Special issue: Leading ICT technologies in the Information Explosion

Research Paper

Academic linkage: A linkage platform for large volumes of academic information

Akiko AIZAWA[1], Atsuhiro TAKASU[2], Daiji FUKAGAWA[3], Masao TAKAKU[4], and Jun ADACHI[5]
[1,2,3,5]*National Institute of Informatics*
[4]*National Institute for Materials Science*

ABSTRACT

We propose a two-layered architecture for information identification that is specifically targeted towards academic information. We first introduce the basic notion of information identification, or *linkage*, that connects fragmented information referring to the same objects or people in the world. We then propose a linkage system that is composed of bibliography and researcher identification layers. As an illustrative example, the results of a coauthor relationship analysis are also shown.

KEYWORDS

Academic linkage, identification of database records, research community mining, name disambiguation

1 Introduction

In our daily lives, we reference and use various types of information. This information, which is stored on individuals' disk drives or on the Internet, forms an information space suited to how the person uses it. However, with the information explosion, this space is violently expanding. Duplicate and/or irrelevant information creeps in, burying the desired information, making it difficult to find, and increasing the cost of obtaining it.

In our research, we are studying ways to connect the fragmented information in this sort of very-large information space. This research centers on identifying descriptions that reference the same objects in the real world, such as people or things. Through this sort of identification, new links connecting separate information sources are generated, and we call this *information linkage*. In the database field, searching for duplicate records in a database is called record linkage, and records have clearly defined attributes. With information linkage, however, this is expanded so the items being linked are texts with no clearly defined attributes.

Information linkage consists of two steps: (1) Determining identity by applying matching functions to candidate pairs, and (2) Analyzing and summarizing the information based on the links extracted from it. Since different types of objects that are related to each other do exist, the information extracted in step (2) for one object type can be further utilized for identifying another object type in step (1). We propose a layered architecture for information linkage that is based on this.

Fig. 1 shows our current implementation of the linkage platform for academic information. With our framework, Layers 1 and 2 are associated with a bibliography and the authors, respectively, and the identification results in Layer 1 are propagated to Layer 2. This means that if two bibliographic records are identical to each other, then the authors of the two records are also considered to be identical even when the notation does not exactly match. Based on this, the bibliographic identification results can be utilized, for example, for name alignment or disambiguation of the authors.

In the following sections, we present our current scheme and implementation of an "Academic Linkage Platform" linking researchers and research papers. First, in Sec. 2, we introduce a fast bibliography-

Received September 24, 2008; Revised December 1, 2008; Accepted December 8, 2008.
[1] *aizawa@nii.ac.jp*, [2] *takasu@nii.ac.jp*, [3] *daiji@nii.ac.jp*,
[4] *TAKAKU.Masao@nims.go.jp*, [5] *adachi@nii.ac.jp*
DOI: 10.2201/NiiPi.2009.6.5

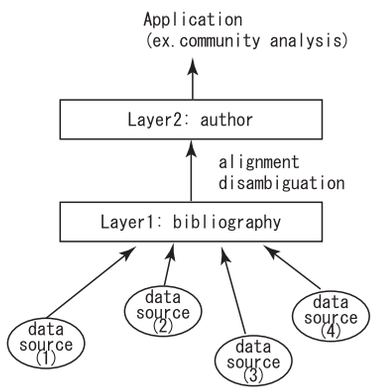


Fig. 1 Two-layered architecture of the academic linkage platform.

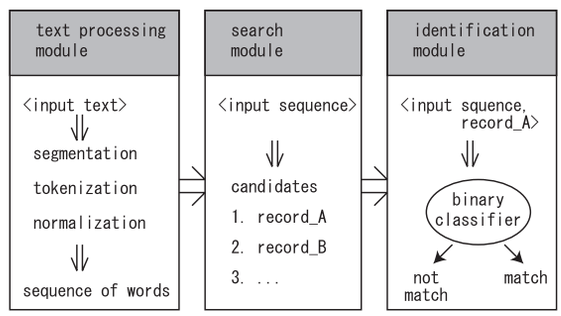


Fig. 2 Proposed bibliography linkage engine.

linkage system that extracts identical relationships between research papers. Next, in Sec. 3, we present an author-linkage system that links the authors of papers to a researchers' database. Then, in Sec. 4, we present an example of a community analysis using the proposed layered architecture.

2 Layer1: Bibliography linkage system

In this section, we focus particularly on the developing technology for highly-reliable information linkage using a large-scale database. We introduce our base technology for connecting a central database with other databases and information on the Web, and thus summarizing this scattered information.

2.1 Data source

The central database we used for the linkages was the bibliography database that the National Institute of Informatics (NII) provides as one of its services [1]. As of 2008, this database publishes bibliography data on approximately 11 million papers and contains over 50 million records, including unpublished citation data. Each bibliography record in the database includes the following field information.

$$\{document\_id, authors, title, journal, volume, number, pages, year\}.$$

2.2 Proposed linkage system

With bibliography linkages, we need a flexible identification method that can handle any variance and/or errors in the text collected from different information sources. In addition, in cases such as inputting documents directly from the Web, the unpredictability of the attribute structure as well as the missing values must also be considered.

Our proposed high-speed linkage system is based on a data structure called a suffix array, and can perform highly-flexible searches at a high-speed by managing a virtual, transposed file of variable-length word sequences [2]. Intuitively, we reduce the search cost and handle the noise by using word strings that are highly identifiable within the text. With this method, the output records only partially match rather than matching the input text entirely, and the results are records that have been determined to match, rather than a list of records ranked by level of similarity.

Fig. 2 presents an outline of the proposed bibliography linkage engine. The engine consists of three modules. The first is a text processing module that converts input text into a sequence of words after segmentation, tokenization, and normalization. The second is a search module that selects the specified number of candidates after a simple dynamic-programming based similarity calculation. The third is an identification module that contains a binary classifier to decide if a given pair of input text and a candidate record matches or not.

2.3 Implementation of linkage system

We are currently developing *i-linkage*, a bibliography linkage engine based on this method and linking text to records in a bibliography database. Input for the *i-linkage* system can take a variety of forms. The examples that follow are some we are currently working on.

- Records extracted from other databases
- Text documents from the Web or local disk
- OCR output from digital library images [3]
- Search results from libraries on the Web
- Inter-Library Loan (ILL) records

[1] http://ci.nii.ac.jp/

We have recently released a prototype server that includes the approximately 11 million bibliography records in NII's research paper database, and we are proceeding with the development of an upgraded version with improved search and performance functions. In the current implementation, attributes like the author and title in the input text must be entered in a specific order, although more than one may be specified beforehand. We are also looking for effective ways of computing similarity for tree-structured data in order to provide more flexibility [4], and applying these methods is another future task.

2.4 Discussion

We investigated the performance of the implemented bibliography linkage system by manually checking 3,000 samples randomly selected from the input text used in our later experiments. Using a simple string similarity-based matching function, we obtained a 95% accuracy for the estimated miss ratio of about 90-95%.

An example of the output from the *i-linkage* system is shown in Table 1. Here, the input text was automatically identified by OCR and has a segmentation error where two independent records are connected into a single sequence. As is shown in the table, our linkage system successfully detects both of the original records from the database. Note that conventional similarity measures do not work well with such errors.

Table 1 Linkage example.

[Input text]
Adachi, K. (2002). Optimal quantification of a longitudinal indicator matrix: Homogeneity and smoothness analysis. Journal of Classification 19, 215-248.Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control 19, 716-723.Benzecri, J.-P. (1979).
[Identification result using a database of academic papers] (two candidates)
AUTH: H. Akaike
TITL: A new look at the statistical model identification
JRNL: IEEE Transactions on Automatic Control
VOLN: 19, 6
PAGE: 716-723
YEAR: 1974
DOI: 10.1109/TAC.1974.1100705
AUTH: Kohei Adachi
TITL: Optimal Quantification of a Longitudinal Indicator Matrix: Homogeneity and Smoothness Analysis
JRNL: Journal of Classification
VOLN: 19, 2
PAGE: 215-248
YEAR: 2002
DOI: 10.1007/s00357-001-0044-8

3 Layer 2: Author linkage system

In our database, each bibliography record includes the title and journal as well as the authors' names and affiliations, with information on a total of over 110 million authors. The attributes and registration of this information is not necessarily fully consistent. For authors in particular, there are instances where different people have the same name string, or there are several different name entries referring to the same author. This is also the case for many other large libraries provided by publishing and search-engine companies [1]. The objective of author linkage is to assign a unique researcher number to an individual author of academic papers.

3.1 Data source

As a base for the experiment, we constructed a database of roughly 150,000 Japanese researchers from a scientific-research funding database [2]. Here, the original database was organized on a project-by-project basis and each project has information on its participating researchers with unique research numbers. By using the researcher number as the keys, we reorganized the database so that each research number contains the following information.

{ *Researcher number*, *Japanese name*, *English name*, *Japanese affiliation*, *English affiliation*, *List of cooperating researchers*, *List of related publications* }

In the above list, "cooperating researchers" are the researchers who joined the same research project in the past, and "related publications" are the publications of each research project. Since the publications are stored as just a plain text without any connection to any outside bibliography databases, the bibliography linkage engine in the previous section is used to extract corresponding bibliography records from the publication list. Note that no research number is attached to the authors of the publication and all that is known is that the publication was written by at least one of the participating researchers.

3.2 Proposed linkage system

Conventional approaches for person name disambiguation are based on the clustering of mentions with

[2] http://seika.nii.ac.jp/

their name strings being the same but referring to different persons. On the contrary, we formulated the problem as a process of generating a 'coreference' network. Here, the nodes of the network correspond to the individual authors of a bibliography database and the edges to the identical relationship between the nodes.

The proposed identification network operates in two distinctive modes. The first is a link generation mode where new links are sequentially added between the identified node pairs. For this purpose, promising node pairs are iteratively selected, and the existence of a coreference link is tested using a binary classifier. The second is an integration mode where the information in connected nodes is collected and summarized for presentation. For example, Japanese and English spellings of a researcher's name, together with their notation variations, are collected across different bibliography nodes. Such information can be further utilized for improving the identification result in a link generation mode.

3.3 Implementation

Due to the large scale of data, it is not possible to enumerate all the combinations of authors for the identification test, and it is crucial to exploit the information from outside sources for candidate pair selection. The following are used in our current implementation:

(1) Titles of papers
Using a string similarity search, titles within a given similarity distance are enumerated and used as candidates for author identification.

(2) Co-authors and cooperating researchers
When the co-authors and cooperating researchers have other already identified bibliography records, the authors of the bibliography become new candidates for author identification.

(3) Self-citation
Also, the authors of citing and/or cited papers are considered candidates for author identification.

(4) Bibliography cooccurrences
The authors of papers listed within the same document, such as project reports, become candidates for author identification.

For item (4), we also use our bibliography linkage server presented in the previous section. This enables us to even utilize documents on the Web, the examples of which include publication lists of researchers or research labs. This is a powerful information source for the authors without co-authors.

In link generation mode, a binary classifier is applied that is based on the string similarity of researcher names and affiliations. Note that we have information both in Japanese and English so that the names can be disambiguated even when the names are written in different (ex. Japanese and English) languages. Once one of the authors of a paper is linked, then that information can be propagated between the co-authors.

In integration mode, in addition to providing basic information, content from various information sources is gathered. Based on this, the platform provides an overview of researchers including not only information about financially-supported research projects, but also information obtained through links to other bibliography databases and information sources on the Web. The following basic fields are also included.

> { *Researcher number*, *Japanese name*, *English name*, *Japanese affiliation*, *English affiliation*, *List of publications*, *List of coauthors*, *Keywords*, *Citations* }

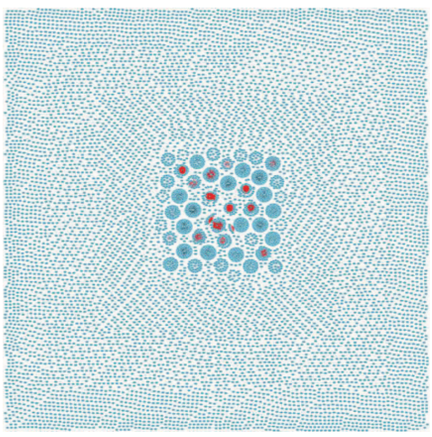
From the publication information, it is possible to obtain the history of researchers' affiliations with different facilities, and variants on their names. Also, using information registered in the database, such as abstracts, keywords and conference names, various types of analyses are possible. For example, the terminology in a specialized field can be extracted, or related researchers can be found using a correspondence analysis.

3.4 Discussion

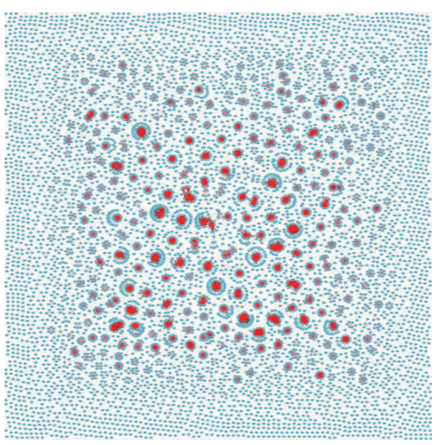
As before, we investigated the performance of the implemented system by manually checking 826 randomly-selected identification results. The accuracy value was 0.4% and we conclude that the identified results were mostly reliable. Note that the accuracy value is better for author identification since the misidentification of the bibliography system often includes confusions caused by similar articles by the same author group. However, this sort of confusion does not affect the author identification.

Fig. 3 illustrates the evolution of the identification network. Fig. 3 (a) is the initial status where the publication lists of 67 researchers were imported as seeds. Each node in the figure represents an individual author of an academic paper, and the red colored links are the identical relationships between nodes. The connected node groups correspond to 67 distinctive researchers. Fig. 3 (b) is after expanding the network using co-authorship relations. It can be seen that the initial identification results were propagated and new links were added to the network. Table 2 lists examples of notation variations collected from the linked nodes on the network. The extracted variety of expressions can be utilized for further identification.

For searches using just a name, problems due to people with the same name or using character or spelling variations are not dealt with, so some way to eliminate these types of ambiguities is still needed. The



(a) Isolated author nodes are connected using information from outside database as a seed



(b) New links are added based on co-author relationship

Fig. 3 Example: evolution of the identification network.

Table 2 Example: notation variations collected on the identification network.

Ministry of Education, Culture, Sports, Sci. and Technol., National Inst. Informatics, JPN
National Institute of Informatics
Nii, Tokyo, Jpn
National Inst. Informatics (nii), Toyko, Jpn
Kokuritsujohogakuken
Nii
Nii(National Institute of Informatics)
National Institute of Informatics(Nii)
National Instisute of Informatics

method we have used here, of combining information in a database with other information on the Web, can be used to handle these problems.

4 Application to researcher-community analysis

As an example of linkage between a database and the Web, we gathered and analyzed information on 319 researchers participating in the project entitled "New IT Infrastructure for the Information-explosion Era," a MEXT Grant-in-Aid for Scientific Research on Priority Areas [3]. As seed information, we used the publication lists from the project's annual report.

4.1 Bibliography and author identification

First, we used the bibliography linkage engine to link entries in the publication list to our bibliography database, and then extracted the papers authored and co-authored by the 319 targeted researchers. Based on the results, we generated an initial identification network for the researchers.

Next, in order to expand the identification network, we generated queries based on the extracted list of papers and submitted them to the search engine using the API, thus obtaining URLs that contained many of the papers by each author. The examples of such pages include lists of publications from individual researchers or research groups. Then, the targeted html documents were obtained and imported to our system through the bibliography linkage engine. Finally, together with the citation and co-authorship relationships already in the database, the identification network was updated and the whole process was iteratively repeated.

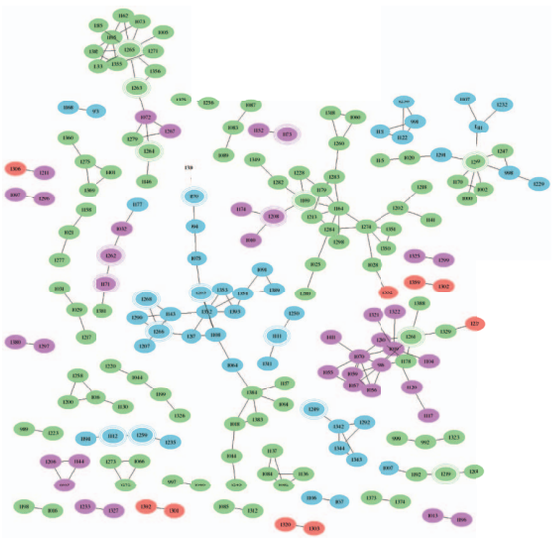
Through this process, we were able to automatically increase the original total of 1,216 papers to 17,011 papers. During the execution, the outside sources (1)-(4) were iteratively used for the expansion. Out of those, the information obtained from the Web (4) was the most useful. Without it, the number of the identified papers would have only been 2,182.

The initial network only contained publications in a specific domain for the period of 2006–2007, without linking back to each researcher, but the final reference list is organized by researcher, and includes publications all of the way back to the 1980s.

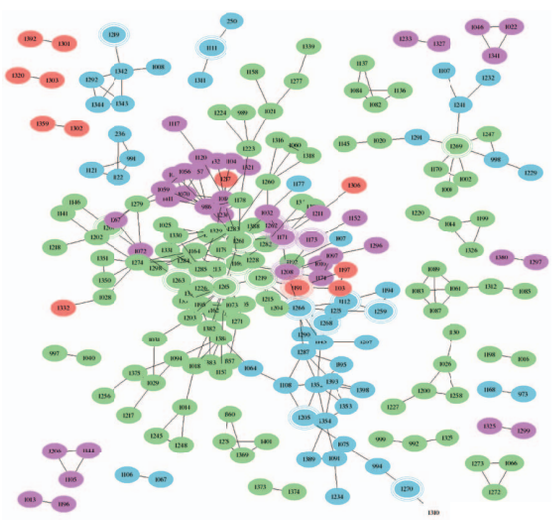
4.2 Analysis of co-authorship in the community

As an example, in Fig. 4 we present the change over time of the co-authorship network in a specific domain. The nodes are the researchers, and the links are the co-authorship relationships. The node colors indicate the different research groups within the domain. Fig. 4 (a) is based on the results from 2005, while Fig. 4 (b) is from December, 2007. This figure shows how, after a domain begins to develop, the creation of the connections across fields within the domain is activated. Ex-

[3] http://www.infoplosion.nii.ac.jp/info-plosion/



(a) co-author network in 2005.



(b) co-author network in 2007.

Fig. 4 Co-author network in 2005 and 2007.

amining newly added links suggests that collaborative research between supporting groups and movement of younger researchers are the causes.

The field of Scientometrics uses relationships, such as co-authorships, and citations to obtain a quantitative analysis of the scientific-production activity, but it is not easy to automatically generate the data required for this analysis. Our method provides a promising way to greatly reduce the amount of human work required to cleanup the data used for this sort of analysis.

5 Future developments

This paper described the information-linkage technology elements that connect a cross-section of information identified with specific people or things, and introduces an academic linkage platform that is currently under development.

One goal of academic linkage is to relate the information on papers and authors in the database to information in other databases and on the Web. By doing this, it will be possible to accomplish tasks like identifying authors and facilities or gathering information on related projects. Based on the policies of the various journals, the bibliography data also links to other information such as keywords, abstracts, and URLs for digital libraries, so an analysis of this rich text data is also very promising. In the future, we plan to use the results of this research to devise more practical systems.

In our study using academic databases, we showed that the proposed linkage system enables the reorganization of databases by assigning unique IDs to originally non-identifiable attribute values. We also showed that outside information resources, particularly ones obtained from the Web, are helpful for identification tasks. We believe the proposed linkage scheme combined with fine grained natural language processing techniques to extract entities and their relations will be one of the key techniques for future Web-database integration.

Acknowledgement

Finally, we would like to express our appreciation to Prof. Kitsuregawa for his constant, valuable comments and discussion, to the members participating in the NLP/IR explosion, to Prof. Keizo Oyama and Asst. Prof. Masashi Inoue of NII.

References

[1] D. LEE, J. Kang, P. Mitra, C. L. Giles, and B. On: "Are your citations clean?" *Commun. ACM* vol.50, no.12, pp.33–38, 2007.

[2] A. Aizawa, M. Takaku, and K. Oyama: "Proposal and implementation of a linkage system making use of large-scale databases," *DBSJ Letters*, vol.6, no.4, pp.17–20, 2008 (Japanese).

[3] A. Takasu: "Bibliographic attribute extraction from erroneous references based on a statistical model," *Proc. Of ACM & IEEE joint conference on digital libraries*, pp.49–60, 2003.

[4] A. Takasu, D. Fukagawa, and T. Akutsu: "Statistical learning algorithm for tree similarity," *IEEE ICDM*, pp.67–72, 2007.

Akiko AIZAWA

Akiko AIZAWA graduated from the Department of Electronics at the University of Tokyo in 1985 and completed her doctoral studies in electrical engineering in 1990. She was a visiting researcher at the University of Illinois at Urbana-Champaign from 1990 to 1992. At present, she is a professor at National Institute of Informatics and also an adjunct professor of the Graduate School of Information Science and Technology, University of Tokyo. Her research interests include statistical text processing, linguistic resources construction, and corpus-based knowledge acquisition.

Atsuhiro TAKASU

Atsuhiro TAKASU received B.E., M.E. and Dr. Eng. from the University of Tokyo in 1984,1986 and 1989, respectively. He is a professor of National Institute of Informatics, Japan. His research interests are database systems and machine learning. He is a member of ACM, IEEE, IEICE, IPSJ and JSAI.

Daiji FUKAGAWA

Daiji FUKAGAWA received the B.E. degree in Computer Science in 2001, M.S. degree in Informatics in 2003, and Ph.D. degree in Informatics in 2006, all from Kyoto University. Since 2006, he has been working for National Institute of Informatics as a Project Researcher. His research interests include the theory of combinatorial optimization for trees and graphs. He is a member of Information Processing Society of Japan, the Institute of Electronics, Information and Communication Engineers, and the Database Society of Japan.

Masao TAKAKU

Masao TAKAKU is a senior engineer of National Institute for Materials Science. He was working for the large-scale information linkage project as an post-doctoral researcher at National Institute of Informatics (NII) until August 2008. He received B.S. and M.S. degrees from University of Library and Information Science in 1998 and 2000 respectively, and received Ph.D. in Information Science from University of Tsukuba in 2004. His current interests include digital library, information retrieval and information seeking behavior.

Jun ADACHI

Jun ADACHI is Professor in the Digital Content and Media Sciences Research Division, National Institute of Informatics (NII), Japan. He is also the Director of the Cyber Sceince Infrastructure Development Department of NII. His professional career has largely been spent in research and development of scholaly information systems, such as NACSIS-CAT and NII-ELS. He is also an adjunct professor of the Graduate School of Information Science and Technology, University of Tokyo. His research interests are information retrieval, text mining, digital library systems, and distributed information systems. Adachi received his BE, ME and Doctor of Engineering in Electrical Engineering from the University of Tokyo in 1976, 1978, and 1981, respectively. He is a member of IPSJ, IEEE, and ACM.