A Multi-Criteria Hybrid Citation Recommendation System Based on Linked Data

Fattane Zarrinkalam, Mohsen Kahani
Web Technology Lab., Dept. of Computer Engineering
Ferdowsi University of Mashhad
Mashhad, Iran
zarrinkalam.fattane@stu-mail.um.ac.ir, kahani@um.ac.ir

***Abstract*—Citation recommendation systems can help a researcher find works that are relevant to his field of interest. Currently, most approaches in citation recommendation are based on a closed-world view which is limited to using a single data source for recommendation. Such a limitation decreases quality of the recommendations since no single data source contains all required information about different aspects of the literature. This paper proposes a citation recommendation approach based on the open-world view provided by the emerging web of data. It uses multiple linked data sources to create a rich background data layer, and a combination of content-based and multi-criteria collaborative filtering as the recommendation algorithm. Experiments demonstrate that the proposed approach is sound and promising.**

Keywords-Citation Recommendation; Linked Data; Enrichmen; Recommender Systems; Multi-criteria

I. INTRODUCTION

In the domain of research, since the number of publications has exponentially increased, the major task of every researcher is to acquire appropriate knowledge about current state of his research area. Finding related publications can be complex and time consuming. One approach is to start with an important related work and trace its citing and cited papers. Another approach is using traditional keyword-based search engines such as Google. Both of the approaches provide a long list of publications to be studied, and they need manual filtering which is very tedious and inefficient.

A solution for these problems is to use a publication recommendation system which inputs user's interest and recommends publications that are related to his field of interest. A citation recommendation system, as a publication recommendation system, takes a text document as input, and recommends a list of publications which should be cited in different places of the input document. If such a system is available, the researcher can write an essay about his idea and then use the system to find recommended citations, which are publications related to his essay.

This paper proposes a citation recommendation approach which utilizes Linked Data to improve the background data and uses a new recommendation algorithm based on a combination of content-based and multi-criteria collaborative filtering.

The rest of the paper is organized as follows. Section II presents the necessary background concepts and motivation of the current work. Section III gives a brief summary of the related work. The proposed architecture is described in Section IV. Section V is dedicated to evaluation of the results and experiments. Finally, Section VI concludes the paper and presents some directions for the future works.

II. BACKGROUND AND MOTIVATION

From the mid-90s, recommender systems are successfully deployed in different domains, e.g. entertainment, e-commerce, e-learning, and research.

Adomavicius et al. [1] describe a recommender system as "an information filtering technique that seeks to identify a set of items that are likely of interest to users". Burke [2] describes that every recommender system is composed of three essentials elements: (1) background data: the data that the system has before the recommendation process begins, (2) input data: the data that is given to the system in order to get recommendations from it, (3) recommendation algorithm that processes the background and input data, and generates recommendations.

Adomavicius et al. [1] describe that the main techniques employed by recommendation algorithms are: (1) Content-based filtering: The algorithm recommends items that their content is similar to the contents of the items which the target user has used in the past. (2) Collaborative filtering: The algorithm recommends items that are used by people who have similar preferences with target user. (3) Hybrid approaches: The algorithm uses both collaborative and content-based methods. Later, they introduce a technique named multi-criteria rating, for extending capabilities of recommender systems. In this technique multiple criteria is used for calculating predictions and generating recommendations [3].

In citation recommendation systems, the background data includes bibliographic data, e.g. data about publications, authors, and venues. Here, like in any other recommender system, richness of the background data has great effect on the quality of recommendations.

There are a number of bibliographic datasets available on the web, each containing data about some publications. For instance some publications are indexed by *IEEE*, but not by *DBLP* or *ACM*, and vice versa. It is probable that a publication is included in a dataset, but some publications that are related to it, e.g. its references, are not. It is worth noting that even if a publication is included in two datasets, they may contain different data about its details. Keywords

of a paper might be indexed in a digital library, but not in another one. As a result, relying on a single bibliographic dataset as the background data means that relations are not sufficiently covered and data of publications is subject to missing values. These lost relations and missing values directly reduces quality of recommendations.

The motivation of this paper is to use Linked Data to tackle with these problems for developing a citation recommendation system.

Linked Data is a data publishing approach based on a set of four rules defined by Tim-Berners Lee as the Linked Data principles [4]. These rules describe how the data that is to be published must be addressed, represented and accessed. They have the great benefit that since they use semantic web technologies for describing data, data becomes more formally expressed. Semantic of data, and also its relation to other data is explicitly defined and therefore it becomes more machine-processable. This consequently reduces problems of data integration across different heterogeneous datasets.

It seems interesting to have a citation recommendation system which, instead of relying on a single local dataset, uses multiple Linked Data sources of bibliographic data. This enables the recommender system to utilize the benefits of Linked Data for providing rich background data in terms of both the data of publications and their relations.

III. Related Work

In this section, the related work on Linked Data driven recommendation systems, as well as publication recommendation is briefly reviewed.

A. Linked Data Driven Recommendation Systems

While almost all current recommender systems use central data sources, Passant et al. [5] introduces the new generation of recommender systems that, by utilizing the benefits of Linked Data, are not limited to a single central data source.

Heitmann and Hayes [6] discuss using Linked Data to build open collaborative recommender systems. They demonstrate the validity of their approach by augmenting the data of a real collaborative music recommender system with data from Linked Data sources. It is shown that using Linked Data improves the quality of recommendations.

Passant [7], introduces *dbrec*, a music recommendation system built on top of DBpedia. Shabir and Clarke [8] discuss using Linked Data as a basis for recommending learning resources in e-learning environment and describe the questions that arose around the provenance, sustainability, licensing and reliability of today's Linked Data cloud.

B. Publication Recommendation Systems

The works focusing on publication recommendation systems, based on their purpose, can be divided into two categories. In the first category, the system takes a user's profile as an input and recommends the publications that are similar to the user's profile. Basu et al. [9] studied the problems of publication recommendation systems in the context of submitting conference paper to the reviewers according to their interests and backgrounds. They used an automatic method in order to collect information of reviewer's interest from the web.

Bogers and van den Bosch [10] used *CiteULike*, a social reference management system, to recommend scientific publications to users, according to their reference library.

Chandrasekan et al. [11] and Pudhiyaveetil et al. [12] proposed a recommender system based on similarities between concepts of user profiles and the concepts of each document in the background data. They are different in terms of creating user profile. Chandrasekan et al. [11] created profile of a user according to his authored publications in the CiteSeer database, while Pudhiyaveetil et al. [12] created user profiles based on their previously viewed documents.

In the second category of works, known as citation recommendation systems, the input is a document, and the output is a list of publications that are related to the input document. In this category, McNee et al. [13] proposed an approach which uses collaborative filtering by analyzing the citation graph of documents, and building different rating matrices.

Torres et al. [14] presented different techniques for building a hybrid recommender system that uses collaborative and content-based filtering algorithms. They concluded that hybrid algorithms generate better recommendations than non-hybrid ones.

Tang and Zhang [15] used a topic based approach which uses a two-layer Restricted Boltzmann Machine model to discover the topic distributions of papers and the citation relationships, simultaneously.

Strohman et al. [16] presented an approach based on the linear combination of text features and citation graph attributes like citation count and Katz distance. Bethard et al. [17] introduced a algorithm which employs new features like topic similarity and author behavioral patterns, in addition to features like citation count and publication year.

He et al. [18], proposed an approach which recommends citations for specific locations in the input document. The recommendation algorithm works on the basis of a non-parametric probabilistic model. This model measures the relevance of two documents based on their contexts.

TABLE I. shows current works in citation recommendation and also the technique used by each work. As it is shown by this table, currently all citation recommendation systems have the limitation that they use a local data source as the background data. This means they work based on the implicit assumption that all the data requirements of the recommender algorithm will be satisfied by that source.

The goal of this paper is to assess the benefits of using Linked Data and using advantages of hybrid and multi-criteria recommender systems for improving the performance of citation recommendation systems. To do so, a new approach is presented for utilizing multiple Linked Data sources for enriching the background data. Further, the proposed approach uses a combination of multi-criteria collaborative and content-based filtering algorithms for generating recommendations.

TABLE I. CURRENT WORKS ON CITATION RECOMMENDATION

	Content-based filtering	Collaborative filtering		Using Linked Data in background
		Single criteria	*Multi criteria*	
[13]		✓		
[14]	✓	✓		
[15]	✓			
[16]	✓			
[17]	✓			
[18]	✓			
Proposed Approach	✓		✓	✓

IV. THE PROPOSED ARCHITECTURE

The proposed system is a hybrid recommender system, since it uses multi-criteria collaborative filtering and content-based filtering techniques. Its architecture is illustrated in Fig. 1. Two main components of the architecture are preparation and recommendation units. These components are described in the following sections.

A. Preparation Unit

The goal of this unit is to setup the background data required for the recommendation algorithm. It includes two processing phases. In the first phase, the local bibliographic dataset is enriched by an enrichment process which utilizes Linked Data sources. In the second phase, multiple user-item matrices are generated. These matrices are essential for the multi-criteria collaborative filtering method employed by the proposed system.

1) Phase1: enriching local dataset

The local dataset initially contains data about N publications P_i $(1 \le i \le N)$, where:

$P_i = (text_i, refList_i, citList_i, authList_i, year_i)$

$text_i$: { T_i, abs_i, $keyList_i$ }

T_i: title of P_i

abs_i: abstract of P_i

$keyList_i$: list of the keywords of P_i

$refList_i$: { P_j | P_i cites P_j }

$citList_i$: { P_j | P_i is cited by P_j }

$authList_i$: list of the authors of P_i

$year_i$: publication year of P_i

This dataset is enriched by utilizing multiple Linked Data sources which publish bibliographic data on the LOD cloud. The enrichment process is performed in an offline mode, i.e. before the recommendation process is initiated, however, in order to keep the background data up-to-date, it can be performed periodically.

Briefly speaking, the enrichment algorithm consists of two steps. In the first step, for each publication from the local dataset, a set of predefined external Linked Data sources are searched to see if they also define the same publication. Results of this step are captured as a set of equivalence relations between the publications from the local dataset and the publications from the external Linked Data sources.

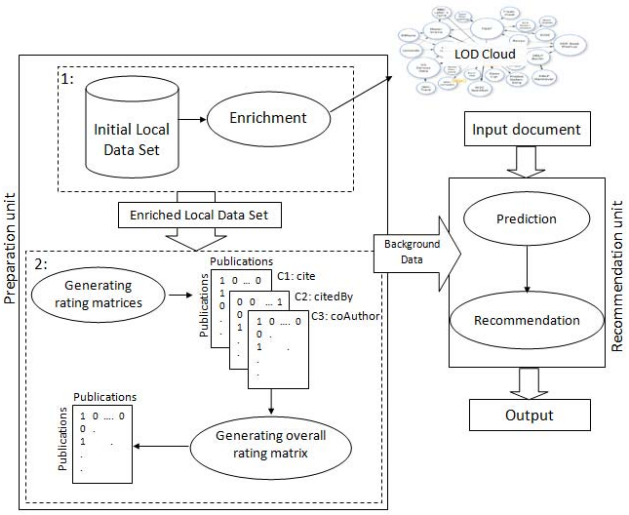

Figure 1. The proposed system architecture

In the second step, for each publication in the local dataset, its missing values in the background data are eliminated by the use of its external equivalences. Details are described in [19].

2) Phase 2: generating user-item rating matrices

User-item rating matrix is the essential element of the collaborative filtering technique. In the proposed architecture, this phase, which also consists of two steps, is done over the enriched local dataset.

In the first step, since the proposed recommendation system is based on multi-criteria collaborative filtering [3], three criteria, i.e. C_1: *cites*, C_2: *citedBy*, C_3: *co-author*, are chosen and three publication-publication matrices are generated: M_1, M_2 and M_3.

The value of the $M_k[i, j]$ (i.e. the element at row i and column j of matrix k), $1 \le k \le 3$, is defined as below:

$$M_1[i,j] = \begin{cases} 1 & if\ P_j \in refList_i \\ 0 & otherwise \end{cases}$$

$$M_2[i,j] = \begin{cases} 1 & if\ P_j \in citList_i \\ 0 & otherwise \end{cases}$$

$$M_3[i,j] = \frac{|authList_i \cap authList_j|}{|authList_i \cup authList_j|}$$

where $1 \le i \le N$, and $1 \le j \le N$.

For each matrix M_k, the value of $M_k[i, j]$ shows the relatedness of publication P_i and P_j, in terms of criteria C_k.

These criteria are based on the following ideas. When a publication cites, or is cited by, another publication, the two publications must be semantically related to each other. This is reflected in two criteria *cites* and *citedBy*. Further, since publications of an author are usually focused on a specific domain, they can be considered as related. This is the idea behind criterion *co-author*.

In the case of criterion C_3: *co-author*, it is important to note that relatedness of two publications, from the point of view of common authors, is directly related to the number of their common authors, and also inversely related to the total

number of their authors. For instance, *2 common authors out of 3 authors*, in comparison to *2 common authors out of 6*, indicates a stronger relatedness from the point of view of common authors.

In the second step, an overall single-rating matrix $M_{overall}$ is generated from the three rating matrices. The value of the $M_{overall}[i, j]$ is defined as:

$$M_{overall}[i,j] = \sum_{k=1}^{3} M_k[i,j]$$

B. Recommendation Unit

This unit is responsible for generating an ordered list of recommended citations that are related to the input document. To do so, a hybrid multi-criteria recommender is used. Based on the definitions presented by [2], this hybrid recommender system is of type feature combination.

Like all techniques based on collaborative filtering, it operates in two sequential steps: prediction, and recommendation.

1) Prediction

In this step, in order to predict the citations for the input document, first its neighbors are identified by the use of a content-based filtering method.

In this method, since the input document, i.e. *input*, is only a text, and not a structured content, for each publication P_i $(1 \leq i \leq N)$ in the local dataset, its textual similarity to *input* is computed by calculating *cosine similarity* of *TF/IDF* vectors of P_i and *input*. These vectors are generated over $text_i$ and text of *input*. Then, top *C* publications with the most textual similarity to *input* are selected as its neighbors.

Next, for each publication P_j $(1 \leq j \leq N)$ in the columns of matrix M_{overal}, its weight is calculated by the following formula.

$$weight_j = \sum_{\substack{P_i \text{ is a neighbor} \\ \text{of input}}} M_{overall}[i,j]$$

2) Recommendation

In this step, first the publications P_j $(1 \leq j \leq N)$ are ordered by $weight_j$ and then by $year_j$, both in decreasing order (i.e. first ordering is performed based on the weight, but if two citations have the same weight, the one with more recent publication year becomes first). Finally the first *K* elements of this ordered list are selected as the recommended citations.

V. Exprimental Evaluation

The proposed architecture is implemented in Java and evaluated through experiments described in next sections.

A. Experimental Setup

1) Experimental Parameters

As mentioned in Section IV.B.1), top *C* publications with the most similarity to *input* are selected as its neighbors. Through initial experiments, it was identified that a value of 50 for *C* appropriately preserves balance between execution time and performance of the system.

For the value of *K* (Section IV.B.2)), ten different values ranged over [15, 150] were used to measure the performance of the proposed approach in different ranges.

2) Data Set

To generate the initial background data, information of about 30000 publications were retrieved from *CiteSeerX*. This was done by developing a special harvester that uses OAI service[1] of *CiteSeerX* for collecting its information.

After collecting this data, it was identified that there are publications that values of some of their attributes, e.g. abstract or title, are missing. The collected data was filtered to remove unpromising publications, i.e. those with no title and no abstract. This caused removal of about 30% of publications.

Then, the collected data was enriched by the enrichment algorithm mentioned in section IV.A.1). Three Linked Data sources ACM[2], DBLP[3], and IEEE[4] have been used in this step.

After this enrichment process, publications published after 2007 were removed from the background data and used as the input data. Further filtering processes led to about 12000 publications for local data set and about 600 publications as the input data for testing the proposed system.

3) Evaluation Metrics

For evaluating the proposed approach, an automatic evaluation approach is used. In this approach, for every input publication, its text is given to the recommender system as input, and its list of references is considered as the excepted output.

Different metrics can be used to measure quality of recommendations. In this paper three metrics *Recall*, *Co-cited probability* and *NDCG* (Normalized Discounted Cumulative Gain) are used.

Recall is defined as the percentage of input references that appear in the top *K* recommended citations. The recommendations that are not in the reference list of the input publication cannot be considered totally unrelated to it. Therefore, *Co-cited probability* is used as a metric for measuring quality of such recommendations. *NDCG* is used for evaluating the order of the recommendations in the output. More details about these metrics is presented in [18].

4) Comparing Methods

In the experiments, three different recommendation methods are compared with each other. They are different in terms of their recommendation algorithm and the background data.

Method1: It is an implementation of the approach proposed in this paper, which uses both content-based filtering and multi-criteria collaborative filtering. It is executed on a background data which is enriched by the proposed enrichment algorithm.

[1] http://citeseerx.ist.psu.edu/oai.html
[2] http://acm.rkbexplorer.com
[3] http://dblp.rkbexplorer.com
[4] http://ieee.rkbexplorer.com

Method2: The recommendation algorithm of this method is the same as Method1, but no enrichment process is included in preparing the background data.

Method3: The background data is the same as Method2, but the recommendation algorithm only uses content-based filtering technique, by using textual similarity to generate recommendations.

B. *Experimental Results*

The three recommendation methods described in the previous section have been compared and the evaluation results are shown in Fig. 2 to Fig. 4 in terms of evaluation metrics.

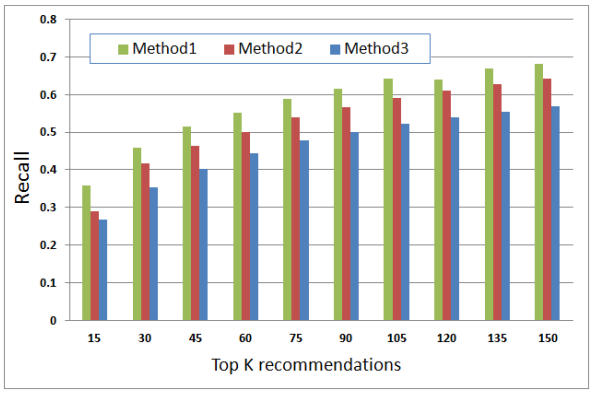

Figure 2. Comparison results in terms of *Recall*

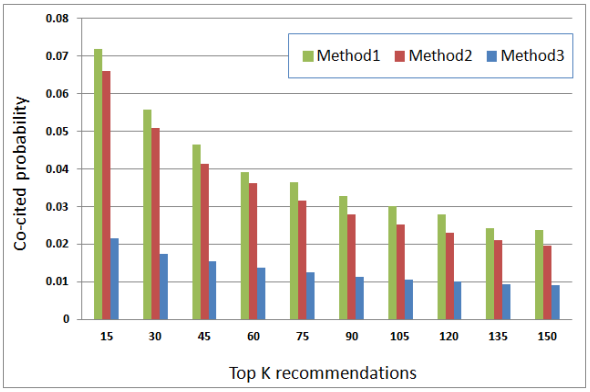

Figure 3. Comparison results in terms of *Co-cited probability*

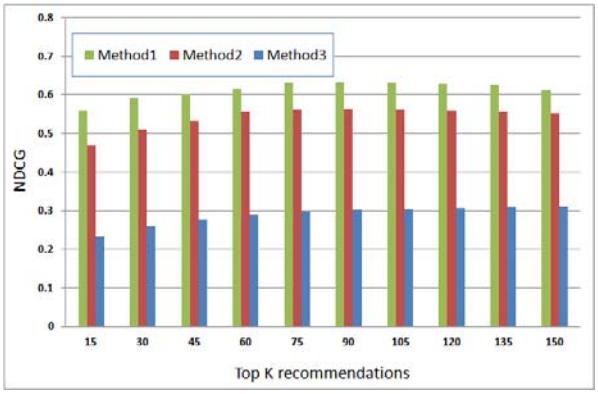

Figure 4. Comparison results in terms of *NDCG*

By comparing Method2 and Method3, it is possible to evaluate the effectiveness of using the proposed recommendation algorithm in comparison with the base method that only uses textual similarity for the recommendation. As illustrated in Fig 2. to Fig. 4, Method2 considerably outperforms Method3 in terms of all three metrics. Therefore the idea of using a hybrid approach has been effective.

Further, in order to evaluate the role of Linked Data in the proposed approach, Method1 and Method2 are compared. As illustrated in Fig. 2 to Fig. 4, the Linked Data driven enrichment process has improved the quality of recommendations but it isn't as much as expected. Its reason is discussed below.

Through manual investigations, it was identified that currently, data sources that publish bibliographic information on the LOD cloud, do not yet provide adequately rich and high-quality data, compared to what these data sources provide on the web of documents, i.e. the traditional Web, and they contain high level of missing data. This fact reduces the effectiveness of the proposed Linked Data driven recommendation system.

Nevertheless, the point of view is that with fast evolution of the web of data, sooner or later, the situation becomes better and publishers of bibliographic datasets provide more quality data on the LOD cloud. The more high quality data exists on the LOD cloud, the more effective the proposed approach can be realized.

VI. CONCLUSION

In this paper, a new architecture for citation recommendation systems is proposed in which the recommendation algorithm is a combination of content-based and multi-criteria collaborative filtering. Linked Data is used in preparing the background data, to release the recommender system from the limitation of relying on a single central dataset. It was shown that by utilizing different sources from the LOD cloud, it is possible to enrich the background data of the system and provide better recommendations. Experimental results prove the soundness and effectiveness of the presented approach.

Future work includes improving the recommendation algorithm by using more criteria, identifying the effect of each criterion on the quality of the recommendations, and using more bibliographic Linked Data sources to enrich the local dataset.

REFERENCES

[1] G. Adomavicius, A. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," J. IEEE Trans. Knowl. Data Eng. Vol. 17, no 6, pp. 734-749, 2005.

[2] R. Burke, "Hybrid Recommender Systems: Survey and Experiments," J. User Modeling and User-Adapted Interaction, vol. 12, no. 4, pp. 331-370, 2002.

[3] G. Adomavicius, N. Manouselis, Y. Kwon, "*Multi-Criteria Recommender Systems,*" in P. B. Kantor, F. Ricci, L. Rokach, B. Shapira (Eds.). Recommender Systems Handbook: A Complete Guide for Research Scientists and Practitioners Chapter 24. Springer, 2011.

[4] T. Berners-Lee, “Linked Data. Design Issues for the World Wide Web,” World Wide Web Consortium. http://www.w3.org/DesignIssues/LinkedData.html, 2006.

[5] A. Passant, B. Heitmann, and C. Hayes, “Using linked data to build recommender systems,” RecSys ‘9, New-York, NY USA, 2009.

[6] B. Heitmann, and C., Hayes, “Using Linked Data to build open, collaborative recommender systems,” AAAI Spring Symposium'Linked Data Meets Artificial Intelligence, pp. 76-81, 2010.

[7] A. Passant, “dbrec - Music Recommendations Using DBpedia,” International Semantic Web Conference (2), pp. 209-224, 2010.

[8] N. Shabir, and C. Clarke, “Using Linked Data as a basis for a Learning Resource Recommendation System,” 1st International Workshop on Semantic Web Applications for Learning and Teaching Support in Higher Education (SemHE'09), ECTEL'09, Nice, France, 2009.

[9] C. Basu, H. Hirsh, W. Cohen, and C. Nevill-Manning, “Technical paper recommendations: a study in combining multiple information sources,” J. Artificial Intelligence Research, vol. 1, pp. 231-252, 2001.

[10] T. Bogers, and A. Bosch, “Recommending scientific articles using citeulike,” ACM conference on Recommender systems, ACM New York, NY, USA. pp. 287-290, 2010.

[11] K. Chandrasekaran, S. Gauch, P. Lakkaraju, and H.P. Luong, “Concept-Based Document Recommendations for CiteSeer Authors,” 5th international conference on Adaptive Hypermedia and Adaptive Web-Based Systems (July 29-August 01). Hannover, Germany, 2008.

[12] A.K. Pudhiyaveetil, S. Gauch, H.P. Luong, and J. Josh Eno, “Conceptual recommender system for CiteSeerX,” *RecSys'2009*. pp. 241-244, 2009.

[13] S. McNee, L. Albert, D. Cosley, P. Gopalkrishnan, S. Lam, A. Rashid, J. Konstan, and J. Ried, “On the Recommending of Citations for Research Papers,” ACM conference on Computer supported cooperative work. New York, NY, USA, pp. 116-125, 2002.

[14] R. Torres, S. McNee, M. Abel, JA. Konstan, J. Riedl, “Enhancing digital libraries with TechLens,” IEEE/ACM Joint Conference on Digital Libraries (ACM/IEEE JCDL’2004). Tuscon, AZ, USA, pp. 228-236, 2004.

[15] J. Tang, J. Zhang, “A Discriminative Approach to Topic-Based Citation Recommendation,” PAKDD, pp. 572-579, 2009.

[16] T. Strohman, W.B. Croft, D. Jensen, “Recommending citations for academic papers,” 30th Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pp. 705-706, 2007.

[17] S. Bethard, D. Jurafsky, “Who should I cite? Learning literature search models from citation behavior,” ACM Conference on Information and Knowledge Management, pp. 609-618, 2010.

[18] Q. He, J. Pei, D. Kifer, P. Mitra, C.L. Giles, “Context-aware Citation Recommendation,” 19th International World Wide Web Conference (WWW), pp. 421-430, 2010.

[19] F. Zarrinkalam, M. Kahani, “Improving Bibliographic Search through Dataset Enrichment Using Linked Data,” International Conference on Computer and Knowledge Engineering (ICCKE2011), pp. 265-270, 2011.