RSDC'08: Tag Recommendations using Bookmark Content

Marta Tatu, Munirathnam Srikanth, and Thomas D'Silva

Lymba Corporation, Richardson, TX, 75080, USA
marta,srikanth,tsilva@lymba.com

Abstract. A variety of factors contribute to a tag being assigned by a user to a document that he or she bookmarked. Textual information present in a URL's title, a user's description of a document, or a bibtex field associated with a scientific publication are sources for automatically recommending tags relevant for a given bookmark. Lymba's submission for the RSDC'08 Tag Recommendation task uses document and user models derived from the textual content associated with URLs and publications by social bookmarking tool users. This paper describes our natural language understanding approach for producing tag recommendations and provides some initial results of our internal evaluation.

1 Introduction

Social bookmarking tools enable people to label content of interest with descriptions from their own vocabulary which facilitate easy recall. When shared in a collaborative setting, the bookmark labels enable users to discover new content and other users with similar or shared interests. While user interests and their vocabulary play a significant part in what is assigned as a label to a given document, the content of the document is the driver of the bookmarking event and the basis of the assigned labels and hence it provides automated systems important clues about the tags that can be assigned to a given document. Among other advantages, automatic tag recommendations help bookmarking systems manage the tag space and direct it towards the standardization of the labels that users assign to documents in order to describe them.

For the RSDC'08 Tag Recommendation task, we used the textual content associated with bookmarks to model documents (web pages and publications) and users based on their tagging and suggest tags for new bookmarks. A combination of statistical and semantic features are used to build document and user models. Section 2 presents the different "spaces" in which documents, bookmarks and users can be modeled and the process that generates these models. The different methods used to recommend tags for bookmarks are discussed in Section 3 followed by a description of the data processing that was performed on the RSDC'08 training and test data sets to build a representation of bookmarks,

documents and users. The results of our experiments and our observations follow.

1.1 Previous Work

Automatic labeling of documents based on predefined categories has been explored in document classification systems. The labels or categories are modeled based on the terms extracted from content or document metadata. In social tag prediction, the set of labels is not fixed and the explicit association of users to documents adds additional dimensions to the classification or prediction problem. User's preferences for tags based on their interests, workflow, etc. play an essential part in the tags associated with content. Tags like *myown* are relevant to a particular user and do not add value in collaborative settings.

A number of approaches to social tag prediction have used information retrieval methods to identify similar documents and recommend their tags for a given document. AutoTag [1] suggests tags to weblogs based on the tags associated with other similar weblogs in a given collection. While this approach depends on the content of the document, no new tag (from the content) is suggested. Some approaches used the collaborative or social aspect of the data to recommend tags based on user similarity. Recently, Heymann et al. [2] use a large dataset of del.icio.us[1] links to evaluate the effectiveness of using text and inlinks information to model and predict tags for documents. A bag of words model with TFIDF weighting of terms is used to represent the features and solve tag prediction as a text classification problem. We use textual content associated with bookmarks to model users and documents and suggest tags not only from the existing tag space (generated from the training data), but also from the textual content associated with bookmarks.

2 Document and User Models

A bookmark b_i is identified by the triple $(u_i, d_i, \{t_{ij}\}_j)$ corresponding to "user u_i bookmarked document d_i by assigned it the set of tags $\{t_{ij}\}_j$". A bookmark b_i can be associated with its corresponding textual metadata. For instance, the metadata of a web document d_i bookmarked by a user u_i can include the actual URL, its title and any user-given description. For scientific publications, this metadata can include the title and author of the paper, the journal where the paper was published, etc. In addition to metadata information provided by users for their bookmarks, the textual content of bookmarked documents (d_i) can be exploited. For scientific publications, this is the textual content of the paper.

[1]

Our system uses the textual content associated with bookmarks and documents to model users and the documents they bookmark.

Given a bookmark that associates document d_0 with user u_0, the tag recommendation problem can be solved by estimating the likelihood $P(t_i|d_0, u_0)$ of a tag t_i being assigned to the bookmark (u_0, d_0). The recommended tags can exist in the tag space derived from the training data. However, concepts derived from the content associated with both d_0 and u_0 can also be suggested as tag recommendations. Joint modeling of documents and users in a common feature space provides the desired tag ranking for our recommendation system. For the RSDC'08 data, the textual content of the metadata provided for each (user,document) pair is used to represent each bookmark. Therefore, the representation of a document d_0 becomes the union of the representations of all its bookmarks $\{(u_i, d_0)\}_i$ (a document is modeled using the descriptions given to the document by all users that bookmarked the document). Similar combinations of bookmark representations provide the features needed to create user models.

For the RSDC'08 Discovery Challenge, we used and evaluated different feature representations depending on the data type and the adopted tag recommendation approach. A suite of natural language processing tools were used to understand the textual content associated with each bookmark (u_i, d_i). We extracted important concepts (nouns, adjectives and named entities) from the textual metadata associated with each bookmark and used semantic analysis to generate normalized versions of the concepts. The normalization process makes use of different lexico-semantic resources, e.g., WordNet[2] to stem the concepts and link synonyms. For instance, the concepts *European Union*, *EU*, and *European Community* (in all their case variations) are normalized to the same concept *european_union*. By representing bookmarks, documents and users in the *concept space* created by the concepts identified in all textual metadata as collections of normalized concepts – each associated with a corresponding weight, our tag recommendation system is able to suggest as tag recommendations concepts that do not belong to the existing tag space.

In order to derive a common feature space to compare documents, users, bookmarks and tags, we developed a conflation method for grouping tags into semantically related groups of tags referred to as *conflated tags*. This process of tag normalization takes care of spelling mistakes, abbreviations, joined concepts and provides a common representation for synonyms. For instance, *weblog* is one of the conflated tags derived for the RSDC'08 data. It conflates the following list of space-separated tags: *blog Blog weblog blogs Weblog weblogs blog, BLOGS Blogs Weblogs bloga bloging blogs, weblogs, Blog. we-*

[2] http://wordnet.princeton.edu

blogs_weblog_blogs_blog weblog_blog to_blog webology bl;ogs Blogs! ??blog blogr Blogs, BLOG Blog"> Blog: blogblogs. The set of conflated tags define a *tag space* that can be used to model bookmarks, documents and users. For this purpose, we use the mapping between normalized concepts and conflated tags provided by the underlying ontology (WordNet) to create a bookmark, document and user representation within the existing tag space.

We note that our mechanism for normalizing concepts and conflating tags was developed for the English language and has not been customized to handle multi-lingual data. This is reflected in our tagging results for non-English content and tags.

3 Recommending Tags for Bookmarks

Given the bookmark, document and user representations detailed in Section 2, let us consider the following notations: for a bookmark $b_i = (u_i, d_i, \{t_{ij}\}_j)$

- $TH(b_i) =$ the set of tag conflations derived for $\{t_{ij}\}_j$ – the tags assigned by user u_i to document d_i; $TH(b_i) \subseteq TagSpace$,
- $TC(b_i) = b_i$'s representation in the concept space (the set of normalized concepts evoked by b_i's textual content); $TC(b_i) \subseteq ConceptSpace$, and
- $TT(b_i) = b_i$'s representation in the tag space (the set of conflated tags evoked by b_i's concepts); $TT(b_i) = TC(b_i)$'s projection to the $TagSpace$; $TT(b_i) \subseteq TagSpace$.

For a document d_0, $TH(d_0) = \cup_i TH(b_i)$, $TC(d_0) = \cup_i TC(b_i)$, and $TT(d_0) = \cup_i TT(b_i)$ where $b_i = (u_i, d_0, \{t_{ij}\}_j)$ is a bookmark provided in the training data. Similar definitions can be derived for a user u_0. We note that TC and TT are independent of the tags that user u_i assigns to document d_i. They are solely based on the textual content that the user associates with the document. Let us also denote $TT(b_i) \cup TH(b_i)$ by $THT(b_i)$ – the set of tags that can be associated with bookmark b_i. It includes the tags assigned by humans as well as tags derived from the textual content of the bookmarked document. $THT(b_i) \subseteq TagSpace$.

3.1 Recommending Existing Tags

Given the models defined in Section 2 and the notations introduced above, tag recommendations derived from the existing tag space for a given bookmark (u_0, d_0) are generated by $[THT(d_0) \cap THT(u_0)] \cup TT(d_0, u_0)$. These are existing tags evoked by the textual content of both d_0 and u_0. They may include tags previously assigned to d_0 by other users (if d_0 is part of the training data,

$TH(d_0) \neq \emptyset$) or by u_0 to other documents (if u_0 is part of the training data, $TH(u_0) \neq \emptyset$).

3.2 Suggesting New Tags

In order to recommend tags that do not currently exist in the tag space, we make use of the document and user representations in the concept space. Therefore, for a given bookmark (u_0, d_0), recommendations are generated using $[THT(d_0) \cap THT(u_0)] \cup TC(d_0, u_0)$. Lymba's submission for the RSDC'08 Challenge made use of this tag recommendation mechanism.

4 Data Preparation and Processing

4.1 Experimental Data

For the RSDC'08 Challenge, we received bookmarking data from BibSonomy[3], a web-based social bookmarking system that enables users to tag web documents as well as bibtex entries of scientific publications. Brief statistics of the data can be found in Table 1.

Table 1. RSDC'08 bookmarking data

	Bookmarks	Publications
Training	176,147	92,545
Average no. of tags	3.38	2.37
Test	16,195	43,348
Average no. of tags	2.12	2.27

Each bookmark was described by its URL (e.g., *http://www.bibsonomy.org/*), a description of the URL (e.g., *BibSonomy::home*) that usually maps to its title and an extended description of the bookmark (e.g., *BibSonomy is a system for sharing bookmarks and lists of literature.*). We note that the user that is bookmarking a URL has complete control over the bookmark's descriptions.

Each bookmarked publication is associated with values of bibtex fields such as *title*, *author*, *booktitle*, *journal*, *series*, *editor*, *publisher*, *volume*, *number*, *pages*, etc. In addition to this information, the *entrytype*, *bibtexKey*, *bibtexabstract*, and *URL* can be specified. The user can also input their own description of the publication. Miscellaneous information is collected in the *misc* field. This may include user comments, non-standard bibtex fields, e.g., *isbn* (1532-0626

[3] http://www.bibsonomy.org

(print), 1532-0634 (electronic)), *doi* (10.1002/cpe.607), and *bibdate* (Mon Feb 25 14:51:24 MST 2002). For instance, one of the publications bookmarked by user *26* is described by the information shown in Table 2.

Table 2. Bookmarked publication in BibSonomy

Title	Temporal and real-time databases: a survey
Author	G. Ozsoyoglu and R. T. Snodgrass
Journal	Knowledge and Data Engineering, IEEE Transactions on
Volume	7
Number	4
Pages	513–532
Year	1995
EntryType	article
BibtexKey	ozso95
BibtexAbstract	A temporal database contains time-varying data. In a real-time database transactions have deadlines or timing constraints. In this paper we review the substantial research in these two previously separate areas. First we characterize the time domain; then we investigate temporal and real-time data models. We evaluate temporal and real-time query languages along several dimensions. We examine temporal and real-time DBMS implementation. Finally, we summarize major research accomplishments to date and list several unanswered research questions
URL	http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=404027
Misc	- I'm not clear why temporal databases are used. Does Amazon use them? I suspect they just annotate the row with a timestamp as needed. - paper surveys variety of research, quite extensive. Tries to combine temporal dbs with real-time dbs.
Description	sdasda

4.2 Data Normalization

Bookmarks are described by their URL. Thus, we planned to use the *url_hash* information associated with each bookmark to identify the unique list of bookmarked webpages. However, multiple URLs can pinpoint to the same webpage. Table 3 lists the different URLs found in the training data that identify the *BibSonomy* main webpage. Therefore, our data preparation process included an URL normalization step that unified the bookmarked URLs.

As a result of this step, several bookmarks were merged into a single bookmark to which we assigned the descriptions of the given bookmarks. For instance, seven URLs bookmarked by user *1339* point to the same webpage (Table 4). Clearly, user *1339* intended to bookmark different sections of the webpage (he also bookmarked the entire page), however, using existing social bookmarking tools, users cannot explicitly bookmark portions of documents. In Ta-

Table 3. URLs describing BibSonomy

http://www.bibsonomy.org/
http://www.bibsonomy.org
http://bibsonomy.org/
http://bibsonomy.org

Table 4. Duplicate bookmarks were merged

User Id	URL	Tag(s)
1339	http://www.gehspace.com/arte26a30.htm#26	arte
1339	http://www.gehspace.com/arte26a30.htm#27	arte
1339	http://www.gehspace.com/arte26a30.htm#28	arte
1339	http://www.gehspace.com/arte26a30.htm#inicio	arte
1339	http://www.gehspace.com/arte26a30.htm#29	arte
1339	http://www.gehspace.com/arte26a30.htm#30	arte
1339	http://www.gehspace.com/arte26a30.htm	arte
1339	http://www.gehspace.com/arte26a30.htm	arte

ble 4's last row, we display the bookmark that unifies the seven (URL,tags) pairs.

A similar unification process was performed on the set of bookmarked publications. For publications, we used the *simhash1* field information. In Table 5, we display one example.

Table 5. Duplicate publications were merged

User Id	Publication	Tag(s)
82	Title: *How triadic diagrams represent conceptual structures*, Author: *Klaus Biedermann*, Series: *LNAI*	triadic formal concept lattices analysis fba begriffsanalyse forschungsgruppe tu ag1 formale darmstadt fca
82	Title: *How Triadic Diagrams Represent Conceptual Structures*, Author: *K. Biedermann*, Series: *Lecture Notes in Computer Science*	FCA OntologyHandbook
82	Title: *How Triadic Diagrams Represent Conceptual Structures*, Author: *K. Biedermann*, Series: *LNAI*	FCA OntologyHandbook
82	Title: *How triadic diagrams represent conceptual structures*, Author: *Klaus Biedermann*, Series: *LNAI, Lecture Notes in Computer Science*	triadic formal concept lattices analysis fba begriffsanalyse forschungsgruppe tu ag1 formale darmstadt fca ontologyhandbook

A second preprocessing stage was performed for all publications whose *misc* field was comprised of (field,value) pairs. A manual review of all bib-

tex fields stored in *misc* by one of the authors of this paper resulted in a map that augments the existing information stored for a publication. For instance, the values stored within *misc* for the field name *englishtitle* were added to the *title* information. Similarly, the *confname* values augmented the *booktitle* information. However, not all bibtex fields stored within *misc* could be mapped to already existing columns. Therefore, new fields were generated and populated with the corresponding values as stored in the *misc* field. Among the new columns, we have *classification, contents, keywords, review, subject, topic*, and others.

Given that we had this rich meta information for only some of the publications, we attempted to fill the missing information by exploiting the data that digital libraries such as *ieeexplore.ieee.org, portal.acm.org, or sciencedirect.com* provide for the publications that they host. Thus, for any publication whose URL pointed to one of these websites, we downloaded the html files, parse them, extracted any piece of information that belonged to the publication's metadata and stored them in the corresponding bibtex fields. We note that the html files downloaded from these resources have a uniform structure and can be easily parsed.

An additional data preparation step was required for the RSDC'08 data. Most of the textual information that is associated with scientific publications contains LaTeX markings. By removing these markings, each publication was associated with plain text that can be accurately processed by our natural language processing tools.

4.3 Data Processing

As mentioned in Section 2, each bookmark is associated with the textual content of the metadata that accompanies it and certain concepts are derived from each piece of information. However, the importance of the identified concepts differs from column to column. For instance, the concepts derived from a publication's *address* field are less important than one ones derived from the *journal* information which are less important than the *title* concepts when it comes to the concept's likelihood to become a tag assigned to the document. Therefore, when we aggregated the concept information for a bookmark for the purpose of deriving the bookmark's representation in the concept space ($TC(b_i)$), we used various field weights that denoted the importance of the field. For Lymba's RSDC'08 submission we used a set of heuristics to assign weights to the different fields that describe the URLs and the publications. However, we plan to learn these weights from the training data within a machine learning framework.

5 Experiments and Observations

For the RSDC'08 challenge, we normalized the test data according to the processing described in Section 4.2. For the testing bookmarks/publications that mapped to bookmarks existing in the training data, we returned the tags assigned to these bookmarks according to the training data. 602 bookmarks/publications were tagged by this process with a maximum average F-measure of 96.95% (at top 1) and a minimum average F-measure of 96.70% (at top 2).

5.1 Results

In Table 6, we show the performance of the system on the RSDC'08 testing data. The highest F-measure (21.33%) is achieved for the top 5 tag recommendations with the highest precision at top 1 and highest recall at top 10. Our measures of the system's performance on the two types of bookmarks (URLs and publications) – also shown in Table 6 – releaved a significant difference between the quality of the tag recommendations made for a publication (highest F-measure – at top 5 – is 27%) and the quality of the recommendations made for a bookmark (highest F-measure : 7%).

Table 6. Performance on the test data (average recall, precision and f-measure over the testing bookmarks/publications for the top N tag recommendations – $N \in \{1, \ldots, 10\}$). Performance numbers are shown for the entire test data, only for bookmarks, and only for publications.

Top N	Recall	Precision	F-measure
1	0.2062 : 0.0700 / 0.2572	0.2062 : 0.0700 / 0.2572	0.2062 : 0.0700 / 0.2572
2	0.2089 : 0.0656 / 0.2625	0.1892 : 0.0629 / 0.2365	0.1986 : 0.0643 / 0.2488
3	0.2515 : 0.0665 / 0.3207	0.1749 : 0.0586 / 0.2185	0.2063 : 0.0623 / 0.2599
4	0.3005 : 0.0684 / 0.3873	0.1635 : 0.0537 / 0.2046	0.2118 : 0.0601 / 0.2678
5	0.3468 : 0.0717 / 0.4497	0.1540 : 0.0502 / 0.1929	0.2133 : 0.0591 / 0.2700
6	0.3894 : 0.0745 / 0.5072	0.1460 : 0.0469 / 0.1830	0.2123 : 0.0575 / 0.2690
7	0.4266 : 0.0792 / 0.5565	0.1389 : 0.0457 / 0.1737	0.2096 : 0.0579 / 0.2648
8	0.4592 : 0.0816 / 0.6004	0.1322 : 0.0437 / 0.1653	0.2053 : 0.0569 / 0.2592
9	0.4865 : 0.0844 / 0.6369	0.1257 : 0.0424 / 0.1569	0.1998 : 0.0564 / 0.2517
10	0.5073 : 0.0860 / 0.6649	0.1193 : 0.0412 / 0.1485	0.1932 : 0.0557 / 0.2428

A closer look at the set of 16,194 bookmarks given in the test dataset revealed 9,183 bookmarks were assigned a single tag (*indexforum*) which was not among the top 10 tag recommendations returned by our system. These bookmarks are various articles from *http://forum.index.hu*. Table 7 displays three of these bookmarks. The tag assigned to these URLs can be derived from their common URL. However, our system recommends the concepts it identified in

the bookmark's content without joining them. For instance, our tag recommendations for the second URL shown in Table 7 are *tft, monitorok, forum, article, la, index, 9156398, hu,* and *78013830.*

Table 7. Quality of out-of-tag-space recommendations and in-tag-space recommendations measured on the test data

URL	Title
http://forum.index.hu/Article/showArticle?t=9096213&la=73045864	tomor, szines programozasi nyelv
http://forum.index.hu/Article/showArticle?t=9156398&la=78013830	TFT monitorok
http://forum.index.hu/Article/showArticle?t=9013231&la=34610778	Temetni jottem a Linuxot, nem dicserni!

We evaluated the system's performance on the test data from which we eliminated the 9,183 URLs from *http://forum.index.hu*. The performance, as it can be seen in Table 8, improves by 4% (F-measure of top 5 tag recommendations). Our immediate plans are to expand our system to include joined concepts in the tag recommendations that it makes.

Table 8. System performance on the test data excluding the *http://forum.index.hu* bookmarks

Top N	Recall	Precision	F-measure
1	0.2062 / 0.2439	0.2062 / 0.2439	0.2062 / 0.2439
2	0.2089 / 0.2470	0.1892 / 0.2238	0.1986 / 0.2348
3	0.2515 / 0.2974	0.1749 / 0.2069	0.2063 / 0.2440
4	0.3005 / 0.3554	0.1635 / 0.1934	0.2118 / 0.2505
5	0.3468 / 0.4101	0.1540 / 0.1822	0.2133 / 0.2523
6	0.3894 / 0.4605	0.1460 / 0.1726	0.2123 / 0.2511
7	0.4266 / 0.5045	0.1389 / 0.1642	0.2096 / 0.2478
8	0.4592 / 0.5431	0.1322 / 0.1563	0.2053 / 0.2428
9	0.4865 / 0.5754	0.1257 / 0.1486	0.1998 / 0.2363
10	0.5073 / 0.6000	0.1193 / 0.1411	0.1932 / 0.2285

The difference in the system's performance on the bookmark data as opposed to the test publications (Table 6) may also be cause by the difference in textual content that the system associated with each bookmark/publication. The bibtex entries were much richer in textual content when compared with the URL descriptions.

5.2 Impact of Tag Space Expansion

Our approach to tag recommendations (described in Section 3) enables us to limit our suggestions to the existing tag space as well as expand the set of tag recommendations to include concepts the were not previously assigned as tags. In this section, we show the impact that the process of recommending out-of-tag-space concepts had on the performance of our system. In Table 9, we show the performance of the system when we limited the tag recommendations to the tag space generated based on the training data. The highest F-measure, 10.04% – achieved for the top 4 tag recommendations, is significantly smaller when compared with the system's F-measure when it recommends concepts that do not exist in the current tag space. The understanding of the bookmark/publication's content (the textual information that accompanies the URL/publication) doubles the quality of the tag recommendations.

Table 9. Quality of out-of-tag-space recommendations and in-tag-space recommendations measured on the test data

Top N	Recall	Precision	F-measure
1	0.2062 / 0.0927	0.2062 / 0.0927	0.2062 / 0.0927
2	0.2089 / 0.0935	0.1892 / 0.0868	0.1986 / 0.0900
3	0.2515 / 0.1137	0.1749 / 0.0842	0.2063 / 0.0968
4	0.3005 / 0.1345	0.1635 / 0.0801	0.2118 / 0.1004
5	0.3468 / 0.1498	0.1540 / 0.0748	0.2133 / 0.0998
6	0.3894 / 0.1600	0.1460 / 0.0693	0.2123 / 0.0967
7	0.4266 / 0.1660	0.1389 / 0.0639	0.2096 / 0.0922
8	0.4592 / 0.1696	0.1322 / 0.0593	0.2053 / 0.0879
9	0.4865 / 0.1718	0.1257 / 0.0557	0.1998 / 0.0842
10	0.5073 / 0.1732	0.1193 / 0.0531	0.1932 / 0.0813

6 Conclusion

In this paper, we briefly describe Lymba's approach for generating tag recommendations for bookmarks/publications. We exploit the textual content that can be associated with bookmarks, documents and users and generate models within the concept space or the existing tag space. Using these rich models, our tag recommendation system is able to suggest various concepts as tags for a given bookmark. The quality of the tag recommendations generated from concept space (without being constrainted to the tag space produced by the training data) exceeds by far the appropriateness of the tags suggested from the current tag space.

Further analysis is required to identify which features contribute and by how much, towards a particular tag. The semantic relation between the tags in the tag space can be exploited to obtain better weighting of tags and improved models for documents and users.

Our future work will focus also on designing an “adaptive tag recommendation” system which uses the chronology of the bookmarking events to grow the tag space and learn from all previously stored bookmarks.

References

1. Mishne, G.: Autotag:a collaborative approach to automated tag assignment to weblog posts. In: Proceedings of WWW’06, Edinburgh, Scotland (2006)
2. Heymann, P., Ramage, D., Garcia-Molina, H.: Social tag prediction. In: Proceedings of SIGIR’08, Singapore (2008)