Despite this drawback, the prolific and cheap nature (e.g. compared to explicit human labeling) of such data has seen its continued use, with research looking at ways of re-weighing judgments to counter the identified biases, most recently [3] [6]. As such the notion that clickthrough data can provide relative relevance is more commonly accepted.

In particular it is proposed that clickthrough data from image searches is more accurate that that of text searches, as discussed in the next section.

2.2 Clickthroughs to Images

In contrast to clickthrough data from traditional web page search, clickthrough data from web image search has seen little evaluation. While a recent study [9] indicated the higher accuracy of image search clickthrough than web search clickthrough data, the exact presence and impact of the biases identified in web search clickthrough data is not clear. It is also notable that proposed research for re-weighting clicks cannot be directly applied due to the different presentation of search results - image search results are often presented in a grid with no obvious "best" or "first" result.

However, when looking to create highly accurate labels of images (as opposed to re-ranking) the presence of various forms of bias is of less concern. Such a goal has applications to techniques such as query and image clustering [2] [10] and image concept learning [11]. In prior work, [1] the ability to accurately label images by a number of different methods was evaluated, including two based on different applications of clickthrough data. In the first of these two methods, the *direct* method, image search clickthrough data was used to effectively filter Google image search results. Using this method, high levels of accuracy were reported, however, the method returned a low number of images due to the sparsity problem inherent in clickthrough data [13]. In contrast a *transitive* method, based on mining images from clicked web pages from a web search clickthrough log and associating the clicked page's query with the image, was evaluated. Such an approach suffered less from the sparsity problem as significantly more text searches are performed than image searches (only 5%

as noted earlier) and hence more pages are clicked than images. However, the accuracy of the classifications generated by the transitive was somewhat less, with completely relevant image/label pairs from the transitive method occurring around 63% of the time, compared to the Google image search's 80%. This paper extends this evaluation work, identifying the cause of this loss in annotation accuracy in order to develop larger sets of images for a given concept (search keyword(s)).

3 Experiment to Ground-Truth Clickthrough Accuracy

3.1 Experiment Description and Motivation

The main comparison to make is between the precision of images classifications/labels created with clickthroughs and the precision of their containing websites, also labelled with clickthroughs. Having established in the earlier experiment [1] that the accuracy of image/label pairs generated with the direct method (around 84%) is greater than the image/label pairs generated by the transitive method (around 63%), it is necessary to identify the cause of this discrepancy.

There are two potential causes of the discrepancy:

i) The labels belonging to the Web pages themselves were inaccurate. This seems feasible given the doubts in the literature about the reliability of clickthrough data for classification purposes (as discussed in section 2);

ii) The filtering mechanisms employed to remove "non-content" images, such as advertisements and formatting, were too coarse. This also seems feasible as the filtering algorithms were derived prior to any analysis of the relevance of contained images, and were largely speculative. The assumption is that images are by default relevant to their containing pages, unless they are "non-content" as defined here.

It may be that both of these causes affected the accuracy of the transitive method, so the experiment set out to ground-truth the accuracy of clickthroughs on the selected set of Web pages from which the images were extracted and labelled.

In summary, the experiment sets out to measure the accuracy of clickthrough as content classification judgement on the set of Web pages from which images were extracted and likewise classified, and to measure the impact of this accuracy.

3.2 Experimental Set Up

There have been two experiments set up to ground-truth the accuracy of labels applied as a result of clickthrough data. In the earlier image label evaluation experiment [1], six different image labelling methods were evaluated for their precision, these being the direct labelling method (a derivate of Google's Image Search as described above), the transitive image labelling method (also described above with implemented filters summarized in Table 1), the human-provided image labels from the Google Image Labeller game[2] (based on the ESP game [12]), the flickr image hosting site[3] and Getty

[2] http://images.google.com/imagelabeler/

[3] http://www.flickr.com/

Images[4], plus the Google Image Search facility[5]. Using a mostly canonical set of query terms, a collection of image/query term pairs was generated, selecting the top 11 or 12 for each query as tendered to Google Image Search, flickr, Getty Images, the direct method and the transitive method, along with a random selection of around 1000 Google Image Labeller images, screen-scraped from the game interface.

These six methods were evaluated by comparing the implied label or classification of each over a total of 4693 images, and we found that the direct labelling method was 4% more accurate than Google Image Search from which it was derived, while the transitive method was incrementally more accurate than the remaining methods, although significantly behind the direct method and Google Image Search [1].

The poor showing of the transitive method needed further analysis, as it was not clear whether it was due to inaccuracy in the filtering process, or due to the Web pages being poorly-labelled themselves. To analyse this we set up the second experiment to firstly assess the validity of the web pages from which the transitive images were drawn, and secondly to compare the accuracy of the labels of the images versus their containing web pages. This second experiment uses the same approach as the original image-assessing experiment, and combining results from the two allows both an assessment of accuracy of web page labelling based on clickthrough, and a measurement of the liability of poor page labelling in the image labels.

Some variables were fixed as far as possible:

- fixed: we used the original clickthrough data from the same dataset for both images and websites;
- fixed: we used the same set of 71 queries for both the image search clickthroughs (from which the direct method created image labels) and for the text-based search (from which the web page labels were derived using the direct method and subsequently the image labels were derived using the transitive method);
- semi-fixed: ground truthing of both images and webpages was done by many of same people. 12 people have ground-truthed all of the websites compared to 10 have ground-truthed all of the images, and 7 of these did both complete sets.

Table 1. The transitive filtering functions as implement in [1]

Filter	Reject condition	Filter	Reject condition
repeated	Img repeated in page	tooSmall	Img height or width < 50 pixels
aspectRatio	Img aspect ratio > 3:1	logos	Img path contains text 'logo'
		advertisement	Img in blacklist 'Rick752's EasyList'[6]

The clickthrough data was extracted from around three years worth of anonymised Web logs provided by the University of Teesside School of Computer Science.

The original image/label ranking experiment showed each image in a frame with the associated label above it, below a set of 4 options for the evaluator to select:

[4] http://www.gettyimages.com/

[5] http://images.google.com/

[6] http://easylist.adblockplus.org/

0 - image did not load or user did not recognise it;
1 - image was not relevant to the label;
2 - image was partially relevant to the label, e.g. a car tyre, labelled 'car'
3 - image was completely relevant to the label.

Each evaluator was presented all images from all methods in a random order, preventing identification of the exact method used for an individual image.

The web page ranking experiment was then set up allowing evaluators to evaluate the relevance of the source website against the same search terms.

The second experiment was done similarly to the original image rating web application, but where the website was loaded within the page itself via an inline frame, and the search term used to find images from that website placed above the frame. The same Likert rating scale was used, with the following interpretations of the ratings made:

0 - Website did not load, or the user did not know what the website was about;
1 - Website not relevant to the search terms used to find images on the site;
2 - Website partially relevant to the search terms, i.e. some content but not all;
3 - Website completely relevant to the search terms used to find images on the site.

This latter experiment relevance-ranked the 184 sites from which the transitively-labelled images were drawn, and hence the labels being evaluated were the same as those being evaluated in the image experiment.

4 Experimental Results

In this section, the data collected is described, then results are given on firstly the relevance of the web pages to their search term from which the transitive images were drawn, and secondly on the accuracy of the labels of the images versus their containing web pages.

4.1 The Data Generated

The data generated by the two experiments now yields a set of relevance ranking data for each of the three sets, the clickthrough-classified web pages, the clickthrough-classified images and the transitively-classified images, as follows:

- a set of images and a set of web pages, both of which were classified using exactly the same clickthrough method (the direct labelling method). These objects were chosen from the same main set of web log data, so as to have coincident classification labels. It is not certain but its is believed likely that the participants generating the original clickthroughs overlapped, since the clickthrough data is from a relatively small population of staff and students in a university school, with strong association of search activity with undergraduate assignments and similar research tasks.
- A third set of objects (the transitively-labelled images) has classification labels derived from its parent objects (the set of web pages).

There are two subcategories of this third set, the first subcategory being the "non-content" images which were those that the filtering process filtered out as being not relevant (due to a wrong aspect ratio, being too small or from a known advertisement source), while the second subcategory were the "content" images, which were those not filtered out. Importantly, both the content and non-content images were relevance ranked by the human assessors, as this gave information about both the false positives of the filtering process (filtered out but should not have been, i.e. those identified as non-content but which were in fact genuinely relevant to their labels) and the false negatives (those not filtered out but which should have been).

4.2 Relevance of Web Sites to Search Terms (Clickthrough Accuracy)

A sample of the data showing the raw data generated from the website relevance ranking experiment is given in table 1. The *precision* is the number of people ranking the Website as 3 (fully relevant) divided by the total number of non-zero rankings (zero rankings are excluded as the evaluator was unable to make a judgement on their relevance). The *partial precision* was the number of rankings of either 3 or 2, divided by the total number of non-zero rankings.

Table 2. Sample data from Website ground-truthing experiment

Site	Precision	Partial Precision	3s	2s	1s	0s	Average Ranking	TotalNo. Rankings
Site 1	0.64	0.73	7	1	3	1	2.36	12
Site 2	0.92	1.00	11	1	0	0	2.92	12
Site 4	0.54	1.00	7	6	0	0	2.54	13
Site 5	1.00	1.00	12	0	0	0	3.00	12
Site 6	1.00	1.00	12	0	0	0	3.00	12
Site 7	0.58	0.92	7	4	1	0	2.50	12
Site 8	1.00	1.00	13	0	0	1	3.00	14
Site 9	0.92	1.00	11	1	0	0	2.92	12
Site 10	1.00	1.00	12	0	0	0	3.00	12
Site 11	1.00	1.00	12	0	0	0	3.00	12

We furthermore found that there were 25 sites that were unanimously ranked 3 by all users, giving them a 100% precision. Another 97 sites had all rankings 2 or 3, i.e. a partial precision of 100%. There were no sites with unanimous 2 or 1 rankings. 57 of the sites (including the 25 unanimously-ranked ones) showed a very consistent ranking with a relative standard deviation (RSD) across all evaluators of not more than 10%. A further 77, making up to 134 sites, showed an RSD of less that 25% - examples include a number of sites with 8 3-rankings and 4 2-rankings whose precision was 2.67, well within the "relevant" range. A small number, 5 sites, showed highly variable rankings where the RSD was over 50%, where presumably the evaluators found the label contentious.

We grouped sites according to their relevance to the search term associated with the clickthrough as follows:

i) relevant to the search term: any site with an average ranking of 2.5 or above, i.e. having at least half of its rankings at 3;

ii) peripherally relevant to the search term: any site with an average ranking of 1.8 to under 2.5;

iii) not relevant to the search term: any site with an average ranking below 1.8.

With this grouping, we find that the following results:

Table 3. Table of site relevance to search term

Relevance	**number of sites**	**proportion of sites**
relevant	133	76%
peripherally relevant	33	18%
not relevant	11	6%

These figures suggest that the relevance of clicked sites to the search term is not as low as indicated by other studies (see section 2). This could be due to the relatively small number of sites evaluated (184), or that the sites selected were not necessarily typical of all searches in a more general context or population[7]. Alternatively it may be due to the distinct type of relevance assessment. Some prior studies assessed user satisfaction with results returned for a given search term [5] [8] where there is implicitly a meaning assigned by the user to the search term in these circumstances. In contrast, the evaluation here is not of the relevance of the website to a given meaning, but rather its relevance to *any* meaning of the search term, i.e. regardless of any ambiguity in the search term, and any meaning the searcher may have had in mind.

4.3 The Match between HTML Page Labels and Contained Image Labels

Having established that the accuracy of text search clickthrough data is, at least in this context, performing better than the image relevance where classifications were transitively-generated, we now consider the accuracy of each image compared with the page it occurred in.

The following table shows the number of each of the possible pairings of image relevance with containing website relevance. Each such pairing is represented as a pair of numbers, with for example (3, 2) representing the case where the site was ranked 3 and the image was ranked 2. The images are also separated out into their distinct filtered type, with Content referring to images that the filtering algorithm thought were relevant to the page content, Too Small being images under 50x50 pixels, Wrong AR (aspect ratio) being an image too narrow, Logo meaning an image for some company and Advertising being an image from a known advertisement-supply site. Note that all of the transitively-labelled images were ranked, including those that were categorised as non-content. The purpose of this was to assess the impact of both false positives (images wrongly judged to be content, i.e. not filtered out but should

[7] On inspection, it seemed apparent that a large proportion of searches, especially multiple searches on the same term, from these weblogs were performed by students seeking information for preparing reports and assignments.

have been) and false negatives (images wrongly judged to be non-content) from the filtering process.

The transitive method had been applied to generate 445 unique image/label pairs. Filtering algorithms were then applied to exclude those images that were superficially deemed to be "non-content", resulting in 185 of the 445 images being categorised as not being content relevant to the web page. Of the non-content, 150 were categorised as "too small", 13 as "wrong aspect ratio", 5 as "logo" and 17 as "advertisement".

Table 4. Paired relevance rankings of transitively-labelled images and sites, separated according to image filtering

(site, image)	Content	Too Small	Wrong AR	Logo	Advertising
3, 3	71.02%	8.05%	6.02%	0.00%	2.27%
3, 2	12.22%	3.82%	7.09%	0.00%	4.55%
3, 1	13.00%	80.00%	64.17%	100.00%	48.18%
2, 3	2.00%	0.24%	0.00%	0.00%	0.00%
2, 2	0.38%	0.38%	0.40%	0.00%	0.00%
2, 1	0.99%	5.28%	13.24%	0.00%	15.00%
1, 3	0.01%	0.02%	0.00%	0.00%	0.00%
1, 2	0.17%	0.06%	0.27%	0.00%	0.00%
1, 1	0.20%	2.41%	8.82%	0.00%	30.00%

Table 2 makes it clear that the filtering process leaves much to be desired. The filtering rules were derived by inspection only, and until this experiment had not been evaluated for their accuracy. Also, it was necessary to understand the extent to which the underlying web pages had been accurately classified so that the influence of the filtering algorithms could be understood and this factor considered separately from the filtering algorithm.

We consider now primarily the correlation between relevant sites (ranked 3) and images labelled or excluded from them. In the non-relevant sites (ranked 1), there were almost no relevant sites, as would be expected, as since the site is not relevant to the label, it is unlikely that any contained images would be, except by chance.

Having evaluated the outputs of the filtering, we can make the following interpretations of the results:

– In the Content category, the majority (71.02%) of site and image rankings combined were positive results for both (3,3). However from the relevant sites, there was a further 12% of images only partially relevant, and another 13% of images not at all relevant, and thus should have been excluded but were not blocked by the filtering algorithm. *In total, there was a false-positive rate (image ranked 1 when site was ranked 3 or 2) of almost 14%.*

– In the Too Small category, the majority of images (80%) that were ranked as being not relevant were deemed to come from an appropriate site. This supports the notion that many of the images were graphical artifacts such as bullet points, icons, avatars etc. despite their placement on a relevant site. There are however almost 12% of Too Small images filtered out but which were ground-truthed as being relevant (8.05%) or partially-relevant (3.82%). The filtering algorithm will need to be refined to better detect the potentially relevant but small images.

– In the Wrong Aspect Ratio category, most of the images that were excluded due to a wrong aspect ratio were ranked as not relevant by the evaluators. There is however a small but significant proportion of false negatives, excluded by the filtering algorithm but rated by the evaluators as being relevant (6%) or partially relevant (7%). Any refinement of the filtering algorithm will need to correct this.

– In the Logo category, all of the excluded images were both ranked as being not relevant to the label (100%) while having come from an appropriate site. Note however that in some cases, a logo might be considered relevant to the site, such as if it was the site belonging to the logo holder, however such sites were not among the pages and images assessed in this experiment.

– In the Advertisement category, very few of the images rejected because of their advertising provenance were either relevant (2.27%) or partially-relevant (4.55%) to the search term. Over 93% were not relevant, with over 63% correctly rejected by the filtering algorithm. There is some scope to correct for the wrongly excluded images so more sophisticated rules, such as correlating the image provenance with the content of both the image and site, could refine these results.

There is scope for improving the classification accuracy with better filtering algorithms. These modified filtering algorithms would be easy to assess as we would merely need to assess the new filtered outputs against the already ground-truthed images and websites. Possible improvements for refinement of the filtering algorithm by reducing false positives and negatives include:

- Aspect ratio filtered images were based on dividing the width by height by length and comparing this result to a predetermined ratio, which initially was 3:1. This ratio can be refined incrementally, for example, 3.5:1, working in 0.5 increments to monitor whether the amount of filtered images increases or decreases based on the adjustments and eventually fixing on the ratio with the lowest error rate;
- Logos can be filtered differently, including potential for automated text analysis in the image. The current filtering mechanisms simply look for the text *logo* in the filename. No consideration is made for language specifics, alternatives of the word (eg. banner) or potential for character matching in larger words that may affect accuracy (eg. logoff, logorrhea);
- Too small image sizes can be adjusted to be smaller, resulting in a decrease in the volume of images filtered out, potentially reducing the number of false negatives;
- Advertisements are a difficult category, as many of the sites were relevant to advertisement images due to site-specific tailoring strategies. Up-to-date "adblock" lists may improve this area of the filtering mechanism.

5 Discussion

So how much of the inaccuracy of the transitive method was due to the poor labelling of the original Web pages and how much was due to the filtering algorithms?

In section 4.2, the relevance of the websites to the search term (i.e. the accuracy of the clickthrough data) was measured, and it was found that 76% of the assessed sites were fully relevant, while a further 18% were peripherally relevant, and hence that

94% of the sites evaluated were are least partially relevant to the search term. However this was not reflected in the transitively-labelled images. The precision of the transitive method was much lower, with only 63% of the transitively-labelled images being relevant, or around 80% of images being partially-relevant. This suggests that the problem lay mainly with the filtering algorithm since in both cases, relevant and partially-relevant, the site relevance rankings were well above the image relevance rankings.

When we look at the correlated image/site rankings in table 3, we find both false positives (images that should have been excluded) and false negatives (images that were wrongly excluded). The false positives have a real impact on the precision of the transitive method, wrongly labelling images with the site's label and associating irrelevant images with the label. Less pressing in terms of the precision of the transitive method but still worth further investigation is the level of false negatives, where images are excluded but are still relevant.

From the point of view of the accuracy of the transitively-generated image labels, there is a false positive rate of almost 14%, with the vast majority of these (13% of the total, or nearly 93% of the false positives) being contained in relevant (ranked 3) sites. 14% of the images labelled by the transitive method are not relevant to the label but are not excluded by the filtering algorithm. This false positive rate accounts for the majority of the discrepancy between the precision of the transitive method (over 63%) and the precision of the direct method (over 83%).

However there is a smaller but still sizeable proportion that is not due to the false positives of the filtering algorithm, but can be attributed to the inaccurate labelling of sites containing images, where those sites were labelled themselves based on clickthrough data. If over 30% of the sites are only partly or not at all relevant to the label, this will propagate to a similar proportion of mislabelled images. In table 2, there were 18% of sites partially-relevant to the label, and 6% not relevant at all. Regardless of the relevance of the site to the image, or the accuracy of the filtering algorithms, there will be a proportion of mislabelled images arising from the inaccurate site labelling, the number depending on how many images the site contains.

Dealing with the wrongly-labelled sites depends more on managing the accuracy of clickthrough data. It might be argued that insisting on numerous clickthroughs before labelling a site would assist here, although we found in the earlier experiment [1] that there was very little improvement in click relevance when the click threshold was raised. However this observation was based on image clickthrough data and should be investigated within text searches to confirm.

6 In Conclusion

In this paper we have considered how the relevance of clickthrough data to text searches differs from that of image searches, and whether this affects the accuracy of the transitive classification method. It appears that the main hurdle for the transitive method lies not in any problem with the accuracy of clickthroughs on Web pages, but rather on the filtering mechanisms that exclude images that are not pertinent to the web page itself. There is undoubtedly a level of inaccuracy in the clickthrough data that classifies the underlying Web page but the filtering algorithms present the greatest potential for improvement at this stage.

Acknowledgements

We would like to extend our thanks to everyone who helped with programming, evaluating or anything else, most especially Mark Truran for the data that underlies this work, and to Christoph Donner, Rebecca Frith, Eric Rebelos and Jan-Felix Schmakeit for their programming work.

References

1. Ashman, H., Antunovic, M., Donner, C., Frith, R., Rebelos, E., Schmakeit, J.-F., Smith, G., Truran, M.: Are clickthroughs useful for image labelling? In: Proc Web Intelligence (September 2009)
2. Beeferman, D., Berger, A.: Agglomerative clustering of a search engine query log. In: Proceedings of ACM SIGKDD, Boston, Massachusetts, US, pp. 407–416 (2000)
3. Chapelle, O., Zhang, Y.: A Dynamic Bayesian Network Click Model for Web Search Ranking. In: Proceedings of ACM WWW, pp. 1–10 (2009)
4. Craswell, N., Zoeter, O., Taylor, M., Ramsey, B.: An Experimental Comparison of Click Position-Bias Models. In: Proceedings of ACM WSDM (2008)
5. Fox, S., Karnawat, K., Mydland, M., Dumais, S., White, T.: Evaluating implicit measures to improve web search. ACM TOIS 23, 147–168 (2005)
6. Guo, F., Liu, C., Kannan, A., Minka, T., Taylor, M., Wang, Y., Faloutsos, C.: Click Chain Model in Web Search. In: Proceedings of ACM WWW, pp. 11–20 (2009)
7. Joachims, T., Granka, L., Pan, B., Hembrooke, H., Gay, G.: Accurately interpreting clickthrough data as implicit feedback. In: Proc. SIGIR, ACM, Brazil, pp. 154–161. ACM, New York (2005)
8. Scholer, F., Shokouhi, M., Billerbeck, B., Turpin, A.: Using clicks as implicit judgments: Expectations versus observations. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 28–39. Springer, Heidelberg (2008)
9. Smith, G., Ashman, H.: Evaluating implicit judgements from Web search interactions. In: Proceedings of the 1st Web Science Conference, Athens (2009)
10. Truran, M., Goulding, J., Ashman, H.L.: Co-active Intelligence for Information Retrieval. In: Proc. Multimedia 2005, pp. 547–550. ACM, New York (2005) `http://doi.acm.org/10.1145/1101149.1101273`
11. Tsikrika, T., Diou, C., de Vries, A.P., Delopoulos, A.: Image annotation using clickthrough data. In: Proceedings of the 8th ACM International Conference on Image and Video Retrieval, July, Santorini, Greece, July 8-10 (2009) (to appear)
12. von Ahn, L., Dabbish, L.: Labelling Images with a Computer Game. In: Proc. SIGCHI CHI 2004, vol. 1, pp. 319–326. ACM, New York (2004)
13. Xue, G., Zeng, H., Chen, Z., Yu, Y., Ma, W., Xi, W., Fan, W.: Optimizing web search using web clickthrough data. In: Proceedings of ACM ICIKM, pp. 118–126 (2004)