Tagging and Navigability

Adish Singla
Bing Search, Microsoft
13511 Commerce Parkway
Richmond, BC, Canada V6V 2J8
adishs@microsoft.com

Ingmar Weber
Yahoo! Research Barcelona
Avinguda Diagonal 177
08018 Barcelona, Spain
ingmar@yahoo-inc.com

ABSTRACT

We consider the problem of optimal tagging for navigational purposes in one's own collection. What is the best that a forgetful user can hope for in terms of ease of retrieving a labeled object? We prove that the number of tags has to increase logarithmically in the collection size to maintain a manageable result set. Using Flickr data we then show that users do indeed apply more and more tags as their collection grows and that this is not due to a global increase in tagging activity. However, as the additional terms applied are not statistically independent, users of large collections still have to deal with larger and larger result sets, even when more tags are used as search terms. We pose optimal tag suggestion for navigational purposes as an open problem.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Search process
General Terms: Human Factors, Measurement
Keywords: tagging, navigability, data organization

1. INTRODUCTION & RELATED WORK

We investigate the relation between tagging and navigability in a personal collection. In a collection of multimedia content, such as images on Flickr, users depend on the tags (or title) to find an object, as content-based image search is still difficult. One of the *main motivations* for a user to tag in this context is to make it possible for her to later find (= navigate to) this object (c.f. "self-organization" in [1]). In the following paragraphs we discuss work related to tagging and navigability. Section 2 presents mathematical models to capture the notion of "navigational power". Experimental results showing that (i) users tag more as their collection size grows and that (ii) despite more tags larger collections are harder to navigate are given in Section 3.

Note that on Flickr (i) images (and recently videos) are labeled, and so the importance for navigation is high, and (ii) exclusively the *owner* of an image tags it, so there is no "collaborative" aspect to tagging. A taxonomy of *why* people annotate their own images can be found in [1]. In their classification, we are concerned with "self-organization (adding tags for later retrieval)", which they claim to be one of the *main* motivations for tagging images. In [3] the authors analyze the evolution over time of (document, tag) pairs applied by Delicious users. Note there is no direct connection to navigability as their findings only indicate that (i) people tag more and more different documents and (ii) their choice in vocabulary remains limited. However, even rarely tagged documents might still be found easily using only their tags. The "retrievability" of a document is defined in [2] to essentially mean "how often will the document be seen in the top k results for queries", depending both on the ranking function and the query distribution. A low retrievability for a document does, however, not necessarily mean that it could not be navigated to. It rather indicates that the document's general topic is unpopular in the query distribution.

2. MATHEMATICAL MODELS

We consider a user with a collection of n objects, where n is growing. The user tags objects and can search using these tags. We analyze the situation where she later wants to navigate to a *specific* object. Let $f(t)$ be the fraction of documents labeled as t. Let $p(t)$ be the probability that the user will (correctly) remember that she applied term t. We do *not* model the case where a user incorrectly "remembers" a label, which was not actually present. We define the *navigational power* (NP) of term t as $NP(t) = p(t) \cdot (1 - f(t))$, which is the expected "zooming in" power of term t. For a set $T = \{t_1, \ldots, t_m\}$ of *independent* terms, the navigational power is defined as $NP(T) = 1 - \prod_{i=1}^{m}(1 - NP(t_i))$. What are the best labels in terms of their navigational power? The answer depends on the memory of the user.

Perfect memory model. Suppose that the user has *perfect* memory of the labels applied, i.e. $p(t) = 1$ for all t. Then terms with the lowest frequency $f(t)$ have the highest navigational power. E.g. a unique ID has a navigational power of $1 - 1/n$.

Random navigator model. Suppose that $p(t) = f(t)$. So a user is more likely to remember frequently used terms. Then a tag with $f(t) = p(t) = 0.5$ has the highest navigational power of 0.25. This model shows a sensible qualitative behavior, where neither extremely rare terms (which will be forgotten) nor extremely common terms (which have no discriminatory power) are optimal.

Limited memory model. Consider a weighted combination of the previous models. Let $p(t) = \lambda \cdot 1 + (1 - \lambda)f_t$. Here λ is a measure for the memory power of the user. For a fixed λ, $NP(t)$ will be maximized with respect to $f(t)$ for $f(t) = \max\{(1 - 2\lambda)/(2 - 2\lambda), 1/n\}$. This again shows a very plausible qualitative behavior. Namely, when the user's memory is sufficiently strong ($\lambda > 1/2$) terms with the lowest possible frequency are best. If her memory is worse, then it is preferable to use terms which are easier to remember, though they have a lower discriminatory power.

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

We want to know when $n \cdot (1 - NP(T)) = n \cdot \prod_{i=1}^{m}(1 - NP(t_i)) < k$ for a constant result set size k and growing n. If you assume that $NP(t)$ is bounded away from 1 for all t, you obtain $n \cdot (1 - NP(T)) \geq n \cdot NP_{\max}^{m}$. Hence the number of correctly remembered terms m needs to grow as $\Omega(\log(n))$ to keep the result size smaller than k. This also implies that the number of labels applied to any object needs to grow at least as $\Omega(\log(n))$, where the hidden constant is smaller for users with good and larger for users with bad memory.

3. EXPERIMENTS

We used a random subset of 110k Flickr users who uploaded an image on July 16 2008. For 80k of these users there were a total of 30M public images with at least one non-award tag (see below). Only images with at least one tag were used. Tags such as "flickrsbest" express a recognition by other Flickr users and do not deliberately serve a navigational purpose from the owner's point of view. As she is not free to add them, we removed the 15 most frequent such award-related tags. At the end, 179M tag occurrences were left for the 30M images.

To verify if, as predicted by Section 2, users add more tags to images as their collection size grows, we assigned an *index* to each image. An image has index x if at the time it was uploaded $x - 1$ images had already been uploaded. Images were bucketed into index ranges of size 50. Figure 1 shows in blue a plot of the average number of tags vs. the index of an image. We also normalized the number of tags applied to a picture by dividing by the average number of tags applied by the corresponding user (shown in red).

We used a Chi-squared test with $p = .01$ to test (i) if the distribution of tag counts for a given index is the same as without the index condition (no), (ii) if the distribution of tag counts given both the time of the upload and the index is the same as given only the index (no), and (iii) if the distribution of tag counts become the same when the time since the first picture upload is factored in (no). This supports the claim that users tag more as their collection grows even when correcting for a global increase in tagging over time and the experience of the user in the system.

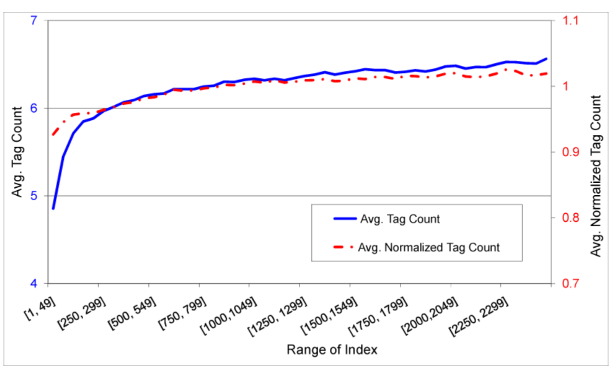

Figure 1: This graph shows an increase in the number of tags applied to an image vs. its index, both for the *absolute* and the *normalized* number of tags applied to an image, averaged across all images in the corresponding bucket.

But does tagging more help with navigability? To quantify the navigability of a given collection we did the following for each of the 80k users. First, for each of a user's images we supposed that the user wants to navigate to this particular image. Then, from the chosen image we iteratively sampled the tags present, again using different sampling strategies. For the set of i tags chosen up to a particular point, we then looked at the percentage of images in the user's collection, which satisfy the corresponding query. The percentage of matching images is a decreasing function in i. The percentage will decrease slowly if the tags present often co-occur, and it will decrease faster if the tags are independent. Obviously, a smaller percentage is better for navigating. To model that "popular" images might be searched for more often, we also weighted the percentage by $1 + \{\text{view count}\}$. As this gave similar results, we only report numbers for the setting with an equal weight for all images. As for order of the tags, we tried sampling uniformly at random (in blue in Figure 2) and sampling according to frequency in the user's collection (in red). We also experimented with a greedy algorithm (in green). Here, at each step the term that leaves the smallest result set is added next.

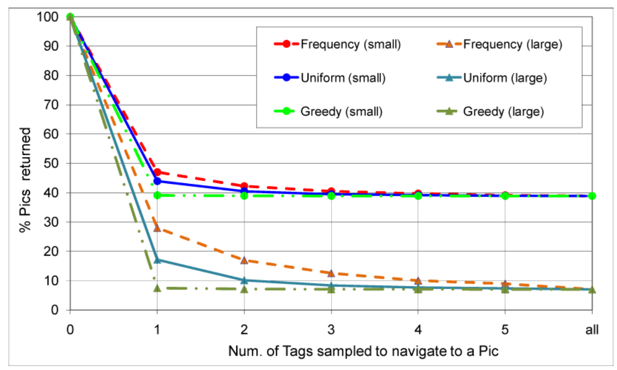

Figure 2: This plot shows the fraction of images remaining for small ($[1, 50]$ images) and large ($[251, \infty$ images) collections as more and more search terms are added.

Whereas the first query term typically already removes 72% of possible images for large collections in the "random navigator" model (triangle, red), the second terms only removes additional 11% of the collection and the third a mere 5%. The reason for this effect is a strong, positive correlation between terms.

Even though search tags for users with $[251, \infty)$ images lead to a smaller fraction of images being returned, they are still left with over 50 results after entering 5 terms. Especially users with a large collection should therefore enter *independent* tags. As current tag suggestion methods [4] reinforce the use of *correlated* tags, algorithms with the objective to optimize navigability should be explored.

4. REFERENCES

[1] Ames and Naaman. Why we tag. In *CHI'07*, pages 971–980, 2007.

[2] Azzopardi and Vinay. Retrievability: An evaluation measure for higher order information access tasks. In *CIKM'08*, pages 561–570, 2008.

[3] Chi and Mytkowicz. Understanding the efficiency of social tagging systems using information theory. In *HYPERTEXT'08*, pages 81–88, 2008.

[4] Garg and Weber. Personalized, interactive tag recommendation for flickr. In *RecSys'08*, pages 67–74, 2008.