Managing the Quality of Person Names in DBLP

Patrick Reuther[1], Bernd Walter[1], Michael Ley[1],
Alexander Weber[1], and Stefan Klink[2]

[1] Department of Databases and Information Systems (DBIS),
University of Trier, Germany
{reuther, walter, ley, aweber}@uni-trier.de
http://dbis.uni-trier.de

[2] Institute of Applied Informatics and Formal Description Methods,
Universitt Karlsruhe (TH), Germany
Stefan.Klink@aifb.uni-karlsruhe.de
http://www.aifb.uni-karlsruhe.de

Abstract. Quality management is, not only for digital libraries, an important task in which many dimensions and different aspects have to be considered. The following paper gives a short overview on DBLP in which the data acquisition and maintenance process underlying DBLP is discussed from a quality point of view. The paper finishes with a new approach to identify erroneous person names.

1 Introduction

The amount of information is growing exponentially. This counts also for scientific domains where one can observe a fast growth in publications. Scientific publications are the appropriate means to communicate results and new insights. Besides on a more personal level and enhanced by the often cited publish or perish mentality publications are a sort of collecting credit points for the CV. Using bibliographic statistics is more and more the first choice to evaluate scientists on an institutional level. It is obvious that all the mentioned aspects build on reliable collection, organization and access to publications.

Of utmost importance for any provider of bibliographical content is the quality of the service they offer. Quality management is ubiquitous and plays a central role in nearly any domain. For services offering access to scientific publications data quality management, a part of quality management in general, is the central challenge. Data quality comprises many different dimensions and aspects. Redman, for example, presents a variety of dimensions such as the completeness, accuracy, correctness, currency and consistency of data as well as two basic aspects to improve quality: data-driven and process driven strategies [4].

The remainder of this paper gives an overview on DBLP and its data acquisition and maintenance process, focussing on quality problems, especially problems connected to personal names. The paper ends with the presentation of a social network based approach to identify erroneous person names.

J. Gonzalo et al. (Eds.): ECDL 2006, LNCS 4172, pp. 508–511, 2006.

2 DBLP and Quality Management

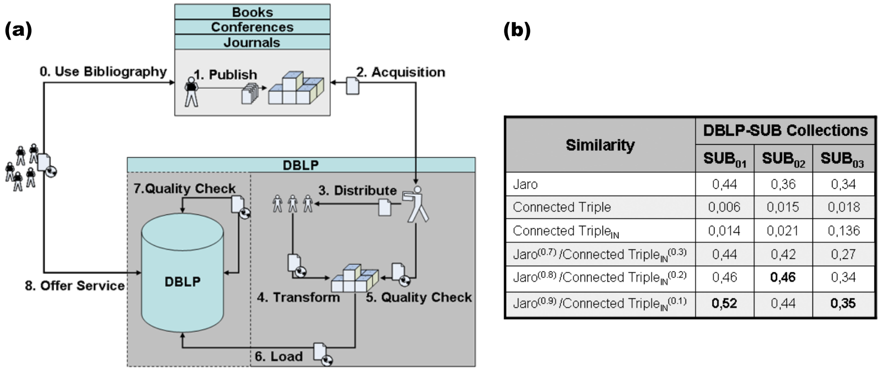
DBLP (*Digital Bibliography & Library Project*) [2] is an *internet newcomer* offering access to scientific publications. Today (May 2006) DBLP indexes more than 750.000 publications published by more than 450.000 authors.

Building a bibliographic database always requires decisions between quality and quantity. For DBLP we decided to prefer the quality of the records we offer to the quantity. It is easy to produce a huge number of bibliographic records disregarding quality aspects like standardization of journal names or person names. However, as soon as you try to guarantee that an entity is always represented by exactly the same character string and no entities share the same representation, data maintenance becomes very expensive.

Traditionally this process is called *authority control.* In DBLP the number of different journals and conference series is a few thousands so that guaranteeing consistency is not a serious problem. In contrast, authority control for person names is much harder due to the magnitude of $> 450k$ and the fact that available information is often incomplete and contradictory.

On a high level representation the data acquisition and maintenance process of DBLP shown in Fig. 1(a) can be seen as a ETL-process often found in data warehousing in which data is **E**xtracted from outside sources, **T**ransformed to fit business needs and finally **L**oaded into the database for further usage. The data of interest for DBLP which is extracted are publications authored by scientist and published in either journals, conference proceedings or more general, scientific venues. For DBLP there is a broad range of primary information sources. Usually we get electronic documents but sometimes all information has to be typed in manually. In some cases we have only the front pages (title pages, table of contents) of a journal or proceedings volume. The table of contents often contains information inferior to the head of the article itself: Sometimes the given names of the authors are abbreviated. The affiliation information for authors often is missing. Many tables of contents contain errors, especially if they were produced under time pressure like many proceedings. Even in the head of the article itself you may find typographic errors. A very simple but important policy is to enter all articles of a proceedings volume or journal issue in one step. In DBLP we make only very few exception from this *all or nothing policy.* For data quality this has several advantages over entering CVs of scientists or reference lists of papers: It is easier to guarantee complete coverage of a journal or conference series. There is less danger to become biased in favor of some person(s).

After the acquisition of data it is distributed in order to transform the data efficiently into the internal representation. Up to now, this is mainly the work of student assistants extra hired to accomplish the time consuming task of transformation. After the transformation in which first consistency checks are naturally applied the data is subject to a more thorough quality analysis. In this stage problematic cases not handled in transformation as well as further error detection are the main tasks. This work is mainly done by M. Ley for which he makes use of small tools such as the DBL-Browser [1] and scripts to automatically identify erroneous data. Here data edits are integrated into the process or



Similarity	DBLP-SUB Collections		
	SUB_{01}	SUB_{02}	SUB_{03}
Jaro	0,44	0,36	0,34
Connected Triple	0,006	0,015	0,018
Connected $Triple_{IN}$	0,014	0,021	0,136
$Jaro^{(0.7)}$ /Connected $Triple_{IN}^{(0.3)}$	0,44	0,42	0,27
$Jaro^{(0.8)}$ /Connected $Triple_{IN}^{(0.2)}$	0,46	**0,46**	0,34
$Jaro^{(0.9)}$ /Connected $Triple_{IN}^{(0.1)}$	**0,52**	0,44	**0,35**

Fig. 1. (a)Aquisition/Maintenance process (b)Evaluation of Avg. Precision

information chain making them a part of the process driven quality management. Example data edits in use are simple rules like firing a warning if two person names have a string distance which is smaller then a predefined threshold or if formatting conventions are not met. After the quality assurance the new data is loaded into the main DBLP database. At a typical working day we add about 500 bibliographic records. It is unrealistic to belief that this is possible without introducing new errors and without overlooking old ones. It is unavoidable that care during the input process varies. Therefore even after integrating the new records into the live system data quality is checked regularly. This data driven quality management is again supported by simple scripts and small tools. The loop of the data acquisition and maintenance process for DBLP closes when researchers use the system, especially the new entered bibliographical records and use them to produce new publications which some day will most likely be integrated into the DBLP system. From a data quality point of view improvements for data quality can only be made for the stages (2. Acquisition) to (7. Quality Check). The primary information creation and publishing (1. Publishing) is not in our area of responsibility and therefore can not be ameliorated, although improvements such as implementing an International Standard Author Number (ISAN), analogously to the ISBN known for publications, would confine the problems connected to names dramatically [5].

3 Personal Name Matching with Co-author Networks

From reflecting on how we find errors and inconsistencies concerning person names, we designed new similarity measurements based on a co-author network G in which authors are represented as vertices V, and co-authorship builds the edges E. For two person names a simple way to determine their similarity is to count the amount of Connected Triples they are part in. A Connected Triple $\wedge = \{V_\wedge, E_\wedge\}$ can be described as a subgraph of G consisting of three vertices with $V_\wedge = \{A_1, A_2, A_3\} \subset V$ and $E_\wedge = \{e_{A_1,A_2}, e_{A_1,A_3}\} \in E, \{e_{A_2,A_3}\} \notin E$. The

Connected Triple similarity of two names i and j is then calculated by $\frac{|C_{\wedge_{ij}}|}{|C_{\wedge_{max}}|}$ where $|C_{\wedge_{ij}}|$ is the number of Connected Triples between i and j and $|C_{\wedge_{max}}|$ the maximal number of Connected Triples between any two authors. This simple similarity function can be systematically improved by considering the amount of publications which lead to the number of Connected Triples as well as the distribution of authors in these publications. Therefore the edges in the co-author network will be weighted according to Liu et al. [3]. With $V = \{v_1, \ldots, v_n\}$ as the set of n authors, m the amount of publications $A = \{a_1, \ldots, a_k, \ldots a_m\}$ and $f(a_k)$ the amount of authors of publications a_k the weight between two authors v_i and v_j for publications a_k is calculated by $g(i,j,k) = \frac{1}{f(a_k)-1}$. Thereby the weight between two authors for one publication is smaller the more authors collaborated on this publication. Considering the amount of publications two authors i and j collaborated on together, an edge between these authors is calculated with $c_{ij} = \sum_{k=1}^{m} g(i,j,k)$ which leads to higher weights the more publications the two authors share. Applying a normalisation the weight between two authors i and j considering the amount of co-authors and publications is calculated by $w_{ij} = \frac{c_{ij}}{\sum_{r=1}^{n} c_{ir}}$ leading to a directed co-author graph. The similarity of two authors using Connected Triples can consequently be either calculated on incoming edges $(ConnectedTriple_{in} = \sum_{\forall c \in V with\ e_{ci}, e_{cj} \in E, e_{ij} \notin E} w_{ci} + w_{cj})$ or outgoing edges $(ConnectedTriple_{out} = \sum_{\forall c \in V with\ e_{ic}, e_{jc} \in E, e_{ij} \notin E} w_{ic} + w_{jc})$. Evaluations (see Fig. 1(b)), show that the sketched approaches lead to a reasonable precision especially when combined with syntactical criteria.

4 Conclusion

Managing data quality plays an increasing role for service providers in the internet such as DBLP which offers access to bibliographic records. The most time consuming quality task is the task of offering consistent data. Especially person names are error prone and hard to deal with. To confine the problems connected with person names we constantly develop new similarity measures which we evaluate on a specially designed framework making use of new test collections. The promising approaches are then integrated into the data acquisition and maintenance process of DBLP to guarantee a high quality of the data.

References

1. S. Klink, M. Ley, E. Rabbidge, P. Reuther, B. Walter, and A. Weber. Browsing and visualizing digital bibliographic data. In *VisSym 2004*, pages 237–242, 2004.
2. M. Ley. The dblp computer science bibliography: Evolution, research issues, perspectives. In *SPIRE*, pages 1–10, 2002.
3. X. Liu, J. Bollen, M. L. Nelson, and H. V. de Sompel. All in the family?, 2005. online: http://public.lanl.gov/liu_x/trend.pdf.
4. T. C. Redman. *Data Quality for the Information Age.* Artech House, 1996.
5. P. Reuther. Personal name matching: New test collections and a social network based approach. *Universität Trier, Technical Report*, 06-01, 2006.