Collaborative Datasets Retrieval for Interlinking on Web of Data

HaiChi Liu
College of Computer, National University of Defense Technology
Changsha, Hunan 410073 P.R. China
liuhaichi@nudt.edu.cn

JinTao Tang
College of Computer, National University of Defense Technology
Changsha, Hunan 410073 P.R. China
tangjintao@nudt.edu.cn

DengPing Wei
College of Computer, National University of Defense Technology
Changsha, Hunan 410073 P.R. China
dpwei@nudt.edu.cn

PeiLei Liu
College of Computer, National University of Defense Technology
Changsha, Hunan 410073 P.R. China
plliu@nudt.edu.cn

Hong Ning
College of Computer, National University of Defense Technology
Changsha, Hunan 410073 P.R. China
hning@nudt.edu.cn

Ting Wang
College of Computer, National University of Defense Technology
Changsha, Hunan 410073 P.R. China
tingwang@nudt.edu.cn

ABSTRACT
Dataset interlinking is a great important problem in Linked Data. We consider this problem from the perspective of information retrieval in this paper, thus propose a learning to rank based framework, which combines various similarity measures to retrieve the relevant datasets for a given dataset. Specifically, inspired by the idea of collaborative filtering, an effective similarity measure called *collaborative similarity* is proposed. Experimental results show that the *collaborative similarity* measure is effective for dataset interlinking, and the learning to rank based framework can significantly increase the performance.

Categories and Subject Descriptors
I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods –*Semantic networks.* H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval –*Search process.*

General Terms
Algorithms, Measurement, Performance, Experimentation.

Keywords
Linked data, Dataset interlinking, Collaborative datasets retrieval.

1. INTRODUCTION
Data linking [1], i.e., predicting links between different Linked Data datasets is an important problem in Linked Data. However, with the rapidly growth of Linked Data datasets, how to find relevant datasets becomes the first non-trivial problem [2] for data linking.

There are few works focusing on the relevant datasets identification. For instance, Nikolov et al utilized semantic web services [2], while Lopes et al. [3] used a Bayesian classifier, and the work in literature employed the link prediction technique in Social Network Analysis (SNA) area. However, these methods are not intuitive for the most common scenario in data linking: when a new dataset is published, how to link the dataset into the Web of Data?

In this paper, we consider the problem from the perspective of Information Retrieval. That is, a given dataset is considered as the query and the datasets available on Web of Data is considered as the candidate documents. Based on this idea, we adopt and develop various similarity measures by exploiting the features of *content, links* etc. Specifically, inspired by the idea of collaborative filtering, a simple but effective measure called *collaborative similarity* is proposed and combined with other measures by a learning to rank framework. Experimental results show that the *collaborative similarity* is effective and learning to rank can further improve the performance.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2567948.2577290

2. PROPOSED APPROACH
2.1 Learning to Rank Framework
Linked data is a structured and semantic data source, which has various features that can be exploited to rank the relevance among datasets. Learning to rank is a data driven approach, which effectively incorporates a bag of features in a model for any ranking task. In this paper, we use the available datasets and their interlinking information to train a ranking model. Given a dataset as query and other datasets are considered as candidates. The candidates interlinked with the query are tagged as relevant, or tagged as irrelevant otherwise. Then a ranking model is trained by applying the algorithm of learning to rank. When a new query is given, the learned ranking model will be used to compute the relevant score for all the candidate datasets.

2.2 Ranking Functions
Since the learning to rank algorithm was employed, the key problem is turned to how to define ranking measures to compute the probability that the query dataset will be interlinked with the candidate datasets. Previous works have already proposed various ranking measures according to the *contents*, *links*, and *popularity* etc., all of which were adopted in this paper by the learning to rank algorithm. We also develop a novel ranking function, *collaborative similarity* that measures the similarity between query dataset and the datasets, which linked to candidate dataset.

Collaborative filtering recommendation is one of the most successful techniques in Recommender Systems. The motivation for collaborative filtering comes from the idea that people often get the best recommendations from someone with similar tastes. Applying this idea to our problem, if the query dataset is similar with the datasets, which were already linked to the candidate dataset, it is reasonable to infer that the query dataset also can establish links to the candidate dataset. Based on this idea, we developed a ranking function called *collaborative similarity*.

The set of datasets linked to candidate dataset c is mentioned as the in-link set s, thus the feature vector of s can be represent as $V_s=(w_1, w_2, ..., w_n)$. The adapted *tf-idf* value is used to measure the features' weight w_i:

$w_i = tf_i \cdot log(|S|/|S_i|)$, while tf_i is the frequency of feature i, $|S|$ is the number of all in-link sets, $|S_i|$ is the number of in-link sets have feature i.

The query dataset q can be presented as a 0-1 feature vector $V_q=(w_1, w_2, ... w_n)$, w_i=1 if q has feature i, otherwise w_i=0. The ranking score of the candidate dataset c for query dataset q is the dot product of the feature vectors of query dataset q and in-link set S of candidate dataset c: $score_{col}(q, c) = V_q \bullet V_s$.

3. EXPERIMENTS

3.1 Dataset

The data extracted from the Data Hub catalog [3] is used in our experiments, which contains 295 datasets and 697 links. Some datasets don't have vocabulary feature, which can be added by grouping all the namespace of their class and property URIs.

There are 122 datasets have out-linked to other datasets, which can be seen as query datasets. For each query dataset, the other 294 datasets are considered as candidate datasets, and the datasets linked to the query dataset are relevant datasets. We conducted *10-fold cross validation* to relieve the over-fitting problem. Using this strategy, all the links between datasets can be tested once. The Coordinate Ascent algorithm [4] in RankLib is adopted to learn the rank model. Standard metrics Mean Average Precision (MAP) and Precision@N (P@N) were adopted to measure the performance.

3.2 Results

The Link Predication method and Bayesian based method [3] were implemented with the best configuration to compare with our proposed approach. The effectiveness of *collaborative similarity*, and the learning to rank model combining measures of all feature types are evaluated.

Table. 1 MAP(%) Performance of Different Methods

Link Predication	Bayesian Method	Collaborative Similarity-Vocabulary	Combine all Feature Types
54.68	55.43	**59.43**	**62.38**

The *collaborative similarity* with *vocabulary* feature measure outperforms previous works. The MAP is around 59%, which shows that the idea inspired by collaborative filtering is quite promising. Using learning to rank algorithm to combine ranking functions of different feature types is very effective, which can further improve the performance. The highest MAP 62.38% is obtained by combining *collaborative similarity* with *vocabulary, class* and *property* features.

Furthermore, the most common scenario is how to recommend some datasets when a new dataset wants to link into Web of Data. The P@N value measure the precision at top n results, which are suitable to evaluate the performance in this scenario. Figure 1 shows the P@N curves of the aforementioned systems, which demonstrated that the learning to rank method is much more effective when considering the top n results. Especially, with the smaller n, the proposed method has better results, which is ideal for recommending the relevant datasets to the new dataset.

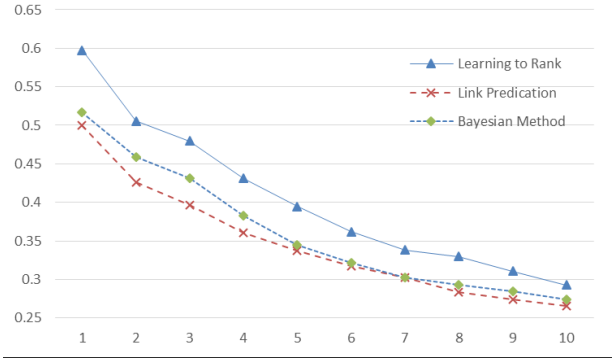

Figure 1. P@N Performance compared with previous works

4. CONCLUSION

In this paper we proposed a novel approach to identify the relevant dataset for data linking on Web of Data. The problem is considered from the perspective of Information Retrieval, thus a learning to rank algorithm is adopted to incorporate various ranking measures according to the contents, links, and popularity. Experimental results show that the *collaborative similarity* is quite effective and the learning to rank algorithm can further improve the performance.

Acknowledgments

This material is based on work supported by the National Natural Science Foundation of China (61200337, 61202118, 61472436).

5. REFERENCES

[1] Ferrara, A., Nikolov, A., & Scharffe, F. 2011. Data Linking for the Semantic Web. International Journal on Semantic Web and Information Systems (IJSWIS), 7(3), 46-76.

[2] Nikolov, Andriy and d'Aquin, Mathieu. 2011. Identifying relevant sources for data linking using a semantic web index. In: Linked Data on the Web Workshop at 20th International World Wide Web Conference (WWW 2011), India.

[3] Lopes G. R., Leme L.A.P.P, Nunes B.P., et al. Two Approaches to the Dataset Interlinking Recommendation Problem. 2014. In: 15th International Conference on Web Information System Engineering (WISE 2014).71-74.

[4] D. Metzler and W.B. Croft. Linear feature-based models for information retrieval. 2007. Information Retrieval, 10(3): 257-274.