Monte D Evans. A New Approach to Journal and Conference Name Disambiguation through K-Means Clustering of Internet and Document Surrogates. A Master's Paper for the M.S. in I.S degree. April, 2009. 54 pages. Advisor: Catherine Blake

Bibliometrics has a long history in Information Science. The validity of any bibliometric analysis depends on accurate citations. We introduce an approach that combines author names and Internet document surrogates with K-means clustering to disambiguate journal and conference titles automatically. To evaluate the quality this approach we used records from the Digital Bibliography & Library Project (DBLP). We found there are 2.54±1.52 authors per articles. A manual analysis of 125 articles selected at random from the 1.18 million DBLP citations revealed only seven article pairs from the same publication venue. We describe the changes in cluster properties as the number of articles increases from 100 to 25,000. Our findings suggest that additional features are required to disambiguate journal and conference names accurately. As 60.86% of the DBLP articles are published at conferences future efforts should focus on conference name disambiguation.

Headings:

Citation Analysis

Publication Analysis

Citation Disambiguation

A NEW APPROACH TO JOURNAL AND CONFERENCE NAME DISAMBIGUATION THROUGH K-MEANS CLUSTERING OF INTERNET AND DOCUMENT SURROGATES

by
Monte D. Evans

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

April 2009

Approved by

_____

Catherine Blake

Table of Contents

Table of Figures **2**
Table of Equations **3**
Table of Tables **4**

1 Introduction **5**
1.1 Definitions **6**

2 Background **8**

3 Related Work **10**
3.1 String Matching and Word Sense Disambiguation Methods **11**
3.2 Machine Learning Algorithms **13**
3.3 Internet surrogates **14**

4 Methodology **17**
4.1 Candidate Features **17**
4.1.1 Primary Author Name 18
4.1.2 Co-authors 19
4.1.3 Base URL 19
4.1.4 Article title 20
4.2 Pre-processing and Data Collection **21**
4.2.1 Collecting citation information 21
4.2.2 Representing the Host Frequency 25
4.2.3 Collecting Base URL Features 27
4.3 Clustering **28**
4.3.1 Methods 28
4.3.2 Clustering Vectors 29

5 Results **30**
5.1 Manually Evaluated Clusters **30**
5.2 Varying the number of articles **39**
5.3 Visualizing clustering performance **42**

6 Discussion **44**
6.1.1 Limitations and Future Work 45

7 Conclusions **46**

8 References **49**

Table of Figures

Figure 1: DBLP Entity Relationship Schema 21
Figure 2: XML Import Parser Example 23
Figure 3: Oracle Feature Selection 29
Figure 4: Example of Entity Encoding Issues 35
Figure 5: Sample 100 elements with K = 10 43
Figure 6: Sample 100 elements with K = 100 43
Figure 7: Sample 2,500 elements with K = 10 43
Figure 8: Sample 2,500 elements with K = 100 43
Figure 9: Sample 25,000 elements with K = 10 43
Figure 10: Sample 25,000 elements with K = 100 43

Table of Equations

Equation 1: Document Recall 16
Equation 2: Inverse Host Frequency Equation 27

Table of Tables

Table 1: Citation Distribution Information in DBLP 19
Table 2: DBLP Distinct Format Distribution 20
Table 3: DBLP Most Prolific Authors 21
Table 4: Sample A - Manual Analysis of Clusters from a Random Set of Journals and Conferences 31
Table 5: Sample B - Manual Analysis of Clusters from Same Journal 36
Table 6: Manual Analysis Summary Statistics 39
Table 7: Cluster Performance Metrics for K = 100 40

1 Introduction

In a letter to Robert Hooke, Sir Isaac Newton stated “If I have seen further it is by standing on the shoulders of giants” (Merton 1993). Scientific research and discovery is largely based upon a gradual refinement of others work and ideas, a process Kuhn calls "normal science" (Kuhn 1970). Therefore, having accurate citation information is a matter of the utmost importance to a researcher and his or her ability to have the greatest possible breadth of knowledge available.

Bibliometrics, a term coined by Alan Pritchard in 1969, is the application of quantitative analysis and statistics to publications such as journal articles and their accompanying citations (Pritchard 1969; Hauptly 2008). One key challenge in bibliometrics is the accurate identification of journal titles. Garfield, arguably the creator of the field of bibliometrics and the inventor of the ISI impact factor for journal publications, indicated the difficulties with journal names to create a corpus of citations in his 1972 paper on citation analysis:

> An immensely irksome problem was the inconsistency with which different authors and editors abbreviate journal titles in their references... Some idea of the work involved in this standardization can be had from the fact that there were more than 100,000 different abbreviations for the 12,000 individual journal titles cited in the 3-month sample.
> Journals merge; they split into new journals, or into “sections” that may be published separately or together. They change titles, with or without continuing their numbering of volumes and issues. Some journals appear in one or more translations; some such translations are complete, other selective, and some are similar, other differently, numbered. (Garfield 1972)

Garfield's analysis directly addresses the difficulty in maintaining the consistency and accuracy of journal names in citations. He shows that his corpus consists of an average of eight different journal title permutations for each journal (inconsistencies exacerbated by a factor of eight) in 1972. The Web of Science has over 23,000 journals, 110,000 conference proceedings, and 9,000 websites (Thomas Reuters 2008) thus, manual disambiguation to resolve incorrect journal and conference names is untenable. More than thirty years ago, Garfield stated that the problem of journal name correction and disambiguation is simply insoluble and authors should begin using the full journal title whenever possible. Despite this recommendation, authors have continued to use abbreviated journal titles. For example, the abbreviation is Astron. J is for both the Astronomical Journal and the Astronomy Journal. Additionally, the Annals of Physics and Annalen der Physik are abbreviated Ann. Phys.

Our goal is to explore the effectiveness of clustering documents to disambiguate journal titles via their author (see Section 4.1.1) and co-author features (see Section 4.1.2) as well as the use of Internet surrogates and the associated Internet address of the server that hosts the surrogates (see section 4.1.3).

1.1 Definitions

The following terms provide the conceptual framework for the methodology and analysis in this study. We accompany any deviations from these definitions presented within the text with a parenthetical reference to the modifications.

- Article/document - A single instance of a written paper from a journal or conference proceeding. The article or document can include editorials, reviews, proceeding notes, research findings, or any other information located within the venue of a journal or conference.
- Bibliometrics - The application of mathematics and statistical methods to books and other media of communication (Pritchard 1969). Bibliometric measures show citation rates, author impact, and journal impact (De Solla Price 1976).
- Bibliometric coupling - A single item of reference used by two papers is defined as a unit of coupling between the two papers (Kessler 1963). Bibliometric coupling defines the similarity between two documents as the number of references that two articles have in common.
- Citation - A reference to a published or unpublished source of information such as a book, article, web page, or other document. Citations include information pertaining to: the author, location of the document, volume, page range, and year of publication.
- Co-citation analysis – The study of links between pairs of documents as indicated by a competent specialist, namely the author of the article. Co-citation defines similarity between two documents as the number of articles that cite both documents. If two articles are cited in the same paper, they are likely related to each other because they belong to the same topic area or the topic areas are closely connected. Many co-citations may be unrelated in an

individual article but a sufficiently large sample of cited articles mitigates this random "noise" (Schildt, Zahra and Sillanpää 2006).

- Journal and conference name: The unified title for a serials collection. Journal listings as defined by the Online Computer Library Center (OCLC) specification for the Machine Readable Cataloging (MARC) 130 field, created by the primary publisher, or the official name for a conference, not including the year, as provided by the hosting institution or publisher.

2 Background

Bibliometrics is the application of mathematics and statistical methods to books and other media of communication that is used to discover patterns and trends over a corpus and evaluate how authors impact specific disciplines. However, the quality of bibliometric analysis depends on accurate journal and author names. Without a coherent and systematic approach to disambiguation of journal and author names, bibliometrics may misrepresent an author's impact. Since scholars depend on the correctness of information obtained through bibliometric analysis, including the process of bibliometric coupling -- measuring the number of references two papers have in common to test for similarity -- having accurate data is paramount (Turnbull 1998).

The field of bibliometrics is expansive, so we refer the reader to Morris and Van der Veer Martens (2008) for an overview of the research within this field. This abridged set of papers outlines a subset of problems and techniques involved in the process of author disambiguation (Elmacioglu, et al. 2007; Tan, Kan and Lee 2006;

Torvik, et al. 2005; Yang, et al. 2006). We applied disambiguation techniques similar to those in the aforementioned works, but to journal and conference names rather than author names. A citation can consist of several features including, but not limited to, authors, year of publication, name of journal or conference, page ranges, format, and location of resource. Scientific disciplines exclude certain information, for instance, chemistry citations may exclude article titles, and literature published in Medline before 2002 did not include the full representation of the author's name (Torvik, et al. 2005).

This study explores the use of Internet document surrogates obtained through online search engines coupled with author and co-author features to create clusters of related articles that have the same journal and conference names automatically. Document surrogates include electronic PDF versions of the article, the article's abstract page, or related websites such as the author's homepage.

Disambiguation systems rely on accurate journal or conference title to enable the system to perform the clustering process (Han, Xu, et al. 2005; Torvik, et al. 2005; Smith 2004). Inaccuracies in the journal or conference name negatively influence the performance of clustering algorithms. Even industry-standard bibliometric data tools, such as Thomas Scientific's ISI Journal Citation Report, are susceptible to these citation errors. Kan and Tan (2008) do not recommend the use of context-free manual or semi-manual equivalency tables of journal and conference names to disambiguate names for two reasons. Firstly, creating the tables is time intensive and does not correct subtle inaccuracies such as misspellings and truncated phrases. Secondly, disambiguation may combine unrelated journal and

conference names, for example, the abbreviation "Ann. Phys." Applies to three distinct journal names, the Annals of Physics, Annales de Physique, and Annalen der Physik (Garfield 1972). A simple translation table would be unable to resolve these ambiguities. It is necessary to locate the article's original source and examine the surrounding context to evaluate the name of the article's journal or conference accurately.

3 Related Work

Journal and conference disambiguation is an example of the more general disambiguation problem that has been studied since the 1960's. Disambiguation and identification techniques are numerous; and include string matching and word sense disambiguation, machine learning algorithms, and the use of external information surrogates. We will briefly discuss the first two implementations and then focus extensively on the last implementation. The reader is advised to refer to the referenced articles for additional background information on the former two techniques.

The study of bibliometrics is one central facet of research in the information and library science community. Bibliometric analysis emerged from the initial work of Eugene Garfield and his creation of the Science Citation Index in 1960. Garfield describes citations as "brief representations of the documents they identify" (Garfield 1964). He states, "only a small number of reference citations are needed to isolate uniquely a particular document in the collection from all others" (Garfield 1964). Bibliometric analysis allows researchers to create maps of scientific

innovation and identify the evolution within sub-groups of what Kuhn refers to as "Revolutionary Science" (Kuhn 1970).

Small's work in co-citation analysis shows that circular citation references between authors (Small 1973) reflect peer recognition and indentify "microevolutions" within sub-disciplines (Small 2003). Researchers also utilize automatic techniques and agglomerative clustering algorithms, creating linkages between citations to show relationships between authors (Morris and Van der Veer Martens 2008). The section below describes three key methods to disambiguate authors: word sense disambiguation, machine learning algorithms, and Internet surrogates.

3.1 String Matching and Word Sense Disambiguation Methods

Researchers reduce ambiguity in natural language through the creation of automatic disambiguation techniques known as "lexical association" and "lexical preference". These two methods tokenize sentences into a structured context and allow researchers to create equivalency tables based on the placement of the words within the sentences (e.g. nouns phrase, verbs, past participles). These techniques rely on the use of thesauri and meta-dictionaries to create the equivalency tables (Brill and Resnik 1994; Baker, et al. 1994).

Brill created a greedy search using WordNet's noun database to resolve prepositional phrase attachments. Through the aid of a supervised training dataset and a comprehensive set of rules, he could accurately predict and annotate ambiguous phrases and provide the appropriate context for each phrase in question.

Baker used the KANT (Knowledge-based, Accurate Natural-language Translation) natural-language translator in an attempt to produce sentences devoid of ambiguity through transformations that did not require any post-processing. She used a standard markup language to annotate the composition of each sentence and then used heuristic measures to reduce ambiguous phrases. For example, if the system encountered the phrase "The parts must be put back together" it would transform the sentence to "The parts must be reassembled", thus removing ambiguity.

The annotation approach provides researchers with numerous advantages. A subset of the English language follows a set of grammatical rules that allows tokenization-based systems to identify and disambiguate more than 50% of the article (Baker, et al. 1994). Knowing that a term could be a noun or a verb allows researchers to disambiguate it more readily. The term tokenization process is replicable and requires only a limited understanding of the process. Additionally, the system's output is understandable and interpretable without the aid of computers, reducing the complexities involved in decoding information stored in machine-structured formats.

The disadvantage of this approach is that the rules and regular expressions are manually encoded and are difficult to maintain when transitioning to a new corpora. This approach relies on equivalency tables that require subject-specific external datasets that may not provide an exhaustive list of journal and conference proceeding names. Furthermore, applying these techniques to articles in a different language is difficult since not all languages follow the same grammatical patterns

found in English. Tokenization and tagging systems are appropriate for relatively known and predictable corpora, however they are not an elegant a solution for corpora with large amounts of unknown data or non-structured data such as citations.

3.2 Machine Learning Algorithms

Another method of bibliometric disambiguation is the use of machine learning algorithms. These algorithms use features from a citation such as the journal title or conference name, co-authorship, and domain of research to automatically group, or cluster, similar articles together (Torvik, et al. 2005). Learning algorithms include Bayesian classifiers, hidden Markov Models, spatial clustering, and vector analysis.

The input dataset affects the performance of machine learning algorithms. For instance, Bayesian algorithms employ simple probabilistic classifiers and exhibit better performance when using a limited training corpus with a new and unseen textual dataset. In contrast, K-means clustering partitions data into a user-defined number of clusters based on the similarity of the data, using the smallest distance between two vectors as a partitioning metric.

Researchers in disciplines ranging from computer science to medicine use different machine learning approaches to extract salient pieces of information from their discipline-specific documents (Metzler and Croft 2005; Lawrie, Croft and Rosenberg 2001; Rosen-Zvi, et al. 2004). Machine learning algorithms enable researchers to use an unsupervised approach to aid in the disambiguation process

by allowing the algorithm to use similarity metrics to detect patterns within the dataset.

3.3 Internet surrogates

Researchers have introduced methods that use Internet surrogates to supplement a machine-learning algorithm's primary corpora. These researchers use a combination of features from the documents, such as the article's title, to collect surrogate documents from online search engines and other external data sources. These surrogates introduce new information to the learning algorithms thus allowing them access to data outside the scope of the original corpus.

Gideon and Yarowsky (2003) create an unsupervised personal name disambiguation system that can distinguish the real world referent of a given name in context. The system uses a language independent bootstrapping process to collect bibliographic facts about the authors using Internet search engines. The system uses the results from the search engines to identify bibliographic patterns pertaining to an author such as their date of birth, location of residence, or field of study. Gideon et al clusters these bibliographic patterns together, producing clusters of keywords (occupation, age, field of study) and the corresponding referent (such as Jim Clark - Netscape Founder versus Jim Clark - Car Salesman in Kansas) (Gideon and Yarowsky 2003).

Lee and colleagues (2005) research two problems in bibliographic databases: namely the mixed citation problem where different scholars' citations are conflated because of name similarities and the split citation problem where the same author appears under a different name variant (Lee, et al. 2005). Using the citation dataset

provided by the Internet Data Base and Logic Programming website (now known as the Digital Bibliography & Library Project), given an author A_i they can accurately identity false citations by another author A_j, even if the authors share identical name spellings. Lee's approach is to create relational models based on terms within the citation and to combine data mining approaches, such as Naive Bayes, String-based Distance Metrics, and Support Vector Machine, to indentify unique authors within a citation.

This study of journal and conference name disambiguation is closely based upon research presented in a paper by (Tan, Kan and Lee 2006). Like the preceding author, Tan used the DBLP database to solve the issue of mixed citation. However, unlike Lee who uses the Internet corpora exclusively, Tan submits the title of each article to a search engine. He then uses the relevant URL weighted by its Inverse Host Frequency (see Section 4.1.3) as features to compute the similarity between two citations using cosine similarity. Finally, he uses hierarchical agglomerative clustering to derive the k-clusters that represent the disambiguated authors. Tan uses Internet search surrogates as the training set for his unsupervised classification algorithm.

The research of (Yang, et al. 2006) reflects and uses the works of Lee and Tan to create an author disambiguation method to address the problems of information scarcity and noise in citations. This research uses a slightly modified version of the Tan Internet surrogate search by excluding URLs that were contained in the DBLP dataset. Yang uses the co-author, title, and venue of the article along with the URL features obtained through the search surrogates as a feature for the K-ways

clustering implementation. The Yang author disambiguation approach, including the use of web surrogates, was statistically better than the Tan approach when comparing the same citations and using the same recall evaluative metrics.

The advantages to these approaches are that over time the corpus transforms, allowing for new data and article permutations. The learning algorithms are exposed to new information as soon as it becomes available to the online search engines, thus affording the classifiers the possibility to determine similarities in a greater number of documents. This enables researchers to expand the domain of their initial study, to allow for the retrieval of documents from external disciplines or subject areas. Additionally, the corpus will "evolve" over time, provided that as new research emerges within a field, it becomes accessible to the electronic source and the learning classifiers.

$$r = \frac{| D_{relevant} \cap D_{retrieved} |}{D_{relevant}}$$

Equation 1: Document Recall

However, Internet surrogates are not without problems. The use of Internet surrogates may increase the recall measure for a particular query. Recall is the number of relevant documents retrieved by the system divided by the total number of existing relevant documents (see Equation 1). An increased recall measure can introduce irrelevant information into the corpus, thus affecting decisions based upon the quality of the dataset. Internet surrogates can also make the application of certain statistical methods difficult due to the possible increased number of false

positives. Furthermore, certain clustering techniques such as K-means rely on a user-defined cluster size and an increase in spurious information could make clusters less meaningful.

Internet surrogates must be carefully obtained, evaluated, and added to the corpus to minimize false positives. The aforementioned articles use features from the original corpus that are most likely to produce relevant surrogates. Researchers have been using data from these surrogates to supplement meta-information about documents already in their corpus, using search engines such as Google and Yahoo. It is imperative that the use of Internet surrogates to enhance or supplement a data collection does not introduce additional ambiguity.

4 Methodology

Our goal is to explore the effectiveness of document clustering to disambiguate journal title via their author and co-author features as well as the base Internet Protocol (IP) address of the server that hosts these articles (see Section 4.1.3). In contrast to the works presented in Section 3.3 that explore author name disambiguation, the process of journal and conference name disambiguation poses additional challenges described in the discussion section (see Section 6).

4.1 Candidate Features

The selection of features for use in any clustering algorithm impacts the partitioning quality of subsequent clusters. The process of journal and conference name disambiguation is no different. Below, we define the features that we

considered to introduce to the clustering process, including features that we did not use in this study.

4.1.1 Primary Author Name

The use of the primary author's name is a useful clustering feature because an author typically specializes in a distinct area of the sciences or humanities. An author will usually write articles within the same subset of journals or conferences, thus allowing the clustering algorithm to reduce the vector space of ambiguous journal and conference names and mitigate the complexities in selecting the appropriate journal name. The primary author's name as a clustering feature is unusual in disambiguation tasks since these studies (Tan, Kan and Lee 2006; Yang, et al. 2006; Lee, et al. 2005) concerns lie with disambiguating the primary author and not the journal or conference name. However, using the primary author's name as a feature without using an authority control record should be met with caution, because a single author entry could represent several different authors. For example, the entry “Yu Chen” in the Digital Bibliography & Library Project (DBLP) database references three different authors, verified by the fact that the authors do not share the same institutions for articles published in the same month of the same year in three distinct journals. The first "Yu Chen" reference is from the University of California, Los Angeles, the second is a Microsoft Beijing researcher, and the third is the senior professor from Renmin University of China (Han, Zha and Giles 2005). Another study showed that in a selection of common surnames from the United Kingdom, 92% of the names chosen resulted in at least two different authors whose

publication were incorrectly merged into one author entry (Jaffri, Glaser and Millard 2008).

4.1.2 Co-authors

Disambiguation studies by Yang and Han suggest that authors appear to write articles with a small group of other authors, and this co-author relationship is an important feature in disambiguating a primary author (Han, Zha and Giles 2005) and (Yang, et al. 2006). We can use this co-author relationship as a clustering feature and improve the data partitioning process because vectors with highly similar co-authors are geometrically similar. Co-authors have the same ambiguity issues as that of the primary author, however the Yang study suggests this is a non-issue. Moreover, Torvik, et al. (2005) suggests that some journals, such as articles in Medline until 2002, did not record the full first name of any author of an article, so the inclusion of co-author information may negatively affect clustering performance.

4.1.3 Base URL

Tan et al. introduced the idea of using metadata from servers that host articles as a clustering feature (Tan, Kan and Lee 2006). As authors begin to publish their articles electronically, the system described by Tan can use URLs to disambiguate authors by showing relatedness of similar articles published in the same journal. For example an article published in the Journal of American Society for Information Science and Technology (JASIST) follows a set of journal submission guidelines. These guidelines ensure consistency between the articles within JASIST making it unlikely that an article not related to information science or technology

would appear. Using this factor of article similarity in journals the base URL should predict related articles.

The base URL suffers from issues that can adversely affect the performance of a machine-learning algorithm. For example, web resources can have URLs such as http://www.acm.org and http://delivery.acm.org that appear different, but may provide identical content. Special care is necessary when including URL information (see section 4.2). Although the effectiveness of using URLs remains unclear, several recent studies indicate positive results (Han, Zha and Giles 2005; Lee, et al. 2005; Elmacioglu, et al. 2007). The Elmacioglu article was able to achieve a purity measure of 0.73 when using URLs of author's personal homepages for the purpose of disambiguation.

4.1.4 Article title

The article's title is a useful metric in determining the relationship between authors and co-authors (Torvik, et al. 2005). Title information often appears in multiple citations along with a list of attributes, such as the author's name. Information retrieval techniques, such as term weighing and frequency, could indicate candidate terms that normally co-occur within a journal or conference. The primary objective of this study is to disambiguate journal and conference names automatically with limited or no human assistance and using articles titles should aid in this aspect. Due to time constraints, we did not include the article's title as a feature for clustering in this study.

4.2 Pre-processing and Data Collection

We describe the process of collecting lists of citations to introduce to the K-means clustering algorithm. We also present the tools and techniques we use to pre-process the data to ensure uniformity.

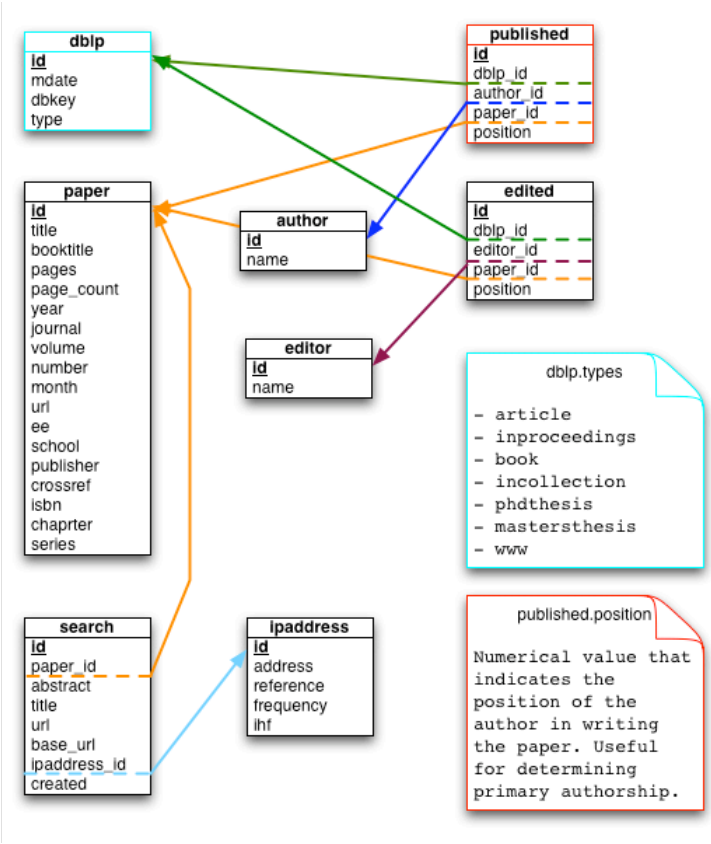

Figure 1: DBLP Entity Relationship Schema

4.2.1 Collecting citation information

The citations used for this study are provided by the Digital Bibliography & Library Project (DBLP) database. The DBLP provides access to citations from hundreds of journals, books, conferences proceedings, and even PhD theses. We used the February 2009 version of the database containing 1,180,280 unique citations from 692,534 distinct authors. This dataset does not explicitly

disambiguate author names, so this figure does not account for author conflation. The DBLP dataset is available in a compressed 550 megabyte XML file that includes most of the structure necessary to reassemble the bibliometric components found in the online version of the DBLP (http://www.informatik.uni-trier.de/~ley/db/). For this study, we created a database to separate the DBLP into citations, papers, authors, editors, and publication records (see Figure 1 for the complete schema).

We created a series of command line PHP scripts to extract the XML data, parse the DBLP database's structure, and populate a MySQL database with the resulting information. This process allows the data to exist in a relational format, suitable for logical querying with standard SQL-commands. The information in the MySQL database includes articles, conference proceedings, books, collections, PhD and master theses, as well as online resources, such as home pages and departmental websites.

The DBLP database encodes all non-ASCII characters, such as accent marks, circumflexes, and umlauts, with their HTML-equivalent representation. We converted these HTML-equivalent characters into their UTF-8 Unicode-equivalent formats, thus enabling the XML parser to function properly. Our script converted 440,891 instances (or 3.2%) of the original DBLP non-ASCII records into a parse-able format in the MySQL database.

After completing the extraction process, the following grouping of citations emerged (see Table 1), where 60% of all citations used during our random selection process are likely citations from conference proceedings. To produce a collection of citations that would form clusters based on the journal or conference name requires the selection of a non-biased random data sample of all of the citations.

Table 1: Citation Distribution Information in DBLP

Publication Type	Records	Percentage
Journals	439,171	37.21%
Proceedings	718,349	60.86%
Books	1,425	0.12%
Collections	8,419	0.71%
PhD Thesis	90	0.01%
Master Thesis	8	0.00%
WWW Sites	12,818	1.09%
Total	1,180,280	100.00%

```
$ php select_article.php
Please enter the number of articles to select: 100
100 random articles selected. (Showing co-authors and URLs):
1 (21015) syntax-based semi-supervised named entity tagging
2 (771260) a locally-organized parser for spoken input
3 (927006) an nc algorithm for the clique cover problem in cocomparability graphs and
its application
4 (602981) facetransfer: a system model of facial image rendering
5 (734008) home page
6 (985385) polynomial decomposition algorithms
```

Figure 2: XML Import Parser Example

We created a script to select random citations without replacement, from the DBLP database (see Figure 2 for import example). We used these random citations as input to the clustering algorithm.

For each random article, we extracted the following features: the article's unique identifier; the title of the article; and the first five authors by the order they are listed in the article's citation. The distribution of unique journal, conference, etc. that an article is published is listed in Table 2. The collection contains documents with as few as one author and no co-authors to as many as 115 co-authors in a single article. We determined that there are 2.54 ± 1.52 authors per citation in the DBLP collection.

Table 2: DBLP Distinct Format Distribution

Publication Type	Records	Percentage
Journals	713	4.37%
Proceedings	2,586	15.83%
Books	78	0.48%
Collections	41	0.25%
PhD Thesis	90	0.55%
Master Thesis	8	0.05%
WWW Sites	12,818	78.47%
Total	16,334	100.00%

Based on this information, we decide to use the first five authors, anticipating that this would be sufficient for accurate clustering without introducing noise into the dataset due to empty or null co-author values (e.g. most citations have a maxium of two authors). Although we acknowledge that one distinct author entity may represent several different individuals, the system uses the author's first initial and complete last name. We then standardize these names to lower case to maintain consistency and re-order the name such that "Smith, John F." becomes "j smith".

4.2.2 Representing the Host Frequency

As stated in Section 4.1, these experiments explore the degree to which author and co-author names along with the base URL of the servers that host the article can be used with a clustering algorithm to predict the journal or conference name. The article's base URL presented several challenges in accurately selecting the most appropriate, representative, and informative URLs to represent the article. Akin to any classical information retrieval problem, we must ensure that we did not collect a list of the URLs resulting in increased recall that introduces "noise" into the algorithm. Noisy data would include URLs that do not provide useful information as to the primary location of the servers that hosts the article online. We modeled a modified form of the term frequency x inverse document frequency (tf-idf), where we divide the document frequency of the URL selection by the log of the inverse URL frequency. However, instead of using terms from the corpus, we are using the unique Internet Protocol address of the server that hosts the article.

Table 3: DBLP Most Prolific Authors

Rank	Name	DBLP	Evans
1	Yu, Philip	535	535
2	Wang, Wei	482	481
3	Chin-Chen Chang	480	469
4	Elisa Bertino	468	466
5	Thomas S. Huang	458	459
6	Edwin R. Hancock	437	438
7	Grzegorz Rozenberg	433	431
8	Sudhakar M. Reddy	431	428
9	Wen Gao	416	416
10	Alberto L. Sangiovanni-Vincentelli	414	415

Tan, Elmacioglu, and Yang were the first to propose URLs, while attempting to determine which websites act as harvesting agents collecting a substantial number of articles (articles, conference proceedings, etc) but were not the primary hosting source for these documents (Tan, Kan and Lee 2006; Elmacioglu, et al. 2007; Yang, et al. 2006). We have adapted a similar approach, but offer additional features to reduce the ambiguities that exist in uniquely identifying a URL. Simply using the URL of an article is insufficient, to a computer (and a casual observer) because although the URLs http://www.acm.org and http://delivery.acm.org look different, both URLS are hosted on the same machine.

We can perform a reverse Domain Name Service (DNS) lookup to produce the IP address of the machine that hosts both of these websites, each of which returns 63.118.7.37. Although the process of reverse DNS resolution is not without issues, we assert that the use of DNS resolution should be sufficient to detect duplicate URLs adequately.

We query the DBLP database for the 100-most prolific authors (see Table 3 for the top ten authors). We then query the database for every paper written by these prolific authors. Table 3 contains the number of articles found in the online DBLP database and in our database implementation using the February version of the dataset. We submit the title of each citation to the Yahoo! Boss search engine (see section 4.2.3). We limit the number of websites returned by the search engine to ten, the first full page of results, and then perform a reverse DNS lookup for each website URL. For each successful DNS resolution, we record the website in the database and then continue to the next item.

Web hosting companies employ a feature called round-robin load balancing where they use of multiple IP address to enable users to view one-of-many servers that host the website and deal with web traffic efficiently. For instance, when we query the IP address of the server that hosts http://www.google.com, it returns three entries for the server's location: 74.125.67.100, 74.125.45.100, and 209.85.171.100. Each of these addresses is valid, displays the Google.com homepage, and is distributed according to their geographic proximity to the computer's IP address.

Resolving the ambiguity involved with load-balanced websites requires querying for the available IP addresses to the reverse DNS lookup call. We order each of these unique IP address numerically and then search the database to locate a previous match. If no matches exist, we add the first address in numerical order to the database. Unlike the referenced articles above, this DNS resolution alteration should reduce some of the noise associated with modern load-balancing practices. We performed 297,952 searches that resulted in 19,770 unique base IP addresses. Finally, for a hostname *h* that has a frequency *f(h)*, we calculated the inverse host frequency to be where:

$$h = \log_2 \frac{\max_h f(h) + 1}{f(h) + 1}$$

Equation 2: Inverse Host Frequency Equation

4.2.3 Collecting Base URL Features

We wanted to use a randomized collection of articles to test the performance of our clustering approach. We used the title for each article in our collection of

random citations from the DBLP database (see section 4.2.1) to submit as a query to the Yahoo search engine via an automated programming interface (API). The Yahoo search returns a listing of the top-ten searches matching our query, including the URL for each site. We resolve each of these URLs to their base components (e.g. http://www.website.edu/path/file.html base URL is www.website.edu), perform a reverse DNS lookup with the website's IP address, and consult our inverse host frequency database to calculate the "rarity" of the server, based on data within our corpus. For each address that does not have an entry in the inverse host frequency table, we assign it an arbitrarily low ranking value of one. We then sort the array by the inverse host frequency, adding "rare" domains first and then save this data to an XML file for later processing by a K-means clustering implementation. We perform several random citation trials consisting of sample datasets of 100, 250, 1000, 2500, and 25,000 unique citations, and two independent samples with 125 articles used for a manual inspection and analysis of the clusters (see Section 5.1).

4.3 Clustering

K-means clustering operates over ten features: the first five authors of an article and the first five base IP addresses of the unique server hosting the article. Our study uses the Oracle Data Miner implementation of the K-means clustering algorithm.

4.3.1 Methods

Using the XML citation datasets described in Section 4.2.3, we import this information into an Oracle database. To ensure that the citation trials remained

unbiased, the only identifiable information we included during the import process was the MySQL internal database primary key, which does not correlate with the journal title. This allowed us to evaluate the performance of the clusters, without influencing the clusters with external document metadata.

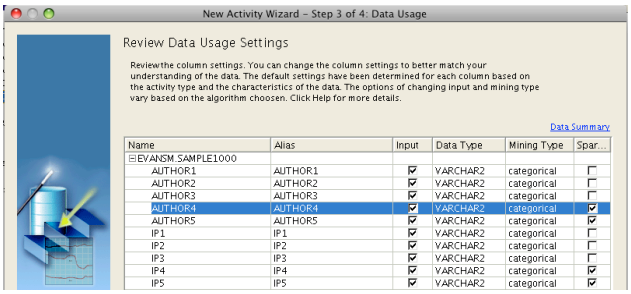

Figure 3: Oracle Feature Selection

4.3.2 Clustering Vectors

As stated in section 4.2.3, we used five different sample datasets consisting of 100, 250, 1000, 2500, and 25,000 unique citations. Since each of the datasets use the same randomized selection process, each datasets should contain the same proportion of citation from each of the sources listed in Table 1. Each of these five datasets included the five authors and the five IP address feature. Attributes "Author4" and "Author5" are labeled as sparse because 20% or more of the data within these two attributes contain null values (see Figure 3), which is not surprising when the mean number of authors within the corpus is 2.54. The following parameters were set in the K-means algorithm: create up-to-100 clusters, use the Euclidean-based distance, and iterate through the dataset 20 times. Finally, we used a size-based split criterion to ensure that one cluster did not dominate during the partitioning phase.

5 Results

We present the clusters produced by the Oracle Data Miner as well as a spatial representation of the overall system's clustering performance. We manually inspected two different samples of 125-articles clustered by the K-means algorithm using a user-specified k-value of fifty (see section 5.1). We review these sample datasets, including encoding representations, data consistency, and observed trends in sections 5.2 and 5.3.

5.1 Manually Evaluated Clusters

The first sample, sample A (see Table 4) used the random sampling method presented in Section 4.2.1 to select the candidate articles. The algorithm created fifty clusters using the clustering features described in Section 4.3.2. We manually examined each of the K-means algorithm's clusters. If the clustering approach worked, each cluster should include a set of articles that have the same journal or conference name.

The values bolded in Table 5 represent an exact match for an author or IP address and values in italics indicate highly likely candidates because of a high level of data similarity. For instance, the IP address 171.64.68.10 and 171.64.75.45 are not the same address, but share many of the same base address characteristics; they are within 65,534 addresses of each other and have the same class B subnet.

Sample A contained 65 articles (54%) from conferences and 56 articles (46%) from journals, two articles were omitted from this sample due to database importing issues. As stated above, we selected the articles in this sample at random,

resulting in seven article pairs that were from the same journal or conference (denoted in orange); the algorithm did not have access to the journal or conference names. The clustering algorithm did not place any of the matching articles in the same cluster. Overall, this sample A contains few articles with highly likely candidate features (less than 12%), but we anticipated this lack of similarity because the remaining articles are drawn different journals and conferences.

Table 4: Sample A - Manual Analysis of Clusters from a Random Set of Journals and Conferences

IP A	*IP B*	*IP C*	*Author A*	*Author B*	*Journal / Conference Name*
165.123.34.126	*130.82.101.38*	128.2.203.164	*j yang*	q li	International Journal of Computer Vision
134.84.135.153	129.97.86.229	**141.51.167.67**	d cuesta-frau	m hern¿¿¿ndez-fenollosa	conf/iciar
193.2.123.5	**136.199.54.125**	*130.82.43.1*	m kannen	m leischner	*1 HMD - Praxis Wirtschaftsinform.
193.194.158.174	64.79.161.47	**136.199.54.125**	j cowie	l oteniya	conf/iceis
130.82.101.132	**141.51.167.67**	**136.199.54.125**	p wognum	r jardim-gon¿¿¿alves	conf/ispe
146.48.87.136	143.167.100.186	94.23.23.174	p clough	a al-maskari	conf/clef
208.69.40.118	64.14.68.65	134.76.74.100	d wishart	*r yang*	In Silico Biology
130.203.133.36	129.130.10.48	128.220.13.101	t amtoft	a banerjee	Sci. Comput. Program.
			c louwrens	s solms	conf/sec
202.41.92.139	128.2.203.164	137.132.80.51	j jouannaud	e kounalis	conf/lics
136.199.55.186	139.6.138.20		s schierholz	e windisch	IWBS Report
157.193.140.25	195.134.65.118	140.98.194.135	a barbieri	g colavolpe	IEEE Transactions on Wireless Communications
132.241.82.63	128.171.194.70	72.246.48.32	m walji	j zhang	*2 conf/hicss
144.124.0.99	158.125.1.136	128.101.35.199	q meng	m lee	Neurocomputing
78.46.52.79	72.246.48.32	137.226.34.227	g gui	h kienle	conf/iwpc
72.9.156.208	150.65.5.208	130.64.20.7	h yoshida	t shigenobu	conf/kes
69.5.195.211	128.220.138.120	98.131.133.89	m roula	a bouridane	conf/isbi
149.132.176.38	159.149.130.205	148.4.2.231	h van	a trentini	conf/seke
152.78.189.29	209.164.14.187	*152.78.68.142*	l wang	t kazmierski	conf/iastedCCS
202.161.41.198	128.125.163.169	128.250.37.111	s nutanong	e tanin	conf/dasfaa

130.203.36.210	130.49.220.23	114.240.114.60	s kim	n vijaykrishnan	conf/islped
146.193.39.30	130.161.37.202	128.32.244.168	j fernandes	m silva	conf/iscas
129.10.32.93	203.159.0.13	129.81.224.51	v deligiannis	s manesis	conf/etfa
140.112.18.7	216.47.152.246	202.141.68.6	a shrivastava	m kumar	conf/vlsid
169.237.114.218	212.201.44.37	128.120.246.26	s park	l linsen	IEEE Trans. Vis. Comput. Graph.
155.98.65.24	132.67.252.100	128.2.203.164	h connamacher	m molloy	conf/focs
152.66.70.16	130.161.254.58	137.132.80.57	g patan¿¿¿¿	m russo	Inf. Sci.
131.114.9.224	209.62.47.32	140.116.82.34	j verd¿¿¿¿	j garc¿¿¿¿a	SIGARCH Computer Architecture News
192.150.18.101	96.7.106.53	171.67.22.33	a laursen	j olkin	conf/compcon
130.237.225.198	78.47.80.59	128.171.224.100	m palm¿¿¿¿r	a naeve	conf/iccs
204.111.14.150	130.235.64.101	165.123.34.126	s crudge	f johnson	*3 JASIST
128.248.155.210	130.126.140.41	128.105.7.26	r iyer	z kalbarczyk	conf/dsn
64.207.133.151	195.82.124.124	158.182.9.1	y fung	c li	conf/iat
128.230.109.13	63.84.220.233	141.217.48.33	w shi	z tang	conf/hipc
210.32.0.229	132.198.19.37	130.233.215.199	g ge	c lam	J. Comb. Theory, Ser. A
78.136.19.25	64.34.197.170	216.87.188.9	c greco		conf/cmg
			j segen		conf/ijcai
			s agostino		Theor. Comput. Sci.
129.107.52.7	130.239.40.24	74.220.219.64	s hegner		conf/pods
193.63.84.78	130.75.87.35	93.93.131.33	w adams		Artif. Intell. Law
12.155.161.151	171.64.73.43	94.23.23.174	d gay	p levis	ACM Trans. Embedded Comput. Syst.
128.9.160.27	209.216.212.21	64.170.98.32	i bisio	m marchese	Computer Networks
209.195.157.80	64.74.98.80	194.9.84.183	d cojocaru	a karlsson	Advances in Engineering Software
208.97.177.125	128.30.76.82	129.64.2.21	d abadi	d carney	conf/sigmod
140.127.112.21	205.178.152.3	137.189.90.239	c huang	j pan	conf/iih-msp
192.103.19.5	64.170.98.32	130.75.87.35	h kim	b oh	conf/cms
144.214.6.167	211.222.57.208	140.126.3.102	j fan	x jia	Algorithmica
129.10.68.74	193.136.138.3	136.165.40.9	s tari	j shah	conf/iccv
132.170.108.1	128.125.4.76	140.98.193.112	n haering	r qian	*4 conf/icmcs
67.205.27.87	129.132.80.110	152.2.131.244	j anderson		Acta Inf.
128.187.48.9	141.142.2.216	171.66.120.77	m ando		IEICE Transactions
155.246.66.29	128.59.48.24	68.181.201.23	w chen	u mitra	conf/icc
69.61.60.58	207.235.4.158	209.237.233.125	h fr¿¿¿¿hlich	m fellmann	Bioinformatics
128.91.40.49	128.135.72.38	130.126.108.21	r ghrist	d koditschek	CoRR
137.226.34.227	128.105.121.60	136.199.54.125	h thoma	h mayr	Angewandte Informatik
155.207.48.20	*136.199.54.125*	*136.199.55.186*	x wang	null	conf/accv
65.61.12.151	128.30.52.51	132.65.16.18	n tishby	null	conf/dis
140.141.2.5	202.161.41.198	132.74.10.59	l agussurja	h lau	conf/iat

208.113.208.243	139.140.14.91	141.218.143.20	l arge	l toma	conf/spaa
128.100.48.11	160.45.137.85	143.239.201.140	j beck	m fox	Artif. Intell.
128.32.48.151	205.157.169.120	203.92.211.161	y chang	y kang	conf/ijcnn
216.10.195.47	38.113.1.102	131.96.101.178	p howard	null	conf/fjcc
137.148.142.62	128.97.92.177	171.67.22.33	w mccartney	n sridhar	conf/sensys
216.245.180.101	66.220.18.178	203.248.159.9	f eisenbrand	f grandoni	*5 conf/icalp
212.67.202.199	128.232.233.16	209.73.187.137	p math¿¿¿¿	null	J. Complexity
141.66.176.200	91.198.174.203	213.212.74.227	a maiti	m maiti	Int. J. Comput. Math.
209.132.201.31	65.214.43.44	78.129.155.6	p chow	w jia	conf/pdpta
128.240.150.127	206.180.225.34	129.97.7.159	j fetzer	null	*6 Commun. ACM
130.65.86.46	208.237.178.123	128.30.52.51	f bry	m kraus	conf/ah
66.33.196.210	160.45.117.200	171.64.74.243	c gr¿¿¿¿pl	h pr¿¿¿¿mel	Discrete Applied Mathematics
130.73.108.4	128.120.246.26	130.203.135.66	x zhang		Int. J. Hum.-Comput. Stud.
207.136.10.135	139.78.113.1	128.178.156.38	p cremonese	s giordano	conf/imsa
193.136.19.20	129.7.174.100	67.18.147.42	¿¿ ribeiro	j fernandes	conf/ecbs
192.18.97.62	192.150.18.101	72.5.124.55	a zamulin		conf/dagstuhl
143.229.6.44	164.76.102.53	143.89.44.246	r krovetz		conf/sigir
128.2.108.203	137.104.129.136	152.2.1.217	b raphael		conf/egice
157.1.32.51	146.176.222.142	*128.200.9.26*	i satoh		Cluster Computing
213.191.194.4	198.65.11.82	143.210.72.22	d guelev		Electr. Notes Theor. Comput. Sci.
146.186.90.90	129.186.52.80	130.232.203.6	t laihonen		Eur. J. Comb.
205.157.169.118	155.245.93.1	130.194.64.145	k jr.		IEEE Computer
208.68.167.134	67.196.156.31	128.112.132.86	j warner		*3 JASIST
170.149.173.130	96.7.97.62	66.235.120.98	v agrawal		J. Electronic Testing
69.20.66.162	*69.20.70.239*	169.145.6.65	j jeng		conf/wecwis
81.169.145.86	69.49.101.51	207.173.206.25	a prior		J. Symb. Log.
131.130.1.78	128.84.158.74	128.135.11.125	v plisko		Ann. Pure Appl. Logic
64.13.192.193	209.34.241.68	198.45.25.111	e prabhakar		conf/lisa
128.200.64.26	192.88.209.244	129.7.240.35	c kahn		conf/nspw
216.235.79.13	64.191.203.30	8.5.0.172	r charette		*6 Commun. ACM
134.2.14.42	192.150.186.14	128.119.240.19	a rosenberg		conf/awoc
141.51.167.67	202.96.51.220	136.199.54.125	t schraml	e schoop	*1 HMD - Praxis Wirtschaftsinform.
80.82.137.233	206.131.241.137	*171.66.120.76*	t suda	y yemini	Computer Communications
128.100.11.60	128.135.8.186	137.189.97.85	m ashihara	s abe	conf/icann
66.255.97.26	64.225.158.79	62.128.138.93	v vatsa	s sural	conf/iciss
125.141.224.207	128.255.44.51	171.64.22.133	z manna	r waldinger	*5 conf/icalp
70.87.146.55	209.133.21.164	140.185.15.228	f henley	h choi	conf/itc
204.152.149.5	128.119.244.5	124.16.137.58	d lime	o roux	conf/rtss
96.7.107.9	147.65.1.22	193.63.84.78	l velho	j gomes	conf/sibgrapi

141.211.144.188	209.237.233.125	*171.66.120.20*	k li	b sanctuary	Journal of Chemical Information and Computer Sciences
208.70.245.189	65.79.173.117	128.197.153.21	m jafari	m zhou	IEEE Transactions on Systems, Man, and Cybernetics, Part B
128.195.1.83	63.252.83.127	130.65.150.51	t kunert	h kr¿¿¿¿mker	conf/hci
209.132.230.51	131.215.229.145	76.12.178.82	p dechpichai	p davy	conf/aiprf
69.5.195.195	208.113.243.93	130.159.187.223	n shulman	m feder	IEEE Transactions on Information Theory
192.12.69.4	66.218.77.68	128.148.32.110	c friedrich	m houle	conf/gd
134.117.27.24	192.20.225.32	128.205.32.53	s akl	b bhattacharya	*7 Parallel Algorithms Appl.
193.136.166.90	129.120.87.240	128.192.251.7	a kishimoto	n sturtevant	conf/atal
91.210.88.245	143.106.12.174	207.56.179.232	r carr	w hart	conf/gecco
65.55.194.74	149.169.31.10	206.130.107.51	c lucarz	m mattavelli	*4 conf/icmcs
140.208.31.101	63.118.7.100	217.115.194.84	b bergen	g wellein	*7 Parallel Algorithms Appl.
137.45.3.1	128.238.24.12	93.93.131.33	j chase	e oakes	conf/sigcse
192.87.172.73	156.56.94.2	141.58.125.71	v huynh	m ryoke	conf/ifsa
216.218.185.154	174.36.28.11	210.150.254.122	z wu	h li	conf/aina
128.194.146.101	128.174.244.220	128.220.13.101	d challou	d boley	conf/icra
128.32.192.116	129.49.108.11	128.196.27.130	m dror	y lee	Int. J. Comput. Geometry Appl.
130.237.32.143	130.243.85.68	129.88.43.46	m t¿¿¿¿rngren	d chen	conf/euromicro
129.174.69.30	*129.174.1.15*	*129.174.1.13*	r michalski	l kerschberg	J. Intell. Inf. Syst.
12.180.48.226	134.214.142.10	198.81.200.2	c grasset-simon	g damiand	Pattern Recognition
64.46.130.10	*140.98.194.139*	89.31.1.164	e falletti	f sellone	Wireless Communications and Mobile Computing
137.110.119.52	75.126.86.8	*140.98.194.146*	t pande	d love	conf/globecom
131.204.2.251	148.129.75.8	128.197.153.21	p balasubramanian	g wyner	*2 conf/hicss
133.25.90.34	192.1.100.20	114.240.114.60	r huang	j ma	conf/ispan
128.111.221.123	128.120.246.26	206.131.241.137	f jr.	m kubo	conf/vr

The function of the second sample, sample B (see Table 5) was to measure how well the K-means clustering approach partitioned articles from the sample journal. The algorithm, similar to sample A, created fifty-clusters using the clustering features described in Section 4.3.2. The only alteration from sample A

was that we selected the articles at random from the journal Parallel Computing. Again, we performed a manual inspection of each of the final clusters, and indicated exact matches from an author or IP address in bold and likely features used for clustering in italic. Seven different article pairs exists from the same journal or conference, the mappings are as follows:

*1 HMD - Praxis Wirtschaftsinform
*2 Hawaii International Conference on System Sciences
*3 Journal of the American Society for Information Science and Technology
*4 International Conference on Multimedia Computing and Systems
*5 International Colloquium on Automata, Languages and Programming
*6 Communications of the ACM
*7 Parallel Algorithms and Applications

The question marks in the author or journal and conference names indicate a Unicode conversion error during the data pre-processing phase (see Figure 4 for example and refer to Section 6.1.1 for implications).

```
<incollection mdate="2003-12-04" key="books/idea/siau2003/TrujilloLS03">
<author>Juan Trujillo</author>
<author>Sergio Luj&#225;n-Mora</author>
<author>Il-Yeol Song</author>
<title>Applying UML For Designing Multidimensional Databases And OLAP
Applications.</title>
<pages>13-36</pages>
<year>2003</year>
<booktitle>Advanced Topics in Database Research, Vol. 2</booktitle>
<url>db/books/collections/Siau2003.html#TrujilloLS03</url>
</incollection>
```

Figure 4: Example of Entity Encoding Issues

Sample B created fewer one article clusters (42% of all clusters) than Sample A, which had more individual article clusters (60%). Sample B also produced 18-exact matches used to assign articles into the same clusters, as opposed to only 2-exact matches in Sample A, because articles in Sample B are more likely to be written by the same author. In contrast to Sample A, seven clusters in sample B contain articles on related topics.

Table 5: Sample B - Manual Analysis of Clusters from Same Journal

% Prob	*IP Address A*	*IP Address B*	*IP Address C*	*Author A*	*Author B*
0.9928	129.118.162.211	130.39.186.110	171.66.120.125	m balduccini	e pontelli
0.9853	87.236.232.147	146.164.34.2	128.180.120.39	m angelaccio	m colajanni
0.9927	130.192.9.200	193.63.84.78	202.96.51.220	c anglano	c casetti
0.9949	130.82.101.132	198.81.200.2	74.220.219.64	m ashworth	f foelkel
0.9841	141.51.167.67	**136.199.54.125**	null	e adamides	p tsalides
0.9887	171.66.120.79	12.104.88.66	137.132.80.51	s akl	k qiu
0.9873	130.15.1.11	129.177.16.249	155.101.98.136	s akl	h schmeck
0.9826	209.73.219.100	128.183.61.67	131.215.145.41	g aloisio	m cafaro
0.9942	150.65.5.208	**136.199.54.125**	69.93.12.187	t altman	y igarashi
0.9960	195.221.162.126	**129.175.15.11**	*171.64.68.10*	e bampis	c delorme
0.9961	130.251.61.252	194.42.16.25	128.59.18.180	f ancona	s rovetta
0.9923	209.132.213.141	128.111.234.156	141.217.43.45	e aarts	j korst
0.9916	137.151.27.1	**198.81.200.2**	*171.64.75.45*	m bessenrodt-weberpals	h weberpals
0.9990	74.54.1.132	132.208.138.223	**129.175.15.11**	j allouche	f haeseler
0.9923	129.12.4.59	128.192.251.7	**198.81.200.2**	g bader	e gehrke
0.9987	131.175.1.159	193.51.208.78	**212.189.136.200**	w andreoni	a curioni
0.9963	130.34.184.58	142.137.245.69	146.87.255.31	h abbas	m bayoumi
0.9962	38.108.68.66	65.79.173.117	155.69.254.74	m aref	m tayyib
0.9998	130.104.62.18	129.2.56.181	**212.189.136.200**	a averbuch	l ioffe
0.9981	144.174.16.100	**160.36.58.108**	141.51.167.67	r ii	l storc
0.9962	146.83.7.3	131.155.70.190	**160.36.58.108**	e ch.	m kiwi
0.9960	91.198.174.203	208.113.243.93	128.30.2.79	m bahi	j miellou
0.9987	129.127.43.96	129.12.4.59	128.186.122.19	s aluru	g prabhu
0.9994	128.174.231.193	128.227.205.212	160.36.58.108	p amestoy	i duff
0.9983	128.178.159.110	128.101.190.11	128.55.6.34	r ii	l storc
0.9985	150.214.108.158	62.108.136.30	62.128.138.93	e alba	g luque
0.9980	140.172.12.69	*140.221.9.215*	*140.221.9.85*	c baillie	j michalakes
0.9999	*129.34.20.108*	128.174.252.84	*129.34.20.3*	v bala	j bruck
0.9989	130.207.222.95	134.88.14.211	155.101.98.136	e bampis	j k√∂nig
0.9998	18.85.45.88	131.114.3.18	204.14.91.24	b bacci	m danelutto
0.9995	128.138.249.54	129.119.70.169	128.101.35.204	m amer	b abdel-hamida
0.9980	128.149.128.145	128.148.160.10	128.174.231.193	d balsara	c norton
1.0000	171.64.163.184	130.238.168.34	193.146.115.82	p aumann	h barnewitz
0.9969	195.176.176.154	74.125.45.137	147.96.1.15	s bandini	m magagnini
0.9999	12.176.28.10	204.121.6.21	128.172.10.65	**s bandini**	**g mauri**
0.9999	209.132.213.141	140.177.205.55	141.211.189.46	**s bandini**	**g mauri**
0.9999	130.37.20.20	193.48.96.20	140.221.9.85	o aumage	l boug√?
0.9996	216.146.212.152	147.83.30.101	128.6.4.24	e ayguad√?	j garcia
0.9983	128.197.15.10	130.207.7.208	128.2.203.164	h azaria	y elovici
0.9994	132.68.115.2	160.36.58.108	155.101.98.136	j andersen	g mitra

0.9999	129.125.14.65	152.1.24.47	128.193.4.112	a attanasio	j cordeau
0.9995	152.81.144.29	193.136.28.36	194.9.84.183	j bahi	s contassot-vivier
0.9999	130.216.27.139	193.255.135.33	128.101.168.25	a amoura	e bampis
0.9995	66.39.124.177	63.166.183.125	72.44.51.239	d audet	y savaria
0.9994	66.207.207.52	209.61.228.48	128.178.33.38	i ahmad	y he
1.0000	130.18.14.28	130.18.208.30	128.174.239.11	i banicescu	r cari√?o
0.9997	130.64.1.83	153.106.4.23	134.76.74.100	m atiquzzaman	p srimani
1.0000	128.175.14.182	128.174.231.193	128.200.85.19	a averbuch	e gabber
0.9999	130.74.120.3	129.72.2.182	150.214.109.5	a averbuch	m israeli
1.0000	164.67.86.89	204.134.131.27	129.177.16.246	s altekar	a ray
0.9998	128.150.4.107	160.91.4.41	142.58.111.32	p altevogt	a linke
0.9995	17.112.152.32	198.9.3.30	131.243.2.154	m ashworth	a lyne
1.0000	*131.193.181.116*	*131.193.78.84*	194.9.84.183	w allcock	j bester
0.9999	144.37.1.95	209.143.129.164	144.202.252.20	m aboelaze	d lee
0.9999	87.236.232.169	72.5.124.61	132.239.51.65	m atiquzzaman	m banat
0.9995	128.83.68.3	213.52.141.23	128.252.153.11	z du	f lin
1.0000	192.43.228.130	192.5.53.208	156.56.104.10	g antoniu	l boug√?
1.0000	94.124.120.11	165.123.34.126	128.59.66.9	i ahmad	s akramullah
0.9997	66.84.34.170	205.155.65.42	129.7.240.35	i ahmad	m dhodhi
1.0000	96.7.103.107	128.172.12.202	128.46.154.95	g balboni	g cabodi
0.9997	137.151.45.6	134.197.40.3	160.36.56.64	p amodio	l brugnano
0.9999	194.81.203.9	137.151.45.6	**193.136.28.36**	m angelaccio	m colajanni
0.9999	131.215.105.115	149.28.120.34	**193.136.28.36**	e babolian	l delves
0.9999	**130.161.210.5**	129.177.16.246	128.101.191.158	p amodio	n mastronardi
0.9999	**130.161.210.5**	128.100.4.14	13.1.64.42	j ag√?√≠	j jim√?nez
1.0000	132.175.81.3	129.132.46.11	132.175.81.4	**p arbenz**	m becka
0.9999	131.193.32.20	165.112.6.70	131.123.41.85	**p arbenz**	w gander
1.0000	129.12.4.59	212.138.39.90	152.78.68.142	**c askew**	**d carpenter**
0.9995	128.255.45.58	202.141.25.100	129.244.40.44	**j allwright**	**d carpenter**
1.0000	155.247.166.60	128.59.66.9	132.66.48.13	c arapis	s gibbs
0.9998	202.57.163.117	212.189.136.200	147.96.1.15	j baker	m shirel
0.9889	146.164.34.2	137.151.45.6	160.36.58.108	g alaghband	
0.9997	**129.72.2.182**	150.214.109.5	**141.51.167.67**	m alef	
0.9998	**129.72.2.182**	128.36.229.30	**141.51.167.67**	m alef	
1.0000	128.227.74.66	140.221.8.232	192.20.225.32	r aiex	s binato
1.0000	169.229.131.81	155.98.27.201	144.214.130.198	g al-rawi	j cioffi
1.0000	130.209.240.1	171.66.122.240	193.63.84.78	j al-sadi	k day
0.9966	151.189.20.30	141.211.144.27	132.68.32.15	i bar-on	
0.9982	159.226.92.9	155.247.28.2	81.19.179.36	z bai	
0.9983	130.39.187.21	134.193.2.78	147.96.1.15	m alsuwaiyel	
1.0000	129.63.176.210	130.104.62.18	128.6.29.77	r aggarwal	d dellwo
1.0000	62.108.136.30	157.182.209.202	64.225.158.79	a awan	r ferreira
1.0000	74.208.30.134	131.193.181.116	128.183.61.67	a ananthanarayan	r balachandran
1.0000	147.96.1.15	132.175.81.4	208.215.179.146	**r aversa**	**b martino**

1.0000	128.55.6.34	160.91.4.41	194.9.84.183	**r aversa**	**b martino**
1.0000	129.6.13.40	216.47.152.246	198.82.184.164	m alabdulkareem	s lakshmivarahan
1.0000	144.174.16.100	192.18.99.187	129.24.24.13	m ayed	j gaudiot
1.0000	64.82.97.56	128.6.68.133	195.83.132.161	d baz	
1.0000	193.52.245.34	193.230.3.106	194.225.73.180	h ahrabian	a nowzari-dalini
1.0000	132.194.10.4	141.217.48.33	208.110.160.59	j annot	
1.0000	150.214.108.33	150.214.108.158	130.82.101.38	e alba	f luna
1.0000	128.42.17.41	69.72.138.172	165.124.180.106	h amman	
1.0000	129.24.244.30	17.254.2.129	141.142.2.216	o axelsson	v eijkhout
1.0000	198.128.246.10	212.189.136.200	160.91.4.41	i ahmad	
1.0000	128.95.22.12	89.105.124.116	129.177.16.246	l adams	e ong
1.0000	64.202.163.202	130.237.232.226	198.128.246.10	g almasi	
1.0000	131.120.251.40	128.32.31.195	129.128.206.32	z baolin	l wenzhi
1.0000	65.79.173.117	198.82.185.31	128.101.35.207	g alaghband	
1.0000	131.202.244.5	128.9.176.20	129.115.28.4	w amme	e zehendner
0.9843	137.158.59.4	209.217.33.166	137.222.102.8	s bangay	j gain
0.9775	143.84.24.63	206.210.75.203	202.141.25.96	v annamalai	c krishnamoorthy
0.9638	129.59.1.212	140.177.205.52	67.18.199.2	e adamides	p tsalides
0.9890	134.245.248.200	152.78.189.29	129.12.4.59	c addison	v getov
0.9842	147.251.3.47	128.42.205.122	152.2.1.217	r aversa	a mazzeo
0.9773	208.109.122.176	130.37.20.20	128.174.239.11	f arbab	p ciancarini
0.9680	67.192.251.145	216.128.29.26	209.40.98.58	g almasi	g paul
0.9680	128.32.63.27	128.148.32.110	129.110.10.36	h alnuweiri	v prasanna
0.9680	209.21.91.170	63.118.7.17	130.126.139.25	n bahoshy	d evans
0.9653	149.28.120.34	129.177.16.246	203.255.181.238	c baillie	g pawley
0.9731	128.83.68.134	129.114.58.17	129.177.16.249	m baker	k bowler
1.0000	130.245.142.129	132.177.4.32	192.101.104.50	c baillie	
0.9863	128.248.155.51	192.138.151.104	129.127.43.96	j bakker	
0.9865	130.238.168.34	64.170.98.32	63.84.220.237	b arafeh	
0.9824	212.189.136.200	128.197.26.35	142.58.111.32	d barth	
0.9658	192.48.178.165	128.30.2.140	132.68.32.15	p arbenz	
0.9865	128.156.250.69	74.205.45.163	129.6.13.90	d banks	
0.9854	128.4.10.31	81.252.67.151	129.237.125.27	j feo	
0.9865	192.203.218.58	136.142.82.188	130.126.142.6	h barada	
0.9847	129.64.2.21	198.81.200.2	128.180.120.39	s agostino	
0.9650	192.5.53.208	209.242.166.3	129.34.20.3	t axelrod	
0.9865	76.12.178.82	96.7.100.187	129.7.240.35	m dow	

Table 6 summary statistics show that on average, clusters from sample A (a random selection of articles) and B (articles from the same journal) should contain 2.42 articles. However, the mode indicates that a majority of clusters from both samples

A and B have only one article. The maximum measure shows that the clustering algorithm created a large cluster for both sample, with Sample B having fewer articles in its largest cluster (11 articles) than Sample A (15 articles).

Table 6: Manual Analysis Summary Statistics

Sample Name	Min	Max	Mode	Average	Standard Deviation
A: Random Selection	1	15	1	2.42	3.40
B: Same Journal	1	11	1	2.42	2.22

5.2 Varying the number of articles

The two samples described in section 5.1 indicate how the clustering algorithm performs on a small dataset with a low K-value We performed five trials of K-means clustering using the Oracle 10g Data Miner software package with a sample size of 100, 250, 1000, 2500, and 25,000 articles. Our goal in these experiments is to observe the performance of the K-means clusters as we increase the value of k and the number of input articles. Table 7 shows the statistical distribution for each of the five different samples. The table contains information about the following clustering facets:

- Sample Size - The number of articles retrieved from the Citation and XML selection process (see section 4.2.1 for selection process). Cases - The actual number of articles clustered. There are several reasons why the clustering algorithm would not add an article, including that some articles have incomplete or missing data (e.g. a missing IP address for the server hosting the article) or the article citation contains too much sparse data resulting in the exclusion of the article from the clustering process.

- Leaves - The final number of clusters produced. This number indicates the highest number of clusters extending from the root node.
- Min - The smallest number of articles within a cluster.
- Max - The largest number of articles within a cluster.
- Mode - The middle number of articles within a cluster if the clusters were sorted in ascending order by size.
- Average - The average number of articles that were partitioned into most clusters.
- Standard Deviation - This metric is based on the average number of articles partitioned into a cluster. This produces a good evaluative metric to show the average quality of the nodes within the clusters.

Table 7: Cluster Performance Metrics for K = 100

Sample Size	Cases	Leaves	Min	Max	Mode	Average	Standard Deviation
100	96	96	1	1	1	1	0
250	243	100	1	27	1	2.43	4.6
1000	969	100	2	142	2	9.69	24.34
2500	2398	100	3	314	5	23.98	61.5
25000	23,947	100	27	4010	39	239.47	654.6

Three of the five trials (1000, 2500, and 25,000) produce minimum clusters with a low number of similar articles, containing between 2 - 27 articles (0.2% - .16% of all articles, respectively). This suggests that articles in smaller clusters have little or no geometric similarity. These minimum clusters include 0.1% to 0.4% (excluding the 100-article sample) of the entire dataset. In the sample of 100

articles, it is understandable that the maximum case size would be one article per cluster since the sample size is 100 articles with a k-value of 100, out of a random selection of over one million articles. However, as the sample size increases so does the maximum number of articles per cluster. Samples with 250 - 25,000 articles have 10.8% - 16.0% of all of the articles clustered within the largest cluster, respectively. This indicates that the algorithm had problems determining feature characteristics that would allow it to create much smaller and thus similar clusters.

On average, each cluster should contain approximately 1% of the total number of articles in the sample, since the algorithm partitioned the data into approximately one hundred clusters. However, the mode suggests that the majority of clusters account for 0.16% - 0.4% (excluding the 100-data item sample) of the total number of articles in the samples. This measure indicates that the data is skewed towards the largest clusters. The clusters show an unequal partitioning of articles, indicating that the features used to cluster the articles, first five authors and first five base IP address, lack internal similarity.

As shown in Table 7, no dataset performed exceptionally well. Using the values for the mode and maximum number of elements partitioned within a cluster coupled the average and standard deviation measures of all of the clusters provides a statistical representation of how the data skews towards a small number of very large clusters. These large clusters represent greater than 10% of all of the data within the sample. For example, in our sample of 2,500 articles, the average number of articles per cluster is approximately 24 with a mode of five. However, the middle clusters accounts for only five articles, but with an average of 24 articles per cluster

and the largest cluster containing well over 300 articles, the data shows the articles are unbalanced towards the largest clusters. The standard deviation of 61.5 further emphasizes this variation. This pattern of poorly distributed clusters is present in the all of the samples, excluding the 100-elements sample, and suggests poor K-means clustering performance or that the features selected are inadequate to reflect the journal or conference title.

5.3 Visualizing clustering performance

The K-means clustering algorithm partitions the dataset based on commonalities between the article features. Table 7 shows the statistical features represented in the datasets that contain sample sizes of 100, 2500, and 25,000 articles. We provide two views into the same dataset: one view shows the data partitioned with a K-value of 10 (Figure 5, Figure 7, and Figure 9), the other (Figure 6, Figure 8, and Figure 10) show the same dataset partitioned with a K-value of 100. The graphs are color-coded, where each color represents one unique feature from the dataset: the first and second author and the first three IP address of the server hosting the article. The features not included in this analysis do not affect the overall clustering performance or data representation. We provide the graphical representations of samples with a K-value of 10 to show the overall distribution of the articles in the samples and feature similarities.

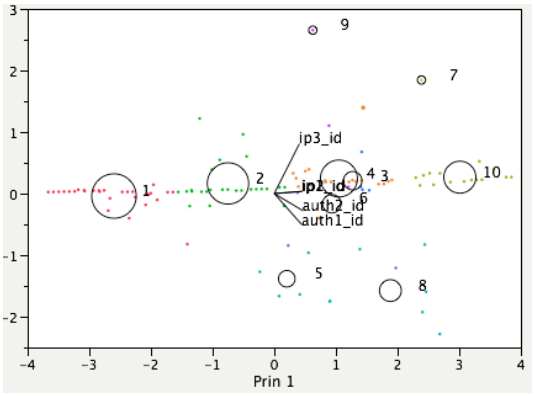

Figure 5: Sample 100 elements with K = 10

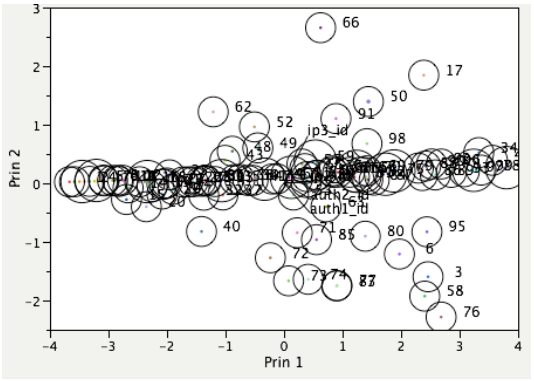

Figure 6: Sample 100 elements with K = 100

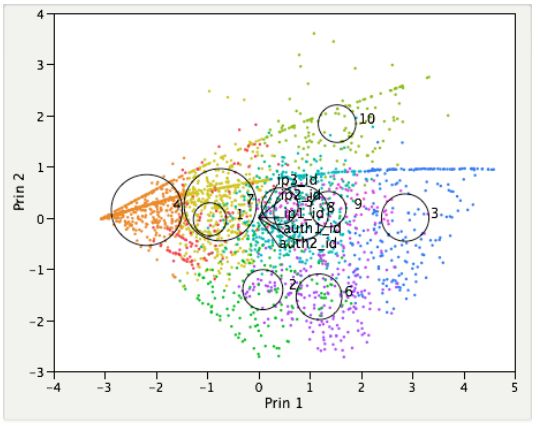

Figure 7: Sample 2,500 elements with K = 10

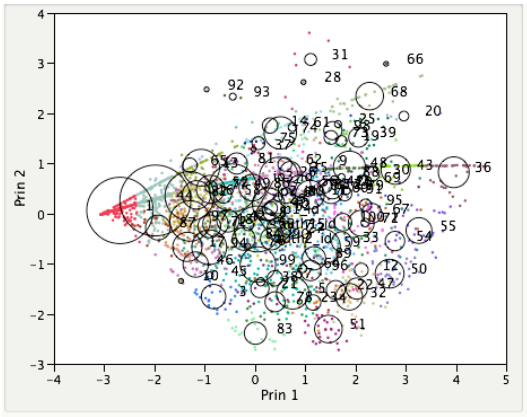

Figure 8: Sample 2,500 elements with K = 100

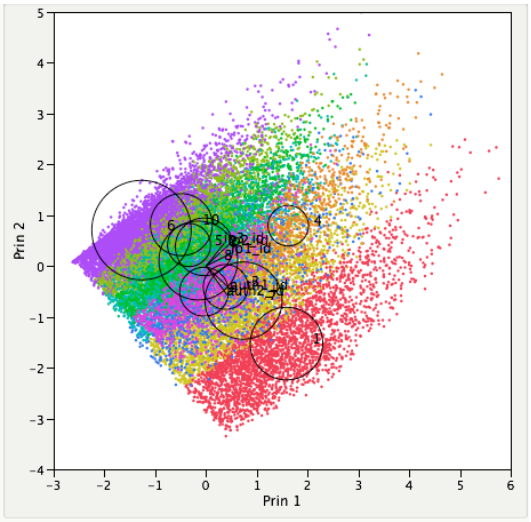

Figure 9: Sample 25,000 elements with K = 10

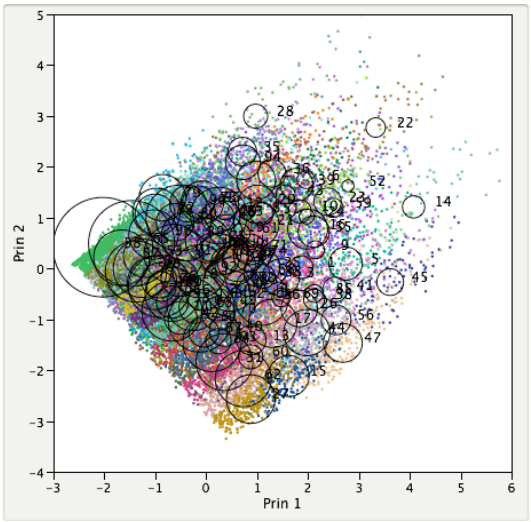

Figure 10: Sample 25,000 elements with K = 100

6 Discussion

Using a small number of articles, less than 1500, can present a challenge to the k-means clustering due to the low probability that articles will be drawn from the same journal or conference. However, selecting two samples with 125-articles each enabled us to increase our understanding of clustering performance through manual inspection. The data suggests that the K-means algorithm does not partition articles into clusters based on the title of the journal or name of the conference that the article was published. However, the resulting clusters do suggest areas where the K-means partitioning did correctly cluster a subset of articles and provides indications where future research can improve the clustering feature selection and general clustering techniques.

In **Figure 5** with a user-defined maximum partition of 10-clusters, there are four larger clusters (clusters 1, 2, 3, and 10) that do not contain more than one feature, but cluster 3 contains articles partitioned using more than one clustering feature. The colors of the data points represent unique features, such as Author1, IPAddress1, etc. As we increase the number of articles beyond 1,500 articles, the clustering algorithm uses more than one feature (clusters 1, 2, 7, and 8 in **Figure 7** and 6, 5, 7, 8, and 10 in **Figure 9**). Visual inspection suggests that the distribution of clusters remain relatively unchanged as we increase the value of K from ten to one hundred.

The samples that we manually inspected (see Table 4 and Table 5) suggest that the clustering features IP Address C (an IP address with a lower relevancy ranking as provided by the search engine) followed by Author A, IP Address A, IP

Address B, and Author B are the most likely features that determines how the articles are partitioned. The features Author A and Author B, first and second author in article citation reference, were most representative in correctly clustering the articles by the correct journal and conference names. An empty Author B appears to be responsible for cluster 99 in Table 5

6.1.1 Limitations and Future Work

One possible source of error is character translations, such as string case conversions that were not Unicode sensitive which may have introduced non-ASCII information conversion errors. Although the search engine supported Unicode queries, the conversion process included only those search results with the same converted Unicode representations. Translation errors may explain why some searches produce no results, despite our expectation that that an entry exists in at least one source on the Internet, namely the DBLP database.

Another potential problem is from the initial paper selection process. The majority of articles, 706,697 (60.47%), are from the 2,586 conference proceedings, whereas only 439,171 (37.58%) are from the 713 published journals. Conference proceedings occur frequently (see Table 1), often annually, and the conference name may change to indicate the year or theme of the conference. This makes the citation less reliable, resulting in an inability to retrieve a sufficiently large corpus of unique articles for a particular conference.

We used the Internet Protocol (IP) address of the source publication website as one of the features in the clustering algorithm. The Internet does not have a system to enforce how a domain owner chooses their domain name. To reduce

some of the associated ambiguity, we rely on the IP address instead of the URL, however, we did not truncate the name of the server that hosts the article before the reverse DNS lookup. Our concern is that there is no guarantee that the base URL has a web addressable IP address. For instance, when presented with the address http://papers.published.com, if we further truncate the URL "papers.published.com" to its base domain of "published.com", we have no guarantee that published.com has a valid IP address. Sub-domains with different IP addresses would also be problematic.

We also would like to investigate if using only the primary author increases clustering performance. We should also exclude documents from the database that are websites, patents, and other extraneous documents that are not articles from journals. Applying simple author disambiguation techniques might reduce partitioning errors during the clustering process due to the lack of any name authority record control for authors and co-authors.

7 Conclusions

Bibliometric research relies on accurate citations. The name of the journal or conference in which an article is published is one of the most important features used in bibliometrics. Journal or conference names that are inaccurate or ambiguous result in errors in citation analysis. We have presented a system that combines Internet-based document surrogates and the first five author names with K-means clustering to disambiguate the name of journal or conference. The system weighs the URL based on an inverse host frequency index that uses the most prolific

authors in the field.

To evaluate the quality of this approach we collected 1.18 million citations from the Digital Bibliography & Library Project (DBLP) data during February 2009. A descriptive analysis revealed that 60.86% of authors disseminate their work in conferences compared with 37.58% in journals. The number of authors per paper varied widely between one and 115, with an average of 2.54.

We conducted a comprehensive manual analysis of clusters produced with two samples of 125-articles. The first sample included a set of articles selected at random with no clusters containing an article from the same journal or conference. The majority (60%) of all clusters contained only one article resulting in the creation of a few large clusters having up to 15 articles. The k-means algorithm did not place any of the seven article pairs from the same venue in the same cluster. The second sample considered articles selected from a single journal. This sample created seven clusters of articles by the same author or co-author. This sample had fewer one article clusters (42%) than the first sample.

We conducted experiments to improve the K-means clustering performance, by varying the number of samples from 100, 250, 1000, 2500, and 25,000 articles (see Section 5.2). The results show that the articles were not evenly distributed among the clusters and that the algorithm assigned a large number of articles to a small number of clusters.

Our results suggest that additional features are required to disambiguate journal and conference names accurately. As more than 60% of the DBLP articles are published at conferences future disambiguation efforts should focus on

conference names. Such work is critical to support future bibliometric analyses.

8 References

Baker, Kathryn L, Alexander M Franz, Pamela W Jordan, Teruko Mitamura, and Eric H Nyberg. "Coping With Ambiguity in a Large-Scale Machine Translation System." *Fifteenth International Conference on Computational Linguistics (COLING-1994).* Kyoto, Japan, 1994. 90-94.

Brill, Eric, and Philip Resnik. "A Rule-Based Approach to Prepositional Phrase Attachment Disambiguation." *Fifteenth International Conference on Computational Linguistics (COLING-1994).* Kyoto, Japan, 1994.

Broadus, R. N. "Toward a definition of “bibliometrics”." *Scientometrics* (Akadémiai Kiadó, co-published with Springer Science and Business Media B.V.) 12, no. 5-6 (November 1987): 373 - 379.

De Solla Price, Derek. "A general theory of bibliometric and other cumulative advantage processes." *Journal of the American Society for Information Science* (Wiley Periodicals, Inc.) 27, no. 5 (1976): 292-306.

Elmacioglu, Ergin, Yee Fan Tan, Su Yan, Min-Yen Kan, and Dongwon Lee. "PSNUS: Web People Name Disambiguation by Simple Clustering with Rich Features." *4th Int’l Workshop on Semantic Evaluation (SemEval).* Association for Computational Linguists, 2007.

Garfield, Eugene. "Can Citation Indexing Be Automated?" *Statsitical Assocristiora Methods for Mechamked Documentation.* Washington: National Bureau of Standards Miscellaneous Publication, 1964. 189-192.

Garfield, Eugene. "Citation Analysis as a Tool in Journal Evaluation: Journals can be ranked by frequency and impact of citations for science policy studies." *Science* 178 (November 1972): 471 - 479.

Garfield, Eugene. ""“Science Citation Index”- A New Dimension in Indexing." (Science) May 1964: 649-6S4.

Gideon, Mann S, and David Yarowsky. "Unsupervised personal name disambiguation." *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003.* Edmonton, Canada: Association for Computational Linguistics, 2003. 33 - 40.

Han, Hui, Hongyuan Zha, and C Lee Giles. "Name disambiguation in author citations using a K-way spectral clustering method." *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries.* Denver, CO: ACM, 2005. 334 - 343.

Han, Hui, Wei Xu, Hongyuan Zha, and C. Lee Giles. "A Hierarchical Naive Bayes Mixture Model for Name Disambiguation in Author Citations." *Symposium on Applied Computing (SAC '05).* Santa Fe, New Mexico: ACM, 2005. 1065 - 1069.

Hauptly, Denis. "Using Bibliometrics: A guide to evaluating research performance with citation data." *Knowledge Link Newsletter*, July 1, 2008: 1 - 11.

Jaffri, Afraz, Hugh Glaser, and Ian Millard. "URI Disambiguation in the Context of Linked Data." *In Proceedings of the 1st Workshop on Linked Data on the Web at WWW2008.* Beijing, China, 2008.

Kan, Min-Yen, and Yee Fan Tan. "Record Matching in Digital Library Metadata." *Communications of the ACM* (ACM) 51, no. 2 (February 2008): 91 - 94.

Kessler, M M. "Bibliographic Coupling Between Scientific Papers." *American Documentation* (The American Documentation Institute) 14 (January 1963): 10 - 25.

Kuhn, Thomas. *The Structure of Scientific Revolutions.* 2nd edition, with postscript. Chicago, IL: University of Chicago Press, 1970.

Lawrie, Dawn, W Bruce Croft, and Arnold Rosenberg. "Finding Topic Words for Hierarchical Summarization." *In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01).* ACM, 2001. 349 - 357.

Lee, Dongwon, Byung-Won Oh, Jaewoo Kang, and Sanghyun Park. "Effective and Scalable Solutions for Mixed and Split Citation Problems in Digital Libraries." *2nd international workshop on Information quality in information systems.* New York, USA: IQIS, 2005. 69 - 76.

McRae-Spencer, Duncan M, and Nigel R Shadbolt. "Also By The Same Author: AKTiveAuthor, a Citation Graph Approach to Name Disambiguation." *Joint Conference on Digital Libraries (JCDL 2006).* Chapel Hill, North Carolina: ACM, 2006. 53 - 54.

Merton, Robert King. *On the shoulders of giants: a Shandean postscript.* Reprint. New York, New York: University of Chicago Press, 1993.

Metzler, Donald, and W Bruce Croft. "A Markov Random Field Model for Term Dependencies." *SIGIR.* Salvador, Brazil: ACM, 2005. 472 - 479.

Morris, Steven A, and Betsy Van der Veer Martens. *Mapping Research Specialties.* Vol. 42, in *Annual Review of Information Science and Technology*, by Blaise Cronin, 213 - 293. Medford, New Jersey, 2008.

Pritchard, Alan. "Statistical Bibliography or Bibliometric?" *Journal of Documentation* (Aslib) 25, no. 4 (1969): 348 - 349.

Rosen-Zvi, Michael, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. "The Author-Topic Model for Authors and Documents." *Proceedings of the 20th UAI Conference.* 2004. 487 - 494.

Schildt, Henri A., Shaker A. Zahra, and Antti Sillanpää. "Scholarly Communities in Entrepreneurship Research: A Co-Citation Analysis." *Entrepreneurship Theory and Practice* (Baylor University) 30, no. 3 (2006): 399-415.

Small, Henry. "Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents." *Jounal of the American Society for Information Science* 23, no. 4 (1973): 265 - 269.

Small, Henry. "Paradigms, Citations, and Maps of Science: A Personal History." *Journal of the American Society for Information Science and Technology* 54, no. 5 (2003): 394 - 399.

Smith, Aida Marissa. "An examination of PubMed's ability to disambiguate subject queries and journal title queries." *Journal of the Medical Library Association* (Medical Library Association), January 2004: 97 - 100.

Tan, Yee Fan, Min-Yen Kan, and Dongwon Lee. "Search Engine Driven Author Disambiguation." *Joint Conference on Digital Libraries.* Chapel Hill, USA: ACM, 2006. 314-315.

Thomas Reuters. *ISI Web of Knowledge Fact Sheet.* November 3, 2008. http://www.thomsonreuters.com/content/PDF/scientific/Web_of_Knowledge_factsheet.pdf (accessed March 13, 2009).

Torvik, Vetle I, Marc Weeber, Don R Swanson, and Neil R Smalheiser. "A Probabilistic Similarity Metric for Medline Records: A Model for Author Name Disambiguation ." *Jounal of the American Society for Information Science and Technology* 56, no. 2 (2005): 140-158.

Turnbull, Don. *Bibliometrics and the World Wide Web.* January 01, 1998. http://www.ischool.utexas.edu/~donturn/research/bibweb.html (accessed February 10, 2009).

Yang, Kai-Hsiang, Jian-Yi Jiang, Hahn-Ming Lee, and Jan-Ming Ho. "Extracting Citation Relationships from Web Documents for Author Disambiguation."

Technical Report No.TR-IIS-06-017, Institute of Information Science, Academia Sinica, Taipei,Taiwan, 2006.