Probabilistic Representations for Integrating Unreliable Data Sources

David Mimno, Andrew McCallum and **Gerome Miklau**
Department of Computer Science
University of Massachusetts
Amherst, MA 01003

Abstract

Databases constructed automatically through web mining and information extraction often overlap with databases constructed and curated by hand. These two types of databases are complementary: automatic extraction provides increased scope, while curated databases provide increased accuracy. The uncertain nature of such integration tasks suggests that the final representation of the merged database should represent multiple possible values. We present initial work on a system to integrate two bibliographic databases, DBLP and Rexa, while maintaining and assigning probabilistic confidences to different alternative values in merged records.

Introduction

The development of machine learning and web mining technology has led to the introduction of large automatically generated databases. These databases often overlap with manually generated databases, which tend to be more accurate but also smaller and more narrowly focused. Integrating these resources has important benefits for scope and reliability, but such integration is difficult and can often only be done approximately. We present initial work towards a method for merging such asymmetric databases while maintaining alternatives with differing levels of confidence.

We focus specifically on two bibliographic databases: DBLP[1] and Rexa.[2] DBLP consists of manually entered records for papers published in Computer Science journals and conferences. The Rexa digital library consists of Computer Science papers downloaded from the web, with information such as author and title identified using a conditional random field (CRF) extractor (Peng & McCallum, 2004). Due to the nature of the two databases, we expect that many of the records in Rexa represent the same paper entities, for example due to PDF and PostScript versions of the same paper, while we expect, with high confidence, that records in DBLP will refer to unique entities.

Integrating curated and extracted databases offers advantages for both source databases. Automatically extracted databases provide larger volumes of data, both in terms of the number of records and the amount of data in each record. For example, the Rexa database includes workshop papers, technical reports, and theses, which are not indexed by DBLP, while also including abstracts, lists of references, and analysis from data mining tools. Such data would be prohibitively expensive and time-consuming to enter by hand. The data in manually curated databases, on the other hand, is significantly more accurate and reliable than that in automatically created databases due to inevitable inaccuracy in automated information extraction. In addition, manually entered databases often contain information that is difficult to extract such as publication venues and page numbers, which may not be present in PDF documents.

Data integration is known to be a difficult problem, even in cases in which the source databases are considered to be reliable. Given the uncertain nature of both the data sources and the integration process, any system that simply chooses the single best alternative is bound to be less reliable than one that represents uncertainty over possible instances. We therefore use a probabilistic representation of the integrated database. By maintaining the uncertainty of the data, we avoid forcing the integration system to make choices based on insufficient information. A probabilistic representation of merged records may also alleviate common user complaints about the inaccuracy of information extraction (IE) systems. By presenting multiple alternatives, an IE-based database could allow a more accurate representation the system's uncertainty while also potentially supporting a simple multiple-choice mechanism for soliciting user corrections.

Integration in bibliographic databases is somewhat different from typical integration application domains such as census records because of the relational nature of the data. A record for a paper contains one or more authors, so we must not only merge DBLP records with one or more Rexa records, but also align the authors

[1] http://www.informatik.uni-trier.de/ley/db/

[2] http://rexa.info

- (DBLP 1) *Integrating Databases with Machine Learning.* Deborah Anderson.
- (DBLP 2) *Data Integration from Information Extraction with Confidence.* Albert Smith, Bernard Müller.

Figure 1: Example records representing a human-curated bibliographic database.

listed for the DBLP record with the authors extracted by Rexa.

Merging Records

An XML version of the DBLP database is available from the DBLP web page. The Rexa data used in this paper consists of several million instances, including full text PDF and PS documents as well as citations to other papers from the references sections of those documents.

The Rexa system merges DBLP records with papers and bibliographic references from the web using string distance metrics. This process begins by selecting a "seed" title and looking for other papers with the exact same title and the same last name and first initial for the first author. The system next looks for other papers within a preset string distance from the seed title. We present a running example in Figures 1 and 2, which is illustrative of typical variations in extracted records such as truncated titles, variant spellings and extraneous or missing authors. In the example we focus on title and author information, because those fields are most likely to be included in both Rexa and DBLP representations of a paper.

Based on our knowledge of how each database was constructed, we have different expectations for the quality of records in each source. Specifically, we do not expect that any paper entity will be represented in DBLP in more than one way. The same is not true for Rexa: although the system will not process the same bitwise-identical file more than once, papers frequently appear in more than one format. In addition, if information from the references section of papers is used, we naturally assume that many papers will appear frequently.

As a result of this assumption, we align multiple records from the Rexa database with single records in DBLP, but not the other way around. Furthermore, we make no attempt to merge records in Rexa that do not correspond with any DBLP record, such as example record Rexa 4.

Probabilistic Representation of Merged Records

We store merged records in a probabilistic representation similar to that used by the TRIO project (Sarma *et al.*, 2006). In this representation, each row in a table represents a real-world entity that may or may not exist and may have one or more possible alternative sets of attributes. For example, the title of an integrated record derived from DBLP 2 and Rexa 2 has two alternatives: the full title and the truncated title. We store both the probability that a tuple exists in the database and the probability that a tuple takes on a particular set of attribute values, given that it exists.

- (Rexa 1) *Data Integration from Information Extraction with Confidence.* Albert H. Smith, Bernard P. Mueller.
- (Rexa 2) *Data Integration from Information Extraction.* Albert Smith, Bernard Müller.
- (Rexa 3) *Data Integration from Information Extraction with Confidence.* Albert Smith, Bernard Müller, Cathy Jones.
- (Rexa 4) *Information Extraction with Conditional Random Fields.* Cathy Jones.

Figure 2: Example records automatically extracted from web documents. Such records often contain variations in author names (either mistakes or normal variation, as in Rexa 1), incorrect field segmentations (as in the title of Rexa 2), and extraneous fields (such as the third author in Rexa 3).

We represent two types of entities: papers and authors. Note that we are not currently attempting to represent coreference between authors on different papers, although this is an area for future work. We represent each entity with two tables, one for entities and one for entity attributes. The first contains a tuple for each merged record and a probability indicating confidence that the record exists. The second contains one or more rows for each row in the entities table, each representing alternative sets of attributes along with the conditional probability that the entity takes on those attributes, given that it exists. The probabilities for each table are defined as follows:

Merged paper entities. We define two constant parameters, α_{DBLP} and α_{Rexa}, where $0 < \alpha_{Rexa} < \alpha_{DBLP} \leq 1$. These constants represent our *a priori* confidence in the output of each database: we believe that DBLP data is more reliable than Rexa data, so we place higher confidence on records that appear in DBLP. The probability is α_{Rexa} if a record is based on Rexa alone and α_{DBLP} if it is based on DBLP alone. For records that are based on a DBLP record a and Rexa records $b_1, ..., b_k$, the probability is

$$\alpha_{DBLP} + (1 - \alpha_{DBLP}) \left(1 - \alpha_{Rexa}^{k}\right).$$

By raising α_{Rexa} to the kth power, we assume that the multiple Rexa records each provide independent evidence. Thus, the addition of increasing amounts of supporting evidence (i.e. matching papers from the web) increases our confidence that a paper exists proportional to our confidence in the automated extraction.

Merged authors. The probability value is calculated in almost the same way as for paper entities, using confidence parameters γ_{DBLP} and γ_{Rexa}, defined in the same way as the α parameters. The difference is that while we make no assumption that a paper entity is in one database given that it is in the other, we do expect merged records to agree on the author lists. Therefore the probability of an author mention that appears in Rexa but not DBLP is $\gamma_{Rexa}(1-\gamma_{DBLP})$. In the case where a DBLP author is missing from Rexa, on the other hand, we use γ_{DBLP}

Once we have merged paper instances into distinct entities, the next step is to create a distribution over possible values for each attribute field. For paper entities these attributes include title, abstract, year of publication, and publication venue. We present two methods for constructing probability distributions over distinct values for a given attribute based on the number of contributing data sources that have that value and the confidence we place on those data sources.

The first method, which we call *constant*, involves dividing the probability mass between DBLP and Rexa values in a constant proportion regardless of the number of distinct sources in Rexa.

Merged paper attributes (constant). For paper entities that are derived from only one database, there is one alternative with probability 1. For entities derived from both databases, the probability is based on constant parameters β_{DBLP} and β_{Rexa} such that $\beta_{DBLP}+\beta_{Rexa}=1$. The probability mass assigned to β_{Rexa} is divided uniformly between the attribute values of all merged records from Rexa. The final probability of a distinct set of attribute values is thus the weighted sum of the source records that contain exactly those values. In the example, DBLP 2, Rexa 1 and Rexa 3 contain the same title, so the probability of the untruncated title is $\beta_{DBLP}+2\frac{\beta_{Rexa}}{3}$ and the probability of the truncated title in Rexa 2 is $\frac{\beta_{Rexa}}{3}$

Merged author attributes (constant). Probabilities for author name alternatives are defined as for paper attributes, using constant parameters ζ_{DBLP} and ζ_{Rexa}, defined in the same way as the β parameters.

Note that if $\beta_{DBLP} > \beta_{Rexa}$, the DBLP value will always have higher probability than any other value, regardless of the number of extracted references discovered by Rexa. The second method, which we call *proportional*, allows a sufficient number of extracted instances to "override" the DBLP value.

Merged paper attributes (proportional). For paper entities derived from both DBLP and Rexa sources, the probability of the DBLP value i of a given attribute a is

$$\frac{k_{DBLP}+N(a,i)}{k_{DBLP}+\sum_i N(a,i)} \tag{1}$$

where $N(a,i)$ is the number of Rexa instances of the paper that contain value i for attribute a. The parameter k_{DBLP} represents the weight we put on a value from DBLP relative to a single reference extracted by Rexa. The probability of a value i that does not occur in DBLP is therefore

$$\frac{N(a,i)}{k_{DBLP}+\sum_i N(a,i)}. \tag{2}$$

For example, if $k_{DBLP}=5$, there must be at least six Rexa instances with a single value for the DBLP value not to have the highest probability.

Table 1: A probabilistic representation of integrated paper entities using the constant method with $\alpha_{DBLP}=0.97$, $\alpha_{Rexa}=0.9$, $\beta_{DBLP}=0.6$ and $\beta_{Rexa}=0.4$.

ID	Title	Pr
1	*Integrating Databases with Machine Learning*	0.97
2	{ *Data Integration from Information Extraction with Confidence* (0.86) | *Data Integration from Information Extraction* (0.14) }	0.98
3	*Information Extraction with Conditional Random Fields*	0.90

Merged author attributes (proportional). Distributions over author names are calculated in the same way as paper attributes.

Probabilistic Queries

Once we have constructed a probabilistic representation for uncertain data extracted from web documents, it is necessary to consider how we will query the probabilistic representation. Given a query, we want to return any record that has a possible attribute value matching that query, ordered by the probability that the record actually takes on that value. Strategies for matching fuzzy queries on probabilistic databases are presented in Dalvi & Suciu (2004). Possible queries include selecting papers with some title and selecting papers by some author. In both cases, we wish to return results in a merged representation such as that shown in Tables 1 and 2, sorted in descending order by probability. In the first case, we need to find matching attribute records and multiply their attribute probability by the existence probability. For a given record we also want to list all the author alternatives in publication order, with alternatives. The second case is similar, but the probability (spelling given author, author given paper existence, paper existence) is slightly more complicated.

Although one of the primary goals of this work is to represent multiple alternative values for individual fields, there are many situations in which it is necessary to present a single answer to users. For each paper and author entity we therefore mark one set of attributes as the "canonical" view. For each field in each entity, we choose the value with the highest probability, breaking ties by preferring longer values. Note that if we use the constant proportion method for constructing probability distributions over attributes, the DBLP value will always be chosen as the canonical value if it is present. Under the proportional method, it is possible for the DBLP value to be overridden if a sufficient number of

Table 2: A probabilistic representation of merged authors using the constant method. The γ and ζ parameters are equal to α and β, respectively.

ID	Name	Pr
1	Deborah Anderson	0.97
2	{ Albert Smith (0.86) | Albert H. Smith (0.14) }	0.98
2	{ Bernard Müller (0.86) | Bernard P. Mueller (0.14) }	0.98
2	Cathy Jones	0.03
3	Cathy Jones	0.9

Rexa instances include the same value, different from the DBLP value.

Experimental Results

We have built a probabilistic representation of a version of the DBLP database integrated with the Rexa digital library. The input consists of 6,581,228 Rexa instances extracted from web documents and 602,738 DBLP records. The merged database contains 2,259,220 merged paper records, of which 217,470 are derived from both data sources. We are interested in determining whether merging a relatively small curated database with a much larger amount of possibly redundant web data can provide a useful representation. The addition of curated DBLP data to extracted Rexa data clearly provides greater reliability for those merged records derived from both sources. We also find that the addition of Rexa data improves DBLP data by expanding the range of attributes available for each record and by providing additional possibilities for attributes that may have several equally correct values.

Full-text Data

Manually constructed databases depend on data entry, which is expensive. Each record in a database such as DBLP or the Library of Congress catalog tends to be relatively short: even keying in a short abstract for each record would be prohibitively expensive. Automatically constructed databases, on the other hand, can easily handle large quantities of data such as the full text of a PDF document. Of the 217,470 records derived from both data sources, 67,462 (31%) include an abstract. Such full-text attributes are useful in giving digital library users a better sense of the content of a paper than titles alone.

Expanded Author Names

The widespread practice of using first initials only in bibliographic references can add considerable uncertainty to coreference and other data integration efforts. Abbreviations can be a particular problem for authors with common last names. Any source of information that can help disambiguate such abbreviated names will be of considerable value.

In constructing bibliographies, authors are often familiar with the authors they cite, so we expect that full names in references will be generally accurate. We looked at instances where the DBLP record listed only the first initial of an author name (for example *J. Smith*), but at least one Rexa record for the same paper included an author with the same last name and a fully spelled out first name with that same first initial (*Jebediah Smith*). Of the 217,470 merged paper entities that contained data from both data sources, 2021 (0.91%) met this criterion. This number seems quite low, but given the difficulty of author coreference, the potential benefit of adding even a small amount of information could be significant.

Expanded Data Fields

Another potential benefit of the merged database is the potential for expanded attribute values. As pointed out before, manually created databases such as DBLP are very sensitive to data entry costs, so the fields that are present in records, such as publication venues, tend to be abbreviated. Bibliographies in research literature are also subject to some abbreviation, but it is not unusual to find fully spelled out attributes such as publication venues.

If we use the proportional method for defining probability distributions on paper attribute values, it is possible for a sufficient number of Rexa instances that share the same value to override the DBLP value. If several data sources independently "vote" for the same value, we place increasing weight on that value. The k_{DBLP} parameter allows us to control how much we trust extracted records. The effect of this parameter is shown in Table 3. As expected, larger values of k_{DBLP} result in fewer overrided values.

We also looked at the effect on attributes of replacing DBLP attribute values with the most represented values in Rexa instances. Randomly selected examples of overrided values can be found in Table 4. Apart from a few differences in punctuation in titles, most of these values are expansions of abbreviated publication venue names. Venues, unlike titles, are frequently shortened, often in unpredictable ways as words can be abbreviated to different degrees (for example *C. ACM*, *Comm. ACM*, and *Commun. of the ACM*). At $k_{DBLP} = 5$, 85.6% of updated values are longer than the values they replace. The distribution of the number of characters added in these updated values is shown in Figure 3(a). Most expanded values added between 1 and 25 characters, although there is a "long tail" of values that had larger expansions. Figure 3(b) shows the same data by the ratio between the length of the DBLP value and the length of the expanded value.

There are several benefits in offering users access to multiple correct values of a given attribute. Researchers familiar with a subject area may prefer a more terse representation of venue names, while students and researchers reading outside their field may prefer longer versions. Having many different alternatives for a given value may also help in automatic disambiguation and integration.

Table 3: The effect of the k_{DBLP} parameter on attribute values in merged records. Under the proportional probability model, if there are more than k_{DBLP} Rexa instances contributing to a given merged paper entity that all have the same value for some attribute, that value will have the largest probability, possibly "overriding" the DBLP value. There are a total of 661,821 distinct attribute values in the 217,470 merged entities.

k_{DBLP}	Updated Values	Percent Updated
1	181249	27.3%
3	62518	9.4%
5	32030	4.8%
7	19621	3.0%
9	13489	2.0%

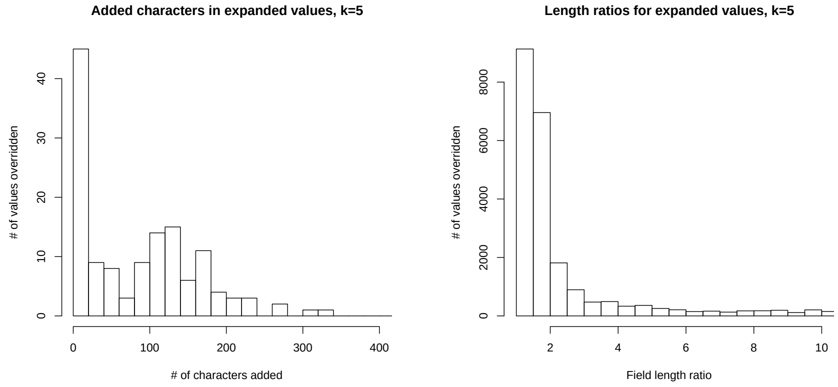
Table 4: Randomly selected samples of DBLP values overridden by Rexa values, $k = 5$

DBLP	**Rexa**
IEEE Trans. Pattern Anal. Mach. Intell.	IEEE Transactions on Pattern Analysis and Machine Intelligence,
IJCAI	In Proceedings of the Eleventh International Joint Conference on Artificial Intelligence,
IEEE Trans. Software Eng.	IEEE Transactions on Software Engineering,
RE	In Third IEEE International Symposium On Requirements Engineering RE'97,
Commun. ACM	Communications of the ACM,
IEEE Trans. Pattern Anal. Mach. Intell.	IEEE Transactions on Pattern Analysis and Machine Intelligence,
A Network-Conscious Approach to End-to-End Video Delivery over Wide Area Networks Using Proxy Servers	A network conscious approach to end-to-end video delivery over wide area networks using proxy servers
Inf. Process. Lett.	Information Processing Letters,
ICRA	In Intl. Conf. on Robotics and Automation,
IEEE Trans. Software Eng.	IEEE Transactions on Software Engineering,
J. Log. Comput.	Journal of Logic and Computation,
OOPSLA	In Proceedings of the 14th Annual Conference on Object-Oriented Programming Systems, Languages and Applications,
Transformational Programming - Applications to Algorithms and Systems	Transformational programming – applications to algorithms and systems
ACM Trans. Program. Lang. Syst.	ACM Transactions on Programming Languages and Systems,
SC	In Supercomputing
DAC	In Design Automation Conf.,
IEEE Trans. Pattern Anal. Mach. Intell.	IEEE Transactions on Pattern Analysis and Machine Intelligence,
POPL	In Proceedings of the 26th Annual ACM Symposium on the Principles of Programming Languages,
SIGPLAN Notices	ACM SIGPLAN Notices,
Design of Logical Topologies for Wavelength-Routed All-Optical Networks	Design of logical topologies for wavelength-routed optical networks
UAI	In Proc. 15th Conf. on Uncertainty in Artificial Intelligence,

Related Work

The use of probabilistic representations in information integration has been explored for many years both in the database and information retrieval communities (Fuhr & Rolleke, 1997). Florescu, Koller, & Levy (1997) present a system that prioritizes queries to information sources based on statistical information. Recent work in probabilistic databases such as Dalvi & Suciu (2004) and Sarma *et al.* (2006) has renewed interest in this area. Menestrina, Benjelloun, & Garcia-Molina (2006) present the Koosh algorithm, which uses several properties of record comparison algorithms to identify match and merge orderings that minimize the number of comparisons. The authors also examine the use of thresholds based on cheap distance functions to limit the number of evaluations of more expensive distance functions. Shen *et al.* (2007) and Doan *et al.* (2006) apply data integration techniques to bibliographic databases, including DBLP, in the DBLife system. The focus of their work, however, is on choosing the best matching algorithms for comparing different data sources, and not on maintaining a probabilistic representation of the results of integration.

Gupta & Sarawagi (2006) propose a probabilistic representation for the distributions over segmentations output by a CRF entity extractors for a single text string. Rexa also uses a CRF to extract named entities from unstructured text, but in this paper, rather than representing the probabilities of multiple segmentations of the exact same string, we represent the single best segmentation of multiple distinct but coreferent strings. In the first case, the CRF extractor implicitly provides the probability distribution over segmentations, so it is not necessary to define one. We present two models for the second case, in which there is no existing distribution over sources of data such as records from structured bibliographic databases and multiple references to the



(a) (b)

Figure 3: In cases where DBLP attribute values were overridden by longer Rexa values, most expanded attribute values added between 1 and 20 characters.

same paper. Incorporating a more sophisticated representation of the output of a CRF on a single string is an interesting area for future work.

Conclusions and Future Work

The goal of this work is to combine web mining data, which has broad coverage but inconsistent accuracy, with manually created data, which has limited coverage and high accuracy, into a database that is both broad and reliable. Although this project is still in its preliminary stages, we have demonstrated several ways in which a probabilistic representation of information extracted from web pages can enhance the usefulness of the merged database. Attributes such as abstracts and lists of references, which are impractical to enter manually, are available for a substantial portion of the DBLP records. In addition, in many cases there are multiple valid values for a given attribute, such as full and abbreviated venue names.

One interesting area for further work is in integrating multiple CRF segmentations of single data sources and taking into account the confidence values generated by the CRF. Other possible applications include expanded search functions, in which canonical values might be returned even if alternative values match the query, and user interfaces that take advantage of alternative values, for example by allowing users to select longer or shorter venue strings in bibliographic entries or by presenting multiple choices when soliciting users corrections.

Acknowledgments

We would like to thank Adam Saunders for many helpful discussions and his tireless work on the Rexa digital library.

This work was supported in part by the Center for Intelligent Information Retrieval, in part by The Central Intelligence Agency, the National Security Agency and National Science Foundation under NSF grant #IIS-0326249, in part by NSF grant # CNS-0551597, in part by NSF Nano # DMI-0531171, and in part by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under contract number NBCHD030010. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

Dalvi, N., and Suciu, D. 2004. Efficient query evaluation on probabilistic databases. In *VLDB*.

Doan, A.; Ramakrishnan, R.; Chen, F.; DeRose, P.; Lee, Y.; McCann, R.; Sayyadian, M.; and Shen, W. 2006. Community information management. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*.

Florescu, D.; Koller, D.; and Levy, A. 1997. Using probabilistic information in data integration. In *VLDB*.

Fuhr, N., and Rolleke, T. 1997. A probabilistic relational algebra for the integration of information retrieval and database systems. *ACM Transactions on Information Systems* 15(1):32–66.

Gupta, R., and Sarawagi, S. 2006. Creating probabilistic databases from information extraction models. In *VLDB*.

Menestrina, D.; Benjelloun, O.; and Garcia-Molina, H. 2006. Generic entity resolution with data confidences. In *VLDB Workshop on Clean Databases*.

Peng, F., and McCallum, A. 2004. Accurate information extraction from research papers using conditional random fields. In *HLT-NAACL*.

Sarma, A. D.; Benjelloun, O.; Halevy, A.; and Widom, J. 2006. Working models for uncertain data. In *ICDE*.

Shen, W.; DeRose, P.; Vu, L.; Doan, A.; and Ramakrishnan, R. 2007. Souce-aware entity matching: A compositional approach. In *ICDE*.