Using Tags for Query by Multiple Examples

Tomáš VESTENICKÝ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
vestenicky@icloud.com

Nowadays, the most widely used approach for searching information on the web is keyword-based. The main disadvantage of this approach is that users are not always efficient in keyword choice. This is why we aim to help them with query formulation or offer them different, more natural approach.

Query by example paradigm is often utilized in image and music search [1], because we can easily determine the topic and assess similarity. In the digital libraries domain, documents can have multiple topics; one may be more important or more relevant than the other. This is why we decided to represent the document by tags and use these to aid searching process.

Our method is based on building the query by choosing positive examples of documents or typing keyword-based query as a starting point and then specifying information radius of results by selecting positive or negative examples utilizing explicit relevance feedback as a query refinement method. We use user-added tags and author keywords to determine document similarity. Users add tags to documents which aids searching process, because users choose what is relevant for them from the particular topic. Therefore, tags can be used for more fine-grained query refinement by enabling users to see and/or remove tags from documents selected as positive or negative examples. In cases, where documents have not yet been tagged, we use keywords provided by the document's author.

While searching, users can label documents (displayed as search results) as positive or negative, based on their preference. Labelling the document means labelling all its associated tags or keywords, which are then added to the corresponding set. Tags and keywords can be separately removed from these sets for further query refinement. Labelling as positive increases the document (or tag) relevance and vice versa. After labelling the first document we disregard the textual query and display results using only our method.

Keywords, which are used when user-added tags are not available, are extracted by AlchemyAPI[1] service. We display top five keywords according to the relevance

* Supervisor: Róbert Móro, Institute of Informatics and Software Engineering
[1] http://www.alchemyapi.com/

Spring 2014 PeWe Workshop, March 21, 2014, pp. 45–46.

score provided by AlchemyAPI. Tag relevance score for the purposes of our method is computed using TF-IDF scheme [2]. Each tag has its score based on relevance to the corresponding document:

$$TFIDF(t,d,D) = \frac{n(t,d)}{\sum_{w \epsilon d} n(w,d)} \times \frac{\log |D|}{|d : t \in d|}$$

where *t* represents tag, *d* document, *w* word and *D* stands for the set of all documents. Function *n(x, d)* counts every occurrence of *x* in document *d*.

Labelled tags will be shown in the form of tag cloud (where the word size represents its relevance score) in the left sidebar alongside the results.

In extreme situations, our method follows these rules:

1. In cases, when user labels the same tag (or keyword, hereinafter referred to as "tag") differently during one session, this tag will be kept in both sets. Each tag will keep its score while reducing the other one's effect (removing tags is left to the user)
2. If user labels the same tag as positive/negative more than once, it will be represented by the highest value

Search results are ordered by their score, which is recalculated after every change user makes in positive or negative set of tags. The score for each document is calculated by following formula:

$$relevance = \sum_{t \epsilon DPT} s(t)d(t) \quad - \sum_{t \epsilon DNT} s(t)d(t)$$

where *DPT* is document's positive tags (document's tags present in positive set), *s(t)* stands for tag's search score and *d(t)* is tag's document score (result from TF-IDF).

We focus on the domain of digital libraries of research articles and we evaluate our proposed method in bookmarking service Annota[2] by means of a user study.

Acknowledgement. This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

[1] Rasiwasia, N., Vasconcelos, N.: Image retrieval using query by contextual example. In: *MIR '08: Proc. of the 1st ACM Int. Conf. on Multimedia Information Retrieval*, ACM Press, (2008), pp. 164-171.

[2] Falessi, D., Cantone, G., Canfora, G.: A comprehensive characterization of NLP techniques for identifying equivalent requirements. In: *ESEM '10: Proc. of the 2010 ACM-IEEE Int. Symposium on Empirical Software Engineering and Measurement*, ACM Press, (2010).

[2] http://annota.fiit.stuba.sk