iTag: A Personalized Blog Tagger

Michael Hart
Stony Brook University
mhart@cs.sunysb.edu

Rob Johnson
Stony Brook University
rtjohnso@cs.sunysb.edu

Amanda Stent
Stony Brook University
stent@cs.sunysb.edu

ABSTRACT

We present iTag, a personalized tag recommendation system for blogs. iTag improves on the state-of-the-art in tag recommendation systems in two ways. First, iTag has much higher precision and recall than previously proposed tagging algorithms. For example, iTag achieved over 60% precision and recall on a set of 1000 blog posts selected at random from a WordPress[4] RSS feed in April 2009, whereas the previously-developed TagAssist[10] achieved less than 10% precision and recall on our data. Second, iTag performs just as well when trained on a single user's blog as when trained on a large corpus of blogs. Thus, iTag can be deployed as a global, non-personalized tag recommendation system, or as a personalized tag recommender. Our experiments and survey of tagging behavior suggest that bloggers use tags idiosyncratically, so personalized tagging is an important option.

Keywords

Tagging, Blogs, Machine Learning

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.5 [**Pattern Recognition**]: Applications

1. INTRODUCTION

Tags (words and phrases that annotate content) are widely used in web-logs (blogs). Blog tags are used in several ways: the blogger may use them for post categorization/topic identification; the reader may use them to get a quick idea of the content and orientation of a blog and for post search and retrieval, and the blogging system may use them for blog layout, post indexing, and blog recommendation services.

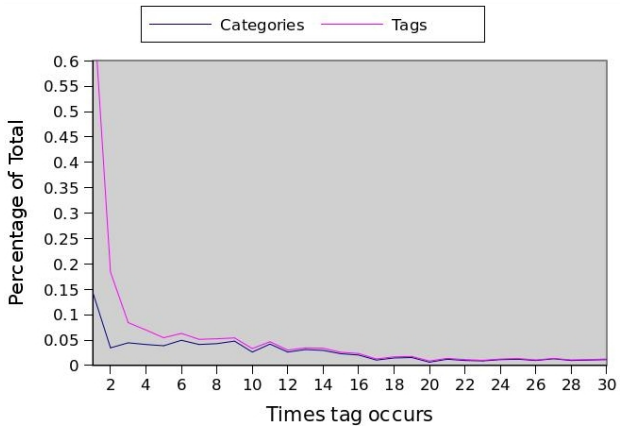
Despite the popularity of blogs and prevalence of tagging, the most popular blogging services do not offer tag suggestion features. Researchers have developed tools for tag suggestion, but most of these tools focus on *social tag prediction*. Social tags are the most interesting or informative tags derived (via aggregation, scoring and filtering) from all tags assigned by multiple users of an online community (e.g. del.icio.us, StumbleUpon, Digg) to a content item. Social tags allow members of the online community to share and interact more effectively [8]. However, they do not reflect the mental model of the content author.

There are two general approaches used in tag suggestion tools for blog posts. The first extracts interesting terms from the post itself [5]. This approach is useful for post clustering, but often author-assigned tags consist of terms not in the post itself (e.g. the date of creation of the post, a content category, a location for a piece of content in the post). The second approach uses search and scoring over a large collection of posts [9]. For example, the TagAssist tool tags a post by constructing search queries from the post content, searching a collection of blog posts using those queries, extracting the author-assigned tags from the retrieved posts, and scoring and filtering those tags [10]. In a series of evaluations on blog data, TagAssist was shown to perform well. However, it is not personalized or localized: it weights tags preferentially if they occur on more popular posts or if they occur on a larger number of posts, but it does not give preference to posts or tags by the same author, or to posts or tags that occur near each other in time.

Many blog post tags are not contained in the post text. Bloggers may label their posts with category tags ("Categories" in WordPress), topical tags, dates, and locations. For example, on the widely-read "Get Rich Slowly" blog[1], there are some category tags ("Administration", "Ask the Readers"), and some topical tags ("Cars", "Credit Cards"). However, most bloggers reuse tags. In this work, we present a tag suggestion tool that focuses on *personalized and localized tag recommendation*. Our method suggests only tags previously used by this blogger. It also incorporates temporal information when selecting tags to suggest.

Our experiments demonstrate that a personalized tagger can substantially outperform a non-personalized tagger. We compared the performance of iTag and TagAssist on a random subset of posts drawn from the "Growing Blogs" Wordpress feed. This feed includes blogs that have had a recent increase in popularity, and therefore tends to contain reasonably well-written posts and few spam blogs. On this data set, iTag achieved precision and recall scores over 60%, while TagAssist had scores below 10%.

The rest of this paper is structured as follows: In Section 2 we present what is to the best of our knowledge the first



analysis of tag usage on a per-blogger basis. In Section 3 we describe a novel and lightweight machine learning technique for tag suggestion that uses blogger identity and temporal information as features. In Section 4 we describe an evaluation of our approach and compare its performance with that of TagAssist. We observe dramatic increases in both recall and precision on tagging posts, even when tags that only occur once are taken into consideration. This suggests that personalized approaches to blog post tag suggestion are more effective than global approaches.

2. TAG USAGE ON INDIVIDUAL BLOGS

We investigated tag usage on individual blogs by downloading 1246 blogs from the Wordpress "Growing Blogs" feed[2] during April 1 to April 25. The "Growing Blogs" feed[2] collects blogs that have experienced a surge in viewership over the last 24 hours. We chose this feed because it rarely contains "spam blogs". We chose WordPress because it allows bloggers to separate tags into two types: tags and categories. Intuitively, bloggers may separate their tags so that the more frequent or taxonomic tags are designated as categories. 84.85% of the blogs in our dataset utilized categories.

We measured the frequency of use of each tag and category by individual bloggers, then aggregated across all the blogs we analyzed. The resulting distribution of tag frequencies is shown in Figure 1. As one would expect, bloggers assign categories more frequently than tags. We note that most tags occur only once per blog (58.82%). On the other hand, 51.2% of categories occur more than ten times per blog. On average, bloggers assign 10.2 tags to each post, but only 2.2 categories. Although most tags occur only once, most posts contain 9.37 old tags, i.e. all but one of the tags are re-used. Bloggers also reuse 2.03 categories on average, introducing a new category about once every 6 posts. 81.4% of posts use only pre-existing categories, whereas 23.8% of posts use only pre-existing tags.

We also measured individual bloggers' consistency in tag use, as this is an indicator of how well a tag suggestion tool can perform. For each post, we determined the minimum number of previous posts such that the set of tags aggregated from these previous posts provides total recall for the tags that have occurred at least once before and are assigned to the target post. If bloggers use tags haphazardly, we would expect many posts to be required to provide total recall. For this analysis, we selected a subset of 100 blogs that each contained 50 to 150 posts.

Approach	Average Number of Posts	Recall
All previous posts	1.38	1.0
5 most similar posts	1.20	0.91
10 most similar posts	1.27	0.94
10 most recent posts	1.22	0.92
20 most recent posts	1.28	0.95

Table 1: Consistency in tag assignment

Finding the minimum number of posts to provide total recall is an instance of the set-cover problem. We used a linear programming solver. We investigated several different strategies for previous post set selection. First, we selected all previous posts regardless of order. Second, we selected the n most recent previous posts (by posting date order). Third, we selected the n most similar previous posts (as determine using cosine similarity).

Results are shown in Table 1. We see that for most bloggers use of tags is highly local, both with respect to topic and with respect to time.

Our analysis of tag usage on individual blogs suggests that a personalized and localized tag suggestion tool can be highly effective. First, more than 40% of tags and 85% of categories are reused by individual bloggers. In fact, the majority of tags and categories on each post have been used before. Our data suggests that a local approach can have good recall even though most tags are unique and it suggests only previously used tags. Second, a relatively small post history (10 posts) gives high recall for previously used tags, indicating that bloggers cluster related posts together in time. At the same time, a high recall for previously used tags based solely on similar posts by the same blogger indicates that document similarity can continue to be a useful feature for personalized tag suggestion.

3. iTag DESCRIPTION

iTag uses a modified form of the approach taken in TagAssist [10]. We start by summarizing TagAssist's search and score strategy. We then highlight key differences in iTag.

3.1 TagAssist

When TagAssist is presented with a target post, it selects tags to suggest using a *search and score* method on a large set of training data.

Training Data TagAssist uses a corpus of blog posts and tags indexed with Lucene[3]. As part of indexing, TagAssist normalizes tags by trimming white space and punctuation, stemming each word in each tag, and ordering the words in multi-word tags alphabetically. It then clusters the normalized tags using tag co-occurrence information, to get a minimal set of semantically distinct tags.

Tag Retrieval TagAssist generates a query of up to 30 unigrams and bigrams from the target post that have high TFIDF scores in its training corpus. It retrieves up to 35 result posts from its index for this query, and retrieves the tags for each result post. Tags that occur on only one result post are discarded.

Tag Scoring TagAssist scores the retrieved tags using a weighted sum over the following features: frequency (in the bag of retrieved tags), text occurrence (in the target post), tag count (frequency in the training corpus), rank (popularity of the blog containing the retrieved post labeled with the

retrieved tag), and co-occurrence with other retrieved tags (in the training corpus).

Tag Selection TagAssist suggests tags that score above the average of all tag scores.

3.2 iTag

iTag also uses a *search and score* method. However, there are significant differences from TagAssist in each part of the system.

Training Data iTag only considers previous posts written by this blogger, and consequently, only tags previously used by this blogger.

Tag Retrieval iTag generates a query in the same way that TagAssist does. However, iTag only retrieves the 10 most similar previous posts by this blogger (which achieved recall of .94 in the analysis described previously). iTag does not remove any retrieved tags from consideration, regardless of their absolute frequency of occurrence.

Tag Scoring iTag does not use the rank, tag count, or co-occurrence features used by TagAssist. Instead, it uses the following features which we have found provide the best information for tag suggestion with respect to previous posts:

- **Count** : the number of times the retrieved tag appears in the retrieved tag set.
- **Highest rank** : the rank of the retrieved post labeled with this retrieved tag that is most similar to the target post.
- **Contained** : 1.0 if the retrieved tag appears in the target post, 0.0 otherwise.
- **Tag.ITF** : a variation on TFIDF, this feature is the number of times the retrieved tag appears in the retrieved tag set, divided by the number of times the retrieved tag is used to label previous posts by this blogger.
- **Last Recently Used** : distance in number of posts since this tag was last used by this blogger. This feature draws on the work of Cattuto et al. [7] which rewards tags that have occurred recently, and is motivated by the analysis in Section 2.

Tag Selection We devised two methods for tag selection. Our first method, *adaptive co-occurrence* tag selection, is a modification to the TagAssist tag selection method. Our second method, *classification-based* tag selection, uses a binary classifier trained on the features described above.

Adaptive Co-occurrence Tag Selection iTag suggests any tag that scores above average. It also suggests any tag a that strongly co-occurs for this blogger with a tag b that scores above average, as long as

$$\frac{co-occurence(a,b)}{min(count(a),count(b))} \geq 0.35$$

Our testing indicates that a co-occurrence threshold of 35% provides good tag suggestions. Every time a blogger makes a change in tag assignments, the co-occurrence frequencies for that blogger are updated.

Classification-Based Tag Selection We apply binary classifiers, using the above features, for tag suggestion using the C4.5 decision tree implementation provided by WEKA [11] (for our task, decision trees outperformed SVMs, Naive Bayes, and nearest neighbor classification algorithms). The label for each tag was 1 if the tag was used by the blogger for his/her post, and 0 otherwise.

iTag takes two approaches to classification-based tag suggestion. In the *pre-trained* approach, we train the classifier on all the tags in our training data regardless of blogger of origin. In the *locally-trained* approach, by contrast, we train a separate classifier for each blog. The pre-trained approach is faster since the classifier only has to be trained once. The locally-trained approach must be retrained each time the blogger makes a change in tag assignments. Note that even when using the pre-trained classifier, iTag is still personalized since it only chooses tags from the blogger's previous posts.

If the classification-based method produces no tag suggestions, iTag backs off to the adaptive co-occurrence method.

4. EVALUATION AND RESULTS

Our evaluation uses the same set of blogs described in Section 2. We eliminated blogs that contained fewer than 30 posts, since these blogs were likely created by new bloggers not yet familiar with post authoring and tagging. We eliminated posts that contained fewer than 10 non-stopwords because they had little information or contained only photos or links. We did not distinguish between tags and categories, but we removed multiple occurrences of a single tag on a single post.

We created our own implementation of TagAssist to compare with our approach. We trained our implementation using the ICWSM 2009 dataset [6]. This dataset consists of of 6.9 million posts with 1.7 million unique tags and 1.4 million TagAssist-normalized tags. We chose this data set because it is the only data set we could find with the popularity information required by TagAssist.

We set aside 400 randomly-selected blogs for testing data. We used the remaining blogs to create the pre-trained classifier. When evaluating the locally-trained classifier on a post, we train it on all the preceding posts in the same blog.

As testing data, we used the 400 blogs set aside earlier. We then sampled 1000 posts from the last 20% of posts in each blog, with no blog contributing more than 3 posts.

For tag suggestion, we report average precision and recall across all 1000 testing posts. Precision and recall for each post are normalized using the number of tags applied to the post (for recall) and suggested (for precision). We report results separately for all tags, and for tags the blogger uses more than once. Our evaluation results are presented in Figure 1.

4.1 Discussion

iTag achieves high precision and recall both for all tags, and for those the blogger uses more than once. This indicates that it is a useful method for tag suggestion which could be deployed almost immediately.

Both classification-based approaches outperform the adaptive co-occurrence method. Surprisingly, the pre-trained approach performs almost as well as the locally-trained approach. We note that precision and recall for tags that occur more than once when the classifier suggested at least one tag were 77.88% and 67.77% for the locally-trained approach and 79.17% and 63.36% for the pre-trained approach. This means that although the classification-based method gives higher precision, it may fail altogether, and the adaptive co-occurrence method gives some robustness.

iTag performs substantially better than our TagAssist implementation in terms of both precision and recall. However, the precision and recall scores of our TagAssist implementa-

Tag Frequency	All Tags	Adaptive	Classification-based		TagAssist
		Co-occurrence	Locally-trained	Pre-trained	
Every tag on the post	0.34 (0.77)	0.48 (0.57)	0.67 (0.59)	0.64 (0.56)	0.02 (0.07)
Occur more than once	0.41 (0.84)	0.52 (0.62)	0.67 (0.63)	0.66 (0.60)	0.02 (0.08)

Figure 1: Precision and recall for tag suggestion method with respect to 1000 tagged WordPress posts from April 2009. All Tags refer to the entire set of tags returned by our search results.

tion are less than half those reported by the TagAssist creators. There are several possible explanations for this. First, many of the tags suggested by our TagAssist implementation were reasonable. However, as our tag usage analysis showed, tag assignment is highly personal. Second, in the original TagAssist evaluation, a set of contemporaneous blogs was used for training and testing, and some of the same blogs were used for training and testing. We did not include other posts by the same blogger or from the same time period in the training corpus for our TagAssist implementation. Third, the blogs used in the original TagAssist evaluation came from Technorati and appear to be news and technology oriented, while many of the blogs in our dataset were personal and had a wider variety of themes. Fourth, many normalized tags occurred relatively rarely in the ICWSM dataset (63% of the tags on the test posts occurred on 10 posts or fewer in the ICWSM dataset). Consequently, the removal of tags that occurred only once in the retrieved tag set may have adversely affected the precision and recall of TagAssist. Curiously, we also found that TagAssist's tag normalization yielded clusters that were not semantically similar. For example, "Cheney", "Cheetos", "Chi" and "Children" all normalize to **Ch??**.

Both iTag and TagAssist may face problems of scale. For TagAssist, the problem is related to storing tag co-occurrence information. The 1.7 million tags in the ICWSM dataset yielded 20 million instances of tag co-occurrence. Without aggressively caching this data, searching for co-occurrence data can be incredibly costly. For iTag, the problem is related to storing classifiers or adaptive co-occurrence features for each blog; however, the tag co-occurrence problem is reduced when only per-blog tag co-occurrences need to be stored.

5. FUTURE WORK

We aim to improve our system in the following ways. First, we will try to better suggest tags that occur once or very infrequently. By only focusing on tags seen before in an individual blog, we cannot anticipate new tags. Therefore, we are looking to find a way to associate local tags with tags used elsewhere in the blogosphere. Also, the actual word space of a blog is relatively small when considering the size of most corpora used in natural language processing and machine learning. A new post will most likely contain many words that do not exist in previous posts. Using dimensional reduction techniques to capture latent connections between new and old vocabulary could improve local search dramatically. Lastly, we are interested in learning tag suggestions from the the blogger's behaviors. Bloggers may be partly influenced in their tagging suggestions by current events, their friend's/popular blogs or tagging styles. Also, feedback from the blogger could prove useful in crafting recommendations.

6. CONCLUSIONS

We have described iTag, a personalized and localized tag suggestion tool motivated by analysis of bloggers' post tagging behavior. We have demonstrated that iTag outperforms taggers trained on large multi-blog, multi-tag data sets. We also demonstrated that incorporating machine learning leads to improved tag suggestion with minimal cost.

7. REFERENCES

[1] Get rich slowly. `http://www.getrichslowly.org/blog/`.

[2] Growing blogs – wordpress.com. `http://botd.wordpress.com/growing-blogs/`.

[3] Welcome to lucene! `http://lucene.apache.org/`.

[4] Wordpress. `http://wordpress.com/`.

[5] Christopher H. Brooks and Nancy Montanez. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 625–632, New York, NY, USA, 2006. ACM.

[6] K. Burton, A. Java, and I. Soboroff. The ICWSM 2009 Spinn3r Dataset. In *Proceedings of the Third Annual Conference on Weblogs and Social Media (ICWSM 2009)*, 2009.

[7] Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero. Collaborative tagging and semiotic dynamics. *CoRR*, abs/cs/0605015, 2006.

[8] Paul Heymann, Daniel Ramage, and Hector Garcia-Molina. Social tag prediction. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 531–538, New York, NY, USA, 2008. ACM.

[9] Gilad Mishne. Autotag: a collaborative approach to automated tag assignment for weblog posts. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 953–954, New York, NY, USA, 2006. ACM.

[10] Sanjay Sood, Sara Owsley, Kristian Hammond, and Larry Birnbaum. TagAssist: Automatic Tag Suggestion for Blog Posts. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007)*, 2007.

[11] Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition, 2005.