

Evaluating implicit judgements from Image search interactions

Gavin Smith

ACRC Computer Science
University of South Australia
gavin.smith@unisa.edu.au

Helen Ashman

ACRC Computer Science
University of South Australia
helen.ashman@unisa.edu.au

Abstract

Click-through data derived from traditional document search has recently been closely examined with some doubts cast on its perceived usefulness as a source of absolute relevance judgements. Conceptually similar, and holding many potential uses, image search click-through data has not seen such close examination, despite vastly different properties. Addressing this we detail and report findings from a user study involving over 65 participants. Aimed at determining the reliability of image search click-through data, factors identified in previous studies in document search click-through data are examined and results compared. In addition, additional factors uniquely present in web based image search are motivated and examined.

1. Introduction

The by-product of search engine interactions - the records of what users have clicked as a result of a query - known as “click through data” is an increasingly popular resource of implicit feedback. Recently, however, the validity of such feedback has been questioned [11, 18, 21]. Previously advocated for use in things such as: learning user models; query suggestion; improving search; and creating query hierarchies, a number of papers have shown greater consideration needs to be taken when evaluating the validity of such feedback. However, the research has largely focused on traditional document web search, for which users are only presented with short documents excerpts to evaluate the resource before clicking - and creating the “implicit feedback” between the query and resource. In contrast, other search types vary significantly with respect to what the user is presented with, and hence what is used to pass judgement and create this “implicit feedback”. Such variations are important as even small differences in such “caption features” can have significant impacts on user behaviour [7]. Of note is web-based image search interactions where the user is presented with a reduced thumbnail of the whole image.

Image search click-through data has also been shown to have many potential uses in recent works [6, 8, 22, 26], however, unlike traditional web search interactions there has been no previous research that has explicitly looked into

its validity as implicit judgements. This work presents a study resulting in a rich data-set enabling comparison with what has been discovered for traditional document search [11, 18, 21]. In addition, motivated by their impact on the thumbnail and hence “caption”, searches for different types of images are evaluated. Such an evaluation is motivated by the impact of small caption features in traditional web search [7]. While conceptually similar, the results of previous studies based on web search interactions can not be directly transferred to image search interactions due to the aforementioned difference in the summary displayed and used by the user in the decision to click.

2. Related work

Implicit feedback in the form of click-through data was just exploited in 1995 with Lieberman [20] using it to help the proposed system determine relevant links for the user. Since then it has been proposed to aid a number of problems including training data for learning user models [2, 4] and document retrieval functions [16], for improving search [1, 28], for query suggestion when comparing complex question style queries [27], clustering related webpages and queries [5] and more recently in auto-generating query hierarchies [24]. However, the theoretical foundation of the use of click-through data as a source of implicit human relevance judgement has recently been questioned. In 2005 Fox *et al.* [11] examine click-through behaviour in regard to standard web search. They found that, used in conjunction with other implicit feedback (such as dwell time), judgements that correlated well with explicit judgments could be determined. Used in isolation, however, such a correlation did not occur, with users satisfied with the returned document only 39% of the time after clicking a link. Also in 2005 Joachims *et al.* [17] reject the use of click-through data for absolute relevance judgement, instead indicating its usefulness as relative judgements, i.e. clicked data is more relevant than un-clicked data. More recently in 2008 Shokouhi *et al.* [25] found proposed search engine document reordering techniques based on click-through data do not always lead to improvements in search quality and may even have a detrimental effect. Also in 2008 Scholer *et al.* [21] presented more evidence against the use of click-through data reporting only a 52% correla-

tion between clicked documents and those the users thought relevant and a 58% agreement rate between the click-data and TREC judgements for the collection used in the study.

All of the above mentioned studies were conducted using click-through data from document web search interactions. However, click-through data is not solely generated from such searches, with image search being one such alternative. Such searches vary from traditional web search in a number of ways. Typically image search results are a set of thumbnails, sometimes with short captions. Such summary information is intuitively more complete than the two or three line summary information displayed for traditional web search with a number of researchers agreeing with such an intuition. The primary argument put forward is that thumbnail summaries contain both more information and that the information can be absorbed faster, resulting in less poorly informed clicks than traditional web search [6, 8]. As such, results based on web document search click-through data should not be arbitrarily generalized to all forms of click-through data. Addressing this, the present work details a study focusing on image search click-through data based on similar studies into traditional web search click-through data [11, 18, 21]. The focus on image search click-through data is motivated primarily by two factors: firstly due to its provision in popular commercial search engines such as Google and hence the existence of such data; secondly due to recent work in the research community using such data.

Image search click-through data has been proposed for a number of uses. In 2005 Truran *et al.* [26] proposed the use of click-through data as a way of removing the noise introduced by polysemy. In 2006 Cheng *et al.* [6] proposed the use of click-through data as relevance feedback for a technique across visual and textual features in image search. In 2007 [8] Craswell and Szummer looked at a technique for addressing the sparsity problem within click-through logs. They succinctly highlight some of the major uses for click-through data, these being improving search, query suggestion, resource annotation and relevance feedback. None of the work, however, adequately addresses the underlying question of the implicit judgments' accuracy, either in the general case or under the varying conditions associated with search results and query types. Cheng *et al.* motivate their choice of click-through data based on intuition alone and validate their results using artificially generated click-through data. Truran *et al.* neither use real world data, nor cover many types of queries. Finally, while using significant quantities of real world data, Craswell and Szummer do not focus on measuring the accuracy of such feedback in general, but rather seek to improve it.

3. Potential factors affecting image click-through data

The presented study evaluates four factors with the potential to impact user click behaviour in web image search and

hence impact the accuracy of the associated click-through data. The first is derived from observations and studies into document search based click-through data. In document based search, Joachims *et al.* [18] highlighted the impact of the quality of the system on accuracy of click-through data and while the results from [21] do not seem to support such an assumption, they do not evaluate true extremes of precisions of the underlying system. Regardless such a factor has the potential to affect image click-through data and as such is included in this study.

The remaining three are motivated by the impact of “caption features” on click-through data as reported with regard to document search based click-through data [7] and the vastly different properties image queries can have [10]. Since image search results typically are returned in the form of a series of thumbnails, “caption features”, as referred to in traditional document search, can be seen to be qualities of the image. We acknowledge that in some image search systems results are returned accompanied with text snippets, however, this is not always the case. In this instance we choose to examine the systems without text. The four qualities of the image that we highlight as potential influencing factors have been motivated by research in both image and image query classification [3, 9, 13, 14, 15, 19, 23].

The first quality selected as a factor is based on categorization proposed initially by [23] and reiterated and refined more recently in [3, 14]. According to this, three categories are identified, *general*, *specific* and *abstract*. In defining the *generic* category we adopt the definition as stated by [14] with queries falling into this category which require only “general everyday knowledge” to recognize the objects or scenes. In contrast, queries falling into the *specific* category refer to things that can be identified and named [14]. As such they require a greater level of knowledge (even though this knowledge might be common knowledge). An example includes the query *person* at the generic level and then *Kirsten Dunst* at the specific level. When a user searches they may or may not have this knowledge and this forms another potential factor affecting image search behaviour and hence click-through data. As such we evaluate the two extreme cases, where the user has the knowledge (known) and when the user does not (unknown). We investigate this due to the potential difference in search behaviour, and hence click-through data, that may occur as a user uses the search as a way or learning as well as acquiring images. Such use of an image search engine has not been previously investigated, and hence can not be discounted as atypical use. These facets do not apply to the category of *generic of* as by definition it only requires “general everyday knowledge” [14]. The third category, “About”, as proposed originally by [23] is not evaluated due to its interpretative and subjective nature, which prevents accurate ground truthing. Rather, the scenarios described by researches in image categorization are described to a level of detail which then fit into one of the former two

categories. The motivation behind this is two-fold. Firstly, as mentioned, objective evaluation is not possible due to the subjective nature. Secondly interpretation and subjectivity can be seen to simply lead a lower/variable mean average precision of the search engine that returns a singular and uninformed interpretation of the users intention which is amplified in the case of subjectivity. For example happiness, once the subjectivity is removed by a concrete description of imagery, such as “any image containing a person smiling” can then be seen to be part of the Generic Action category.

The third factor is drawn from a further break down of the above categories as proposed by [9, 14] who make a distinction between objects and actions and scenes. These subcategories are investigated as they change the visual pattern that a user is looking for when searching.

A final factor that, while present in the literature, we do not investigate is the distinction between visual (such as colour and texture attributes) and conceptual descriptions of images. The reason for this is the lack of prevalence as a category of actual image search queries with [15] reporting that less than 2% of web based images fall into this category.

4. Experimental Design

Web document search click-through data has been evaluated from a number of angles using a variety of approaches. [11] developed a plugin in order to record several types of user behaviour, explicitly requesting user feedback about each selected result in two real world search engines. Based on the explicit judgements they were able to calculate the overall accuracy of their click-through data as 39%. Potential reasons for accuracy levels of click-through data, however, were not investigated. In contrast [18] specifically examine the reliability of the implicit feedback given, determining that click-through data is affected by both the quality of the search engine and the trust the user has in the system. Their system was also backed by a commercial search engine for which they used a proxy server to manipulate results. Differing from that of Fox *et al.* they attempt to both control factors they considered to influence search behaviour, and hence user clicks, and also use a laboratory setting. More recently [21] investigated click-through data in a more controlled fashion, generating and manipulating the results based on ground truths derived from the TREC WT10g collection. Such control was used to ensure levels of system performance in order to investigate how click behaviour varies as the quality of the underlying search system changes.

As a primary goal is to examine the impact of the quality of the system on click-through accuracy under differing levels of system accuracy we explicitly fix the precision levels. In a similar fashion to [21] this is achieved by creating a controlled retrieval system using images given a ground truth in advance, ensuring such variance. While reversing result orderings to vary the system precision (as performed by [18]) is a simpler approach and allows real world search

results to be used, the two-dimensional ordering of image search results coupled with the lack of previous research indicating a degradation of performance with such an approach (as was available and cited by [18] with regard to document based search) makes the validity of such an approach open to debate. Where possible images with ground truths from the IAPR TC-12 Benchmark dataset [12] were used. Additional images were selected and given ground truths in the same manor as was originally employed for generating the dataset. The complete dataset used in the study consisted of 2880 images (1440 relevant and 1440 irrelevant) enabling 8 pages per topic/AP combination to be displayed to the subjects.

The additional factors identified as potential influencing factors (see section 3) were then encapsulated in six categories (shown in table 1) and evaluated under two fixed levels of system accuracies. These fixed levels of accuracy (precision levels) were set at extremes of 16.6% and 83.3% which represents 2 out of 12 and 10 out of 12 correct images per page respectively.

For each category 3 topics were selected. Topics were ImageCLEF style topics consisting of both a title and a narrative. *Generic* category topics were selected from the ImageCLEF topics by first categorizing all topics with more than 100 available ground truths and then randomly selecting a topic for each group for which one or more topics were identified. *Specific* category topics were developed to maximize the known/unknown response in the pre-topic questionnaire, taking into account typical query types as reported by [15]. The selected topics are shown in table 1. For each topic each participant was presented with a pre-topic questionnaire to establish their familiarity with the topic in order to enable proper processing of the known/unknown facets being examined.

5. Discussion and comparison with recent web search findings

The study ran for approximately one week as a 1 hour voluntary online study which was advertised primarily to university students and researchers. As an incentive a prize draw was conducted with an entry given to participants who completed the study. The study was able to be stopped at any time and, if desired, resumed later. 116 people participated, of whom 67 completed the entire study. As in [21] a pre-experiment question was administered aimed at establishing the participants familiarity with online image searching. Most participants considered themselves to have average experience at using image search engines, with a mean rating of 3.5 on a scale of 1 (no experience) to 5 (a large amount of experience). The mean rating for performing image searches was only 2.9 times per week. Participants reported a mean enjoyment rating of 3.4 also on a 1 (dislike) to 5 (high enjoyment) scale. Such results differ somewhat from those reported by the participants in [21] with regard to traditional

Category	Topics
Generic Scene	group in front of mountain landscape tennis player on tennis court winter landscape
Generic Object/Action	straight road bird flying photos of dark-skinned girls
Specific Scene (Known)	Paris, France London, England Sydney, Australia
Specific Object (Known)	George W. Bush Coca Cola branded can Eiffel Tower
Specific Scene (Unknown)	Baku, Azerbaijan Quito, Ecuador Tbilisi, Georgia
Specific Object (Unknown)	Shwezigon Pagoda Ali Abdullah Saleh (President of Yemen) Ushabti/Shabti/Shawabti

Table 1. The six categories evaluated and their corresponding topics

document search who indicate mean ratings of 4.7, 7+ and 4.2 for experience, number of searches per week and enjoyment respectively.

The results of the study show an increased level of accuracy of the click-through data with users indicating the clicked result relevant 88.1% of the time. This is markedly higher than the 39% satisfaction with clicked document search results reported by [11] and the 52% click-through document relevance reported by [21]. Interestingly the accuracy found by looking at the proportion of images with relevant ground truths is slightly lower (84%). At the category level, the reason for this difference becomes more apparent with the results showing the accuracy is not uniform across search categories and knowledge levels of users when system precision is low, as shown in table 2. Since the difference in perceived accuracy of the clicks is greater for the categories where the subjects did not think they could visually identify what was described by the topic, one possible explanation is that when the system precision was low the participants had a hard time learning the difference between object/scenes and that while similar, are strictly not correct. An alternate explanation is that participants simply clicked and marked images as relevant to simply complete the study. Such an explanation is unlikely for a number of reasons. Firstly to complete the study participants did not need to click any images. Rather they could simply indicate that they had finished the task. Secondly if this was the case it would be expected that the difference between participant indicated, and ground truth relevance, would be somewhat constant over all categories at low system precision which

does not seem to be the case. In addition a manual examination of 20 random images selected by participants as relevant but not indicated as such by the ground truth for unknown topics lends partial evidence to former explanation, with the majority being able to be explained in this way. The presence of such results cast some doubt on the reliability of subject detected relevance methods when examining the true accuracy of click-through data.

In comparison when the precision of the system is high the participant indicated relevance and ground truths are much closer (table 3). A possible explanation is that when users see large numbers of similar images they quickly learn the concept by placing trust in the system.

The final facets evaluated by the study, specific *vs.* generic and object *vs.* scene show little impact although further analysis is needed an planned for future work.

Category	P(U)	P(G)
Specific of scene (unknown)	0.836	0.587
Specific of object (unknown)	0.804	0.610
Specific of scene (known)	0.859	0.738
Generic of scene	0.902	0.752
Generic of object	0.886	0.835
Specific of object (known)	0.890	0.836

Table 2. Mean click-through relevance proportions for system precision of 16.67%. P(U): The proportion based on the number of results clicked the subject explicitly deemed relevant *vs.* the total number of clicks. P(G): The proportion based on the number of results clicked judged as relevant by the ground truth judges *vs.* the total number of clicks.

Category	P(U)	P(G)
Specific of scene (unknown)	0.904	0.949
Specific of object (known)	0.914	0.951
Specific of object (unknown)	0.878	0.961
Specific of scene (known)	0.903	0.966
Generic of object	0.887	0.967
Generic of scene	0.900	0.969

Table 3. Mean click-through relevance proportions for system precision of 83.33%. P(U): The proportion based on the number of results clicked the subject explicitly deemed relevant *vs.* the total number of clicks. P(G): The proportion based on the number of results clicked judged as relevant by the ground truth judges *vs.* the total number of clicks.

6. Conclusions

Addressing the reliability of web image search click-through data, in comparison to traditional web based document search, we motivate and present an in-depth user study. Based on similar studies into document search click-throughs by [11, 18, 21] we examine factors shown to undermine the

reliability of such data [21] as well as additional factors uniquely present in image search. The results of this study indicate that image search click-through data is considerably more accurate in general than document based search click-through data, although potential issues exist surrounding users searching with low levels of knowledge when coupled with poor performance levels of the search engine. Future work will seek to explore this and other properties in greater detail. Additionally, work needs to be done to determine how often users perform image searches with limited knowledge of their task. Such a property was not able to be investigated in this study due to the limited, pre-fixed image search content. It is possible that given such a task users in the real world would more likely do a document search first to understand what they are looking for.

References

- [1] Eugene Agichtein, Eric Brill, and Susan Dumais. Improving web search ranking by incorporating user behavior information. In *SIGIR*, pages 19–26, Seattle, Washington, USA, 2006. ACM.
- [2] Eugene Agichtein, Eric Brill, Susan Dumais, and Robert Ragno. Learning user interaction models for predicting web search result preferences. In *SIGIR*, pages 3–10, Seattle, Washington, USA, 2006. ACM.
- [3] Linda H. Armitage and Peter G.B. Enser. Analysis of user need in image archives. *Journal of Information Science*, 23:287–299, August 1997.
- [4] R. Baeza-Yates, C. Hurtado, M. Mendoza, and G. Dupret. Modeling user search behavior. In *Web Congress, 2005. LA-WEB 2005.*, page 10, 2005.
- [5] Doug Beeferman and Adam Berger. Agglomerative clustering of a search engine query log. In *SIGKDD*, pages 407–416, Boston, Massachusetts, US, 2000. ACM.
- [6] En Cheng, Feng Jing, Lei Zhang, and Hai Jin. Scalable relevance feedback using click-through data for web image retrieval. In *ACM-MM*, pages 173–176, Santa Barbara, CA, USA, 2006. ACM.
- [7] Charles L. A. Clarke, Eugene Agichtein, Susan Dumais, and Ryen W. White. The influence of caption features on click-through patterns in web search. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 135–142, Amsterdam, The Netherlands, 2007. ACM.
- [8] Nick Craswell and Martin Szummer. Random walks on the click graph. In *SIGIR*, pages 239–246, Amsterdam, The Netherlands, 2007. ACM.
- [9] J.P. Eakins. Techniques for image retrieval. *Library & information briefings*, pages 1–15, 1998.
- [10] P. G. B. Enser, C. J. Sandom, J. S. Hare, and P. H. Lewis. Facing the reality of semantic image retrieval. *Journal of Documentation*, 63:465–481, 2007.
- [11] Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. Evaluating implicit measures to improve web search. *ACM TOIS*, 23:147–168, 2005.
- [12] Michael Grubinger. *Analysis and Evaluation of Visual Information Systems Performance*. PhD, Victoria University, 2007.
- [13] L. Hollink, A. Th. Schreiber, B. J. Wielinga, and M. Worring. Classification of user image descriptions. *International Journal of Human-Computer Studies*, 61:601–626, November 2004.
- [14] A. Jaimes and S. F. Chang. A conceptual framework for indexing visual information at multiple levels. *IS&T/SPIE Internet Imaging*, 3964:2–15, 2000.
- [15] Bernard J. Jansen. Searching for digital images on the web. *Journal of Documentation*, 64:81 – 101, 2008.
- [16] Thorsten Joachims. Optimizing search engines using click-through data. In *SIGKDD*, pages 133–142, Edmonton, Alberta, Canada, 2002. ACM.
- [17] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR*, pages 154–161, Salvador, Brazil, 2005. ACM.
- [18] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Trans. Inf. Syst.*, 25:7, 2007.
- [19] C. Jorgensen. *Image Retrieval: Theory and Research*. Scarecrow Press, 2003.
- [20] H. Lieberman. Letizia: An agent that assists web browsing. In *IJCAI*, volume 14, pages 924–929, Montreal, Quebec, Canada, 1995.
- [21] Falk Scholer, Milad Shokouhi, Bodo Billerbeck, and Andrew Turpin. Using clicks as implicit judgments: Expectations versus observations. In *ECIR*, volume 4956, pages 28–39. Springer, 2008.
- [22] A. Sharma, Gang Hua, Zicheng Liu, and Zhengyou Zhang. Meta-tag propagation by co-training an ensemble classifier for improving image search relevance. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on*, pages 1–6, 2008.
- [23] S. Shatford. Analyzing the subject of a picture: A theoretical approach. *Cataloging and Classification Quarterly*, 6(3):39–61, 1986.
- [24] D. Shen, M. Qin, W. Chen, Q. Yang, and Z. Chen. Mining web query hierarchies from clickthrough data. In *AAAI*, volume 22, page 341, 2007.
- [25] Milad Shokouhi, Falk Scholer, and Andrew Turpin. Investigating the effectiveness of clickthrough data for document reordering. In *ECIR*, volume 4956/2008, pages 591–595. Springer, 2008.
- [26] Mark Truran, James Goulding, and Helen Ashman. Co-active intelligence for image retrieval. In *ACM-MM*, pages 547–550, Hilton, Singapore, 2005. ACM.
- [27] Ji-Rong Wen, Jian-Yun Nie, and Hong-Jiang Zhang. Clustering user queries of a search engine. In *WWW*, pages 162–168, Hong Kong, China, 2001. ACM.
- [28] Gui-Rong Xue, Hua-Jun Zeng, Zheng Chen, Yong Yu, Wei-Ying Ma, WenSi Xi, and WeiGuo Fan. Optimizing web search using web click-through data. In *CIKM*, pages 118–126, Washington, D.C., USA, 2004. ACM.