

# Social Reference: Aggregating Online Usage of Scientific Articles in CiteULike for Clustering Academic Resources

Jiepu Jiang, Daqing He  
School of Information Sciences,  
University of Pittsburgh  
{jjj29, dah44}@pitt.edu

Chaoqun Ni  
School of Library and Information Science,  
Indiana University Bloomington  
chni@indiana.edu

## ABSTRACT

Citation-based methods have been widely studied and applied for clustering of academic resources and mapping science. Although effective, these methods suffer from citation delay. In this study, we propose to use a novel alternative source, i.e. use of literatures in social academic web. We coin the term “social reference” to refer to the reference of literatures in online social academic community environment, which is meant to be the counterpart of bibliographic reference (citation). Social reference data can be meaningful for bibliometrics studies from two aspects: first, it is timely data source and publicly accessible usage data; second, it may reflect novel perspectives of scholarly communication other than academic publishing. We experiment for journal clustering and author clustering using social reference data and compare with citation-based methods. Our experiments indicate: first, connections among literatures reflected from social reference data are comparable in clustering effectiveness to those reflected from citation; second, the sparseness of social reference data (at current stage and at least for CiteULike) makes it less effective than citation in clustering as a general, while timeliness makes it more effective than citation in clustering new resources.

## Categories and Subject Descriptors

H.2.8 [Database Applications]: Scientific Databases.

## General Terms

Algorithms, Performance, Experimentation.

## Keywords

Social reference, clustering, CiteULike, citation analysis.

## 1. INTRODUCTION

Bibliographic references provide important clues for connections among scientific literatures, which have been used for clustering of academic resources (e.g. articles, journals, and authors), and mapping disciplines. In spite of its popularity, citation analysis is argued by many researchers because of unclear and diverse cites’ motivation and citation delay. Recently, some researchers began to focus on online scholarly resources. An emerging topic is usage bibliometrics [1], which makes use of large-scale web usage data from web server logs for bibliometrics studies. Web usage data benefits from its large scale and timeliness, but is limited in its anonymous nature and limited accessibility. Another thread of works is to make use of data from online social communities [2]. Compared with current studies that focus on how scholars behave on social web [3], we focus on how academic resources are used

on social web environment.

In this study, we propose to cluster academic resources by usage data from social web. We coin the term “social reference” here to refer to the reference of literatures in social web environment, which is meant to be the counterpart of bibliographic reference. In specific, we use data from CiteULike. Similar websites include Bibsonomy, Mendeley, etc. In these websites, users can manage their collections of literatures as personal libraries. In such scenario, we can cluster academic resources by their shared users. Our assumption here for clustering is that resources used by similar users (assuming they are mostly scholars) are related. We experiment journal clustering and author clustering using data from CiteULike and compare with citation-based clustering. Next section will introduce experiment details and results.

## 2. EXPERIMENTS

### 2.1 CLUSTERING METHODS

For an academic entity (in our case, entity is journal or author, but it can also be article), we define two types of feature vectors can be created from social reference: occurrence based feature vector (OC) and co-occurrence based feature vector (COOC).

For an entity  $e$ , we define its occurrence based feature vector as  $(ef_1, ef_2, \dots, ef_n)$ , in which  $ef_i$  is the frequency of entity  $e$  used by user  $i$ . In the case of CiteULike,  $ef_i$  is frequency of articles from a journal entity (in journal clustering) or written by an author entity (in author clustering) in a CiteULike user  $i$ 's personal library. This method is similar to bibliographic coupling and direct citation. In order to normalize the feature vectors, we applied frequently used methods in bibliometrics, including binary vector (BV), TF, IDF, TFxIDF, and popular retrieval models (replacing term frequency to entity frequency), including BM25 and language modeling with dirichlet smoothing (LM-DIR).

The co-occurrence based feature vector of an entity  $e$  is defined as  $(p(e_i|e))$ , where  $p(e_i|e)$  is the probability of  $e_i$  being used by users given we know the user used  $e$ . In the case of CiteULike,  $p(e_i|e)$  is the probability of  $e_i$  in users’ libraries given we find  $e$  in a library. This method is similar to co-citation analysis. The estimation of  $p(e_i|e)$  is described in formula (1), where:  $L$  is each user’s library in CiteULike;  $p(L|e)$  is the probability that  $e$  is in  $L$  given  $e$  is seen;  $p(e_i|L, e)$  is the probability of seeing  $e_i$  in  $L$  if  $e$  is seen in  $L$ ;  $p(e_i)$  is the frequency of  $e_i$  in the whole collection. Estimation of  $p(L|e)$  and  $p(e_i|L, e)$  is described in (2), where  $ef(e, L)$  is the frequency of  $e$  in  $L$ ,  $|L|$  and  $|e|$  are the total frequency of  $L$  and  $e$ . Parameters are tuned by maximizing mean silhouette value (MSV) of clusters.

$$\hat{p}(e_i | e) = (1 - \lambda) \sum_L p(e_i | L, e) \times p(L | e) + \lambda p(e_i) \quad (1)$$

$$p(L | e) = \frac{ef(e, L)}{|e|}, \quad p(e_i | L, e) \approx p(e_i | L) = \frac{ef(e_i, L)}{|L|} \quad (2)$$

## 2.2 EXPERIMENT SETTINGS

Two citation datasets are used for experiments: WOKJ, a dataset contains articles of top journals from 40 disciplines; MSAS-CS, a dataset contains articles of top authors from 24 fields of computer science.

WOKJ dataset is created as follows: we select 20 science and 20 social science disciplines by categories in Web of Knowledge; for each discipline, top 20 journals (by JIF in 2009) are selected. The original selection includes 743 journals, but some are removed: 66 that did not consistently publish for over 10 years from 1960 to 2010 (which is for other studies); 92 that belong to multiple fields (because we use hard clustering evaluation in experiments); 108 that cannot be found in CiteULike. We use rest 477 journals and articles from 2006-2010 for our experiments.

MSAS-CS dataset is created as follows: we select top 600 authors in computer science domain from Microsoft Academic search; the authors are assigned to 24 fields by their highest ranking in each field given by Microsoft Academic Search; articles of 600 authors from 2006 to 2010, including citations and references of articles, are crawled from Microsoft Academic Search. For 57 authors, we cannot find any of their articles in CiteULike, but we still include these authors in experiments. Data are collected in January 2011.

Note that we removed the 108 journals that are not in CiteULike from WOKJ but kept the 57 authors in MSAS-CS. This is because: by removing the 108 journals, the experiment setting for WOKJ excludes the influence of data sparseness in CiteULike, and is fair for a pure evaluation on the quality of connections reflected from social reference; by keeping the 57 authors, we can evaluate the influence of data sparseness in CiteULike in a practical case.

Articles in CiteULike and the associated users' posting behaviors from 2004 to 2010 are collected as social reference data (CULSF). CULSF contains 87174 users, 3877 groups, and 1223690 articles (99% of all articles posted by users by Dec 31, 2010). Articles in WOKJ and MSAS-CS are mapped to CiteULike articles by title, first author, and publish year.

## 2.3 EVALUATION

In this section, we evaluate clustering using social reference data in WOKJ and MSAS-CS datasets and compare the effectiveness with citation-based clustering. We use the 40 disciplines in WOKJ and 24 fields in MSAS-CS as groundtruth, and evaluate clustering by normalized mutual information (NMI) and adjusted rand index (ARI). KMeans is used for clustering. To reduce influence of start points, we use the same 20 random points for tuning parameters, and the same 100 random start points for all experiments. Metrics reported are average value of results from the 100 points. The 57 authors in MSAS-CS are randomly assigned to grouped clusters.

We first experiment for citation-based methods for both datasets. Three citation-based relations are used to create feature vectors: bibliographic coupling (BC), co-citation (CO) and cross-citation (CR). For each of the vectors, normalization methods in 2.1 are experimented. Best methods by NMI are selected as citation-based methods baselines: for WOKJ dataset (journal clustering), cross-citation normalized by binary vector (BV) is selected; for MSAS-CS dataset (author clustering), co-citation and BM25 normalization is selected. The evaluation results of baselines are reported in Table 1 (\* and \*\* means significant at 0.05 and 0.01).

Then, we experiment for the two methods (occurrence-based and co-occurrence based method) for social reference in WOKJ and MSAS-CS, and compare with citation-based methods baselines.

Table 1 reports the results (top 3 methods are reported; for each method, only the best normalization is reported). For clustering of journals (WOKJ), we find the best social reference based methods (OC+BM25) are comparable to the baseline methods (CR+BV); OC+BM25 is slightly worse in NMI while slightly better in ARI, while differences in both metrics are not significant at 0.05 level. Because we exclude the influence of CiteULike data sparseness in WOKJ, results in WOKJ indicate connections among literatures reflected from social reference is comparable in quality to those reflected from citation. For clustering of authors (MSAS-CS), we find significant better performance of citation-based methods than social reference, which indicates the sparseness of data, at current stage and at least in CiteULike will influence the effectiveness of social reference clustering. However, whether the influence (10% for both metrics) is practically significant is left as a future work.

Considering social reference is timely data compared with citation, we select only articles published in 2010 for experiments. Table 2 shows the results: for WOKJ social reference based methods have lightly better results than citation based methods (not significant); for MSAS-CS, social reference based methods are significantly better than citation based methods. Compared with table 1, results in table 2 indicate: citation delay does influence the effectiveness of citation-based clustering (such influence is less significant for journal because of the large scale of journal data); social reference is a timely data source and outperforms citation in clustering new resources (which is most significant for author clustering).

**Table 1. Results for social reference-based methods.**

Dataset	Method	Norm	Evaluation Metrics	
			NMI	ARI
WOKJ	Cross-citation	BV	<b>0.645</b>	<b>0.277</b>
		Raw	0.620	0.281
	Occurrence-based	<b>BM25</b>	<b>0.624</b>	<b>0.294</b>
		LM-DIR	0.623	0.270
Journal Clustering	Co-occurrence	--	0.613	0.275
MSAS-CS	Co-citation	<b>BM25</b>	<b>0.701**</b>	<b>0.599**</b>
		TFxIDF	0.633	0.548
	Occurrence-based	BM25	0.637	0.555
		<b>LM-DIR</b>	<b>0.640</b>	<b>0.552</b>
Author Clustering	Co-occurrence	--	0.630	0.498

**Table 2. Results for clustering new resources (<=1 year).**

Dataset	Method	Norm	Evaluation Metrics	
			NMI	ARI
WOKJ	cross-citation	BV	0.609	0.246
	Occurrence-based	<b>LM-DIR</b>	<b>0.614</b>	<b>0.254</b>
MSAS-CS	cross-citation	BM25	0.509	0.207
	Occurrence-based	<b>LM-DIR</b>	<b>0.532*</b>	<b>0.264**</b>

## 3. ACKNOWLEDGMENTS

This work was supported in parts by the National Science Foundation under grant IIS-1052773.

## 4. REFERENCES

- [1] Kurtz, M. J. and Bollen, J. Usage Bibliometrics. 2010. *Annual Review of Information Science and Technology*, 44, 3-64.
- [2] Priem, J. and Hemminger, B. M. 2010. Scientometrics 2.0: Toward new metrics of scholarly impact on the social Web. *First Monday*, 15, 7.
- [3] Priem, J. and Costello, K. L. 2010. How and why scholars cite on Twitter. In *Proceedings of the American Society for Information Science and Technology*, 47, 1-4.

