Review of Author Name Disambiguation Techniques for Citation Analysis

인용분석에서의 모호한 저자명 식별을 위한 방법들에 관한 고찰

Hyunjung Kim(김현정)*

ABSTRACT

In citation analysis, author names are often used as the unit of analysis and some authors are indexed under the same name in bibliographic databases where the citation counts are obtained from. There are many techniques for author name disambiguation, using supervised, unsupervised, or semisupervised learning algorithms. Unsupervised approach uses machine learning algorithms to extract necessary bibliographic information from large-scale databases and digital libraries, while supervised approaches use manually built training datasets for clustering author groups for combining them with learning algorithms for author name disambiguation. The study examines various techniques for author name disambiguation in the hope for finding an aid to improve the precision of citation counts in citation analysis, as well as for better results in information retrieval.

초 록

서지 데이터베이스를 이용한 인용분석연구를 진행하기 이전에 이루어져야 할 과정 중 하나가 모호한 저자명의 식별이라고 할 수 있다. 대부분 서지 데이터베이스에는 저자의 성(姓)과 이름의 이니셜만을 표기하는 경우가 많은데, 중국이나 한국 등 아시아 국가 출신의 연구자들은 같은 성을 가진 사람이 매우 많고, 이름의 이니셜까지 같은 경우도 상당히 많아서 이름검색만으로 찾고자 하는 저자를 식별해내기가 쉽지 않기 때문이다. 아시아 국가 출신의 학자들이 유난히 많은 연구분야들에서는 이러한 문제들이 더더욱 큰 문제가 되며, 인용분석 뿐만 아니라 일반적인 정보검색에서도 매우 중요한 요인이 될 수 있다. 모호한 저자명을 식별해내는 방법에는 자동화된 알고리듬을 이용하여 각각의 저자를 식별해내는 방법과 저자 클러스터링을 얻어내기 위해 일일이 수작업으로 데이터셋을 구축하는 방법, 그리고 두 가지 방법을 혼용한 반자동화된 방법 등이 있다. 본 연구는 "모호한 저자명 식별"을 위해 개발된 여러 가지 방법들을 고찰해보기로 한다.

키워드: Author Name Disambiguation, Citation Analysis, Information Retrieval, Algorithm, Homonymous Names
모호한 저자명 식별, 인용분석, 정보검색, 알고리듬, 동명이인

* 서울여자대학교 사회과학대학 문헌정보학과 초빙강의교수(hk48@swu.ac.kr)
논문접수일자 : 2012년 9월 18일 논문심사일자 : 2012년 9월 19일 게재확정일자 : 2012년 9월 24일

www.kci.go.kr

1. Introduction

Many bibliographic studies, including citation analysis, use author names as the unit of analysis. They identify and collect bibliographic information of the works published by an author by using databases or digital libraries, such as Web of Science and CiteSeerx. For example, Web of Science offers a function of "author finder" to find a specific author under investigation using subject areas or Web of Science categories as an aid to identify the author. However, common names like John Smith or foreign names like J Wang are very difficult to isolate the works by the intended author from works by other authors with the same name. Smalheiser and Torvik (2009, 287) addressed four distinct challenges of author name disambiguation: (1) a single individual may publish under multiple names, (2) many different individuals have the same name, (3) the necessary metadata are often incomplete or lacking entirely, and (4) an increasing percentage of scholarly articles are not only multi-authored, but represent multi-disciplinary and multi-institutional efforts. White and Griffith (1981) suggested Author Cocitation Analysis (ACA) as a technique to minimize such errors of including works by authors other than those who were intended to be analyzed. White and McCain (1998) emphasized the effectiveness of ACA that cocitations between multiple names would be able to pair two particular authors that were meant to be paired, and the analysis would result in a map with oeuvres of authors for visualizing a discipline. However, author name ambiguity still remains as problematic in bibliometrics as well as information retrieval for both authors with the same name and authors published and/or indexed under multiple names.

Strotmann and Zhao (2012, 1821) indicated "the increasing globalization of research" as one of the causes that have affected this author name ambiguity since Asian names tend to have the same last name with the first initial for many individuals. For example, a search for "Kim, H" will result in many works published by authors with the same last name and the first initial, while "Kim" is the last name of about 20% of the whole Korean population.[1] In the study of author cocitation analysis on the editorial board of Journal of Communication (Kim 2008) using cocitation counts obtained from Web of Science database, such ambiguity caused delay in data collection and might have affected the result even though cocitation mapping reduced the errors of including works by authors indexed as the same name. The study started with Author Finderin Web of Science to find the indexed name of a particular author. In general, a search for an author results in citations in various disciplines. Restrictions on subject fields or Web of Science categories would help reduce the number of documents allegedly written by the intended author. However, using the Author Finder is not enough for identifying the particular author in many cases, especially with common names and

1) The year 2000 data shows 9,925,949 "Kim"s among the population of 45,985,289. 〈http://kostat.go.kr/portal/korea/index.action〉.

www.kci.go.kr

Asian names.

In addition, as Han et al. (2004, 296) suggested, ambiguous names "can decrease the performance of information retrieval and web search." When a user of a bibliographic database or digital library needs to find all references written by a particular author, he or she wants to be able to identify whether the name of the author has any variations or any other author shares the same name. Only the accurate representation of the author's indexed name would result in the proper set of search result. In an attempt to provide users with such accurate information, digital libraries and bibliographic databases would need to make an effort to find a way to disambiguate such cases. Researches on methods for disambiguating author names have been done in the field of computer science, mostly, and information science.

2. Review of Author Name Disambiguation Techniques

There are various techniques for author name disambiguation. A group of techniques use machine learning algorithms (Veloso et al. 2012; D'Angelo et al. 2011; Cota et al. 2010; Treeratpituk and Giles 2009; Kang 2008). Levin et al. (2012, 1031) summarized the techniques using machine learning algorithms into three classes; (1) supervised machine learning classifiers, which require large scale training sets, (2) unsupervised learning algorithms, which apply "an unsupervised clustering" based on similarities between pairs of articles, and (3) semi-supervised and weakly supervised methods, which use additional hand-labeled training set to train the large unlabeled dataset. Supervised machine learning algorithm may not be effective for large scale databases because hand-labeled training sets are not available for such cases. Unsupervised approaches use similarity-based algorithms to form clusters of documents that are linked to a single author instead of using manually annotated datasets. Semisupervised methods apply supervised training datasets to unsupervised or unlabeled large scale datasets to form author clusters. Author groups can be clustered by using co-authorship data (Ferreira et al. 2010; Dai and Storkey 2009; Kang et al. 2009b; Masada et al. 2007), ontology-based method using properties (Kim et al. 2011; Kim and Park 2009), and author profiling (Ferreira et al. 2012b).

Ferreira et al. (2012a, 18-19) also investigated the taxonomy of author name disambiguation methods, focusing on automatic techniques. According to the type of approach, they classified (1) author grouping methods and (2) author assignment methods. For author grouping methods, similarity functions are used to determine the similarity between pairs or groups of references for each individual author, either using predefined functions or learning a similarity function. After defining a similarity function, the author grouping methods use author clustering techniques to form clusters of authors with ambiguous names. For author assignment methods, classification and clustering are main techniques used for author name disambiguation and they don't need training datasets because they "directly assign each reference

www.kci.go.kr

to a given author by constructing a model that represents the author using either a supervised classification technique or a model-based clustering technique." Based on the type of evidence for author name disambiguation, they indicated that citation information and web information are main sources for extracting bibliographic information in most cases, but sometimes implicit evidence such as "the latent topics of a citation" can also be used to calculate the similarity between references.

2.1 Author disambiguation using unsupervised algorithm

One of the fully automated procedures for author name disambiguation, Velosoet al. (2012) proposed a method called associative name disambiguation under the assumption of the existence of strong association between bibliographic features and specific authors. For example, names of coauthors as well as titles of articles and journals in which their works are published are good candidates for training data used by the associative name disambiguation approach. Their proposed method extracts bibliographic data from large-scale databases and used learning algorithms to identify individual authors in a cost-effective approach. These types of author name disambiguation method can be useful for large-scale databases and digital libraries, but may not be very useful for authors who are not so high prolific and not included in many citations.Treeratpituk and Giles (2009) also suggested an automatic machine learning classification algorithm called "Random Forests", which is an ensemble combining a collection of decision trees. Just as building a forest, they build a decision tree out of different bootstrap samples, then the collection of decision trees can be used to construct a forest.

2.2 Author disambiguation using semisupervised algorithm

Yang et al. (2008) used topic and web correlations for citation analysis of authors. Topic correlation refers to a method of measuring similarities between topics of two citations to see if the two citations with the same author name are actually related to the specific author. Web correlation measures the number of co-occurrences using pair-wise grouping algorithm and similarity metrics as well as binary classifier and cluster filter for accuracy. Topic-based name disambiguation was also used by Song et al. (2007), employing unsupervised machine learning algorithm. Their models include one for Probabilistic Latent Semantic Analysis (PLSA) for document-name-word association, and another for Latent Dirichlet Allocation (LDA) for topic-document-word-name association. Yet, they found out using topic-name matrices created by both algorithms were not enough for name disambiguation because it is possible that some homonymous authors belong to the same topic group. Therefore, topic-based learning method should be accompanied by further hierarchical clustering of authors.

Peng et al. (2012, 10523) used web correlations to improve traditional author name disambiguation

www.kci.go.kr

techniques as well as authorship correlations. They generated pairs of citations using web and authorship correlations, created two datasets of training data and testing data, applied the binary classifier to label the matching pairs, then grouped the citations into clusters using the labels. The web correlation refers to the procedure of searching for citation titles, recording URLs for each citation, making publication list based on the collected URLs, and compiling titles with the same author names on the same publication list as a database of works by an individual author. They also employed authorship correlation with name popularity measure and citation grouping with binary classifier to increase the effectiveness of matching the citations for creating author clusters. Their experiments were evaluated with "similarity metric analysis, performance evaluation using combinations of different similarity metrics, and baseline performance evaluation" (Peng et al. 2012, 10527) and their evaluation indicated that combining web and authorship correlation for author name disambiguation performed better than traditional techniques.

Ferreira et al. (2012b, 46-47) proposed a system called SyGAR (Synthetic Generator of Authorship Records) that generates citation records based on author profiles. The procedure consists of three main steps, (1) taking in real authorship records as an input, (2) inferring author publication profiles, generating synthetic citation records, and modifying citation attributes, and (3) producing synthetic authorship records as an output. Their proposed synthetic design uses "author publication profiles" to generate citation records using author names, work titles, and publication venue titles. Authors with ambiguous names will be analyzed and used for synthetic data generation. The process of generating the synthetic authorship records can be used to evaluate existing author name disambiguation methods.

Kang et al. (2008) used a large-scale test set for Korean in an analysis of the individual features'effect on author disambiguation. They manually identified authors for 8,675 articles published in 29 conference proceedings of 9 IT-related associations from 1999 to 2006. Their bibliographic metadata included 23,177 names of authors as well as other properties such as titles of their works, journal, publication year, their email addresses, and their affiliations. The manual identification process gives different identifiers for each author with the same name, then they create clusters of works published by each identified author. As a result, for a group of 5,332 authors with same names, they found 9,133 'real' individual authors. Their test showed using more than one properties would result in statistically significant difference than using only one property. Especially, co-authorship data and author's affiliation and email are appeared to be more effective than other properties.

Later, Kang et al. (2009a) built a large-scale test set for Korean author name disambiguation using manually identified author names. The set consists of 41,673 author name entity records for 881 author names, and it includes 6,921 real world author identifiers. Their manually constructed author identifier set includes author names for initial identification and their corresponding real world author identifiers, and it complements the unsupervised learning algo-

www.kci.go.kr

rithms for large scale databases. The set also includes basic bibliographic information such as title of the article, names of coauthors, year of publication, and the name of the journal. The main difference between their newly built system and existing datasets is the size, which includes 41,673 author entity records with 881 ambiguous author names and 6,921 identified individual authors. It means on average one ambiguous author name has to be identified among 12.7 authors with the same name. The test showed the full names are the best property when used alone, then name of co-authors, title of the article, year of publication, and name of the published journal are also found to be good identifying properties.

For self-supervised learning algorithm, Levin et al. (2011) presented a self-trained two-stage method, including (1) bootstrapping via rule-based clustering and (2) supervised classification and clustering. The first stage applies a small set of rules to identify works by a particular author, such as self-citation rules, subject area rules, and coauthorship rules. Then the second stage involves training of a classifier and using the classifier as a similarity metric. Citation features they used for classifiers include Cited Article IDs, Citing Article IDs, and Cited Journal Titles in Web of Knowledge database. These features are used for measuring similarity between works supposedly written by a particular author. Their proposed method can be useful for large scale databases like Web of Knowledge, as they performed an experiment on it. Self-supervised method shares advantages of both unsupervised and supervised approaches because it doesn't need manually labeled datasets but enables generating a large training set of features.

D'Angelo et al. (2010, 259-260) proposed a heuristic approach for author name disambiguation in bibliometric datasets. They denoted the shortcomings of supervised approaches as the need of expensive training datasets, which also requires proper maintenance if any changes occur. Unsupervised approaches also have some disadvantages of having "very high precision but low recall" because of the articles outside the author's citation network. Unlike those approaches, their proposed approach integrated structured data with the bibliometric database using a reference external source of information, following three steps of "database integration, mapping generation, and filtering." The process of integrating data with external information sources enables obtaining more information on authors which assist the author name disambiguation. Their filtering process employs (1) the address filter, (2) the WoS-SDS (Web of Science-Scientific Disciplinary Sectors) filter, (3) the shared SDS filter, and (4) the maximum correspondence filter. The filters are used to "analyze each cluster and select the correct pair based on an educated guess derived from the distribution of the previously disambiguated data" (D'Angelo et al. 2010, 264). Although their research focused on Italian author names, their heuristic approach is worth noticing in terms of providing morein depth information on authors. Cota et al. (2010) utilized heuristic-based hierarchical clustering method for name disambiguation using features of citations, such as coauthor names, title of the work, and publication venue title. They performed a series of experiments

www.kci.go.kr

using (1) the list of coauthor names in the bibliographic database, (2) the work title attribute in the bibliographic database, (3) the publication venue title attribute in the bibliographic database, and (4) all attributes together in the bibliographic database. Using the list of coauthor names, the first three approaches produced very fragmented clusters, but using the work title attributes produced the purest clusters. Using all attributes together produced very pure and less fragmented clusters than other cases. Their research showed the value of bibliographic data in author name disambiguation methods; especially those did not require training datasets.

Han et al. (2004, 297) suggested two supervised learning approaches for author name disambiguation, including one based on a generative statistical model called "the naïve Bayes approach" and the other based on a discriminant model, which acts as a classifier, called "the SVM (Support Vector Machine) approach." The naïve Bayes approach uses positive training citations for recognizing author's writing patterns, and assigns citations to a specific author based on probabilities. The SVM approach learns from both positive and negative training citations, and "considers each author as a class, and trains the classifier for each author class" (Han et al. 2004, 299). They also used three types of citation attributes including coauthor names, title of the article, and title of the journal, and coauthor names appear to be the strongest property for obtaining better accuracy of identifying a particular author.

2.3 Author disambiguation using supervised method

Onodera et al. (2009, 681-684) suggested a two-step filtering method for eliminating articles by authors with the same name from large scale document databases. The first stage eliminates articles based on affiliation addresses and citation relationships between the journal and the article under investigation. The second stage uses discrimination functions utilized by manual judgment on the retrieved set of articles. Their source articles were from 24 journals in six subject fields, including (1) condensed matter physics, (2) inorganic and nuclear chemistry, (3) electric and electronic engineering, (4) biochemistry and molecular biology, (5) physiology, and (6) gastroenterology. Then 60 source articles were sampled from each journal and eventually 2,595 source authors were selected for the author search. The author search was done with Web of Science database and retrieved 629,000 articles. They used (1) coauthors of source and retrieved articles, (2) affiliation addresses of sources and retrieved articles, (3) citation relationships between the journals of source and retrieved articles, (4) title words of source and retrieved articles, (5) citation of retrieved articles by source articles, (6) cocitation between source and retrieved articles, (7) interval between the years of publication of source and retrieved articles, and (8) whether the source author's affiliation country is the specified one or not as the information for discrimination. For filtering purposes, they employ three features including (1) common coauthor(s) between source and retrieved

www.kci.go.kr

articles, (2) similarity of affiliation addresses of source and retrieved articles, and (3) journal citation relationships. The first stage focuses on eliminating false articles, then the second stage focuses on finding further false articles by using “manual judgment of sample articles” and “modeling discrimination functions based on logistic regression” (p. 683). Their result shows very high precision rate because the first stage eliminates as many false articles as possible, even if it means some ‘true’ articles are also eliminated. However, their methodology of using citation data as discriminant factor appears to be more effective than other traditional simplified name disambiguation methods.

Qian et al. (2011) proposed a supervised author name disambiguation method combining machine learning and human judgment in, pointing out the shortcomings of the fully automated author name disambiguation methods. They indicated that the fully automated unsupervised techniques might not be able to identify a particular author out of ambiguous names in digital libraries because complete information on authors might not be available in many cases. As an alternative approach, they applied a framework called LOAD (Labeling Oriented Author Disambiguation), which consists of clustering methods with high precision and high recall as well as top dissimilar clusters selection and ranking. The combination includes a supervised learning algorithm that trains the similarity functions and a clustering algorithm that generates author clusters.

3. Author Name Disambiguation in relation to Author Cocitation Analysis

A recent study by Strotmann and Zhao (2012, 1823) explored if the choice of author name disambiguation method affected citation-based author ranking and mapping by author cocitation analysis. They compared two sets of author citation ranking and author citation analysis mapping result to see if the choice between sophisticated author name disambiguation and simplified approach of author name disambiguation that uses simply last-name-plus-first-initial combination for mapping author clusters makes any difference. Author cocitation analysis is known to immune to the issue of ambiguous author names because the process of mapping pairs of authors, only the right author will be co-occurred (White and McCain 1998). In reality, it is extremely difficult to manually verify if the search result includes references published by the particular author under investigation because in most cases author cocitation analysis deals with many authors and a matrix created for pairs of authors includes half of the number of authors squared. For example, an author cocitation analysis about 145 authors will require a matrix of more than 10,000 cells of cocitation counts. Even with a small number of authors, like 20, the cocitation matrix will include 200 cells with cocitation counts, which means each cell contains the number of cocited documents by the pair of authors. Most cells are zeros but some cells have more than 100, if both authors have many publications and their subject

www.kci.go.kr

relatedness is high. In such cases, it is almost impossible to see if individual documents are published by the right author. However, cocitation mapping of authors still offers overall visualization of a discipline when applied for a group of authors related to the field, even with certain errors of including author names that are not supposed to be included.

Strotmann and Zhao (2012, 1824) used author cocitation mapping for comparisons between traditional simplified approach and their sophisticated approach for author name disambiguation. They compare two sets, including a reference data set and a test data set. The reference data set uses a sophisticated automatic author name disambiguation method to obtain author clusters, and the test set uses the traditional simplified method for author name disambiguation. Then they compare the result for first-author-based ranking, all-author-based ranking, and last-author-based ranking as well as all author cocitation analysis mappings between the two approaches. One of the observations from citation rankings show:

Many Chinese and Korean names made it to the top 100 lists by first- and all-author citations but only one to the top 100 list last-author citation. Although it appears that Chinese and Korean research programs are quite significant and successful in the field of stem cell research, there are surprisingly few Chinese or Korean researchers among the established laboratory heads, which comprise the highly cited last-author set (Strotmann and Zhao 2012, 1826).

Then they compared author cocitation maps with the sophisticated author name disambiguation method and with the traditional simplified author name disambiguation method to see how well they matched. The cocitation mapping of authors selected by the simplified author name disambiguation method tends to have more Chinese and Korean names than the mapping of authors selected by the sophisticated author name disambiguation, and it shows more distortion than author mapping with the sophisticated author name disambiguation. Using the sophisticated author name disambiguation method, the cocitation map of authors appear to be properly represented because the disambiguation method indeed identifies the actual highly cited authors as the central actors in the map. However, their results for the simplified disambiguation method show it doesn't work well for author cocitation analysis for many ambiguous names, especially "the massively ambiguous Chinese and Korean author names" (Strotmann and Zhao 2012, 1828).

4. Unique Author Identifier: ResearcherID

In an effort to enable author name disambiguation even before entering the author's name into databases, a unique author identifier can be provided for every author. ResearcherID is one of the examples of a portal that enables authors to enter information about themselves and gives them unique IDs to be used with their citation data. The IDs can be integrated into Web of Science, which makes it easier for users to identify works by a particular author, not confusing

www.kci.go.kr

with any other who share the same name. Authors can update their profiles when any changes occur. However, the service is currently restricted to the invited authors and the identifier information appears to be only used by Web of Science and EndNote. In addition, some services are restricted to subscribers of either Web of Knowledge or Web of Science, even though basic searches are available for non-subscribers. If these types of services are available for many other bibliographic databases and digital libraries and should every author be given a unique identifier, the issue of ambiguous names will no longer cause any problems in bibliographic data collection for citation analysis and in information retrieval based on author names.

5. Discussion

The author name disambiguation techniques found in the literature are mostly using machine learning algorithms to train systems to identify individual authors with ambiguous names in large scale databases and digital libraries. Indeed it seems reasonable to explore such techniques and try to improve the accuracy and speed of finding references by a particular author, but since most of them are using automated procedure of extracting bibliographic information about documents, it may not result in a complete dataset for training the algorithm and some of the necessary information can be left out. On the other hand, manual disambiguation can offer much more precise result, but it is very expensive and "a surprisingly hard and uncertain process, even on a small scale" (Smalheiser and Torvik 2009, 293). In addition, since it takes a long time to build such systems manually, it may not be able to keep up the bibliographic data of the newly acquired materials. So it seems the best way to solve the shortcomings of the two techniques can be combining some strong points of the two, such as using heuristic approach or semisupervised methods.

However, it still leaves some question on effectiveness and efficiency of such systems' developments. Are these techniques going to really function as they are supposed to? Is it worth to invest so much for a system that may or may not be so useful for the users? Also, it is difficult to imagine how users can utilize these functions when they need to find references published by authors with ambiguous names. Are these large scale databases and digital libraries currently using or going to use these algorithms to make the search more precise? How are they going to let the users know about this feature and utilize it when they need to?

Strotmann and Zhao (2012) pointed out that Romanized Asian names like Chinese and Korean names were extremely ambiguous but most studies, including those done by Korean scholars, were dealing with author name disambiguation methods in general, not in specific for Asian names. Those designs suggested by Korean scholars may have had training sets designed specifically for Korean names, but it still looks very similar to other author name disambiguation techniques developed for English names except Kang et al.'s 2009 research (2009b)

www.kci.go.kr

on an author disambiguation method using co-authorship data. More studies should be done for Korean names, whether they are in Korean or in English.

6. Conclusion

At first, the study was carried out to find a way to achieve a better precision related to ambiguous author names in both citation analysis and information retrieval. However, current literature of the topic shows those techniques for author name disambiguation are mostly studied in the field of computer science, utilizing machine learning algorithms whether it is fully automatic procedure or involves manually constructed training datasets. Hopefully, these studies would help build a system that can actually be effective in improving the precision of finding references published by a particular author. Other than these built-in methods of identifying ambiguous author names, a different approach may be more effective. For example, Thompson Reuters' Researcher ID (http://www.researcherid.com) can be very helpful to identify individual authors by giving them unique identifiers even before they are indexed into bibliographic databases like Web of Science. If there was a unique identifier for each author in bibliographic databases and digital libraries, collecting bibliographic data for analysis or searching for works by a particular author could be more effective and the process of eliminating false references could be avoided.

References

강인수. 2008. 한글 저자명 중의성 해소를 위한 기계학습 기법의 적용. 『정보관리학회지』, 25(3): 27-39.

강인수, 김평, 이승우, 정한민, 류범종. 2009a. 저자 식별을 위한 대용량 평가셋 구축. 『한국콘텐츠학회논문지』, 9(11): 455-464.

강인수, 이승우, 정한민, 김평, 구희관, 이미경, 성원경, 박동인. 2008. 저자 식별을 위한 자질 비교. 『한국콘텐츠학회논문지』, 8(2): 41-47.

김제민, 박영택. 2009. 저자명 모호성 해결을 위한 개념망 기반 카테고리 유틸리티. 『정보처리학회논문지』, 16B(3): 225-232.

김태홍, 정한민, 성원경, 김평. 2011. 대표속성을 이용한 저자개체 식별. 『한국콘텐츠학회논문지』, 12(1): 17-29.

Cota, Ricardo G., Anderson A. Ferreira, Christiano Nascimento, Marcos Andrew Goncalves, and Alberto H.F. Laender. 2010. "An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations." *Journal of the American Society for Information Science and Technology,* 61(9): 1853-1870.

Dai, Andrew M. and Amos J. Storkey. 2009. "Author disambiguation: A nonparametric topic and

www.kci.go.kr

co-authorship model." *NIPS Workshop on Applications for Topic Models: Text and Beyond,* December 11, 2009, Whistler, Canada.

D'Angelo, Ciriaco Andrea, Christiano Giuffrida, and Giovanni Abramo. 2010. "A heuristic approach to author name disambiguation in bibliometrics databases for large-scale research assessments." *Journal of the American Society for Information Science and Technology,* 62(2): 257-269.

Ferreira, Anderson A., Marcos Andre Goncalves, and Alberto H.F. Laender. 2012a. "A brief survey of automatic methods for author name disambiguation." *SIGMOD Record,* 41(2): 15-26.

Ferreira, Anderson A., Marcos Andre Goncalves, Jussara M. Almeida, Alberto H.F. Laender, and Adriano Veloso. 2012b. "A tool for generating synthetic authorship records for evaluating author name disambiguation methods." *Information Sciences,* 206: 42-62.

Ferreira, Anderson A., Adriano Veloso, Marcos Andre Goncalves, and Alberto H.F. Laender. 2010. Effective self-training author name disambiguation in scholarly digital libraries. In *Proceedings of the 2011 Joint International Conference on Digital Libraries (JCDL '10).* New York: ACM Press.

Han, Hui, Lee Giles, Hongyuan Zha, Cheng Li, and Kostas Tsioutsiouliklis. 2004. Two supervised learning approaches for name disambiguation in author citations, In *Proceedings of the 2004 Joint International Conference on Digital Libraries (JCDL '04),* June 7-11, Tucson, AZ, USA.

Kang, In-su, Seung-Hoon Na, Seungwoo Lee, Hanmin Jung, Pyung Kim, Won-Kyung Sung, and Jong-Hyeok Lee. 2009b. "On co-authorship for author disambiguation." *Information Processing and Management*, 45: 84-97.

Kim, Hyunjung. 2008. Author cocitation analysis using social network analysis. In *AMCIS 2008 Proceedings,* Americas Conference on Information Systems, August 2008, Toronto, Canada.

Levin, Michael, Stefan Krawczyk, Steven Bethard, and Dan Jurafsky. 2012. "Citation-based bootstrapping for large-scale author disambiguation." *Journal of the American Society for Information Science and Technology*, 63(5): 1030-1047.

Masada, Tomonari, Atsuhiro Takasu, and Jun Adachi. 2007. Citation data clustering for author name disambiguation. *The Second International Conference on Scalable Information Systems (INFOSCALE '07),* June 6-8, Suzhou, China.

Onodera, Natsuo, Mariko Iwasawa, Nobuyuki Midorikawa, Fuyuki Yoshikane, Kou Amano, Yutaka Ootani, Tadashi Kodama, Hiroyuki Tsunoda, and Shizuka Yamazaki. 2011. "A method for eliminating articles by homonymous authors from the large number of articles retrieved by author search." *Journal of the American Society for Information Science and Technology,* 62(4): 677-690.

Qian,Yanan, Yunhua Hu, Jianling Cui, Qinghua Zheng, and Zaiqing Nie. 2011. Combining machine learning and human judgment in author disambiguation, In *Proceedings of the*

www.kci.go.kr

20th ACM international conference on information and knowledge management (CIKM '11), 1241-1246.

Smalheiser, Neil R. and Vetle I. Torvik. 2009. "Author name disambiguation." *Annual Review of Information Science and Technology,* 43: 287-313.

Song, Yang, Ian Huang, Isaac G. Councill, Jia Li, and C. Lee Giles. 2007. Efficient topic-based unsupervised name disambiguation. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '07),* 342-351.

Strotmann, Andreas, and Danzhi Zhao. 2012. "Author name disambiguation: What difference does it make in author-based citation analysis?" *Journal of the American Society for Information Science and Technology,* 63(9): 1820-1833.

Treeratpituk, Pucktada and C. Lee Giles. 2009. Disambiguating authors in academic publications using random forests. In *Proceedings of the 2009 Joint International Conference on Digital Libraries (JCDL '09).* New York: ACM Press.

Veloso, Adriano, Anderson A. Ferreira, Marcos Andre Goncalves, Alberto H.F. Laender, and Wagner Meira Jr. 2011. "Cost-effective on-demand associative author name disambiguation." *Information Processing and Management,* 48: 680-697.

White, Howard D. and Belver C. Griffith. 1981. "Author cocitation: A literature measure of intellectual structure." *Journal of the American Society for Information Science,* 32(3): 163-171.

White, Howard D. and Katherine W. McCain. 1998. "Visualizing a discipline: An author co-citation analysis of information science, 1972-1995." *Journal of the American Society for Information Science,* 49(4): 327-355.

Yang, Kai-Hsiang, Hsin-Tsung Peng, Jian-Yi Jiang, Hahn-Ming Lee, and Jan-Ming Ho. 2008. Author name disambiguation for citations using topic and web correlation. In *Proceedings of the European conference on research and advanced technology for digital libraries,* 14-19.

www.kci.go.kr

www.kci.go.kr