ELM: An Extended Logic Matching Method on Record Linkage Analysis of Disparate Databases for Profiling Data Mining

Colin Conrad*, Naureen Ali†, Vlado Kešelj‡, Qigang Gao§
Faculty of Computer Science
Dalhousie University
Halifax, Canada
Email: *colin.conrad@dal.ca, †naureen.ali@dal.ca, ‡vlado@cs.dal.ca, §qggao@cs.dal.ca

***Abstract*—As predictive marketing and customer profiling solutions have become more sophisticated, they have increasingly become dependent on data from external sources. In order to utilize this data, records must be linked to internal records without the use of unique identifiers. The Extendable Logic for Matching (ELM) performs probabilistic matching from disparate sources and classifies matches according to discrete values reflective of their utility. Sets of matching rules are evaluated based on their performance on supervised classification tasks. High performance on a classification task is indicative of congruity with the real-world entity concerned, giving a sense of matching quality without the use of a gold standard. A set of matching rules generated using name and address was compared to a set which was matched using exact string comparison. We conclude that exact string comparison is a superior method for matching on highly sparse demographic data from disparate sources.**

***Keywords*-data engineering, business analytics, predictive marketing, prospect mining, record linking**

I. INTRODUCTION

Record linkage can be broadly defined as the procedure of identifying when a set of records describe the same entity in reality. Record Linkage is a subject within Data Matching, which also concerns other subjects such as entity resolution and duplicate detection. Though duplicates and entity resolution are common business problems, historically, record linkage concerned a relatively narrow set of circumstances where it was necessary to determine whether two database records referred to the same individual. Often, this was the case with historical census data or medical records [1] [2]. More recently however, record linkage has been of interest in natural language processing, entity disambiguation [3], and has gained interest in data mining for electronic commerce.

There are a large number of public and proprietary databases that contain demographic information, historical purchasing records or social data. Organizations hoping to utilize this data to create comprehensive customer profiles are faced with an immediate problem: this data cannot be matched by means of a common unique identifier, which forces matching procedures to utilize data features for matching. Behavioural record databases, purchase data, for instance, might contain names and purchase receipts, while demographic and wealth summaries might contain considerable details about individuals name, address, income and family, but contain fewer numbers of records. These disparate records must be matched and integrated in order to perform effective information extraction necessary to develop optimal business process and marketing practices. Different databases might contain different levels of data cleanliness, such as missing records or entry errors, and inconsistent data format across databases. Importantly, in many of these situations there is no objective "gold standard" with which to measure the matching quality between databases.

To solve these problems, marketing research professionals often adopt industry standards for matching across databases and performing data integration. Current industry matching practices are varied, developed ad-hoc and often utilize a number of matching techniques to integrate data from distributed sources. In the case of predictive marketing or customer prospect research the cost of a false positive match is high. Matching is often performed on a prospect by-prospect basis, which results in a tuple-centric integration process. Matching techniques are often rule-based and are often validated by end users through trial and error. Organizations offering data services thus run the risk of alienating end users, as users are forced to validate the matching quality by risking their business practices.

Our research concerns record linkage from disparate sources for the purpose of online customer profiling. Our challenge is to develop a method for performing matches on a tuple-by-tuple basis and develop a consistent method for evaluating matching techniques without burdening end users. As we discuss in this paper, standard record linkage techniques are not necessarily well designed for marketing and prospect research from unrelated databases, as they are built using examples where ground truth can be assessed. Our approach is to try multiple matching techniques simultaneously and evaluate them using extrinsic criteria unique to the marketing prospecting problem. By utilizing behavioural data that are unique to marketing research databases, we are able to demonstrate a data-driven technique for evaluating matching quality based on supervised classification performance.

The remainder of the paper is organized as follows. In section 2, we describe the record linkage literature, examine approaches and existing open-source data matching tools. In section 3, we analyze the proposed data matching methodology and its architecture. Section 4 discusses the experiments performed using the proposed methodology and its evaluation. In section 5 we conclude and recommend future research.

II. RECORD LINKAGE SYSTEMS

Record linkage is not a new subject. Record linkage problems date back to the early 20th century and attempts to reconcile US Census data and health records [1]. In an era when most records were maintain by hand, the process of identifying matching pairs was often tedious especially in light of errors in human hearing. To compensate, Soundex algorithms were developed and later adopted by census bodies [4]. In addition to Soundex, researchers such as Howard Newcombe developed an odds approach to identify data features that were indicative of likely matches [1] [5]. Record matching by comparison of string features such as name, address and birthdate was quickly adopted by industry. We call this procedure the *deterministic* rule-based approach. Deterministic rule-based approaches could include matching based on data features, such as name, address or phone number. These are often used in industry to perform record linkage tasks, while specific rules are often tailored to the nature of the data at hand [6]. In contrast to the deterministic rule-based approach, *probabilistic* approaches determine a match based on probabilistic values. Records are considered as match, if data features match within some tolerance and are considered not-match, if they differ [7]. Various string comparison algorithms are used to compute the similarity of the data features of the records. Matching rules based on the features similarity in the probabilistic approach can be predefined or it can be generated using machine learning approach.

A. Machine Learning Approaches

In contrast to rule-based approaches, machine learning can be used to generate data-driven rules that identify possible matches. Record linkage problems can be overcome by solving supervised learning tasks, such as classification problems. Classification process is a process where system learned matching rules from a labelled training dataset. On the basis of learned rules, unlabelled data are classified into predefined classes, and then evaluated according to the labels [6]. In order to perform supervised learning tasks, data labels must therefore be present. This is not often the case with disparate sources, as there is rarely criteria for judging the truth value of the integrated data without referring to the datas real-world reference.

Without extensive labelled data, one needs to consider options for unsupervised learning. Unsupervised learning approaches to record linkage problem start with identifying the possible clusters of match, or possible matches. These clusters can then be used to label the records. Classification can then be performed on these labelled data to generate the matching rules [6]. Theoretically, clustering can be used in case of training dataset not being present but these features are not currently supported by open-source matching tools. Clustering is instead used for indexing and blocking to reduce the number of possible matches to optimize the matching process.

B. Available Solutions

There are many open source matching solutions available in the market. However, none of these solutions were effective for the unique nature of our challenges. Some tools such as OYSTER [8], D-Dupe [9] and Duplicate Detection (DuDe) [10] provide entity resolution or deduplication, and not record linkage at our desired scale. Active Atlas [11] and Multiply Adaptive Record Linkage with Induction (MARLIN) [3] require quality training dataset to identify matching rules. As discussed, this is often not available when mining from disparate sources with unlabelled matches.

Some tools do perform record linkage, but have data source restriction of using dataset of particular format. For example MergeToolBox (MTB) [12] accept comma separated values (CSV) and STATA format, Freely Extensible Biomedical Record Linkage (FEBRL) [13] supports CSV, Tabulator separated values (TAB) and COL 3 (column oriented values with fixed-width fields). In our case, the scale of the data requires a tool that can support relational databases, such as MySQL. Some tools impose restrictions of using only two datasets. Marketing profiling often requires the ability to compare and integrate records from multiple sources, in our case, up to seven. These and other difficulties led us to come up with our own tool named ELM, which is optimized for matching and integrating large numbers of records using multiple matching rules, across multiple databases.

C. Performance Evaluation

The most salient challenge of our data matching problem is the performance evaluation. In the case of disparate personal information, there is no effective method for data labelling short of contacting the marketing prospects and asking for their address information [1]. Not only is this impractical for data at a large scale, it is also detrimental to the business interests of many prospect marketing businesses.

There have been attempts to work around this problem by testing matching success on pre-validated data [14], but in many cases, a record linkage model cannot be tested this way, as there is no pre-validated data available. With no pre-validated data, such as the case of matching profiling data from disparate sources, there is no way to verify the performance of the record linkage system. Metric-based evaluation methods could be developed based on distance criteria or rules [15]. However, when we are in the process of testing the performance of these very rules, such metrics cannot be used to evaluate their own performance. There is clearly a need to develop an alternative method for measuring the performance of matching specifically for marketing profiling problems.

III. EXTENDABLE LOGIC FOR MATCHING (ELM)

Given the lack of strong identifiers across datasets and need to minimize false positives on matches, we determined that the best course of action would be to design an Extendable Logic for Matching (ELM). The Extendable Logic for Matching (ELM) is a Java-based tool used to perform record linkage

tasks by comparing tuples from a source table with n others. Unlike other approaches, ELM is optimized to perform database record linkage and match evaluation using multiple methods simultaneously. The ELM approach can then evaluate multiple techniques using extrinsic task-dependent criteria, such as predictive accuracy.

The overall architecture of ELM is depicted in Figure 1. Disparate datasets and a set of rules defined in XML format are the inputs to the developed tool. After cleaning and matching processes of the data, output results are stored in the database table. ELM maximizes linkage performance while ensuring extendibility. The tool is extendable in the sense that users can add any number of rules in the XML file without impacting the overall program.

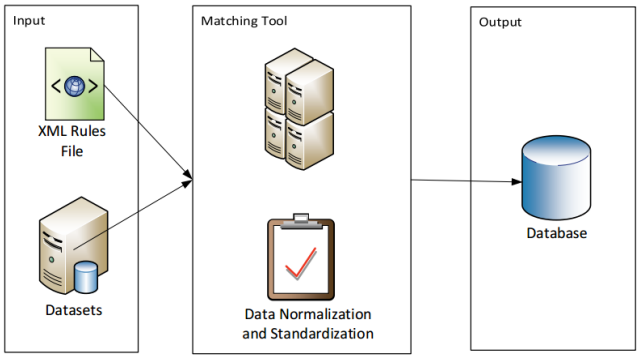

Fig. 1. ELM Architecture

ELM data matching ELM utilizes five steps to match the data. Though these steps are common with most record linkage systems, the specific implementation of these steps will be discussed in turn.

A. Data Preprocessing

Disparate data sources are often unclean and not prepared for data matching. The ELM data preprocessing routine determines whether datasets are in the same format. In order to match names and address, first important step is to clean and prepare the data. If nicknames are not considered in name matching then it can result in errors. For the experiments, a dictionary of nicknames was created and inserted in the datasets. Similarly, some dataset has complete state name and others has state code. For consistent data, we created a dictionary of states and cleaned the datasets to have state code in all datasets. We can avoid typing errors in names by using a soundex algorithm. Just like nicknames, we have inserted additional columns in the datasets where last name soundex needs to be compared. To keep things simple, we are using soundex mainly where we have address or city/state/zip code available for matching. If a dataset does not have address fields then we have not used soundex, in an effort to reduce false positives.

B. Blocking

Blocking is the phase to reduce the number of executed matches by blocking on specified parameters. For our experiments, we used blocking technique to fetch only those record sets whose firstname and lastname are same instead of matching each record with each other record that is if we are matching a dataset with N rows with a dataset with M rows then blocking improves the performance by reducing the number of comparison from NxM comparisons to NxQ where Q will be less than M. These matching record sets are then evaluated according to the record comparison rules. Records in the dataset that is not matching as per the rule get discarded and records that are matching are inserted in the output database.

C. Record Comparison

Blocked tuples are compared using different comparison functions and comparison rules. Different matching criteria and rule sets are specified through an XML rule sheet. The rule sheet guides specifies the rules for the program to follow through the comparison routine. In the aforementioned test on personal data test, the disparate datasets did not always include addresses. In our specific case, two databases only contained names and addresses, so one rule using exact name matching was performed, while a second rule performing exact name matching while considering nicknames and last name soundexes was compared. A third rule set using addresses were also compared, and this set used exact name matching and an approximate address string matching function using the Levenstein distance algorithm. Levenshtein distance is special case of the edit distance, defined as the smallest number of alterations of single characters (insertions, deletions and substitutions) required to convert one string from another [1] [16]. In this way, we demonstrated how ELM can be utilized to implement multiple matching rules simultaneously over the same matching task.

D. Record Classification

In Record Classification phase, records are classified as match, non-match or partial match according to each defined rule. The discrete execution of matching rules is used to classify records according to ELMs classification function. Records are classified as these discrete matching values according to the criteria specified by the user. User criteria could include matching rules derived from clustering or similar machine learning, though our experiments were restricted to sets of simple logics described above. In our case, match status was restricted to comparisons that attained exact string matches on name and an address comparison with 95% similarity on Levenshtein distance assessment. Though record classification techniques can vary, ELM does not support probabilistic record integration. The XML rule sheet must utilize comparison rules and is not compatible with fuzzy techniques.

The following pseudocode describes how the XML rules and matching criteria are defined. In this example, a political donations database is matched with internal company records, and three rulesets are created for evaluation.

DataSet: Political

– Rule 1
—- Criteria: FirstName (exact string)
—- Criteria: LastName (exact string)
– Rule 2
—- Criteria: AlternateName (exact string)
—- Criteria: LastName (exact string)
– Rule 3
—- Criteria: FirstName (exact string)
—- Criteria: LastName (exact string)
—- Criteria: Address (Levenshtein)

By defining multiple rules as such we can easily apply evaluation to determine the appropriate matching algorithm. If rules are satisfied, a record for that rule satisfaction is kept in a separate table, which can be used to link the two database records using the different criteria. Evaluation can then be applied using the linked data.

E. ELM Extrinsic Evaluation

As stated, there is no established method for evaluating matching performance for our problem, and there is a clear need to evaluate the success of the different rules. Broadly defined, evaluation is the process to determine the correctness and completeness of the matched data. We opted to develop a method of external evaluation for the matching routine. External evaluation is the analysis of performance based on reference to a larger task that includes matching as one of the components. In our case, we are developing marketing prediction models. We therefore measure performance with reference to performance of the predictive model.

This is not without precedent. External evaluation is regularly performed when measuring success in clustering, and is less common among other tasks. Like record linkage problems, clustered data frequently lacks truth value labels. Instead, clusters can be evaluated based on how they conform to the model, or the ground truth [17]. For instance, data clusters formed from Fischer's iris data might be expected to form clusters that reflect the four iris species [18]. Likewise, ELM matching performance is measured externally by the performance of the marketing profiling task. Assuming that the matched data contains predictive power, better performance on profile classification is indicative of better matching quality. Thus, if two datasets are prepared for predictive marketing using different sets of matching rules, matching could be evaluated by measuring differences in classification accuracy or precision at the same classification task. This, of course, assumes that the marketing classification model is in fact reflective of the ground truth. In the case of many common marketing tasks, such as psychographic profiling, customer labels might be better understood as useful fictions [19].

Concerns might be raised about the hidden structures in the marketing data. In clustering, unsupervised techniques might identify traits that are mathematically more significant than the normative data categories identified by humans [20]. When using support vector machines for hand-written character recognition (for example), significant noise might result in clusters that are far from the normative categories. If marketing profile categories are useful fictions, might the predictive performance be underdetermined by the profiling model?

This is an important concern when considering data mining for marketing research, though in our case, we need not be concerned. Though classification performance might be underdetermined by the profiling model (eg. income, psychographics, etc.), matching performance can be evaluated so long as the model remains the consistent for each evaluation. Though models might be inaccurate or fail to represent the present state of affairs, changes in performance of the same model on the same set of actual prospects would need to be determined by the data quality. If the predictive model is constant during the comparison, only the matched records, or information gained changes. One matching model can thus be determined to gain more information by better performance on the classification task.

IV. Predicting Political Affiliation

To demonstrate our approach, we designed a customer profiling experiment using an industry partners proprietary data. Our task was to identify a US customers political affiliation based on their most recent political donation. We sampled 827 295 records from a demographic database. These records contained data from all of the available prospects residing in a USA state, complete with name, gender, address and employment history data, when available. We then sampled records from two disparate databases. The first contained 211 761 political donation records, while the second contained 1 603 680 charitable donation records. The political dataset contained donations, the political candidate, the candidate party, along with the name and address of the donor. Charitable donations consisted of donation record, the charitable cause, along with first and last names of the donors. All records were processed into separate tables of a single MySQL database processed on an Apache server.

We ran two matching routines on this data through ELM. The first routine (R1) opted to match on the basis of name and address. While name matching consisted of exact string comparisons, address matching utilized the Levenshtein distance algorithm with a 95% similarity threshold. The second routine (R2) was instructed to match the disparate records on the string comparison of first and last names alone. Table 1 (below) summarizes the number of matches for each routine

	Political	Charitable
R1	51 103	320 894
R2	211 815	1 291 699

TABLE I
Number of Records in Each Test

R1 identified 51 103 matches on political donations and 320 894 matches on charitable donations. R2 identified 211 815 political donations and 1 291 699 charitable donations. This significant difference could be attributed to the data quality, as missing address fields were common and discounted from the Levenshtein address comparison. Following the matching

routines, the records were summarized integrated into a single MySQL table. The integrated table was then cleaned using the R programming language and exported to Weka, a java-based data mining tool.

A. Evaluation and Results

In order to predict political affiliation, we opted to perform supervised classification using the C 4.5 decision tree algorithm. We opted to clean redundant or noisy features used during the matching process, such as names and addresses. We thus opted to model the tree to predict political affiliation using five data features mentioned in Table 2.

Feature	Type
Gender	Nominal
Recent Election Donations (sum)	Numeric
Recent Donation Year	Nominal
Largest Charitable Gift	Numeric
Number of Charitable Gifts	Numeric
recent Political Affiliation	Nominal

TABLE II
FEATURES USED IN CLASSIFICATION

In order to facilitate weight classification, the experiment was restricted to prospects with unique recent political affiliation records. Both R1 and R2 contained 30 422 unique political affiliations, though contained different gender and election donation records. Decision trees were made using the C 4.5 algorithm and evaluated using 10 fold cross validation. The results are summarized in Table 3 below.

	Accuracy	Precision	Recall	F-Measure
R1	69.6%	0.699	0.696	0.692
R2	71.3%	0.713	0.713	0.709

TABLE III
FEATURES USED IN CLASSIFICATION

The decision tree made from exact string name matching had higher accuracy, precision and recall compared to the tree derived from the set matched using names and addresses. We thus have grounds for choosing R2 as the optimal matching rules for this prospect marketing task. Of course, different tasks might require different criteria for evaluation and have different matching algorithms. This is why we opt for an extendable system, so that we can evaluate multiple matching logics to determine a suitable one for the prospecting task.

V. CONCLUSION

As predictive marketing based on comprehensive customer profiling becomes increasingly common, there is a clear need for effective, timely and measurable matching from disparate sources that do not have gold standards. This paper has demonstrated a framework that satisfies these demands while providing a means to quickly test different sets of matching rules to find optimal solutions. As far as we are aware, this is the first demonstration of the use of an extendable matching approach specifically optimized for market profiling from disparate sources. It is unique in the sense that it accounts for external evaluation unique to the profiling task.

Future work on this approach should consider incorporating matching rules derived from unsupervised learning methods. When designed specifically to account for features unique to the problem data set, machine learning methods could realize higher quality matching, as reflected by the profiling accuracy. Our preliminary results also reveal that there could be significant advantages to reinvestigating industry record linkage practices, especially when performing customer profiling. The information lost by taking more conservative matching approaches, such requiring prospect addresses, might result in lower classification precision and accuracy, outweighing the perceived benefits of the matching rule.

REFERENCES

[1] P. Christen, Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution and Duplicate Detection, Canberra: Springer, 2012.

[2] D. Dey, V. Mookerjee and D. Liu, "Efficient Techniques for Online Record Linkage," IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 3, pp. 373-387, 2011.

[3] M. Bilenko, S. Basu and M. Saami, "Adaptive Product Normalization: Using Online Learning for Record Linkage in Comparison Shopping," in Proceedings of the Fifth IEEE International Conference on Data Mining, 2005.

[4] M. K. Odell, "The profit in records management," Systems, vol. 20, 1956.

[5] H. B. Newcombe and J. Kennedy, "Record linkage: making maximum use of the the discriminating power of identifying information," Communications of the ACM, vol. 5, no. 11, pp. 563-566, 1962.

[6] X. Wang, Matching Records in Multiple Databases Using a Hybridization of Several Technologies, Louisville, KY: Department of Industrial Engineering, University of Lousiville, 2008.

[7] I. P. Fellegi and A. B. Sunter, "A Theory for Record Linkage," Journal of the American Statistical Association, vol. 64, no. 328, pp. 1183-1210, 1969.

[8] W. Yancey, "BigMatch: A program for extracting probable matches from a large file for record linkage," Tech. Rep. RRC2007/01, US Bureau of the Census, 2007.

[9] H. Kang, L. Getoor, B. Shneiderman, M. Bilgic and L. Licamele, "Interactive entity resolution in relational data: A visual analytic tool and its evaluation," IEEE Transactions on Visualization and Computer Graphics, p. 9991014, 2008.

[10] U. Draisbach and F. Naumann, Dude: The duplicate detection toolkit, Singapore: Workshop on Quality in Databases, held at VLDB, 2010.

[11] S. Tejada, C. Knoblock and S. Minton, "Learning domain-independent string transformation weights for high accuracy object identification," in Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 350359, 2002.

[12] R. Schnell, T. Bachteler and S. Bender, "A toolbox for record linkage," Austrian Journal of Statistics, vol. 33, no. 1 & 2, p. 125133, 2004.

[13] P. Christen, "Development and User Experiences of an Open Source Data Cleaning, Deduplication and Record Linkage System," SIGKDD Explorations, vol. 11, no. 1, p. 3948, 2009.

[14] R. D. Wilson, "Beyond Probabilistic Record Linkage: Using Neural Networks and Complex Features to Improve Genealogical Record Linkage," in International Joint Conference on Neural Networks, San Jose, CA, 2011.

[15] S. Ivie, B. Pixton and C. Giraud-Carrier, "Metric-Based Data Mining Model for Genealogical Record Linkage," in Information Reuse and Integration, Las Vegas, 2007.

[16] G. Navarro, "A guided tour to approximate string matching.," ACM Computing Surveys, vol. 33, no. 1, pp. 31-88, 2001.

[17] S. Gunnemann, I. Farber, E. Muller, I. Assent and T. Seidl, "External Evaluation Measures for Subspace Clustering," in CIKM '11, Glasgow, UK, 2011.

[18] R. A. Fisher, "The use of multiple measurements in taxonomic problems," Anals of Eugenics, vol. 7, no. 2, pp. 179-188, 1936.

[19] W. D. Wells, Life Style and Psychographics, USA: American Marketing Association, 1974.
[20] I. Farber, S. Gunnemann, H.-P. Kriegel, P. Kroger, E. Muller, E. Schubert, T. Seidl and A. Zimek, "On Using Class-Labels in Evaluation of Clusterings," in International Workshop on Discovering, Summarizing and Using Multiple Clusterings, Washington, DC, 2010.