Measuring Vertex Centrality in Co-occurrence Graphs for Online Social Tag Recommendation

Iván Cantador, David Vallet, Joemon M. Jose

Department of Computing Science
University of Glasgow
Lilybank Gardens, Glasgow, G12 8QQ, Scotland, UK
{cantador, dvallet, jj}@dcs.gla.ac.uk

Abstract. We present a social tag recommendation model for collaborative bookmarking systems. This model receives as input a bookmark of a web page or scientific publication, and automatically suggests a set of social tags useful for annotating the bookmarked document. Analysing and processing the bookmark textual contents - document title, URL, abstract and descriptions - we extract a set of keywords, forming a query that is launched against an index, and retrieves a number of similar tagged bookmarks. Afterwards, we take the social tags of these bookmarks, and build their global co-occurrence sub-graph. The tags (vertices) of this reduced graph that have the highest vertex centrality constitute our recommendations, which are finally ranked based on TF-IDF and personalisation based techniques.

Keywords: social tag recommendation, co-occurrence, graph vertex centrality, collaborative bookmarking.

1 Introduction

Social tagging systems allow users to create or upload resources (web pages[1], scientific publications[2], photos[3], video clips[4], music tracks[5]), annotate them with freely chosen words – so called *tags* – and share them with others. The set of users, resources, tags and annotations (i.e., triplets user-tag-resource) is commonly known as *folksonomy*, and constitutes a collective unstructured knowledge classification. This implicit classification is then used by users to organise, explore and search for resources, and by systems to recommend users interesting resources.

These systems usually include tag recommendation mechanisms to ease the finding of relevant tags for a resource, and consolidate the tag vocabulary across users. However, as stated in [7], no algorithmic details have been published, and it is assumed that, in general, tag recommendations in current applications are based on suggesting those tags that most frequently were assigned to the resource, or to similar resources.

---

[1] Delicious – Social bookmarking, http://delicious.com/
[2] CiteULike – Scholarly reference management and discovery, http://www.citeulike.com/
[3] Flickr – Photo sharing, http://www.flickr.com/
[4] YouTube – Video sharing, http://www.youtube.com/
[5] Last.fm – Personal online radio, http://www.last.fm/

Recent works have proposed more sophisticated and accurate methods for tag recommendation. These methods can be roughly classified into content-based and collaborative approaches. Content-based techniques [3, 4, 10, 16] analyse the contents and/or meta-information of the resources to extract keywords, which are directly suggested to the user or matched with existing tags. Collaborative strategies [6, 7, 17], on the other hand, exploit folksonomy relations between users, resources and tags to infer which of the tags of the system folksonomy are most suitable for a particular resource. Hybrid techniques, combining content and collaborative features, have been also investigated [5, 15].

In this paper, we present a hybrid tag recommendation model for an online bookmarking system where users annotate online web pages and scientific publications. The model receives as input a bookmark, analyses and processes its textual contents – document title, URL, abstract and description – extracting a set of keywords, and forms a query that is launched against an index to retrieve a number of similar tagged bookmarks. Afterwards, its takes the social tags of these bookmarks, and builds their global co-occurrence sub-graph. The tags (vertices) of this reduced graph that have the highest vertex centrality constitute the recommendations, which are finally ranked based on TF-IDF [14] and personalisation based techniques.

Participating at the ECML PKDD 2009 Discovery Challenge[6], we have tested our approach with a dataset from BibSonomy system[7], obtaining precision values of 42% and 25% when, respectively, one and five tags are recommended per bookmark. As we explain herein, the benefits of our approach are its low computational cost, and its capability of suggesting diverse tags in comparison to selecting the most popular tags matched with each bookmark.

The rest of the paper is organised as follows. Section 2 summarises state-of-the-art tag recommendation techniques. Section 3 describes the document and index models used by our tag recommender. Section 4 explains the stages of the recommendation process. Section 5 describes the experiments conducted to evaluate the proposal. Finally, Section 6 provides some conclusions and future work.

2 Related work

Analogously to recommender systems [1], tag recommendation techniques can be roughly classified into two categories: content-based and collaborative techniques. Whereas content-based approaches focus on the suggestion of keywords extracted from resource contents and meta-data, collaborative approaches exploit the relations between users, resources and tags of the folksonomy graph to select the set of recommended tags. Continuing with the previous analogy, tag recommendation techniques that combine content-based and collaborative models can be called hybrid approaches, and techniques that make tag recommendations biased by the user's (tag-based) profile can be called personalised models.

Based on the previous classification, in this section, we describe state-of-the-art tag recommendation techniques that have been proposed for social bookmarking systems.

---

[6] ECML PKDD 2009 Discovery Challenge, http://www.kde.cs.uni-kassel.de/ws/dc09/

[7] BibSonomy – Social bookmark and publication sharing, http://www.bibsonomy.org/

2.1 Content-based tag recommenders

Mishne [10] presents a simple content-based tag recommender. Once a user supplies a new bookmark, bookmarks that are similar to it are identified. The tags assigned to these bookmarks are aggregated, creating a ranked list of likely tags. Then, the system filters and re-ranks the tag list. The top ranked tags are finally suggested to the user. To find similar bookmarks, the author utilises a document index, and keywords of the input bookmark to form a query that is launched against the index. The tags are scored according to their frequencies in the top results of the above query, and those tags that have been used previously by the user are boosted by a constant factor. Our approach follows the same stages, also using an index to retrieve similar bookmarks. It includes, however, more sophisticated methods of tag ranking based on tag popularity and personalisation aspects.

Byde et al. [3] present a personalised tag recommendation method on the basis of similarity metrics between a new document and documents previously tagged by the user. These metrics are derived either from tagging data, or from content analysis, and are based on the cosine similarity metric [14]. Similar metrics are used by our approach in some of its stages.

Chirita et al. [4] suggest a method called P-TAG that automatically generates personalised tags for web pages. Given a particular web page, P-TAG produces keywords relevant both to the page contents and data residing on the user's desktop, thus expressing a personalised viewpoint. A number of techniques to extract keywords from textual contents, and several metrics to compare web pages and desktop documents, are investigated. Our approach applies natural language processing techniques to extract keywords from bookmark attributes, but it can be enriched with techniques like [4] to also analyse and exploit the textual contents of the bookmarked documents.

Tatu et al. [16] propose to extract important concepts from the textual metadata associated to bookmarks, and use semantic analysis to generate normalised versions of the concepts. For instance, `European Union`, `EU` and `European Community` would be normalised to the concept `european_union`. Then, users and resources are represented in terms of the created conceptual space, and personalised tag recommendations are based on intersections between such representations. In our approach, synonym relations and lexical derivations between tags are implicitly taking into consideration through the exploitation of tag co-occurrence graphs.

2.2 Collaborative tag recommenders

Xu et al. [17] propose a collaborative tag recommender that favours tags used by a large number of users on the target resource (high authority in the HITS algorithm [8]), and minimises the overlap of concepts among the recommended tags to allow for high coverage of multiple facets. Our approach also attempts to take into account tag popularity and diversity in the recommendations through the consideration of vertex centralities in the tag co-occurrence graph.

Hotho et al. [6] present a graph-based tag recommendation approach called FolkRank, which is an adaptation of PageRank algorithm [12], and is applied in the folksonomy user-resource-tag graph. Its basis is the idea that a resource tagged with important tags by important users becomes important itself. The same holds, symmetrically, for users and tags. Having a graph whose vertices are associated to users, resources and tags, the algorithm reinforces each of them by spreading their weights through the graph edges. In this work, we restrict our study to the original folksonomy graph. As a future research goal, PageRank, HITS or other graph based techniques could be applied to enhance the identification of tags with high graph centrality values.

Jäscke et al. [7] evaluate and compare several tag recommendation algorithms: an adaptation of user-based collaborative filtering [13], FolkRank strategy [6], and methods that are based on counting tag co-occurrences. The authors show that graph-based and collaborative filtering approaches provide better results that non-personalised methods, and state that methods based on counting co-occurrences have low computational costs, thus being preferable for real time scenarios. Our approach is computationally cheap because it is based on a simple analysis of tag co-occurrence graphs, and includes a personalisation stage to better adjust the tag recommendations to the user's profile.

2.3 Hybrid tag recommenders

Heymann et al. [5] present a technique that predicts tags for a website based on page text, anchor text, surrounding hosts, and other tags assigned to the website by users. The tag predictions are based on association rules, which, as stated by the authors, may serve as a way to link disparate vocabularies among users, and may indicate synonym and polysemy cases. As a hybrid approach, our tag recommender makes use of content-based and collaborative tag information. Nonetheless, we simplify the process limiting it to the exploitation of meta-information of the contents available in the bookmarks.

Song et al. [15] suggest a tag recommendation method that combines clustering and mixture models. Tagged documents are represented as a triplet (words, documents, tags) by two bipartite graphs. These graphs are clustered into topics by a spectral recursive embedding technique [18]. The sparsity of the obtained clusters is dealt with a two-way Poisson mixture model [9], which groups documents into components and clusters words. Inference for new documents is based on the posterior probability of topic distributions, and tags recommendations are given according to the within-cluster tag rankings.

3 Document and index models

To suggest tags for an input bookmark, our recommender exploits meta-information associated to it. The text contents of bookmarked documents (web pages or scientific publications) could be also taken into account, but we decided to firstly study how accurate tag recommendations can be by only using bookmarking meta-information.

In this work, we test our approach with a dataset obtained from BibSonomy system, whose bookmarks have, among others, the attributes shown in Table 1.

Table 1. Meta-information available in BibSonomy system about two different bookmarks: a web page and a scientific publication.

URL	**http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html**
Description	Folksonomies - Cooperative Classification and Communication Through Shared Metadata
Extended	General overview of tagging and folksonomies. Difference between controlled vocabularies, author and user tagging. Advantages and shortcomings of folksonomies

Title	**Semantic Modelling of User Interests Based on Cross-Folksonomy Analysis**
Author	M. Szomszor and H. Alani and I. Cantador and K. O'hara and N. Shadbolt
Booktitle	Proceedings of the 7th International Semantic Web Conference (ISWC 2008)
Journal	The Semantic Web - ISWC 2008
Pages	632-648
URL	http://dx.doi.org/10.1007/978-3-540-88564-1_40
Year	2008
Month	October
Location	Karlsruhe, Germany
Abstract	The continued increase in Web usage, in particular participation in folksonomies, reveals a trend towards a more dynamic and interactive Web where individuals can organise and share resources. Tagging has emerged as the de-facto standard for the organisation of such resources, providing a versatile and reactive knowledge management mechanism that users find easy to use and understand. It is common nowadays for users to have multiple profiles in various folksonomies, thus distributing their tagging activities. In this paper, we present a method for the automatic consolidation of user profiles across two popular social networking sites, and subsequent semantic modelling of their interests utilising Wikipedia as a multi-domain model. We evaluate how much can be learned from such sites, and in which domains the knowledge acquired is focussed. Results show that far richer interest profiles can be generated for users when multiple tag-clouds are combined.

In our approach, for each bookmark, using a set of NLP tools [2], the text attributes title, URL, abstract and description, and extended description are processed and transformed into a weighted list of keywords. These simplified bookmark representations are then stored into an index, which will allow fast searches for bookmarks that satisfy keyword- and tag-based queries. In our implementation, we used Lucene[8], which allowed us to apply keyword stemming, stop words removal, and term TF-IDF weighting.

[8] Apache Lucene – Open-source Information Retrieval library, http://lucene.apache.org/

4 Social tag recommendation

In this section, we describe our approach to recommend social tags for a bookmark, which does not need to be already tagged. The recommendation process is divided in 5 stages, depicted in Figure 1. Each of these stages is explained in detail in the next subsections. For a better understanding, the explanations follow a common illustrative example.

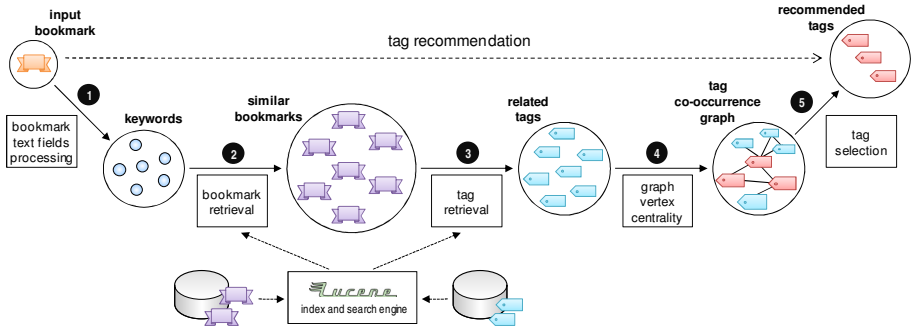

Figure 1. Tag recommendation process.

4.1 Extracting bookmark keywords

The first stage of our tag recommendation approach (identified by label 1 in Figure 1) is the extraction of keywords from some of the textual contents of the input bookmark.

According to the document model explained in Section 2, we extract such keywords from the title, URL, abstract, description and extended description of the bookmark. We made experiments processing other attributes such as authors, user comments, and book and journal titles, but we obtained worse recommendation results. The noise (in the case of personal comments) and generality (in the case of authors and book/journal titles) implied the suggestion of social tags not related to the content topics of the web page or scientific publication associated to the bookmark.

For plain text fields of the bookmark, such as title, abstract and descriptions, we filter out numeric characters and discard stop words from English, Spanish, French, German and Italian, which were identified as the predominant languages of the bookmarks available in our experimental datasets. We also carry out transformations to LaTeX expressions. Finally, we remove punctuation symbols, parentheses, and exclamation and question marks, and discard special terms like `paper`, `work`, `section`, `chapter`, among others. For the URL field, we firstly remove the network protocol (`HTTP`, `FTP`, etc.), the web domain (`com`, `org`, `edu`, etc.), the file extension (`html`, `pdf`, `doc`, etc.), and possible GET arguments for CGI scripts. Next, we tokenise the remaining text removing the dots (.) and slashes (/). Finally, we discard numeric words and several special words like `index`, `main`, `default`, `home`, among others. In both cases, a natural language processing tool [2] is used to singularise the resultant keywords, and filter out those that were not nouns.

Table 2 shows the content of an example bookmark whose tag recommendations are going to be explained in the rest of this section. It also lists the keywords extracted from the bookmark in the first stage of our approach. The bookmarked document is a scientific publication. Its main research fields are *recommender systems* and *semantic web technologies*. It describes a content-based collaborative recommendation model that exploits semantic (ontology-based) descriptions of user and item profiles.

Table 2. Example of bookmark for which the tag recommendation is performed, and the set of keywords extracted from it.

Title	**A Multilayer Ontology-based Hybrid Recommendation Model**
Authors	Iván Cantador, Alejandro Bellogín, Pablo Castells
URL	http://www.configworks.com/AICOM/
Journal title	AI Communications
Abstract	We propose a novel hybrid recommendation model in which user preferences and item features are described in terms of semantic concepts defined in domain ontologies. The concept, item and user spaces are clustered in a coordinated way, and the resulting clusters are used to find similarities among individuals at multiple semantic layers. Such layers correspond to implicit Communities of Interest, and enable enhanced recommendation.
Extracted keywords	multilayer, ontology, hybrid, recommendation, configwork, aicom, ai, communication, user, preference, semantic, concept, domain, ontology, item, space, way, cluster, similarity, individual, layer, community, interest

In this stage, we performed a simple mechanism to obtain a keyword-based description of the bookmarked document (web page or scientific publication) contents. Note that more complex approaches can be performed. For example, instead of only being limited to the bookmark attributes, we could also extract additional keywords from the bookmarked document itself. Moreover, external knowledge bases could be exploited to infer new keywords related to the ones extracted from the bookmark. These are issues to be investigated in future work.

4.2 Searching for similar bookmarks

The second stage (label 2 in Figure 1) consists of searching for bookmarks that contain some of the keywords obtained in the previous stage.

The list of keywords extracted from the input bookmark are weighted based on their appearance frequency in the bookmark attributes, and are included in a weighted keyword-based query. This query represents an initial description of the input bookmark.

More specifically, in the query q_n for bookmark b_n, the weight $q_{n,k} \in [0,1]$ assigned to each keyword k is computed as the number of times the keyword appears in the bookmark attributes divided by the total number of keywords extracted from the bookmark:

$$\mathbf{q_n} = q(b_n) = \{q_{n,1}, \dots, q_{n,k}, \dots, q_{n,K}\}$$

where

$$q_{n,k} = \frac{f_{n,k}}{\sum_{i=1}^{i=K} f_{n,i}},$$

being $f_{n,k}$ the number of times keyword k appears in bookmark b_n fields.

The query is then launched against the index described in Section 2. Thus, we are not only taking into account the relevance of the keywords for the input bookmark, but also ranking the list of retrieved similar bookmarks. The searching result is a set of bookmarks that are similar to the input bookmark, assuming that "similar" bookmarks have common keywords. Using the cosine similarity measure for the vector space model [14], the retrieved bookmarks are assigned scores $w_{n,i} \in [0,1]$ that measure the similarity between the query q_n (i.e., the input bookmark b_n) and the retrieved bookmarks b_i:

$$w_{n,i} = sim(q_n, b_i) = \cos(\mathbf{q_n}, \mathbf{b_i}) = \frac{\mathbf{q_n} \cdot \mathbf{b_i}}{\|\mathbf{q_n}\| \|\mathbf{b_i}\|}$$

For the example input bookmark, Table 3 shows the keywords, query, and some similar bookmarks obtained in the second stage of our tag recommendation model.

Table 3. Extracted keywords, generated query, and retrieved similar bookmarks for the example input bookmark.

Input bookmark: **A Multilayer Ontology-based Hybrid Recommendation Model**

Keywords	multilayer, ontology, hybrid, recommendation, configwork, aicom, ai, communication, user, preference, semantic, concept, domain, ontology, item, space, way, cluster, similarity, individual, layer, community, interest
Query	recommendation^0.125, ontology^0.09375, concept^0.0625, hybrid^0.0625, item^0.0625, layer^0.0625, multilayer^0.0625, semantic^0.0625, user^0.0625, aicom^0.03125, cluster^0.03125, configwork^0.03125, individual^0.03125, interest^0.03125, communication^0.03125, community^0.03125, preference^0.03125, similarity^0.03125, space^0.03125, way^0.03125
Similar bookmarks	• Improving Recommendation Lists Through Topic Diversification • Item-Based Collaborative Filtering Recommendation Algorithms • Probabilistic Models for Unified Collaborative and Content-Based Recommendation in Sparse-Data Environments • Automatic Tag Recommendation for the Web 2.0 Blogosphere using Collaborative Tagging and Hybrid ANN semantic structures • PIMO - a Framework for Representing Personal Information Models

In this stage, we attempted to define and contextualise the vocabulary that is likely to describe the contents of the bookmarked document. For that purpose, the initial set of keywords extracted from the input bookmark was used to find related bookmarks, assuming that the keywords and social tags of the latter are useful to describe the content topics of the former.

4.3 Obtaining related social tags

Once the set of similar bookmarks has been retrieved, in the third stage (label 3 in Figure 1), we collect and weight all their social tags.

The weight assigned to each tag represents how much it contributes to the definition of the vocabulary that describes the input bookmark. Based on the scores $w_{n,i}$ of the bookmarks retrieved in the previous stage, the weight v_n of a tag t for the input bookmark b_n is given by:

$$v_n(t) = \sum_{i:t \in \text{tags}(b_i)} w_{n,i}.$$

At this point, we could finish the recommendation process suggesting those social tags with highest weights v_n. However, doing this, we are not taking into account tag popularities and tag correlations, very important features of any collaborative tagging system. In fact, we conducted experiments evaluating recommendations based on the highest weighted tags, and we obtained worse results that the ones provided by the whole approach presented herein.

Table 4 shows a subset of the tags retrieved from the bookmarks that were retrieved in Stage 2 for the example input bookmark. The weights v_n for each tag are also given in the table.

Table 4. Weighted subset of tags retrieved from the list of bookmarks that are similar to the example input bookmark.

Input bookmark: **A Multilayer Ontology-based Hybrid Recommendation Model**

Related tag	*Weight*	*Related tag*	*Weight*	*Related tag*	*Weight*
recommender	10.538	clustering	2.013	dataset	0.871
recommendation	6.562	recommendersystems	1.669	evaluation	0.786
collaborative	5.142	web	1.669	suggestion	0.786
filtering	5.142	information	1.539	semantics	0.786
collaborativefiltering	3.585	ir	1.378	tag	0.786
ecommerce	3.138	retrieval	1.378	tagging	0.786
personalization	3.138	contentbasedfiltering	1.006	knowledgemanagement	0.290
cf	2.757	ontologies	1.006	network	0.290
semantic	2.745	ontology	1.006	neural	0.290
semanticweb	2.259	userprofileservices	1.006	neuralnetwork	0.290

In this stage, we collected the social tags that are potentially relevant for describing the input bookmarked document based on a set of related bookmarks. We assigned a weight to each tag capturing the strength of its contribution to the bookmark description. However, we realised that this measure is not enough for tag recommendation purposes, and global metrics regarding the folksonomy graph, such as tag popularities and tag correlations, have to be taken into consideration.

4.4 Building the global social tag co-occurrence sub-graph

In the fourth stage (label 4 in Figure 1), we interconnect the social tags obtained in the previous stage through the co-occurrence values of each pair of tags.

The co-occurrence of two tags t_i and t_j is usually defined in terms of the number of resources (bookmarks) that have been tagged with both t_i and t_j. In this work, we make use of the asymmetric co-occurrence metric:

$$co(t_i, t_j) = \frac{\#\{n: t_i \in \text{tags}(b_n) \wedge t_j \in \text{tags}(b_n)\}}{\#\{n: t_i \in \text{tags}(b_n)\}},$$

which assigns different values for $co(t_i, t_j)$ and $co(t_j, t_i)$ dividing the number of resources tagged with the two tags by the number of resources tagged with one of them.

Computing the co-occurrence values for each pair of tags existing in a training dataset, we build a global graph where the vertices correspond to the available tags, and the edges link tags that co-occur within at least one resource. This graph is directed and weighted: each pair of co-occurring tags is linked by two edges whose weights are the asymmetric co-occurrence values of the tags.

We propose to exploit this global graph to interconnect the tags obtained in the previous stage, and extract the ones that are more related with the input bookmark. Specifically, we create a sub-graph where the vertices are the above tags, and the edges are the same as these tags have in the global co-occurrence graph. From this sub-graph, we remove those edges whose co-occurrence values $co(t_i, t_j)$ are lower than the average co-occurrence value of the sub-graph vertices:

$$avg\_co(b_n) = \frac{\sum_{i,j} co(t_i, t_j)}{\#\{(i,j): co(t_i, t_j) > 0\}},$$

where t_i and t_j are the pairs of social tags related to the input bookmark b_n. Removing these edges, we aim to isolate (and later discard) "noise" tags that less frequently appear in bookmark annotations.

We hypothesise that vertices of the generated sub-graph that are most "strongly" connected with the rest of the vertices correspond to tags that should be recommended, assuming that high graph vertex centralities are associated to the most informative or representative vertices. In this context, it is important to note that related tags with high weights v_n do not necessarily have to be the ones with highest vertex centralities in the co-occurrence sub-graph. We hypothesise that a combination

of both measures – local weights representing the bookmark content topics and global co-occurrences taking into account collaborative popularities – is an appropriate strategy for tag recommendation.

Figure 2 shows the resultant co-occurrence graph associated to the tags retrieved from the example input bookmark. The tags with highest vertex in-degree seem to be good candidates to describe the contents of the bookmarked document.

Input bookmark: **A Multilayer Ontology-based Hybrid Recommendation Model**

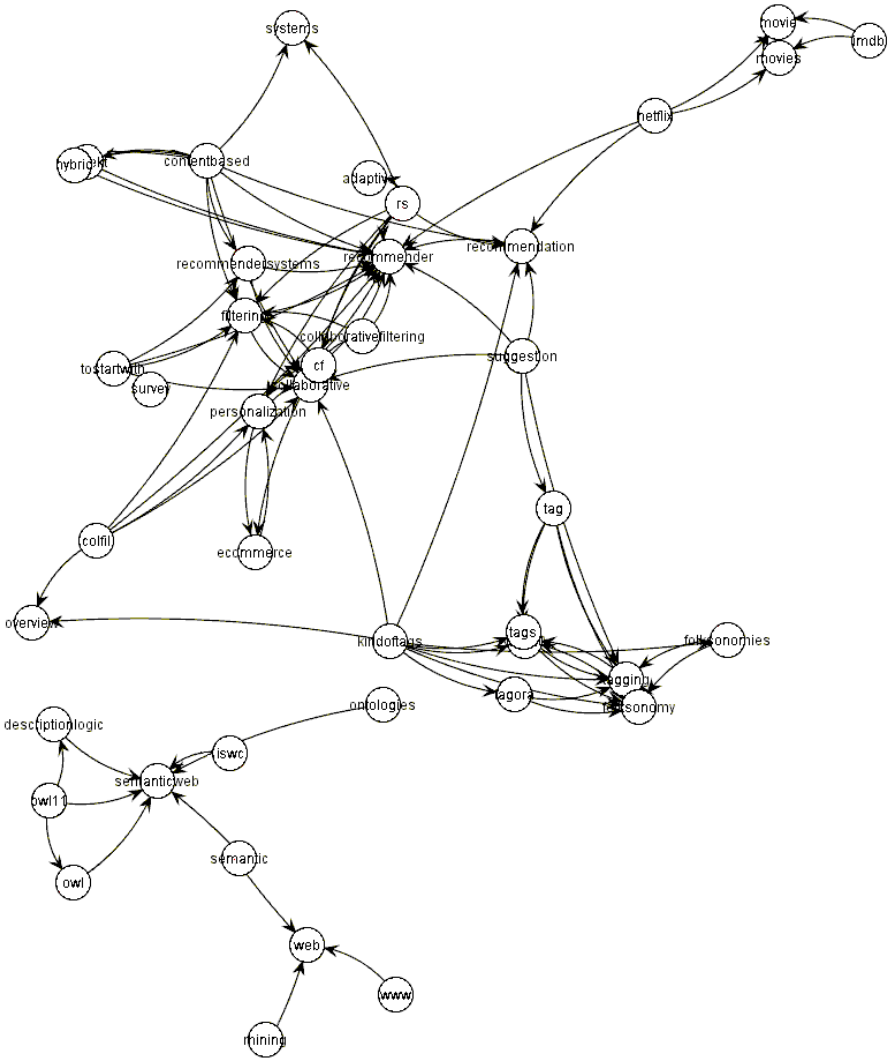

Figure 2. Filtered tag co-occurrence graph associated to the example input bookmark. Edge weights and non-connected vertices are not shown. Two main clusters can be identified in the graph, which correspond to two research areas related to the bookmarked document: recommender systems, and semantic web technologies.

The goal of this stage was to establish global relations between the social tags that are potentially useful for describing the input bookmark. Exploiting these relations, we aimed to take into account tag popularity and tag co-occurrence aspects, and expected to identify which are the most informative tags to be recommended.

4.5 Recommending social tags

In the fifth stage (label 5 in Figure 1), we select and recommend a subset of the related tags from previous stages. The selection criterion we propose is based on three aspects: the tag frequency in bookmarks similar to the input bookmark (stage 3), the tag co-occurrence graph centrality (stage 4), and a personalisation strategy that prioritises those tags that are related to the input bookmark and belong to the set of tags already used by the user to whom the recommendations are directed.

For each tag t, the first two aspects are combined as follows:

$$c_n(t) = in\_degree_n(t) \cdot (v_n(t))^2$$

where $in\_degree_n(t)$ is the number of edges that have as destination the vertex of tag t in the co-occurrence sub-graph built in stage 4 for the input bookmark b_n.

In order to penalise too generic tags we conduct a TF-IDF based reformulation of the centralities $c_n(t)$:

$$r_n(t) = c_n(t) \cdot log\left(\frac{N}{\#\{i: t \in \text{tags}(b_i)\}}\right)$$

where N is the total number of bookmarks in the repository.

Finally, to take into account information about the user's tagging activity, we increase the $r_n(t)$ values of those tags that have already been used by the user:

$$p_{n,u}(t) = r_n(t) \cdot (1 + p_u(t))$$

where $p_u(t)$ is the normalised preference of user u for tag t:

$$p_u(t) = \begin{cases} \dfrac{f_{u,t}}{\max\limits_{i \in \text{tags}(u)} f_{u,i}} & \text{if } t \in tags(u) \\ 0 & \text{otherwise} \end{cases},$$

$f_{u,i}$ being the number of times tag t has been used by user u.

The tags with highest preference values $p_{n,u}(t)$ constitute the set of final recommendations. Both the TF-IDF and personalisation based mechanisms were evaluated isolated and in conjunction with the baseline approach $c_n(t)$ improving its results.

Table 5 shows the final sorted list of tags recommended for the example input bookmark: `recommender`, `collaborative`, `filtering`, `semanticweb`, `personalization`. It is important to note that these tags are not the same as the top tags obtained in Stage 3 (see Table 4). In that case, all those tags (`recommender`, `recommendation`, `collaborative`, `filtering`, `collaborativefiltering`) were biased to vocabulary about "recommender systems", and no diversity in the suggested tags was provided.

Table 5. Final tag recommendations for the example input bookmark.

Input bookmark: **A Multilayer Ontology-based Hybrid Recommendation Model**

Tag 1	recommender
Tag 2	collaborative
Tag 3	filtering
Tag 4	semanticweb
Tag 5	personalization

In the fifth and last stage, we ranked the social tags extracted from the bookmarks similar to the input one. For that purpose, a combination of tag co-occurrence graph centrality, tag frequency, and tag-based personalisation metrics was performed. With an illustrative example, we showed that this strategy seems to offer more diversity in the recommendations than simply selecting the tags that more times were assigned to similar bookmarks.

5 Experiments

5.1 Tasks

Forming part of the ECML PKDD 2009 Discovery Challenge, two experimental tasks have been designed to evaluate the tag recommendations. Both of them get the same dataset for training, a snapshot of BibSonomy system until December 31st 2008, but different test datasets:

- **Task 1**. The test data contains bookmarks, whose user, resource or tags are not contained in the training data.
- **Task 2**. The test data contains bookmarks, whose user, resource or tags are all contained in the training data.

5.2 Datasets

Table 6 shows the statistics of the training and test datasets used in the experiments. Tag assignments (user-tag-resource) are abbreviated as *tas*.

Table 6. ECML PKDD 2009 Discovery Challenge dataset.

		Web pages	Scientific publications	All bookmarks
Training	*users*	2679	1790	3617
	resources	263004	158924	421928
	tags	56424	50855	93756
	tas	916469	484635	1401104
	tas/resource	3.48	3.05	3.32
Test (task 1)	*users*	891	1045	1591
	resources	16898	26104	43002
	tags	14395	24393	34051
	tas	64460	99603	164063
	tas/resource	3.81	3.82	3.82
Test (task 2)	*users*	91	81	136
	resources	431	347	778
	tags	587	397	862
	tas	1465	1139	3382
	tas/resource	3.40	3.28	4.35

5.3 Evaluation metrics

As evaluation metric, we use the average F-measure, computed over all the bookmarks in the test dataset as follows:

$$F\big(tags_p(u,b)\big) = \frac{2 \cdot precision\big(tags_p(u,b)\big) \cdot recall\big(tags_p(u,b)\big)}{precision\big(tags_p(u,b)\big) + recall\big(tags_p(u,b)\big)}$$

where:

$$recall\big(tags_p(u,b)\big) = \frac{\big|tags(u,b) \cap tags_p(u,b)\big|}{|tags(u,b)|}$$

$$precision\big(tags_p(u,b)\big) = \frac{\big|tags(u,b) \cap tags_p(u,b)\big|}{\big|tags_p(u,b)\big|}$$

being $tags(u,b)$ the set of tags assigned to bookmark b by user u, and $tags_p(u,b)$ the set of tags predicted by the tag recommender for bookmark b and user u.

For each bookmark in the test dataset, we compute the F-measure by comparing the recommended tags against the tags the user originally assigned to the bookmark. The comparison is done ignoring case of tags and removing all characters which are neither letters nor numbers.

5.4 Results

The tag recommendation approach presented in this work exploits training bookmark meta-information and tags, but does not analyse document contents, and does not make use of external knowledge bases, to enrich the set of suggested tags. Thus, all our recommended tags belong to the training collection, and our algorithm is only suitable for Task 2 of the ECML PKDD 2009 Discovery Challenge.

Table 7 shows recall, precision and F-measure values for the test datasets provided in the tasks. In task 2, recommending 5 tags, we reach an average F-measure value of 0.3065. We obtain a precision of 42% if we only recommend one tag, and 25% when we recommend 5 tags.

Table 7. Average recall, precision and F-measure values obtained in tasks 1 and 2 of ECML PKDD 2009 Discovery Challenge for different numbers of recommended tags.

	Number of recommended tags	**Recall**	**Precision**	**F-measure**
Task 1	*1*	0.0593	0.1810	0.0894
	2	0.0910	0.1453	0.1120
	3	0.1131	0.1233	0.1179
	4	0.1309	0.1091	0.1190
	5	0.1454	0.0991	0.1179
Task 2	*1*	0.1454	**0.4190**	0.2159
	2	0.2351	0.3477	0.2805
	3	0.2991	0.3059	0.3025
	4	0.3462	0.2716	0.3044
	5	0.3916	0.2518	**0.3065**

6 Conclusions and future work

In this work, we have presented a social tag recommendation model for a collaborative bookmarking system. Our approach receives as input a bookmark (of a web page or a research publication), analyses and processes its textual metadata (document title, URL, abstract and descriptions), and suggests tags relevant to bookmarks whose metadata are similar to those of the input bookmark.

Besides focusing on those tags that best fit the bookmark metadata, our strategy also takes into account global characteristics of the system folksonomy. More specifically, it makes use of the tag co-occurrence graph to compute vertex centralities of related tags. Assuming that tags with higher vertex centralities are more informative to describe the bookmark contents, our model weights the retrieved tags through their centrality values in a small co-occurrence sub-graph generated for the input bookmark. As additional features, the weighting mechanism also penalises tags that are too generic, and strengthens tags that have been previously used by the user to whom the tag recommendations are conducted.

Two are the main benefits of our approach: a low computational cost, and the capability of providing diversity in the recommended tag sets. On one hand, an index of keywords and tags for the available bookmarks, and the global tag co-occurrence graph, are the only information resources needed. On the other hand, the combination of exploiting content-based features, tag popularity and personalisation in the recommendation process allows suggesting tags that not only are relevant for the input bookmark, but also might belong to different domains.

A main drawback of our approach is its limitation to recommend tags that already exist in the system folksonomy. The suggestion of new terms, for example extracted from the bookmarked text contents or from external knowledge bases such as dictionaries or thesauri, is thus an open research line.

More investigation is needed to improve and evaluate the effectiveness of our tag recommender. In this context, the study of alternative graph vertex centrality measures (e.g. [11]), and the exploitation of extra folksonomic information obtained from the user and item spaces (e.g., as done in [6]), represent priority tasks to address in the future. The evaluation has to be also done comparing our approach with other state-of-the-art techniques.

Acknowledgments. This research was supported by the European Commission under contracts FP6-027122-SALERO, FP6-033715-MIAUCE and FP6-045032 SEMEDIA. The expressed content is the view of the authors but not necessarily the view of SALERO, MIAUCE and SEMEDIA projects as a whole.

References

1. Adomavicius, G., Tuzhilin, A. 2005. *Toward the Next Generation of Recommender. Systems: A Survey of the State-of-the-Art and Possible Extensions*. In IEEE Transactions on Knowledge and Data Engineering, pp. 734-749.
2. Alfonseca, E., Moreno-Sandoval, A., Guirao, J. M., Ruiz-Casado, M. 2006. *The Wraetlic NLP Suite*. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006).
3. Byde, A., Wan, H., Cayzer, S. 2007. *Personalized Tag Recommendations via Tagging and Content-based Similarity Metrics*. In Proceedings of the 2007 International Conference on Weblogs and Social Media.
4. Chirita, P. A., Costache, S., Handschuh, S., Nejdl, W. 2007. *P-TAG: Large Scale Automatic Generation of Personalized Annotation TAGs for the Web*. In Proceedings of the 16th International Conference on World Wide Web (WWW 2007), pp. 845-854.
5. Heymann, P., Ramage, D., Garcia-Molina, H. 2008. *Social Tag Prediction*. In Proceedings of the 31st Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR 2008), pp. 531-538.
6. Hotho, A., Jäschke, R., Schmitz, C., Stumme, G. 2006. *Information Retrieval in Folksonomies*. 2006. In Proceedings of the 3rd European Semantic Web Conference (ESWC 2006), pp. 411-426.
7. Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., Stumme, G. 2008. *Tag Recommendations in Social Bookmarking Systems*. In AI Communications, 21, pp. 231-247.

8. Kleinberg, J. 1999. *Authoritative Sources in a Hyperlinked Environment*. In Journal of the ACM, 46(5), pp. 604–632.
9. Li, J., Zha, H. 2006. T*wo-way Poisson Mixture Models for Simultaneous Document Classification and Word Clustering*. In Computational Statistics & Data Analysis, 50(1), pp. 163-180.
10. Mishne, G. 2006. *AutoTag: A Collaborative Approach to Automated Tag Assignment for Weblog Posts*. In Proceedings of the 15th International Conference on World Wide Web (WWW 2006), pp. 953-954.
11. Newman, M. E. J. 2005. *A measure of Betweenness Centrality based on Random Walks*. In Social Networks, 27, pp. 39–54.
12. Page, L., Brin, S., Motwani, R., Winograd, T. 1999. *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report. Stanford InfoLab.
13. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J. 1994. *GroupLens: An Open Architecture for Collaborative Filtering of Netnews*. 1994. In Proceedings of the 1994 ACM conference on Computer Supported Cooperative Work (CSCW 1994), pp. 175-186.
14. Salton, G., McGill, M. J. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York.
15. Song, Y., Zhuang, Z., Li, H., Zhao, Q., Li, J., Lee, W. C., Giles, C. L. 2008. *Real-time Automatic Tag Recommendation*. In Proceedings of the 31st Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR 2008), pp. 515-522.
16. Tatu, M., Srikanth, M., D'Silva, T. 2008. *RSDC'08: Tag Recommendations using Bookmark Content*. In Proceedings of the ECML PKDD 2008 Discovery Challenge (RSDC 2008).
17. Xu, Z., Fu, Y., Mao, J., Su, D. 2006. *Towards the Semantic Web: Collaborative Tag Suggestions*. In Proc. of the WWWW 2006 Workshop on Collaborative Web Tagging.
18. Zha, H., He, X., Ding, C., Simon, H. 2001. *Bipartite Graph Partitioning and Data Clustering*. In Proceedings of the 10th ACM International Conference on Information and Knowledge (CIKM 2001), pp. 25-32.