Forschungsbericht Nr. 06 - 1

Personal Name Matching:

New Test Collections and a Social Network based Approach.

Patrick Reuther

Personal Name Matching: New Test Collections and a Social Network based Approach.

Patrick Reuther
Department for Databases and Information Systems (DBIS)
University of Trier, 54296 Trier, Germany
E-Mail: reuther@uni-trier.de

March 16, 2006

Abstract

This paper gives an overview of Personal Name Matching. Personal name matching is of great importance for all applications that deal with personal names. The problem with personal names is that they are not unique and sometimes even for one name many variations exist. This leads to the fact that databases on the one hand may have several entries for one and the same person and on the other hand have one entry for many different persons. For the evaluation of Personal Name Matching algorithms test collections are of great importance. Therefore existing test collections are outlined and three new test collections, based on real world bibliographic data, presented. Additionally state-of-the art techniques as well as a new approach based on semantics are described.

1 Introduction

Person names are ubiquitous in information systems. They are the only widely accepted method to identify persons. For humans the use of familiar

person names is superior to all artificial identifiers. For information systems person names have many drawbacks compared to synthetically generated identifiers. Person names may not be unique; several persons with the same name may exist. This problem connected to personal names is not new. In England for example after the Norman Conquest (1066) and in Germany at about the 12th A.C. people started to make use of surnames in addition to their given name because the use of only the given name was not sufficient anymore to identify different individuals with the same name. In Zurich(Switzerland) for example during the 13th century there were 692 known people whereas 77 people shared the same name [16], [7]. Some hundred years later, with the evolvement of the internet again a situation where many people come together from different countries, making an identification of individuals problematic, due to the fact that many share the same names, has arisen. A further problem connected with personal names is that a person may change his or her name; often several variations of a person name are used. Variations of names may be caused by abbreviations (for example Jeffrey D. Ullman may become J. D. Ullman, J. Ullman or Jeff Ullman), nicknames (e.g. Michael may become Mike, William / Bill, Joseph / Joe etc.), permutations (e.g. Liu Bin may be the same person as Bin Liu), different transcriptions (e.g. Andrei / Andrey / Andrej may be identical), accents (Stephane vs. Stéphane), umlauts (Muller / Müller / Mueller), ligatures (Weiß / Weiss or Åström / Aastrom ...), case (Al-AAli vs. Al- Aali), hyphens (Hans-Peter vs. Hans Peter), composition (MaoLin Qin vs. Mao Lin Qin), postfixes (Jr./Sr., a number, or a parents name), and by typos. Names may be changed at marriage or emigration to another cultural environment. For DBLP the variations in person names have two consequences: (1) It is unclear how to search persons best, and (2) it may be hard to normalise different spellings of names to get correct author pages. A containment of the problems of personal names can be achieved by two different strategies. A first strategy is standardisation. A standardisation can be achieved by establishing a unique identifying number for each author. However although standardisation has been proposed since the early seventies unique identifying numbers for authors have not been widely accepted [28]. A second strategy discussed in this paper tries to solve the problem connected with personal names on the basis of algorithms and is also discussed under the term Personal Name Matching. Personal Name Matching is a term first used by [6] which in their publication describes the task of finding synonyms among personal names. Personal Name Matching, as I understand it, has a

broader horizon. It describes the task of finding synonym personal names, one person is stored under many different names in a dataset, and homonym personal names, more than one person is stored under the same label. These two tasks have been widely discussed in literature by many different research domains from computer science to history, under a variety of terms such as record linkage [12], [24], duplicate detection [2], merge/purge [14] or name diambiguation [26]. Personal Name Matching can be viewed as a process which consists of five different stages. The first stage is **(1) normalisation**, in which a standardised representation of a person is guaranteed. Different representations can occur because of different standards in which personal data was stored such as BIBTEX and MARC or because of different principles such as a different handling of prefixes or suffixes. The second stage in the process is the so called **(2) blocking**, where a preselection of persons for the coming stages of the process is made. In order to find all synonyms each person would have to be compared to all other person in the dataset leading to $\frac{n(n-1)}{2}$ comparisons, with n as the total number of persons in the database. The same counts for homonyms where each person would have to be analysed thoroughly. In a database the size of DBLP with more than 320.000 authors one would have to make more than 51 billion comparisions only for an analysis of synoyms. This is, in terms of time and space, not to be accomplished efficiently. Diverse techniques such as character based distance meassures and token based similarity meassures were proposed to confine the amount of comparisions. After the preselection of persons during the blocking phase the remaining persons are undertaken a thorough **(3) analysis** in oder to determine whether one person is a homonym or two persons resemble a synonym. Again, as for the blocking phase, many different criteria – in general more complex than the blocking criteria due to less comparisons necessary – for an analysis have been presented in literature (phonetic criteria, edit-distance and clustering based approaches, etc.) each having in common a similarity measure to indicate whether to persons resemble a synonym or one person a homonym. This similarity measure is the basis for the fourth stage of the Personal Name Matching process, the **(4) decision** stage. Here as the name indicates a decision has to be made whether the similarity is high enough to take to persons as synonyms or to postulate that many different people are represented by one record. During the decision stage normally a classification into true matches, possible matches and false matches is obtained. Building up on this classification and error types derived from it, a **(5) performance evaluation** is carried

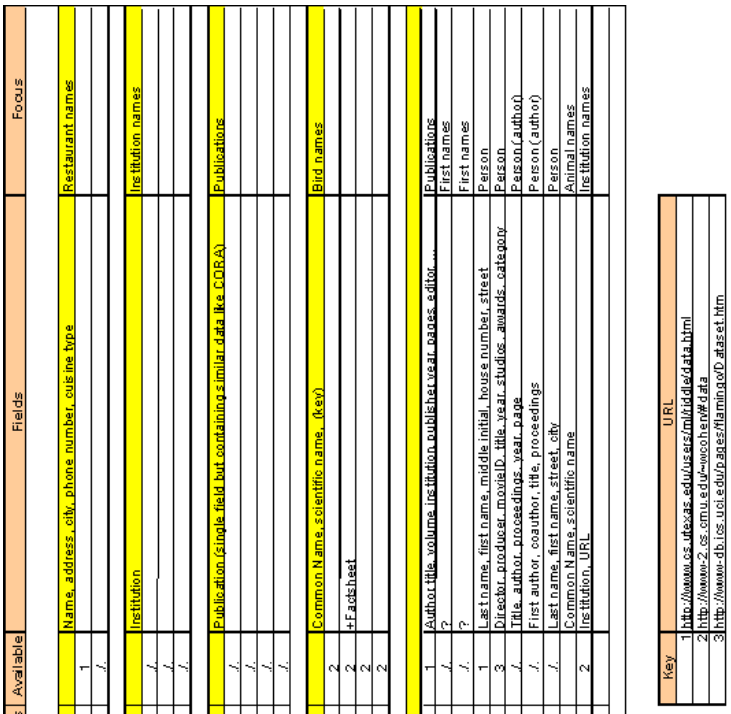
out as the last step in the here described process. The prevailing measure for an evaluation of the effectiveness of Personal Name Matching approaches are recall and precision, well known in the information retrieval community. The focus of this paper will lie on the evaluation of Personal Name Matching algorithms. First existing test collections related to personal name matching are presented and their shortcomings illustrated. Afterwards new collections for the Personal Name Matching task, focusing on bibliographic data are introduced. Before the paper ends with a conclusion and an invitation for participation a brief discussion of existing algorithms will be given.

2 Existing Test Collections

Up to today there has been little work on the evaluation of different Personal Name Matching strategies. According to [4] this is due to the fact that on the one hand there are only few test collections for duplicate detection available publicly. On the other hand they claim that the variety of different measures has hindered accuracy evaluation. Following [4] that Recall/Precision curves are the most appropriate methodology for duplicate detection evaluation, the lack of test collections remains as the main problem of evaluation. The following paper gives an overview of existing test collections and presents three new test collections which will be publicly available, therefore trying to confine the problem of the lack of existing collections.

Test collections in information retrieval in general consist of a variable amount of data from an arbitrary context (documents), example information requests (queries) and ground truth in form of relevant objects to a given query [1]. Test collections for Personal Name Matching or duplicate detection are composed of the same elements. However, the amount of example information request is often restricted to one: 'Find all duplicate objects in the given data set'. The relevant object to this sole query is a list of all duplicate objects in the data set. Obtaining this set of duplicates is a tedious task. Besides the construction of test collections with only on query it is also possible to have more than one query in the test collection. Such test collections have distinct queries for distinct personal names, c.f. 'Find all duplicate entries to Bill Clinton'. Figure 1 gives an overview of existing test collections suitable for the evaluation of identifying database records that are syntactically different but describe the same physical entity.

Figure 1: Overview of existing test collections



Name	Records	Duplicates	Distinct entities	Available	Fields	Focus
Restaurant					Name, address, city, phone number, cuisine type	Restaurant names
-->Zagat_Fodor	864	112	./.	1		
-->Zagat_Dinesite	2154	136	./.	./.		
Affiliation					Institution	Institution names
-->USCD	2824	./.	273	./.		
-->Stanford	5471	./.	379	./.		
-->USCD_UNION_Stanford	?	?	?	./.		
CiteSeer					Publication (single field but containing similar data like CORA)	Publications
-->Reasoning	514	./.	196	./.		
-->Face	349	./.	242	./.		
-->Reinforcement	406	./.	148	./.		
-->Constraints	295	./.	199	./.		
Birds					Common Name, scientific name, (key)	Bird names
Bird1	337	19	./.	2		
Bird2	1060	67	./.	2	+Factsheet	
Bird3	38	./.	./.	2		
Bird4	719	114	./.	2		
Others						
CORA	1916	./.	121	1	Author title, volume, institution, publisher, year, pages, editor, ...	Publications
Bolw	3872	?	?	./.	?	First names
ACM[1]	3076	?	?	./.	?	First names
Census	841	?	?	1	Last name, first name, middle initial, house number, street	Person
Movie	11453	200	./.	3	Director, producer, movieID, title, year, studios, awards, category	Person
Hep-ph	28204	585	./.	./.	Title, author, proceedings, year, page	Person (author)
DBLP[2][3]	12268	305	./.	./.	First author, coauthor, title, proceedings	Person (author)
MAILING	1200	./.	400	./.	Last name, first name, street, city	Person
Animal	5709	298	./.	./.	Common Name, scientific name	Animal names
Business	2139	./.	./.	2	Institution, URL	Institution names

Key	URL
1	http://www.cs.utexas.edu/users/ml/riddle/data.html
2	http://www-2.cs.cmu.edu/~wcohen/#data
3	http://www-db.ics.uci.edu/pages/Naming/Dataset.htm

The Census test collection is a synthetic census-like data set supplied by William Winkler containing several different fields such as last name, first name, middle initial, house number and street [5]. Restaurant describes a

test collection of 864 restaurant names, addresses, telephone numbers and a short description of the cuisine served, containing 112 duplicates obtained from two different sources, the Fodor's and Zagat's restaurant guides ([3],[29], [15]). Besides this collection further analysis of algorithms was undertaken on other restaurant collections containing data from Zagat's restaurant guide and Dinesite resulting in 136 matching records [20] in a total of 2154 records. Animal, Bird 1-4 and Business are collections, used from Cohen within the scope of his publication 'Data Integration Using Similarity Joins and a Word based Information Representation Language'[8]. Each of these collections focuses on a different domain, indicated by the name given to the collection. The Animal collection for example focuses on names of animals. The test collection contains the common name and scientific name of animals collected from different sources. The Business test collection comprises company names and can therefore be used to test algorithms which try to identify duplicate entries among names of companies. Besides the above mentioned collections Cohen also tested his approach to data integration on further different collections containing for example movie data. The collection from [22] is built up from the INSPEC bibliographical database and comprises three sub collections. The first dataset is a collection of affiliation records returned by an INSPEC query that extracted 2824 bibliographical records containing the keywords 'San Diego'and 'Univ.'in their affiliation field. The dataset was created by saving only the affiliation field in these citations. This so called UCSD dataset contains 273 affiliations after exact duplicate affiliations are removed. The second dataset is referred to as Stanford dataset and contains 379 distinct affiliation records gathered from an INSPEC query that retrieved 5417 total records containing the keywords 'Stanford'and 'Univ.'. The third dataset is the union of the UCSD and Stanford dataset [21]. Cora is a test collection based on the Cora research paper search engine. The Cora test collection comprises 1916 citations to 121 distinct papers. About one quarter of these papers have only one or two references to them [19]. The papers were originally collected from postscript files that were automatically converted to ASCII introducing systematic errors into the data set [9]. The BoW data set contains 3872 documents tagged with 5740 author names which was used from [11] for the evaluation of his approach to finding variants in first names, caused by abbreviations. In addition his evaluation was also undertaken on a further dataset constructed from a small fragment of the ACM Portal author index. The collection contains 3076 names where the last name starts with the letters 'Fe'[11]. A data set with a similar focus is the Citeseer dataset.

The Citeseer data set comprises four different subsets which are all single-field data sets extracted from the famous Citeseer publication research index [http://citeseer.ist.psu.edu/]. First this is the Reasoning data set containing 514 citation records that represent 196 unique papers from the computer science area automated reasoning. The second collection is the Face collection containing 349 citations to 242 papers from the face recognition area. Thirdly 406 citations to 148 papers covering the topic of reinforcement learning were extracted into the Reinforcement collection. Last but not least the Constraints collection consisting of 295 citations to 199 papers dealing with constraint satisfaction is the fourth collection of the Citeseer data set ([3], [5]). A test collection containing movie data was used by [17]. This collection consists of 11453 movie records containing 200 potential duplicate records. Furthermore [17] designed two more test collections, Hep-ph and the $\text{DBLP}_{(Lee)}$. Hep-ph contains 28204 publications covering energy physics and particle phenomenology. For this collection 585 potential duplicates were identified. The $\text{DBLP}_{(Lee)}$ collection is a subset of DBLP containing 12258 records which include 305 potential duplicates. Besides the existing test collections also tools to randomly construct test collections are available. One example is a database generator that provides a large number of parameters including the size of the desired database, the percentage of duplicate records and the amount of error to be introduced. The database generator generates records containing several fields such as social security number, first name, initial, last name, address, apartment and zip code [14]. Example collections constructed with the database generator are described in [14] and [2]. A framework that includes a range of string-matching methods from a variety of communities, including statistics, artificial intelligence, information retrieval, and databases is available at [http://secondstring.sourceforge.net/]. From a test collection perspective this framework is of interest because it also includes tools for systematically evaluating performance on test data. For an analysis some of the above mentioned data sets have been integrated into their project and have been subject of research [9].

As one can see, there are several test collections and tools available considering the problem of duplicate detection and object identification. However the awareness of these collections is not high among researchers. Most of the researchers design their own new test collection when testing new algorithms. This is due to the fact that no collection in the duplicate detection community is as popular as the TREC-collections for information retrieval evaluation. Additionally the given overview shows that duplicate detection and identity

uncertainty is a problem in many different domains making it necessary to have collections covering each of these different domains. For Personal Name Matching focussing on author names in bibliographic databases no test collection is available which can be used without adaptation. The Cora, BoW, Hep-ph, $\text{DBLP}_{(Lee)}$ and Citeseer test collections are the ones most similar to the Personal Name Matching context because all deal with bibliographical data. Unfortunately for the Cora, BoW and Citeseer data sets there is no ground truth available that determines the real duplicate authors in the collection therefore making an analysis of the Personal Name Matching techniques in terms of Recall/Precision diagrams impossible. For the $\text{DBLP}_{(Lee)}$ collection access was gained to the test collection, however the ground truth is no longer available, making it useless as an appropriate test collection. Access to the Hep-ph was not possible (March 2005). The following paragraph introduces two new collections for the Personal Name Matching.

3 New collections for name matching

The new data sets developed for the Personal Name Matching tasks are built upon a real world database, the Digital Bibliography & Library Project (DBLP). DBLP is maintained at the University of Trier and currently contains more than 725.000 publications from more than 450.000 authors. The topical focus of the publications lie within the computer science domain. It is frequently used in the scientific community which is indicated by more than 6.000.000 page hits a month. The bibliography is mirrored on servers around the world including the RWTH Aachen, ACM New York, VLDB Darmstadt, UCLA Los Angeles. Additionally current textbooks as well as more than 20 of the leading international organisations dealing with databases (ACM SIGMOD, VLDB Endowment, EDBT Foundation, and IEEE TC Data Engineering) recommend the usage of DBLP. Duplicate detection on the DBLP data is of high importance if one wants to maintain the high quality standard of the database. Without quality management a person search would not work properly making innovative ideas like exploration of interrelationships between scientists on a conference, making use of DBLP data and RFID technology, impossible (c.f. ICDE 2005, Tokyo).

Desirable from a quality management aspect is the construction of a test collection containing all the DBLP data because thereby one could identify all duplicate entries in the database while gathering the ground truth for the test

collection. However achieving this is critical because of complexity problems. With the current size of DBLP an algorithm for duplicate detection on a test collection containing all data and not making use of blocking would have to consider more than 190 billion comparisons of author pairs. Assuming that such a comparison takes 0.1 second the whole comparison step would take more than 609 years. The time necessary for a human to decide whether two records really represent duplicates is however much higher than the assumed tenth of a second. Because the constructed collections only contain a subset of the DBLP database they will be referred to as DBLP-SUB$_{01}$, DBLP-SUB$_{02}$ and DBLP-SUB$_{03}$ in the following.

For DBLP-SUB$_{01}$ 500 authors were chosen randomly from the DBLP database. All the publications and naturally the authors were added to DBLP-SUB$_{01}$. Additionally, all the co-authors of the publications the 500 randomly chosen authors have, were added to the collection. DBLP-SUB$_{02}$ was constructed in a similar way with the difference that now 1000 authors were chosen randomly from the whole DBLP database. As a result the DBLP-SUB$_{01}$ data set contains 2796 authors and 1510 publications, whereas the DBLP-SUB$_{02}$ contains 6351 authors and 4139 publications. DBLP-SUB$_{03}$ was constructed in a slightly different way. Here the data for the collection was gained by comparing two different versions of the DBLP-data for updates in spellings of names. By following this strategy the identification of duplicate entries was integrated into the step of building the collection itself. Applying the described approach to the 16/03/2004 and 16/09/2004 versions of the DBLP data leads to the DBLP-SUB$_{03}$ dataset which contains 21688 authors and 18872 publications. Besides the authors and the title of the publication

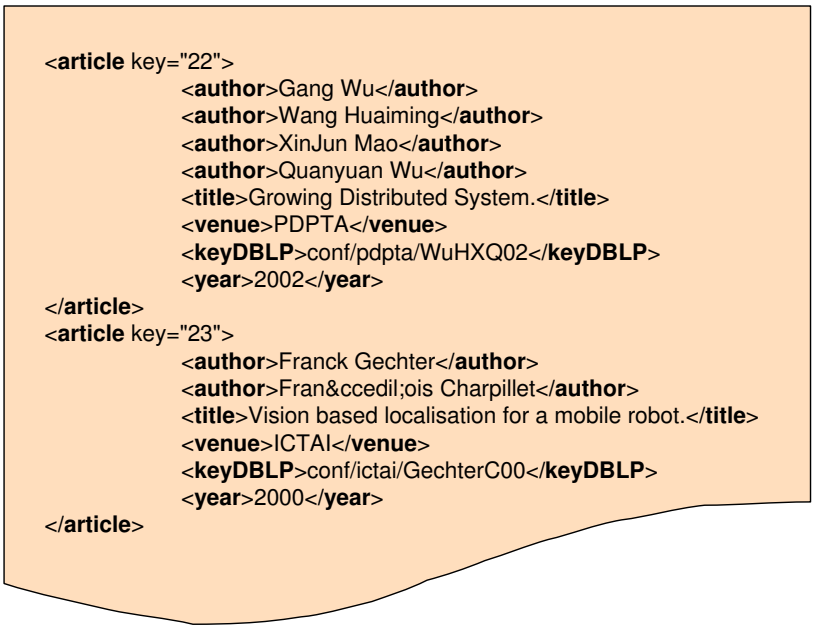
Collection	Authors	Publications	Synonyms
DBLP-SUB$_{01}$	2796	1510	23
DBLP-SUB$_{02}$	6351	4139	73
DBLP-SUB$_{03}$	21688	18872	2020

Table 1: Statistics of DBLP-SUB$_{01-03}$

also the year of publication, the publication venue and the corresponding DBLP-Key is stored in the test collections. Figure 2 shows a sample of the DBLP-SUB collections in a XML representation.

Typical for test collections designed for the task of duplicate detection

Figure 2: Example taken from DBLP-SUB$_{01}$



only one example information request is considered in the the developed collections. The one query included in the collections is 'Who are duplicate authors'. The ground truth for the sole query in all three collections was obtained by pooling [1] and for DBLP-SUB$_{03}$, as mentioned, additionally through making use of already available duplicate information. As described above the identification of duplicates in a dataset normally requires a full comparison of each author with all the other authors stored in the collection. Again this comparison is too costly because for the test collection a manual inspection of all candidate pairs has to be made. Taking 30 seconds for a comparison of two authors considering their aspect of being duplicates, a time often not sufficient, determining the ground truth for DBLP-SUB$_{01}$ would require more than 3 years in total. Following the pooling technique the ground truth for the example information request of all three collections was obtained from a pool of possible relevant candidate pairs of authors. This pool was created by taking the top K candidate pairs in the rankings generated by various duplicate detection algorithms. The candi-

dates in the pool were then shown to human assessors who ultimately decided whether the two candidates really represent synonyms. By applying the described strategy to the DBLP-SUB$_{01}$, DBLP-SUB$_{02}$ and DBLP-SUB$_{03}$ datasets, 23 duplicates were identified for the DBLP-SUB$_{01}$ collection, 73 for DBLP-SUB$_{02}$ and 2020 for DBLP-SUB$_{03}$. All three collections including the ground truth are publicly available and can be downloaded from http://dbis.uni-trier.de/Mitarbeiter/reuther_files/private/reuther.shtml.

4 Personal Name Matching with Social Networks

Personal Name Matching has been of interest for research and practice quite a long time. Therefore it is not surprising that a variety of different techniques have been proposed for this task. The proposed algorithms can be either distinguished by the granularity or by the matching criteria used. Considering the granularity, a differentiation in a letter based comparison and a token based comparison can be made. Letter based comparison treats the author name as a whole without further consideration of the different tokens like first name, initials or family names. The comparison is based on comparing the letters of the author names. In contrast, a token based comparison makes use of the different elements of a name and uses these elements as a basis for the comparison. The token based comparison is also referred to as a vector based approach [3]. The choice of granularity affects the selection of comparison criteria. There are several different comparison criteria, categorized into probabilistic and distance based methods, of which only a few shall be discussed in the following. Probabilistic criteria rely, as the name indicates, on probabilities. An example of such a criterion is the likelihood ratio used in record linkage. A problem connected to probability-based approaches lies in the difficult estimation of probabilities. This is why many distance based methods, which are claimed to be more robust, are proposed [13].

One of the most common distance based comparison criterion is the Levenshtein Distance [18]. Levenshtein measures the similarity of two strings, or, in the context presented here, authors, by calculating the minimum number of edit operations (insertion, deletion and substitution) of single characters required to transform one name into the other. There exist various variations of the basic Levenshtein distance such as allowing the permutation of

adjacent characters or establishing differentiated cost models for the different actions. Another distance-based comparison technique is the concept of n-grams [27]. An n-gram is a sequence of n characters. All n-grams of a word can be obtained by using a stencil of the length n. This stencil is laid over the word starting at the beginning of the word and thereby achieving the first n-gram. Then the stencil is moved right character by character, each time receiving the next n-gram of the word. Two identical words consequently have the same n-grams. The similarity of two authors can be calculated on the basis of the number of similar n-grams and the total number of unique n-grams. If the different parts of a name are treated separately, for example, by building a vector representation of a name a token based comparison technique can be obtained. Similar to the Vector Space Model known from information retrieval, a name is split into its components which receive a somehow calculated weight. The similarity of two authors can be obtained by standard cosine measure or Euclidean distance. Such a comparison technique is claimed to often outperform Levenshtein-distance-based comparison techniques [9]. Besides the mentioned comparison methods many other techniques like typewriter distance, Jaro-Winkler, Monge-Elkan, or phonetic distances could be used. However, most of the approaches only make use of the syntactical characteristics of the given author names and do not pay attention to semantic information available.

A new approach which makes use of semantic information is what I refer to as Personal Name Matching with Connected Triples. This approach makes use of semantic relationships based on social networks. A social network is characterised through a set of people or groups each of which has connections of some kind to some or all the others. In the language of social networks, the people or groups are referred to as actors, the connection between actors as ties. An actor in a social network might be a single person, a group or a company, whereas the tie, correspondingly, can be either friendship between two people, collaboration between two teams, or business relationships between companies [25]. The social networks of interest for the Personal Name Matching task are the so called co-authorship networks. Co-authorship networks are a special kind of social networks in which the actors are authors and a tie is created between authors if there is a publication which both actors collaborated and published together. Consequently, a co-authorship network can also be seen as a graph, where the authors represent the nodes, the shared publications the ties [23]. The question remains how the concept of co-authorship networks can be used for synonym detection. In social net-

works, one can observe a phenomenon that if one person A1 has a friendship with two other people C_1 and C_2 then it is likely that person C_1 will get to know person C_2 or vice versa and, consequently, a friendship could start between these two people as well. Transferred to the context of authors this means that if one author regularly collaborates and publishes with two other authors it is likely that these two other authors start collaborating and publishing together as well. This is what in literature is referred to as transitivity linking [10] or clustering [30]. The existence of a transitivity link, however, is not very promising for the task of finding a social network based approach to Personal Name Matching. In fact, not the existence, but the missing of a transitivity link can be used for the Personal Name Matching task.

Figure 3: Social Network with Connected Triple

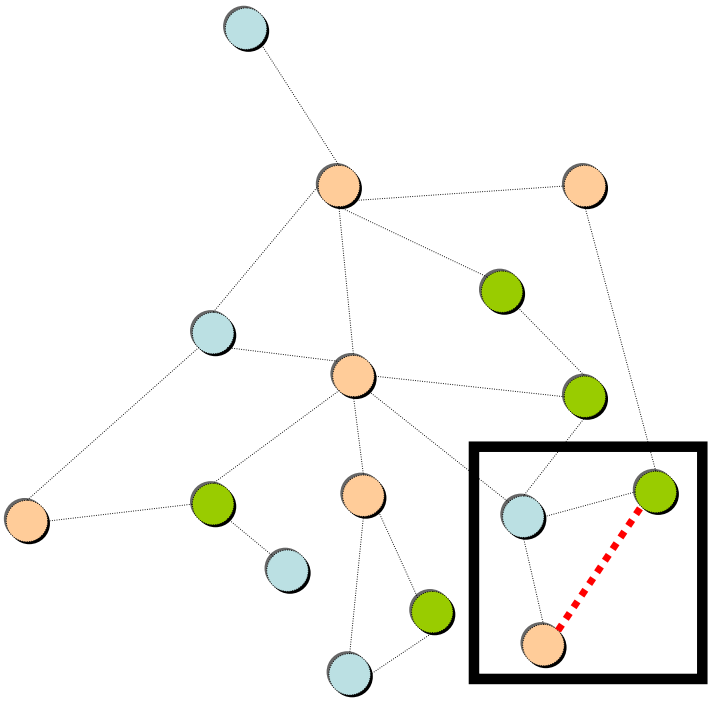
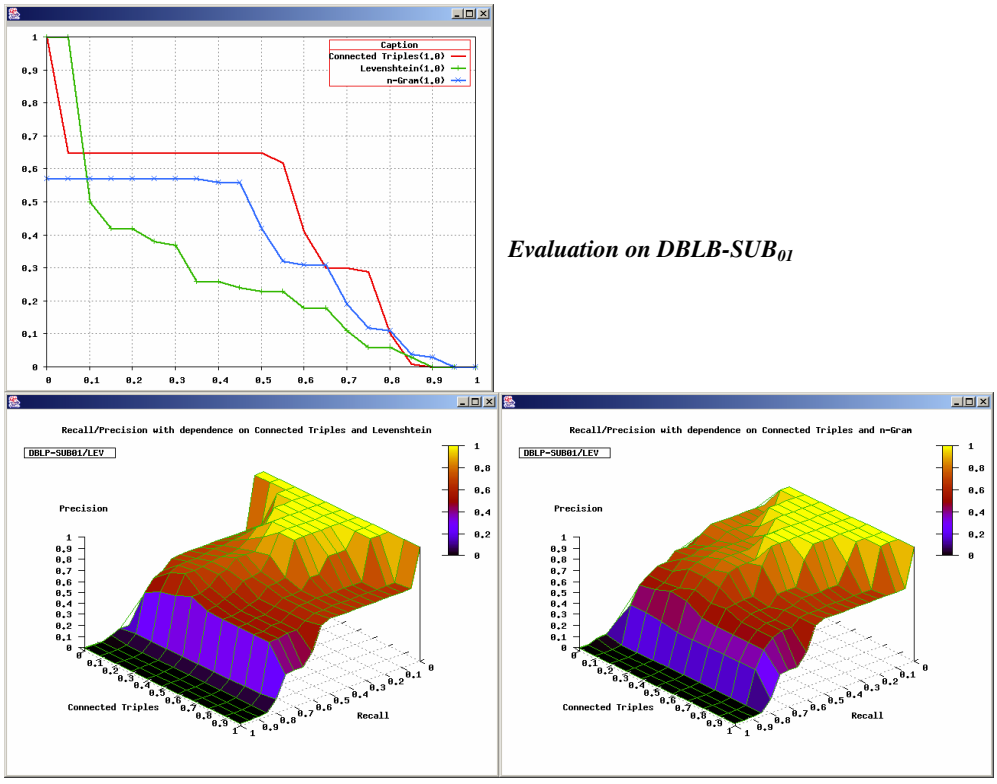


Figure 3 shows an example of a missing transitivity link. In the black-bordered box one can see that the blue node has published with the red as well as with the green node, yet there is no transitivity link between the red

and the green node indicating that the latter have not published a single publication together. The constellation shown in the discussed figure is referred to as Connected Triple. A Connected Triple in a co-authorship network can always be found if two authors A_1 and A_2 each published with an author A_k but have never published together. The approach of identifying duplicates with Connected Triples described so far does not take into consideration aspects of time and aspects of topic. If author A_1 of a Connected Triple published with author A_k in the 1950s, author A_2 with A_k in the late 1990s then it is likely that author A_1 and author A_2 do not know each other and therefore no transitivity link is given. Furthermore, A_1 and A_2 will probably not be duplicates.

Figure 4: Evaluation of Connected Triple Matching on DBLP-SUB_{01}



Evaluation on DBLB-SUB_{01}

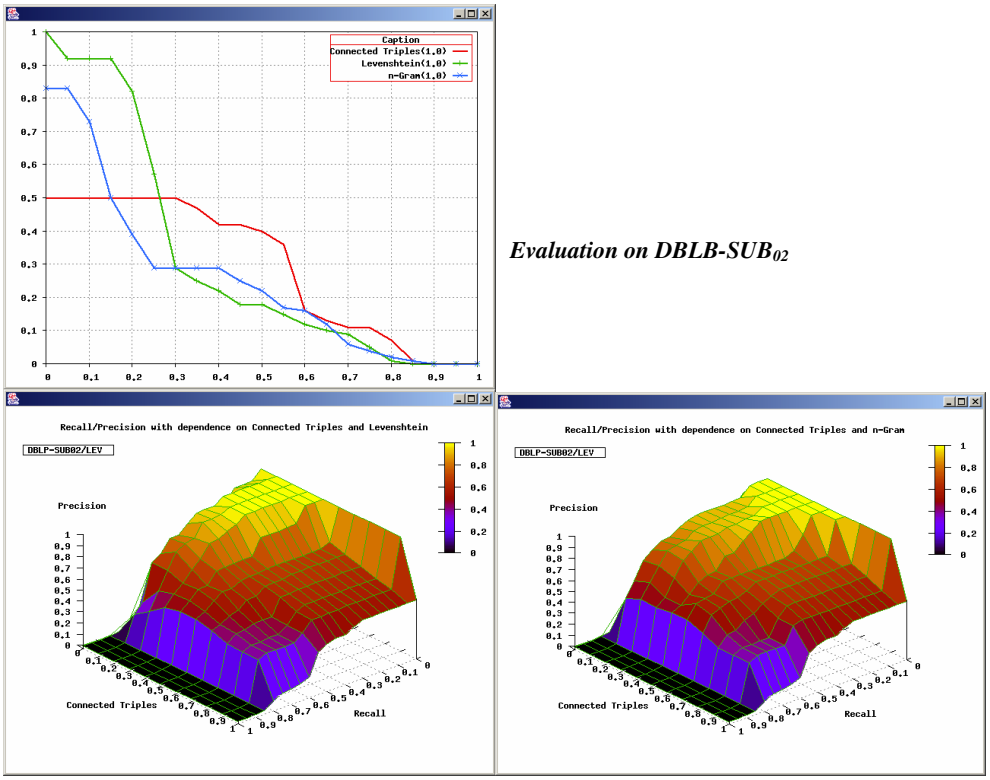
The second mentioned aspect, the missing consideration of topical aspects, aims at the same consequences. It is possible that two authors A_1 and A_2 both publish with author A_k without publishing together because of topical aspects. If author A_1 always publishes with A_k on databases and A_2 with A_k on biology then it is apparent why there is no transitivity link and a Connected Triple. In order to consider both aspects of time and topic it is possible to make use of further information available. The consideration of time can be regarded by including the publication date of the papers the authors in question published. By doing so, it is, for example, easy to see that B. Ludwig and A. Weber in the DBLP database are not duplicates although there is a Connected Triple in the co-authorship graph because B. Ludwig published in the early 1990s and A. Weber in the early 2000s with B. Walter. The identification of Connected Triples due to topical differences is more complicated than the identification of Connected Triples caused by time aspects. Topical aspects cannot be reduced to one single number as easily as the date of publication. Rather it is necessary to try to identify the topics of the papers' authors A_1 and A_2 published with author A_k. This is normally achieved through clustering the different publications by using keywords, titles, abstracts or full texts if available.

An evaluation of the proposed algorithm based on the newly developed test collections indicates that the use of semantic information can enhance the Personal Name Matching task. For the evaluation the approach of Connected Triples was used in its purest form, making use only of the amount of Connected Triples. As a blocking criterion the Levenshtein distance was used, thereby implicitly adding some syntactical considerations to the Connected Triple approach.

Figure 4 gives a graphical representation of the evaluation task using Recall/Precision diagrams as a basis for evaluation. The top diagram shows the performance of the algorithm based on Connected Triples in comparison to two syntactical approaches, one based on the Levenshtein distance, the other on n-grams for the DBLP-SUB_{01} data test collection.

As one can see, making use of only semantic information is a solution to the Personal Name Matching task. The precision making use of Connected Triples for recall values up to 10% is lower than the one obtained when making use of Levenshtein, but higher than using n-grams. From 10% recall level to about 80% the precision of the approach based on Connected Triples remains higher than the precision values of both other approaches. From 80% recall on, all three algorithms lead to similar recall levels. On the whole, the se-

Figure 5: Evaluation of Connected Triple Matching on DBLP-SUB$_{02}$



mantic based approach leads to the highest average precision when compared to the other two syntactical approaches. Figure 4 also shows that not all relevant entries can be identified if a blocking criterion is used. In the presented example all candidate authors whose names have a Levenshtein distance of more than six were excluded from further analysis. Blocking all these candidates makes it possible to find only about 90% of the duplicate entries present in the data test collection. This is indicated in Recall/Precision diagram by the fact that the precision is zero at a recall level of approximately 90%. Having a look at the upper Recall/Precision diagram in figure 5 one can see, that on the DBLP-SUB$_{02}$ data test collection both the Levenshtein distance and n-grams outperform the approach based on Connected Triples for recall

levels up to about 30% (for n-grams 15%). From then on the precision of the semantic approach is higher than the precision of both syntactic approaches. Again, one can see from the fact that precision is zero at about 90% recall that blocking can lead to not finding all duplicates. One reason for the lower performance of the social network based approach for low recall levels on the second data test collection is that this collection contains publications which many authors collaborated on. For two authors who participated only in one of those two collections many Connected Triples are created leading to a high score for the two candidates. Further studies considering detailed information on the origin of Connected Triples have to be done in order to confine the mentioned phenomenon. Besides using syntactical approaches and semantic approaches separately also hybrid approaches combining them have been analysed. The two lower Recall/Precision diagrams in figure 4 and figure 5 show the performance of hybrid approaches once on the DBLP-SUB_{01} test collection and on the DBLP-SUB_{02} data set. All four diagrams show that a combination of either Levenshtein and Connected Triples or n-grams and Connected Triples leads to an increase in precision. Having a closer look at the bottom-left diagram in figure 5 for example indicates that combining Levenshtein with 60% and Connected Triples with 40% leads to the best Recall/Precision curve. The bottom-right diagram shows that a fifty-fifty consideration of n-grams and Connected Triples results in the best performance. As the above paragraphs show, the accuracy for Personal Name Matching can be increased while making use of semantics. The evaluation with Recall/Precision diagrams on two test collections shows that best results for Personal Name Matching can be obtained when combining syntactic and semantic information. Further research considering machine learning aspects as well as further available information such as the internet will be done, with a bit of luck leading to an even better performance for the Personal Name Matching task.

5 Conclusion and Outlook

This paper gives an overview of different data collections for duplicate detection and object identification. Although there already exist a couple of test collections, none has arisen to a standard for evaluation of duplicate detection algorithms. Scientists in this research area tend to generate new test collections for their studies. This is firstly due to the fact that exist-

ing collections are not well known and secondly because duplicate detection can be a subject in different domains each needing specific test data. By presenting an overview of existing data sets this article tries to increase the awareness of already existing data sets which can be used by researchers for testing their algorithms without having to invest their time in creating their own new collections.

Additionally three new collections were developed which can be used for the Personal Name Matching task described in the introduction of this paper. Each of the constructed datasets contains bibliographical data and a list of duplicate authors in the dataset. Such test collections based on high quality data give researchers the possibility to test Personal Name Matching algorithms on real life data which makes it easier to draw conclusions for real life scenarios that have to cope with personal names.

Besides presenting an overview of test collections and introducing new collections a short sketch on existing approaches to personal name matching was given. A new approach making use of semantic information extracted from a social network of co-authors was illustrated and evaluated. Although there exist a variety of Personal Name Matching approaches, Personal Name Matching is still an active research area. Especially the exploitation of semantics offers new possibilities to invent better algorithms. Besides the use of semantics for finding duplicates hardly any work can be found on the second part of the Personal Name Matching definition, the identification of persons who share the same name but are stored under the same label in the database. Hopefully this paper aroused the reader's interest to participate in an interesting research area dealing with something of everyday life, personal names.

References

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Publishing Company, 1999.

[2] Mikhail Bilenko and Raymond J. Mooney. Learning to combine trained distance metrics for duplicate detection in databases. Technical Report AI 02-296, Artificial Intelligence Lab, University of Texas at Austin, Austin, Texas, USA, 2002.

[3] Mikhail Bilenko and Raymond J. Mooney. Adaptive duplicate detection using learnable string similarity measures. In Lise Getoor, Ted E. Senator, Pedro Domingos, and Christos Faloutsos, editors, *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 24 - 27, 2003*, pages 39–48. ACM, 2003.

[4] Mikhail Bilenko and Raymond J. Mooney. On evaluation and training-set construction for duplicate detection. In *Proceedings of the KDD-2003 Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, pages 7–12, Wasington, DC, 2003.

[5] Mikhail Bilenko, Raymond J. Mooney, William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. Adaptive name matching in information integration. 18(5):16–23, 2003.

[6] Christine L. Borgman and Susan L. Siegfried. Getty's synoname and its cousins: A survey of applications of personal name-matching algorithms. *Journal of the American Society for Information Science (JASIS)*, 43(7):459–476, 1992.

[7] William Dodgson Bowman. *The Story of Surnames*. George Routledge & Sons, LTD., London et. al, 1931.

[8] William W. Cohen. Data integration using similarity joins and a word-based information representation language. *ACM Transactions on Information Systems (TOIS)*, 18(3):288–321, 2000.

[9] William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. A comparison of string distance metrics for name-matching tasks. In Subbarao Kambhampati and Craig A. Knoblock, editors, *Proceedings of IJCAI-03 Workshop on Information Integration on the Web (IIWeb-03)*, pages 73–78, Acapulco, Mexico, 2003.

[10] Holger Ebel, Jrn Davidsen, and Stefan Bornholdt. Dynamics of social networks. *Complexity*, 8(2):24–27, 2002.

[11] Dror G. Feitelson. On identifying name equivalences in digital libraries. *Information Research*, 9(4), 2004.

[12] Ivan P. Fellegi and Alan B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, December 1969.

[13] Lifang Gu, Rohan A. Baxter, Deanne Vickers, and Chris P.Rainsford. Record linkage: Current practice and future directions. Technical Report 03/83, CSIRO Mathematical and Information Sciences, Canberra, Australia, 2003.

[14] Mauricio A. Hernández and Salvatore J. Stolfo. The merge/purge problem for large databases. In Michael J. Carey and Donovan A. Schneider, editors, *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, pages 127–138, San Jose, California, USA, 1995. ACM Press.

[15] Craig A. Knoblock and Steven Minton. Building agents for internet-base supply chain integration. In *Proceedings of the Workshop on Agents for Electronic Commerce and Managing the Internet-Enabled Supply Chain*, 1999.

[16] Konrad Kunze. *dtv-Atlas Namenkunde: Vor- und Familiennamen im deutschen Sprachgebiet*. 5. durchgesehene und korrigierte auflage edition, 2004.

[17] Mong-Li Lee, Wynne Hsu, and Vijay Kothari. Cleaning the spurious links in data. 19(2):28–33, 2004.

[18] Vladimir I. Levenshtein. Binary codes capable of correcting spurious insertions and deletions of ones (original in russian). *Russian Problemy Peredachi Informatsii*, 1:12–25, 1965.

[19] Andrew McCallum, Kamal Nigam, and Lyle H. Ungar. Efficient clustering of high-dimensional data sets with application to reference matching. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 169–178, Boston, MA, USA, 2000.

[20] Martin Michalowski, Snehal Thakkar, and Craig A. Knoblock. Exploiting secondary sources for unsupervised record linkage. In *Proceedings of the 2004 VLDB Workshop on Information Integration on the Web*, Toronto, Canada, 2004.

[21] Alvaro E. Monge. *Adaptive detection of approximately duplicate database records and the database integration approach to information discovery.* PhD thesis, University of California, San Diego, 1997.

[22] Alvaro E. Monge and Charles Elkan. The field matching problem: Algorithms and applications. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 267–270, Portland, Oregon, USA, 1996.

[23] Peter Mutschke. Enhancing information retrieval in federated bibliographic data sources using author network based stratagems. In Panos Constantopoulos and Ingeborg Sølvberg, editors, *ECDL*, volume 2163 of *Lecture Notes in Computer Science*, pages 287–299. Springer, 2001.

[24] H.B. Newcombe, J.M. Kennedy, S.J. Axford, and A.P. James. Automatic linkage of vital records. *Science*, 130(3381):954–959, 1959.

[25] Mark E. J. Newman. Who is the best connected scientist? a study of scientific coauthorship networks. *Complex Networks*, pages 337–370, 2004.

[26] Byung-Won On, Dongwon Lee, Jaewoo Kang, and Prasenjit Mitra. Comparative study of name disambiguation problem using a scalable blocking-based framework. In Mary Marlino, Tamara Sumner, and Frank M. Shipman III, editors, *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, JCDL 2005*, pages 344–353, Denver, CA, USA, 2005. ACM.

[27] James L. Peterson. Computer programs for detecting and correcting spelling errors. *Communications of the ACM (CACM)*, 23(12):676–687, 1980.

[28] M. M. M. Snyman and M. Jansen van Rensburg. Revolutionizing name authority control. In *ACM Digital Libraries 2000 (ACM DL)*, pages 185–194. ACM, 2000.

[29] Sheila Tejada, Craig A. Knoblock, and Steven Minton. Learning object identification rules for information integration. *Information Systems*, 26(8):607–633, 2001.

[30] Duncan J. Watts. *Six Degrees: The Science of a Connected Age*. W. W. Norton & Company, New York, 2004.

Trierer Forschungsberichte
Mathematik / Informatik

Nr. 06 - 1 Patrick Reuther
Personal Name Matching: New Test Collections and a Social Network based Approach

Nr. 05 - 5 Subhendu Bikash Hazra
Reduced Hessian Updates in Simultaneous Pseudo-Timestepping for Aerodynamic Shape Optimization

Nr. 05 - 4 M. Tsurutani, M. Umezawa, Y. Yamamoto
Impossibility and Possibility Theorems of Social Choice Function with Restricted Alternative Set

Nr. 05 - 3 Ditmar Erdmann
Convex Homogeneous Functions and Inequalities

Nr. 05 - 2 S. B. Hazra, V. Schulz, J. Brezillon
Simultaneous Pseudo-Time Stepping for 3D Aerodynamic Shape Optimization

Nr. 05 - 1 A. Kaplan, R. Tichatschke
Bergman-like functions and proximal methods for variational problems with nonlinear constraints

Nr. 04 - 9 A. Kaplan, R. Tichatschke
Some results about proximal-like methods

Nr. 04 - 8 C. Frougny, V. Brattka, N. Müller
6th Conference on Real Numbers and Computers, Dagstuhl 2004

Nr. 04 - 7 S. B. Hazra
An Efficient Method for Aerodynamic Shape

Nr. 04 - 6 S. B. Hazra, V. Schulz
Simultaneous Pseudo-Timestepping for Aerodynamic Shape Optimization Problems with State Constraints

Nr. 04 - 5 Volker Klotz, Christoph Meinel
10 Jahre ECCC – Eine Digitale Bibliothek in weltweiter Benutzung

Nr. 04 - 4 Mikail Gevantmakher, Christoph Meinel
TI-jPACS - eine frei verfügbare leistungsfähige Plattform zur medizinischen Bildverarbeitung und -visualisierung

Nr. 04 - 3 Mikail Gevantmakher, Christoph Meinel
Medizinische Bildverarbeitung - eine Übersicht

Nr. 04 - 2 S. B. Hazra, V. Schulz, J. Brezillon, N. R. Gauger
Aerodynamic Shape Optimization using Simultaneous Pseudo-Timestepping

Nr. 04 - 1 Hannes Frey, Johannes K. Lehnert, Daniel Görgen, Peter Sturm
A Generic Background Dissemination Service for Mobile Ad-Hoc Networks

Nr. 03 - 6 Michael Schmitt, Christoph Meinel
Design and Implementation of a PHP-based Web Server for the Tele-Lab IT Security

Nr. 03 - 5 Feng Cheng, Paul Ferring, Christoph Meinel
Lock-Keeper Technology - A New Network Security Solution

Nr. 03 - 4 Christoph Meinel, Volker Schillings
tele-TASK – Teleteaching praxistauglich für den Universitätsalltag

Nr. 03 - 3 Wei Zhou, Christoph Meinel
Implement Role-Based Access Control with Attribute Certificates

Nr. 03 - 2 Ditmar Erdmann
A New Proof of the Beckenbach Inequalities