Information Processing and Management 45 (2009) 84–97

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

On co-authorship for author disambiguation

In-Su Kang [a], Seung-Hoon Na [c], Seungwoo Lee [b], Hanmin Jung [b], Pyung Kim [b], Won-Kyung Sung [b,*], Jong-Hyeok Lee [c]

[a] School of Computer Information, Kyungsung University, 314-79, Daeyeon-dong, Nam-gu, Busan 608-736, South Korea
[b] Korea Institute of Science and Technology Information (KISTI), Daejeon, South Korea
[c] Pohang University of Science and Technology (POSTECH), Pohang, South Korea

ARTICLE INFO

Article history:
Received 24 March 2008
Received in revised form 21 May 2008
Accepted 12 June 2008
Available online 20 August 2008

Keywords:
Author name disambiguation
Co-authorship
Author clustering

ABSTRACT

Author name disambiguation deals with clustering the same-name authors into different individuals. To attack the problem, many studies have employed a variety of disambiguation features such as coauthors, titles of papers/publications, topics of articles, emails/affiliations, etc. Among these, co-authorship is the most easily accessible and influential, since inter-person acquaintances represented by co-authorship could discriminate the identities of authors more clearly than other features. This study attempts to explore the net effects of co-authorship on author clustering in bibliographic data. First, to handle the shortage of explicit coauthors listed in known citations, a web-assisted technique of acquiring implicit coauthors of the target author to be disambiguated is proposed. Then, a coauthor disambiguation hypothesis that the identity of an author can be determined by his/her coauthors is examined and confirmed through a variety of author disambiguation experiments.

1. Introduction

As a tremendous amount of person-related information is available on the web or other electronic media, people want to find persons of interest frequently using the names as queries (Guha & Garg, 2004). However, there exists many-to-many mapping relationships between persons and their names. A person may have multiple names, and different persons may share the same name. For example, a person named *David S. Johnson* can write his name as *David S. Johnson*, *David Johnson*, *D. S. Johnson*, or *D. Johnson*, etc., and there may exist two or more different persons named *David S. Johnson*. When a single person is viewed as an object having a unique meaning, different names signifying the same person can be considered as synonyms, and the same names indicating different persons homonyms.[1]

The above many-to-many mapping between persons and their names may severely deteriorate the effectiveness of the person search. When a person name is submitted to a web people search system (Wan, Gao, Li, & Ding, 2005) that finds persons in web pages, the web pages that have synonymous names of the input query may not be retrieved, lowering the recall of the person search, and the retrieved pages may contain numerous homonymous names of the input query, decreasing the precision of the search system. Moreover, synonymous and/or homonymous person names may hinder the correct representation of person-related social networks such as coauthor and citation networks.

Thus, identifying synonymous person names and resolving homonymous ones are crucial to any person-related search or representation systems. Thus far, much research has been conducted to attack the identification of synonymous person names under the terms of record linkage, merge/purge, duplicate detection and elimination (Elmagarmid, Ipeirotis, & Verykios, 2007;

* Corresponding author. Tel.: +82 42 828 5011; fax: +82 42 828 5198.
E-mail address: wksung@kisti.re.kr (W.-K. Sung).

[1] This usage follows Reuther (2006).

0306-4573/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2008.06.006

Gu, Baxter, Vickers, & Rainsford, 2003; Winkler, 2006). Matching of synonymous names employs similarity measures which mainly rely on name-internal features such as overlapping characters or tokens (first, middle, last name) of the two names to be compared. On the other hand, disambiguation of homonymous names has relatively recently received a growing attention with the advent of the semantic web and social networks. It is highly dependent on name-external features: domain-independent biographic features (Guha & Garg, 2004; Mann & Yarowsky, 2003; Wan et al., 2005) such as birth data/place and e-mail/postal addresses, and/or domain-specific contextual features such as coauthors and paper titles in case of bibliographic data.

This study addresses the resolution of homonymous author names appearing in citation data. As disambiguation features, previous works have employed coauthors, titles of articles, titles of publications, and years of publications that constitute basic citation data. Titles of articles may epitomize research areas of their authors, thus under the assumption that namesakes do not heavily share their research areas, title similarity between articles may be employed in resolving homonymous author names. The same also applies to the case of titles of books. In addition to the above citation-internal features, some have utilized citation-external features such as abstracts, self-citations, and citation URLs. When the full text of articles is available, additional features such as e-mail addresses, affiliation, and keywords can be extracted and applied to the author name disambiguation.

We believe that of the above features, co-authorship is the most reliable and decisive from the viewpoint of discriminating the identities of authors, since it implies real-world acquaintances among authors. For instance, when one citation contains *D. S. Johnson* and *C. J. Date* as its authors, we can say that a *D. S. Johnson* in the citation indicates *the D. S. Johnson whom C. J. Date knows*. When another citation has also *D. S. Johnson* and *C. J. Date*, we can further state that "*D. S. Johnson*"s in the two citations are the same individual, namely, *the D. S. Johnson whom C. J. Date knows*, under the assumption that "*C. J. Date*"s in the two citations are not different persons. Other citation features are not as much person-related as co-authorship.

This study concentrates on investigating the sole effect of co-authorship information on the resolution of homonymous author names in bibliographic data. In particular, we start with a hypothesis stating that the identity of an author is characterized by his/her coauthors. Next, a web-based technique of gathering coauthors is proposed to supplement implicit coauthors not found in known citations. Then, our hypothesis is investigated using large-scale test data from the following viewpoints:

- Is it helpful to enrich coauthors with web-assisted collaborators in disambiguating namesakes?
- Is a person identified more accurately by more coauthors?
- How probable is it that co-authorship succeeds in disambiguating authors?
- Is it beneficial to use coauthors of coauthors in resolving authors?

The scope of co-authorship that this study handles is limited to academic co-authorship. In addition, the evaluation of co-authorship is performed exclusively using a test set in Korean. The remainder of this article is organized as follows. Section 2 summarizes related works. Section 3 describes the usage of co-authorship for author name disambiguation. Next, a web-assisted method of expanding coauthors is proposed in Section 4. Then, evaluations and concluding remarks are given in Sections 4 and 5.

2. Related works

2.1. Overview

Research on the processing of person names can be divided into personal name matching and personal name disambiguation. The former handles the identification of synonymous person names, while the latter involves the discrimination of homonymous ones. One may argue that person name disambiguation inherently includes the problem of personal name matching, since there may exist many namesakes who have a variety of name variants. This study, however, assumes that personal name matching precedes personal name disambiguation so that we can separate the two problems.

Synonymous names are normally found by computing similarities between two candidate person names, using mainly internal features from the person name string itself such as commonness/rareness of name tokens (first/middle/last name), overlapping ratio of name characters, and pronunciation affinity. Similarity measures employed for synonymous name matching include edit-distance, soundex, Jaro (1989), cosine, etc. Given a huge number of person names, the time of pairwise comparisons required to find synonymous names could be costly. To reduce the time, many have proposed a variety of blocking methods which limit the set of candidate names to be compared. Readers may refer to many other literatures (Elmagarmid et al., 2007; Gu et al., 2003; Winkler, 2006) that provide an excellent survey of synonymous name matching.

Unlike synonymous name matching, homonymous name resolution relies mainly on name-external features. Web people disambiguation (Guha & Garg, 2004; Mann & Yarowsky, 2003; Wan et al., 2005), which groups person names in web pages into real individuals, depends largely on domain-independent biographic features such as birth date/place, e-mail/postal address, affiliations, job title, phone numbers as well as collocational term features, and neighboring person names in web pages. Author name disambiguation, which is the focus of this study, resolves authors appearing in the domain of bibliography, and primarily utilizes domain-specific features such as co-authorship, titles of articles, titles of publication, year of publication, topics of articles, etc.

2.2. Author name disambiguation

Author disambiguation divides the namesakes appearing in citation data into their real individuals. This problem is normally attacked by clustering approaches based on pair-wise author similarities. Pair-wise author similarities, the degree to which two author instances appearing on two different articles refer to the same person, rely commonly on coauthors, paper titles, and publication titles, on the assumption[2] that during a certain period of time, coauthors do not tend to change radically, and researchers normally attack the same or similar research areas, and submit papers to conferences or journals related to their research areas. In addition, other features proposed for pair-wise author similarities include first/middle names (Culotta, Kanani, Hall, Wick, & McCallum, 2007; Kanani & McCallum, 2007; Kanani, McCallum, & Pal, 2007), emails/affiliations (Culotta et al., 2007; Huang, Ertekin, & Giles, 2006; Kanani & McCallum, 2007; Kanani et al., 2007), author commonality/rarity (Culotta et al., 2007; Han, Xu, Zha, & Giles, 2005; Han, Zha, & Giles, 2005; Kanani & McCallum, 2007; Kanani et al., 2007), the number of overlapping coauthors (Culotta et al., 2007; Kanani & McCallum, 2007; Kanani et al., 2007; Torvik, Weeber, Swanson, & Smalheiser, 2005), topics (Song, Huang, Councill, Li, & Giles, 2007), self-citations (McRae-Spencer & Shadbolt, 2006), project membership (Alani, Dasmahapatra, O'Hara, & Shadbolt, 2003) and Web evidences (Aswani, Bontcheva, & Cunningham, 2006; Kanani & McCallum, 2007; Kanani et al., 2007; McRae-Spencer & Shadbolt, 2006; Tan, Kan, & Lee, 2006; Yang, Jiang, Lee, & Ho, 2006).

When a typical citation record is viewed as being comprised of authors, a paper title, and publication information (e.g., title of journals or proceedings, year of publication, volume/issue, pages, etc.), the above features can be categorized into citation-internal and citation-external ones. All internal features may more or less contribute to resolving author identities. From the viewpoint that the identity of a person can be strongly determined by his/her social relationships, however, the most influential and reliable is the coauthor feature. Some earlier empirical evaluations (Han, Giles, Zha, Li, & Tsioutsiouliklis, 2004; Torvik et al., 2005; Yang et al., 2006) have also observed its good disambiguation capability.

Among citation-external features, emails/affiliations and self-citations have a high potential for author disambiguation. Email addresses may play the role of personal identifiers. Affiliations may provide biographic physical locations to which individuals belong. The use of self-citations is supported by the assumption that researchers tend to cite their own earlier work (McRae-Spencer & Shadbolt, 2006). These features, however, are not readily available since they should be extracted from full-text papers. Project membership information is also difficult to obtain although it supplies another type of researcher's social context. Thus, of the external features, the most accessible and reliable ones are Web evidences. The usage of Web is to check out if a pair of citation records is authored by the same individual either by searching the Web for personal publication pages that contain both of the two citation records (Aswani et al., 2006; Kanani & McCallum, 2007; Kanani et al., 2007), or by comparing two URLs (Uniform Resource Locators) that link to web pages containing each citation record (McRae-Spencer & Shadbolt, 2006; Tan et al., 2006; Yang et al., 2006).

Pair-wise author similarities may be calculated by either supervised or unsupervised methods using the above features. From the training data, supervised methods learn the binary classifier that determines whether two author appearances are co-referent or not, using machine learning techniques such as SVM (Huang et al., 2006; Yang et al., 2006), maximum entropy (Kanani & McCallum, 2007; Kanani et al., 2007), and error-driven training (Culotta et al., 2007). Then, classification results or confidence scores yielded by the classifiers are used as pair-wise author similarities for later clustering. Unsupervised approaches normally define a learning-free similarity function between two authors. Some of them (Aswani et al., 2006; Lee, On, Kang, & Park, 2005) represented the similarity function as a linear combination of weighted feature-similarity values. Torvik et al. (2005) instead presented a model for estimating the probability that a pair of author names refer to the same individual, from automatically generated reference sets consisting of pairs of articles authored by the same versus different persons. Whereas supervised methods would have difficulties in manually compiling training data in terms of its quantity and balance, unsupervised ones need to develop a method to integrate each feature similarity.

After obtaining pair-wise author similarities, the author disambiguation module clusters authors by viewing author similarities as edge weights between author nodes. Previously employed clustering techniques mainly include agglomerative clustering (Culotta et al., 2007; Song et al., 2007; Tan et al., 2006), stochastic graph partitioning (Kanani & McCallum, 2007; Kanani et al., 2007), and DBSCAN method (Huang et al., 2006).

Instead of pair-wise author similarities, one group of researchers (Han et al., 2003, 2004; Han, Xu et al., 2005; Han, Zha et al., 2005) had attempted to create author-specific models by either supervised or unsupervised techniques. They constructed one author-specific model for each individual, assuming that the number of real persons corresponding to namesakes is known in advance. Although their method can be applied and evaluated over a prepared test set since the number of different authors is known from the test set, it would be nontrivial to generalize such author-centric approach to the resolution of unknown real authors.

3. Co-authorship

In general, coauthors who are listed in the same paper know each other. This acquaintance among coauthors may provide clues on disambiguating homonymous author names. For example, suppose that we want to discriminate appearances of '*A. Cohen*' as authors in the five citations C1 through C5.

[2] This assumption was first introduced by Han, Giles, and Zha (2003), but not for the calculation of pair-wise author similarities.

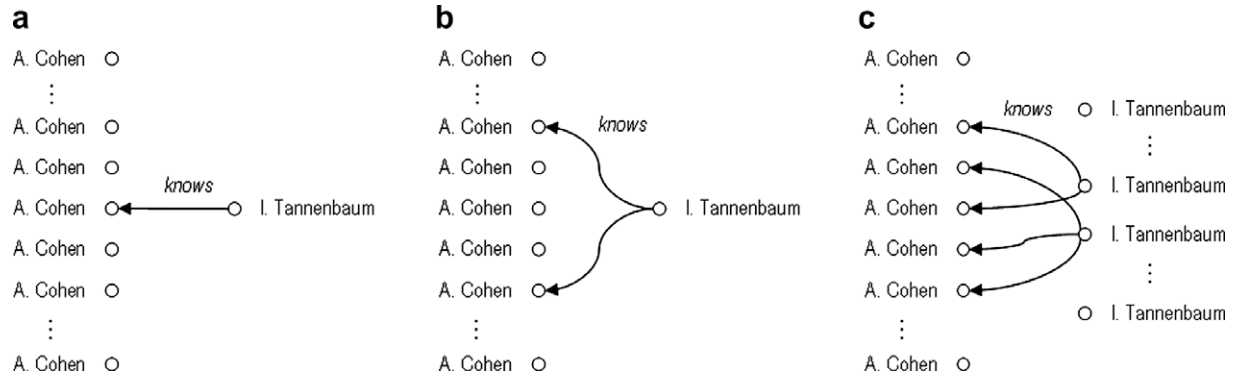

Fig. 1. Curse of common person names.

C1:	*A. Cohen*, S. Draper, E. Martinian, G. Wornell (2006). Stealing bits from a quantized ...
C2:	*A. Cohen*, S. Draper, E. Martinian, G. Wornell (2002). Source requantization: successive ...
C3:	*A. Cohen*, J. Siegel, P. Rozin (2003). Faith versus practice: different bases for ...
C4:	*A. Cohen*, A. Malka, P. Rozin, L. Cherfas (2006). Religion and unforgivable offenses ...
C5:	*A. Cohen*, I. Tannenbaum (2001). Lesbian and bisexual women's judgments of ...

Then, co-authorship of '*A. Cohen*' would suggest that (C1, C2), (C3, C4), and (C5) have been authored respectively by three different persons named '*A. Cohen*', from the fact that (C1, C2) and (C3, C4) respectively share the same coauthor.

The disambiguation capability among co-appearing authors may be related to the distributional hypothesis in linguistics stating that words occurring in the same contexts tend to have similar meanings (Harris, 1954). In the same vein, Firth (1957) stated that a word is known by the company it keeps. For example, when we see the authors '*A. Cohen, I. Tannenbaum*' in C5, '*A. Cohen*' is believed to be the person who '*I. Tannenbaum*' knows, of numerous persons named '*A. Cohen*', and vice versa. However, this resolution may be complicated as shown in Fig. 1. The best case (a in Fig. 1) is when the person named '*I. Tannenbaum*' is unique, and there is only one person named '*A. Cohen*' whom the '*I. Tannenbaum*' knows. In general (c in Fig. 1), however, there may be several persons named '*I. Tannenbaum*', and each person named '*I. Tannenbaum*' may know many different persons named '*A. Cohen*'. In that case, '*A. Cohen*' would not be known by its co-appearing author '*I. Tannenbaum*'. Fundamentally, the level of such ambiguity is affected by how common each of the two names is, since when the name '*I. Tannenbaum*' is rare, (c) would be simplified to (b)-like cases, and when the name '*A. Cohen*' is also rare, (c) would be further reduced to (a)-like cases. In other words, the commonality would weaken mutual discrimination power among co-mentioned authors, which we call the curse of common person names. Thus, some earlier works adopted commonality/rarity of author names (Culotta et al., 2007; Han, Xu et al., 2005; Han, Zha et al., 2005; Kanani and McCallum, 2007; Kanani et al., 2007) to resolve author names.

To attack the curse of common person names in resolving author names, on the other hand, the aforementioned distributional hypothesis in linguistics may be anew employed in the form of the number of overlapping coauthors (Culotta et al., 2007; Kanani & McCallum, 2007; Kanani et al., 2007; Torvik et al., 2005), under the assumption that the more coauthors in common two author instances have, the more probable the two authors refer to the same individual. For instance, in the above '*A. Cohen*' example, grouping (C1, C2) would be more accurate than grouping (C3, C4), since '*A. Cohen*'s in C1 and C2 share more acquaintances than those in C3 and C4.

Another problem in applying co-authorship to author name disambiguation is the scarcity of co-authors. For instance, although the '*A. Cohen*' who authored papers C3 and C4 is the same person as the one in C5, co-authorship information known from C1 through C5 does not give any clue. However, if we newly know that the '*I. Tannenbaum*' is one of collaborators of the '*A. Cohen*' in C3 and C4, then three 'A. Cohen's in C3, C4 and C5 probably belong to the same person. This broad interpretation on coauthors that include implicit collaborators not found in known citation records would help us to understand the influence of co-authorship on disambiguating namesakes.

On the other hand, the identity of an author may be determined by coauthors of his/her coauthors.

C6:	*A. Smith*, D. Bell, K. McKinley (2005). Topic-based representation of ...
C7:	G. Segev, *A. Smith*, H. Silverman (2006). Information loss analysis for ...
C8:	D. Bell, *J. Lafferty*, K. McKinley (2004). Automatic sub-categorization for ...
C9:	*J. Lafferty*, G. Segev, H. Silverman (2005). Word sense disambiguation for ...

Consider the above four artificial citations. It is difficult to view the authors named '*A. Smith*' in C6 and C7 as the same individual since they do not share coauthors. The fact that each of their (at least two or more) coauthors shares other coauthor '*J. Lafferty*' in C8 and C9, however, weakly supports that they ('*A. Smith*'s in C6 and C7) refer to the same individual. This casts us a question of whether coauthors of coauthors are useful to disambiguating authors or not.

To explore the distributional hypothesis of linguistics from the perspective of disambiguating author names, this study sets out the following coauthor disambiguation hypothesis.

Coauthor disambiguation hypothesis: The identity of an author is characterized by his/her coauthors.

In the above hypothesis, coauthors of an author are meant to include all collaborators whom the author has written articles in cooperation with, rather than only explicit coauthors listed in currently known citation records. It is then crucial to devise a strategy to acquire all real world collaborators for an author. For this, this study proposes a web-based method of gathering collaborators.

4. Web-based acquisition of coauthors

When we collect coauthors of a person named '*T. Mitchell*', we need to specify a particular '*T. Mitchell*' among numerous persons named '*T. Mitchell*'. For this, we use a known coauthor of the particular '*T. Mitchell*'. For instance, '*T. Mitchell*' in the following artificial citation C10 can be specified as the '*T. Mitchell*' known to '*R. Niculescu*' or '*R. Rao*' or '*K. Patrick*'.

C10: R. Niculescu, *T. Mitchell*, R. Rao, K. Patrick (2006). Bayesian network learning with parameter constraints. Journal of Machine Learning Research 7, pp.1357-1383.

In other words, '*T. Mitchell*' in the above citation can be represented as ⟨'*T. Mitchell*', '*R. Niculescu*'⟩ where the name of a person whose coauthors are to be gathered is followed by the name of a known coauthor. Such representation however still may not pinpoint the '*T. Mitchell*' above, due to the possibility of one or more persons named '*R. Niculescu*' knowing different '*T. Mitchell*'s. This study, however, assumes that the occurrence of such ambiguity is less probable.

Next, a pair of author names, for example, ⟨'*T. Mitchell*', '*R. Niculescu*'⟩, is converted into a pair of last names, for example, ⟨*Mitchell*, *Niculescu*⟩. Then, the pair of last names is submitted as a query to web search engines to retrieve documents that contain both author names. Within the highly ranked retrieved documents, we locate sequences of author names each of which matches original author names ⟨'*T. Mitchell*', '*R. Niculescu*'⟩. From each of such sequences, new author names are gleaned as coauthors of the original author-pair. The above procedure is generalized into Algorithm 1.

Loop step 4 of Algorithm 1 may require fine-tuning, since retrieved web pages may not be about publications, like alumni pages. For more accurate extraction, thus, advanced techniques would be needed to create a publication-page classifier or a citation-record extractor. This study does not explore such techniques, but uses a simple regular expression matching to discover sequences of person names.

Algorithm 1

Web-based acquisition of coauthors

Input:	
	a: the name of an author whose coauthors are to be gathered
	$C = \{c_1, \ldots, c_k\}$: a seed set of k known coauthors of a ($k \geqslant 1$)
Initialize:	
	$C_{new} = C$
Loop:	
1	$W_{new} = \emptyset$
2	For each $c_i \in C_{new}$
3	Search the Web for documents containing both last names of a and c_i
4	Extract a set W of coauthors of a and c_i from top-n retrieved documents
5	$W_{new} = W_{new} \cup (W - C)$
6	End For
7	Exit Loop if $W_{new} = \emptyset$
8	$C_{new} = W_{new}$
9	$C = C \cup W_{new}$
Output:	
	C: a set of expanded coauthors of a

There are two ways of expanding coauthors of '*T. Mitchell*' in the above citation through Algorithm 1. The first expansion method is to execute the algorithm using $\boldsymbol{a}$ = '*T. Mitchell*', and seed set $\boldsymbol{C}$ = {'*R. Niculescu*', '*R. Rao*', '*K. Patrick*'} as its input. The second is to iteratively execute the first expansion method for coauthors except for '*T. Mitchell*', and merge the expanded coauthors. In other words, the second way is to run the algorithm iteratively over each of ⟨$\boldsymbol{a}$ = '*R. Niculescu*', $\boldsymbol{C}$ = {'*R. Rao*', '*K. Patrick*'}⟩, ⟨$\boldsymbol{a}$ = '*R. Rao*', $\boldsymbol{C}$ = {'*R. Niculescu*', '*K. Patrick*'}⟩, and ⟨$\boldsymbol{a}$ = '*K. Patrick*', $\boldsymbol{C}$ = {'*R. Rao*', '*R. Niculescu*'}⟩. For convenience, coauthors augmented by the first and second expansion methods are respectively called *author-centric implicit coauthors* (*aIC*), and *coauthor-centric implicit coauthors* (*cIC*). clC represents coauthors of coauthors.

To specify more precisely the '*T. Mitchell*' in C10, two or more (or all) coauthors could be used as restrictions like '*T. Mitchell*' known to '*R. Niculescu*' and '*R. Rao*' (⟨'*T. Mitchell*', '*R. Niculescu*', '*R. Rao*'⟩), or '*T. Mitchell*' known to '*R. Niculescu*'

Table 1
Statistics of the test set

Number of papers	Number of same-name person groups	Number of real individuals	Number of author occurrences	Avg. number of author occurrences	Avg. number of author ambiguity
8675	5332	9133	23,177 of a total of 28,042	3.23 (=28,042/8675)	1.71 (=9133/5332)

and '*R. Rao*' and '*K. Patrick*' (⟨'*T. Mitchell*', '*R. Niculescu*', '*R. Rao*', '*K. Patrick*'⟩). More restrictions however may reduce the possibility of acquiring new coauthors not found in known citations.

5. Evaluation setup

5.1. Test set

As a test set to assess namesake resolution, we created a gold standard for a total of 8675 IT-related conference papers which were published in Korean during 1999–2006. Korean does not suffer from name-variant problems. In other words, a person's name is written in Korean in a single form, that is, a surname followed by a given name without delimiters or middle names. Owing to this feature, Korean author disambiguation system may skip matching of synonymous names. To manually discriminate the same-name authors into real individuals, two persons for three months have fastidiously confirmed the identities of authors by utilizing coauthors, e-mails, affiliations, project-related information from full-texts of papers as well as by looking up publication lists of author homepages on the Web.

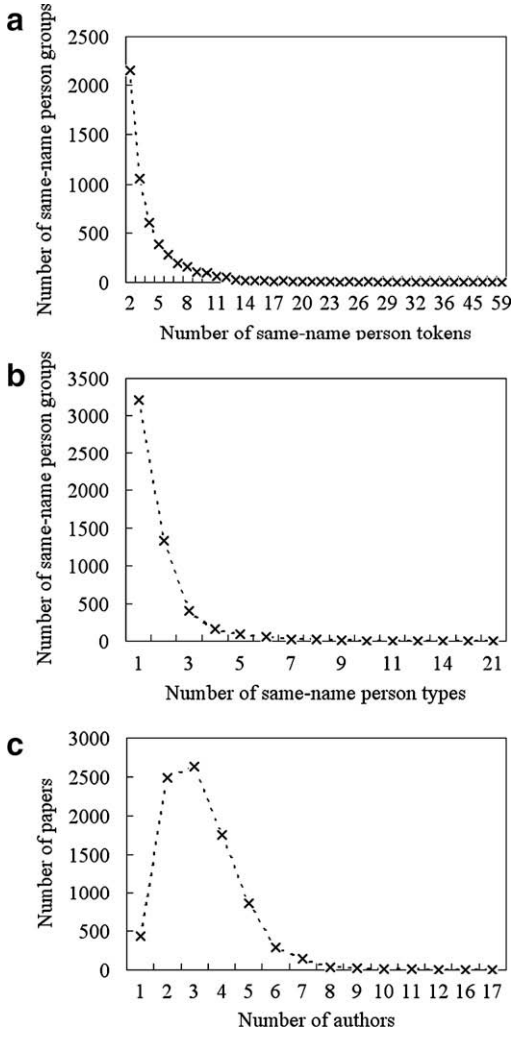

Fig. 2. Characteristics of the test set. (a) and (b) were obtained for 23,177 author occurrences and 5332 person groups, and (c) for 8675 papers and 28,042 author occurrences.

Table 1 summarizes the statistics of the test set. Originally, there were 28,042 author appearances within 8675 papers. Then, after English names or single-occurrence author names were excluded, the test set has 23,177 author occurrences and 5332 same-name author groups which were manually disambiguated to obtain 9133 real individuals. The average number of authors per paper was 3.23, and average author ambiguity 1.71.

From (a) of Fig. 2, it is known that about 60% of a total of 5332 same-name person groups have two or three author occurrences, and same-name person groups that have more than 10 occurrences were less than 6%. In addition, (b) of Fig. 2 indicates that 60% of a total of 5332 same-name person groups are unambiguous, and most (80%) of ambiguous persons has two or three namesakes. (c) of Fig. 2 explains that most papers (79.2%) in our test set were collaborated by two to four persons. Fig. 3 shows the token-type distribution for same-name person groups found in the test set. Up to 10 person tokens as the size of a same-name person group, person groups exhibit various numbers of person types, and beyond 10 person tokens, the number of person types within a person group does not attain to the size of the person group. These factors imply that larger same-name person groups do not always cause higher ambiguities. Expectedly, lower ambiguity person groups distribute over more diverse sizes, while showing rare occurrences of high-ambiguity person names.

5.2. Features, clustering method

This study divides coauthors of author A to be disambiguated into different types: *EC*, *IC*, *aIC*, and *cIC*. *Explicit coauthors* (*EC*) mean coauthors explicitly listed in particular citation records of author *A*. *Implicit coauthors* (*IC*) indicate coauthors newly found from Web. *IC*s are further classified into *author-centric* and *coauthor-centric IC*. As exemplified in Section 4, *Author-centric IC*s (*aIC*) are obtained using Algorithm 1 by submitting a pair of author A and one of *EC* of A as a query to web search engines. *Coauthor-centric IC*s (*cIC*) are gathered through Algorithm 1 by creating a pair of arbitrary two *EC*-type coauthors of author A as a web query.

We created an author disambiguation system that divides the same-name author occurrences in citation data into different clusters, each of which are expected to correspond to a real individual. This is equal to grouping papers having the same-name author appearances. Our author disambiguation system performs single-link agglomerative clustering (Jain, Murty, & Flynn, 1999; Sneath & Sokal, 1973) as shown in Algorithm 2. It takes a list of same-name author occurrences in which each name is represented by a set of his/her coauthor names. Clustering starts by making each of author appearances a singleton cluster. Next, the clustering algorithm iterates to merge the most similar two clusters of which similarity equals to or exceeds a threshold into a larger cluster, until the largest similarity between the two clusters is less than the threshold. Cluster similarity $CSim(c_i,c_j)$ for two clusters (c_i,c_j) is defined as the author similarity (a_x,a_y) between the most similar author-pair (a_x,a_y). Author similarity $ASim(a_x,a_y)$ is defined as the count of matched coauthors between two authors, where exact matching is used since our test set in Korean does not suffer from the synonymous author name problem. Note that a new cluster generated by merging is represented as a union of the two smaller clusters to be combined. A cluster-merging threshold θ means that author clusters are merged only if they have an author-pair sharing at least θ coauthor(s).

As baselines to be compared with coauthor-based author resolution, two methods of clustering namesakes are used: singleton-clustering and single-clustering. Singleton-clustering (ST) views that each occurrence of the same name belongs to different persons, and single-clustering (SC) considers that every occurrence of the same name belongs to one and only one person. Thus, ST partitions each of the same-name author instances into a separate cluster, while SC groups all of the same-name author instances into a single cluster.

Algorithm 2

Agglomerative Clustering for Same-name Author Occurrences

Input:	
	$a_1,\ldots,a_n$; same-name author occurrences
	$a_i = \{v_{i1},\ldots,v_{im}\}$; each name occurrence a_i has a set of m ($m \geqslant 0$) his/her coauthor names
	θ; a cluster-merging threshold
Initialize:	
	$c_i = \{a_i\}$; consider each name occurrence a_i as an element of cluster c_i
Loop:	
1	DO
2	For each cluster-pair (c_i, c_j), calculate $CSim(c_i,c_j)$
3	$CSim(c_i,c_j) = \max(ASim(a_x, a_y))$, $\forall a_x \in c_i$, $\forall a_y \in c_j$
4	$ASim(a_x,a_y) = \lvert a_x \cap a_y \rvert$
5	Find the most similar cluster-pair (c_u, c_v)
6	$(c_u,c_v) = \text{argmax}\ CSim(c_i, c_j)$
7	IF $CSim(c_u,c_v) \geqslant \theta$ THEN
8	$c_{u\_v} = c_u \cup c_v$; merge c_u and c_v into a new larger cluster c_{u_v}
9	ENDIF
10	WHILE $(CSim(c_u,c_v) \geqslant \theta)$
Output:	
	Clusters of author occurrences: $\{c_k\}$

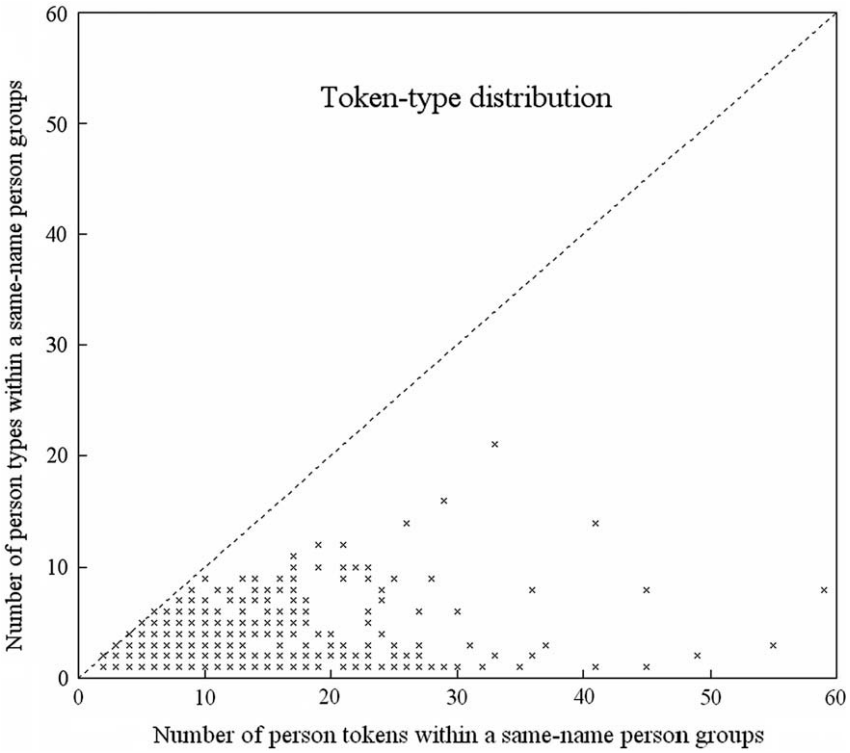

Fig. 3. Token-type distribution for same-name person groups in the test set.

5.3. Evaluation metrics

Since resolving namesakes is a clustering problem, it is required to calculate the degree of agreement between a set of system-output partitions and a set of true partitions. In general, the agreement between two partitions is measured for a pair of entities within partitions. The basic unit for which pair-wise agreement is assessed is a pair of entities (authors in our case) which belongs to one of the four cells in Table 2. Let *M* be the set of machine-generated clusters, and *G* the set of gold standard clusters. In Table 2, then, *a* is the number of pairs of entities that are assigned to the same cluster in each of *M* and *G*. *d* is the number of pairs of entities that are placed in different clusters in each of *M* and *G*. *b* is the number of pairs of entities that are in the same cluster in *M*, and are in different clusters in *G*. *c* is the number of pairs of entities that are in different clusters in *M*, and are in the same cluster in *G*. In other words, *a* and *d* are interpreted as agreements, and *b* and *c* disagreements.

When Table 2 is considered as a confusion matrix for a two-class prediction problem, the standard 'Precision', 'Recall', *F1* measure, and 'Accuracy' are defined as follows.

$$\text{Precision} = a/(a+b)$$
$$\text{Recall} = a/(a+c)$$
$$F1 = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall})$$
$$\text{Accuracy} = (a+d)/(a+b+c+d)$$

'Precision' is the ratio of correct predictions among all pairs of entities predicted to belong to the same cluster in machine-generated partitions. 'Recall' is the ratio of correct predictions among all pairs of entities assigned to the same cluster in gold standard partitions. *F1* is the harmonic mean of 'Precision' and 'Recall'. 'Accuracy' calculates the fraction of agreements

Table 2
Matching matrix for the agreement between two sets of clusters

		Gold standard clusters (*G*)	
		Match	Mismatch
Machine-generated clusters (*M*)	Match	a	b
	Mismatch	c	d

among all pairs of entities between machine-generated and true partitions. For the evaluation of the author disambiguation, most researchers (Culotta et al., 2007; Huang et al., 2006; Kanani & McCallum, 2007; Song et al., 2007) have employed *F1*, and a few others (Aswani et al., 2006) have utilized the 'Accuracy'. The difference between 'Accuracy' and *F1* occurs from whether the count of pairs of authors correctly placed in different clusters is considered or not. McCallum, Nigam, and Ungar (2000) argued that 'Accuracy' would be less informative than *F1* since it may overestimate the clustering performance due to the overwhelming number of pairs of entities that are correctly assigned to different partitions. Adopting the argument of McCallum, this study uses *F1*.

As additional measures, we use two types of clustering errors: over-clustering and under-clustering. They are obtained from the matching matrix of Table 2, where *b* and *c* respectively indicate the counts of pairs of entities that are over-clustered and under-clustered. In other words, *Over-clustering* and *Under-clustering errors* are computed as follows:

$$\text{Over-clustering error} = b/(a + b + c + d)$$
$$\text{Under-clustering error} = c/(a + b + c + d)$$

6. Evaluation results and discussion

Table 3 shows the statistics of coauthor features. On average, 2.85 explicit coauthors were found in the test set, and 19.75 aICs and 17.65 cICs were augmented by the web-based coauthor expansion algorithm which was executed with top 20 documents retrieved through the Google search engine. In order to block the inclusion of erroneous coauthors from irrelevant web pages, the algorithm was actually stopped after at most 50 new coauthors were gathered for each pair of author names. We have evaluated the performance of the web-based coauthor expansion procedure for 100 random pairs of author names. Its recall, precision, and *F*1 were 86.55, 83.52, and 85.01, respectively.

Table 4 compares clustering performances over combinations of different coauthor features using the number of overlapping coauthors set to 1. EC means that a name occurrence a_i in Algorithm 2 is represented by a set of only EC-type coauthors of a_i. EC + aIC indicates that a name occurrence a_i is represented by a union of a set of EC-type coauthors of a_i and a set of aIC-type coauthors of a_i. The other combinations of different coauthor features can be similarly interpreted.

EC (explicit coauthor feature) contributed to resolve author identities, significantly improving two baselines. EC combined with aIC (author-centric implicit coauthor feature) achieved the best performance. cIC corresponding to coauthors of coauthors was also helpful when it was combined with EC. These explain that author identities can be more clearly discriminated by revealing their implicit coauthors through a web-assisted method, supporting a coauthor disambiguation hypothesis. Mixing EC + aIC with cIC slightly decreased the performance, showing that aIC and cIC are not complementary.

EC and aIC represent direct acquaintances among persons since they are direct coauthors of an author to be disambiguated, while cIC indicates indirect acquaintances since they are coauthors of coauthors of an author to be resolved. Table 4 shows that indirect acquaintances are beneficial to the determination of author identities, but their resolution power is weaker than that of direct acquaintances. Precisely, cIC in this study is a first-order indirect acquaintance. When cIC is extended to include coauthors of coauthors of coauthors (of an author to be disambiguated), cIC becomes a second-order indirect acquaintance. A higher-order indirect acquaintance is likely to disambiguate authors much weakly than smaller-order ones. The higher-order cases are not further dealt with in this study.

Table 3
Statistics of coauthor features

Average number of explicit coauthors (EC)	Average number of author-centric implicit coauthors (aIC)	Average number of coauthor-centric implicit coauthors (cIC)
2.85	19.75	17.65

Table 4
Clustering performances over different coauthor features (with the number of overlapping coauthors set to 1)

Features	Recall	Precision	*F*1	Under-clustering error	Over-clustering error
SC	0.7846	0.7017	0.7271	0.0000	0.3239
ST	0.2154	0.2063	0.2082	0.6761	0.0000
EC	0.8279	0.8725	0.8358	0.2303	0.0105
aIC	0.7770	0.7512	0.6994	0.0983	0.0722
cIC	0.6418	0.6308	0.5506	0.2498	0.0692
EC + aIC	0.8692	0.8820	0.8645	0.0900	0.0601
EC + cIC	0.8490	0.8743	0.8494	0.1478	0.0357
aIC + cIC	0.7901	0.7960	0.7550	0.1000	0.0873
EC + aIC + cIC	0.8693	0.8755	0.8609	0.0814	0.0848

SC and ST indicate two baselines: single-clustering and singleton-clustering, respectively.

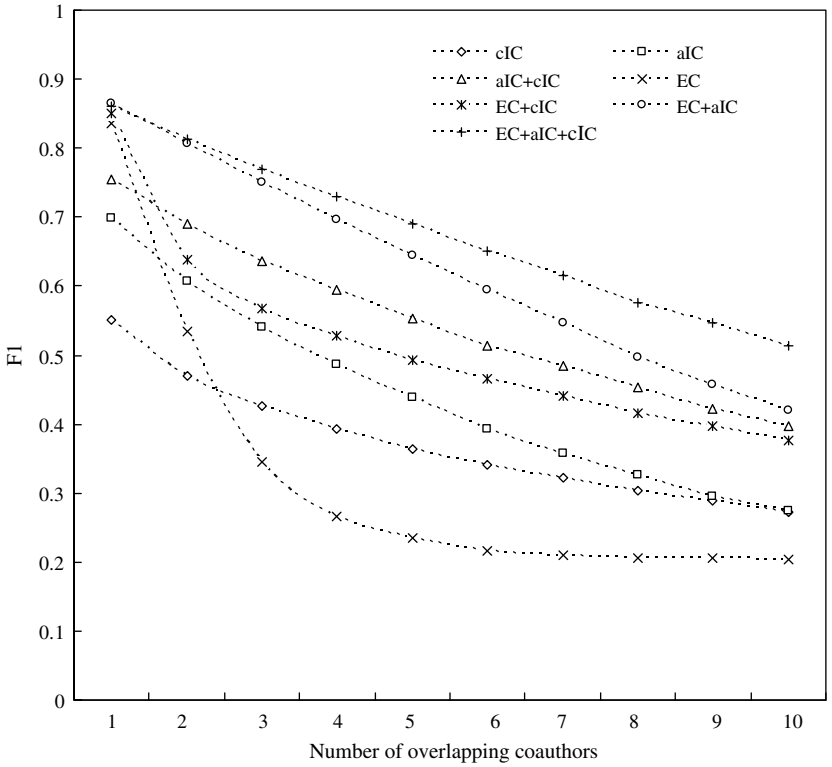

Fig. 4. Effect by the number of overlapping coauthors for different combinations of coauthor features.

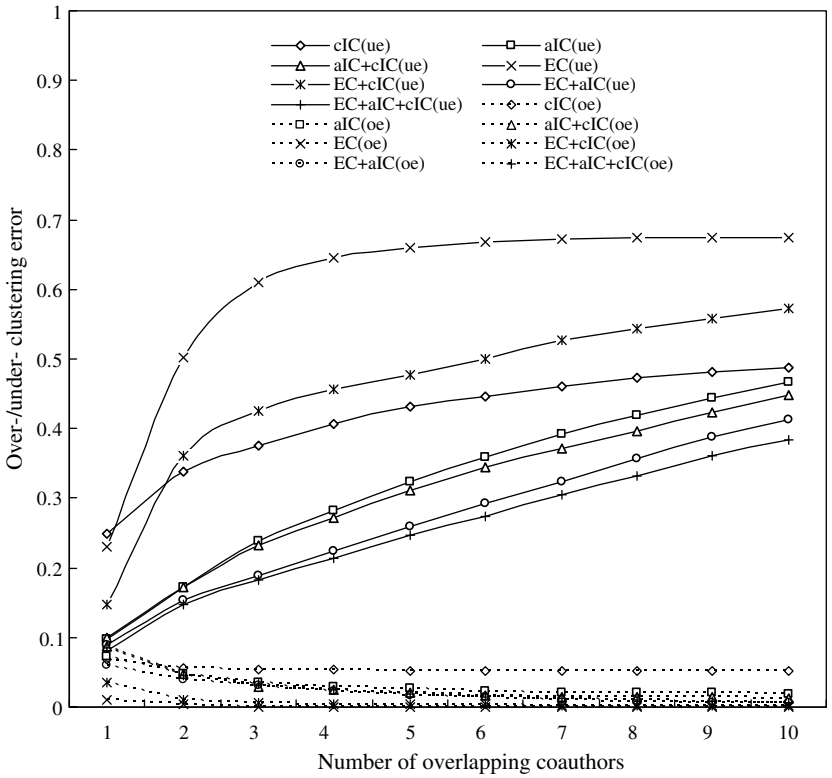

Fig. 5. Effect of the number of overlapping coauthors on two types of clustering errors (ue: under-clustering error, oe: over-clustering error).

Fig. 4 shows the effect of the number of sharing coauthors on author disambiguation for different combinations of coauthor features. A sharp decline of EC-based clustering, which was caused by the small number of explicit coauthors (see Table 3) in citation records, is mitigated by the additional introduction of different types of coauthor features from the web. Overall, the use of more overlapping coauthors as a cluster-merging condition monotonically decreases the clustering performance. The reason for this can be explained by Fig. 5 which shows the effect of the number of overlapping coauthors on two types of clustering errors. Expectedly, as the system requests more sharing coauthors to confirm whether two author instances refer to the same individual, it reduces the risk of mistakingly assigning same identity to different author occurrences and simultaneously increases the risk of not merging author appearances referring to the same person. Fig. 5 indicates that the request for more sharing collaborators to merge authors aggravates the latter under-clustering in larger degrees than it alleviates the former over-clustering. Thus, total errors are not reduced even the count of overlapping coauthors increases. Additionally, Fig. 5 implies that the identities of most authors can be determined by their two to four collaborators, showing that more acquaintances are either superfluous or not available.

Employing the count of sharing coauthors as a decision on the person identity does not consider author-specific numbers of coauthors. Some authors may write papers with numerous collaborators, and others may publish articles with a few coworkers. Hence, it would be more reasonable to use the ratio of coauthors in common in order to merge two same-name author clusters into a larger one. The coauthor ratio $CSimR(\boldsymbol{c}_i,\boldsymbol{c}_j)$ between two author clusters $(\boldsymbol{c}_i,\boldsymbol{c}_j)$ are defined as follows.

$$CSimR(\boldsymbol{c}_i,\boldsymbol{c}_j) = \max(ASimR(\boldsymbol{a}_x,\boldsymbol{a}_y)),\ \forall\ \boldsymbol{a}_x \in \boldsymbol{c}_i,\ \forall \boldsymbol{a}_y \in \boldsymbol{c}_j$$
$$ASimR(\boldsymbol{a}_x,\boldsymbol{a}_y) = |\boldsymbol{a}_x \cap \boldsymbol{a}_y| / \min(|\boldsymbol{a}_x|, |\boldsymbol{a}_y|)$$

An author-clustering algorithm using the coauthor ratio is obtained by replacing $CSim(\boldsymbol{c}_i,\boldsymbol{c}_j)$ and $ASim(\boldsymbol{a}_x,\boldsymbol{a}_y)$ in Algorithm 2 with $CSimR(\boldsymbol{c}_i,\boldsymbol{c}_j)$ and $ASimR(\boldsymbol{a}_x,\boldsymbol{a}_y)$. Fig. 6 shows the effects of the ratio of overlapping coauthors on author clustering using EC + aIC features. As shown in Fig. 6, as the number of overlapping coauthors increases, the ratio of common coauthors keeps the overall clustering performance relatively more stable, compared to the use of the count of sharing coauthors. However, the use of more sharing coauthors by the coauthor ratio as well could not bring about a higher performance, due to its low applicability in spite of its more accurate disambiguation.

Fig. 7 shows the performance of disambiguating authors over the varying sizes of same-name person groups, using EC or EC + aIC features and comparing them with two baselines ST and SC. Fig. 7 was prepared to answer the question of whether coauthor features are helpful to the resolution of authors over varying numbers of same-name author instances. For the var-

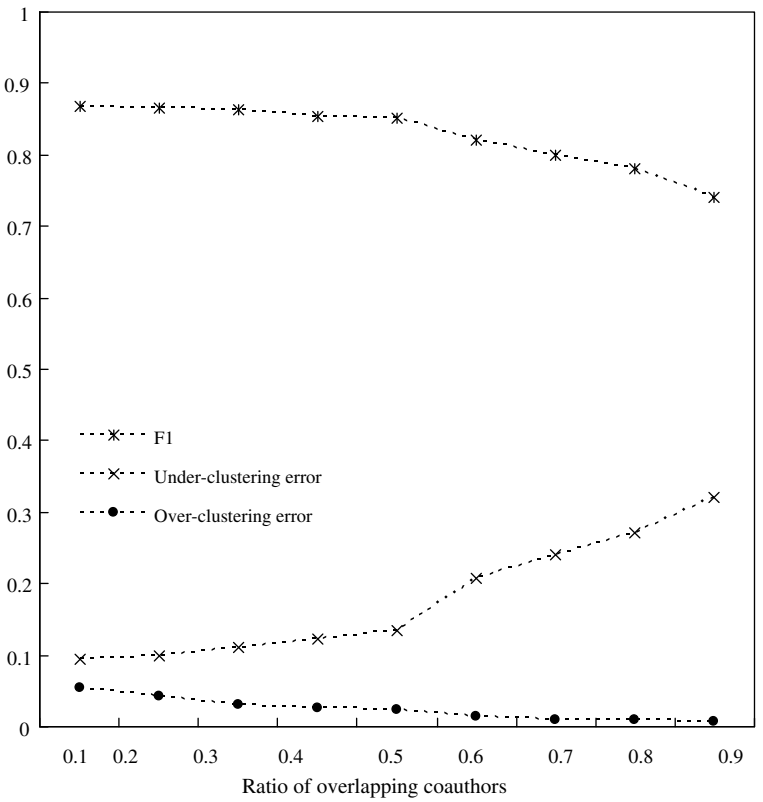

Fig. 6. Effect of the ratio of overlapping coauthors on author disambiguation using EC + aIC.

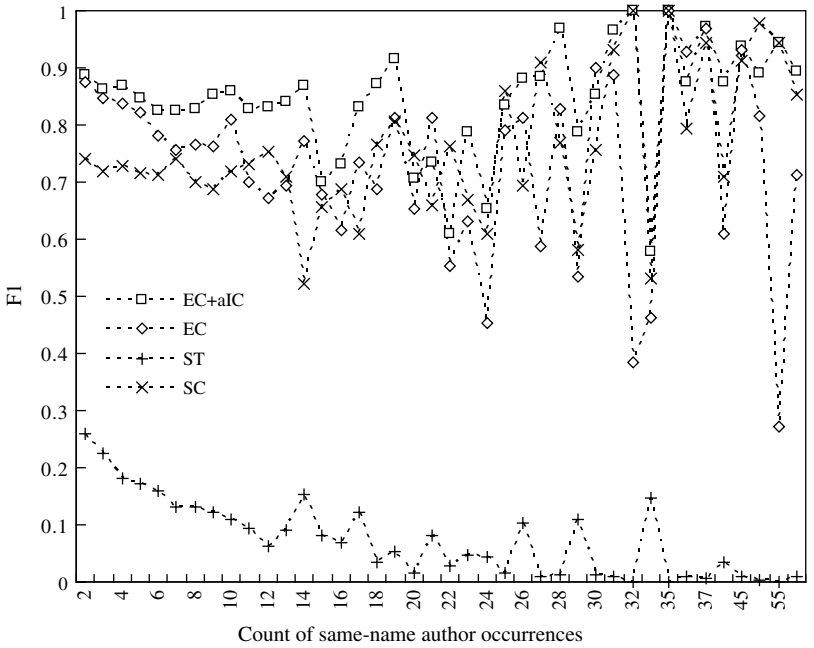

Fig. 7. Clustering performance by the count of same-name author occurrences (with the number of overlapping coauthors set to 1 as a cluster-merging condition).

ious sizes of same-name person groups, EC + aIC-based performances are kept stable up to group size 14, exceeding EC-based ones. In particular, additional aIC features assist to prevent the performance of EC-only clustering from lessening. In the test set, person groups from group size 2–14 cover 85.9% of a total of 5332 same-name person groups, and 97.2% of a total of 23,177 author occurrences. This implies that applying EC + aIC coauthor features to author disambiguation are not sensitively influenced by the various sizes of the same-name person groups. Beyond the size 14 of person groups, the performance fluctuates. However, it may be hazardous to generalize the result, since the average count of same-size person groups is less than 6 beyond group size 14, and is about 3 after group size 20.

In summary, for most same-name person groups, implicit collaborators hidden from known citation records successfully contributed to resolving author identities, independently of their size. Moreover, co-authorship features collectively resolve more than 85% of ambiguities. This largely supports the coauthor disambiguation hypothesis. On the other hand, however, this means that co-authorship alone is not sufficient as an author-disambiguation feature, and raises the need for the adoption of additional features. As another strong biographic feature, emails or affiliations could become good candidates. However, their use requires the availability of full-text for citations. Readily available identity-related features include titles of papers and/or publications which could suggest author-specific research areas. Author occurrences showing highly similar research areas can be merged even though they do not have any coauthors in common. Strong irrelevance of research topics between the same-name authors can be used to prevent them from being grouped even if they share coauthors, only if names of sharing coauthors are not rare.

7. Conclusions

This study attempted to investigate the influence of co-authorship on author name disambiguation. To deal with the scarcity of explicit coauthors appearing in known citation records, a web-rendered technique of gleaning implicit coauthors not listed in known citations was proposed. Under the assumption that a pair of person names mutually determines the identities of each other, a string of two author names in known citations was submitted to a web search engine to gather their unrevealed coauthors from retrieved web pages. The precision of the coauthor-expansion technique, however, can be reduced due to the possibility that even a pair of two person names that have authored the same articles may indicate different pairs of real individuals. Probably, such a problem can be attacked by utilizing commonality/rarity of author names.

Findings from experiments largely supported a coauthor disambiguation hypothesis. More than 85% of author ambiguities were resolved by co-authorship. In addition, the performance of author disambiguation relying on co-authorship alone remained roughly stable without large fluctuations for most ranges of sizes. The use of more coauthors in common helps to uncover the identity of authors more accurately, but not for five or more sharing coauthors, due to their low applicability and/or superfluous-ness. Indirect acquaintances represented by coauthors of coauthors (of an author to be resolved) also contributed to the resolution of author names, but its disambiguation capability was less than that of direct acquaintances which correspond to explicit/implicit coauthors of an author to be resolved.

The use of Korean person names in coauthor-expansion and evaluations may limit the generalization of our findings to the case of English person names. However, note that Korean is one of best suitable languages for this study, since it does not suffer from the matching problem of synonymous person names. We plan to adapt the coauthor-expansion technique for English author names, and repeat the experiments for previous author resolution test sets in English to confirm our results. In addition, this study needs to be extended to include the analysis of relationships among co-authorship and other author-resolving factors such as titles of papers/publications, emails/affiliations, etc.

References

Alani, H., Dasmahapatra, S., O'Hara, K., & Shadbolt, N. (2003). Identifying communities of practice through ontology network analysis. *IEEE Intelligent Systems, 18*(2), 18–25.

Aswani, N., Bontcheva, K., & Cunningham, H. (2006). Mining information for instance unification. In *Proceedings of the 5th international semantic web conference (ISWC)* (pp. 329–342). November 5–9, GA, USA.

Culotta, A., Kanani, P., Hall, R., Wick, M., & McCallum, A. (2007). Author disambiguation using error-driven machine learning with a ranking loss function. In *Proceedings of the 6th international workshop on information integration on the Web (IIWeb-07)*. July 23, Vancouver, Canada.

Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering, 19*(1), 1–16.

Firth, J. R. (1957). *A synopsis of linguistic theory 1930–1955. Studies in linguistic analysis*. Oxford: Philological Society. pp. 1–32.

Gu, L., Baxter, R., Vickers, D., & Rainsford, C. (2003). Record linkage: current practice and future directions. Technical Report 03/83, CSIRO Mathematical and Information Sciences, Canberra, Australia.

Guha, R., & Garg, A. (2004). Disambiguating people in search. In *Proceedings of the 13th international conference on World Wide Web (WWW)*. May 17–20, NY, USA.

Han, H., Giles, C. L., & Zha, H. (2003). A model-based *k*-means algorithm for name disambiguation. In *Proceedings of semantic web technologies for searching and retrieving scientific data*. October 20, Florida, USA.

Han, H., Giles, C. L., Zha, H., Li, C., & Tsioutsiouliklis, K. (2004). Two supervised learning approaches for name disambiguation in author citations. In *Proceedings of the ACM/IEEE joint conference on digital libraries (JCDL)* (pp. 296–305). June 7–11, AZ, USA.

Han, H., Xu, W., Zha, H., & Giles, C. L. (2005). A hierarchical naive Bayes mixture model for name disambiguation in author citations. In *Proceedings of the ACM symposium on applied computing (SAC)* (pp. 1065–1069). March 13–17, New Mexico, USA.

Han, H., Zha, H., & Giles, C. L. (2005). Name disambiguation in author citations using a *k*-way spectral clustering method. In *Proceedings of the ACM/IEEE joint conference on digital libraries (JCDL)* (pp. 334–343). June 7–11, CA, USA.

Harris, Z. (1954). Distributional structure. *Word, 10*(23), 146–162.

Huang, J., Ertekin, S., & Giles, C. L. (2006). Efficient name disambiguation for large scale databases. In *Proceedings of the 10th European conference on principles and practice of knowledge discovery in databases (PKDD)* (pp. 536–544). September 18–22, Berlin, Germany.

Jain, A., Murty, M., & Flynn, P. (1999). Data clustering: A review. *ACM Computing Surveys, 31*(3), 264–323.

Jaro, M. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association, 84*(406), 414–420.

Kanani, P., & McCallum, A. (2007). Efficient strategies for improving partitioning-based author coreference by incorporating Web pages as graph nodes. *In Proceedings of the 6th international workshop on information integration on the Web (IIWeb-07)*. July 23, Vancouver, Canada.

Kanani, P., McCallum, A. & Pal, C. (2007). Improving author coreference by resource-bounded information gathering from the Web. In *Proceedings of the 20th international joint conference on artificial intelligence (IJCAI)* (pp. 429–434). January 6–12, Hyderabad, India.

Lee, D. W., On, B. W., Kang, J. W., & Park, S. H. (2005). Effective and scalable solutions for mixed and split citation problems in digital libraries. In *Proceedings of the international workshop on information quality in information systems (IQIS)* (pp. 69–76). June 17, Maryland, USA.

Mann, G. S., & Yarowsky, D. (2003). Unsupervised personal name disambiguation. In *Proceedings of the 7th conference on computational natural language learning (CoNLL)* (pp. 33–40). May 31–June 1, Edmonton, Canada.

McCallum, A., Nigam, K., Ungar, & L. H. (2000). Efficient clustering of high-dimensional data sets with application to reference matching. In *Proceedings of the 6th ACM SIGKDD international conference on knowledge discovery and data mining (KDD)* (pp. 169–178). August 20–23, 2000, MA, USA.

McRae-Spencer, D. M., & Shadbolt, N. R. (2006). Also by the same author: AKTiveAuthor, a citation graph approach to name disambiguation. In *Proceedings of ACM/IEEE joint conference on digital libraries (JCDL)* (pp. 53–54). June 11–15, NC, USA.

Reuther, P. (2006). Personal name matching: New test collections and a social network based approach. Technical Report 06-01, Mathematics/Computer Science, University of Trier.

Sneath, P., & Sokal, R. (1973). *Numerical taxonomy*. London, UK: Freeman.

Song, Y., Huang, J., Councill, I., Li, J., & Giles, C. L. (2007). Efficient topic-based unsupervised name disambiguation. In *Proceedings of the ACM IEEE joint conference on digital libraries (JCDL)*. June 18–23, Vancouver, Canada.

Tan, Y. F., Kan, M. Y., Lee. D. W. (2006). Search engine driven author disambiguation. In *Proceedings of ACM/IEEE joint conference on digital libraries (JCDL)* (pp. 314–315). June 11–15, NC, USA.

Torvik, V. I., Weeber, M., Swanson, D. R., & Smalheiser, N. R. (2005). A probabilistic similarity metric for Medline records: A model for author name disambiguation. *Journal of the American Society for Information Science and Technology (JASIST), 56*(2), 140–158.

Wan, X., Gao, J., Li, M., Ding, B. (2005). Person resolution in person search results: WebHawk. In *Proceedings of the 14th ACM international conference on information and knowledge management (CIKM)* (pp. 163–170). October 31–November 5, Bremen, Germany.

Winkler, W. E. (2006). Overview of record linkage and current research directions. Research Report Series #2006-2, Statistical Research Division, US Census Bureau.

Yang, K. H., Jiang, J. Y., Lee, H. M., & Ho, J. M. (2006). Extracting citation relationships from Web documents for author disambiguation. Technical Report, TR-IIS-06-017, Institute of Information Science, Academia Sinica, Taipei, Taiwan.

In-Su Kang received his M.S. and Ph.D. degrees in Computer Science from POSTECH in 1999 and 2006. Currently, he teaches students in Kyungsung University. His research interests include IR, NLP, and semantic web.

Seung-Hoon Na received his M.S. and Ph.D. degrees in Computer Science from POSTECH in 2003 and 2008. Currently, he works for POSTECH as a post-doctoral researcher. His research interests include IR and NLP.

Seungwoo Lee received his M.S. and Ph.D. degrees in Computer Science from POSTECH in 1999 and 2005. Currently, he works for KISTI as a senior researcher. His research interests include IR, NLP, and semantic web.

Hanmin Jung received his M.S. and Ph.D. degrees in Computer Science from POSTECH in 1994 and 2003. Currently, he works for KISTI as a senior researcher. His research interests include information extraction, IR, NLP, and semantic web.

Pyung Kim received his M.S. and Ph.D. degrees in Computer Science from Chungnam National University in 1999 and 2004. Currently, he works for KISTI as a senior researcher. His research interests include IR, text mining, and semantic web.

Won-Kyung Sung received his Ph.D. degrees in Linguistics from University of Paris 7 in 1996. Currently, he works for KISTI as a senior researcher. His research interests include NLP and semantic web.

Jong-Hyeok Lee received his M.S. and Ph.D. degrees in Computer Science from KAIST in 1982 and 1988. Currently, he is a full professor in POSTECH. His research interests include machine translation, IR, and NLP.