Are clickthrough data reliable as image annotations?

Theodora Tsikrika [1], Christos Diou [2,3], Arjen P. de Vries [1,4], Anastasios Delopoulos [2,3]
[1] CWI, Amsterdam, The Netherlands
[2] Multimedia Understanding Group, ECE Dept., Aristotle University of Thessaloniki, Greece
[3] Informatics and Telematics Institute, Centre for Research and Technology Hellas
[4] Delft University of Technology, Delft, The Netherlands
Theodora.Tsikrika@cwi.nl, diou@mug.ee.auth.gr, arjen@acm.org, adelo@eng.auth.gr

Abstract

We examine the reliability of clickthrough data as concept-based image annotations, by comparing them against manual annotations, for different concept categories. Our analysis shows that, for many concepts, the image annotations generated by using clickthrough data are reliable, with up to 90% of true positives in the automatically annotated images compared to the manual ground truth. Concept categories, though, do not provide additional evidence about the types of concepts for which clickthrough-based image annotation performs well.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.1 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation, Performance

Keywords

Image annotation, concepts, search logs, clickthrough data, collective knowledge, implicit feedback

1 Introduction

The application of text-based techniques in the retrieval of images requires their annotation with textual metadata, typically *captions* or *tags* that tend to describe both the content and context of the images they accompany [11], and/or semantic *concepts* that aim at unambiguously describing their visual content. Whereas captions and tags are commonly manually assigned, an active research field investigates the automatic *concept-based image annotation* through methods based on machine learning approaches. However, the supervised machine learning approaches that are widely employed in the automatic annotation of images with semantic concepts require the availability of labelled samples to be used as training data; these are typically generated manually, a laborious and expensive endeavour.

To compensate for the high cost in manually labelling training samples, research has recently moved towards the use of alternative data sources that are automatically acquired from the Web in order to be used for training concept classifiers [15, 8, 1, 14]. Our recent research [13] in the

context of the VITALAS European project (`http://vitalas.ercim.org/`) has proposed and investigated the use of a previously untapped source for acquiring such examples: the *clickthrough data* logged by retrieval systems. Such data consist of users' queries, together with the images in the retrieval results that these users selected to click on. This information can be viewed as a type of users' *implicit feedback* [5] that provides a "weak" indication of the relevance of the image to the query for which it was clicked on [2]. We refined the notion of relevance in this assumption by considering that the queries for which a image was clicked provide in essence a "weak" description (or *annotation*) of the image's visual content and used these "weakly annotated" images as training samples to build classifiers for 25 concepts. Our experimental results indicated that the contribution of search log based training data is positive; in particular, the combination of manual and automatically generated training data outperforms the use of manual data alone.

However, this recent work has not examined in detail the reliability of clickthrough data as concept-based image annotations; to this end, this paper explores how accurately clickthrough-based image annotations correspond to explicit concept-based manual annotations. Previous research has examined the reliability of clicks as relevance assessments, both in text [4, 10] and image [12] retrieval, but to the best of our knowledge no previous work has investigated their reliability as concept-based image annotations.

2 Approach

This section presents our methods for automatically annotating images with concepts based on clickthrough data; each concept is considered to correspond to a clearly-defined, non-ambiguous entity, represented by its *name*.

The simplest clickthrough-based method for selecting the images to annotate with a concept is to consider the images clicked for queries that *exactly match* the concept's name; we denote this method as *exact*. Given the sparsity of clickthrough data [2], we also need to apply methods with less stringent criteria for matching the queries in the clickthrough data to concepts' names. For each image, we use the terms in the queries for which the image has been clicked to create a description for it (similar to [9]). These textual descriptions can then be used to index and retrieve images in response to text queries. To this end, we employ a *language modelling* (LM) approach [3] to retrieve the indexed images using the concept name as the query.

In this approach, a language model φ_I is inferred for each image I. Given query Q, the images are ranked by estimating the *likelihood of the query*:

$$P(\mathbf{q}|\varphi_I) = P(q_1, q_2, \ldots, q_k|\varphi_I) = \prod_{j=1}^{k} P(q_j|\varphi_I) \tag{1}$$

assuming that each q_j is generated independently from the previous ones given the language model of the image's textual description. The simplest estimation strategy for an individual term probability is the *maximum likelihood estimate* (*mle*), which corresponds to the relative frequency of a term t_j in the textual description of an image.

Given that Equation 1 assigns zero query likelihood probabilities to images missing even a single query term from their description, we apply *smoothing* techniques to address this sparse estimation problem. We use a mixture model of the language model of the image's textual description with a background model (the collection model in this case), a technique well-known in text retrieval as Jelinek-Mercer smoothing [3]:

$$P(\mathbf{q}|\varphi_I) = \prod_{j=1}^{k} (1-\lambda) P_{mle}(q_j|\varphi_I) + \lambda P_{mle}(q_j|\varphi_C) \tag{2}$$

where λ is a smoothing parameter (typically set to 0.8), and $P_{mle}(t_j|\varphi_C)$ is the document frequency of the term t_j in the collection.

The aim of these LM-based selection strategies is to increase the number of clickthrough-based annotated images by progressively relaxing the strictness of the matching criteria. We apply 4 variants of this approach: unsmoothed LM without stemming (*LM*), unsmoothed LM with stemming (*LM stem*), LM with Jelinek-Mercer smoothing and no stemming (*LMS*), and LM with Jelinek-Mercer smoothing and stemming (*LMS stem*).

Our final method exploits the *clickgraph* on the premise that the visual contents of images clicked for the same query are likely to pertain to similar semantic concept(s). For each concept, an initial image set is formed with the images selected using the *exact match* method. If this method does not produce any results, we consider instead the images clicked for the most similar query to the concept name (using *LM* as our retrieval model). This initial image set is then expanded with the images accessible by a 2-step traversal of the graph. First, each image i in this initial set is added to a final set. For each such i, we first find the queries for which this image was clicked, and then add to the final set the images (other than the ones already there) clicked for that query. Alternative approaches that exploit the clickgraph are iterative methods, such as the random walk models employed in [2].

3 Experiments

3.1 Data

Belga news agency (`http://www.belga.be`) provided us with 101 days of image search logs that contain 96,420 images clicked for 17,861 unique queries that have been "lightly" normalised, including conversion to lower case and removal of punctuation, quotes, and the term "and". These photographic images cover a broad domain and are accompanied by textual captions written by the agency's professional archivists. Given that these search log data are obtained from a commercial portal, they are much smaller in size, compared to those collected by general purpose search engines [2]. On the other hand, given that this agency provides services to professional users, mainly journalists, we expect their search log data to be relatively less noisy. For each of the 111 VITALAS concepts listed in the first column of Table 2 (at the end of this paper), we applied each of the 6 methods presented in Section 2. This resulted in 6 sets of images per concept *automatically* annotated with that concept. We considered only the sets that contained at least 10 images; Table 2 lists the number of images per method for each of the 111 conceps. Belga then created a reliable ground truth by *manually* annotating the per-concept image sets with that concept, assuming the presence of each concept in an image to be binary.

3.2 Clicks vs. manual annotations

We evaluated the accuracy of the automatic clickthrough-based annotations by comparing them to the manual annotations and measuring their agreement (i.e., finding the true positives in the clickthrough-based annotations). Figure 1 shows for each method the total number of concepts for which that method produced results and the distribution of those concepts across various levels of agreement. The figure indicates that the level of agreement between manual and clickthrough-based annotations varies greatly across concepts. Across methods, though, there exists a number of concepts, around 20% of the total number of concepts for each method, that reach agreement of at least 0.8. This observation leads us to the question: could we characterise the *types* of concepts that can be reliably annotated using clickthrough data?

3.3 Concept categories

We determine the types of concepts by assigning them to categories based on those proposed by LSCOM [7], which have been previously applied to the categorisation of over 1000 concepts [6]. The difference is that we provide explicit descriptions of these categories (see the 2nd column of Table 1) and also add *animals* as a separate category. The concept categorisation in Table 1 has been agreed by 3 subjects.

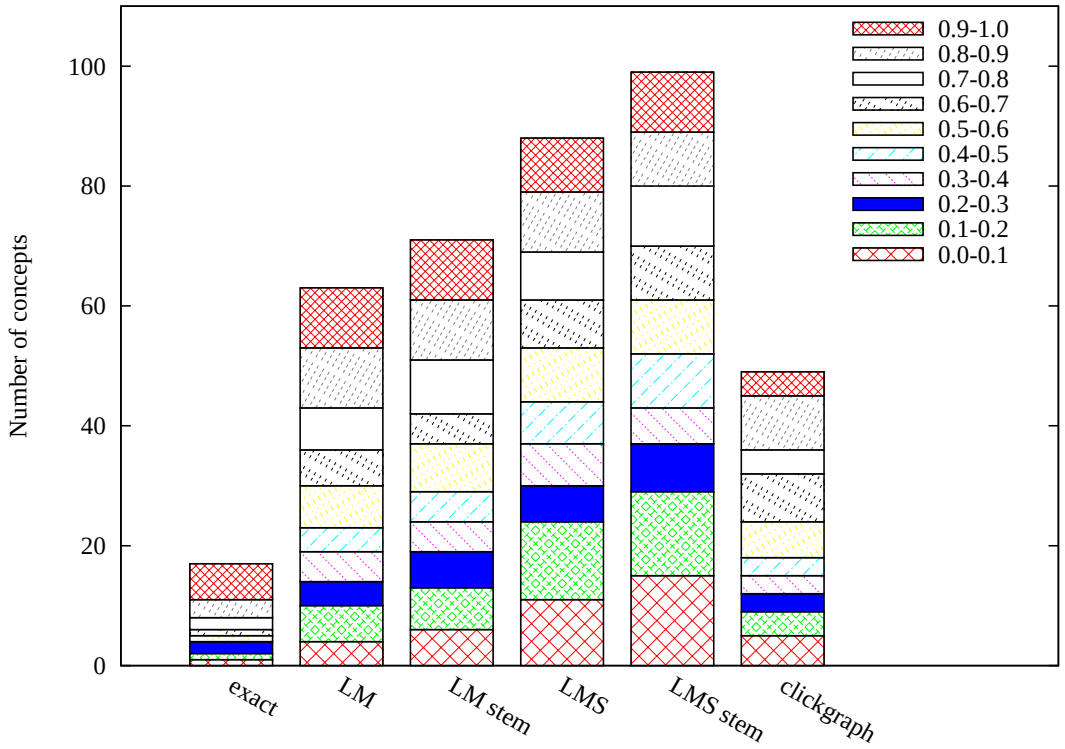

Figure 1: Agreement per method.

Table 1: Concept categories, their descriptions, and the categorisation of 111 concepts.

Category	**Category description** (i.e., a concept is classified under this category if it describes ...)	**Concepts**	**# of concepts**	
image theme	a broad area of interest	classic, european_union, rally_motorsport, soccer, etc.	11	9.91 %
setting/scene/site	a specific place or land site or the environment in which something is set	airport, airport_terminal, amusement_park, art_gallery, atomium, bank, beach, bridge, church, court, damaged_building, disco, european_parliament, gas_station, highway, hospital_room, house, internet, mountain, parliament, road, school, stock_exchange, theatre_building, traffic	25	22.52 %
people	a human being, a group of human beings, or human body parts	abused_child, ac_milan_soccer, agricultural_people, anderlecht, arsenal_fc, artist, baby, belgium_royal, bush, child, club_brugge, factory_worker, fashion_model, federer, finger, girl, government, jacques_chirac, king, lawyer, parent, pope_benedict, queen, red_devils, sex, teenager	26	23.42 %
animals	an animal or a group of animals	animal, cat, dog, horse, lion	5	4.51 %
objects	a physical, tangible, and visible entity, excluding people,animals, & entities occupying land sites	apple, belgian_flag, boat, bus, car, computer, european_flag, fish, flag, street_sign, television, usa_flag	12	10.81 %
activities	a specific behaviour or action taking place	beach_leisure, car_racing, children_care, children_playing	4	3.60 %
events	something that happens at a given place and time, including natural events (e.g., fires and avalanches) and social events (e.g., shows, social functions, contests, and competitions)	airplane_crash, australian_open, award, cannes_festival, car_accident, champions_league, davis_cup, flood, demonstration, earthquake, election, explosion, fashion_show, festival, fire, flood, formula_one, gala, goal, hurricane_typhoon, memorial_services, olympic_games, parade, roland_garros, storm, tour_of_flanders	25	22.52 %
graphics	any form of artificially generated visual content	cartoon, illustration, logo	3	2.71 %
		Total	**111**	**100.00%**

Figure 3 shows the percentage of concepts in each category that reach, for at least one of the 6 methods, agreement over 0.6, 0.7, 0.8, or 0.9. A substantial number of concepts, around 50% in each category (apart from *activities*), reach an agreement of at least 0.6, a first indication of the usefulness of clickthrough-based image annotations. The small number of concepts in the *animals* and *graphics* categories do not allow us to reach reliable conclusions, and are not considered further. The concepts with the highest levels of agreement (over 0.9) are those categorised as *setting/scene/site*, *people*, and *events*. Surprisingly, the *objects* type concepts do not perform that well; this could be attributed to the textual metadata not including the visually obvious. This motivates research in the vision community for developing automatic image annotation for such types of concepts.

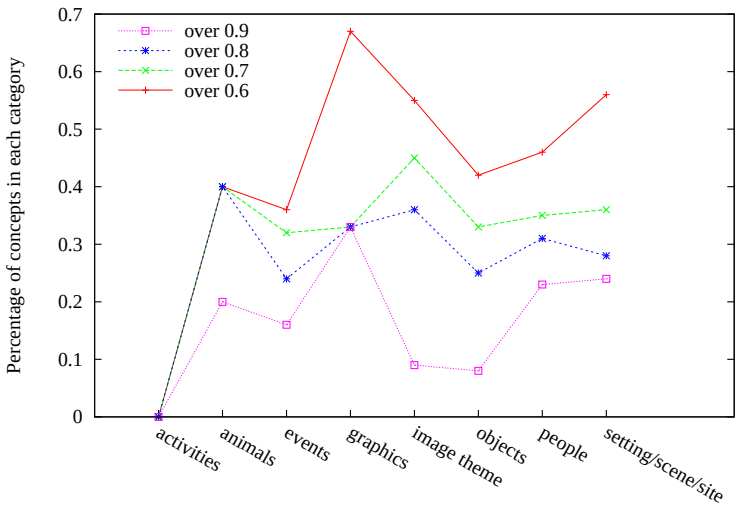

Figure 2: Agreement per category for all 111 concepts.

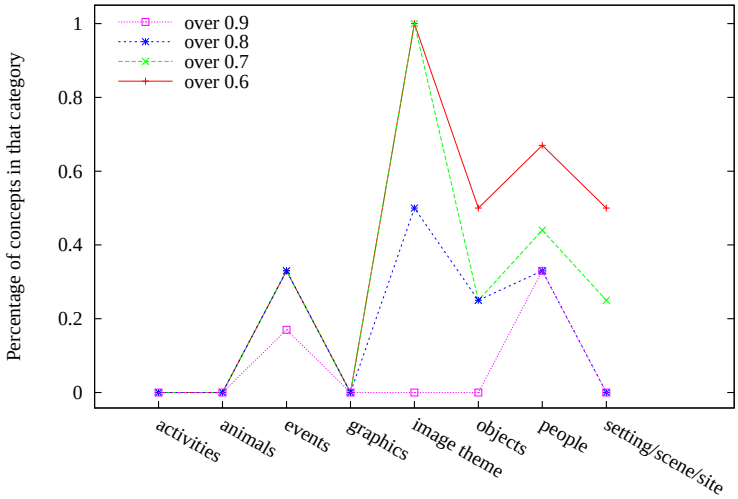

Figure 3: Agreement per category for the 25 concepts with the most clicked images in the search logs.

Given that the reached agreement values are more reliable when more images are included in the evaluation, we perform a similar analysis by considering only the concepts with the highest number of clicked images. Figure 3 shows the results when restricting our analysis to the top 20% of concepts with the highest number of clicked images in at least one method, resulting in 25 of the 111 concepts. Here, only *people* and *events* reach agreement over 0.9, with *image_theme* performing well over lower agreement thresholds. Overall, though, categories do not appear to be good predictors of concepts' performance.

4 Conclusions

Clickthrough data have the advantage of being gathered unobtrusively and in large quantities in search logs during the users' search-related interactions. Despite their sparsity, inherent noise, and the fact that logged queries (as well as tagging data) tend to describe not only the visual content, but also the context of multimedia resources, the use of large amounts of "noisily labelled" data that encode the collective knowledge of multiple past users might be the key in dealing with their quality gap to the reliable manual annotations. Our analysis shows that many concepts, particularly frequently queried ones, can be reliably annotated using clickthrough data, up to levels of 90% agreement. Unfortunately, concept categories do not provide additional evidence about the types of concepts that reach high agreement.

Acknowledgements

The authors are grateful to the Belga news agency for providing the images and search logs used in this work. This work was supported by the EU-funded VITALAS project (FP6-045389). Christos Diou is supported by the Greek State Scholarships Foundation (`http://www.iky.gr`).

References

[1] S.-F. Chang, J. He, Y.-G. Jiang, E. E. Khoury, C.-W. Ngo, A. Yanagawa, and E. Zavesky. Columbia University/VIREO-CityU/IRIT TRECVID2008 high-level feature extraction and interactive video search. In *Proceedings of TRECVID 2008*, 2008.

[2] N. Craswell and M. Szummer. Random walks on the click graph. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 239–246. ACM Press, July 2007.

[3] D. Hiemstra. A linguistically motivated probabilistic model of information retrieval. In *Proceedings of the 2nd European Conference on Research and Advanced Technology for Digital Libraries*, volume 1513 of *Lecture Notes in Computer Science*, pages 569–584. Springer, September 1998.

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

[5] D. Kelly and J. Teevan. Implicit feedback for inferring user preference: A bibliography. *SIGIR Forum*, 37(2):18–28, 2003.

[6] W.-H. Lin and A. Hauptmann. Which thousand words are worth a picture? Experiments on video retrieval using a thousand concepts. In *Proceedings of the 7th IEEE International Conference on Multimedia and Expo (ICME 2006)*, pages 41–44, 2006.

[7] M. R. Naphade, L. Kennedy, J. R. Kender, S.-F. Chang, J. R. Smith, P. Over, and A. Hauptmann. A Light Scale Concept Ontology for Multimedia Understanding for TRECVID 2005. Technical Report RC23612, IBM, 2005.

[8] A. Natsev, W. Jiang, M. Merler, J. R. Smith, J. Tešić, L. Xie, and R. Yan. IBM Research TRECVID-2008 Video Retrieval System. In *Proceedings of TRECVID 2008*, 2008.

[9] B. Poblete and R. A. Baeza-Yates. Query-sets: using implicit feedback and query patterns to organize Web documents. In *Proceedings of the 17th Conference on the World Wide Web*, pages 41–50, 2008.

[10] F. Scholer, M. Shokouhi, B. Billerbeck, and A. Turpin. Using clicks as implicit judgments: Expectations versus observations. In *Advances in Information Retrieval, Proceedings of the 30th European Conference on IR Research*, pages 28–39, 2008.

[11] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th Conference on the World Wide Web*, pages 327–336, 2008.

[12] G. Smith and H. Ashman. Evaluating implicit judgements from image search interactions. In *Proceedings of the Web Science Conference: Society On-Line (WebSci 2009)*, 2009.

[13] T. Tsikrika, C. Diou, A. P. de Vries, and A. Delopoulos. Image annotation using clickthrough data. In *Proceedings of the 8th International Conference on Content-based Image and Video Retrieval (CIVR 2009)*, 2009.

[14] A. Ulges, M. Koch, C. Schulze, and T. M. Breuel. Learning TRECVID'08 high-level features from YouTube™. In *Proceedings of TRECVID 2008*, 2008.

[15] X.-J. Wang, W.-Y. Ma, and X. Li. Exploring statistical correlations for image retrieval. *Multimedia Systems*, 11(4):340–351, 2006.

Table 2: Number of clicked images for each of the 111 concepts.

	exact	**LM**	**LM$_{stem}$**	**LMS**	**LM$_{stem}$**	**clickgraph**
# concepts per method	17	63	71	88	99	49
abused_child		20	20	20	20	12
ac_milan_soccer				69	69	
agricultural_people					22	
airplane_crash				17	17	
airport		49	49	49	49	
airport_terminal				44	44	
amusement_park						13
anderlecht	289	309	309	309	309	399
animal		10	14	10	14	
apple						18
arsenal_fc				23	25	
art						20
art_gallery						24
artist			14		14	
atomium		12	12	12	12	17
australian_open		18		18		
award			10		10	
baby		38	38	38	38	37
bank		28	28	28	28	
basketball		11	11	11	11	
beach		26	26	26	26	41
beach_leisure				22	22	
belgian_flag	13	21	27	88	97	85
belgium_royal			18	84	121	156
boat		30	30	30	30	
bridge		26	26	26	26	10
bus		21	21	21	21	32
bush	19	51	51	51	51	38
cannes_festival					12	
car		56	57	56	57	
car_accident				44	45	
car_racing				18	19	
cartoon						10
cat			10		10	
champions_league		12	12	18	20	
child	17	37	37	37	37	26
children_care				49	49	
children_playing				53	54	
church		17	17	17	17	

Continued on next page

Table 2 – continued from previous page

	exact	LM	LM_{stem}	LMS	LM_{stem}	clickgraph
classic						13
club_brugge	55	74	74	134	134	252
computer		21	33	21	33	20
court		11	11	11	11	
damaged_building		10		10	10	
davis_cup					13	
demonstration		10	10	10	10	
disco						44
dog		21	25	21	25	
earthquake	20	28	28	28	28	103
election			36		36	
european_flag				29	33	
european_parliament				42	42	
european_union				10	10	
explosion		24	24	24	24	
factory_worker					12	
fashion_model				31	31	
fashion_show			45	45	45	
federer	20	26	27	26	27	38
festival		13	13	13	13	52
finger		10	10	10	10	
fire		32	40	32	40	52
fish		10	12	10	12	
flag	25	59	71	59	71	78
flood	39	64	110	64	110	93
formula_one		15	21	21	21	47
gala		15	15	15	15	
gas_station				15	15	
girl			17		17	
goal						19
government						14
highway		10	12	10	12	
hockey	47	65	65	65	65	80
horse		25	26	25	26	74
hospital_room			63	63	63	
house		18	19	18	19	12
hurricane_typhoon				38	38	
illustration		17	17	17	17	
internet		15	15	15	15	
jacques_chirac				11	11	
king		123	123	123	123	165
lawyer		10	10	10	10	12
lion		11	11	11	11	
logo		12	12	12	12	14
memorial_services				143	143	
mountain		10	10	10	10	
nuclear		14	14	14	14	
olympic_games				46	72	30
parade	12	34	34	34	34	47
parent			11		11	
parliament		38	38	38	38	45
pope_benedict		30	30	30	30	18
queen		63	63	63	63	124
rally_motorsport		14	14	14	14	
red_devils	94	120	120	129	129	257
road		11	11	11	11	
roland_garros				11	11	
school	28	81	81	81	81	97
sex		13	13	13	13	
soccer	21	42	42	42	42	167
stock_exchange	17	45	45	74	85	44
storm			13		13	13
street_sign					35	
teenager						31
television		10	10	10	10	
tennis						78
theatre_building				17	17	
tour_of_flanders				21	21	
traffic	18	107	107	107	107	32
uefa		10	10	10	10	
usa_flag				61	73	
volleyball	28	76	77	76	77	141
mean	44.82	35.86	36.27	39.11	38.69	66.2
median	21	21	24	25.5	25	38
min	12	10	10	10	10	10
max	289	309	309	309	309	399