# DLConnector: Connecting a publication list to scholarly digital library

Jen-Ming Chung[1], William W.Y. Hsu[1], Cheng-Yu Lu[1*], Kuo-Ping Wu[2], Hahn-Ming Lee[2], and Jan-Ming Ho[1]
[1]*Institute of Information Science, Academia Sinica, Taiwan*
{*jenming,wwyhsu,cylu,hoho*}*@iis.sinica.edu.tw*
[2]*Dep. CSIE, National Taiwan University of Science and Technology, Taiwan*
{*wgb,hmlee*}*@mail.ntust.edu.tw*
[*]*corresponding author*

***Abstract*—Researchers and graduate students spend a great deal of time on searching and reading papers. However, they may usually lack of experience on using proper keywords and finding the relevant papers can be challenging. To keep track on the newest activities of key authors and venues is important as well. In addition, some researcher present publication list on his/her home pages. To maintain the up-to-date information of the publication list for the web pages (i.e., citation number) is labor-intensive. In this paper, we incorporate our previous work and construct a Web 2.0 platform, namely *Digital Library Connector* (*DLC*). *DLC* provides services to facilitate the tedious research processes. Researchers can easily search relevant papers and subscribe other author's academic activity. Moreover, researchers can easily construct their Web 2.0 web pages to present their profile, publication list, and recent academic activity by using the service on *DLC*. The users show highly satisfactory feedback on using *DLC*.**

***Keywords-Key Paper Searching; Key Author Recommendation; Expertise Extraction; Web 2.0 Scholarly Platform;***

## I. INTRODUCTION

A serious literature survey of a new research topic for a beginning graduate student is an essential work. However, students may have limited experience on his/her research topic, and usually spend much time on searching and reading academic papers. This process, namely *Research Cycle* [21], is described as follow: *Using keywords to digital library*, *Reading papers from the returned results*, *Focus on key papers*, *Keeping track on important authors and venues*, and *Focus on research society*. One of the scenarios is that a student submits a list of queries to scholarly search engines to retrieve relevant papers. However, using proper keywords can be challenging, especially for beginners who are lack of experience and domain knowledge. To find relevant papers from huge returned results is time-consuming which may lead to much wasted effort to make determination of importance of the papers. Furthermore, it is necessary to keep track of the latest research, activities from the academic society.

In addition, experienced scholars usually need to maintain the self-owned publications for sharing the latest achievements and describing their research contributions. The most common way is to make a home-page or use the template provided by their institute. These web pages may usually not provide academic information for audience, i.e., citation number that partly represents the impact of the research topic.

In this paper, we incorporate with our recent [1], [2], [3], [4], [5] and some of previous work [6], [7], [8], [9], [10], [11] to develop a platform for scholarly use, namely *Digital Library Connector* (*DLC*), to facilitate the process (research cycle). The *DLC* provides researchers services as follows: (*i*) *Key paper searching*: searching the publishing lecture about the target conceptual entity; (*ii*) *Key author recommendation*: recommending the authors with the closely relevance with the given topic; (*iii*) *Expertise extraction*: extracting expertise by analyzing publications; (*iv*) *Web 2.0 platform construction*: providing services for researchers sharing and maintaining the publication data, users subscribe researchers and papers. The remainder of this paper is organized as follows. In Section 2, we review the state-of-the-art research, related work, and our previous research results. We first formulate a topic model for multiscale dynamics, and describe its online inference procedures. In Section 3, the platform *Digital Library Connector* is proposed, we elaborate the system architecture and components. In Section 4, we demonstrate the effectiveness of the proposed method by analyzing the dynamics of real document collections. Finally, discussion and conclusion are presented.

## II. RELATED WORK

To develop an academic platform to support the research cycle requires to keep track on (1) Relevant papers retrieval. Researchers need focus on not only the newest papers but also the classic papers of a topic (2) Key authors and venues localization; Key authors usually present advanced research on famous conference/journals (3) A platform that provides these services.

### A. Key Paper Searching

Chen *et al*. [4] proposed a citation-network-based methodology, namely Citation Authority Diffusion ($CAD$), to rapidly mine the limited key papers of a topic, and measure the novelty on literature survey. A defined Authority Matrix ($AM$) is used to standardize duplication rate of authors and to describe the authority relation between the citing and

 ISBN: 978-1-61208-200-4 

the cited papers. Based on $AM$, our $CAD$ methodology leverages the Belief Propagation to diffuse the authority among the citation network. Therefore, $CAD$ transforms the citation network to a novelty paper list to researchers. The experimental results show $CAD$ can mine more novelty papers by using real-world cases.

### B. Key Author Localization

Wu *et al*. [3] presented an approach applying Web Mining to recommend key authors of a topic. The authors designed a measure, namely *p*-index, for the ranking of researchers. Users can use academic keywords such as "Data Mining", the service returned a list of ranked authors.

### C. Expertise Extraction

Lu *et al*. [2] and Yang *et al*. [8] analyzed researcher's publication list and turned out their expertise. They both adopt *Wikipedia* as ontology which contains many fashion terms such as "Cloud Computing" which has not yet been recorded in current existing ontology (e.g., Wordnet).

### D. Related Service and Platform

There have been numerous developments in references management tools makes a great assistance in building personal library. Zotero [33] covers thousands of sites that senses content automatically to allow users have a personal library. CiteULike [32] and Connotea [30] provide online service for managing and discovering scholarly references as well as personal library building. Several desktop softwares, including Biblioscape, EndNote, Mendeley and BibDesk with the similar abilities.

Hoang *et al.* [13] proposed a bowser extension and web service called ***Scholarometer*** for academic impact analysis. Also an alternative named *Publish or Perish* [34] that is a software to retrieve and analyze citation information from Google Scholar to present several academic statistics. Their common characteristic is that the individuals of which they utilize the academic citations from Google Scholar to present a fast way to obtain the impact analysis in many aspects. However, users without the rights to make modification to these presented data such as remove the publications which not belonging to self permanently.

Some systems such as Arnetminer [27] and Microsoft Academic Search [26], are committed to provide extensive search and mining services for scholarly communities and network. Both of them make efforts in aggregating academic data from multiple sources and automatically build profiles. In addition, Odysci [28] pays attention to contribute a place where can express opinions on articles. Moreover, Google Scholar Citations [29] provides an easy way for researchers to realize the citation number of the publications.

Based on the analysis of existing state-of-the-art work, it appears that a sophisticated platform with services, such as Key paper searching, Key author localization, is important. In the next chapter, we propose a Web 2.0 platform, *DLC*, which provides several essential services for the scholarly use. There are several distinctions of *DLC*firstly offers a simply way to users to create their own citation repository; DLC recommends key papers and impact scholars by analyzing user's expertise. We also adopt crowdsourcing to take responsibility to maintain the correctness of publications and profiles as well as *Wikipedia*, all the registered users are allowed to edit the publication records on the system.

## III. DLC: Digital Library Connector

In this section, we describe the main features of proposed *DLC* platform, which is currently accessible at http://dlc.iis.sinica.edu.tw. The framework of DLC is composed of three conceptual modules: data integration, scholar recommendation and expertise exploration. Figure 1 presents the architecture of DLC. The back-end database is integrated of existing online scholarly repositories which are Google Scholar, DBLP, and user-provided data, such as, researcher's personal publication web page. Digital libraries such as DBLP and CiteSeerX have provided web-services to provide data and metadata, they dump the database into XML format and is compliant with the Open Archives Initiative Protocol. We aggregate attributes (i.e., metadata) from different data sources for different uses. In the *DLC* interface, *DLC* offers various information of author, expertise, publication list, co-author, related venue, and citation number. To provide the information, several components are designed to fulfill the needs. We elaborate the main components "Data Collection", "Object Aggregation", and "Object Recommendation" in the subsections.

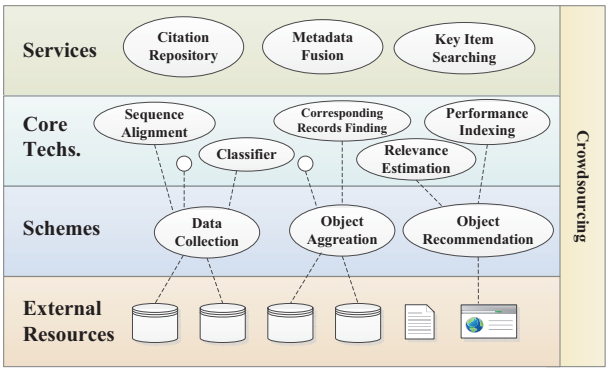


Figure 1. System Architecture

### A. Data Collection

In order to improve the satisfactory coverage, DLC integrates Google Scholar, DBLP, and Citeseer. The integrated data are refined and integrated to provide further implementation and services, such as direct answer. We apply the object schema presented by Nie *et al*. [19]. We also employ entity-attribute-value data model to refer the real

world entities and relationship among objects based on connectivity. This part is considered the following aspects, including publication data collection, object aggregation, and maintenance scheme. These components and their usage are described below.

This phase commences establishing publications for each scholar by conducting an import task of DBLP XML records, which is formatted and organized bibliographic entries on major computer science venues and available online [23]. *DLC* periodically synchronizes the data from the original data sources, these data includes publication list and citation number. So far we have construct a skeleton structure of our repositories. Each entity extends the metadata (i.e., abstract, reference) from CiteSeerX [25] and Google Scholar [24].

However, in our own observation, the coverage of publications among those scholarly repositories is not perfect. Even though Google Scholar has satisfactory coverage of disciplines and citation information, but name ambiguous problem has not yet been solved. Furthermore, the abbreviation policy of publication in DBLP database causes parsing problem. Previous studies [7], [17], [18] show that the citation records are usually using almost the same sequence of HTML tags and under one parent node. We use our previous work for the extraction of citation records [7]. Furthermore, DLC provides function to import personal publication list which is described by well-organized BibTEX file.

The DLC also applied *Citation Record Extractor* (CRE) [7] for users to provide their personal publication list web page. CRE identifies candidate citation patterns within pages in the DOM tree structure, and then filters out irreverent patterns by using a length-distribution-based classifier. Before returning to client, the BibPro [1] is responsible for citation parsing process. It is a sequence-alignment based citation parser designed to extract components of citations in arbitrary formats.

## IV. Object Aggregation

### A. Object Aggregation

The aggregation of attributes among those physical scholarly repositories has been performed by using the similarity measure on title field of each citation string. *DLC* regards well-known DBLP dataset as our authorized publication material which contains more than 1.7 million publication records and is easy to be stored with tabular format in a DBMS. *DLC* employs *Edit Distance* to locate the corresponding data which is under a given threshold and the year value must be the same. Hence, *DLC* harvests abstract value from CiteSeerX and citation number from Google Scholar respectively. Note that the aggregation procedure requires time *DLC* establishes multiple query strategies to achieve the satisfactory coverage.

Table I
Query strategies for retrieving the corresponding candidates from Google Scholar

| Strategy | Query Composition | Mode |
|---|---|---|
| Author | author: <*name*> (with quotes) | Online |
| Title | allintitle: <*title*> (without quotes) | Queue |

### B. Query Strategies

*DLC* regards Google Scholar as the metadata source because the coverage of Google is satisfactory on scholarly lecture. However, several constrains make it difficult to access easily, such as the limited number of queries for a single client and the lack of convenient application programming accessing interface. In practice, *DLC* employs Web proxy technology for cross-domain XMLHttpRequest calls in JavaScript to access the external resources. To guarantee a short response time and save the query number, we have proposed the following query strategies as shown in Table I.

In this scheme, system will not raise the whole update procedure for each client's visit but update those authors who have not yet been updated. According to the observation, we find that the most fluctuations of citation number are centralized in top-$k$ cited papers. Therefore, for a specific author without updating over our defined threshold, the $author$ strategy will be triggered immediately for obtaining the approximately citation number. The number of queries range between 1 and 10 according to the name alphabet order. The similar results with [14] show the authors with unusual names have effective and precise return results. However, those who have the overlap part of names are insufficient for corresponding records and are accompanying with the ambiguous problem. Consequently, totally matched is not guaranteed that merely by using *name* strategy.

The remaining unmatched publications will be delivered to the *title queue* for next query. The *title* strategy just take the whole title without quotes as keyword send to the search engine. Each query will perform the *Edit Distance* measurement to the title of the first returned snippet. In the end, we mark these data to remind the clients to check if any typo exists.

## V. Object Recommendation

### A. Key Paper Recommendation

For a graduate student, to survey and study papers is important. To realize a details of a topic, it is required to read classic papers and recent presented papers. However, the students may not have experience on knowing which the important papers are? So, one of the component of *Object Recommendation* is *key paper survey* which returns a list of classic and new papers for a specific domain [4].

Searching and identifying the key papers can be regarded as a *recursive* process. Users usually run the process for couple times so that they can have part of important papers.

 ISBN: 978-1-61208-200-4 

They usually use some possible keywords, review the returned results, modify the keywords, and use a new query in each iteration. As long as one most relevant paper has been focused. Paper-to-paper can be applied to obtain a key paper by using the perspective on *link* to focus on more key papers. Briefly reviewing a paper can obtain information, including the authors, keywords, and references. We obtain a series of papers of a research topic from the same authors. The information of a paper contains categories and subject description, general terms, keywords, references and context, these information can be processed for the future search. Moreover, cross validating the references from these candidate papers can discover other interesting topic and related technologies. The process usually repeats several time so that user can focus on the research topic.

However, scholarly search engines have developed their ranking mechanisms by processing the user-input. The most common way is to use keyword matching method and to calculate the citation number. Based on the results from search engine, user mostly focuses on the top-*k* highly cited papers as their survey materials. This ranking algorithm brings the challenge of how to skip the well-known papers which we have already learned, to identify the novel papers becomes necessary. Generally, the novel papers are usually lack of enough exposure opportunities [4], which means that they usually receive few citations. Our previous work *Citation Authority Diffusion* (*CAD* in short)is deployed [4], *CAD* is a citation-network-based method to discover potential co-relations to reveal the critical papers. Hence, this unit recommends classic papers and avoids cold start problem.

The whole procedure is triggered by a target research paper $tr$ which was provided by the user. The *information collection* module then analyzes the title, abstract, and keywords of the input paper, and generates the key phrases. The key phrases are extracted by using part-of-speech tagging, linguistic filtering, and *C-value* [15] method. The key phrases are regarded as the input for the scholarly search engine in order to collect the related survey materials $sm$. In order to retrieve the potential papers and to construct the citation network of $sm$, the *authority propagator* and *believe propagation* method are applied. The most relevant bibliographies could easily be estimated by using the following Equation (1).

$$rel(s,d) = \frac{InDeg(s \mid d)}{InDeg(d)} \tag{1}$$

where $d \in sm$, current paper of survey material, $s$ is one of the siblings (i.e., references) belongs to $d$. $InDeg(d)$ represents the amount of citation number of $d$, and $InDeg(s \mid d)$ means the paper $s$ is cited by $d$'s citer. Moreover, the harmonic mean is calculated and is regarded as the threshold to filter the irrelevant candidates, the detailed equation is described as follows,

$$f(d) = \frac{|sib(d)|}{\sum_{s \in sib(d)} rel(s,d)^{-1}} \tag{2}$$

where $sib(d)$ is the reference set of $d$, and $|sib(d)|$ is the size of $sib(d)$. Only the $rel(s,d)$ greater than the threshold will be added into the citation network. The main idea is that the more common references they shared, the higher correlation they gained. Citation network is regarded as the input of the authority propagator to identify the key publication list. We leverage the *belief propagation* [16] with our potential function named *authority matrix*. The *belief propagation* focuses on initial weight setting and state transition to diffuse the authority as the belief. The belief propagation is based on the Equation (3) and Equation (4) used to infer the probabilities about maximum likelihood state from each paper in citation network.

$$m_{ij} = \sum_{\sigma'} \Psi(\sigma', \sigma) \prod_{n \in N(i) \setminus j} m_{ni}(\sigma') \tag{3}$$

$$b_i(\sigma) = k \prod_{j \in N(i)} m_{ji}(\sigma) \tag{4}$$

In the above formulas, $m_{ij}$ is the message vector sent by the paper $i$ to $j$ and $N(i)$ is the set of papers citing $i$, and $k$ is a normalization constant. An authority matrix $\Psi(\sigma', \sigma)$ is exploited on prior state assignment for each citation pair to standardize author duplication rate in citation network from 0 to 1. Generally, a paper holds higher authority if it is cited by another paper who is also having high authority. So, authority propagator regards $\Psi(\sigma', \sigma)$ as a diffusion factor and dynamically updates the authority for each pair. Finally, a converged network with certain authority is expected, i.e., the key paper list in novelty aspect.

### B. Expertise Extraction

Locating a specific researcher's expertise is always an important and essential task among scholar repositories. This information may facilitate people search which addressed on similar research domains. The expertise extraction procedure composed of two stages, including key term extraction and Wikipedia ontology inference. We incorporate our previous results [2] and [8] to extract researcher's expertise by analyzing their publication list.

## VI. CROWDSOURCING

In order to guarantee a satisfactory coverage and to collect more author-publication records, *DLC* uses crowdsourcing mechanism to let registered users to edit. As long as more users join the editing task, the less typos and omissions can be avoided. Registered users are allowed to upload their own publications, we also allow their co-authors and registered users to edit the records. This mechanism is as same as in Wikipedia. Therefore, part of meta-data management in

 ISBN: 978-1-61208-200-4 

our proposal framework is delegated to the public. Some unique social identifier such as OpenID, Facebook ID are recognized to apply for the ownership of a specific or an interested scholar in system to entitle an applicant the right to maintain the records, including author's profile and publication list. We express more details in publication maintenances in the following.

We propose an additional *verification bit* mechanism in each articles whether they be created by crawlers in advance or those be established recently by volunteers. Those users with authorities for management can perform the *lock* and *unlock* operation to the verification bit. A unlocked publication means that every registered user is allowed to modify. As long as a publication list has been verified, the authorized user can lock this citation to avoid the destruction with evil intension. This manner makes it easier on the management task to have the information with the interested publications, statistics of contributors and the number of the uncertified records.

## VII. Conclusion

In this paper, we incorporate our previous work and develop a Web 2.0 platform, namely *Digital Library Connector* (*DLC*). *DLC* provides services to facilitate the tedious research processes. Researchers can easily search key papers, key authors and subscribe key author's academic activity. Moreover, researchers can easily construct their Web 2.0 web pages to present their profile, publication list ,and recent academic activity by using the service on *DLC*. The users show their high satisfactory on using *DLC*.

## Acknowledgment

This work is partly supported by the National Science Council of Taiwan, under grant NSC 99-2221-E-011-075-MY3, NSC 100-2631-H-001-012, NSC 100-2811-E-001-003 and NSC 101-2631-H-001-006.

## References

[1] C.-C. Chen, K.-H. Yang, C.-L. Chen and J.-M. Ho, "BibPro: A Citation Parser Based on Sequence Alignment," IEEE Transactions on Knowledge and Data Engineering, volume 24, issue 2, pp. 236–250, 2012.

[2] C.-Y. Lu, S.-W. Ho, J.-M. Chung, H.-M. Lee and J.-M. Ho, "Mining Fuzzy Domain Ontology Based on Concept Vector from Wikipedia Category Network," Web Intelligence, 2011.

[3] C.-J. Wu, J.-M. Chung, C.-Y. Lu and J.-M. Ho, "Using Web-Mining Approach for Scholar Measurement and Recommendation," Web Intelligence, 2011.

[4] C.-H. Chen, C.-Y. Lu, H.-M. Lee and J.-M. Ho, "Novelty Paper Recommendation Using Citation Authority Diffusion," TAAI, 2011.

[5] J.-M. Chung, C.-Y. Lu, H.-M. Lee and J.-M. Ho, "Automatic English-Chinese Name Translation in Digital Library Management by Using Web-Mining and Phonetic Similarity ," IEEE IRI, 2011.

[6] K.-H. Yang, T.-L. Kuo, H.-M Lee and J.-M. Ho, "A reviewer recommendation system based on collaborative intelligence," Web Intelligence, pp. 564–567, 2009.

[7] K.-H. Yang, S.-S. Chen, M.-T. Hsieh, H.-M. Lee and J.-M. Ho, "CRE: An automatic citation record extractor for publication list pages," Proc. WMWA, 2008.

[8] K.-H. Yang, C.-Y. Chen, H.-M. Lee and J.-M. Ho, "EFS: Expert finding system based on Wikipedia link pattern analysis," Systems, Man and Cybernetic, pp. 631–635, 2008.

[9] K.-H. Yang, J.-M. Chung and J.-M. Ho, "PLF: A Publication list Web page finder for researchers," Web Intelligence, 2007.

[10] K.-H. Yang, H.-T. Peng, J.-Y. Jiang, H.-M. Lee and J.-M. Ho, "Author name disambiguation for citations using topic and web correlation," Research and Advanced Technology for Digital Libraries, pp. 185–196, 2008.

[11] K.-H. Yang and J.-M. Ho, "Parsing Publication Lists on the Web," Web Intelligence, pp. 444–447, 2010.

[12] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang and Z. Su, "ArnetMiner: extraction and mining of academic social networks," ACM SIGKDD, pp. 990–998, 2008.

[13] D.-T. Hoang, J. Kaur and F. Menczer, "Crowdsourcing Scholarly Data," Proc. of Web Science Conference: Extending the Frontiers of Society On-Line (WebSci), 2010.

[14] A. Thor, D. Aumueller and E. Rahm, "Data integration support for mashups," Proceedings of the Sixth International AAAI Workshop on Information Integration on the Web, pp. 104–109, 2007.

[15] K. Frantzi, S. Ananiadou and J. Tsujii, "The c-value/nc-value method of automatic recognition for multi-word terms," Research and Advanced Technology for Digital Libraries, pp. 520–520, 1998.

[16] J. Pearl, "Probabilistic reasoning in intelligent systems: networks of plausible inference," 1988.

[17] B. Liu, R. Grossman and Y. Zhai, "Mining data records in Web pages," Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 601–606, 2003.

[18] C.H.A. Hong, J.P. Gozali and M.Y. Kan, "FireCite: Lightweight real-time reference string extraction from webpages," Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, pp. 71–79, 2009.

[19] Z. Nie, J.R. Wen and W.Y. Ma, "Object-level vertical search," Third Biennial Conference on Innovative Data Systems Research, pp. 235–246, 2007.

[20] A. Thor and E. Rahm, "MOMA - A Mapping-based Object Matching System," CIDR, pp. 247–258, 2007.

[21] Keshav, S., "How to read a paper," ACM SIGCOMM Computer Communication Review, 2007.

[22] Digital Library Connector. http://dlc.iis.sinica.edu.tw.

[23] DBLP XML records. http://dblp.uni-trier.de/xml/.

[24] Google Scholar. http://scholar.google.com.

 ISBN: 978-1-61208-200-4 

[25] CiteSeerX. http://citeseerx.ist.psu.edu.

[26] MS Academic Search. http://academic.research.microsoft.com.

[27] Arnetminer. http://www.arnetminer.org.

[28] Odysci. http://www.odysci.com.

[29] Google Scholar Citations. http://scholar.google.com.

[30] Connotea. http://www.connotea.org.

[31] Scholarometer. http://scholarometer.indiana.edu.

[32] CiteULike. http://www.citeulike.org.

[33] Zotero. http://www.zotero.org.

[34] Publish or Perlish. http://www.harzing.com/pop.htm.

 ISBN: 978-1-61208-200-4 