Searching Consumer Image Collections Using Web-based Concept Expansion

Mark D. Wood
Eastman Kodak Company
Rochester, NY, USA
mark.d.wood@kodak.com

Alexander Loui
Eastman Kodak Company
Rochester, NY, USA
alexander.loui@kodak.com

Stacie Hibino
Eastman Kodak Company
San Jose, CA, USA
stacie.hibino@kodak.com

ABSTRACT

As consumers accumulate more and more personal imagery, searching for specific images has become increasingly difficult. Consumers typically provide little or no annotations, and automated classifiers and concept tagging tools are limited in their scope and vocabulary. This work addresses this sparsity of semantic information by leveraging domain-specific information provided by online photo-sharing communities. Such information enables improved search by allowing user-provided search terms to be expanded into a set of semantically related concepts, using relevant semantic relationships provided by millions of users. Our system first extracts metadata using a modest number of image and event-based semantic classifiers, as well as any meaningful file or folder names. When users pose text-based queries, our system retrieves images from their personal image collections by leveraging Flickr's tag dataset for concept expansion. This approach enables users to search their collections without having to manually annotate their pictures. We compare the retrieval performance of using a Flickr-based concept expander with the performance obtained without concept expansion and with using a WordNet-based concept expander. The results demonstrate that common sense knowledge gleaned from online photo sharing communities can enable meaningful image search on consumer image collections, searches that would be impossible using only the available image metadata.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval.

General Terms

Algorithms.

Keywords

concept expansion, Flickr, image search, multimedia search, multimedia retrieval, semantic search, WordNet

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.
Copyright 2010 ACM 978-1-4503-0099-5/10/10...$10.00.

1. INTRODUCTION

As consumer collections of personal multimedia grow, so grows the challenge of finding a particular, or even a relevant, asset. A consumer wishes to find a particular picture they vaguely remember, perhaps to reminisce, or to include in a special creative work. Or, the consumer is interested in searching for types of pictures—from a certain place, of a particular person, etc. Without automated tools for image retrieval, the consumer must potentially browse through their entire collection, examining each asset to see if it matches what they seek.

The ability to search and retrieve assets hinges upon two key pieces of functionality: the ability to express or describe what one is looking for in the form of a query, and the ability to characterize how well each asset satisfies the specified query. Query interfaces may take different forms, including both graphical and textual modes. Graphical or visual interfaces can be effective for certain applications, but require some sort of visual example. Text-based interfaces allow users to describe what they are looking for, using terms from their natural language. Consumer image search terms typically can be separated into who, what, when, or where categories. The query may be expressed as a list of keywords, such as "florida vacation," or it may be a more complex expression, such as "pictures of the crocodiles we saw while vacationing in the Everglades." The latter form requires complex natural language parsing to extract the semantic concepts, such as "where=Everglades, Florida," "what=crocodiles," and "when=vacation." Matching these queries against a set of pictures requires having the corresponding semantic information either available or readily inferable for each asset.

Few people invest the time and energy required to tag their images, although some might group related photos into folders with a descriptive name; for photos of particular importance, they might even rename the file. Metadata recorded by the capture device, such as the capture date and time, provides basic "when" information. Although still relatively rare, geospatial information is increasingly becoming available and recorded at capture time, providing valuable location information. Automatic extraction of semantic information using people and object recognition, along with scene and material classifiers, can greatly increase the amount of available semantic information, although the performance of such classifiers varies.

This work focuses on text-based queries and specifically considers mechanisms for bridging the concepts used by consumers to describe their search target with the concepts used by the system to describe the assets it has indexed. In this work, a consumer's image collection is first indexed using various semantic indexers or classifiers to produce a high-level characterization of each asset. The semantic indexers include image understanding algorithms such as scene, material, and color classifiers, as well as algorithms that interpret image Exif data. Optionally, the indexing process may also include user-provided content such as image captions or even the image pathname. The results from all of these indexers are combined to form a unified characterization for each asset in the form of a bag of concepts, such as {“vacation”, “Easter”, “Palm Springs”, “family”, “baby”, “beach”, ...}.

To search against the collection, the user enters in a text-based query. Query terms are run through a concept expander to produce a more complete set of related concepts. A common approach for concept expansion is to use the WordNet [3] lexical database. The linguistic mappings performed by WordNet are just that—linguistic mappings. They do not capture common sense relationships such as the fact that fireworks are often associated with July, or that flowers are often red or yellow. An important part of this work is to leverage such common sense relationships whenever possible to improve the search results.

Numerous online photo sharing communities such as Flickr, Photobucket, and the Kodak Gallery enable users to tag pictures with labels, which can provide a rich source of knowledge about commonly related concepts. Flickr is one of the larger such communities, containing over four billion uploaded photos [6]. This site is popular with photo enthusiasts as well as consumers simply interested in sharing their photos. Flickr permits individual photos to be tagged with up to 75 distinct tags. Our system uses the Flickr getRelated API [7] to map a search term to a set of terms that are likely to co-occur as labels in the Flickr database. The Flickr getRelated function returns for a given tag, the set of tags that are most frequently used together with that tag in tagging images. Such co-occurring terms may not be linguistically related. However, because they are likely to appear together as labels on images, and as the goal of this work is to search consumer image collections, their co-occurrence as Flickr tags makes them reasonable expansion terms. This work compares the quality of the search results using the raw search terms as provided by the consumer with concept expansion based on Flickr and WordNet.

Our work focuses on taking the user from an unindexed collection to initial search results, by combining semantic indexing with Flickr-based concept expansion. Given the sparse and imprecise semantic characterization of assets used by our system, we do not expect every search result to be relevant. However, people have become accustomed to searching for information using tools such as Google, which return back a pageful of responses, with the user visually skimming those results for relevance. Once a relevant item is found, the user can go directly to that result, and perhaps use that result as a launching pad to other relevant results. In the same manner, this system returns back a set of ranked results, which should be viewed as just the first step in information retrieval. As long as at least one returned result is relevant, the system has given the user a starting point into their collection. A complete system would then provide the user with other tools for branching out from that image to other images. Secondary navigation strategies might include retrieving other images from the same event, featuring the same person, or having similar visual features.

Related work is reviewed in the next section. Section 3 discusses background research conducted to understand the types of queries consumers wish to conduct. Section 4 describes the system architecture, and Section 5 presents the details of our concept expansion methodologies. We present the experimental results of our system in Section 6 and close with a final discussion in Section 7.

2. RELATED WORK

Representing concepts as terms and applying statistical text retrieval techniques for image retrieval is not new. In [17], the authors propose a retrieval system that represents both multimedia documents and queries as sets of feature terms using Boolean vectors, and uses statistical text retrieval models to perform relevance calculations. The work described here, while using a statistical retrieval model, leverages the Indri [19] text-based information retrieval system with a simple, keyword-based representation of image concepts, in order to allow us to focus on the primary purpose of this research—concept expansion for consumer image searches.

Past work has looked at a variety of means for concept expansion, including common sense knowledge. For example, [14] uses semantic concept expansion by leveraging the Open Mind Common Sense (OMCS) Knowledge Base as its source of real-world knowledge. Rules are derived from the OMCS knowledge base and represented as a type of semantic network, specifically, a weighted graph of concepts. Concepts are then expanded using spreading activation. The approach described here also seeks to use common sense knowledge for its concept expansion, but leverages Flickr as its basis for that common sense knowledge. In particular, in this work, the co-occurrence of tags as used to tag images within Flickr serves as a source of common sense knowledge. Although this form of common sense knowledge is weaker than OMCS or a formally constructed knowledgebase such as Cyc [12], it is directly related to the subject matter of consumer photography and is readily available in an already-analyzed form.

Alternative approaches for concept expansion rely on lexical techniques. In [8], the authors use the lexical database WordNet’s synonym, hypernym, and hyponym relationships to expand queries for images on the Web. Using a lexical expansion mechanism such as WordNet can easily result in irrelevant concepts being used as part of the expansion. For example, words commonly used in English often have multiple meanings or senses. Using WordNet terms for an unintended sense of the word may result in noise and decreased precision in the search results. To address this problem, [8] filters search terms by using a term semantic network (TSN) that is specific to their collection. The term semantic network is constructed by using an association mining algorithm to establish one-to-one relationships between words. The strength (or weakness) of these relationships is then used to filter out expansions suggested by WordNet. This type of approach may be applicable to both the Flickr and WordNet expansion mechanisms used in our work, but further work would be needed to determine how this approach might apply to this domain. Unlike [8], our work searches over a user’s personal image collection, and the vocabulary associated with the collection is likely to be much smaller than would be present in a Web-based search system. Our system can also incorporate WordNet’s synonym, hypernym, and hyponym relationships as part of its expansion mechanism; however, such relationships are rather limited. The language-based expansion provided by

WordNet is not likely to indicate that beaches are often associated with vacations, or that manatees are associated with Florida, but such common sense associations are available from Flickr. The wealth of information available via sources such as Flickr has previously been harnessed for the related problem of annotating images; for example, in [18], the authors describe a system for suggesting additional tags for pictures based upon a limited snapshot of Flickr's database. An earlier work [1] provided the user with a way to iteratively search through a subset of images downloaded from Flickr by either specifying tags, or by searching for visually similar images.

The work described in [2] considered two semantic similarity measures for video retrieval, one based on the Lin metric for word similarity, and the other based on a metric of pointwise mutual information for information retrieval, PMI-IR. The better performing PMI-IR approach computes similarity between any two concepts by measuring the likelihood of text co-occurrence in web documents as reported by Yahoo search; the Lin metric uses WordNet as its knowledgebase. Although our implementation is currently for still imagery, our technique should be readily applicable to video; likewise, their technique should transfer to still images. Our work uses a domain-specific knowledgebase (Flickr) for determining co-occurrence rather than arbitrary web documents.

A different approach to this problem is adopted in [13], where the authors also implement text-based queries on consumer image collections. In [13], the authors use the query to first identify relevant and irrelevant images from a set of 1.3 million images obtained from the Photosig.com online community, which are then used to train k-nearest neighbor and decision stump classifiers, which are then used to rank the consumer images. Although our approach requires precomputing both semantic descriptors for each asset as well as tag co-occurrence metrics, these computations are done in advance and they do not impact the retrieval performance.

This work is distinguished from the prior work in that it provides a model for a consumer-centric system for fast real-time image retrieval, where the consumer is not required to perform any further tagging tasks beyond what they might normally do. By leveraging web-accessible domain-specific semantic information obtained from the Flickr photo-sharing community, the system is able to significantly enhance the value of available metadata, enabling the system to accurately retrieve assets from the collection.

3. QUERY NEEDS

We conducted an online survey to characterize how consumers would like to query their personal image and video collections. As part of the survey, we asked users to submit five to ten top text queries that they would like to pose to their image and video collections. A total of 4813 queries from 932 different American-based users were analyzed. The length of user queries was an average of two words, with a standard deviation of 1.3 words and a median of two words. This is about one word less than that reported for general web searches [11] and about two words less than that reported for web-based image searches [9][10], although the latter case includes words such as "image" or "jpeg" as part of the query.

In terms of query content, each query was translated into a sequence of query codes and the query codes were organized into a query ontology. High-level results show that users query most frequently by when, then by who, other, what, and where. The top "when" query terms included event references (e.g., Thanksgiving) and year. The top "who" query terms referenced a person by first name. The top "other" query terms consisted of connector references such as "and", "of", and "at". The top "what" terms referred to things such as pets or nature. The top "where" terms referred to cities. The top query term pairs consisted of when-when queries (e.g., Christmas 2008) and who-when queries (e.g., Mom's birthday). The survey results were used to inform the development and evaluation of our image retrieval system described herein.

4. SYSTEM OVERVIEW

4.1 System Components

The system consists of several major components illustrated in Figure 1. The system runs indexers on the user's personal image collection to produce descriptive conceptual data, represented as one descriptor document per image. These descriptor documents are then indexed using the Indri information retrieval system. The query processor takes user queries and expands them into queries for Indri, using Flickr and WordNet (not shown) to obtain an expanded set of concepts. The expanded search terms are supplied to Indri, with the resulting image descriptors returned to the user interface to display the appropriate image thumbnails.

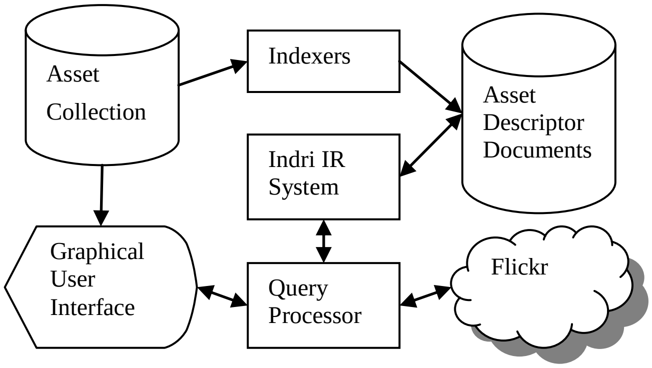

Figure 1. System Components

The asset descriptor documents are simple XML-based documents containing semantic descriptors. This type of document structure was used to enable the use of the Indri information retrieval system for indexing. Figure 2 illustrates the body of an asset descriptor, which is described further in Section 4.3.2.

Figure 3 provides a screenshot of the system in action. A separate mode was used for testing and collecting ground-truth data. In the example, the user had entered the search term "manatee" and requested that the system perform the search using only the Flickr-based concept expansion methodology. The lower left shows the returned results by file name, along with their respective scores; the lower right shows image thumbnails. In the example, all four pictures in the first row of pictures and one picture in the second row portray manatees.

```
<text>
      <path>\data\users\l411946\images\FamilySubset\2005\Florida\
100_0432.JPG</path>
      <label>Valentine's Day 2005</label>
      <temporal>February 14, 2005 winter afternoon</temporal>
      <type>Sports</type>
      <topType>Vacation</topType>
      <address>Homosassa, Florida, US</address>
      <featureType>park</featureType>
      <nearbyFeatures>Homosassa Springs Wildlife State Park: park
Homosassa Springs: spring(s)
      </nearbyFeatures></text>
      <classes>beach beach beach urban sunset mountain field
foliage foliage foliage foliage rock rock rock rock sand sand grass
water</classes>
      <classes>gray gray gray gray gray</classes>
</text>
```

Figure 2. Sample Asset Descriptor document

4.2 Image Indexing

A key goal of this work is to enable image retrieval in the absence of user-provided annotations. To this end, the system employs as many conceptual level indexers as feasible. These indexers use advanced image understanding and recognition techniques to automatically label images or video clips with semantic concepts typically associated with consumer imagery.

To determine which indexers to run, we must trade off the cost of running the indexer with the expected return, i.e., the quality of the conceptual information provided by the indexer. The system currently includes a variety of customary scene and material classifiers (e.g., mountain, beach, and water) as well as a probabilistic event classifier [3], which classifies an event into one of four types: vacation, family moment, party, or sports. The event classifier uses both image-level features such as people present, indoor/outdoor, and type of scene detected (e.g., nature, urban, beach), and event-level features including inter-event time and the time of day.

Face detectors are combined with age and gender estimators to infer high-level conceptual descriptors of the people portrayed in an image, using terms such as baby, boy, girl, man, and woman. Related concepts provided by Flickr often include color names; the system uses a color analyzer to generate image color labels from a list of twelve common color names [4]. Running all the various pixel-based image indexers may take up to a second per image to run, or more, depending on the speed of the machine, but such classifiers are expected to be run in the background, in advance of any retrieval operations.

Most cameras record the time of capture with the image, and this information can be a valuable source of semantic information. An auto-event labeler maps the date to the name of a holiday, such as Christmas. The current auto-event labeler is primarily aware of American holidays, but could be easily adapted to incorporate different calendars for other localities and cultures; it could also be extended to be aware of personal calendar items such as birthdays. Similarly, the date and time are mapped to time of day as well as the season of the year. Some cameras record the latitude and longitude of where an image was captured; alternatively, such information is often added afterwards by the user. If geospatial information is available, the system uses the online geonames.org web service to reverse geocode the latitude and longitude to the place name as well as to an expected geospatial feature type, using terms from the geonames.org vocabulary, such as park or school.

Finally, should the user have provided the system with any type of annotations, such information is also included in the image descriptor. While users seldom provide captions, they often do give the file or containing folder a descriptive name.

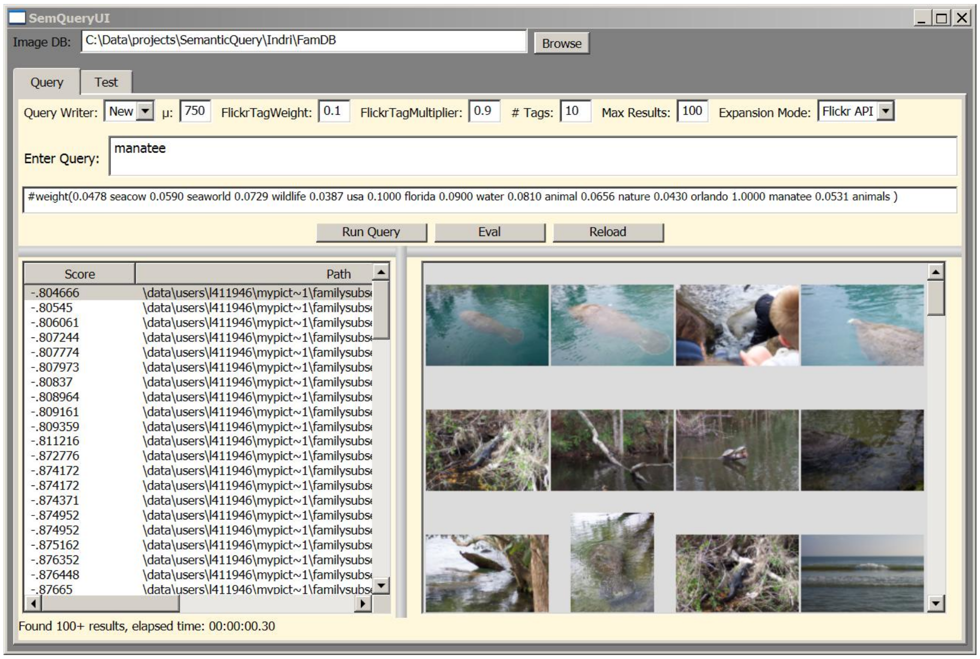

Image DB: C:\Data\projects\SemanticQuery\Indri\FamDB

Query Writer: New μ: 750 FlickrTagWeight: 0.1 FlickrTagMultiplier: 0.9 # Tags: 10 Max Results: 100 Expansion Mode: Flickr API

Enter Query: manatee

#weight(0.0478 seacow 0.0590 seaworld 0.0729 wildlife 0.0387 usa 0.1000 florida 0.0900 water 0.0810 animal 0.0656 nature 0.0430 orlando 1.0000 manatee 0.0531 animals)

Score	Path
-.804666	\data\users\l411946\mypict~1\familysubs
-.80545	\data\users\l411946\mypict~1\familysubs
-.806061	\data\users\l411946\mypict~1\familysubs
-.807244	\data\users\l411946\mypict~1\familysubs
-.807774	\data\users\l411946\mypict~1\familysubs
-.807973	\data\users\l411946\mypict~1\familysubs
-.80837	\data\users\l411946\mypict~1\familysubs
-.808964	\data\users\l411946\mypict~1\familysubs
-.809161	\data\users\l411946\mypict~1\familysubs
-.809359	\data\users\l411946\mypict~1\familysubs
-.811216	\data\users\l411946\mypict~1\familysubs
-.872776	\data\users\l411946\mypict~1\familysubs
-.874172	\data\users\l411946\mypict~1\familysubs
-.874172	\data\users\l411946\mypict~1\familysubs
-.874371	\data\users\l411946\mypict~1\familysubs
-.874952	\data\users\l411946\mypict~1\familysubs
-.874952	\data\users\l411946\mypict~1\familysubs
-.875162	\data\users\l411946\mypict~1\familysubs
-.876352	\data\users\l411946\mypict~1\familysubs
-.876448	\data\users\l411946\mypict~1\familysubs
-.87665	\data\users\l411946\mypict~1\familysubs

Found 100+ results, elapsed time: 00:00:00.30

Figure 3. Example Screenshot

4.3 Indexing and Retrieval

Our system uses the Indri information retrieval system to rank and retrieve asset descriptor documents. Corresponding images can then be easily retrieved and displayed to the user, as shown in Figure 3. In this section, we provide an overview of Indri and how it is used within our image retrieval system.

4.3.1 The Indri Information Retrieval System

The Indri information retrieval system [19] is a hybrid document search engine, based upon a combination of language modeling and inference network retrieval [15]. A language modeling system is one in which the system computes an estimated probability for generating a given query, given a particular language model for a document. The document language model is not the document itself, but can be thought of as a set of probabilities, or priors, for the words in a given vocabulary. Using this model, the probability that a given word is in a document can be estimated. Using the simple maximum likelihood assumption, the probability p_{ml} of a given term *w* occurring in document D is defined as follows [16]:

$$(1) \quad p_{ml}(w \mid D) = \frac{tf_{w,D}}{|D|}$$

where $tf_{w,D}$ is the term frequency of w in D, and $|D|$ is the number of words in the document. The maximum likelihood assumption is seldom used in practice. If a word is not present in a given document, the maximum likelihood assumption results in a zero probability. To avoid this problem, most retrieval systems apply a smoothing function. We used Indri's default, the Dirichlet smoothing function. Using this smoothing function, the probability of a given term *r*, given a document D, is defined as:

$$(2) \quad P(r \mid D) = \frac{tf_{r,D} + \mu \cdot P(r \mid C)}{|D| + \mu}$$

where $tf_{r,D}$ is the term frequency of r in D, $P(r|C)$ is the probability of r given the entire collection of documents C, and $|D|$ is the number of words in the document. The constant μ provides the smoothing function; if μ is 0, then the formula degenerates to the maximal likelihood assumption (1). The default value for μ in Indri is 5000. Preliminary testing of our system indicated that the retrieval performance was improved by using a smaller value of μ. This is consistent with previous research on smoothing [20], which indicates that if the queries are small, as is expected in our system, then smaller values of μ provide better results. Larger values of μ result in more smoothing, and consequently decrease the precision of the results. We used μ = 750 in our tests.

The term $P(r|C)$ represents the probability of a given term *r* for a given collection of documents C. This probability may be modeled by considering the relative frequency of each term appearing in the collection. In this case, note that (2) still returns a value of zero for terms that do not appear in any document in the collection.

The Indri inference network model applies Bayesian logic to define how complex queries may be composed from simpler terms. The leaves of the inference network are the representation nodes, corresponding to the probability of a given term appearing in a document, using the language model. Various representation nodes may be combined using belief nodes, using both weighted and unweighted belief operators [16]. The #combine operator is an example of an unweighted belief operator, defined as follows:

$$(3) \quad b_{\#combine} = \prod_{i=1}^{n} b_i^{\left(\frac{1}{n}\right)}$$

The b_i are the values of the constituent *n* belief nodes, which may be either the base probabilities for the representation nodes—the individual search terms—or other belief nodes. The weighted equivalent is #weight, defined as

$$(4) \quad b_{\#weight} = \prod_{i=1}^{n} b_i^{\left(\frac{w_i}{W}\right)} \text{ where } W = \sum_{i=1}^{n} w_i$$

The w_i specify the weights to be associated with each belief node. Indri defines a relatively rich set of belief operators that allow arbitrarily complex inference networks to be constructed.

Indri combines the inference network model with the underlying language model to compute for a given query, the probability that a given document model satisfies that query. The document model is a smoothed multiple-Bernoulli distribution representation of a document. The overall probability that a given document satisfies a query is represented by the document node, which combines the various document models into a single probability. The probabilities from a set of documents from a document collection may be used in order to compute a rank ordering.

Queries are made in Indri using a high-level query language, which supports operators such as #combine and #weight, as well as additional operators, including proximity operators and filtering operators.

4.3.2 Metadata Representation

Indri is designed as a document retrieval system. As such, its model was not a perfect match for using it to match probabilistic data generated by some of our image indexers. For example, our scene classifiers generate numeric scores. The asset document creator normalizes this score and then outputs the concept term a variable number of times, in proportion to how highly the term scored. This introduces round-off error, and the computed probability will depend upon the smoothing function.

Indri provides a way to specify tables of priors, and this approach was originally used to model the probabilistic data associated with the scene and material classifiers. However, when queries contain a mixture of terms, some associated with data for which priors are specified, and some not, the priors had an undue impact on the outcome. For example, consider the term "mountain." Our scene classifier included a mountain scene detector, so it was possible to generate a table of priors specifying for each image some prior probability that the document contained this term, based on the output of the scene detector. An asset descriptor document might contain the term mountain independent of the scene detector. For example, the geospatial indexer may have reverse geocoded the GPS coordinates to determine that the picture was taken on a mountain. Or the user may have provided the term mountain as part of a caption or the filename. These other sources of information could even be more accurate than the scene classifier. Ideally the two sources of information should be seamlessly woven together to form an overall probability that the mountain concept belonged to an image. However, merging these two sources of data proved to be problematic. Although the asset document generator could scale the priors as needed, it appeared that it would require considerable document analysis to determine the scaling function. Instead, we observed a major discontinuity in the computed probabilities between those based on a prior value and those based solely on the presence of the word in the document. Without a way to smooth the differences between the two types of probabilities, the system returned anomalous scores. To avoid this problem, we adopted the simple approach of making the frequency of such terms in the asset descriptor dependent upon their probability or score. For example, Figure 2 shows a sample asset descriptor document where the corresponding image has a higher probability of portraying rock than water.

5. CONCEPT EXPANSION

5.1 Flickr-Based Concept Expansion

Flickr provides a means for people to tag pictures with one or more keywords; it also supports a flexible API for developers to access Flickr content and summarizing data. This API includes the method getRelated, which returns a list of tags related to a specified tag, based upon cluster analysis performed by Flickr. The details of the algorithm used to compute related tags has not been published. However, it appears that the returned list of tags is the set of keywords most likely to co-occur with the specified tag. We use the getRelated API in this work as a source of common sense information. This interface provides us with a set of related concepts, and one tied to the domain of interest: consumer photography. Although the Flickr community includes a disproportionate number of prosumer and professional content, it also contains a considerable amount of "everyday" photos—activities, events, and places that could apply to almost anyone.

Some examples will illustrate the type of content provided by this interface. For the keyword "vacation," the getRelated method returns the following list:

> *travel, beach, summer, water, ocean, sky, nature, blue, sun, trip, sunset, sand, landscape, sea, clouds, fun, green, family, waves, trees, holiday, boat, people, city, europe, florida, island, light, architecture, night, art, street, mexico, red, tree, coast, canon, color*

For the keyword "turkey," getRelated returns the following:

> *istanbul, mosque, sea, thanksgiving, bluemosque, blue, islam, sultanahmet, bosphorus, travel, hagiasophia, food, muslim, bridge, architecture, street, turchia, dinner, bird, church, people, europe, ayasofya, palace, topkapi, camii, minaret, asia, cami, turquia, sunset, night, estambul, turkiye, boat, galata, turkei, city, water, sky, light, turkish, bw, red, sun, market, deniz, bazaar, cat, turquie, museum, taksim, nature, portrait, canon, ottoman, ship, holiday, tower, hijab, man, window, animal, constantinople, family, white, woman, goldenhorn, black, grandbazaar, mosaic, clouds, nikon, sophia, reflection, summer, sultan, stuffing, old, lights, dome, art, byzantine, birds, fish, religion, bosporus, blackandwhite, ferry, green, flag, topkapipalace, hagia*

The latter list indicates the ambiguous nature of the term—it could be referring to the country of Turkey, or the bird. As keeping with the prosumer bent of the Flickr audience, the set of tags also includes terms specific to the photo taking activity, such as "nikon" and "blackandwhite," which tell us nothing about the concept.

The tags returned by Flickr appear to be in order of co-occurrence frequency. In using these tags for expansion, the system takes only the first *n* tags, where *n* by default is ten. The expanded query consists of the original keyword plus the keywords from Flickr. Weights are applied as in the case of WordNet. The terms provided by Flickr are weighted 0.1, relative to the weight of 1.0 assigned to the user-specified term. One limitation of the Flickr getRelated interface is that it does not return compound words as such; instead, such tags are made into single words by removing spaces. Hence, "New York" becomes "newyork" when returned by the getRelated method. To compensate for the specific but common case of state names, the system recognizes and substitutes the full state name for the abbreviated form of states having names composed of compound words.

5.2 WordNet-Based Concept Expansion

WordNet groups words into sets of cognitive synonyms, or synsets. A word may have multiple synsets, corresponding to different senses of the word. Rather than use all possible senses of a word, we limit ourselves to the two most common senses, as reported by WordNet. Moreover, a word may be used both as a noun and as a verb, with the noun and verb forms each having separate synsets. For the purposes of this work, we consider only the noun synsets for a word. When a user provides a list of search terms, it is generally not readily possible to determine the intended part of speech for each term. We make the simplifying assumption in this work that search terms are intended as nouns, and limit ourselves to the use of noun synsets. We expect that people are more likely to provide nouns or noun forms of verbs as search terms than verbs, and this assumption was borne out by the results of our online survey described in Section 3, where people rarely used verbs as search terms.

WordNet also links words related by hypernym/hyponym relationships, as well as other relationship types not used in this work. A word is a *hypernym* of another word if the first word represents a more general class than the second word: for example, ballgame is a hypernym of baseball. A word is a *hyponym* of another word if it represents a more specific instance of a more general class; baseball is a hyponym of sport. This work uses the synonym, hypernym and hyponym relationships defined by WordNet to identify related words. As with synsets, we limit the expansion to noun senses of the word only.

WordNet can return hypernym chains, for example, sport is a hypernym of athletic game which in turn is a hypernym of outdoor game, a generalization of field game, a generalization of ballgame, a generalization of baseball. As the classes become more general, they are less useful for the purposes of concept expansion. Consequently, our system only considers two levels of hypernyms, and that only for the two primary senses of the word. For hyponyms, the system again only considers the hyponyms for the two primary senses of the word, and only the immediate hyponyms are used.

The system expands a term using WordNet by constructing a bag of terms, using the terms provided by WordNet's synset, hypernym, and hyponym relationships. We expect that the user-provided term should have a higher weight than synonyms, which in turn should have a higher weight than the hypernyms and hyponyms. After experimenting with various weights, we adopted the following weight assignments:

- The word itself, 1.0
- Synonyms, 0.25
- Hypernyms, 0.05
- Hyponyms, 0.05

6. EXPERIMENTAL RESULTS

To assess the relative performance of Flickr-based concept expansion, we compared its retrieval performance against not using any expansion and against using WordNet-based expansion. All testing was first-party testing; test participants generated meaningful queries for their personal collections, and provided ground-truth results for each query. Our study included five test participants with collections of varying sizes, as shown in Table 1.

Table 1. Participant Collections

Participant	Collection Size	Number Geotagged
L	1099	0
D	1908	0
W	1245	63
G	4291	0
H	3055	226

Collections were as provided by the user, with file and directory names as they were in the user's environment. While the majority of images did not have meaningful filenames, many image files resided in a directory with a name corresponding to a date, activity or place. Two of the test participants had some geotagged images in their collections; one collection had approximately 20 images with associated Exif captions. Most collections spanned several years. Some collections included duplicates of some images, at different resolutions, where the participant had resized some images for the purposes of sharing; this was especially true for Participant G. Because of the difficulty of obtaining ground-truth data, two of the authors participated as test participants.

Table 2. Test Query Results (With User Metadata)

		Precision					Recall			
Part.	User Query	Flickr	WORD	COMBO	NONE		Flickr	WORD	COMBO	NONE
W	christmas	100.0%	95.0%	100.0%	95.0%		32.3%	30.6%	32.3%	30.6%
W	christmas presents	45.0%	55.0%	55.0%	60.0%		18.4%	22.4%	22.4%	24.5%
W	florida vacation	100.0%	100.0%	100.0%	100.0%		37.0%	37.0%	37.0%	37.0%
W	camping adirondacks	35.0%	50.0%	40.0%	0.0%		6.0%	8.5%	6.8%	0.0%
W	manatee	35.0%	0.0%	35.0%	0.0%		100.0%	0.0%	100.0%	0.0%
W	canoeing	35.0%	100.0%	35.0%	100.0%		87.5%	62.5%	87.5%	62.5%
D	mountain vacation	95.0%	85.0%	95.0%	90.0%		6.5%	5.8%	6.5%	6.2%
D	alberta vacation	100.0%	100.0%	100.0%	100.0%		6.7%	6.7%	6.7%	6.7%
D	christmas	95.0%	100.0%	100.0%	95.0%		38.0%	40.0%	40.0%	38.0%
D	christmas presents	85.0%	90.0%	90.0%	85.0%		77.3%	81.8%	81.8%	77.3%
D	flowers	0.0%	25.0%	5.0%	0.0%		0.0%	6.8%	1.4%	0.0%
D	ocean beach	45.0%	10.0%	25.0%	5.0%		11.7%	2.6%	6.5%	1.3%
D	dance recital	10.0%	20.0%	10.0%	0.0%		4.3%	8.7%	4.3%	0.0%
L	cello recital	40.0%	80.0%	40.0%	80.0%		100.0%	100.0%	100.0%	100.0%
L	cruise	80.0%	0.0%	75.0%	0.0%		10.6%	0.0%	9.9%	0.0%
L	natural history museum	40.0%	0.0%	40.0%	0.0%		100.0%	0.0%	100.0%	0.0%
L	boating	0.0%	5.0%	0.0%	0.0%		0.0%	1.6%	0.0%	0.0%
L	disney princess	30.0%	15.0%	30.0%	15.0%		28.6%	14.3%	28.6%	14.3%
H	golf	45.0%	45.0%	45.0%	100.0%		100.0%	100.0%	100.0%	100.0%
H	mickey mouse vacation	5.0%	0.0%	10.0%	0.0%		16.7%	0.0%	33.3%	0.0%
H	birthday beijing	0.0%	0.0%	0.0%	0.0%		0.0%	0.0%	0.0%	0.0%
H	birthday	5.0%	0.0%	5.0%	0.0%		1.3%	0.0%	1.3%	0.0%
H	waterslide	0.0%	0.0%	0.0%	0.0%		0.0%	0.0%	0.0%	0.0%
H	daughter grandma	40.0%	0.0%	45.0%	0.0%		12.1%	0.0%	13.6%	0.0%
H	son soccer	5.0%	0.0%	0.0%	0.0%		20.0%	0.0%	0.0%	0.0%
H	violin recital	0.0%	0.0%	0.0%	0.0%		0.0%	0.0%	0.0%	0.0%
H	grand canyon	10.0%	0.0%	10.0%	0.0%		5.0%	0.0%	5.0%	0.0%
G	wedding	0.0%	0.0%	0.0%	0.0%		0.0%	0.0%	0.0%	0.0%
G	christmas	100.0%	100.0%	100.0%	100.0%		16.9%	16.9%	16.9%	16.9%
G	graduation	0.0%	0.0%	0.0%	0.0%		0.0%	0.0%	0.0%	0.0%
G	letchworth camping	5.0%	0.0%	5.0%	0.0%		1.4%	0.0%	1.4%	0.0%
G	hot air balloon	0.0%	0.0%	0.0%	0.0%		0.0%	0.0%	0.0%	0.0%
G	swimming pool	10.0%	5.0%	15.0%	0.0%		28.6%	14.3%	42.9%	0.0%
G	mom and dad	30.0%	25.0%	15.0%	0.0%		50.0%	41.7%	25.0%	0.0%

In keeping with the types of queries we found that consumers typically wish to perform, the queries were one- to three-word when, who, where, or what type queries, where the what type queries included both objects and activities. The complete set of queries and the associated precision/recall results for the datasets with user-provided metadata are shown in Table 2. Some subjectivity is involved in determining the ground truth. In the most extreme case, one participant searched for “mom and dad,” expecting pictures of the participant’s mother and father, but the participant also accepted results of the participant and his wife, if the picture was taken at a time when he had children. However, the vast majority of the queries did not have such ambiguous interpretations. For two-word queries, participants expected the search results to satisfy both terms. The raw collections as provided by the test participants represent the type of collections real consumers would have: some user-provided metadata, largely in the form of meaningful directory and file names, but not much. In addition, as noted in Table 1, some participants had manually or automatically geotagged images. To provide a worse-case scenario, each participant collection was cloned, and all directories and files renamed before the images were indexed. In addition, geospatial information was discarded in the cloned copy. Although an increasing number of capture devices automatically record geospatial information, most consumer imagery is still not geotagged. The cloned collection was also indexed, to assess the performance of the system when leveraging only typical image capture-time Exif or pixel data. As would be expected, the vocabulary associated with the metadata-reduced, cloned collection was much smaller than the vocabulary for the corresponding original collection. For example, for participant W, the meaningful vocabulary for the original collection consisted of approximately two thousand words, while the corresponding vocabulary for the metadata-reduced collection consisted of approximately seventy words.

Each test participant was instructed to supply five to seven queries, searching for events, places, activities, and objects, in line with the types of queries our consumer research indicated as being most important to consumers. For each query, the query was expanded using each of the tested expansion methods, with the top twenty results from each method taken and presented to the user. The user was instructed to indicate whether or not the results were correct. The test was repeated with the participant’s cloned collection, where user-provided metadata had been deleted.

The results were analyzed to determine the precision and recall of the various expansion methodologies. In addition, we considered the overall utility of the system, by considering what the likelihood was that the system would return at least one relevant result that would rank in the top twenty results.

Figure 4 illustrates the expected probability of returning at least one relevant result, given asset descriptors that may include user-provided metadata. The crosshair indicates the proportion of searches that returned at least one correct result. To assign confidence values to these proportions, we assumed a beta distribution, and computed the confidence interval for α = 0.1. As can be seen from the graph, the Flickr-based concept expansion provides significantly better results than using no concept expansion. Combining the Flickr and WordNet expansion methodologies provided nearly identical results in this case. Using WordNet alone as the expansion methodology did not provide significantly better results than using no expansion methodology, although the observed proportion of successful results was higher.

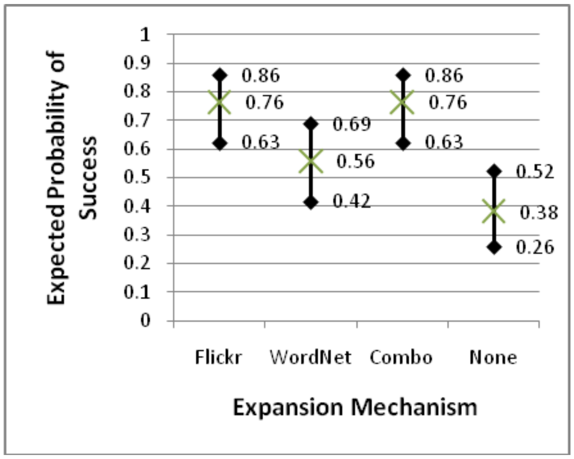

Figure 4. Expected Probability of Success, with User Metadata

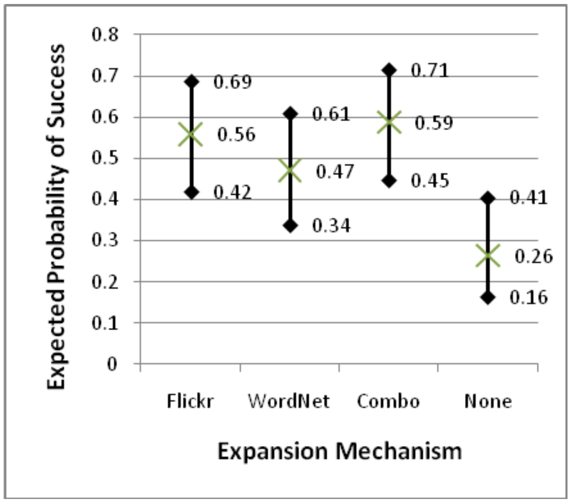

Figure 5. Expected Probability of Success, No User Metadata

In Figure 5, we show the same type of data, but this time as applied to the cloned collections, lacking the user-provided and geospatial metadata. The same conclusions apply, although here the difference between Flickr and no expansion is barely statistically significant. Combining Flickr with WordNet strengthens the separation.

For reference, the average precision using the combined approach for our test queries was 0.36 compared to 0.30 with no expansion; these numbers drop to 0.25 and 0.16 respectively when the queries were executed against the collection with no user-provided metadata. The average precision using just Flickr-based expansion was similar to the precision using the combined approach. To confirm our expectation that the presence of user-provided metadata improved the quality of the search results, we performed a paired t-test to compare the precision for each query as returned by the two methods. The presence of user-provided metadata significantly improved the results, with a one-tail p = 0.004 (mean change = 0.11, st. dev. = 0.04, t-stat(34) = 2.78).

Fast response time is a critical requirement for any practical search system. We measured the time to execute each of the queries as shown in Table 2, using the combined Flickr plus WordNet expansion mechanism, and taking the first twenty results. The average response time was 0.63 seconds with a standard deviation of 0.23, running on a Intel Xeon processor (2.67 GHz, high-speed Internet connectivity, 32-bit Windows XP). No special effort was made to optimize the execution time.

7. DISCUSSION

In considering the query results, we note both some characteristics that resulted in more successful results, as well as areas that merit further research. As most consumers do not want to review more than one page of search results, our data analysis only considered the performance of the system relative to the top twenty search results.

As would be expected, the presence of user-provided metadata results in better precision: the more annotations provided by the user, the better the expected results. Even with no user-provided metadata, our Flickr-based concept expansion mechanism significantly improved search performance compared to not using an expansion mechanism. Leveraging information such as pathnames and geospatial tags significantly improved performance. Meaningful file and folder pathnames are often available for free, in that the consumer has provided such information as part of their asset organization; geospatial information is increasingly becoming available automatically.

In looking at specific queries, some queries included terms that were directly returned by one of the asset indexers or were part of user-provided data; for example, our holiday recognizer tagged some images with the tag "Christmas," and several of the queries did involve this tag. Likewise, the vacation tag was also directly returned by our event classifier, although in this case the classification was not always accurate. The more interesting queries were ones such as manatee, illustrated in Figure 3. For that particular term, the concept expander expands the term to the following Indri query when using Flickr as the concept expansion method:

> #weight(0.1 seacow 0.1 seaworld 0.1 wildlife 0.1 usa 0.1 florida 0.1 water 0.1 animal 0.1 nature 0.1 orlando 1 manatee 0.1 animals)

This query worked with 100% recall, even though the term manatee was not part of the collection vocabulary. In this case, the relevant images were geotagged, and the latitude and longitude information was mapped by the geospatial indexer to tags such as Wildlife State Park and Florida. The pixel-based water classifier further improved the expansion, resulting in four of the manatee pictures being the top-ranked results. Without the water tag, those pictures still ranked in the top twenty, but further down the list. However, in the absence of user-provided and geospatial metadata, this query is essentially hopeless, returning no valid matches in the top twenty results, as the water tag by itself was insufficient to identify the correct images.

As another example, the user query "golf" had moderately high precision, as shown in Table 2, but this is hardly surprising as the relevant images were in a folder named "golf." Due to space limitations, we do not show the test query results data for the case where there is no user metadata. In that case, the precision for the golf query dropped to 5% (one image) using the Flickr-based expansion mechanism, but it is worth noting that the Flickr-based mechanism was the only one that returned any valid images for that query. The retrieved image matched based on the results of the scene and color classifiers. An area for future research is to identify which classes of asset descriptor tags are most useful for query expansion. This research would indicate where improvements in indexer accuracy would most improve retrieval results, as well as potentially indicate additional indexer types that could be helpful.

We note that both Flickr and WordNet expansion could result in adding irrelevant terms. If these terms were part of the collection vocabulary, then these terms could negatively impact the search results. This is illustrated in Table 2 where for a few queries the Flickr results are actually worse than using no expansion. For these queries, the user-provided search terms typically matched at least one of the more reliable asset descriptor terms—terms provided by the user or by the date-based autolabeller. The Flickr-based expansion, while improving recall in one case, did so at the expense of decreasing precision.

We conducted some informal experiments to determine how the Flickr-based algorithm performed when varying the number of terms used to do the expansion. We postulate that the number we used, ten, might be generally too low; however, using twenty or more seemed to add more noise and more error in the results.

Techniques such as the term semantic network used in [8] could possibly be adapted to the work here, to filter the expanded search terms. Such filtering could be applied in several different ways. The WordNet expansion could be filtered by using Flickr to determine the strength of the suggested associations. In addition, if the user supplies multiple search terms, filtering could be used to ensure that the expanded terms are consistent with all terms. For example, our current expansion algorithm will include the word ocean for the query "Vermont vacation," since ocean is related to vacation according to Flickr. Since oceans are not typically associated with Vermont, dropping the term ocean from the expansion will improve the precision.

We also note that the behavior and semantics of the Indri information retrieval system did not always match our intended semantics. Queries consisting of multiple terms were mapped to a single Indri expression using a weighted combination operator. This resulted in a weighted disjunction, although our test participants considered the results to be erroneous unless all terms were satisfied. For example, the "birthday beijing" query readily returned pictures from Beijing, as those pictures had been geotagged, but it did not succeed in retrieving the birthday pictures. A more complex rewriting of queries into one or more Indri expressions would be required to more closely match user intent, and should result in improved precision. Considerably more effort would be required to develop an information retrieval system that more closely matched our data model, and it is unclear how much that would improve system performance. Our method of representing probabilistic values resulted in a loss of precision. For many queries, multiple assets had the same score, and because we only considered the top twenty results, the cut-off in some cases was indiscriminate. A more precise means to represent probabilistic scores would have provided better differentiation in the ranking of results, and we may do more work in this area in the future. We anticipate consumer applications will have a growing need for information retrieval systems that can readily handle a mix of probabilistic and non-probabilistic metadata.

Our implementation makes real-time calls to the Flickr getRelated interface to expand searches. This highly available interface typically provides results very quickly, making it suitable for real-time interactive use. Our experimental system often returns results within one second, including the calls to Flickr, the WordNet expansion, and finally the Indri queries. No real-time pixel-level image processing is required, as all semantic information is extracted in advance, enabling fast response times. Performance could be further improved by caching common keyword expansions, although real-time invocations of the Flickr service ensure that our system's expansion tracks current usage. We considered using other sources of knowledge such as Wikipedia, but a dataset such as Flickr is a better fit for our search space.

Although the currently collected data is limited, the results indicate that meaningful image search is possible on consumer image collections without requiring the consumer to provide metadata beyond that which they might normally provide. Our initial success in this domain leverages the shared knowledge of millions of Flickr users as to what concepts are likely to go together in the context of consumer imagery. This readily-obtained knowledge provides us with a fast way to expand consumer queries into sets of semantic concepts that can be readily extracted from consumer imagery. By combining concept expansion with image semantic classifiers, our system provides consumers with real-time interactive search and retrieval capabilities for their unlabeled or sparsely labeled personal image collections.

8. ACKNOWLEDGEMENTS

Madirakshi Das, Dhiraj Joshi, Jiebo Luo and Tony Scalise provided helpful insights and observations on this work. We also thank the test participants for providing us with real user data and the anonymous reviewers for their helpful feedback.

9. REFERENCES

[1] Aurnhammer, Melanie, Peter Hanappe, and Luc Steels. "Integrating Collaborative Tagging and Emergent Semantics for Image Retrieval," In Proc. 2006 International World Wide Web Conference (Edinburgh, UK, May, 2006), ACM Press.

[2] Ayata, Yusf, Mubarak Shah and Jiebo Luo, "Utilizing Semantic Word Similarity Measures for Video Retrieval," In Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Miami Beach, Florida, June, 2009).

[3] Das, Madirakshi, and Alexander Loui. "Event Classification in Personal Image Collections," In Proc. 1st IEEE Workshop on Media Information Analysis for Personal and Social Application (New York, NY, June 2009, in conjunction with ICME '09).

[4] Das, Madirakshi, R. Manmatha, and Edward M. Riseman, "Indexing Flower Patent Images using Domain Knowledge," IEEE Intelligent Systems, 14, 5 (September 1999), 24-33.

[5] Fellbaum, Christiane (ed.), *Wordnet: An Electronic Lexical Database,* Cambridge, M.A.: Bradford Books, 1998.

[6] Flickr, "4,000,000,000," *Flickr Blog,* http://blog.flickr.net/en/2009/10/12/4000000000/.

[7] Flickr, "flickr.tags.getRelated," http://www.flickr.com/services/api/flickr.tags.getRelated.htm.

[8] Gong, Zhiguo, Chan Wa Cheang, and Leong Hou U, "Web Query Expansion by WordNet," DEXA 2005, Lecture Notes in Computer Science, Springer-Verlag (2005), 166-175.

[9] Goodrum, Abby and Amanda Spink, "Image Searching on the Excite Web Search Engine," Information Processing & Management, 37, 2 (2001), 295-311.

[10] Jansen, Bernard J., Amanda Spink, and Jan Pedersen, "An Analysis of Multimedia Searching on AltaVista," In Proc. SIGMM MIR 2003, ACM Press (2003), 186-192.

[11] Kamvar, Maryam, Melanie Kellar, Rajan Patel, and Ya Xu, "Computers and iPhones and Mobile Phones, Oh My! A Logs-based Comparison of Search Users on Different Devices," In Proc. 2009 International World Wide Web Conference (Madrid, Spain, April, 2009), 801-810.

[12] Lenat, "CYC: A Large-scale Investment in Knowledge Infrastructure," Communications of the ACM, 38, 11 (November, 1995).

[13] Liu, Yiming, Dong Xu, Ivor W. Tsang, and Jiebo Luo, "Using Large-Scale Web Data to Facilitate Textual Query Based Retrieval of Consumer Photos," In Proc. ACM Multimedia 2009, (Beijing, 2009), 55-64.

[14] Liu, Hugo and Henry Lieberman, "Robust Photo Retrieval Using World Semantics," In Proc. LREC2002 Workshop: Using Semantics for IR, (Canary Islands, 2002), 15-20.

[15] Metzler, Donald and W. Bruce Croft, "Combining the Language Model and Inference Network Approaches to Retrieval," Information Processing and Management Special Issue on Bayesian Networks and Information Retrieval, 40(5), (2004), 735-750.

[16] Metzler, Donald, "Indri Retrieval Model Overview," http://ciir.cs.umass.edu/~metzler/indriretmodel.html.

[17] Ren, Reede, Martin Halvey and Joemon M. Jose, "Aggregative Query Generation," IEEE International Conference on Multimedia and Expo (ICME 2009, New York, June, 2009), 850-853.

[18] Sigurbjörnsson, Börkur and Roelof van Zwol, "Flickr Tag Recommendation Based on Collective Knowledge," In Proc. 2008 International. World Wide Web Conference, (Beijing, China, April, 2008).

[19] Strohman, Trevor, Donald Metzler, Howard Turtle, and W. Bruce Croft, "Indri: A Language-model Based Search Engine for Complex Queries," In Proc. International Conference on Intelligence Analysis (McLean, VA, 2005).

[20] Zhai, Chengxiang and John Lafferty, "A Study of Smoothing Methods for Language Models Applied to Information Retrieval," ACM Transactions on Information Systems, 22, 2, (April 2004), 179-214.