AutoTag: A Collaborative Approach to Automated Tag Assignment for Weblog Posts

Gilad Mishne
ISLA, University of Amsterdam
Kruislaan 403, 1098SJ Amsterdam, The Netherlands
gilad@science.uva.nl

ABSTRACT

This paper describes AutoTag, a tool which suggests tags for weblog posts using collaborative filtering methods. An evaluation of AutoTag on a large collection of posts shows good accuracy; coupled with the blogger's final quality control, AutoTag assists both in simplifying the tagging process and in improving its quality.

Categories and Subject Descriptors

H.3.4 [**Information Systems**]: Systems and Software

General Terms

Human Factors, Languages

Keywords

Blogs, tags

1. INTRODUCTION

Tagging is an old-new method for organizing data by assigning descriptors to documents and other sources of information. These descriptors, or "tags", are typically short textual labels, which provide an easy way to categorize, search, and browse the information they describe. The annotation of documents with keywords is nothing new by itself, but a collaborative form of this method with some unique properties is attracting a lot of attention on the web in recent years. The main characteristics of collaborative tagging differentiating it from traditional keyword annotation are its open-vocabulary, non-hierarchical nature, and the fact that tags are assigned by authors and users of the information rather than by professional annotators, with no rules or limitations [1, 2]. Tagging is particularly popular in some web mediums such as photo sharing websites (e.g., Flickr) and the blogosphere, where tags are often assigned to weblog posts to facilitate categorization and filtering.

This paper addresses the task of automatic assignment of tags to weblog posts; while some work on weblog classification exists, we are not aware of published work about tag discovery. To this end, we describe a system—AutoTag—that, given a weblog post, offers a small number of tags which seem useful for it; the blogger then reviews the suggestions, selecting those which she finds instrumental. More than just simplifying the tagging process, AutoTag also improves its quality: first, by increasing the chance that weblog posts will be tagged in the first place, and second by offering relevant tags that may have not been applied otherwise. This in turn improves the tasks for which tagging is aimed at, providing better search and browse capabilities.

WWW 2006, May 22–26, 2006, Edinburgh, Scotland.
ACM 1-59593-332-9/06/0005.

2. TAG ASSIGNMENT

Our basic approach to automated tag assignment is that of collaborative filtering, or recommender systems [3]. In a nutshell, a recommender system helps users find desirable products or services by analyzing their profile and matching it with profiles of other users similar to them, or by finding products that are similar to the ones they expressed interest in; it is assumed that similar users share similar tastes. Amazon, TiVo, Netflix and others are among the many successful applications of commercial recommender systems.

An application of collaborative filtering methods to automated tag discovery becomes clear when the "user" and "product" concepts are examined from a different angle. In AutoTag, the blog posts themselves take the role of users, and the tags assigned to them function as the products that the users expressed interest in. In traditional recommender systems, similar users are assumed to buy similar products; AutoTag makes the same assumption, and identifies useful tags for a post by examining tags assigned to similar posts. Just as with traditional recommender systems, the recommendations are then further improved by incorporating external knowledge about the bloggers, the posts, or the tags.

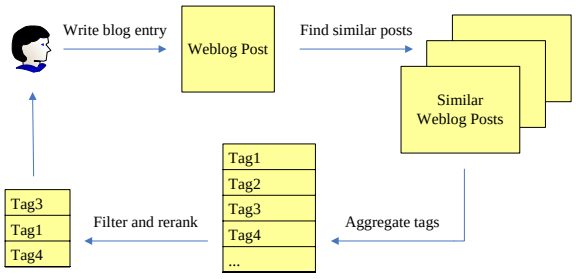

Figure 1: Flow of information in AutoTag

The different stages of the tag suggestion process in AutoTag are shown in Figure 1. Once the user supplies a weblog post, posts which are similar to it are identified. Next, the tags assigned to these posts are aggregated, creating a ranked list of likely tags. In the next stage, AutoTag filters and reranks this tag list; finally, the top-ranked tags are offered to the user, who selects the tags to attach to the post. We follow with additional details about each step.

Finding Similar Posts. AutoTag uses Information Retrieval measures to estimate the similarity between weblog posts. In practice, this means a large collection of posts is indexed by an IR engine, and a query generated from the original post is submitted to this engine. The most similar posts are then taken to be the highest-ranking ones retrieved from the collection, using some retrieval model.

We experimented with a number of methods for generating queries from a post, including using the entire text of the post and using links in it to locate cocitations. The best results were obtained by using a "distinctive term" query: standard corpus comparison methods are first used to derive the most distinctive terms in the vocabulary of a post (compared to the general vocabulary in the corpus); the top-ranking terms are in turn used as the query.

A Tag Model. AutoTag uses simple heuristics for composing the ranked list of tags from the top-retrieved posts: each tag is scored according to its frequency in the top results. Experimenting with more complex ways of scoring the tags, taking into account the retrieval rank or score, yielded only minor improvements in accuracy.

Filtering and Reranking. One clear source of information we have about a blogger is the tags she already used prior to writing the post being analyzed. Therefore, if such previously-used tags appear in the ranked list, AutoTag boosts their score by a constant factor.

3. EVALUATION

For evaluating our method, we used the corpus distributed by Intelliseek for the 3rd Weblogging Workshop.[1] The corpus contains 10M weblog posts collected during a period of 3 weeks; of these, 1.8M posts are tagged, with a total of 2.3M tags. For indexing and retrieval, we used the open source engine Lucene which uses a rather simple vector space retrieval model; text was stemmed with an English stemmer.

Two methods were used to evaluate the effectiveness of the tag suggestion methods. First, we manually examined the tags assigned to a random subset of 30 posts from our collection; for each tag, we decided whether the tag was indeed a relevant label for the post. This is the preferred method of evaluation, but due to its cost it can only be applied to a small number of posts; additionally, it is difficult to use non automated methods to tune and improve a system. Because of this, we used an automated method to evaluate a much larger subset of posts: AutoTag was used to tag 6000 of the "tagged posts" in our corpus - the posts which were assigned tags by their authors (we used only posts with 3 or more tags). Then, AutoTag's output was compared to the actual post tags. To account for minor differences in tags ("blogs" and "blogging"), we used string distance to compare the tags rather than exact string match. Even so, the automated precision scores are lower than manual ones, since tags which are useful for a post but were not originally used by its author are mistakenly taken to be incorrect. To demonstrate this, we evaluated the small test set with the automated method as well, resulting in substantially lower scores; this indicates that the actual performance of AutoTag on the large set is likely to be much better than reported by the automated evaluation.

For the manual evaluation, we measured precision at 10: the fraction of tags out of the top-10 suggestions by AutoTag which were judged as appropriate for the post by a human; only the first 10 results are checked because it is assumed that users are unwilling to examine long result lists. For the automated evaluation, we measured recall at 10 as well: the fraction of tags offered by AutoTag in the top-10 suggestions which were also assigned by the blogger out of the total number of tags assigned by her.

Test Set	Evaluation	Precision@10	Recall@10
30 posts	Automated	0.38	0.47
30 posts	Manual	0.59	—
6000 posts	Automated	0.40	0.49

Table 1: Auto-tagging accuracy

Results are shown in Table 1; an example of tags offered by AutoTag is given in Table 2. On average, 4 to 6 suggestions out of AutoTag's top-10 suggestions are either considered useful by the blogger, or were actually used by her for the given post. Cursory examination of posts for which AutoTag scores low shows many non-English posts (for which much less data exists in the corpus, entailing lower success of data-driven methods), and many tags which are highly personal and used by few bloggers (such as names of family members).

`http://www.stillhq.com/diary/000959.html`
On pitching products to bloggers Anil comments on how to pitch a product to him as a blogger, and makes good suggestions as Lindsay agrees before getting distracted by how this applies to press releases. I have to wonder though how much of this promotional pitching actually happens. I certainly haven't ever had a product pitched to me for this site. I've had people pitch advertising, and other spammy things, but not products. Does it really happen to us normal people bloggers?
Suggested tags: PR, blogging, weblogs, marketing, net, gripes, email, small business life, Anil Dash, PR pitching *Original tags: blog, pitch, product, marketing, communications*

Table 2: Sample tags suggested for a post

4. CONCLUSIONS

We proposed and evaluated AutoTag, a tool for tagging weblog posts based on a collaborative filtering approach. AutoTag offers suggestions for tags based on tags assigned to other, similar posts; the final decision about a tag is left to the blogger. Despite a relatively small corpus for this type of task, AutoTag shows good results, and has the potential to benefit both the bloggers and others making use of tags assigned to weblog posts.

The different components of AutoTag provide fertile ground for further work: identifying effective ways to generate queries from a post and successful retrieval models to use; improving the aggregation of tags from the retrieved posts; and various methods for filtering and reranking the lists produced by AutoTag. In addition to the collaborative approaches described in this paper, we are currently investigating a "local" approach to tag suggestion, in which suggestions for tags are made without access to the entire blogosphere as is the case with AutoTag, but using deeper analysis of the contents of the post and the blog it belongs to.

5. REFERENCES

[1] S. Golder and B. A. Huberman. The structure of collaborative tagging systems. *J. Inf. Science*, 2006.

[2] A. Mathes. Folksonomies: Cooperative classification and communication through shared metadata. *Computer Mediated Communication*, 2004.

[3] P. Resnick and H. R. Varian. Recommender systems (special section). *Comm. of the ACM*, 40(3), 1997.

[1] http://www.blogpulse.com/www2006-workshop