Analyzing hidden semantics in social bookmarking of open educational resources

Julià Minguillón[1]

[1] Universitat Oberta de Catalunya, Rambla Poblenou 156,
08018 Barcelona, Spain
jminguillona@uoc.edu

Abstract. Web 2.0 services such as social bookmarking allow users to manage and share the links they find interesting, adding their own tags for describing them. This is especially interesting in the field of open educational resources, as delicious is a simple way to bridge the institutional point of view (i.e. learning object repositories) with the individual one (i.e. personal collections), thus promoting the discovering and sharing of such resources by other users. In this paper we propose a methodology for analyzing such tags in order to discover hidden semantics (i.e. taxonomies and vocabularies) that can be used to improve descriptions of learning objects and make learning object repositories more visible and discoverable. We propose the use of a simple statistical analysis tool such as principal component analysis to discover which tags create clusters that can be semantically interpreted. We will compare the obtained results with a collection of resources related to open educational resources, in order to better understand the real needs of people searching for open educational resources.

Keywords: tags, delicious, social bookmarking, delicious, open educational resources, principal component analysis, repositories

1 Introduction

Open educational resources (OER) have been a hot topic in the recent years, as more and more educational institutions and individuals are making their assets and collections available to the whole community through Internet. These resources are organized and published through open repositories such as MERLOT or OpenCourseWare, for instance, but other kinds of resources that can be also considered educational under some circumstances are also available, such as Wikipedia or LearningSpace, among others. In fact, the term OER can be used for a very wide range of resources, formats, granularity, etc. Therefore, it would be interesting to develop a methodology for establishing the meaning of words related to OER at different levels, improving the descriptions of both learning objects and collections in open repositories.

Since the explosion of Web 2.0 services and applications, it has never been so easy to create and share resources using the web as the platform. Blogs, wikis, LMSs,

CMSs, etc., there are plenty of open source tools and online services for such purposes. But discovering and organizing all these contents is also an interesting subject from the perspective of each user. Social bookmarking, and more specifically delicious[1], is a way of organizing such resources by tagging them with words closer to the final user rather than the resource creators. Hundreds of thousands of users tag and share the same links, creating a huge folksonomy [1] of terms related to a given concept. This information can be further analyzed and exploited in order to better understand the way people use certain words to describe common terms which are widely used, such as the term "open" in "Open Educational Resources".

In this paper we discuss the possibilities of using the crowd-sourcing phenomenon of social bookmarking for extracting semantics from the tags added by delicious users, which describe links related to open educational resources. In Section 2 we outline the Open Educational Resources movement and the way of organizing content through open repositories and other Web 2.0 based platforms. Section 3 describes the methodology used for extracting useful information from delicious tags related to a given search term, in our case "Open Educational Resources". Section 4 presents the results obtained for such search term and the semantics hidden in the relationships between similar users tagging similar resources, following the abovementioned methodology. Finally, Section 5 summarizes the main conclusions that can be drawn from this work and identifies the topics that should be addressed in order to improve the proposed methodology.

2 Open Educational Resources and Open Repositories

The concept of "Open Educational Resource" was first firstly adopted at UNESCO's 2002 Forum on the Impact of Open Courseware for Higher Education in Developing Countries, according to Wikipedia. The main idea in the educational field was to reproduce the amazing success of the Open Source movement, trying to create a true community involved in open education. Since then, the OER movement has become a strong reality with thousands of open educational resources available, ranging from small chunks of knowledge (i.e. a definition or a formula) to whole web sites devoted to a specific topic (i.e. DLESE for earth sciences). Setting the granularity at course level, the OpenCourseWare consortium offers 3771 courses from 46 different sources in 7 languages (as July 2010).

In this sense, and taking advantage of the evolution of digital libraries, more and more learning object repositories are currently now available. Nevertheless, learners do not have an easy way to organize such learning objects accordingly to their own particularities and preferences yet. The concept of repository itself is very different depending on the point of view (institutional vs. personal, for instance [2]), as institutional needs (preservation, dissemination) are not exactly the same than learners needs (organization, learning). Nevertheless, and with the advent of the web 2.0, many other sites such as flickr or youtube can be also considered open repositories managed by users themselves.

[1] http://delicious.com

2.1 METAOER: resources for understanding the OER movement

As part of the activities of the UOC UNESCO Chair in e-Learning, a collection of resources related to the OER movement has been created and shared using delicious[2]. It is not another collection of open educational resources; it is a collection of open resources that describe key issues related to the OER movement. This collection is always "under construction", as more and more links are added every week; currently now there are around 80 resources but we expect to reach a few hundreds.

All these resources are tagged as "#metaoer" with the idea that this collection will continue growing with collaborations from individuals (experts or not) interested in supporting this project. Each resource tagged as "#metaoer" will be analyzed by a group of experts and, if possible, it will become part of an open repository about open educational resources, named METAOER, which will be part of the UOC institutional repository[3]. In order to do so, each link is tagged according to several tag bundles created for such purpose, including information about its authors, file format (PDF, Word, OpenOffice and so), license, organization / affiliation, type (paper, report, blog post, etc.) and, specially, its category. This last bundle defines what we think it is necessary to understand in order to become an active member of the OER movement, as shown in Table 1.

Table 1. Categories used for cataloguing the documents tagged as "#metaoer".

Category	Explanation
Awareness	Documents related to the philosophy behind the OER movement, institutional declarations, policies and so.
Format	Documents related to the best file formats and standards used to support OER documents.
License	Documents for understanding which license is more appropriated and the restrictions imposed by each one.
Metadata	Documents about standards and specifications for describing learning objects using metadata, vocabularies and taxonomies.
Repository	Documents about how to build and maintain an open repository, as well as documents about best practices.
Software	Documents about software tools related to the other categories (format conversion, repositories, etc.)

In order to improve this classification and the collection of selected links, we will compare it with the results obtained when applying the methodology described in the following section. This methodology helps us to build semantic clusters starting from the tags used by people tagging open educational resources in delicious. We expect these clusters to reveal information about what people think (and probably need) about open educational resources, from a user centered perspective. In fact, user tags can be seen as "open descriptions" of open educational resources, assuming that users searching for resources only tag and share those that they find really valuable.

[2] http://delicious.com/uocunescochair/#metaoer
[3] http://openaccess.uoc.edu

3 Extracting and analyzing delicious tags

As abovementioned, delicious is a Web 2.0 social bookmarking service that allows users to manage and share links in a standardized way. It improves the idea of "favorite links" implemented by web browsers, allowing users to manage them from any computer, using their own tags for describing links and sharing them with other users, if desired. It is very popular (ranked as the 326 most visited web site by Alexa, as July 2010), although preliminary studies showed that is not very used by higher education online students for managing educational resources (only 11% of students used delicious to manage their links). Delicious is a very good example of "the wisdom of crowds" [3], as it reflects the independent and diverse opinions of a group of individuals, although common tags are provided by default by the system, thus introducing an accumulation effect for the most common tags.

Table 2 shows the main results obtained when searching for "Open Educational Resources". A total of 3621 results are obtained (as July 2010), the first 10 are shown. For each resource, only the most important 5 tags are shown (according to the delicious search engine). Notice that "education" is a common tag for most resources, for example, as well as "learning" or "opensource". This is not remarkable, though, as it is the expected behavior when thousands of users are tagging the same resources and the recommendation system probably offers such tag as an option. On the other hand, we are more interested in discovering the "long tails" [4] related to each resource, as well as the coincidences (or no-coincidences) among tags between users, that is, how tags are grouped to create clusters which can be considered natural sub-domains of the field, extending the analysis performed in [5].

Table 2. 10 first page results obtained from searching for "Open Educational Resources" in delicious (search performed in July 2010).

Resource	Number of times saved	Five most important tags
OER Commons	3417	education, opensource, learning, curriculum, resources
MIT OCW	15058	education, mit, learning, free, online
MERLOT	3775	education, teaching, elearning, multimedia, resources
Academic Earth	22808	education, video, lectures, learning, university
LearningSpace	3192	education, learning, free, elearning, university
Connexions	5078	education, opensource, learning, courseware, collaboration
OCW Consortium	3977	education, opencourseware, learning, opensource, free
Wikipedia	37546	reference, encyclopedia, Wikipedia, wiki, research
DOAJ	6245	journals, research, reference, openaccess, science
Moodle	12941	education, opensource, moodle, elearning, cms

Notice that the first seven sites are perfect examples of learning object repositories, while the other three sites (Wikipedia, DOAJ and Moodle) are of different nature, especially Moodle, which is an open source learning management system not directly related to content. On the other hand, Wikipedia has become a very popular resource because of its internal structure of links, which makes it to appear on top of searching engines results, while the Directory of Open Access Journals is a particular case of repository with a target audience, namely the scientific community.

3.1 Methodology

Our goal is to analyze all the tags used to describe resources in a given domain, trying to discover not only the most common tags but also the relationships between tags, users and the links described using such tags.

In order to retrieve all the information stored in delicious for a given concept, we propose to use the following methodology:

1. Retrieve the first N links related to a given concept using the delicious search engine.
2. For each link, retrieve the first M users that have bookmarked such link.
3. For each pair {link, user}, generate a list of the tags used by such user to describe such link. In case of an empty list, remove such pair.
4. Generate a list of all the tags for all valid pairs.

Applying the first three steps we obtain a variable length data set with, at most, N * M entries (i.e. all the valid pairs), each one up to 45 tags (the maximum shown by delicious). In step 4 we obtain a set that can be used to determined the most important tags. Nevertheless, some data cleansing is needed in order to improve the quality of tags before they are analyzed. For example, some users may have used “elearning” as a tag while others may have used “e-learning”. Misspelling is also a common problem, as well as using the same word in plural or singular (i.e. “tools” vs “tool”), or using words that can be considered to be equivalent (i.e. “education” and “educational”). This step could be partially automated using a system such as WordNet, for example [6].

As expected, the set of tags generated in step 4 is also a long tail that must be cut at a certain point. Two main options are available: a) to specify a number T of desired tags; b) to specify an accumulated probability p (i.e. a threshold) and select the first T tags that achieve such threshold. Anyway, this step can be simplified as follows:

5. Generate a reduced tag set containing the most important T tags and a conversion table which specified which valid tag is used for every tag in the original data set, ranked according to tag importance.

Then we can proceed to clean the results obtained in step 3, as follows:

6. For each pair {link, user} obtained in step 3 replace every tag by its valid version according to the tag set generated in step 5, removing those pairs with no valid tags.

Finally, when a set of tags has been determined, we binarize the data set in order to obtain a matrix saying whether a given pair {link, user} has been tagged using $\{tag_i\}$, as follows:

7. For each pair {link, user} replace column i by “1” if the i-th tag in the tag set was used, “0” otherwise.

This generates a data set of dimension T which can be further analyzed using statistical or data mining techniques. For exploratory purposes, we propose to use Principal Component Analysis, as follows.

3.2 Principal component analysis

Once the binary data set has been generated, we want to determine two different things: first, which tags describe better the data set in terms of variance and, second, how these tags are related to each other. In order to do so, we propose to use Principal Component Analysis, a well known tool for data exploration analysis. We use maximum likelihood as the criterion for extracting components from the original data set, followed by a Varimax rotation, in order to minimize the number of fields (i.e. tags) needed to interpret the obtained components. Finally, we only consider those factors larger or equal to 0.3 for explaining each component, thus reducing the number of factors in each component.

We expect each component to gather a reduced set of factors (i.e. tags) which, hopefully, will be related to each other, discovering natural sub-domains of the top level concept used for building the analyzed data set. Notice that in the case of binary variables, maximum variance is obtained when mean is exactly 0.5, which is not the expected behavior except maybe for the most common tags, which do not provide useful information as they are, usually, too general (i.e. “education”).

3.3 Main methodology drawbacks

This methodology has some well known drawbacks, although they are not really important for experimentation purposes. The main drawback is that links retrieved in steps 1 and 2 depend on delicious internal search engine, which can be biased. This can be partially avoided by including as many links and users as possible, but due to the necessity of attending thousands of concurrent requests, it is not possible to retrieve all the links and all the users for a given search term, as delicious uses bandwidth throttling protection schemes. Nevertheless, it is possible to perform an incremental search (every few hours), thus improving the quality of the results. Another important drawback is that the same web site can be accessed through different links, depending on whether the URL specified by the user includes the name of the web page or not (that is, “index.php” may be part of the URL or not).

In the following section we will test the methodology as a preliminary experiment for understanding what users mean when they tag links related to the OER movement. This methodology can be combined with the method described in [7] in order to create ontologies for representing domain concepts.

4 Results and discussion

Following the example shown in Table 2, we applied the methodology described in Section 3.1 in order to analyze the way users are tagging links related to “Open

Educational Resources". Steps 1 and 2 generate a set with 65545 pairs {link, user} from 100 links tagged by 38298 different users, containing a total of 13929 tags (6632 of them are used only once). Once the data cleansing step described in step 6 is performed, we set the number of desired tags to T=100, so the final data set contains 47674 different {link, user} pairs from 28292 different users, containing 100 different tags. The first ten tags according to their importance are: education (used 17871 times), free (9704), resource (7432), learning (6682), reference (5768), video (5530), opensource (5483), web20 (4782), elearning (4333) and tool (3903). The last tag in this data set (cms) has been used 289 times. As expected, tags follow a power-law distribution with a very long tail.

It is interesting to analyze also the joint distribution of the pairs {link, user}. For each link, we have between 249 and 844 users that have tagged it at least with one valid tag. A visual inspection of the histogram of this distribution shows that it is not normal, but a combination of two normal distributions, the largest centered in the lowest values and another centered in the highest ones, but smaller. This is consistent with the fact that there are some links extremely popular (i.e. Wikipedia). On the other hand, each user tags between 1 and 25 sites, with a distribution that can be very well approximated by a geometric discrete with p=0.593447. We tried also to use a Zipf's law distribution but it does not fit as well as the geometric one. This information, combined with the quality of tags, could be used to rank users as novice or expert for a given search term, although this is out of the scope of this paper.

4.1 Principal component analysis

Then, we proceed to perform a principal component analysis, as described in Section 3.2. We obtain 33 components which eigenvalue (i.e. explained variance) is larger than one, explaining the 49.8% of the original data set variance (the largest explains only 3.0%, showing a lack of strong structure among tags). Table 3 shows the first ten components obtained and the tags that define the factors used in each component.

Table 3. 10 first components obtained after PCA is applied to the data set.

Component	Tags involved in the component
C1	images, photos, photography, photo, stock
C2	free, learning, online, course, university, college
C3	book, ebook, library, reading, literature
C4	programming, kids, animation, games
C5	opensource, tool, software, freeware
C6	research, journals, openaccess
C7	tutorial, howto, diy[4]
C8	graphics, clipart
C9	curriculum, lessonplans
C10	web20, collaboration, community

[4] Common acronym for "do it yourself".

Notice that due to the nature of the Varimax rotation applied after the principal component analysis, the same tag is not used in two different components (at least for the first 10 shown, but it is also true for the first 15 computed components). This allows us to create separate clusters with respect to tagging strategies but, as we will see, clusters will probably show some semantic relationship among them.

4.2 Interpretation of results

As described in [8], the components generated by PCA can be seen as an initial solution for any clustering algorithm based on computing centroids (or, in the case of binary vectors, medoids). Therefore, we can infer that the tags that appear in the same component are the basis for a given cluster and that they are strongly related to each other. In fact, results in Table 3 show a natural aggregation of tags around different concepts which reveal some hidden semantics:

- C1 and C8 deal with images, one of the most common assets needed by people creating resources, specially blog posts and wiki entries. It is interesting to observe that there is a clear separation between real world images and computed generated cliparts. Therefore, from a semantic point of view, it is important to tag graphic resources according to their nature (real / synthetic). Sites such as flickr allow users to specify such category for a given image.
- C2 and C9 are related to courses which are available through Internet. These two components are directly related to the concept of OER from a consumer's perspective. C2 is more oriented towards open content, while C9 is more related to organizational issues. This shows one of the well known problems in the OER movement, the necessity of providing additional support for open content, as "learning is more than just content", quoting D. Wiley, who also stated that "we live in the age of content abundance" and "content is infrastructure". These two levels (content and organization) need to be well described and related one to each other.
- C3 is related to books, but it also includes a reference to e-books, a recent technology which is increasing very fast as more and more users, institutions and companies are adopting it.
- C4 is probably the most inconclusive cluster, as it mixes tags with very different meanings. An inspection of the data set reveals that most of the high values for this component correspond to two web sites, namely zoho (a web suite of online applications) and curriki (an environment for creating educational materials).
- C5 is related to software, including references to the open source movement (such as Moodle), but the appearance of "tool" and "freeware" seems to indicate that users needs have a very wide degree of granularity, ranging from small solutions for PDF creation to full content management systems.
- C6 describes the domain of open access journals, which publish papers that can be considered open resources for research purposes. This component shows the relevance of DOAJ as one of the 10 most important web sites when searching for "open educational resources".

- Finally, C7 and C10 are also two very interesting clusters, as they are directly related to the philosophy of the OER movement, but from the producer's perspective (i.e. creating and sharing), both individually and collectively.

On the other hand, for the 33 computed components, there are several tags that are never used as a factor in any component. More specifically: resource, reference, elearning, teaching, opencourseware, blog, school, academic, lecture, podcast, creativecommons and oer, among others. Notice that the first three were in the list of 10 most used tags, which means that not all the most popular tags are relevant for analysis purposes, mainly because they appear everywhere in combination with others tags which are more representative of a specific concept.

4.3 Improving the METAOER repository

In the light of the results obtained in the previous section, the most interesting clusters for our purposes are, on the one hand, C2 and C9, that is, open courses already available; and on the other hand, C7 and C10, that is, creating and sharing content. C5 is also an interesting cluster as it is directly related to the category "Software" in Table 1, as abovementioned. Therefore, in order to improve METAOER, we will:

- Define a taxonomy for OER granularity, including documents that explain how to use it by means of current standards and specifications for metadata.
- Include a new category for tutorials, that is, documents that explain how to put into practice the techniques explained in the other categories.
- Include a new category for learning communities, related to the "awareness" one but describing communities of learning with best practices on OERs.
- Define a taxonomy for software tools related to the OER movement.

5 Conclusions

Hundreds of millions or even billions of Internet users are interacting everyday with hundreds of millions of potential educational resources. Since the Web 2.0, users are able to create and share resources, so the amount of information available has grown exponentially. On the other hand, these resources need to be organized and described in order to make them visible in such a huge collection as Internet is. This organization is partially created and shared through social bookmarking services such as delicious, for example. Delicious can be used as a huge database for understanding the way users are tagging resources they find to be valuable, at a very high level, with very simple words.

In this paper we have proposed a methodology for extracting information from delicious, retrieving the most important tags used to describe the most relevant links related to a given search term, using PCA for exploratory purposes. We have applied this methodology to the concept of "Open Educational Resource", in order to better understand what people mean when they tag resources according to such term and

which kind of resources (or repositories, in a wide sense of the term) are they tagging, in order to discover their needs and see how the OER movement can be promoted.

Our results show that the clusters obtained from the PCA applied to the collection of links describe different sub-domains related to OER from different perspectives: images, courses, books, software, journals, etc. Therefore, any collection of open educational resources should take into account these categories in order to provide fast access to users trying to find such resources. Furthermore, it is necessary to provide also descriptors that allow users to narrow the range of results obtained when searching for a specific term, such as software related to OER, for instance. We are in the process of setting up an open repository about the OER movement, and the obtained results will allow us to improve the descriptions of the documents stored in such repository, by means of taxonomies and vocabularies.

Current and future research in this topic should include the analysis of other terms commonly used in the OER movement, such as “open education”, for example, and then trying to compare and/or combine the obtained clusters. On the other hand, for each one of the obtained sub-domains, it would be interesting to apply the same methodology in order to create a hierarchical taxonomy of terms. This could be combined with hierarchical clustering, trying to establish the relationships between sub-domains. In order to do so, enlarging the data set used for building the clusters is also necessary, incorporating not only one hundred of links but thousands of them, although this will be time consuming in both information retrieving and analysis. Finally, connecting the obtained clusters with some heuristics could be useful for building taxonomies and vocabularies for a given concept.

Acknowledgments. This paper has been partially supported by the UOC UNESCO Chair in e-Learning, 2010.

References

1. Angeletou, S., Sabou, M., Specia, L., Motta, E.: Bridging the Gap Between Folksonomies and the Semantic Web: An Experience Report. Workshop Bridging the Gap between Semantic Web & Web 2.0 at 4th European Semantic Web Conference, Innsbruck, (2007)
2. Peters, T.A.: Digital repositories: individual, discipline-based, institutional, consortial or national? The Journal of Academic Librarianship, 28(6), pp. 414-417 (2002)
3. Sunstein, C.: Infotopia: How Many Minds Produce Knowledge. Oxford University Press (2006)
4. Anderson, C.: The long tail. Wired, October (2004)
5. Robu, V., Halpin, H., Shepherd, H.: Emergence of Consensus and Shared Vocabularies in Collaborative Tagging Systems. ACM Transactions on the Web, 3(4), no. 14 (2009)
6. Fellbaum, C., ed.: Wordnet: An electronic lexical database. Cambridge, MIT Press (1998)
7. Sugumaran, V., Purao, S., Storey, V.C., Conesa, J.: On-Demand Extraction of Domain Concepts and Relationships from Social Tagging Websites. Proceedings of the 15th International Conference on Applications of Natural Language to Information Systems, pp. 224-232, Cardiff, UK, 23-25 June (2010)
8. Ding, C., He, X.: K-means Clustering via Principal Component Analysis. Proceedings of the International Conference on Machine Learning, pp. 225–232, Banff, Canada, 4-8 July (2004)