Plurality: A Context-Aware Personalized Tagging System

Robert Graham
Dep't of Computer Science
Texas A&M University
College Station, TX 77843
rbgraham@tamu.edu

Brian Eoff
Dep't of Computer Science
Texas A&M University
College Station, TX 77843
bde@cs.tamu.edu

James Caverlee
Dep't of Computer Science
Texas A&M University
College Station, TX 77843
caverlee@cs.tamu.edu

ABSTRACT

We present the design of Plurality,[1] an interactive tagging system. Plurality's modular architecture allows users to automatically generate high-quality tags over Web content, as well as over archival and personal content typically beyond the reach of existing Web 2.0 social tagging systems. Three of the salient features of Plurality are: (i) its self-learning and feedback-sensitive capabilities based on a user's personalized tagging style; (ii) its leveraging of the collective intelligence of existing social tagging services; and (iii) its context-awareness for optimizing tag suggestions, e.g., based on spatial or temporal features.

Categories and Subject Descriptors: H.3.1 Information Storage and Retrieval: Content Analysis and Indexing

General Terms: Algorithms, Design

Keywords: tags, social annotation, context-sensitive, personalization

1. INTRODUCTION

Tags – words or phrases that serve as informal metadata for objects like Web pages, images, and videos – have grown in popularity and purpose in the last few years. Tagging as a phenomenon corresponds with a Web 2.0 mentality that users can create not only content but a richer, more adaptive and responsive way to navigate and search both existing and new media. Widespread social tagging promises better and more intuitive information access through tag-based browsing (e.g., [1]), search (e.g., [4]), and new applications centered around the emergent semantics inherent in the aggregation of the tagging habits of thousands (or millions) of users (e.g., [2]). In contrast to traditional metadata annotation by experts, tagging can overcome less precision in individual tags (e.g., through misspellings, spam tags, and off-topic tags) through the sheer volume of tags that can be generated for an object.

Our research goal is to study how the underlying tag generation processes can be applied in domains either lacking a wide-scale audience (which are typically assumed in Web 2.0 social tagging contexts) or lacking a tagging-savvy audience. How can we take advantage of new approaches to tag browsing, tag search, and emerging tag-based information access approaches over documents currently "left out" of the Web 2.0 social tagging phenomenon? For example, few, if any, of the local documents on a user's desktop are exposed to a Web-scale audience for tagging, and users are typically resistant to go back through their archives to manually apply tags. Similarly, a huge amount of untagged content can be found in internal company email and document sharing networks, archival content in digital libraries, and even on the Web, where most content has yet to be tagged. Even for Web content that has already been tagged, a particular user's personalized view over the Web content may not be reflected in the existing tags.

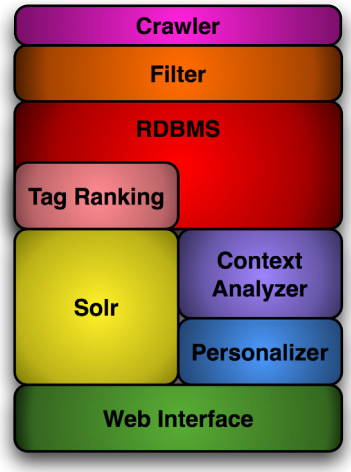

Figure 1: High-level system architecture for Plurality

2. DESIGN AND ARCHITECTURE

With these challenges in mind, we introduce Plurality – an interactive tagging recommendation system (see Figure 1). Plurality is implemented using Apache's Solr – a web services stack built over the Lucene search engine – to provide real-time tag suggestions. Solr is finding increasing use and traction among institutional users seeking to create in-house solutions for indexing large catalogs of content; it is appealing as a tagging solution since it supports web services and lightweight web interfaces for easily adapting Plurality to both institutional and personal settings with only slight modifications. Building on previous work studying

[1] http://faculty.cs.tamu.edu/caverlee/plurality

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

Figure 2: Blog entries tagged using Plurality and compared with the blogger's original tags. Obtained with permission from Jason Kottke, kottke.org.

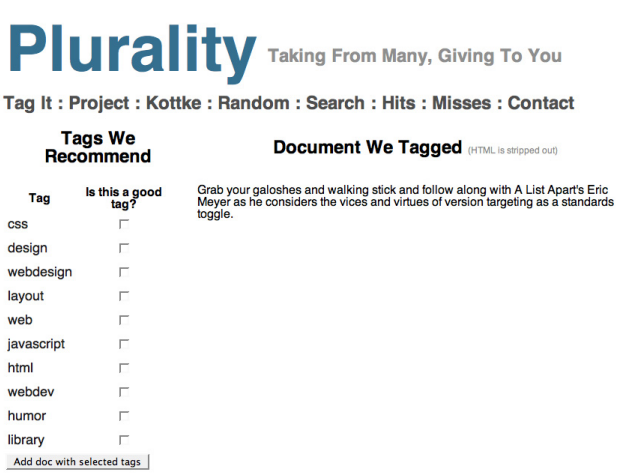

Figure 3: Incorporating user feedback. A snippet of text from A List Apart's RSS feed, alistapart.com.

the automatic generation of tags (e.g., [3],[5]), Plurality is distinguished by three salient features:

1. Leveraging Existing Tags: To bootstrap the tagging process, Plurality can be seeded with already tagged documents from an existing tagging service. In our first prototype, we have crawled the popular social bookmarking site del.icio.us and collected over 280,000 tags. The crawler is built in Python and is designed to allow for flexible adaptations to other tagging resources. The crawled documents are filtered through a spam detector, normalized via stemming and HTML stripping, and indexed by Lucene. In our initial design, tag recommendations for an untagged document are generated by finding the top-10 most similar documents, ranking their tags based on TF-IDF measures across the tag corpus and on the user's tag profile. After the initial bootstrapping, the system can persist and grow without further use of the crawler. Figure 2 shows an example of Plurality's tag suggestions for two posts by the prolific blogger Jason Kottke versus the actual tags assigned by Kottke himself.

2. Self-Learning and Feedback: Relying on existing tags provides a baseline which Plurality refines through user-specific learning and feedback. A user's personal tagging style can vary in syntax (e.g., tagging a document about Al Gore with "al-gore" versus "AlGore"), in viewpoint (e.g., "nobel prize winner" versus "presidential runner-up"), in goal (e.g., "todo", "homework"), and in many other dimensions. In our first prototype of Plurality, we apply traditional information retrieval techniques and specialized rule-based heuristics to model a user's tagging style based on a history of the user's tags and the user's interactions with the system. Figure 3 shows a sample user feedback screenshot. Plurality recommends a set of potentially relevant tags; based on the user's preference, the user can accept the recommended tags, reject the tags, or add their own tags; each of these decisions is reflected in an update to the user's tag profile.

3. Context-Sensitivity: When suggesting tags for a document, the context of both the document and the corpus against which tag suggestions are drawn are critically important. In Plurality, users can manually select a relevant corpus to compare against – e.g., all of the crawled del.icio.us documents, only the user's tagged documents – or select custom context filters based on temporal or spatial features. For example, the Figure 2 blog entry about "the high price of oil" was written in 2005. Plurality's tag suggestions in this case are drawn from a recent crawl of del.icio.us, so some of the tag suggestions are temporally relevant to the original blog entry, e.g., "iraq", "war", and "bush." One of the goals of the Plurality project is to tag archival content; hence, a 1970s document referencing the "high price of oil" could be tagged "jimmy carter" and "opec." To capture this contextual information, Plurality uses custom time and location regular-expressions to extract the creation date of a document and the location information (if available). Based on these contextual cues, Plurality supports tag suggestions based on a user-specific window around a particular date, or with respect to a user-specified geographic region.

3. CLOSING REMARKS

We have presented Plurality, an interactive tagging system that couples the collective intelligence of existing tag-based resources with a personalized context and feedback-sensitive interface. In our ongoing work, we have deployed Plurality internally at Texas A&M, and we are collecting usage and tagging data to evaluate the effectiveness of Plurality both in terms of application-specific tag quality (e.g., for search or browsing) and in terms of user satisfaction. Our research on Plurality continues along several directions. First, we will continue enhancing the capability and efficiency of the system through the incorporation of tag and document clustering, as well as the further personalization of results. We are also interested in continuing to explore context and its effects on tagging and tag selection – what contextual cues are most important to users?

4. REFERENCES

[1] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing web search using social annotations. In *WWW*, 2007.
[2] C. R. Brooks and N. Montanez. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *WWW*, 2006.
[3] P. A. Chirita, S. Costache, S. Handschuh, and W. Nejdl. Ptag: Large scale automatic generation of personalized annotation tags for the web. In *WWW*, 2007.
[4] R. Li, S. Bao, Y. Yu, B. Fei, and Z. Su. Towards effective browsing of large scale social annotations. In *WWW*, 2007.
[5] G. Mishne. Autotag: A collaborative approach to automated tag assignment for weblog posts. In *WWW*, 2006.