Evaluating Tag Recommender Algorithms in Real-World Folksonomies: A Comparative Study

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

Elisabeth Lex
Knowledge Technologies Institute
Graz University of Technology
Graz, Austria
elisabeth.lex@tugraz.at

ABSTRACT

To date, the evaluation of tag recommender algorithms has mostly been conducted in limited ways, including p-core pruned datasets, a small set of compared algorithms and solely based on recommender accuracy. In this study, we use an open-source evaluation framework to compare a rich set of state-of-the-art algorithms in six unfiltered, open datasets via various metrics, measuring not only accuracy but also the diversity, novelty and computational costs of the approaches. We therefore provide a transparent and reproducible tag recommender evaluation in real-world folksonomies. Our results suggest that the efficacy of an algorithm highly depends on the given needs and thus, they should be of interest to both researchers and developers in the field of tag-based recommender systems.

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*

Keywords

tag recommender; recommender evaluation; social tagging systems; accuracy; diversity; novelty; computational costs

1. INTRODUCTION

Since social tagging has become an essential Web 2.0 tool for collaborative content annotation, research on tag recommenders has significantly expanded over the past years. Tag recommender algorithms process folksonomy data in order to assist people in finding descriptive tags for their bookmarked resources. Although a number of tag recommender evaluation studies have been performed (e.g., [6, 5, 13, 11]), most of them have only involved a limited view of the tag recommender evaluation process with respect to the algorithms, datasets and evaluation metrics included. Furthermore, most of these evaluations were conducted only on p-core pruned datasets which does not reflect a real-world folksonomy setting as shown by Doerfel et al. [2]. With that regard, this study aims to provide a transparent and reproducible evaluation of various tag recommender algorithms in real-world folksonomies. Our contributions are as follows:

- We compare the performance not only of classic tag recommender algorithms, such as Collaborative Filtering, FolkRank and Pairwise Interaction Tensor Factorization, but also of novel time-based and cognitive-inspired approaches.
- We conduct our evaluation using unfiltered dataset samples (i.e., no p-cores) gathered from six folksonomies (Flickr, CiteULike, BibSonomy, Delicious, LastFM and Delicious) to demonstrate the performance of the algorithms in real-world settings.
- We investigate the performance of the algorithms via a wide range of evaluation metrics measuring not only the accuracy, ranking, diversity and novelty of the recommended tags but also the computational costs (runtime and memory) of the approaches.
- We calculate all of our results using the open-source tag recommender evaluation framework TagRec [8], which contains implementations of the tag recommender algorithms, evaluation metrics and the protocol used in this study.

In summary, this study may help researchers in the area of tag-based recommender systems and especially tag recommenders to obtain an overview of the performance of state-of-the-art approaches as well as developers of live recommender systems to achieve an understanding of which algorithm could best fit their needs. To the best of our knowledge, this is the first tag recommender study of this kind, which provides a transparent overview of such a wide range of algorithms, datasets and metrics in real-world folksonomy settings.

2. METHODOLOGY

This section describes the methodology used in our study, including descriptions of the algorithms, datasets, metrics and the evaluation protocol. All the evaluations have been conducted via the open-source Java framework TagRec[1] (2015-01-21) [8], except for the results for the Pairwise Interaction Tensor Factorization (PITF) algorithm that were calculated

RecSys'15, September 16–20, 2015, Vienna, Austria.
© 2015 ACM. ISBN 978-1-4503-3692-5/15/09 ...$15.00.
DOI: http://dx.doi.org/10.1145/2792838.2799664.

[1] https://github.com/learning-layers/TagRec/

using the C++ code provided by the University of Konstanz[2] (2011-07-14). Unless stated otherwise in the text, we applied the default parameter settings for the algorithms specified in the frameworks.

2.1 Algorithms

In this study, we used a wide range of folksonomy-based tag recommender algorithms. We focused on folksonomy-based rather than content-based approaches for two reasons: first, freely available social tagging datasets typically do not contain content data about the resources (e.g., title or description) and second, Rendle et al. [13] proved that personalized folksonomy-based approaches outperform the theoretically best unpersonalized method (to which content-based algorithms typically belong). The algorithms for our comparative study were chosen based on their novelty, popularity and effect in the field. Since previous tag recommender studies mostly employed classic approaches, we attempted to expand the coverage by also including novel time-based and cognitive-inspired algorithms.

The simplest approaches that we utilized are the frequency-based *MostPopular*$_r$ *(MP*$_r$*)* and *MostPopular*$_{u,r}$ *(MP*$_{u,r}$*)* algorithms [5] and *Collaborative Filtering (CF)* [12] with a neighborhood size of 20. As for algorithms that apply latent factor models, we chose two types of algorithms: *Latent Dirichlet Allocation (LDA)* [10] with 1000 latent topics and *Pairwise Interaction Tensor Factorization (PITF)* [14] with 256 dimensions of factorization. Another well-known tag recommender approach we chose for this study was *FolkRank (FR)* [5]. With regard to time-dependent tag recommenders, we included four algorithms: *Temporal Tag Usage Patterns (GIRPTM)* [15] that works in a more data-driven way and three that are inspired by models of cognitive science. The first one of this kind, BLL$_{AC}$ [7], is the only algorithm in our study that works solely on the individual level and thus, can only recommend tags used by the target user in the past. Furthermore, we also included BLL$_{AC}$+MP$_r$, which extends BLL$_{AC}$ by also recommending tags that were assigned by other users to the target resource [7]. The third cognitive-inspired algorithm (and last one in this study) is the 3LT+MP$_r$ approach [9].

2.2 Datasets

In this section, we describe the datasets used in our study. We chose a set of six freely available folksonomy datasets: Flickr[3] (2010-01-07), CiteULike[4] (2015-02-03), BibSonomy[5] (2015-01-01), Delicious[3] (2010-01-07), LastFM[6] (2011-05-12) and MovieLens[7] (2009-01-05). These datasets differ in terms of their domain type (i.e., images, URLs, citations, music and movies), size and narrowness degree. For the purposes of this study, we defined the degree of narrowness as the average number of posts assigned to a resource which also correlates with the commonly known definition of narrow and broad folksonomies [4].

Dataset	$\|U\|$	$\|R\|$	$\|T\|$	$\|P\|$	$\|P\|/\|R\|$
Flickr	9,590	856,755	125,119	856,755	1.000
CiteULike	18,474	811,175	273,883	900,794	1.110
BibSonomy	10,179	683,478	201,254	772,108	1.129
Delicious	15,980	963,741	184,012	1,447,267	1.501
LastFM	1,892	12,522	9,748	71,062	5.674
MovieLens	4,009	7,601	15,238	55,484	7.299

Table 2: Summary of the real-world folksonomy datasets used in this study where $|U|$ is the number of users, $|R|$ is the number of resources, $|T|$ is the number of tags, $|P|$ is the number of posts and $|P|/|R|$ accounts for the degree of narrowness.

As outlined in Section 1 above, it was crucial for us to benchmark the algorithms in the unfiltered datasets without p-core pruning to avoid a biased evaluation and to simulate a real-world folksonomy setting (see also [2]). This is especially important for the development of live recommender services. The narrowness degrees of the datasets used (see the last column of Table 2) justifies this approach since the average number of posts assigned to a resource is lower than two in four of the six datasets (Flickr, CiteULike, BibSonomy and Delicious). This means that even a small p-core of two would delete a lot of posts and so, substantially distort the natural structures of these datasets. Hence, the only filtering techniques we applied to our datasets were decapitalizing the tags and excluding all automatically generated tags (e.g., "no-tag" or "bibtex-import"). In order to be able to process the data of Flickr, CiteULike and Delicious, we had to use samples of the whole datasets. Thus, we randomly chose 3% of the complete user profiles (i.e., all posts of a user) in Flickr and Delicious and 15% in CiteULike (see [3]) to maintain the original characteristics of the data. The final properties of our datasets after these steps are summarized in Table 2.

2.3 Metrics

We used a wide set of evaluation metrics known from field of recommender systems to assess the performance of the algorithms. Specifically, we measured the accuracy, ranking, diversity and novelty of the recommended tags, as well as the computational costs in terms of runtime and memory consumption of the algorithms.

Accuracy. In terms of metrics that measure recommender accuracy [11], we report *F1-Score (F1@5)*, which was the main performance metric in the PKDD Discovery Challenge 2009[8], and the ranking-dependent metrics *Mean Reciprocal Rank (MRR@10)*, *Mean Average Precision (MAP@10)* and *Normalized Discounted Cumulative Gain (nDCG@10)*.

Diversity. We measure tag recommender diversity by means of the *Average IntraList Distance (AILD@10)* metric as defined in [1]. In this metric, the dissimilarity of two tags is given by the relative difference between the sets of resources to which the tags were applied. This means that a set of tags is diverse if the tags were used for different sets of resources.

Novelty. The novelty of the recommended tag list is calculated using the *Average Inverse Popularity (AIP@10)* metric. Similarly to [1], we define a recommended tag as novel if it was not previously used to annotate the target resource. Thus, the lower the popularity of a tag for a resource is, the higher its novelty.

[2] http://www.informatik.uni-konstanz.de/rendle/software/tag-recommender/

[3] https://www.uni-koblenz.de/FB4/Institutes/IFI/AGStaab/Research/DataSets/PINTSExperimentsDataSets

[4] http://www.citeulike.org/faq/data.adp

[5] http://www.kde.cs.uni-kassel.de/bibsonomy/dumps

[6] http://files.grouplens.org/datasets/hetrec2011/hetrec2011-lastfm-2k.zip

[7] http://files.grouplens.org/datasets/movielens/ml-10m.zip

[8] http://www.kde.cs.unikassel.de/ws/dc09/

Dataset	Metric	MP_r	$MP_{u,r}$	CF	LDA	PITF	FR	GIRPTM	BLL_{ac}	$BLL_{ac}+MP_r$	$3LT+MP_r$
Flickr	F1@5	-	.371	.453	.178	.350	.365	.455	.470	.470	**.482**
	MRR@10	-	.392	.474	.184	.366	.387	.488	.512	.512	**.525**
	MAP@10	-	.509	.631	.216	.469	.501	.647	.680	.680	**.698**
	nDCG@10	-	.569	.666	.280	.535	.561	.686	.711	.711	**.727**
	AILD@10	-	.789	.975	**.980**	**.980**	**.980**	.789	.789	.789	.670
	AIP@10	-	-	-	-	-	-	-	-	-	-
	Runtime [s]	-	**1**	4,342	1,227	228,868	18,090	2	5	5	10,594
	Memory [MB]	-	4,672	8,488	9,652	**2,502**	9,190	4,974	6,053	6,053	6,942
CiteULike	F1@5	.042	.249	.231	.089	.178	.250	.262	.259	.273	**.277**
	MRR@10	.043	.277	.263	.086	.207	.276	.303	.312	.319	**.321**
	MAP@10	.054	.329	.311	.094	.233	.327	.359	.367	.380	**.383**
	nDCG@10	.063	.392	.359	.138	.294	.392	.420	.422	.438	**.440**
	AILD@10	.152	.916	.961	**.991**	**.991**	**.991**	.916	.902	.916	.893
	AIP@10	.142	.952	.960	.983	**.991**	.958	.953	.985	.951	.953
	Runtime [s]	**1**	2	6,315	10,673	343,181	27,305	3	1,290	1,424	10,796
	Memory [MB]	5,725	5,913	9,301	11,943	**3,030**	9,347	6,631	8,177	8,789	9,474
BibSonomy	F1@5	.068	.281	.260	.145	.215	.279	.291	.279	.298	**.307**
	MRR@10	.054	.268	.248	.143	.218	.269	.282	.278	.289	**.298**
	MAP@10	.073	.337	.310	.162	.257	.337	.356	.346	.365	**.378**
	nDCG@10	.091	.407	.369	.219	.327	.408	.425	.409	.434	**.445**
	AILD@10	.199	.916	.941	.990	**.991**	**.991**	.916	.901	.916	.889
	AIP@10	.182	.939	.954	.966	.973	.944	.940	**.976**	.937	.941
	Runtime [s]	**1**	2	2,797	9,847	219,573	12,549	2	502	601	9,316
	Memory [MB]	4,811	4,972	9,405	14,012	**2,432**	9,494	5,567	8,078	8,307	9,137
Delicious	F1@5	.135	.238	.243	.182	.199	.196	.261	.243	.283	**.284**
	MRR@10	.117	.232	.241	.171	.193	.184	.258	.261	.290	**.291**
	MAP@10	.153	.279	.296	.204	.229	.226	.314	.312	**.358**	.357
	nDCG@10	.187	.358	.356	.271	.302	.292	.393	.374	**.431**	.430
	AILD@10	.353	.968	.972	**.999**	**.999**	**.999**	.968	.955	.968	.946
	AIP@10	.256	.882	.874	.887	.895	.877	.873	**.938**	.863	.874
	Runtime [s]	**1**	3	9,645	15,373	324,737	44,747	4	395	396	12,869
	Memory [MB]	12,198	12,596	35,090	11,894	**3,075**	7,381	13,672	16,469	17,620	59,425
LastFM	F1@5	.199	.258	.226	.258	.276	.270	.263	.251	**.283**	.279
	MRR@10	.186	.251	.208	.254	.276	.257	.255	.260	**.283**	.277
	MAP@10	.226	.301	.252	.306	.336	.313	.310	.312	**.344**	.338
	nDCG@10	.283	.386	.317	.388	.414	.399	.397	.375	**.425**	.421
	AILD@10	.722	.902	.855	.918	**.919**	**.919**	.902	.840	.902	.900
	AIP@10	.604	.730	.761	.741	.797	.728	.736	**.866**	.711	.722
	Runtime [s]	**1**	**1**	6	265	8,657	101	**1**	**1**	**1**	225
	Memory [MB]	**80**	92	214	593	87	237	155	204	301	3,332
MovieLens	F1@5	.135	.153	.124	.141	.156	.153	.159	.086	**.160**	.162
	MRR@10	.211	.260	.198	.233	.264	.243	.251	.183	**.265**	.263
	MAP@10	.223	.269	.209	.242	.275	.253	.262	.188	**.276**	.274
	nDCG@10	.271	.328	.254	.296	.324	.319	.326	.203	**.338**	.336
	AILD@10	.910	.954	.935	**.958**	.957	.957	.954	.726	.954	.954
	AIP @10	.787	.741	.861	.785	.816	.777	.751	**.976**	.756	.755
	Runtime [s]	**1**	**1**	11	206	6,091	90	**1**	**1**	**1**	120
	Memory [MB]	365	375	1,043	761	**96**	833	434	500	501	3,297

Table 1: Summary of the tag recommender results showing the accuracy, diversity, novelty, runtime and memory consumption estimates of the algorithms in the six datasets (bold numbers indicate the best results).

Computational Costs. Since recommendations should not only be accurate, diverse and novel but also be provided in (near) real-time, we determined the computational costs of the algorithms in terms of *Runtime* (in seconds) and *Memory* (in megabytes) required. Both runtime and memory are measured for the complete workflow of the algorithms (including training and testing) using an IBM System x3550 M4 Server with one Intel(R) Xeon(R) CPU E5-2640 v2 @ 2.00GHz and 256GB RAM.

2.4 Evaluation Protocol

We followed a standard evaluation procedure in tag recommender research (e.g., [5]) to split our datasets mentioned in Section 2.2 into training and test sets. To that end, for each user, the set of tags in her most recent post in time were put it into the test set and the remaining posts were then used to train the algorithms. This protocol is a promising simulation of a real-world social tagging environment since it preserves the chronological order of the data and predicts the user's future tag assignments based on the past tag assignments. To compute the metrics from Section 2.3, we compared the top-10 tags an algorithm suggested for a given user and resource pair in the test set with the set of relevant tags actually used in the corresponding post.

3. RESULTS AND DISCUSSION

Table 1 shows an overview of the exact accuracy (F1@5, MRR@10, MAP@10 and nDCG@10), diversity (AILD@10), novelty (AIP@10) and computational cost (runtime in seconds and memory consumption in megabytes) estimates of all the algorithms for the six datasets (highest values per dataset and metric are shown in bold). We merged the results across the datasets and metrics in Table 3 to make it easier to determine the usefulness of the algorithms in respect to the various user needs. The merged results indicate that the two cognitive inspired algorithms $BLL_{AC}+MP_r$ and $3LT+MP_r$ were the best (in the narrow and broad settings) with regard to recommender accuracy.

Algorithm	Accuracy		Div	Nov	Runtime	Memory
	narrow	broad				
MP$_r$	-		-	-	++	+
MP$_{u,r}$					++	+
CF			+			
LDA	-		++		-	-
PITF	-	+	++	+	-	++
FR		+	++			
GIRPTM	+	+			++	+
BLL$_{AC}$	+		-	++	+	
BLL$_{AC}$+MP$_r$	++	++			+	
3LT+MP$_r$	++	++			-	-

Table 3: Summary of the performance of the algorithms in real-world folksonomies showing tag recommender accuracy in narrow and broad settings, diversity (Div), novelty (Nov), runtime and memory consumption. "++" indicates best, "+" good, "-" poor and an empty space average performance.

The difference between the narrow and broad settings are especially of interest when comparing our results with previous studies (e.g., [6, 5, 13, 11]), in which FR and PITF typically had the best recommender accuracy in p-core pruned (i.e., very broad) folksonomies. Our results also indicate a good performance of FR and PITF in broad folksonomies but a fairly poor one in the narrow settings. The opposite is the case for the BLL$_{AC}$ approach, which strictly operates on the individual level and thus, performs well in the narrow setting but is only average in the broad setting.

Furthermore, the most diverse tag recommendations are provided via the classic approaches LDA, PITF and FR. With regard to novelty, the strictly individual BLL$_{AC}$ approach outperforms all the other algorithms (which operate also on the collective level). As for the computational costs, the best runtime results are delivered by the frequency-based methods MP$_r$, MP$_{u,r}$ and GIRPTM and the lowest memory is required by the PITF approach. One reason for the low memory consumption of PITF is surely the fact that it was developed in C++ (the other approaches were implemented in Java). Additionally, the cognitive-inspired BLL$_{AC}$+MP$_r$ approach that has the highest recommender accuracy also provides fair results in terms of the other metrics. Interestingly, this is not the case for the other cognitive-inspired algorithm 3LT+MP$_r$ which has poor runtime and memory consumption estimates since it requires a computationally expensive topic calculation step (see [9]).

4. CONCLUSION

Providing helpful tag recommendations in real-world folksonomies is not a trivial task, which greatly depends on the given user needs, as our results suggest. If recommender accuracy is mostly important, cognitive-inspired algorithms, such as BLL$_{AC}$+MP$_r$, provide the best results. If runtime is crucial, simple frequency-based methods, such as MP$_{u,r}$, should be applied. Although the classic approaches (CF, LDA, PITF and FR) known from most previous studies do not seem to be the best choice in terms of accuracy in these (mostly narrow) real-world settings, they provide the most diverse recommendations. The most novel tags, however, can be recommended via strictly individual methods, such as the cognitive-inspired BLL$_{AC}$ algorithm.

Thus, we believe that our results should be of interest to both researchers and developers in the field of tag-based recommender systems. In the future, we plan to verify these results with an online user study in a live recommender system (e.g., BibSonomy), which would also allow us to assess the real user acceptance of the recommendations.

Acknowledgments: This work is funded by the Know-Center and the EU-IP Learning Layers (Grant Agreement: 318209). The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Ministry of Transport, Innovation and Technology, the Austrian Ministry of Economics and Labor and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency (FFG).

5. REFERENCES

[1] F. Belém, E. Martins, J. Almeida, and M. Gonçalves. Exploiting novelty and diversity in tag recommendation. In *Advances in Information Retrieval*, pages 380–391. Springer, 2013.

[2] S. Doerfel and R. Jäschke. An analysis of tag-recommender evaluation procedures. In *Proc. of RecSys'13*, pages 343–346. ACM, 2013.

[3] J. Gemmell, T. Schimoler, M. Ramezani, L. Christiansen, and B. Mobasher. Improving folkrank with item-based collaborative filtering. *Recommender Systems & the Social Web*, 2009.

[4] D. Helic, C. Körner, M. Granitzer, M. Strohmaier, and C. Trattner. Navigational efficiency of broad vs. narrow folksonomies. In *Proc. of HT'12*, pages 63–72. ACM, 2012.

[5] 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.

[6] R. Jäschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag recommendations in social bookmarking systems. *Ai Communications*, 21(4):231–247, 2008.

[7] D. Kowald, S. Kopeinik, P. Seitlinger, T. Ley, D. Albert, and C. Trattner. Refining frequency-based tag reuse predictions by means of time and semantic context. In *Mining, Modeling, and Recommending'Things' in Social Media*, pages 55–74. Springer, 2015.

[8] D. Kowald, E. Lacic, and C. Trattner. Tagrec: towards a standardized tag recommender benchmarking framework. In *Proc. of HT'14*, pages 305–307. ACM, 2014.

[9] D. Kowald, P. Seitlinger, S. Kopeinik, T. Ley, and C. Trattner. Forgetting the words but remembering the meaning: Modeling forgetting in a verbal and semantic tag recommender. In *Mining, Modeling, and Recommending'Things' in Social Media*, pages 75–95. Springer, 2015.

[10] R. Krestel, P. Fankhauser, and W. Nejdl. Latent dirichlet allocation for tag recommendation. In *Proc. of RecSys'09*, pages 61–68. ACM, 2009.

[11] M. Lipczak. *Hybrid tag recommendation in collaborative tagging systems.* PhD thesis, Dalhousie University Halifax, 2012.

[12] L. B. Marinho and L. Schmidt-Thieme. Collaborative tag recommendations. In *Data Analysis, Machine Learning and Applications*, pages 533–540. Springer, 2008.

[13] S. Rendle, L. Balby Marinho, A. Nanopoulos, and L. Schmidt-Thieme. Learning optimal ranking with tensor factorization for tag recommendation. In *Proc. of KDD'09*, pages 727–736. ACM, 2009.

[14] S. Rendle and L. Schmidt-Thieme. Pairwise interaction tensor factorization for personalized tag recommendation. In *Proc. of WSDM'10*, pages 81–90. ACM, 2010.

[15] L. Zhang, J. Tang, and M. Zhang. Integrating temporal usage pattern into personalized tag prediction. In *Web Technologies and Applications.* Springer, 2012.