

Diversifying Image Search with User Generated Content

Roelof van Zwol, Vanessa Murdock, Lluís Garcia Pueyo, and Georgina Ramirez
Yahoo! Research, Barcelona, Spain

{roelof, vmurdock, lluis, ramirezg}@yahoo-inc.com

ABSTRACT

Large-scale image retrieval on the Web relies on the availability of short snippets of text associated with the image. This user-generated content is a primary source of information about the content and context of an image. While traditional information retrieval models focus on finding the most relevant document without consideration for diversity, image search requires results that are both diverse and relevant. This is problematic for images because they are represented very sparsely by text, and as with all user-generated content the text for a given image can be extremely noisy.

The contribution of this paper is twofold. First, we present a retrieval model which provides diverse results as a property of the model itself, rather than in a post-retrieval step. Relevance models offer a unified framework to afford the greatest diversity without harming precision. Second, we show that it is possible to minimize the trade-off between precision and diversity, and estimating the query model from the distribution of tags favors the dominant sense of a query. Relevance models operating only on tags offers the highest level of diversity with no significant decrease in precision.

Categories and Subject Descriptors

H.3 [INFORMATION STORAGE AND RETRIEVAL]:
Content Analysis and Indexing

General Terms

Algorithms, Experimentation, Performance

Keywords

pseudo-relevance feedback, diversity, image retrieval, Flickr, retrieval performance, ambiguity

1. INTRODUCTION

Millions of images are uploaded every day to photo shar-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MIR'08, October 30–31, 2008, Vancouver, British Columbia, Canada.
Copyright 2008 ACM 978-1-60558-312-9/08/10 ...\$5.00.

ing services like Flickr¹ and Picasa². These services allow users to upload, annotate and share their photos with family, friends and the general public. The photo annotations provided by the user are also referred to as user-generated content. Typically, a user can define a title, description and a set of tags for each of their photos. The annotations provided are essential to making the photos retrievable by the text-based retrieval models, and allow users to formulate keyword-based queries against the photo collection.

Due to the rich nature of the image content, and the limited expressiveness of keyword-based query formulation it is often difficult for a user to precisely formulate his information need. To address this shortcoming, we propose that in the absence of disambiguating information the user should be presented with a diverse set of images that embodies many possible interpretations of the user's query. Hopefully, when presented with results reflecting multiple senses of the query, the user's intention will be represented. As an alternative, the system can include additional steps that address the visual context [1, 11].

When focusing on text-based image retrieval, the diversity of a result set is related to the ambiguity of the query. For instance, is a user searching with the query *apple* interested in the *fruit*, or the *company*? We refer to this type of ambiguity as word-sense ambiguity. When refining the query to *apple company*, a different type of ambiguity occurs. Ideally the search results are still diverse and contain examples of the different apple products, logos, etc. We refer to this as type-specific ambiguity. There are other types of diversity such as visual diversity that are not captured in the textual metadata associated with an image, and are beyond the scope of this paper.

Attempts have been made to incorporate a notion of diversity in textual search, most notably in the TREC Novelty Track [5, 18, 17]. Images offer the additional challenge that the text representing them is extremely sparse and does not always reflect the image content [16]. In our data, an image represented by a description, a title, and a set of tags, has associated with it 32 terms on average - about the length of 1.5 sentences, which is considerably shorter than typical newswire articles. Although there is not an explicit length assumption in a given retrieval model, the models have been designed and benchmarked against the TREC collection, where documents are considerably longer. It has been shown that reducing the length of the document negatively affects retrieval performance [13].

¹Flickr: <http://www.flickr.com/>

²Picasa: <http://www.picasa.com/>

In any reasonable retrieval setting, we would require a model that does not harm precision for unambiguous queries while respecting the topical diversity of ambiguous queries. Relevance models [9], which estimate a model of the query from the distribution of relevant documents in the collection, effectively add terms to the query that are related to the relevant documents. For a query that is topically unambiguous, the effect is to encourage relevant documents to be ranked higher. For topically ambiguous queries, we propose that terms related to multiple senses of the query are included in the query model, and therefore the resulting documents will be more topically diverse.

The tag set associated with an image resembles a query in character, in that a tag set contains content terms not written in natural language. Tag sets are more succinct, and people often tag photos with sets of synonyms and related terms [16, 12]. Although many believe tags to be noisy [3], we claim that they are more effective for encouraging diverse results. In this paper we investigate incorporating tags in the retrieval model, and show how this affects both the relevance and diversity of the result set.

The contributions of this paper are as follows. First we compare the retrieval performance of several different retrieval models in terms of precision on a set of unambiguous topics, and a set of ambiguous topics. We choose two ways to represent the images: with all meta-data, or with just the tag-sets. We investigate the diversity of the search results by measuring the number of different senses present in the ranking at various cutoff points and by analysing the distribution of topic senses in the ranked list.

The notion of diversity in image search has been proposed most often as a post-process to the retrieval model, using query expansion or results re-ranking. Our approach integrates the diversity of the search results directly into the retrieval strategy.

The remainder of the paper is organized as follows. In Section 2 we put our work in the context of previous work. In Section 3 we discuss the proposed retrieval models. The results of retrieval performance on unambiguous topics is presented in Section 4. In Section 5 we present the retrieval results for the ambiguous topics, and we evaluate the diversity of the image search results. Section 6 offers a discussion of the results. Finally we conclude with directions for future work in Section 7.

2. RELATED WORK

Image retrieval has been studied for many years, and the contributions to the field are significant [7, 11, 1]. In the context of this paper, we focus the related work on diversity in search results for image retrieval, and the use of pseudo-relevance feedback models for text and image retrieval.

2.1 Related Work on Diversity in Search

The following three approaches achieve diversity in a post-retrieval process, independent of the retrieval algorithm. This contrasts with our approach in which it is the retrieval model itself which provides the diversity. In Zhang et al. [21] diversity of search results is examined in the context of Web search. They propose a novel ranking scheme named Affinity Ranking to re-rank search results by optimizing two metrics: diversity and information richness. More recently, Song et al. [19] also acknowledge the need for diversity in search results for image retrieval. They propose a re-ranking method

based on topic richness analysis to enrich topic coverage in retrieval results, while maintaining acceptable retrieval performance.

Zeigler studied topic diversification to balance and diversify personalized recommendation lists in order to reflect the user’s complete spectrum of interests [22]. Although their system is detrimental to average accuracy, they show that the method improves user satisfaction with recommendation lists, in particular for lists generated using the common item-based collaborative filtering algorithm. They introduced an intra-list similarity metric to assess the topical diversity of recommendation lists and the topic diversification approach for decreasing the intra-list similarity.

In a different setting, Yahia et al. [20] propose a method to return a set of answers that represent diverse results proportional to their frequency in the collection. Their algorithm operates on structured data, with explicitly defined relations, which differs from our setting, as user-generated content in Flickr has a minimum amount of structure associated with it.

The TREC Novelty Track [5, 18, 17] aimed to encourage research in finding novel sentences in a set of relevant sentences. The task resembles the current work in that the results must be relevant as well as novel, and the systems were acting upon sentence-length data. One difference in this setup, however, was that the data was entirely unstructured, and the sentences appeared in the context of a document. The document context allows for a much richer term distribution, and smoothing from the document improves retrieval results [14].

2.2 Related Work on Implicit Relevance Feedback Models

Our work uses relevance models, which were first proposed by Lavrenko and Croft [9]. Relevance models have been used for a number of different applications, but most relevant to the current work is the extension of relevance models to image annotation [6, 10]. In this work, the task is to automatically assign textual annotations to images, based on the visual content of the image. The data used in their work, the Corel dataset, differs from our data in that it is a very small data set with human-edited annotations. Lavrenko and Croft essentially treat the language of images and the language of annotations as separate languages, and apply a cross-lingual relevance model [8].

Diaz and Metzler [4] estimate the query model as a mixture of models from a set of large external collections. Their application was document retrieval using topics from TREC, where the queries are longer than our queries, and represent multiple concepts. They discover that the system performs better when the concept density in a given external collection is higher for a particular concept represented in the query.

3. ESTIMATING QUERY MODELS FROM IMAGE TAGS

We can imagine when annotating an image, a user selects a few terms from some distribution of terms that could be used to represent the image. The user may also write a short natural language description of the image, and add a title. Although we have these three pieces of evidence about the image, they are a sparse representation. In our data,

for example, images have an average of 32 terms associated with them, including title, description, and tags. There are on average ten tags associated with each image, and the titles are an average of three terms long.

In a similar process, a user querying for an image has in mind some image he would like to see. The user creates a query by sampling from the distribution of terms that might be associated with the hypothetical image. Typical search queries are two to three terms long, and resemble image tags in the sense that they consist mostly of content terms representing concepts embodied in the image, and are less frequently represented by natural language.

This generative process is embodied in the query likelihood retrieval model [15]. In the discussion that follows we call the metadata associated with an image the “document”. Query likelihood estimates the probability that a document was generated from the same distribution as the query:

$$\begin{aligned} P(D|Q) &\approx P(D)P(Q|\theta_D) \\ &= P(D) \prod_i^k P(q_i|\theta_D) \end{aligned} \quad (1)$$

where Q is a query of length k , composed of terms q_i , D is a document, and θ_D is a unigram language model of the document. In practice $P(D)$ is assumed to be uniform, and $P(q_i|\theta_D)$ is estimated by the frequency of q_i in θ_D . We smoothed the model with Jelinek-Mercer smoothing, which redistributes some of the probability mass to unseen events, by estimating the probability of the query term given the model of the collection, θ_C :

$$P(q_i|\theta_D) = \lambda \frac{tf(q_i, D)}{\sum_v tf(v, D)} + (1 - \lambda)P(q_i|\theta_C) \quad (2)$$

In our data the documents are so sparse that the query terms may not be represented even if the image is relevant. For this reason we turn to relevance models [9], which estimate a model of the query by sampling from the distribution of terms that generated the original query. Images are ranked by the KL-divergence between the query model and the model of the image. Typically relevance models [9] estimate the query model from the same collection that is ranked to produce the final result, but there is no reason this need be so.

In our data, the image tags resemble the query in the sense that the tags are a set of content terms representing the image, and typically do not contain natural language. The relevance model chooses a distribution of tags, θ_T , from which to sample, from all possible tag distributions Θ . It then samples a term w from the distribution according to:

$$P(w, q_1, \dots, q_k) = \sum_{\theta_T \in \Theta} P(\theta_T)P(w, q_1, \dots, q_k|\theta_T) \quad (3)$$

Once we have fixed the sampling distribution, the term w is conditionally independent of the query terms, q_1, \dots, q_k , and their joint distribution can be estimated as a product of the marginals:

$$P(w, q_1, \dots, q_k|\theta_T) = P(w|\theta_T) \prod_i P(q_i|\theta_T) \quad (4)$$

Combining the two equations, we repeatedly sample by

choosing a distribution with probability $P(\theta_T)$, and sample a term from the given distribution:

$$P(w, q_1, \dots, q_k) = \sum_{\theta_T \in \Theta} P(\theta_T)P(w|\theta_T) \prod_i P(q_i|\theta_T) \quad (5)$$

In our case, $P(\theta_T)$ is estimated using unigram distribution priors, and the probability of a term, given a model of the tag distribution, is its term frequency in the distribution, smoothed with its term frequency in a general model of tags (in our case, estimated from the collection):

$$P(w|\theta_T) = \lambda \frac{tf(w, T)}{\sum_v tf(v, T)} + (1 - \lambda)P(w|\theta_{CT}) \quad (6)$$

In theory, a query model would be estimated over the entire vocabulary, sampling from all possible models of tag distributions. In practice, we empirically set the number of terms and the number of models as parameters.

4. RETRIEVAL PERFORMANCE

In this section we make three retrieval comparisons on a set of unambiguous topics. First we compare the retrieval performance of relevance models to query likelihood. Next, we compare retrieval from all fields, including tags, descriptions and titles, to retrieval only from the tags. Finally, we compare the estimation of the query model from the distribution of tags to the standard relevance model where the query model is estimated from the same collection that is ranked.

4.1 Experimental Setup

Flickr is an online photo-sharing service that contains hundreds of millions of photos that are uploaded, tagged, and organised by more than 9 million users. Starting with a random subset of queries to Flickr, we eliminated queries with adult content, and queries that were unjudgable (such as queries for person names), and ambiguous topics. Our final topic set consists of 95 queries in English and in Spanish, of an average of one to two terms. We pooled the top 50 results from 5 retrieval approaches, including their parameter settings. We judged a total of 51,000 images for relevance. The images, title and tags were presented in random order, in pages of six. Each image was judged relevant or nonrelevant by one assessor. We selected 20 percent of our topics randomly to be assessed by a second assessor to measure inter-assessor agreement. The inter-assessor agreement was more than 85% for every topic, and for most topics, above 90%.

The collection consisted of a subset of 8.5 million photos that had annotations for each of the three fields: title, description, and tags. The three fields were concatenated and indexed as a single “document” for the purpose of ranking. For experiments with relevance models where the query model was estimated from the tags, we indexed the tags field separately. The data was not stemmed and stopwords were not removed. We used the Lemur Toolkit [2] for both indexing and retrieval.

We evaluated five systems:

- **Query likelihood** where the unit of retrieval is all metadata associated with an image.
- **Query likelihood** where the unit of retrieval is the tag set.

- **Relevance models** where the unit of retrieval is all metadata associated with an image.
- **Relevance models** where the unit of retrieval is the tag set.
- **Dual-index relevance model** where the query model is estimated from the tag set, as described in Section 3, but the unit of retrieval is all metadata associated with an image.

Images are viewed in blocks rather than as a ranked list so users see more images than they might see documents in a traditional retrieval setting. Since users see a number of thumbnails simultaneously, they can pick the relevant subset by scanning the page, rather than scrolling through a list. For this reason, we are most interested in optimizing the number of relevant images appearing in a block of images. We report results for precision at k .

Results reported are the best performing runs for all parameter settings. Query likelihood and relevance models have a smoothing parameter, λ , which was set to 0.3 for all experiments. The parameter settings for the relevance model included the number of documents to compute the query model, and the number of terms to sample. Both were optimized for precision. For all runs, the optimal number of documents was ten. For the standard relevance model, five feedback terms was optimal. For both relevance models creating query models from tags, the optimal number of terms was ten. In addition, a feedback coefficient can be set to adjust the weight given to the original query terms in the expanded query. In our model we found that the best performance was achieved by allowing the model to choose the terms in the query, without enforcing that the original query terms should be included.

4.2 Retrieval results

The first two rows of Table 1 compare query likelihood ranking on all metadata to query likelihood ranking on tags only. The precision at rank one is slightly higher for ranking on tags, but the results overall are comparable. The same can be said of the standard relevance model (the third row) and the relevance model ranking based on tags (the fourth row). The fifth row (indicated with “Dual Index”) shows the result of estimating the query model from the tags, but ranking on all of the metadata. Although the precision at rank one is lower than both retrieval approaches ranking on tags, this approach consistently outperforms the other approaches farther down the ranked list.

We require a retrieval model that encourages diversity for ambiguous queries, but does not harm precision for any query. We can see from Table 1 that none of the results is significantly different in terms of precision.

5. DIVERSITY IN SEARCH RESULTS

In this section, we present the setup and results of the experiments to measure the diversity in the search results using a set of ambiguous topics. The setup of the experiment differs slightly from that of a retrieval performance experiment and is discussed in more detail below. We present the results in three steps. First we report the overall performance of the models on the ambiguous topics in terms of precision at k , and compare the outcome with the results reported in Section 4.2. Second, we report the proportion

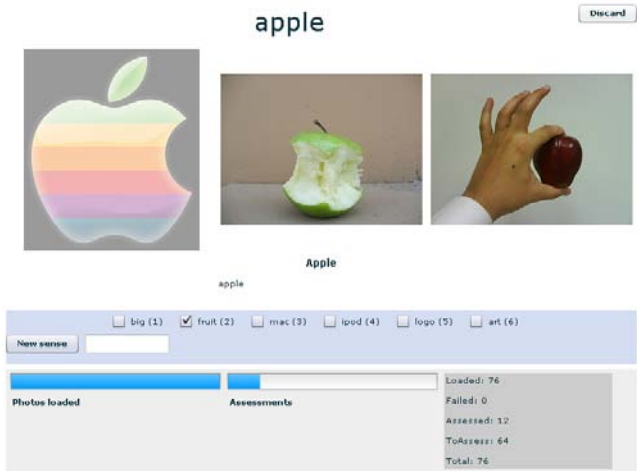


Figure 1: Screenshot of the assessment interface for diversity in search results.

of senses present in the ranking at various cutoff points. Finally, we investigate the distribution of senses represented in the results list.

5.1 Experimental Setup

For the diversity experiment we used the same collection of 8.5 million images from Flickr, and the same set of models described in Section 4.1. We chose 25 topics that are inherently ambiguous, shown in the first column of Table 4. Using a blind review pooling method we pooled the top 50 results of the 5 different systems. The assessors were asked to judge the relevance of each image. If an image was judged relevant, the assessor also had to identify a suitable sense which would allow for disambiguation of the topic. Table 4 shows the ambiguous topics, with their primary, secondary, and other senses, as judged by the assessors. Each topic was judged by one assessor, and the inter-assessor agreement was not computed. As mentioned in the introduction, we consider two types of ambiguity: word-sense ambiguity and type-specific ambiguity. As an example, the topic *apple* exhibits word-sense ambiguity, as assessors identified a primary and secondary sense of *apple* (*logo* and *fruit*). The assessors further identified type-specific senses of *apple* (*art*, *mac*, *ipod*). A screenshot of the assessment interface that was used for the experiment is shown in Figure 1.

5.2 Retrieval Performance

Though we are aiming for a high degree of diversity in the search results, we cannot accept a significant decrease in precision. The results for the retrieval performance with the ambiguous topics are shown in Table 2.

We did not find significant differences in precision between the different systems at the reported positions in the ranking, however we notice that the tags-only relevance model and the dual-index relevance model performs slightly better at the top of the ranking (<10), after which the standard relevance model has the best performance. When comparing the results for the unambiguous topics (Table 1) with the results for the ambiguous topics (Table 2) we find that the performance is comparable, i.e. no significant differences are found between the two sets of topics, in terms of precision.

Model	P@1	P@5	P@10	P@15	P@20	P@25	P@50
Query Likelihood	0.747	0.733	0.733	0.719	0.709	0.701	0.667
Query Likelihood (Tags Only)	0.779	0.749	0.720	0.712	0.703	0.700	0.673
Relevance Model	0.758	0.743	0.720	0.708	0.706	0.699	0.677
Relevance Model (Tags Only)	0.779	0.726	0.717	0.719	0.714	0.710	0.683
Relevance Model (Dual Index)	0.768	0.754	0.739	0.726	0.719	0.716	0.680

Table 1: Comparing results of the query likelihood baseline to the various relevance models. Boldface indicates the best result.

Model	P@1	P@5	P@10	P@15	P@20	P@25	P@50
Query Likelihood	0.680	0.760	0.720	0.725	0.734	0.744	0.734
Query Likelihood (Tags Only)	0.800	0.736	0.732	0.720	0.736	0.736	0.734
Relevance Model	0.720	0.760	0.768	0.784	0.788	0.792	0.778
Relevance Model (Tags Only)	0.840	0.728	0.744	0.741	0.756	0.752	0.735
Relevance Model (Dual Index)	0.720	0.776	0.768	0.755	0.754	0.760	0.763

Table 2: Retrieval performance results on the set of ambiguous topics. Boldface indicates the best result.

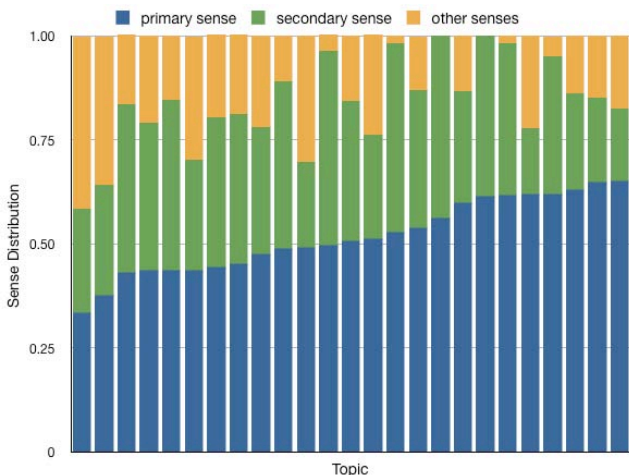


Figure 2: Frequency distribution of senses over the pool of assessed topics, sorted by increased frequency of the primary sense.

5.3 Sense Distribution

Figure 2 plots the frequency distribution of the different senses identified by the assessor over the entire pool of assessed topics. The topics (x-axis) are sorted by increased frequency of the primary sense. Each of the topics has at least two senses. The remaining senses are grouped in “other” senses. The figure shows that for most topics, two prominent senses were always detected, and most topics had at least one other sense represented. The frequency distribution of the primary sense ranges from 0.33 to 0.65 and the distribution of the secondary sense ranges from 0.16 to 0.47.

A first notion of diversity in the search results is obtained by computing the number of distinct senses retrieved by the model divided by the total number of senses identified by the assessors over the pool, averaged over all topics. Table 3 shows the proportion of the distinct senses represented in the top k of the ranking ($S@k$) for each of the retrieval models. It clearly shows that the relevance model has the highest proportion of senses detected at each point in the ranking. The dual-index relevance model, on the other hand,

has the lowest number of senses represented in the ranking. This indicates that the dual-index relevance model has more focused results, which is consistent with its performance on the unambiguous topics.

5.4 Diversity vs. Precision

The results of the previous section provide insight in the proportion of total senses represented in the ranking, however it does not give us an indication of the distribution of the different senses, nor does it allow us to analyse the trade off between precision and diversity. We therefore compute the precision at k for each sense per topic per system, and report the results for the primary sense, secondary sense, and other senses. Figures 3 (a) - (e) plot the precision at each point in the ranking for each of the five systems. Besides the overall precision (red line), the precision curves for the primary sense (blue line), the secondary sense (green line) and the other senses (yellow line) are shown.

The overall performance for the tags-only query likelihood model and the tags-only relevance model show a clear boost at the top of the ranking, which can best be explained by an increased performance for the primary sense. But more importantly, we find that for these two models the distribution of the primary and secondary senses is very balanced, as both curves are almost overlapping, which is not the case for the other systems. The dual-index relevance model exemplifies this point, and shows a clear preference for returning results that reflect the primary sense.

6. OBSERVATIONS AND DISCUSSION

When discussing text-based image retrieval there are several points to consider. First the textual representation of both the query and the image is sparse. Second, the text representing the image is itself imperfect. User-generated content is well-known for being noisy, and suffers from problems such as bulk uploading, where hundreds of photos are tagged with the same tags, regardless of their content. Any retrieval model designed for user-generated content must cope with sparse, noisy data [16].

Relevance models are well-suited to this task, as they enrich the query model with terms from the distribution of relevant documents. The frequent terms in the query model will have more weight in the distribution, and thus docu-

Models	S@5	S@10	S@15	S@20	S@25
Query Likelihood	0.45	0.59	0.70	0.77	0.83
Query Likelihood (Tags Only)	0.45	0.56	0.66	0.73	0.78
Relevance Model	0.48	0.61	0.72	0.80	0.87
Relevance Model (Tags Only)	0.43	0.56	0.69	0.74	0.82
Relevance Model (Dual Index)	0.48	0.55	0.68	0.73	0.80

Table 3: Proportion of total senses represented in the top X, averaged over all topics

ments containing those terms will be ranked higher. Expanding from the distribution of image tags showed a small improvement in the retrieval results for unambiguous topics. For ambiguous topics, the relevance model produced higher precision for the primary sense of the topic, leading to less diversity in the results set. This is evident in Figure 3 (c) where we see that the primary sense has a clearly higher precision than the other senses. We suppose that this is because the initial retrieval results are dominated by the primary sense. Once the query model is estimated, more terms will be sampled that are related to the primary sense. A similar effect is achieved by the dual-index relevance model (Figure 3 (e)).

A different pattern is exhibited by the two models that rank the results based on the tags. In Figure 3 (b) and (d) we see that the retrieval results are more or less equivalent for all senses of the topics. Therefore we can conclude that in the case of diverse topics, the tag-only systems generate a balanced distribution of the primary and secondary sense in the search results. A slight increase in precision is found for the tags-only relevance model, at the cost of computational complexity when compared to the tags-only query likelihood model.

We see from Table 3 that the relevance model represents the greater number of senses, irrespective of their distribution, than query likelihood. This would suggest that the standard relevance model is the best combination of performance in terms of precision, and both measures of diversity.

7. CONCLUSIONS AND FUTURE WORK

The user-generated content associated with an image is a primary source of information about the content and context of an image, and allows users to formulate keyword-based queries against the photo collection. However, due to the limited expressiveness of keyword-based query formulation it is often difficult for a user to precisely formulate his information need. We therefore argue that the user should be presented with a diverse set of images that embodies many possible interpretations of the user’s query. Unlike previous approaches, which encourage diversity as a post-retrieval step, our approach shows that the proper retrieval model itself can encourage diverse results.

We turn to relevance models, which estimates a model of the query by sampling from the distribution of terms that generated the original query. We introduce a dual-index approach, where additional query terms are first sampled from the tag index, and in a second step we retrieve relevant images from full index. This proves especially useful to increase the retrieval performance for unambiguous topics, as it samples terms related to the primary sense in the query model. Alternatively, we find tags-only relevance models to be most useful in the case topics with an ambiguous nature. The evaluation results show the that tags-only systems generate

a balanced distribution of the primary and secondary senses in the search results, with a marginal decrease in retrieval performance.

We deliberately have chosen to limit our research on diversity of search results to text-based retrieval models. We do this primarily because we believe that image retrieval on the Web is naturally initiated by a keyword-based search, after which the user can interactively refine his information need using a combination of text-based and content-based approaches. Furthermore, the ImageCLEF 2008³ initiative also acknowledges the need for diversity in image search results, and dedicated a specific task on this topic. The initiative will certainly encourage further research in this area. For our future work we would like to extend the proposed evaluation framework for diversity in image search by addressing multiple dimensions. In addition, we will expand our research focus by studying diversity in image search results based on a multi-modal approach.

8. REFERENCES

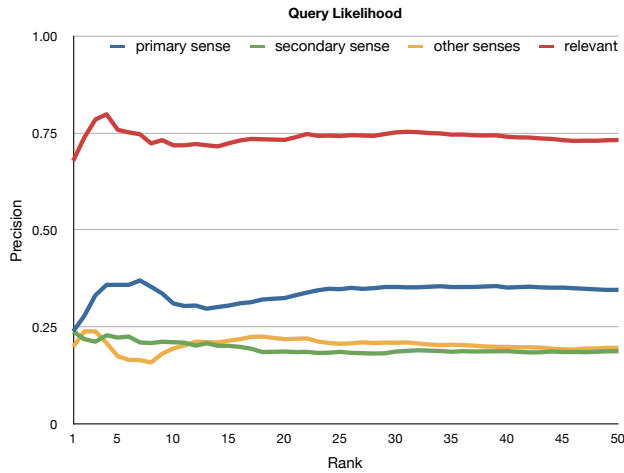
- [1] *Exploratory image databases: content-based retrieval*. Academic Press, Inc., Duluth, MN, USA, 2001.
- [2] J. Allan, J. Callan, K. Collins-Thompson, W. B. Croft, F. Feng, D. Fisher, J. Lafferty, L. Larkey, T. N. Truong, P. Ogilvie, L. Si, T. Strohman, H. Turtle, and C. Zhai. The lemur toolkit for language modeling and information retrieval, 2005. <http://www.cs.cmu.edu/lemur>.
- [3] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):1–60, 2008.
- [4] F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. In *Proceedings of SIGIR*, 2006.
- [5] D. Harman. Overview of the TREC 2002 novelty track. In *Proceedings of the Eleventh Text Retrieval Conference (TREC)*, 2002.
- [6] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th Annual Conference on Research and Development in Information Retrieval (ACM SIGIR)*, 2003.
- [7] M. L. Kherfi, D. Ziou, and A. Bernardi. Image retrieval from the world wide web: Issues, techniques, and systems. *ACM Comput. Surv.*, 36(1):35–67, 2004.
- [8] V. Lavrenko, M. Choquette, and W. B. Croft. Cross-lingual relevance models. In *Proceedings of the 25th Annual Conference on Research and Development in Information Retrieval (ACM SIGIR)*, 2002.

³ImageCLEF 2008 photo retrieval task: <http://www.imageclef.org/2008/photo>

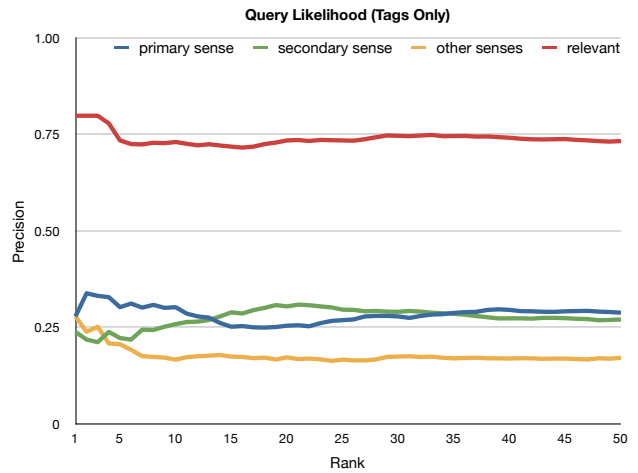
Topic	Primary	Secondary	Others
apple	logo	fruit	art, mac, ipod
beatle	vw	bug	band
belly	dance	pregnant	human, animal, other, piercing/tattoo
canals	venice	amsterdam	england, denmark, brazil, delaware, singapore, thailand, other
carving	wood	ice	stone, food, pumpkin, skiing
castro	cuba	chile	san francisco, algarve, movie, galicia, theater, festival
chihuahua	dog	mexico	
chopper	motorbike	helicopter	bicycle, art, harley
corps	drum	army	peace, soft, art, fife, animal
flames	fire	hotrod	art, flower, tattoo, light
jaguar	car	animal	python
kingdom	disney	united	
kingston	ontario	river	jamaica, other, new york, london, washington
leon	city	person name	flower, animal
mendoza	argentina	river	wine
pet	dog	cat	mouse, bird, reptile, insect, turtle
pillow	fight	couch	bed, other
playoffs	basketball	football	hockey, baseball
riviera	mexico	las vegas	car, hotel, game, france
ruins	roman	mexico	US, greece, india, peru
sandals	shoes	jamaica	chokbay, art
sharks	fish	hockey	other sports team, iconic image, pool shark
sports	soccer	action	car, athletes, basketball, baseball, Volleyball, hockey, water, football
springs	towns	hot	wire
tivoli	copenhagen	italy	theatre, music, okayama

Table 4: Ambiguous topics, with their primary, secondary, and “other” senses.

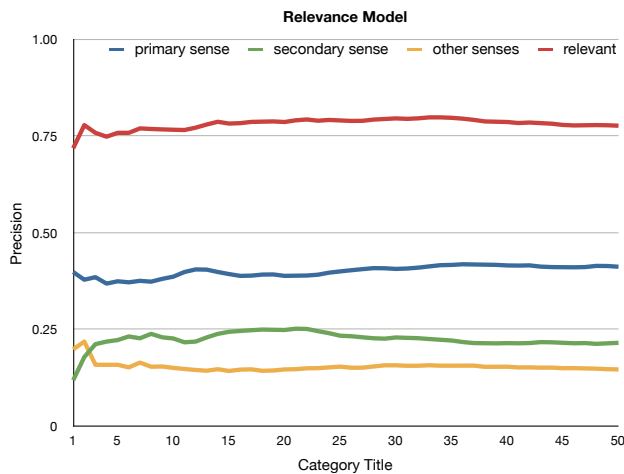
- [9] V. Lavrenko and W. B. Croft. Relevance-based language models. In *Proceedings of the 24th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2001.
- [10] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *Proceedings of the 17th Annual Conference on Neural Information Processing Systems*, 2003.
- [11] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(1):1–19, 2006.
- [12] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPertext '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40, New York, NY, USA, 2006. ACM Press.
- [13] V. Murdock. *Aspects of Sentence Retrieval*. PhD thesis, University of Massachusetts, 2006.
- [14] V. Murdock and W. B. Croft. A translation model for sentence retrieval. In *Proceedings of the Conference on Human Language Technologies and Empirical Methods in Natural Language Processing (HLT/EMNLP)*, 2005.
- [15] J. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual Conference on Research and Development in Information Retrieval (ACM SIGIR)*, 1998.
- [16] B. Sigurbjornsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th International World Wide Web Conference (WWW2008)*, Beijing, China, April 2008.
- [17] I. Soboroff. Overview of the TREC 2004 novelty track. In *Proceedings of the Thirteenth Text Retrieval Conference (TREC)*, 2004.
- [18] I. Soboroff and D. Harman. Overview of the TREC 2003 novelty track. In *Proceedings of the Twelfth Text Retrieval Conference (TREC)*, 2003.
- [19] K. Song, Y. Tian, W. Gao, and T. Huang. Diversifying the image retrieval results. In *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*, pages 707–710, New York, NY, USA, 2006. ACM.
- [20] S. A. Yahia, P. Bhat, J. Shanmugasundaram, U. Srivastava, and E. Vee. Efficient online computation of diverse query results. In *Proceedings of VLDB*, 2007.
- [21] B. Zhang, H. Li, Y. Liu, L. Ji, W. Xi, W. Fan, Z. Chen, and W.-Y. Ma. Improving web search results using affinity graph. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 504–511, New York, NY, USA, 2005. ACM.
- [22] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 22–32, New York, NY, USA, 2005. ACM.



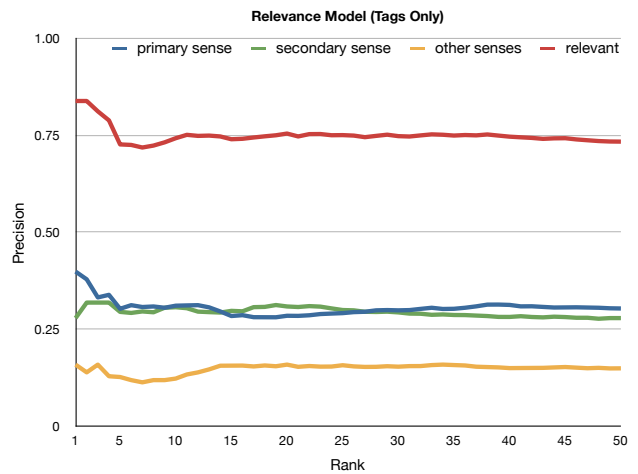
(a)



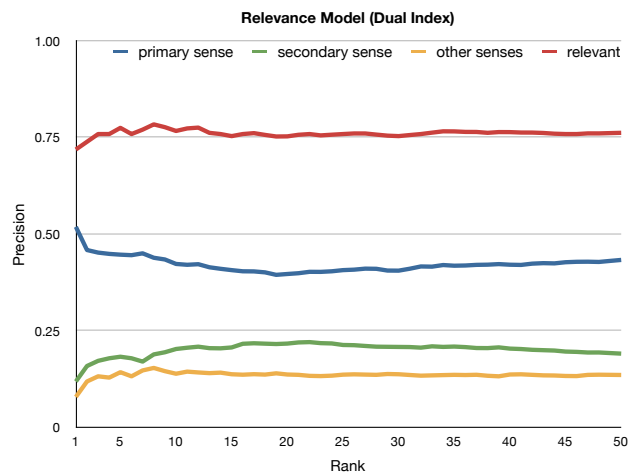
(b)



(c)



(d)



(e)

Figure 3: Precision at rank X for the each of the models compared to the primary, secondary and other senses.