Are clickthroughs useful for image labelling?

Helen Ashman, Michael Antunovic, Christoph Donner, Rebecca Frith, Eric Rebelos, Jan-Felix Schmakeit, Gavin Smith
School of Computer and Information Science
University of South Australia
Adelaide, Australia
e-mail: helen.ashman@unisa.edu.au

Mark Truran
School of Computer Science
University of Teesside
U.K.
e-mail: m.a.truran@tees.ac.uk

***Abstract*—In this paper we look at how images can be labelled as a result of clickthroughs from searches. One approach acts as a filter on image searches specifically, while the other approach propagates labels to images from their containing pages, where those pages were labelled themselves using clickthrough as a filter on text search.**

Then the paper reports on an experiment where users ranked for relevance six methods for labelling images, comparing the two clickthrough-based methods with flickr's amateur explicit labelling, Getty's professional explicit labelling, Google's standard image search, and the new Google Image Labeller. As well as comparing the accuracy of the proposed image labelling methods and discovering that automatic methods outperform explicit human labelling methods, the experiment suggests clickthrough data is reliable with very few clicks for image classification purposes.

I. Introduction

A. *User judgement in content classification*

The value of the judgement of users is increasingly recognised in Web applications for the organisation and categorisation of Web artifacts. We discuss two approaches for classifying and tagging images based on the collective wisdom of users, derived from their interactions with search engines, i.e. clickthrough data.

The work extends image labeling work done previously which recorded clickthrough data from users interacting with search engines [19]. The experiment reported there indicated that images could be aggregated using the selections made by users from search engine results. This earlier experiment was hampered by a lack of data, and the difficulty of ever generating enough clickthrough data to create conclusive and comprehensive topic coverage. Subsequent work has now focused on extracting the same data from Web logs, and we have assimilated well over two years of Web log data from a UK university for this purpose. Result selections and search terms from Web logs are available in 'virtually unlimited quantity' [11], with around 600 million searches performed around the world every day [6].

The reliability of users' judgement on the relevance of search engine results has been disputed, with Joachims et al. [9] suggesting that clickthrough data was still a 'reasonably accurate' method of obtaining implicit relevance feedback, and the purported correlation between search terms and selections [6]. However, clickthrough data raises many doubts about its reliability for absolute document relevance. Joachims et al. [9] conclude that substantial trust and quality bias make reliance on clickthrough data as absolute judgment 'problematic'. Fox et al. [5] report a 39% satisfaction rate for the documents returned after following search result links and Scholer et al. [15] report a user satisfaction rate of only 52% with the results they view directly after clicking a search result, and only a 58% correspondence between clicked documents and those manually labeled as relevant. Finally Shokouhi et al. [16] completely reject using click-though data to reorder documents.

In contrast, we reported elsewhere that clickthrough data, as applied to images only, is significantly more reliable as an indicator of subject than clickthroughs on normal, text-based Web search [18]. Since image queries return results that essentially encapsulate the entire object, as opposed to a small excerpt of text, the clickthrough relevance of image searches was expected to be higher than found by others assessing clickthrough data on text-based searches and this was borne out by the experiment reported there. It also justifies the application of image search clickthrough for image classification and labelling purposes.

One outcome of this paper is to validate the use of clickthrough data derived from image search as implicit judgment with a set of user rankings on images retrieved in this way. This validation is supplemented by comparing the accuracy of clickthrough data for image labelling with other image labelling methods, including those explicitly authored by users (such as on flickr, Google Image Labeller and gettyimages) and automatic, surrounding-text labelling, characterised by Google's traditional image search. In particular we found that directly applying clickthrough data as a relevance filter on Google Image Search improved the results relevance by around 4%.

Clickthrough data, applied either directly to images or to text pages (and inherited by images on those pages), is more accurate than any of the explicit, human-authored methods for labelling images.

B. Image Labelling

Image analysis and classification has historically been done with two approaches: content-based image retrieval (CBIR) algorithms analysing the content of the image in order to detect knowable features; and analysis of surrounding content, as done by Google and other search engines, extracting keywords to label images for subsequent search results. A third major approach, starting to appear in the literature now is the consensual approach to classification, not just of images but of other resources too, generally text. The consensual approach is essentially to let the users do the classification and to record statistically-significant judgements for other users to make use of, say for search purposes. It is manifested in increasing numbers of applications and sites such as flickr and del.icio.us, and recently was worked into Google's image classification schemes, as discussed below.

This paper discusses and evaluates two complementary consensual ways to label images. Direct image labelling is the use of clickthrough data derived from image searches to directly associate search terms with images, search terms that can later be used as labels for images. In contrast, transitive image labelling inherits the labels of HTML pages (previously classified in the same way using clickthrough data on text-based pages) onto their contained images – the images are deemed to be "about" the same thing as their containing pages.

The anticipated outcome was that direct labelling would produce better results, while transitive labelling would produce more results. The evaluation has supported this. The low proportion of image searches (5% of that of standard text searches in our data) means that direct labelling is a less productive method. However transitive labelling cannot be applied indiscriminately, and preliminary work suggested that certain major categories of image should be excluded. These categories, including advertisements, logos and layout images, were not specific to the content and thus ought not be labeled with the labels pertinent to the content.

In section 2, related work on image labelling, primarily consensual methods and non-CBIR methods, is discussed. The direct and transitive image labelling methods are described in section 3. In section 4, we describe the setup of the experiment whose dual purpose is to firstly validate the use of clickthrough data for image labelling and secondly to measure the precision of the direct and transitive image labelling methods versus established alternative methods. Section 5 presents the results of the experiment while section 6 interprets them. Section 7 concludes with remarks about ongoing work.

II. Related Work

A. Image Classification

Content-based image retrieval (CBIR) leverages many different characteristics in order to determine an image's meaning, characteristics such as colour [1], shape [2] and texture [12]. These features are often compared to those features previously detected in a training set of images. A match with the established characteristics of a training image will see the new image labeled accordingly. The key downfall of CBIR is the amount of time required to train the system. CBIR generally does not use human intelligence, although the system may utilise a training set of images labeled with user-developed tags.

B. Object labelling in consensual systems

In the last few years the tagging and labelling of web resources has become an online norm. Examples of web sites encouraging this trend are the image hosting service at flickr and the social bookmarking site known as del.icio.us. The precise mechanism of this labeling behaviour varies (e.g. flickr allows users to tag uploaded images, while del.icio.us allows users to tag bookmarks) but the underlying rationale is fairly constant - tags are provided by users as a side effect of how they organise their personal space and as a form of community service.

Users are not always inclined to explicitly participate in this way. One solution is to make the activity enjoyable, as in the ESP game [20]. Players were asked to tag images. When there was agreement between users, it was added to the tag set. However, a game is unlikely to generate significant quantities of data, nor is making every single task enjoyable feasible, so gathering tag data in an implicit fashion on the back of other activities is the more viable option. Also, the usage of games and other specific feedback-eliciting techniques may bias the participation of the population contributing to the feedback, who may self-select based on their interest in games for example. The more widespread the usage of a given feedback mechanism, the more likely the results are to be representative of the general perception of the meaning or relevance of the results. As search engines are very widely used, they plausibly offer a more representative snapshot of meaning and relevance than any of the voluntary meaning/feedback collecting tools.

One of the authors of [20] has since addressed this issue in a specific application, using human consensus not for labelling, but for digitising printed texts [14]. CAPTCHAs are images of distorted text which are hard to read with optical character recognition (OCR) software, thus requiring humans to read. They are used to deter automatic registrations and to combat spam.

Robu et al. [9] have discussed how social tagging can give rise to categorisation and classification schemes. They used the cosine measure to determine co-occurrence between tags, an approach which went some way towards solving the problem of ambiguity in search results (see also [19]). Both the tags and the co-occurrences in social tagging sites are explicitly added by users on a voluntary basis, while Web log-derived resource labels and co-occurrences are implicit, a byproduct of widespread search-and-select activity.

All of the above, and the methods in this paper are consensual methods that derive labels from user judgement and activity. The major difference between these and traditional classification methods is that consensual methods do not analyse or interpret content. Also they apply equally well to any data type, as demonstrated in [19] where images are labelled and disambiguated without content analysis.

C. *Consensual image classification*

Consensual image classification and labelling can be either explicit, where user participation in the process is deliberate and voluntary, or implicit, where the user participation is a side-effect of some other activity. Explicit labels can be either amateur, as in flickr, or professional, commercially-motivated tags, such as in Getty Images.

However explicit labelling has a lack of distinct labels, plus a lack of users providing them [3]. Also many of the labels provided by explicit labellers are not for defining image content *per se* [13] but relate to the tagger's own organisational purposes, making them less globally-applicable as image labels.

Some consensual image classification approaches such as the ESP game (and concomitantly Google's Image Labeller), and flickr's tagging system rely on explicit user participation. These approaches generate classification data relatively slowly. It was claimed that 40 million images had been tagged with the ESP game since its introduction in 2003[1], yet this remains a drop in the ocean if one considers just how many images there may be on the Web – flickr's front page notes how many images are uploaded in the last minute. Extrapolating from that, we find that it will take under a week for 40 million images to be uploaded[2].

Clearly a more automated approach to image labeling is needed to supplement the existing methods. Domain-specific labels like flickr's apply only to the domain-hosted images, and other voluntary activities cannot keep up with the influx of new images. With search being so widespread, exploiting semantic information gained from search activity can provide larger quantities of classified images, as well as being applicable to any arbitrary Web-addressable image, regardless of its hosting site or ownership, just as long as it is indexed by one of the many search engines.

It is also evident that explicit user tags are often "local" in their scope, as opposed to global. While many tags are globally relevant, many also are of unclear relevance, not obviously pertinent to the content of the image, perhaps only relevant to a specific user group. A typical example would be a photograph labelled with a date or camera type.

Comparatively little research has been undertaken in entirely implicit image tagging. Implicit refers to an implied, but not specifically expressed, rating or meaning. With implicit image labelling, an image is being defined without a user writing a tag or selecting a category. Claypool et al. [3] investigated the use of implicit interest indicators in a Web environment. User actions were analysed including mouse movement, mouse clicks, page scrolling and time spent on a page. The authors noted that although these implicit indicators may be less accurate than explicit ratings, the sheer amount of data provides a very valuable yet rarely used resource. However their analyses were restricted to text searches and did not feature image search clickthrough data.

Joachims et al. [10] investigated the accuracy of clickthrough data as implicit feedback. They determined that using clicks as an absolute relevance judgement was difficult due to 'Trust Bias' where a user trusts the search engine's ordering of results, and 'Quality Bias', which means if the quality of results displayed is less relevant, users will select less relevant results. Xue et al. [21] also noted that users tend to select popular Web pages. After Joachims et al. [10] took this bias into account, they decided that clickthrough data was still 'reasonably accurate' for obtaining relative (if not absolute) implicit feedback. They did not consider images.

III. Using clickthrough data to label images

The two methods proposed here both rely on extracting clickthrough data from searches and associating the search term with the selected results. In the first method, the search term is associated directly with the image, while in the second the search term is associated with a text-based document and from there propagated on to contained images.

A. *Direct image labelling*

The direct image labelling method extracts clickthrough data for images (.gif, .jpg and .png) from raw Web logs. The results are then clustered into single-sense aggregations with membership determined by an image having been selected at least 2 times from the results page for the image search on the search term. The threshold 2 was chosen to match the ESP game [20] and the Google Image Labeller [8]. For each term and image pair, the actual number of selections is recorded so we could analyse whether or not the number of clicks correlated with the relevance of the images selected.

This is a very simple method for labelling images which acts essentially as a multi-user relevance feedback filter over normal image searches. As a filter, it obviously should perform better than the normal image searches over which it operates, and this result was confirmed in the experiment.

B. *Transitive (inherited) image labelling*

The transitive image labelling method labels HTML or PDF pages in the same way as images were labelled in section 3.1. For each page so labelled, this method propagates the label onto images contained within the Web page. However we did not propagate the label to all images - a number of exclusions were identified. The characteristics that seemed common to irrelevant images were as follows:

- **too small**: images with width or height below 50 pixels were usually not relevant to the Web page label. They generally represent punctuation (e.g. bullet points), emphasis (e.g. smileys) or structure (e.g dividers).
- **advertisements**: these were filtered out according to source URL, based on lists of advertisement sources.
- **repeated**: images repeated within a page were usually not pertinent to the content topic, e.g. bullet points.
- **logos**: anything matching *logo* was excluded.
- **aspect ratio**: anything too narrow was excluded.

Images that fell into these categories were filtered out as "non-content", while the remaining images were "content", i.e. relevant to the containing Web page label. Content images were labelled with the containing page's label.

[1] See http://ljiang.wordpress.com/

[2] On 10 September 2008, at one stage there were 4322 uploads in the last minute, prorating to over 40 million images in under a week.

IV. Experimental analysis of image labelling

We evaluated the relevance of both methods from section 3 and compared them with four other established image labelling methods. There was also a seventh group of images composed of the images rejected by the transitive labelling method (non-content images) We included this group to assess the accuracy of our image filtering method.

The six methods comprise:

- **Direct Labelling** – this method uses Web image searches and selections as described in section 3.1.
- **Transitive Labelling** – labelling images via clickthroughs as described in section 3.2.
- **Google Image Search (GIS)**[3] – this is Google's standard image search facility that determines image labels from page metadata or surrounding text.
- **Flickr**[4] – a Web site allowing photo storage and sharing. Images are explicitly tagged with labels.
- **Getty Images**[5] – a Web site licensing the use of professionally authored, tagged images.
- **Google Image Labeller (GIL)**[6] –implemented originally as the ESP game [20].

Obviously the methods are not all entirely independent, as the direct method relies on GIS (although other image search engines formed part of the clickthrough data). It is also possible that GIL is incorporated into GIS although the relatively low quality of labels in the analysis suggests that prudence in using such labels is required.

The original data from which the clickthroughs on both images and text searches were derived was generated from the University of Teesside School of Computer Science complete Web logs from March 2006 to the present.

A. *Experiment set up*

We needed to generate a set of search terms over which to evaluate the methods. This set was constrained by the Web log data we possessed. We selected 71 distinct query terms common to both text- and image searches as the set of all "included terms". We then generated image/label pairs for each of the methods as follows:

- Transitive labelling: for each included term, we used images from the 71 pages that were selected from searches on the search term. There were 445 unique image/label pairs generated this way. Of these, 260 were content and 185 were non-content. Of the non-content, 150 were categorised as "too small", 13 as "wrong aspect ratio", 5 as "logo" and 17 as "advertisement".
- Direct labelling: every selection from an image search on one of the 71 included terms in the Web log data generated a single image/label pair. There were 405 image/label pairs generated this way.
- GIS, Flicker, Getty Images: for each of the 71 search terms, a search was submitted to the site and the first-returned 14 (or fewer if not available) images selected, yielding 966, 958 and 931 image/label pairs respectively.
- GIL: the game is played with random images, so this method cannot be analysed using the same terms. Instead, four participants played approximately 200 rounds each of the game, recording image URLs and any 'off-limit' and 'included' tags. When images were offered a second time, previously-agreed labels had migrated to the "off-limits" list. 988 labels were generated this way.

The final set of image/label pairs to be ranked by users was quite large, 4693. To encourage participation, the ranking software minimised the effort required from users, presenting the image with a label above it, the user needing only to select a number as a ranking. Users could log in and terminate a session at any time, to continue later.

Users ranked each image/label pair on a Likert scale, with 0 meaning "not sure" or that the image was not accessible, 1 meaning "not relevant", 2 meaning partly relevant (e.g. "blue" applied as a label for a blue object) and 3 meaning fully relevant.

B. *Experimental data generated*

Over 20 days, around 100 users ranked varying numbers of image/label pairs. No pair was ranked fewer than 7 times and over 93% of pairs had over 10 rankings.

The ranking counts of all methods is shown here:

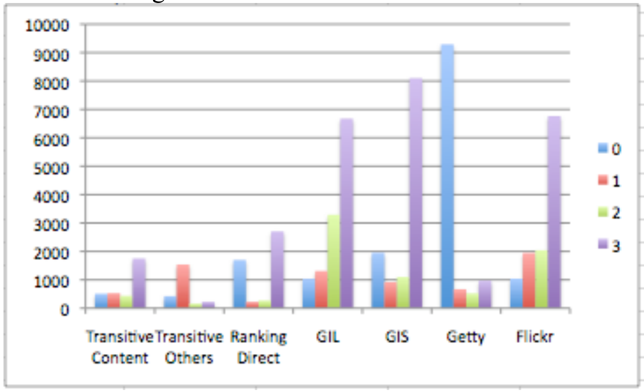

Figure 1. Count of rankings 0, 1, 2 and 3 for each method

The standard deviation and average rankings are:

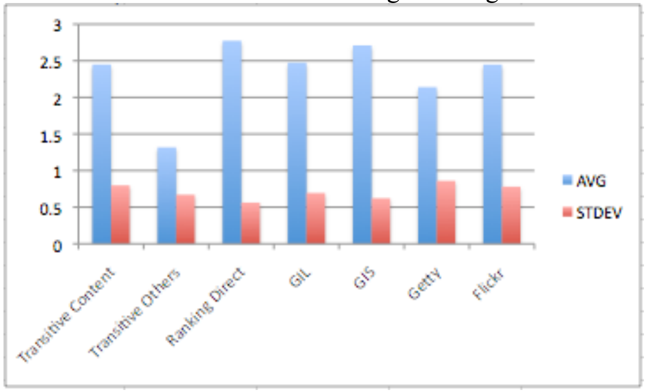

Figure 2. Average ranking and standard deviation across methods

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

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

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

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

V. Analysis

For each method, we calculated *precision* as the number of images ranked 3 divided by the total number of non-0 ranked images. Direct labelling at 0.8444 is 0.043 higher than GIS at 0.8006, which is in turn .16 higher than transitive at 0.6441, with flickr at 0.6293, GIL at 0.5921 and Getty at 0.4493.

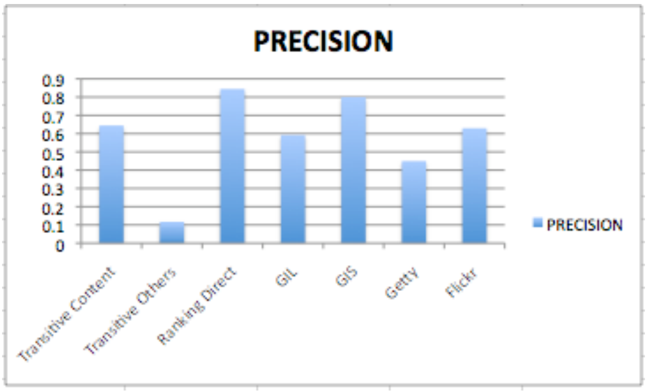

Figure 3. Precision (3 rankings only) of all methods

We also calculated *partial precision* as the number of times images ranked either 2 (partially relevant) or 3 (relevant) divided by the total number of non-0 ranked images. Factoring in the "maybe" rankings changes the relative positions of the methods. It reduces the lead of direct labelling (0.9294) over GIS (0.9094), with GIL at 0.8839 and flickr at 0.8190 overtaking transitive labelling, but transitive labelling closing the gap with GIS at 0.8051, and Getty images with 0.6909.

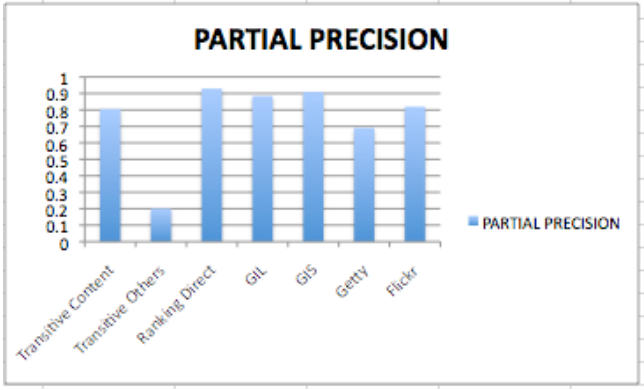

Figure 4. Partial precision (2 and 3 rankings) of all methods

Ignoring the transitive other category (non-content) images where the precision is known to be low (these being the excluded images), there is much less variation between the methods when partial precision is considered rather than full precision. The standard deviation across the methods (not including transitive others) is 0.1442 for precision and 0.0879 for partial precision. Thus including the 2-rankings ("maybe" rankings) significantly reduces the standard deviation, marking a lower level of disparity amongst the methods.

For the transitive labelling method, an exclusion accuracy calculation consisting of the number of images ranked 1 (not relevant) divided by the total number of images in the non-content categories and given a non-0 rating was also calculated:

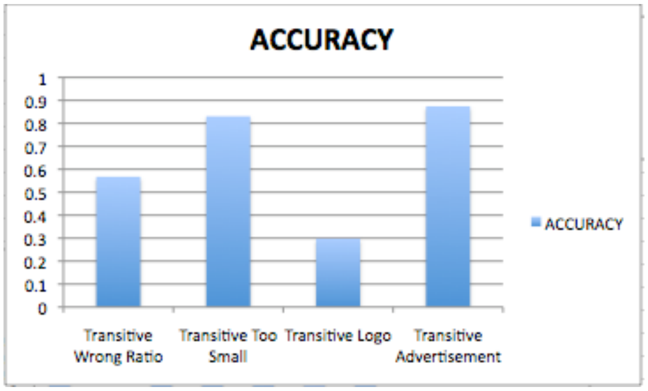

Figure 5. Exclusion accuracy across excluded image types

We could also estimate how many clickthroughs were required for the association between search term and selected object to be reliable. We extracted the number of clicks from which the label was derived ("selection weight") for all directly-labelled images ranked 2 or more overall. The following chart plots the percentage of 3-rankings and 2- and 3-rankings combined versus selection weight in the x-axis:

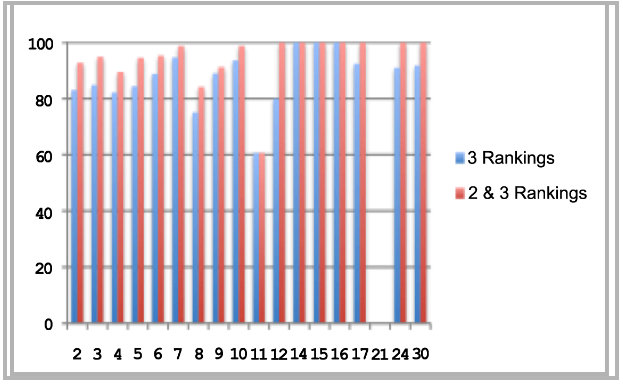

Figure 6. Selection weights versus precision and partial precision

0 rankings were given to images which either failed to load or which the user did not have an opinion regarding label relevance.

Most image/term pairs had been identified immediately prior to the experiment commencement, so a reasonably low level of failed URLs were anticipated. The exception was with direct labelling, as the searches that the labelling was based on dated back over a 2-year period.

Anecdotally however, there were significant numbers of failed URLs during the experiment. The most problematic was the apparent failure of Getty images to serve images on 9,291 occasions out of 11,452 total requests, no doubt due to the requesting site being blocked by Getty. There was a moderate quantity of information remaining on which to base analyses (668 1-ranked, 522 2-ranked and 971 3-ranked), but this problem does suggest that the accuracy of these rankings in this experiment may not be as robust as for the other methods assessed.

VI. Discussion

A. *Reliability of clickthrough data*

Labelling images with clickthroughs is predicated on the accuracy of the clickthroughs themselves - if the clickthrough data underlying either method is unreliable, the label accuracy of the label would be compromised.

In earlier work we established that image clickthroughs are generally much more accurate than text clickthroughs [18]. The experimental results support this finding. Direct labelling is essentially a relevance filter on image search (primarily but not exclusively Google image searches in our data), so it is to be expected that the accuracy would be higher than GIS, in this case by around 4.4% for relevant labels, and 2% for partial plus fully relevant labels.

However, the transitive labelling method was potentially hampered by the lower level of clickthrough accuracy in text-based search. The accuracy of images labelled this way could be compromised either by i) low-relevance clickthrough data, or ii) inadequacies in our content-detection algorithms. It was however a modestly successful method, outperforming Google's Image Labeller, flickr's amateur tagging and Getty Image's professional tagging methods in precision (rankings of 3) if not for partial precision (rankings of 2s and 3s).

To assess how the accuracy of the text-based clickthroughs affected the dependent transitive image labels, we firstly set up a smaller experiment to ground-truth the accuracy of the clickthroughs on the text pages containing the images, using pages from which some of the images analysed had been drawn. The pages were assessed by users with the same Likert scale as for the images (3 being relevant, 2 partially relevant and 1 not relevant, while 0 is unknown). While fewer users performed the analysis, each page received 5 or more human assessments. Site averages were derived, with 77% of sites assessed having an average of at least 2.5 (i.e. mostly 3 values), with another 10% with average value between 1.8 and 2.5, and 13% below 1.8. 94 of the 199 sites were unanimously rated as fully relevant (all 3s). These figures suggest that in the analysed data at least, the clickthroughs were reliable as indicators of content (but see section 6.4), and that perhaps our content-filtering algorithms were at least partly responsible for the poor precision.

We subsequently compared image ratings with their containing site ratings, in particular to know how relevant the labelling was when the site was relevant. We expected that badly-labelled sites would generate badly-labelled images (garbage in, garbage out) and the data showed that well under 0.4% of images were accurately labelled when their containing sites were badly-labelled, while over 71% of accurately-labelled images derived from accurately-labelled sites. However 13% of images were incorrectly labelled as relevant because they occurred in relevant sites - if the filtering could be improved then the transitive labelling accuracy could improve up to the level of the direct labelling. Other adjustments to the filtering are also necessary since many of the "too small" or "too narrow" images filtered out were in fact relevant.

B. *Consensus from two clickthroughs*

The number of clickthroughs did not appear to affect the label accuracy. In figure 6, the precision and partial precision of images labelled with the direct labelling method did not much alter over the number of clickthroughs. That is, better results are not necessarily gained from having more users agree on the click.

This essentially means that if two people agreed on the relevance of a search result to a search term, i.e. clicked through on it, then this was sufficient to establish its relevance generally. That is, the wisdom of crowds can often be represented by a consensus of two.

C. *Accuracy of non-consensual methods*

The precision for GIL is significantly lower than the clickthrough-based methods, although its partial precision brings it within 2% of GIS. Hence GIL offers a greater proportion of marginally relevant labels. One explanation is that GIL users are motivated more by quantity rather than quality of labels, encouraged perhaps by the rules of the game to agree on as many labels as possible. Also in some cases the set of high-quality labels seems to be exhausted, so agreement must be reached on only peripherally relevant tags, such as colours.

In common with GIL, other methods also looked much better when including partial precision. Given the reduced standard deviation from 0.1442 for precision to 0.0879 for partial precision, it is clear that the partial precision is something of an equaliser of the six methods.

Counting fully relevant results only, the precision of direct labelling and GIS were the only strong performers over .80, with transitive, flickr and GIL in the next clump between .59 and .65, and Getty images well down on the third clump at 0.4493.

All three of the implicit labelling methods (GIS, direct, transitive) outperform all three of the explicit methods (GIL, flickr, getty), if only just in some cases. One might think that a carefully-considered labelling process would generate better-quality, more relevant labels than automatically generated ones. However this seems not to be the case, and there are a number of possible reasons for this, such as the rationale for creating labels being at odds with the requirements for indexing and retrieval, or perhaps that the quality of labels generated for recreational purposes is low.

What we do see, however, is a much better partial precision. While Getty still lags behind on 0.69, the other methods all have a partial precision over 0.80. The explicit methods plus the transitive method have far more marginally relevant labels than the implicit methods. For the transitive method this is not surprising, as the only human judgment applied is at the page classification level. However for the explicit methods, what is curious is that deliberately-authored labels are significantly less reliable and relevant than GIS and direct labelling. The explicit methods result in a large number of labels that are not actually wrong but are not exactly right either.

For GIL, this may be due to constraints of the game or its motivation as discussed in 6.3, and for flickr, the labels are often compromised for general use by the limited scope of the tagger's intention. However what of Getty's poor showing? Getty is a professional site whose indexing would have commercial implications for the images on sale. It may be that Getty's searchers use more search terms than is characteristic in search engines generally and that image purchasers might search for very specific characteristics in an image, including dominant colours or emotive descriptors.

D. Other observations

There is scope for poor judgements and misconceptions to propagate more widely with consensual methods, e.g. three axolotl pictures were labelled "fish". This is not restricted to the consensual methods but does indicate that reliance on humans to act as an oracle, even en masse, is risky. The crowd does not always manifest wisdom.

Secondly, we noted that there is little ambiguity in the set of "included terms" in section 4. We noticed this as we are interested in using clickthrough data for disambiguation as done previously in [19] but found little to disambiguate. This may be due to the data having been collected from a single School in a single university, so that the user population has a much smaller set of search interests than the general population. Also the search terms indicated that many searches were goal-oriented, often relevant to student assignments. This may be one reason why our results disagree with for example Fergus et al. [4] (who claimed that in GIS up to 83% of images are labelled incorrectly).

It may also be that low levels of satisfaction with clickthrough data based on text-based searches (see 1.1) are related to the ambiguity of results which are not apparent in the excerpt of text returned. In our experiment, users were asked to rank a label as relevant regardless of what meaning of the word was implied, e.g. "Paris" applied to a picture of the Eiffel Tower was just as relevant as applied to a picture of an actress of that name. So in the other trials discussed in 1.1, the results sets may not have been wrong as such, but perhaps did not meet searcher requirements due to ambiguity of the search term.

VII. Conclusions

The two image labelling methods proposed and analysed in this paper have been shown to be viable. Our use of direct labelling demonstrates that implicit relevance feedback derived from clickthrough can be use to improve relevance while transitive labelling shows promise as an alternative method for achieving the same goal.

We found that all of the implicit image labelling methods (direct, transitive, GIS) are better than all of the explicit image labelling methods (GIL, flickr, Getty) in terms of precision. Deliberate, explicit human labelling seems to come a poor second place.

Finally, the number of clickthroughs creating a label seems to have little effect on its relevance. Consensus is apparently reliable with only 2 people.

References

[1] Almeida, J, Rocha, A, Torres, R & Goldenstein, S, 2008, *Making colors worth more than a thousand words*, Proc. ACM Symposium on Applied Computing, Brazil, 1184-1190.

[2] Chen, Y & Wang, JZ 2004, *Image Categorization by Learning and Reasoning with Regions*, J. Machine Learning Research, 5, 913-939.

[3] Claypool, M, Le, P, Wased, M & Brown, D 2001, *Implicit interest indicators*, Proc. 6th Intl Conf. Intelligent user interfaces, 33-40.

[4] Fergus, R, Fei-Fei, L, Perona, P & Zisserman, A 2005, *Learning object categories from Google's image search*, Proc 10th IEEE International Conference on Computer Vision, Vol. 2, pp. 1816–1823.

[5] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White. *Evaluating implicit measures to improve web search*. ACM TOIS, 23:147–168, 2005

[6] Fuxman, A, Tsaparas, P., Achan, K. & Agrawal, R. 2008, *Using the Wisdom of the Crowds for Keyword Generation*, Proc WWW2008, ACM, 61-70.

[7] Google Inc. 2008, *Google Advertising*, http://www.google.com/ads/indepth.html, 5 June 2008.

[8] Google Inc. 2008b, *Google Image Labeller*, http://images.google.com/imagelabeler/

[9] H. Halpin, V. Robu and H. Shepherd 2007. *The Complex Dynamics of Collaborative Tagging*, WWW2007, 211-220, ACM.

[10] Joachims, T., Granka, L., Pan, B., Hembrooke, H. & Gay, G. 2005, *Accurately interpreting clickthrough data as implicit feedback*, Proc. of SIGIR '05, ACM, 154 – 161.

[11] Joachims, T, Granka, L, Pan, B, Hembrooke, H, Radlinski, F & Gay, G 2007, *Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search*, ACM TOIS, 25(2), p. 7

[12] Liu, Z & Wada, S 2005, *Efficient Feature Extraction for Robust Image Classification and Retrieval*, Proc. 7th IEEE Workshop on Multimedia Signal Processing, China, 1-4

[13] Marlow, C, Naaman, M, Boyd, D & Davis, M 2006, *HT06, tagging paper, taxonomy, Flickr, academic article, to read*, Proc. Hypertext '06, Denmark.

[14] *ReCAPTCHA: Digitising books one word at a time*, http://recaptcha.net/learnmore.html, accessed 2008/01/22

[15] F. Scholer, M. Shokouhi, B. Billerbeck, and A. Turpin. *Using clicks as implicit judgments: Expectations versus observations*. In ECIR, volume 4956, 28–39 Springer, 2008.

[16] M. Shokouhi, F. Scholer, and A. Turpin. *Investigating the effectiveness of clickthrough data for document reordering*. In ECIR, v. 4956/2008, 591–595. Springer, 2008.

[17] Sigurbjörnsson & Zwol (2008), *Flickr Tag Recommendation based on Collective Knowledge*, Proc WWW2008, ACM, 327-336.

[18] G. Smith and H. Ashman, *Evaluating implicit judgments from Web search interactions*, Proceedings of Web Science 2009, Athens, 2009.

[19] M. Truran, J. Goulding and H.L. Ashman, *Co-active Intelligence for Information Retrieval*, Proc Multimedia '05, 547-550, ACM, 2005.

[20] von Ahn, L., Dabbish, L. 2004. *Labelling Images with a Computer Game* Proc. SIGCHI CHI04, 1, 319-326, ACM.

[21] Xue, G-R, Zeng, H-J, Chen, Z, Yu, Y, Ma, W-Y, Xi, W & Fan, W 2004, *Optimizing web search using web clickthrough data*, Proc CIKM 04, ACM, 118 - 126.