# Domain-Specific Query Augmentation using Folksonomy Tags: the Case of Contextual Advertising

Andrei Broder Peter Ciccolo Evgeniy Gabrilovich Bo Pang
Yahoo! Research, 701 First Ave, Sunnyvale, CA 94089.
{broder,gabr,ciccolo,bopang}@yahoo-inc.com

## ABSTRACT

Folksonomies allow users to collaboratively tag a variety of textual and multimedia objects with sets of labels. The largest folksonomy projects, such as FLICKR and DEL.ICIO.US, contain millions of multi-labeled objects, and embed significant amounts of human knowledge. We propose a method for automatically using this knowledge to augment traditional IR systems, using contextual advertising as an application domain. Given a query, we first identify a set of relevant tags, and then use tags that cooccur with them to augment the query. Importantly, our method performs domain-specific query disambiguation, and can actually learn that a query "menu" is likely to have food connotation on FLICKR but user interface connotation on DEL.ICIO.US.

## 1. INTRODUCTION

Folksonomy is a method for assigning user-defined labels to objects stored in public repositories of textual or multimedia content. Examples of popular folksonomies include FLICKR (a photo collection), DEL.ICIO.US (a bookmark sharing project) and YOUTUBE (a video sharing system). Typically users can add tags to any object, whether they "own" it or not. Folksonomies facilitate interaction between Web users and promote knowledge sharing by integrating the user-defined tags in searching and browsing activities. In a sense, folksonomies comprise a competing approach to restricted lexicons, as the numerous labels potentially allow users to achieve higher recall. When the original content creator might not have thought of all the applicable tags, users who subsequently encounter the object are likely to add tags they deem relevant.

Some tags are automatically assigned (e.g., a FLICKR picture can be automatically labeled with the camera model and geographical location of the pictured scene), but the majority of tags are assigned manually by Web surfers. For example, a Flickr photo of an elephant could be labeled with tags such as *Thailand, Asia, colorful* and *sit on elephant back*. While some tags are only meaningful to their creator, many are useful to other users. Consequently, folksonomies encode a cornucopia of human knowledge, and in this paper we propose a method for leveraging this knowledge to achieve better focus in information retrieval.

In particular, we use co-tagging (i.e., tagging of the same object with different tags) to infer tag relatedness *in the context* of individual folksonomies. Prior studies in contextual information retrieval mainly defined context as a fragment of natural language text surrounding the object in question. We propose an alternative definition of context as a collection of tags assigned to or related to an object. Such contexts can be quite different between folksonomies, and can serve for word sense disambiguation. For instance, studying tag cooccurrence reveals that on FLICKR the word "menu" mostly refers to food and restaurants, but on DEL.ICIO.US it often describes elements of graphical user interface.

In this paper, we use sponsored search advertising as our application domain, where our aim is to match search queries in different folksonomies to most relevant textual ads. Besides the obvious commercial incentive in placing more relevant ads, judging the relevance of textual ads to a textual query is simpler than judging the relevance of, say, pictures or movies, and thus the relevance of ads provides a convenient means of validating our approach.

It is also of great interest to study the effect of returning site-specific ads for a given query. A query submitted to FLICKR most likely conveys a different intent of the user than the same query submitted at DEL.ICIO.US. That is, knowing at which site the query is submitted can help identify the search intent of user. Treating the content of the site as the context for queries and matching ads accordingly can potentially improve user experience. In the previous example, FLICKR ads for the query "menu" should ideally include offers from restaurants rather than services of UI experts, which would be more appropriate on DEL.ICIO.US.

Matching ads to short queries is challenging, and in mainstream information retrieval query expansion techniques are often used to augment queries with additional terms or concepts based on some form of relevance feedback [7, 15], dictionary lookup [14], ontological classification [4], or electronic encyclopedias [3]. However, to the best of our knowledge, no prior studies examined the use of folksonomies as an alternative source for query augmentation or explored the effect of using the content of a vertical site as the context for queries submitted to that site.

We propose a way to use tag cooccurrence statistics for site-specific query augmentation. Specifically, we use relevant tags to expand the bag of words for the query, as well as classify those tags to create new taxonomy-based features. Representing queries in this rich feature space results in more relevant ad matches, so that the ads displayed on different folksonomy sites better reflect the intent of their users. We present the results of an initial evaluation of the proposed method. The performance of our method is competitive with that of query expansion based on Web search results, and is superior to it at low recall (i.e., in the high precision region). We also analyze the difficulties of such evaluation, when judges needed to adopt the mindset of users of different folksonomies (e.g., FLICKR and DEL.ICIO.US).

# 2. BACKGROUND

### Folksonomies.

Tagging systems allow users to annotate a variety of resources with textual labels, or tags, which could be individual words or phrases [6, 5]. The term "folksonomy" is a *portmanteau* of "folk" and "taxonomy" and is due to Thomas Vander Wal [12]. Folksonomies provide a scalable way to collect metadata about objects; in fact, one of the first tagging projects, the ESP Game [13], was designed to collect tags to facilitate retrieval of images. Many folksonomies double as social networks, where users are grouped either explicitly by interests or explicitly by their tagging behavior.

### Online textual advertising.

A large part of the Web advertising market consists of *textual ads*. There are two main channels for distributing such ads. *Sponsored search* places ads on the result pages of a Web search engine, where ads are selected to be relevant to the search query. *Content match* places ads on third-party Web pages, which range from individual bloggers and small niche communities to large publishers such as major newspapers.

In this work we focus on sponsored search, where a few carefully-selected paid textual ads are displayed alongside algorithmic search results. Identifying relevant ads is challenging because a typical search query is short and because users often choose terms to optimize Web search results rather than ads. There is a fine but important line between placing ads relevant to the query and placing unrelated ads. Users often find the former to be beneficial as an additional source of information or Web navigation, while the latter annoy the searchers and hurt the user experience.

Sponsored search is an interplay of three entities. The **advertiser** provides the supply of ads; as in traditional advertising, the goal of the advertisers is to promote products or services. The **search engine** provides "real estate" for placing ads (i.e., allocates space on search results pages), and selects ads that are relevant to the user's query. **Users** visit the Web pages and interact with the ads.

Search engines select ads based on their expected revenue, computed as a probability of a click times the advertiser's bid. However, in this paper we focus on ad textual relevance only. Several prior studies examined the textual aspects of relevance in sponsored search. For instance, people have looked into predicting click through rate based on keywords in queries as well as content of ads [9, 10, 8]. To the best of our knowledge, there has not been previous work that considers the site-specific nature of ads placement.

# 3. METHODOLOGY

We now present our methodology for using folksonomies for site-specific query augmentation. The input to our system is a search query, and the output is a set of ads that are relevant to this query. Processing the input query involves two main phases. First, given a query, we identify a set of relevant tags, and then identify tags that cooccur with them. We then pool these tags together in a *context vector*, i.e., a vector of tags whose individual entries are weighted by cooccurrence frequency. Second, we use the context vector to construct an augmented *ad* query, to be executed against a corpus of ads. The features of the ad query include an augmented bag of words and a set of taxonomy classes. We now describe these two phases in detail.

## 3.1 Building context vectors

Tags used to label the same object (an image in FLICKR, or a Web page in DEL.ICIO.US) are often semantically related words or phrases, as they represent different aspects or characteristics of the same object. Tag cooccurrence information aggregated over all the objects in a folksonomy reflects site-specific relatedness as defined (and shared) by its users. In the preprocessing phase, we try to capture this information by analyzing the set of objects in a folksonomy $\mathcal{F}$, and build a tag cooccurrence matrix $M$, where $M(i,j)$ is the number of objects co-tagged with tags $t_i$ and $t_j$. To reduce noise, we ignore all cells such that $M(i,j) < 2$.

To construct the context vector for an input query, we tokenize the query into words, and then map the words into relevant tags. For each tag $t_i$, we look up its cooccurrence vector, namely, a row $M(i)$, and finally sum the retrieved vectors to obtain a single *context vector* $V$ for the query. We decimate the vector entries by retaining only the $n$ most frequently cooccurring tags ($n = 10 \dots 100$). The values of individual vector entries are assigned using the TFIDF scheme [11], with logarithmic term frequency and IDF computed over the ad corpus.

We now address two research questions involved in this process, namely, how to handle multi-word tags and queries.

### Mapping the tag space into the word space.

Many tags contain several words (e.g., "sanfrancisco" or "ToRead"). This does not pose problems for building the tag cooccurrence matrix $M$ as this type of concatenation is a convention of the tagging system (indeed, some folksonomies automatically remove white spaces in phrases for each individual tag). However, it is problematic to use such multi-word tags for query augmentation since such concatenations are not common in the ad corpus, and as a result they are unlikely to improve the ad matching process. To this end, we use a dynamic programming algorithm (based on a unigram language model trained on the ad corpus) to break tags into individual words, and update the counts in $V$ accordingly.

If a tag $t_j$ is segmented into $k$ tokens $t_{j,1}, ..., t_{j,k}$, we need to decide how to distribute the counts aggregated for $t_j$ among these tokens. We considered two different options: each token receives the same count as the original tag, or only a portion thereof. More specifically, we compute a count $c(j,p)$ for token $t_{j,p}$ based on $M(i,j)$. If we consider each of the segmented tokens as a tag in itself, then each of them would have cooccurred with $t_i$ $M(i,j)$ times, which suggests setting $c(j,p) = M(i,j)$. On the other hand, if we consider each tag to have the same importance for a given object, then each of the tokens on its own would not have cooccurred with $t_i$ with the same likelihood, and one way to approximate this is to set $c(j,p) = M(i,j)/k$. Based on examining context vectors in the development set, we implemented the second option in our system.

### Handling multi-word queries.

Building context vectors for multi-word queries is challenging, because some word combinations have meanings that are different from a simple composition of the meanings of constituent words. One possibility is to map each word into the closest tag and consider different ways to combine the context vectors retrieved for these individual tags. If we consider all the words in a query as context for each other, which can be employed to achieve further disambiguation, we should take the intersection of the vectors retrieved to represent the "common" context vector. Alternatively, if we consider each word as enrichment to other words in the query, we can take the sum over all the context vectors retrieved. The dataset we used in this work only contained a few multi-word queries, hence for simplicity we mapped each multi-word query into a single tag by taking out the white spaces. In future work, we are interested in exploring the effect of different strategies of combining context vectors where each constituent words in the query will be mapped into the closest tag instead of being concatenated into one single tag.

## 3.2 Retrieving ads

We now discuss how to use the context vector to construct an augmented *ad* query to be executed against a corpus of ads. Ad queries are represented with two kinds of features. We use feature selection to identify most salient words in the context vector $V$, and use the selected features to augment the bag of words representation of the original (short) query (with stop words removed). We also consider the context vector as a pseudo-document, and automatically classify it with respect to a large commercial taxonomy of over 6000 nodes. Previous work found it beneficial to include class information in ad retrieval [2, 8], as generalizing from individual words to classes allows one to match related queries and ads even though they might use different vocabularies. Furthermore, classifying the query context with respect to an external taxonomy introduces yet another valuable source of external knowledge. We adopted the taxonomy used in [2]; further details on the taxonomy are available therein. The 5 most relevant class nodes for each query, along with their ancestors, comprise a second kind of features. Our experiments confirmed previous work and found class information to be useful in our site-specific setting as well.

We analyze the ad text and construct the same two types of features as for queries, namely, words and classes. In an online advertising system, the number of ads can easily reach hundreds of millions, hence we use an inverted index to facilitate fast ad retrieval. Finding relevant ads for the query amounts to evaluating the scores of candidate ads, and then retrieving the desired number of highest-scoring ads. We compute query-ad scores as a linear combination of cosine similarity scores over the two feature sets.

# 4. EVALUATION

We implemented the above methodology for site-specific query augmentation in a software system called Alexandrite[1].

## 4.1 Editorial evaluation

### *Dataset.*

We evaluated Alexandrite on two actual folksonomies, Flickr and Del.icio.us, while our hypothesis was that taking site-specific tagging patterns into account would allow us to match queries on each site to more relevant ads. We constructed the dataset by taking a set of most frequent queries from each site, as well as a set of queries with most different meaning (as judged by comparing their context vectors $V$ defined in Section 3). After removing duplicates and adult queries, we ended up with 492 queries, of which about 10% contained more than one word. We held out 92 queries as a validation set to tune parameters, and the remaining 400 queries formed the test set.

### *Reference systems.*

We compared Alexandrite with two other systems. The first one was a baseline system that did not use any site-specific information and implemented a generic Sponsored Search (SS) algorithm, which expanded queries with general purpose Web search results [1]. Naturally, this baseline returns the same set of ads for both sites.

We also compared Alexandrite to a "site-aware" system, which used site-specific search results instead of those from general Web search. This approach is akin to so-called Content Match (CM) advertising scenario, where ads are matched to Web pages instead of queries. This system (referred to as CM in the sequel), used the input query to conduct a regular search on either Flickr or Del.icio.us, and then used the results page to build a rich ad query. Similarly to Alexandrite, both SS and CM systems represented ad queries and ads in the space of words and classes.

We implemented Alexandrite with the following parameters. Two parameters control the relative importance of words vs. classes in the augmented vector. We considered emphasizing only classes or words individually, as well as placing equal importance on both types of features. Another parameter controls the number of cooccurring tags to include in the context vector. We augmented queries with up to $n$ most frequently cooccurring tags from $M$, and considered $n = 10, 20, 50, 100$.

### *Judging with* Flickr *or* Del.icio.us *mindset.*

For each system, we matched each query to up to 3 ads for each of the two sites. We obtained human judgments for each query-ad pair on the following numeric scale: Perfect (0), Certainly Attractive (1), Probably Attractive (2), Somewhat Attractive (3), Probably Not Attractive (4), and Certainly Not Attractive (5). To compute the standard metrics of precision and recall, we converted the above judgments to binary by considering the first four as relevant, and the rest as irrelevant.

The query-ad pairs were judged by editors who are trained in conducting relevancy evaluations. They were not aware of the algorithmic details, and all the query-ad pairs were presented to them in random order. In order to evaluate how well our system can capture the site-specific context, we asked the editors to adopt the mindset of a typical Flickr or Del.icio.us user. Search result pages on each site were provided to help the editors better understand the scope of each site. The editors were also instructed to use Web search if they required additional information about the meaning of the query or about the products and services described in the ads.

## 4.2 Pilot study

One potential concern about the validity of our approach is that its utility may be limited by the available ad inventory. Even if our technique does model the site-specific context reasonably well, and the context vector does a good job of capturing site-specific user intent, if the ad inventory does not contain ads that reflect such differences, we will not be able to distinguish between the results produced by the different systems.

To assess the importance of this concern, we first conducted a pilot study with a set of single-word queries that exemplified different user intent in the two sites. Our goal was to verify whether there are any differences in the top ads returned for such queries in the two sites, and if so, whether the differences are consistent with an intuitive interpretation of the intentions of typical Flickr or Del.icio.us users. Table 1 presents a subset of queries with sample ads retrieved by Alexandrite. Indeed, the sample ads seem to be consistent with our intuition about Flickr as a fairly general site and Del.icio.us as a geek-oriented site with more technical content.

## 4.3 Results

Table 2 summarizes the average numeric scores for the different systems we evaluated (lower values correspond to more relevant ads and are better). For Alexandrite, the editorial judgment confirmed our expectation that the best performance is achieved by using both types of features (namely, words and classes), and taking the 50 most frequent tags for the context vectors.

| Site \ Method | SS | CM | Alexandrite |
|---|---|---|---|
| Flickr | 3.88 | 3.95 | 4.09 |
| Del.icio.us | 3.495 | 3.50 | 3.485 |

**Table 2: Average system scores (at maximum recall)**

Based on these preliminary results, Alexandrite performance is competitive with that of the two reference systems. Importantly,

[1] Alexandrite is a semi-precious stone that changes its color under different lighting conditions.

| Query | Ads for FLICKR | Ads for DEL.ICIO.US |
|---|---|---|
| menu (table) | Online Restaurant Menu / Food Service Consultant | Quickly Learn HTML Web Site Design |
| sun | Sun 'n Sea  Sunset Waters Beach Resort | Solaris (Sun) Training  java |
| fly | Fly Fishing Shop | Blue sky air / airplane |
| mouse | Disney Mickey Mouse Items | Wireless Keyboard |

**Table 1: Sample Alexandrite output**

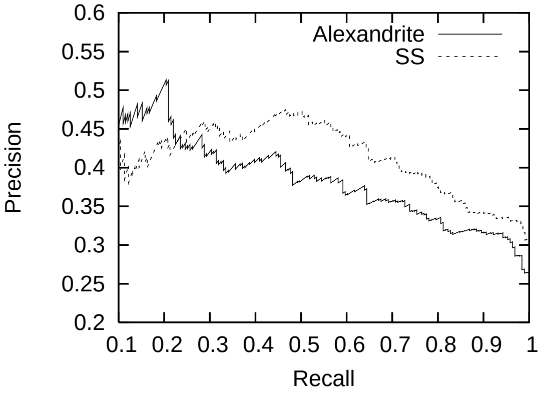

**Figure 1: Alexandrite vs. SS**

Alexandrite performs most tag cooccurrence analysis in the preprocessing phase, and is thus more efficient than both SS and CM, which involve a query-dependent search on the Web or the folksonomy.

Table 2 compares the systems at maximum recall. On a precision-recall graph produced by thresholding the ad retrieval scores (Figure 1), we observed that in the low recall (= high precision) range, the precision of our algorithm is superior to that of SS. We also experimented with different lengths of the context vector ($n = 10 \dots 100$ tags), and predictably found $n = 50$ to yield optimal results. Lower values of $n$ under-utilized available context, and higher values resulted in using less reliable tags owing to noise (we omit the graph for lack of space).

It is essential to note that the editors reported the task of adopting the mindset of FLICKR and DEL.ICIO.US users to be quite difficult, which partly explains why in our preliminary evaluation Alexandrite did not definitively outperform the baselines. For instance, for the query "Antarctica" on DEL.ICIO.US, Alexandrite returned a Web design ad, which was judged as Certainly Not Attractive. However, this particular ad offered the services of Antarctica Media company, which specializes in Web design, and hence should have arguably been scored much better. While FLICKR content is quite general, DEL.ICIO.US caters to the tech-savvy geek community, hence adopting the mindset of DEL.ICIO.US users was particularly difficult.

Also noteworthy is the disparity of scores for the SS system on the two sites (see Table 2). This system expanded queries using general Web search results (without any site-specific information), and hence we would expect its output ads to be more relevant for the more general FLICKR site. However, its ads have been judged more relevant (= lower score) for DEL.ICIO.US, which again reinforces our concern about the difficulty of judgment by adopting a particular mindset. Furthermore, some queries are indeed hard to judge for non-expert users. In our future work, we plan to improve our judgment procedure, and also evaluate the system by conducting an experiment with actual users, measuring the relevance of ads by actual click-through rates.

# 5. DISCUSSION

We proposed a methodology for using folksonomy tags for query augmentation in one IR task (sponsored search advertising). Our approach leverages co-tagging data to capture site-specific query intent and to disambiguate polysemous queries. Although we focused on sites with rich tagging information, the methodology proposed could also be applied to other sites by modeling site-specific distribution of words. Our initial evaluation confirmed that the proposed method is competitive with another system that performs query augmentation based on site-specific search results (CM). We also discussed inherent judging difficulties when editors are asked to adopt mindsets of typical users of particular Web sites. In our future work, we plan to further refine our method and to revise the editorial evaluation, as well as to perform a real-life evaluation of Alexandrite with actual folksonomy users and evaluate the system with user-generated click data.

# 6. ACKNOWLEDGMENTS

We thank Vanja Josifovski, Ravi Kumar, and Malcolm Slaney for fruitful discussions, technical assistance, and pointers.

# 7. REFERENCES

[1] Andrei Broder, Peter Ciccolo, Marcus Fontoura, Evgeniy Gabrilovich, Vanja Josifovski, and Lance Riedel. Search advertising using Web relevance feedback. In *CIKM'08*, 2008.

[2] Andrei Broder, Marcus Fontoura, Vanja Josifovski, and Lance Riedel. A semantic approach to contextual advertising. In *SIGIR'07*, pages 559–566. ACM Press, 2007.

[3] Ofer Egozi, Evgeniy Gabrilovich, and Shaul Markovitch. Concept-based feature generation and selection for information retrieval. In *AAAI'08*, July 2008.

[4] Evgeniy Gabrilovich and Shaul Markovitch. Harnessing the expertise of 70,000 human editors: Knowledge-based feature generation for text categorization. *JMLR*, 8:2297–2345, October 2007.

[5] Scott Golder and Bernardo Huberman. The structure of collaborative tagging systems. Technical report, HP Labs, 2005.

[6] Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, ToRead. In *HT'06*, 2006.

[7] Mandar Mitra, Amit Singhal, and Chris Buckley. Improving automatic query expansion. In *SIGIR'98*, 1998.

[8] Filip Radlinski, Andrei Broder, Peter Ciccolo, Evgeniy Gabrilovich, Vanja Josifovski, and Lance Riedel. Optimizing relevance and revenue in ad search: A query substitution approach. In *SIGIR'08*, 2008.

[9] M. Regelson and D. Fain. Predicting click-through rate using keyword clusters. In *Second Workshop on Sponsored Search Auctions*, 2006.

[10] Matthew Richardson, Ewa Dominowska, and Robert Ragno. Predicting clicks: Estimating the click-through rate for new ads. In *WWW'07*. ACM Press, 2007.

[11] Gerard Salton and Chris Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[12] Thomas Vander Wal. Folksonomy definition and Wikipedia. `http://www.vanderwal.net/random/entrysel.php?blog=1750`, 2005.

[13] Luis von Ahn and Laura Dabbish. Labeling images with a computer game. In *CHI'04*, 2006.

[14] Ellen M. Voorhees. Using wordnet for text retrieval. In Christiane Fellbaum, editor, *WordNet, an Electronic Lexical Database*. The MIT Press, 1998.

[15] Jinxi Xu and W. Bruce Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM TOIS*, 18(1):79–112, 2000.