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

Jiepu Jiang, Daqing He
School of Information Sciences,
University of Pittsburgh
{jij29, 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 employed for clustering academic resources and mapping science. Although effective, these methods suffer from citation delay. In this study, we extend reference and citation analysis to a broader notion from social perspective. We coin the term "social reference" to refer to the references of literatures in social academic web environment. We propose clustering methods using social reference information from CiteULike. We experiment for journal clustering and author clustering using social reference and compare with citation-based methods. Our experiments indicate: first, social reference implies connections among literatures which are as effective as citation in clustering academic resources; second, in practical settings, social reference-based clustering methods are not as effective as citation-based ones due to the sparseness of social reference data, but they can outperform in clustering new resources that have few citation.

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 academic resources and mapping science. In spite of its popularity, citation analysis is argued by many researchers for citation delay and unclear citers' motivations. 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. 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.

We coin the term "social reference" here to refer to the references of literatures in social academic web environment, which extends bibliographic reference and citation analysis to a broader notion from social perspective. As a novel source of information, social reference can be useful for bibliometrics studies: first, it has much less delay than citation; second, it is open accessible information compared with web usage logs; third, it may provide perspectives of scholarly communication other than academic publishing.

This study is an initial step towards social reference-based mining of academic resources and bibliometrics. We propose journal and author clustering methods based on social reference information. We collect social reference data from CiteULike for experiments and use citation data and classification from Web of Knowledge (in journal clustering) and Microsoft Academic Search (in author clustering) for comparison and clustering evaluation. In this study, the comparison between social reference and citation focuses on clustering effectiveness, while differences between clusters of two methods are left for future studies.

JCDL'11, June 13–17, 2011, Ottawa, Ontario, Canada.
ACM 978-1-4503-0744-4/11/06.

2. EXPERIMENTS
2.1 CLUSTERING METHODS
As we defined, "social reference" refers to references of literatures by online users in social academic web environment. As a general, for each user, we can extract a list of resources used by the user and the usage frequencies. Thus, as a general, we can define social reference data as a matrix (uf_{ij}), in which each element uf_{ij} denotes the usage frequency of resource j by user i. In the specific case of CiteULike, uf_{ij} is the frequency of resource j in user i's personal library. Then, for an academic entity (in our case journal or author) to be clustered, we can define two types of feature vectors based on social reference data matrix: occurrence based (OC) vector and co-occurrence based (COOC) feature vector. Then, entities can be clustered by similarities of either OC or COOC vectors.

For an academic entity e to be clustered, we define its OC vector as (uf_{1e}, uf_{2e}, ... , uf_{ne}), in which uf_{ie} is the usage frequency of e used by user i. This method is similar to bibliographic coupling and direct citation. In order to normalize OC vectors, we apply frequently used methods in bibliometrics (i.e. binary vector (BV), TF, IDF, TF×IDF) and some popular retrieval models (i.e. BM25, language modeling with dirichlet smoothing (LM-DIR)). To apply retrieval models, we simply consider entities as words.

The COOC 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. This method is similar to co-citation analysis. Estimation of $p(e_i|e)$ is described in formula (1), where: u is each user; $p(u|e)$ is the probability of selecting a specific user u given we know the user used e; $p(e_i|u,e)$ is the probability that user u used e_i given we know u also used e; $p(e_i)$ is the probability of e_i in the collection. Estimation of $p(u|e)$ and $p(e_i|u,e)$ is described in (2), where $uf(e, u)$ is the frequency of e used by u, $|u|$ and $|e|$ are the total frequency of u and e. Parameters are tuned to maximize MSV of clusters.

$$\hat{p}(e_i \mid e) = (1-\lambda)\sum_u p(e_i \mid u,e) \times p(u \mid e) + \lambda p(e_i) \quad (1)$$

$$p(u \mid e) = uf(e,u)/|e|, \;\; p(e_i \mid u,e) \approx p(e_i \mid u) = uf(e_i,u)/|u| \quad (2)$$

2.2 EXPERIMENT SETTINGS

We collect social reference information from CiteULike (CULSF) for experiments. CULSF dataset includes 87,174 CiteULike users' personal libraries and 1,223,690 articles from 2004 to 2010. Then, we match CULSF articles with articles in citation datasets (WOKJ and MSCS) by titles, first authors, and publishing years.

WOKJ includes articles and citations of selected journals from 40 fields. The dataset is created as follows: we select 20 science and 20 social science ISI categories from JCR 2009; for each category, top 20 journals (by JIF 2009) are selected. The original selection includes 743 journals, but some are removed: 66 journals that did not consistently publish for over 10 years from 1960 to 2010 (this process is for another study [4]); 92 that belongs to multiple fields (because we use hard clustering evaluation); 108 that cannot be found in CiteULike. The final dataset used for journal clustering includes 477 journals and articles of journals from 2006 to 2010. Journal categories from ISI are used as groundtruth for evaluation.

Another dataset, MSCS, contains articles of top authors from 24 fields of computer science, which is created as follows: we select top 600 authors in computer science listed by Microsoft Academic search; authors are assigned to 24 fields by their highest ranking in 24 fields; articles of these authors from 2006 to 2010, including citations and references of the articles, are crawled from Microsoft Academic Search. 57 authors are not found in CiteULike, but they are still included in the dataset. All data are collected in Jan 2011.

Note that we treat journals and authors that cannot be found in CiteULike differently: by removing the 108 journals in WOKJ, we exclude the influence of data sparseness of social reference, and thus WOKJ is fair for an evaluation on the quality of connections among articles implied by social reference data; by keeping the 57 authors in MSCS, the dataset can indicate the influence of data sparseness of social reference in a practical setting.

2.3 EVALUATION

In this section, we evaluate clustering using social reference data and compare with citation-based methods. The 40 ISI categories in WOKJ and 24 fields of computer science in MSCS are used as groundtruth for evaluation. We evaluate clustering by normalized mutual information (NMI) and adjusted rand index (ARI). We use KMeans for clustering. To reduce the influence of start points, we sampled 20 random sets of start points and use the 20 sets of start points for all experiments. Metrics reported are average value of 20 sets of start points. The 57 authors in MSCS that cannot be found in CiteULike are randomly assigned in clustering.

We first experiment for citation-based clustering methods. Three citation-based relations are used: bibliographic coupling (BC), co-citation (CO) and cross-citation (CR). For feature vectors created for each relation, normalization methods in 2.1 are experimented. Best methods by NMI are selected as baselines: for WOKJ dataset (journal clustering), cross-citation normalized by BV is selected; for MSCS dataset (author clustering), co-citation normalized by BM25 is selected. Results are reported in table 1.

Then, we experiment for social reference-based methods (OC and COOC) in journal clustering and author clustering, and compare with citation-based baselines. Table 1 reports the results (for each method, top 3 results using different normalization are reported). For journal clustering in WOKJ, the best social reference-based methods (OC+BM25) are comparable to the baseline (CR+BV): OC+BM25 is slightly worse in NMI while slightly better in ARI. Because we exclude the influence of CiteULike data sparseness in WOKJ by removing journals that cannot be found in CiteULike, results in WOKJ indicate connections among articles implied by social reference data is as effective as citation in clustering. For clustering of authors in MSCS, social reference-based clustering methods have about 10% lower NMI and ARI than citation-based methods, which indicates the sparseness of social reference data (the 57 authors cannot be found in CiteULike) will influence the effectiveness of social reference clustering in a practical setting.

Results reported in table 1 indicate social reference implies high-quality relations among literatures, while the sparseness of data in a practical setting influences the effectiveness of social reference. Considering social reference is timely data source compared with citation, we select only articles published in 2010 for experiments. Table 2 reports the results: for journal clustering in WOKJ, social reference based methods have slightly better results than citation based methods; for author clustering in MSCS, social reference based methods are better than citation based methods. Compared with table 1, table 2 indicates: 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 timely data source and can outperform citation in clustering new resources (which is more significant for author clustering).

Table 1. Results for social reference-based clustering.

Dataset	Method	Norm	Evaluation Metrics	
			NMI	ARI
WOKJ Journal Clustering	**Cross-Citation**	**BV**	**0.645**	**0.277**
	Occurrence-based	Raw	0.620	0.281
		BM25	**0.624**	**0.294**
		LM-DIR	0.623	0.270
	Co-occurrence	--	0.613	0.275
MSCS Author Clustering	**Co-citation**	**BM25**	**0.701**	**0.599**
	Occurrence-based	TF×IDF	0.633	0.548
		BM25	0.637	0.555
		LM-DIR	**0.640**	**0.552**
	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	**BM25**	**0.614**	**0.254**
MSCS	Co-citation	BM25	0.509	0.207
	Occurrence-based	**LM-DIR**	**0.532**	**0.264**

3. ACKNOWLEDGMENTS

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

4. REFERENCES

[1] Kurtz, M. J., Bollen, J. 2010. Usage Bibliometrics. *Annual Review of Information Science and Technology*, 44, 3-64.

[2] Priem, J., Hemminger, B. M. 2010. Scientometrics 2.0: Toward new metrics of scholarly impact on the social Web. *First Monday*, 15, 7.

[3] Priem, J., 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.

[4] Ni, C., Sugimoto, C., Jiang, J. 2011. Degree, Closeness, and Betweenness: Application of group centrality measurements to explore macro-disciplinary evolution diachronically. In *Proceedings of ISSI 2011*, Durban, South Africa.