Recommending Tags with a Model of Human Categorization

Paul Seitlinger
Knowledge Technologies Institute
Graz University of Technology
Graz, Austria
paul.seitlinger@tugraz.at

Dominik Kowald
Know-Center
Graz University of Technology
Graz, Austria
dkowald@know-center.at

Christoph Trattner
Know-Center
Graz University of Technology
Graz, Austria
ctrattner@know-center.at

Tobias Ley
Institute of Informatics
Tallinn University
Tallinn, Estonia
tley@tlu.ee

ABSTRACT

When interacting with social tagging systems, humans exercise complex processes of categorization that have been the topic of much research in cognitive science. In this paper we present a recommender approach for social tags derived from ALCOVE, a model of human category learning. The basic architecture is a simple three-layers connectionist model. The input layer encodes patterns of semantic features of a user-specific resource, such as latent topics elicited through Latent Dirichlet Allocation (LDA) or available external categories. The hidden layer categorizes the resource by matching the encoded pattern against already learned exemplar patterns. The latter are composed of unique feature patterns and associated tag distributions. Finally, the output layer samples tags from the associated tag distributions to verbalize the preceding categorization process. We have evaluated this approach on a real-world folksonomy gathered from Wikipedia bookmarks in Delicious. In the experiment our approach outperformed LDA, a well-established algorithm. We attribute this to the fact that our approach processes semantic information (either latent topics or external categories) across the three different layers. With this paper, we demonstrate that a theoretically guided design of algorithms not only holds potential for improving existing recommendation mechanisms, but it also allows us to derive more generalizable insights about how human information interaction on the Web is determined by both semantic and verbal processes.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information filtering*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

Copyright 2013 ACM 978-1-4503-2263-8/13/10 ...$15.00.

http://dx.doi.org/10.1145/2505515.2505625.

Keywords

personalized tag recommendations; LDA; human categorization; Wikipedia; Delicious

1. INTRODUCTION

There is now broad agreement that in order to support users in tagging resources on the Web, a good understanding of the mechanisms that underlie human tagging behavior is advantageous [4, 6]. Based on models of information theory [6] and human memory theory [4] generative models of social tagging have been developed providing much insight into the emergence of the data observed in social tagging systems. The generative models implement cognitive assumptions about human information processing and provide computational models that predict a tag distribution. Comparing the theoretical to the empirical tag distribution then allows making claims about the validity of the underlying cognitive assumptions. A stricter test of theoretical claims can be provided by controlled experiments as these allow for testing causal relationships more directly. Such studies have been conducted, for instance, by Fu et al. [4] to test the semantic imitation model of social tagging, by Cress et al. [3] to test a social variant of information foraging theory, as well as by our own group to find evidence for a dual-process memory mechanism [22]. These studies, on the other hand, have limitations as they need to necessarily control the setting in which they are conducted. To generalize findings from these lab settings to naturally occurring tagging behavior, the models need to be tested in real-life settings. In this paper, we have devised a recommender mechanism which implements some basic mechanisms of human categorization that is assumed to take place in social tagging environments. When contrasting predicted with observed tag choices, this provides validation of the underlying model. Additionally, this approach allows several algorithms and their underlying models to be compared to each other.

The contributions of the paper are threefold: (1) We present a novel tag recommendation mechanism that is based on psycholinguistic models of categorization and speech production, (2) We demonstrate that such a recommendation mechanism performs significantly better than a standard tag recommendation approach such as LDA, and (3) We demonstrate that a theoretically guided design of a recommender complements data-driven approaches in that it allows for learning something about how humans process information in sensemaking tasks on the Web.

The remainder of the paper is structured as follows: In Section 2 we discuss related work and in Section 3 we present our new approach. In Section 4 we shortly introduce our experiments and the used dataset. Section 5 presents the results of our study. Section 6 concludes the paper and discusses our findings in light of the benefits of connecting data-driven and theory-driven research for recommender systems research. Finally, Section 7 outlines future research directions.

2. RELATED WORK

In contrast to the research on generative models of social tagging mentioned above, recommender systems research has taken a more pragmatist stance, such as helping users discover useful resources on the Web, or improving the overall tag consistency. One of those approaches that have been very successful in predicting and recommending tags for Web resources has been collaborative filtering [7]. The first work describing such a mechanism for the domain of collaborative tagging systems is the work of Xu et al. [26] who introduced a simple The probably first work describing such a mechanism for the domain of collaborative tagging systems is the work of Xu et al. [25] who introduce a simple tag co-occurrance approach to recommend tags to a user. Sigurbjornsson et al. [24] developed a similar approach and showed for the photo tagging system Flickr that it is "essential to take the co-occurance values of the candidate tags into account when aggregating the intermediate results in a ranked list of recommended tags". Hotho et al. [9] presented an algorithm called FolkRank which uses the structure of folksonomies for searching and ranking. These rankings can also be used to recommend tags, resources and users or to build communities of interest from the folksonomy. In [10] Jaschke et al. extended FolkRank to design a graph-based tag recommendation algorithm on top of it and compared it to collaborative filtering based on users, where they achieved better recall and precision values. Another interesting contribution to tag recommender systems was made by Lipczak and Milios [18] who introduced a novel scalable and adoptable system, which can recommend tags based on the resource's title and content and the user's profile and which allows to learn new tags efficiently. Rendle et al. introduced a factorization model PITF (Pairwise Interaction Tensor Factorization) with linear runtime for both learning and recommending tags [21]. Similarly to the work of Lipczak and Milios [18] they addressed the problem of the cubic runtime of Tensor Factorization approaches which have been shown to outperform for instance other tag recommender algorithms such as FolkRank, collaborative filtering, etc. One of the first extensive studies of the tag prediction problem from the rule-mining perspective was performed by Heyman et al. [8], who achieved high-precision results in a number of experiments using tags from the social tagging system Delicious. In a follow-up, Krestel et al. [13]. tested the use of recommending tags in LDA and showed that it delivered significantly better results than the association rules. In [12], they enhanced the performance of LDA by combining it with simple language models based on the most frequent tags of the users and the resources in the bookmarks.

Although these methods of finding algorithms to accurately predict historical user-interaction data are rather efficient, they often lack theoretical background in the cognitive processes that lead to the data that is being predicted. By applying formal models of human semantic memory, the new recommender presented in the next section complements the above-mentioned recommender systems and integrates current cognitive science results into the recommender systems for social tagging.

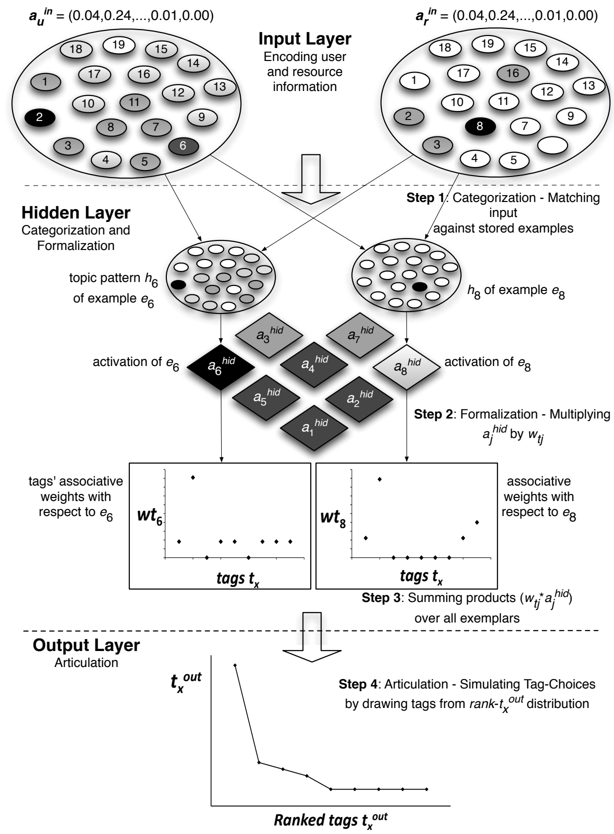

Figure 1: Basic architecture of *3Layers* (Note that only two of the six exemplars at the hidden layer are illustrated completely).

3. APPROACH

Our theoretical focus is on formal memory models explaining word (re-)productions and hence, psycholinguistic processes that we deem to be in play during tag assignments and tag imitations. A number of prominent memory models assume word productions to proceed in different steps on distinct levels of memory. For instance, the Fuzzy Trace Theory (FTT, [2]) postulates an activation of a gist-trace in response to a stimulus, e.g. a Web-resource or a set of associated tags, which contains semantic aspects (concepts, relations, patterns) of the stimulus. The gist-trace in turn reconstructs several, semantically related word forms verbalizing the activated gist. By means of a Markov model derived from FTT, [22] showed that a substantial amount of tag productions can indeed be predicted by a two-step memory retrieval involving both gist-based and verbal processes.

Similarly to FTT, the psycholinguistic theory of Levelt et al. [16] distinguishes between three processes during the production of words: 1) *Categorization* (resulting in a message or gist to be articulated), 2) *Formalization* (accessing the mental lexicon to activate word forms corresponding to the categorization) and 3) *Articulation* (selecting and producing appropriate word forms). The recommender presented here is called *3Layers* and is in line with this proposed translation of latent structures into words. We assume a set of tagged resources, which are at the same time assigned to a category. These categories (hereinafter called "semantic features") are either given a-priori (e.g. because a page is categorized to a wikipedia category) or are derived as LDA topics [5] from the tag assignments. The recommender starts with categorizing a user-specific resource by encoding and processing semantic features true for the user and/or resource, then formalizes the categorization by identifying tag distributions associated with the resource's semantic features and finally, articulates tags by sampling the most appropriate tags from the identified tag distributions.

3Layers is based on ALCOVE [14, 15], a formal model of human category learning. The basic architecture is a feed-forward connectionist network consisting of three layers of nodes realizing a top-down pattern completion process by means of straightforward equations. In response to semantic information on the input layer (two patterns of LDA-topics or external categories, one characterizing a user and one a specific resource), the hidden layer categorizes and formalizes the resource by calculating the input's similarity to already stored exemplars that are unique topic (or category) patterns and associated tag distributions. Finally, the output layer articulates the preceding categorization and formalization processes by sampling tags from the tag distributions of the identified, similar exemplars.

On the input layer, there are two input vectors representing semantic features that are true for the user u, a_u^{in}, and the resource r, a_r^{in}. Within each vector, each of the N nodes represents a single semantic feature f_i (in our case a topic identified by LDA or a category). Its activation (denoted a_i^{in}) indicates the extent to which that feature applies to the user, a_{iu}^{in}, and resource in question, a_{ir}^{in}. a_{iu}^{in} is given by

$$a_{iu}^{in} = \frac{c(f_i, u)}{\sum_{j=i}^{N} c(f_j, u)} \tag{1}$$

where $c(f_i, u)$ represents the counted frequency of the semantic feature in the user's personomy (i.e., her or his bookmark collection). Correspondingly, a_{ir}^{in} represents the association of the semantic feature to the resource and is estimated in a similar way from the counted frequency of the feature f_i in all bookmarks of the resource $r, c(f_i, r)$. The activations across the N input nodes constitute the vectors a_u^{in} = $(a_{1u}^{in}, a_{2u}^{in}, ..., a_{Nu}^{in})$ and a_r^{in} = $(a_{1r}^{in}, a_{2r}^{in}, ...$ $, a_{Nr}^{in})$. In Figure 1, the left semantic feature pattern at the input layer corresponds to the input vector a_u^{in} = $(.04, .24, ..., .01, .00)$ indicating that, for instance, the topics/categories 1 and 2 have the relative frequencies .04 and .24, respectively, across the user u's personomy.

The nodes on the hidden layer store information about exemplars e_j extracted from the training set, that is all previous tag assignments of that user. Figure 1 illustrates two such exemplars (e_6 and e_8) that are composed of unique, semantic feature patterns, h_j = $(h_{j1}, h_{j2}, ..., h_{jN})$, and associative weights w_{tj}. The latter are maintained between each of all m tags t and the unique feature pattern h_j and are illustrated in form of diagrams plotting the weights against the tags. The estimates of each h_{ji} in h_j are calculated in a similar way as the activation of each input feature (either a_{iu}^{in} or a_{ir}^{in}), and the associative weight w_{tj} encodes the relative frequency of each tag t in e_i and is estimated as

$$w_{tj}^{in} = \frac{c(t, e_j)}{\sum_{k \in e_j} c(t_k, e_j)} \tag{2}$$

where $c(t, e_j)$ is the counted frequency of tag t in exemplar e_j.

Step 1 in Figure 1 is based on a simple pattern matching process and results in probability estimates of each exemplar. To perform it, we firstly calculate the distance of a given exemplar e_j to the input vectors a_u^{in} and a_r^{in}, denoted d_j^u and d_j^r, respectively, by applying the cosine similarity measure and subtracting the result from 1, i.e.,

$$d_j^u = 1 - \frac{\langle a_u^{in}, h_j \rangle}{\|a_u^{in}\| \|h_j\|} \tag{3}$$

Correspondingly, d_j^r is calculated by subtracting the similarity between a_r^{in} and h_j from 1. The distances are linearly combined to a single distance, which is then transformed to an activation (or similarity) estimate a_j^{hid} falling exponentially with the distance between the hidden node and the input [23], and yielding a probability estimate for e_j:

$$a_j^{hid} = \frac{exp[-(d_j^u + d_j^r)]}{\sum_k exp[-(d_k^u + d_k^r)]} \tag{4}$$

For example, the Figure 1 schematically illustrates that e_6 receives higher activation than e_8 (illustrated by the black- and grey-filled rhombic form, respectively) since e_6's topic pattern h_6 is more similar to both input vectors a_u^{in} and a_r^{in} than e_8's topic pattern h_8.

We then form response strengths for each of the tags, t_x^{out}. In step 2 (see Figure 1), each hidden node's activation a_j^{hid} is multiplied by the corresponding tags' associative weights, i.e., $a_j^{hid} \cdot w_{tj}$, and in step 3, these products are summed over all hidden nodes, given by

$$t_x^{out} = \sum_j wt_j \cdot a_j^{hid} \tag{5}$$

where each t_x^{out} is a realization of a discrete random variable X since $\sum_x^m Pr(X = t_x^{out}) = 1$.

In a last step 4, we make use of this probability distribution to simulate the user's tag assignments by drawing y random numbers and mapping them into events, i.e. t_x^{out}. Finally, the observed count of tag t_x in the simulation, $c(t_x)$, determines its ranking for being recommended. If the parameter l specifies the number of tags to be selected, the subset of tags to be recommended $RecTags$ is given by

$$RecTags := \{t_x | rank[c(t_x)] \leq l\} \tag{6}$$

4. EXPERIMENTAL SETUP

In order to evaluate our approach, we compared it with a popular tag recommendation approach based on Latent Dirichlet Allocation [12, 13].

Latent Dirichlet Allocation (LDA) is a probability model that helps to find latent topics for documents where each topic is described by words in these documents [13]. This can be formalized as follows:

$$P(t_i|d) = \sum_{j=1}^{Z} P(t_i|z_i = j) P(z_i = j|d) \tag{7}$$

Here $P(t_i|d)$ is the probability of the ith word for a document d and $P(t_i|z_i = j)$ is the probability of t_i within the topic z_i. $P(z_i = j|d)$ is the probability of using a word from topic z_i in the document. In LDA the number of latent topics Z has to be chosen in advance, which defines the level of specialization of the topics.

When using LDA for tag recommendation, documents are resources which are described by tags. This means that each resource, or more specified each bookmark of a resource, can also be represented with the top tags of topics identified by LDA.

We implemented the LDA tag recommendation algorithm with Gibbs sampling using the Java framework LingPipe[1]. Therefore we

[1] http://alias-i.com/lingpipe/

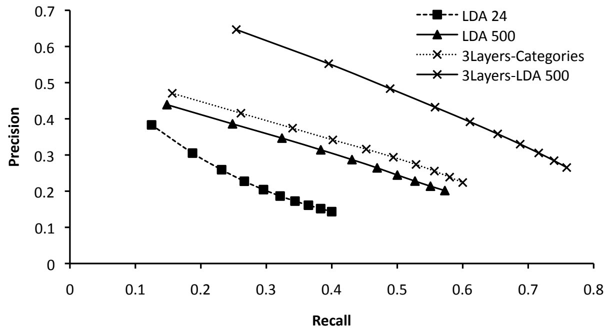

Figure 2: Recall/precision plots for LDA with 24 topics, LDA with 500 topics, *3Layers* with Wikipedia categories and 3 Layers with LDA tags on 1 - 10 recommended tags.

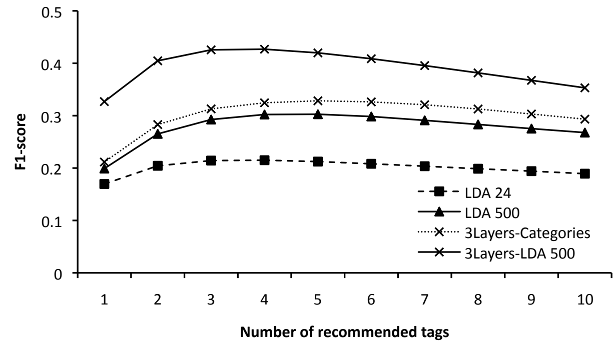

Figure 3: F1-score values for LDA with 24 topics, LDA with 500 topics, *3Layers* with Wikipedia categories and *3Layers* with LDA tags on 1 - 10 recommended tags.

calculated the probability of a tag t $P(t|r,u)$ based on a given user u $P(t|u)$ and based on a given resource r $P(t|r)$ and combined these two values based on the smoothing technique described by Krestel and Frankhauser [12]. This ensures that the two probabilities are weighted according to their importance and that no tag gets a probability value of 0.

4.1 Dataset

For our experimentation we used a large-scale social tagging dataset crawled from Delicious[2] and provided by Arkaitz et al. [26]. It was crawled between 2003 and March 2011 and contains nearly 340 million bookmarks, 119 million unique resources, 15 million unique tags and 2 million unique users. To obtain a dataset where all resources are categorized and freely available, we parsed out all bookmarks of Wikipedia[3] articles, which resulted in 1.7 million bookmarks, 386 thousand unique resources, 361 thousand tags, 304 thousand unique users and 4.9 million tag assignments. This focus on the Wikipedia domain gives us not only the possibility to test our approach with external knowledge such as category information, but also increases the reproducibility of our experiments. In order to get a dense fraction of the dataset we used a p-core pruning technique as proposed by Batagelj and Zaversnik [1]. The final dataset we used for our experiments was a p-core pruned dataset at level 14 and contained 49,691 bookmarks, 2,003 unique resources, 1685 unique tags, 1,968 unique users and 194,584 tag assignments.

In order to extend the resources in our dataset with semantic features that can be used as external knowledge for the input layer of our approach, we fetched the category information of the Wikipedia articles from that time latest Wikipedia dump[4]. Since the articles categories are very specific, we only focused on the 24 Wikipedia top-level categories for each article obtained from the Wikipedia category-taxonomy that we created according to [19].

4.2 Evaluation Method and Metrics

To evaluate the performance of our tag recommender approach we used a 80/20 split to randomly generate 20 different training and test sets. The bookmarks in a training set were used as the input for the algorithms to predict the tags of the bookmarks in the corresponding test set [13].

As evaluation metrics we used different well-established metrics for tag recommendations in order to obtain the performance of our approach compared to LDA [10, 17]. All these metrics are reported for different numbers of recommended tags (1 - 10) and as an average over our 20 training and test sets.

Recall is calculated as the number of correctly recommended tags divided by the number of relevant tags, where t_u denotes the list of recommended tags and T_u the list of relevant tags of a bookmark of user u. This is averaged on all known bookmarks U.

$$Recall = \frac{1}{|U|} \sum_{u \in U} \frac{|t_u \cap T_u|}{|T_u|} \tag{8}$$

Precision is calculated as the number of correctly recommended tags divided by the number of recommended tags.

$$Precision = \frac{1}{|U|} \sum_{u \in U} \frac{|t_u \cap T_u|}{|t_u|} \tag{9}$$

F1-score combines precision and recall into one score [17].

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{10}$$

Mean reciprocal rank (MRR) is the sum of the reciprocal ranks of all relevant tags in the list of the recommended tags. This means that a higher MRR is achieved if the relevant tags occur at the beginning of the recommended tag list [20].

$$MRR = \frac{1}{|U|} \sum_{u=1}^{|U|} (\sum_{t \in T_u} \frac{1}{rank(t)}) \tag{11}$$

Mean average precision (MAP) is an extension of the precision metric that also looks on the ranking of the recommended tags. It is described in the formula below where B_n is 1 if the recommended tag at position n is relevant [20].

$$MAP = \frac{1}{|U|} \sum_{u=1}^{|U|} (\frac{1}{|T_u|} \sum_{n=1}^{|T_u|} B_n \times Precision@n) \tag{12}$$

5. RESULTS

In this section we present the results of our approach compared to LDA based on the previously mentioned evaluation metrics and the Wikipedia dataset.

As reported in Section 4, the number of latent topics for LDA has to be set in advance. When generating recommendations for 24 (corresponding to the number of top-level categories in Wikipedia), 100, 250, 500, 750 and 1000 topics based on 10 recommended tags,

[2]https://delicious.com/
[3]http://en.wikipedia.org/
[4]http://dumps.wikimedia.org/enwiki/20121101/

Algorithm	MRR±STD	MAP±STD
LDA 24	.662±.014	.240±.005
LDA 500	.862±.015	.345±.007
3Layers-Categories	.940±.007	.391±.003
3Layers-LDA 500	**1.200±.005**	**.549±.002**

Table 1: MRR and MAP values with standard deviations for LDA with 24 topics, LDA with 500 topics, *3Layers* with Wikipedia categories and *3Layers* with LDA tags on 10 recommended tags.

Metric	Source	SS	DF	MS	F	p-value
MRR	between groups	2.971	3	.990	7,610	<.001
	within groups	.009	76	.0001		
	TOTAL	2.982	79			
MAP	between groups	.993	3	.331	12,514	<.001
	within groups	.002	76	2.645E-5		
	TOTAL	.995	79			

Table 2: Summary of one-way ANOVA for MRR and MAP on 10 recommended tags with *Algorithm* as the between-subjects factor.

we found that 500 topics produced the best results (MRR = .862 and MAP = .345). We therefore configured our *3Layers* approach with two different data sources for its input layer, (i) Wikipedia's 24 top-level categories as described in Section 4.1 and (ii) tags based on LDA with 24 (corresponding to the 24 Wikipedia categories) and 500 topics. For the second configuration we used the top 10 tags identified by LDA for each bookmark in the training set.

Figure 2 shows the recall/precision plot for LDA with 24 topics, LDA with 500 topics, *3Layers* with Wikipedia categories and *3Layers* with LDA tags calculated for 500 topics on 1 - 10 recommended tags. LDA with 24 topics is used here as a simple baseline based on the number of top-level categories in Wikipedia. Furthermore, Figure 3 also shows the F1-score values for these algorithms on 1 - 10 recommended tags. It can bee seen that both *3Layers* approaches outperform LDA on all values where the maximum values are reached for recall@10 = .758, precision@1 = .646 and F1-score@4 = .426 for *3Layers* with LDA tags identified for 500 topics.

The MRR and MAP values with standard deviations are shown in Table 1 for all the algorithms on 10 recommended tags. Also on these metrics the two *3Layers* approaches outperforms LDA on all values. The maximum values are reached by *3Layers* with LDA tags based on 500 topics for MRR = 1.200 and MAP = .549 (visualized in bold). These estimates clearly imply that independent of the measure the probability estimates vary with the conditions, i.e. the tag recommenders, in a constant ordering.

To check for statistical significance we performed two one-way ANOVAs on MRR and MAP for 10 recommended tags with *Algorithm* as a between-subjects factor. The statistical prerequisites of normal distribution and equal variances were met. The results of both ANOVAs are shown in Table 2 and are well in line with the descriptive pattern of Table 1. In particular, the overall difference between the four recommenders proved highly significant and yielded the large effect sizes of η^2_{MRR} = .997 and η^2_{MAP} = .998. Additionally, pairwise comparisons conducted by means of the Tukey's HSD test corresponded to the ordering described above. First, the difference between the two best performing recommenders, i.e. *3Layers*-LDA 500 and *3Layers*-Categories (MRR: q = 55.38, p < .001; MAP: q = 64.56, p < .001), second, the difference between *3Layers*-Categories and LDA 500 (MRR: q = 21.73, p < .001; MAP: q = 28.05, p < .001) and third, the difference between LDA 500 and LDA 24 (MRR: q = 71.88, p < .001; MAP: q = 97.29, p < .001) all proved large and highly significant.

6. DISCUSSION AND CONCLUSION

In this paper we have presented and evaluated *3Layers*, a model of human categorization implemented in form of a tag recommender. The model takes into account semantic information about a user-specific bookmark, which is either a set of available Wikipedia categories or a set of topics derived by LDA. The semantic information is further processed in a connectionist network of three layers that mimics the user's categorization and formalization of the bookmark to predict the user's tag assignments. We think this has introduced some new perspectives into recommender systems research for social tagging environments.

Our experiments show that the *3Layers*-model holds potential of realizing a strongly performing recommender system. In particular, *3Layers*-LDA that utilizes LDA-topics as input significantly outperforms the LDA-recommender introduced by [13]. The same applies to *3Layers*-Categories, which makes use of Wikipedia categories and therefore, operates independently of the LDA-approach.

Of course, several limitations of these results need to be addressed. As we have only tested the performance in one data set, generalizability to other cases needs to be demonstrated. Also without a doubt, there is nowadays a much larger set of recommender algorithms available than we could take into account in our study.

We take the results as a promising outcome. First of all, the processing of semantic categories (either explicitly given, or latent) can alleviate the cold start problem that other approaches are suffering from (such as Collaborative Filtering or those based on popularity, for instance). Reliance on these categories should also improve the robustness as the algorithm does not only depend on word-level imitation but takes into account shared semantic interpretations (e.g. [4]).

Additionally, our approach significantly enhances the LDA-recommender [13, 12] by further operating on the identified latent topic patterns. We attribute the latter result to the calculation steps of formalization where prior tag distributions are weighted according to the preceding categorization steps. The result is a distribution at the output layer exhibiting fewer ties and allowing for a more accurate selection of relevant tags. Therefore, our approach provides an appropriate theoretical framework and an effective recommender that integrates top-down and bottom-up generated data.

Our approach therefore should transfer well to other related Web interaction paradigms where both top down classification systems and bottom-up categorization co-exist. For example, Web curation is a recent trend in which Web users can create collections of resources and share these collections with others. These usually employ mechanisms of social bookmarking and tagging, but also employ classification systems to which collections are assigned.

With Web interaction paradigms changing quickly, a purely data-driven strategy has its limitations, as the data sets produced within them may differ considerably. It is then more difficult to understand, why certain approaches perform very well in certain datasets, but not very well in others. The reason is that datasets are products of very complex processes [6] and they depend on a number of factors that the models would need to take into account. While the datasets will look different, many of the fundamental processes that underlie the interaction in these new environments (such as human categorization or language production) will be very similar. Hence, the danger of a predominantly data-driven research strategy is that with every new paradigm, we have to start from zero as the earlier algorithms are not directly transferable. With the current work, we have demonstrated how a connection between a data-driven and theory-driven approach can be realized when the algorithms implement well-founded theories of cognitive science.

7. FUTURE WORK

In future work we will address the previously mentioned issues by testing the recommender mechanism in other tagging datasets as well as with other Web interaction paradigms, such as Web curation. Additionally, we will compare *3Layers*' performance to other well-established approaches, such as FolkRank [9] or Collaborative Filtering [25]. A distinctive benefit of our theory-driven approach in designing tag recommendation mechanisms is that it opens up fruitful directions for future research. For instance, we hypothesize that our approach relates to the distinction of categorizers and describers that was introduced in [11] to explain different tagging motivations. We suspect that *3Layers* will especially work well for the categorizers who draw on a more refined system on personal categories when assigning tags.

Acknowledgments: The authors would like to thank Andreas Hotho for his constructive comments on this work. This work is partially funded by the Know-Center and the EU FP7 under the Learning Layers project (http://www.learning-layers.eu).

8. REFERENCES

[1] V. Batagelj and M. Zaveršnik. Generalized cores. *arXiv preprint cs/0202039*, 2002.

[2] C. Brainerd and V. Reyna. Recollective and nonrecollective recall. *Journal of memory and language*, 63(3):425–445, 2010.

[3] U. Cress, C. Held, and J. Kimmerle. The collective knowledge of social tags: Direct and indirect influences on navigation, learning, and information processing. *Comput. Educ.*, 60(1):59–73, Jan. 2013.

[4] W.-T. Fu, T. Kannampallil, R. Kang, and J. He. Semantic imitation in social tagging. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 17(3):12, 2010.

[5] T. L. Griffiths, M. Steyvers, J. B. Tenenbaum, et al. Topics in semantic representation. *Psychological review*, 114(2):211, 2007.

[6] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 211–220, New York, NY, USA, 2007. ACM.

[7] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, Jan. 2004.

[8] P. Heymann, D. Ramage, and H. Garcia-Molina. Social tag prediction. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 531–538. ACM, 2008.

[9] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. In *The semantic web: research and applications*, pages 411–426. Springer, 2006.

[10] R. Jäschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag recommendations in folksonomies. In *Knowledge Discovery in Databases: PKDD 2007*, pages 506–514. Springer, 2007.

[11] C. Körner, R. Kern, H.-P. Grahsl, and M. Strohmaier. Of categorizers and describers: an evaluation of quantitative measures for tagging motivation. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, HT '10, pages 157–166, New York, NY, USA, 2010. ACM.

[12] R. Krestel and P. Fankhauser. Language models and topic models for personalizing tag recommendation. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 82–89. IEEE, 2010.

[13] R. Krestel, P. Fankhauser, and W. Nejdl. Latent dirichlet allocation for tag recommendation. In *Proceedings of the third ACM conference on Recommender systems*, pages 61–68. ACM, 2009.

[14] J. K. Kruschke et al. Alcove: An exemplar-based connectionist model of category learning. *Psychological review*, 99(1):22–44, 1992.

[15] M. D. Lee and D. J. Navarro. Extending the alcove model of category learning to featural stimulus domains. *Psychonomic Bulletin & Review*, 9(1):43–58, 2002.

[16] W. J. Levelt. A theory of lexical access in speech production. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*, pages 3–3. Association for Computational Linguistics, 1996.

[17] M. Lipczak. *Hybrid Tag Recommendation in Collaborative Tagging Systems*. PhD thesis, Dalhousie University, 2012.

[18] M. Lipczak and E. Milios. Learning in efficient tag recommendation. In *Proceedings of the fourth ACM conference on Recommender systems*, RecSys '10, pages 167–174, New York, NY, USA, 2010. ACM.

[19] D. Milne and I. H. Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM, 2008.

[20] M. Rawashdeh, H.-N. Kim, J. M. Alja'am, and A. El Saddik. Folksonomy link prediction based on a tripartite graph for tag recommendation. *Journal of Intelligent Information Systems*, pages 1–19, 2012.

[21] S. Rendle and L. Schmidt-Thieme. Pairwise interaction tensor factorization for personalized tag recommendation. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 81–90, New York, NY, USA, 2010. ACM.

[22] P. Seitlinger and T. Ley. Implicit imitation in social tagging: familiarity and semantic reconstruction. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, pages 1631–1640. ACM, 2012.

[23] R. N. Shepard et al. Toward a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323, 1987.

[24] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 327–336, New York, NY, USA, 2008. ACM.

[25] Z. Xu, Y. Fu, J. Mao, and D. Su. Towards the semantic web: Collaborative tag suggestions. In *Proceedings of Collaborative Web Tagging Workshop at 15th International World Wide Web Conference*, 2006.

[26] A. Zubiaga, V. Fresno, R. Martínez, and A. P. García-Plaza. Harnessing folksonomies to produce a social classification of resources. *IEEE Transactions on Knowledge and Data Engineering*, 99(PrePrints), 2012.