Tess: Using Resource Contents for Tag Suggestion

Bruno Oliveira[1], Pável Calado[1], and H. Sofia Pinto[1]

INESC-ID/IST, Av. Rovisco Pais, 1049-001 Lisboa, Portugal
bruno.santos.oliveira@gmail.com,
pavel.calado@tagus.ist.utl.pt,sofia@inesc-id.pt

Abstract. In this paper, we present an automatic tag suggester, *Tess*. Our system makes recommendations based only on the textual contents of the resource and is independent of existing tags, thus allowing the emergence of novel tags. Preliminary evaluation experiments show that the system is not only able to suggest many useful tags, but also to discover new and relevant tags, not suggested by any of the users.

1 Introduction

The shape of the World Wide Web has been changing, as websites tend to be more interactive and user-centric than in the past. Some of the most successful systems on this "Web 2.0" generation are the so-called social tagging systems [1]. These systems have their content managed by the users, who publish resources along with sets of tags, which are keywords to help organize and describe each resource.

Typically, the user uploads resources and associates a set of tags to them, as well as any additional information the system may require. Some systems, such as *del.icio.us*[1], suggest a set of tags for the user to choose from, making her task easier. This is commonly done by presenting the most popular tags associated to a resource when it was already tagged by other users. Nevertheless, tagging is still mostly a manual process.

In this work we present a novel text-based tag suggesting system, *Tess*. Differently from previous works, the tags suggested by Tess are extracted from the textual contents of the documents. Since there is no dependency on tags that were previously employed by users, *Tess* can keep up with the changes and evolution of the communities, their interests and resources, by suggesting tags that never appeared before. Furthermore, it is able to deal with resources that have never been tagged before, as long as they are text based.

Other works, such as [2] and [3] have also been proposed to suggest tags based on the tags of related resources. Chirita et al. [4] propose a system for annotating Web pages, also based on their content.

2 The Tag Selection Algorithm and Preliminary Evaluation Results

Tess works by examining the documents already present in the system and the new document, for which tags should be suggested (the *query document*).

[1] http://del.icio.us/

The algorithm for tag selection consists of two distinct phases. First, the query document is processed and modified in order to acquire all the words that might be useful for describing it. This step is called *vector displacement*. Following, its words are ranked, according to a given importance measure, and the top-ranked subset is selected as the tags to be presented to the user. This step is called *tag extraction*.

The vector displacement process starts by finding the documents most similar to the query document, using the classic *cosine similarity* formula [5]. Following, the query document is transformed by discarding terms that do not occur in the similar documents and term weights are changed so that the query document vector is displaced to the center of the most similar documents, according to the formula:

$$\boldsymbol{f} = \sum_{s \in S} \frac{1}{|S|}(w_{1,s}, w_{2,s}, ..., w_{n,s}) \quad (1)$$

where $\boldsymbol{f}$ is the final vector, S is the set of similar documents and $w_{i,s}$ is the weight of term i in document s. Five other variations of this method were also proposed.

Once the original document is modified, the words to be suggested as tags are extracted. To that effect, all words in the new vector are ranked, according to some importance measure. To this effect, we use 14 different methods, ranging from a simple term-frequency count [5] through Information Gain [6], plus combinations of all these measures.

Experimental evaluation results, involving 12 human users, show that Tess can obtain an average precision of 60%, which means that the majority of tags it suggests are considered relevant by users. Furthermore, and differently from existing proposals, Tess can suggest new and relevant tags, not previously used in any document. In fact, Tess was able to show the users from 61% to 76% new tags, showing that it will be capable of keeping up with the changes and evolution of the communities, by suggesting tags that never appeared before.

In future work, we intend to explore the graph-like structure of folksonomies, to detect relations between tags, users, and documents.

References

1. Treese, W.: Web 2.0: Is It Really Different? netWorker **10**(2) (2006) 15–17
2. Mishne, G.: Autotag: a collaborative approach to automated tag assignment for weblog posts. In: Proceedings of the 15th international conference on World Wide Web. (2006) 953–954
3. Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., Stumme, G.: Tag recommendations in folksonomies. In: Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases. (2007) 506–514
4. Chirita, P.A., Costache, S., Nejdl, W., Handschuh, S.: P-TAG: large scale automatic generation of personalized annotation tags for the Web. In: Proceedings of the 16th international conference on World Wide Web. (2007) 845–854
5. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill (1983)
6. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proceedings of the 14th International Conference on Machine Learning. (1997) 412–420