Search and Exploration in LinkedCourse

Darina Dicheva, Christo Dichev, Rob Drayton

Winston Salem State University
601 M.L.K. Jr. Dr., Winston Salem, N.C. 27110, USA
{dichevad, dichevc}@wssu.edu, robad77@gmail.com

Abstract

This paper presents a collaborative learning repository that leverages the potential of domain specific social tagging in combination with ontology based classification. It is exemplified in *LinkedCourse* a learning repository prototype for collaborative development, sharing and reuse of resources. The focus of the paper is on the collaborative semantic annotation and searching for similar resources.

Introduction

To address existing needs in instructional resources in emerging disciplines we are developing an environment for collaborative development, sharing and reuse of learning resources – *LinkedCourse* (Dicheva et al, 2009). The work reported here is focused on the collaborative semantic annotation and searching for similar resources. It was motivated by our belief that in a community of practice tagging still has unexploited potential.

Collaborative Semantic Annotation

LinkedCourse aims to support community-based development and sharing of learning resources while acknowledging and preserving the copyright of the authors. The learning material registered in the LinkedCourse repository is distributed and resides on authors' websites. The repository contains only records with metadata for the original resources and their authors. The framework supports resource bookmarking and tagging.

The advantages and disadvantages of ontologies and folksonomies are well known. Our approach for sharing learning content is an attempt to reconcile the two structuring approaches by combining their distinct powers: the usability and flexibility of folksonomies with the standardization and interoperability of ontologies. It is driven by two observations: (1) Based on their experience with personal folders instructors are used to classify their material under courses, and subdivide it by course topics; (2) Tags are inseparable from the context of the community in which they are created and used (Mika, 2007). Based on the first observation, the learning resources in LinkedCourse are divided into course collections.

Course collections are the place where storing, tagging, and searching resources take place. Thus courses are used as both an organizational infrastructure of learning resources and social infrastructure for user interactions and forming course level communities. Course names are provided by the course creators and therefore are subjective. One possible strategy is to allow open course naming but to provide an additional option for tagging courses with 'standard tags'. In LinkedCourse we took this approach where the standard tags come from standard taxonomies such as ACM Computing Classification System (ACM, 1998). This approach implies two modes of tagging: regular and *conformant*. With the term 'regular tagging' we refer to the process of freely choosing words, while 'conformant tagging' refers to the process of choosing terms from a domain taxonomy or ontology. In order to motivate users to use conformant tagging the LinkedCourse interface supports a convenient drag-and-drop term selection from a visualizd ontology. Note that the conformant tagging is more often on a course level (as opposed to a resource level), thus not so frequent.

In LinkedCourse we also reuse existing vocabularies such as Dublin Core for describing learning resources and FOAF for describing contributors. This type of ontological support is in line with the MOAT framework (MOAT, 2001) that aims to provide a way for users to define meaning of their tags using URIs of Semantic Web resources. Users can choose to keep such semantic grouping within their private spaces. If they open it for sharing then it will be visible in the shared space. Inside courses user interaction with LinkedCourse is similar to the ordinary tagging systems – tagging with freely chosen words.

Finding Similar

When we request web pages similar to the one currently on display we typically mean pages (i) *matching the topic* of *the current one* and (ii) *found relevant by people sharing our interests*. This observation implies an informal definition depending on two concepts – "matching topic" and "relevant to the community of users sharing users interests". These observations suggest the following strategy: the level of similarity between two document d_1

and d_2 is a function of the number of shared tags and the number of users that have tagged both d_1 and d_2.

Definition 1. A folksonomy *F* is a tuple *F = (U,T,D,A),* where *U* is a set of users, *T* is a set of tags, *D* is a set of Web documents, and $A \subseteq U\times T\times D$ is a set of annotations.

We focus on tags and users associated with a particular document since we want to exploit these elements for deriving document similarity metrics.

Definition 2. A *tag-user* bipartite graph TU_d is a set of projections on the documents dimension $d\in D$ of a folksonomy *F*: i.e. $TU = \{TU_d|\ d\in D)$, where $TU_d = (U_d,T_d, A_d)$, A_d is the set of tag annotations: $A_d = \{(t,u)/ (u, t, d)\in A\}$, T_d is the set of tags applied by a given user *u*: $T_d = \{t|\ (u, t) \in A_d\}$ and U_d is the set of users applying given tag *t*: $U_d = \{(u|\ (u, t) \in A_d\}$.

The above bipartite graph can be represented in matrix form, capturing the users u_i *(i= 1,2,.., n)* that have applied tag t_j *(j = 1,2,.., k)* to document *d.*

Measuring Resources. We further define several notations.

Let $UT_d = \begin{pmatrix} r_{11} & r_{12}.. & r_{1n} \\ r_{21} & r_{22}.. & r_{2n} \\ ... & .. & \\ r_{k1} & r_{k2}.. & r_{kn} \end{pmatrix}$, where $r_{ij} = 1$ if user *i* has applied tag *j* to document *d*, and $r_{ij} = 0$ otherwise, is a matrix corresponding to the user-tag bipartite graph for document *d.* By $T_d = (r_1, r_2,..., r_n)$ we denote the vector obtained from summing all rows of the matrix UT_d, that is, $r_l = r_{1l}+ r_{2l}+..+ r_{kl}$ *(l= 1,2, ..,n)*. Each component r_l of T_d captures the number of occurrences of tag t_l in document *d.* We call vectors T_d *(d = 1, 2, .., m) tag vectors.*

Let $TU_d = \begin{pmatrix} s_{11} & s_{12}.. & s_{1k} \\ s_{21} & s_{22}.. & s_{2k} \\ ... & .. & \\ s_{n1} & s_{n2}.. & s_{nk} \end{pmatrix}$ where $s_{ij}= 1$ if tag *i* has been applied by user *j* to document *d*, and $r_{ij} = 0$ otherwise, is a matrix transposed to UT_d . Note that TU_d corresponds to the tag-user bipartite graph for document *d.* By $U_d = (s_1, s_2,..., s_k)$, $s_l = s_{1l}+ s_{2l}+..+ s_{nl}$, *(l= 1,2, ..,k)* we denote the vector resulting from summing all rows of matrix TU_d. Each component s_l of U_d captures the number of tags applied by user u_l to document *d.* We call vectors U_d *(d = 1, 2, ..,m) user vectors.*

Thus we describe each document by two vectors T_d and U_d intended as two measures for quantifying the documents' similarity. These to vectors induce two similarity measures between documents based on the classic cosine metrics - *tag* similarity and *user* similarity

$$sim_t(d_i,d_j) = \cos(T_i,T_j) = \frac{T_i\cdot T_j}{\|T_i\|*\|T_j\|}$$

$$sim_u(d_i,d_j) = \cos(U_i,U_j) = \frac{U_i\cdot U_j}{\|U_i\|*\|U_j\|}$$

We will use sometime the terms *tag similar* and *user similar* to make clear which measure was used for quantifying the distance between documents.

The algorithm for selecting similar documents is based on the following insight: document *d* is similar to document *x,* if *d* is both tag similar and user similar to *x*. The following factors are incorporated in the actual algorithm for finding documents similar to document *d.*

1. The set of documents that are *tag similar* to document *d.*
2. The set of documents that are *user similar* to *d.*
3. The popularity of the *tag similar* and *user similar* set of documents among the users tagging the document *d.*
4. The recency of the retrieved documents.

The proposed algorithm is summarized in the following steps. The starting assumption is that the user wants to find documents similar to the document *d* that has been tagged by users $u_1, u_2,.., u_r$..

1. All resources $D_0 = \{d_{i,}|\ sim_t(d_i, d) > c_1\}$ that are tag similar to *d,* where c_1 is a threshold, are retrieved,.
2. All resources $D_1 = \{ d_i \in D_0|\ sim_u(d_i, d) > c_2 \}$, where c_1 is a threshold, are retrieved.
3. For all resources $d_i \in D_1$ a combined similarity $sim(d_i, d) = k.sim_t(d_i, d). sim_u(d_i, d)$, is computed, where *k* is empirical coefficient.
4. Resources $d_i \in D_1$ are ranked based on the combined similarity measure: $R(d_i) > R(d_j)$ if $sim(d_i, d)>sim(d_j, d)$.
5. The documents d_i and d_i with similar ranking computed at step 5, i.e. $|R(d_i) - R(d_j)| < e_1$ are reordered according to the following criteria: $R(d_i) > R(d_j)$ if the number of users that have tagged d_i is greater than the number of users that have tagged document d_j. This criterion favors resources that are more popular among the relevant users.

The above approach for finding similar documents can be adapted for finding similar users. The user-tag-document matrices need to be analyzed now in their tag-document context, assuming tag-document (document-tag) matrices. These two matrices induce two similarity measures between users - *document* similarity and *tag* similarity.

References

Dicheva, D., Dichev, C., Zhu, Y . Sharing Open-Content Learning Resources in Emerging Disciplines. 7th Int'l *Workshop on Ontologies and Social Semantic Web for E-Learning* (SWEL'09), Brighton, UK, July 6-10, 2009.

Mika P. Ontologies are us: A unified model of social networks and semantics, *J. Web Semantics* **5** (1) (2007), pp. 5–15.

ACM CCS: http://www.acm.org/about/class/1998/.

MOAT, http://www.w3.org/2001/sw/wiki/MOAT.