

Relevant Clouds: Leveraging Relevance Feedback to Build Tag Clouds for Image Search^{*}

Luis A. Leiva, Mauricio Villegas, and Roberto Paredes

ITI/DSIC, Universitat Politècnica de València
Camí de Vera, s/n – CPI edif. 8G, 46022 (Spain)
{luileito,mvillegas,rparedes}@{iti,dsic}.upv.es

Abstract. Previous work in the literature has been aimed at exploring tag clouds to improve image search and potentially increase retrieval performance. However, to date none has considered the idea of building tag clouds derived from relevance feedback. We propose a simple approach to such an idea, where the tag cloud gives more importance to the words from the relevant images than the non-relevant ones. A preliminary study with 164 queries inspected by 14 participants over a 30M dataset of automatically annotated images showed that 1) tag clouds derived this way are found to be informative: users considered roughly 20% of the presented tags to be relevant for any query at any time; and 2) the importance given to the tags correlates with user judgments: tags ranked in the first positions tended to be perceived more often as relevant to the topic that users had in mind.

Keywords: Image Search and Retrieval, Relevance Feedback, Tag Clouds

1 Introduction

It is said that a picture is worth a thousand words, though the majority of commercial image search engines require the user to issue a textual query to retrieve images. This may be problematic because formulating the right query is difficult. This is especially true for users searching for uncommon topics or when users are unsure of how to express the query. In these cases, query autocompletion techniques might not be very helpful.

One possibility to improve search experience and increase retrieval performance of image search engines consists in assisting the user by suggesting tags that relate to the issued query. In this regard, tag clouds have been shown to be a useful approach [2,4,10]. For instance, Flickr features “tag clusters”¹ as tag clouds. Then, clicking on one tag within a tag cloud provides the user with semantic zoom, so that the initial image set is refined with images that were annotated with the clicked tag.

Another option to increase the retrieval performance of image search engines is to capitalize on relevance feedback (RF) [8], i.e., presenting the user with a

^{*} Prototype available at <http://risenet.iti.upv.es/rise/tc>

¹ E.g., <http://www.flickr.com/photos/tags/sky/clusters/>

set of images according to the issued query, and letting the user select those images that are relevant and those that are non-relevant, possibly leaving some images unmarked. With this information, the retrieval engine can refine its results, leading to a hopefully better outcome after each RF iteration.

In an image search engine with RF, query suggestions can be derived by exploiting the relevance information given by the user [5]. The idea is that every time the user changes the image selection a new query is suggested, which the user can optionally follow to refine the initial search.

In this paper, we propose an alternative to the RF-based query suggestion approach. Our idea is presenting the user with a tag cloud that gets updated whenever the user selects/deselects images. This way, the tag cloud informs about the relevance of words for the images being selected; so that the most important tags would ideally be the ones that will help to retrieve more images of the kind the user has in mind. We also implemented a simple method and conducted a user study to support our idea.

2 Related Work

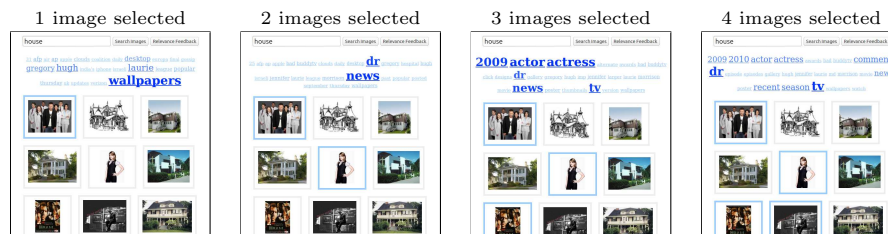
Tag clouds are seen as a “social” way to visualize information [10], and much work has been driven in this direction [4]. In the context of image retrieval, Callegari and Morreale [2] showed that less frequently used words in a tag cloud can significantly increase retrieval speed for the images associated with the tags. However, Zhang *et al.* [12] concluded that this has a mixed effect, as tags may lead the user to select irrelevant terms and introduce thus noise in the retrieval.

Typically, tag clouds are built either from keywords assigned by users [7,9] or from query logs [1,3]. This works well as long as the search engine has a very large user base and the query being searched is relatively popular among the users. Since these assumptions are not always fulfilled, other approaches to build tag clouds should be devised that are no so dependent on these factors.

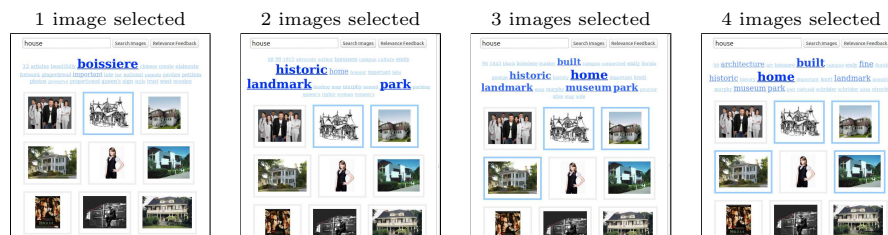
In a different vein, Liu *et al.* [6] proposed an automatic image tag ranking method based on relevance labels. While this could be exploited to build RF-based tag clouds, unfortunately it is not always feasible nor scalable to have relevance labels for all of the crawled images. Moreover, their proposed method is computationally expensive for real-time applications. In the context of this paper, tag clouds should be *reactive*, in the sense that whenever the user indicates which images are considered relevant, a new tag cloud must be generated on the fly (see Figure 1).

3 RF-based Tag Clouds

Our idea is to take advantage of the information obtained from RF to generate tag clouds. In other words, each time that the user modifies the set of relevant images, a tag cloud is updated according to this information. This behavior is illustrated in Figure 1. The goal of our proposal is twofold. First, these tag clouds provide the user with a gist about the subjects that relate to a particular set of images, giving more importance to the relevant (selected) images than the non-relevant (non-selected) ones. Second, these tag clouds give another option



(a) Query "house", subset 1: related to the TV series.



(b) Query "house", subset 2: related to buildings.

Fig. 1: The tag cloud gets updated accordingly to inform the user about the topics that relate most to the selected images *and* less to the non-selected images.

to the user beyond traditional RF. Since tag clicking is optional, the user can alternate between traditional RF iterations and clicking on a tag to refine the presented image set.

This proposal has several lines of action that need to be explored in order to fulfill the underlying purpose, which is helping the user to retrieve the desired images with less effort. First, given a selection of images, words should be selected as candidates to be shown in the tag cloud. There are several resources from which the words can be obtained, e.g., text surrounding the image from web pages, automatic image annotation, image metadata, etc. Then, the candidate words need to be preprocessed with approaches that help to filter out unwanted tags, such as removing noise, stopwords and redundant terms. Once the candidate words are identified and preprocessed, they should be ranked in such a way that the highest scores are assigned to the words that would help the user to retrieve more relevant images. Finally, once the user clicks on a tag, it must be decided how this feedback information will be used for retrieving the next set of images. One example is to use the tag as a word that expands the original text query. Another possibility would be to use the tag to re-rank all the images that were retrieved with the original query. In this work, however, these alternatives are left as an opportunity for future work.

3.1 Proposed Approach

To make our approach scalable and applicable to any image on the Web, images are automatically annotated by using the text near the image from the web pages

that contain such an image. These annotations are weighted depending on word distance to the image, term frequency, and the DOM elements.

Let $\{w_1, \dots, w_n\}$ be the words of the vocabulary, i.e., all of the different words that appear in the associated text of the N images being shown to the user. We denote the set of relevant images as \mathcal{Q}^+ and the set of non-relevant images as \mathcal{Q}^- . Let \mathcal{W} be the set of words w_i that appear in any of the relevant images. Each word $w_i \in \mathcal{W}$ is scored as follows:

$$s(w_i) = \left[\frac{\frac{1}{|\mathcal{Q}^+|} \sum_{j \in \mathcal{Q}^+} t_{ij}}{\frac{1}{|\mathcal{Q}^+|} \sum_{j \in \mathcal{Q}^+} t_{ij} + \frac{1}{|\mathcal{Q}^-|} \sum_{k \in \mathcal{Q}^-} t_{ik}} \right] + \left[\frac{\delta(|\mathcal{Q}^-|)}{|\mathcal{Q}^+|} \sum_{j \in \mathcal{Q}^+} t_{ij} \right] + E, \quad (1)$$

where $\delta()$ is the Kronecker delta function, t_{ij}, t_{ik} are the weights of the word w_i in the relevant image j and irrelevant image k , respectively, $E = |\{\forall j \in \mathcal{Q}^+ : t_{ij} \neq 0\}|$ is the number of relevant images which contain w_i , and $\sum_n t_{n\theta} = 1, \forall \theta$.

4 User Study

To date, we have not found any suitable labeled dataset to perform an automatic evaluation of RF-based tag clouds. Generally, public image datasets have relevance labels but either no associated text (e.g., ImageNet²) or a fairly limited amount of text (e.g., Web Queries³) from which to generate meaningful tags. On the other hand, Flickr has human-generated tags, but this does not extrapolate to every image on the Web. Moreover, manually labeling an image dataset to perform a rigorous evaluation of our proposal is rather difficult. The labeling would imply, for a given query and a series of image subsets for that query, to have a ground truth list of tags for each particular image subset. Therefore, with the intention of shedding light on the value of RF-based tag clouds for image search, we performed a controlled lab study. For future work, we will evaluate the retrieval performance of our proposal.

Materials: We crawled 30 million of images by querying Google, Bing, and Yahoo! using the English dictionary, and for each image the surrounding text from the web page was extracted [11]. Then, we compiled a list of 164 queries by merging the two subtasks of ImageCLEF 2012⁴.

Participants: Fourteen subjects (3 females) in their thirties ($M=31.42$, $SD=5.3$) were recruited via email advertising to participate in the study. All participants were regular users of image retrieval engines. Each participant was assigned 12 queries to evaluate.

Procedure: For each query, participants were presented with a set of the top 10 ranked images according to that query. Then, participants had to select a subset of images for different subtopics from the presented image set. For instance, for the query "hot air balloon" one could select only clipart pictures, only photos showing two or more air balloons, or only photos taken from the inside of the balloon's basket. Then, whenever a relevant image was selected from

² <http://www.image-net.org>

³ <http://lear.inrialpes.fr/~krapac/webqueries/>

⁴ <http://imageclef.org/2012/photo>

the presented image set, a list of the top 10 scored RF-based tags was displayed. A check box attached to each tag allowed participants to indicate which tags were found to be most informative and/or most adequate to the different subtopics they had in mind for each presented image set. Participants had no restrictions on subtopic selection, i.e., no minimum or maximum subtopics per query were imposed, and a subtopic could have any number of relevant images associated.

5 Results and Discussion

In total, participants assessed the relevance of 928 tag lists. They reported that sometimes the tags shown were found to be really useful and beneficial for the current query, but also sometimes they were found to be meaningless. This fact may be explained in part by the noise due to our image indexing procedure, which was completely unsupervised and therefore the tag cloud may contain irrelevant terms for a particular query. The results of this experiment are shown in Figure 2a, where the bars represent the percentage of relevant tags (according to the participants) for each rank position as assigned by Eq. (1). As expected, tags with the highest scores tended to be perceived more often as relevant. Differences between the first ranked tag and the other tags are statistically significant at the 95% confidence level.

Figure 2b depicts the proportion of relevant tags according to the number of selected images. Differences between the number of tags presented when selecting 1 or 2 images with respect to the rest of selections were found to be statistically significant. A couple of observations were derived from this experiment: 1) as more images are selected, the topic overview the tag cloud provides about such a set of images tends to be more general; and 2) the perceived quality of the tags depends highly on the particular query.

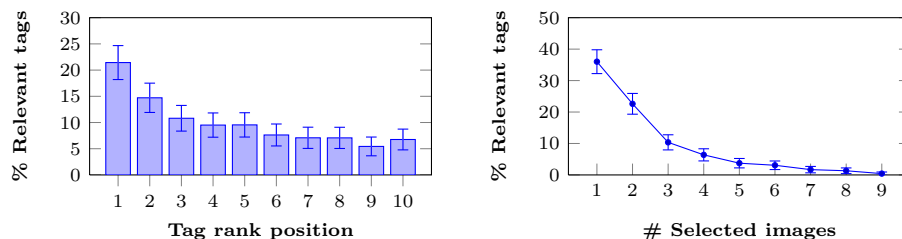


Fig. 2: Evaluation results. Error bars denote 95% confidence intervals.

As observed, when a single image is selected, nearly half of the presented tags are considered to be relevant, since they are specifically tailored to such a single selection. We find it interesting for guiding users to nail down the concept of images they are looking for. On the contrary, selecting many images may be an indicator that the initial search is actually successful, so the associated tags are likely to be seen as less relevant. As a result, when many images are

selected, a different strategy for generating RF-based tag clouds should be devised. Nonetheless, following our approach, 21.49% (SD=10) of all presented tags were considered as relevant at any time.

In general, participants liked the RF-based tag cloud idea. Some of them anecdotally commented that these tag clouds could be useful to decide which tags can lead to better retrieval results. All in all, this study indicates that our approach effectively informs the user about the relevance of the words for the images being selected. Furthermore, the tag cloud provides the user with more options to refine the image search results.

6 Conclusion

We have introduced the idea of generating RF-based tag clouds to improve image search, together with a simple approach that served as a proof of concept. The goal of these tag clouds is not only limited to providing the user with a gist about the underlying contents of the selected images. These tag clouds, in addition, give more options to the user beyond traditional RF. Then, a clicked tag can be used to disambiguate, filter, or re-rank the initial results and retrieve thus hopefully better images. We believe that this has an interesting potential and therefore deserves further research.

Acknowledgements

Work supported by EU FP7/2007-2013 under grant agreements 600707 (tranScriptorium) and 287576 (CasMaCat), and by the STraDA project (TIN2012-37475-C02-01).

References

1. G. Begelman, P. Keller, and F. Smadja. Automated tag clustering: Improving search and exploration in the tag space. In *Collaborative Web Tagging*, 2006.
2. J. Callegari and P. Morreale. Assessment of the utility of tag clouds for faster image retrieval. In *Proc. MIR*, 2010.
3. K. Ganchev, K. Hall, R. McDonald, and S. Petrov. Using search-logs to improve query tagging. In *Proc. ACL*, 2012.
4. Y. Hassan-Montero and V. Herrero-Solana. Improving tag-clouds as visual information retrieval interfaces. In *Proc. InSciT*, 2006.
5. L. A. Leiva, M. Villegas, and R. Paredes. Query refinement suggestion in multi-modal interactive image retrieval. In *Proc. ICMI*, 2011.
6. D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang. Tag ranking. In *Proc. WWW*, 2009.
7. S. Overell, B. Sigurbjörnsson, and R. van Zwol. Classifying tags using open content resources. In *Proc. WSDM*, 2009.
8. Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool for interactive content-based image retrieval. *T Circ Syst Vid*, 8(5), 1998.
9. B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *Proc. WWW*, 2008.
10. C. Trattner, Y.-l. Lin, D. Parra, Z. Yue, W. Real, and P. Brusilovsky. Evaluating tag-based information access in image collections. In *Proc. HT*, 2012.
11. M. Villegas and R. Paredes. Image-text dataset generation for image annotation and retrieval. In *Proc. CERI*, 2012.
12. C. Zhang, J. Y. Chai, and R. Jin. User term feedback in interactive text-based image retrieval. In *Proc. SIGIR*, 2005.