School of Informatics, University of Edinburgh

Institute for Adaptive and Neural Computation

An Expectation Maximisation Algorithm for One-to-Many Record Linkage, Illustrated on the Problem of Matching Far Infra-Red Astronomical Sources to Optical Counterparts

by

Amos J Storkey, Christopher K I Williams
Institute for Adaptive and Neural Computation
School of Informatics
University of Edinburgh
5 Forrest Hill, Edinburgh UK
a.storkey@ed.ac.uk, c.k.i.williams@ed.ac.uk

Emma Taylor, Robert G Mann
Institute for Astronomy
University of Edinburgh
Blackford Hill, Edinburgh UK
elt@roe.ac.uk, rgm@roe.ac.uk

Informatics Research Report 318

School of Informatics **August 2005**
http://www.informatics.ed.ac.uk/

An Expectation Maximisation Algorithm for One-to-Many Record Linkage, Illustrated on the Problem of Matching Far Infra-Red Astronomical Sources to Optical Counterparts

Amos J Storkey, Christopher K I Williams
Institute for Adaptive and Neural Computation
School of Informatics
University of Edinburgh
5 Forrest Hill, Edinburgh UK
a.storkey@ed.ac.uk, c.k.i.williams@ed.ac.uk

Emma Taylor, Robert G Mann
Institute for Astronomy
University of Edinburgh
Blackford Hill, Edinburgh UK
elt@roe.ac.uk, rgm@roe.ac.uk

Informatics Research Report 318
SCHOOL *of* INFORMATICS
Institute for Adaptive and Neural Computation
August 2005

University of Edinburgh, Informatics Technical Report, August 2005

Abstract : The problem of record linkage is often seen simply in terms of making links between data points that might be generated from the same source. However, in many cases the grounds for linking items is itself not certain. In fact it is often desirable to *learn*, in an unsupervised manner, what form linked objects take in different databases. One simple case of this is the "one to many" linkage problem, where each object in one dataset is potentially linked to one of many objects in another dataset, and where the candidate matches are mutually exclusive. We show how the Expectation Maximisation algorithm can be used for this matching problem, both to calculate the probability of a match, and to learn something about the characteristics that matched objects have. The approach is derived for the specific astronomical problem of linking far infra-red observations to optical counterparts, but is generally applicable. This report outlines the theory of this record linkage procedure, but does not discuss its application or any implementational details.

Keywords : Record Linkage, EM algorithm, Astronomical Data Mining

Copyright © 2005 University of Edinburgh. All rights reserved. Permission is hereby granted for this report to be reproduced for non-commercial purposes as long as this notice is reprinted in full in any reproduction. Applications to make other use of the material should be addressed to Copyright Permissions, School of Informatics, University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, Scotland.

1 Introduction

Data integration is becoming increasingly important in many endeavours. One key aspect of data integration is data linkage, the process of making meaningful links between different items in a variety of databases. Due to increasing data sizes it has become critical that automated techniques should play some part in this process. The most common data linkage issue is record linkage: linking together records in different databases that are derived from the same underlying source. This is appropriate for data derived from individual people across different census datasets for example, or for linking similar DNA sequence sections from different organisms, for linking features derived from different images which might contain the same objects, or matching astronomical objects in catalogues from observations at different wavelengths.

Record linkage is a non-trivial exercise. For example in astronomy the same astronomical object might have a very different apparent form at different wavelengths, even discounting the different measurement capabilities we have. In record linkage on census data, people move house, change their name, die, are born, move countries, change jobs, change religious affiliations *etc.* between the census periods. Furthermore the data sizes can be large and the number of potential reasonable matches to any one data point may well be fairly large. There are often matching constraints such as if one record matches to another, it cannot also match to yet another, and it is also likely that the knowledge about what makes a match likely is vague.

Record linkage can reap major benefits. For example it is a vital part of epidemiological study; it was useful in the analysis of the impact of air travel in deep vein thrombosis [6]. The general benefit is that the combined data from linked records allows statements to be made about the data that could not be ascertained from the individual databases alone.

One key research direction in record linkage involves learning the characteristics of linked data. Especially in scientific linkage studies, the characteristics of the linked objects is not necessarily known a priori. It is vital to be able to take steps towards learning what features make a particular record more or less likely to be linked to another record.

1.1 Current Procedure for Record Linkage

Felligi and Sunter [4] developed a framework by Newcombe et al [10] into a mathematical formulation for record linkage, with the emphasis on developing an efficient linkage rule. This model has been subject to significant analysis and development. Thibaudeau [14] notes that the independence across fields assumption results in poor discrimination in more realistic circumstances. A reduction in this assumption (Winkler [16]) creates some improvements. Belin and Rubin [1] note poor false match rates in the Felligi and Sunter Model, and develop a mixture model approach to improve the model, and Larsen and Rubin [7] use maximum likelihood learning amongst candidate models.

A summary of these approaches can be summarised by the following outline:

- Define some similarity measure between records in one dataset and those in another.
- Estimate or specify distributions of this similarity measure in the two cases: where the records match, and where the records are unmatched.
- Use this distribution to classify potential record pairs as either matched or unmatched.

The limitations of this approach are well understood. First it does not take into account any characteristics that matched and unmatched records have. For example in linking astronomical databases, faint objects are much more likely to be unmatched than bright ones. Second it makes assumptions of independence of the pairwise similarity measure. However for the problem of linking two datasets of size m and n, here are mn such similarities for a system with only $m+n$ degrees of freedom. Hence this assumption is invalid and causes failures in cases of, for example, clustered data where within-cluster links are made which should not be. The most important problem of this approach is that, because it does not capture the actual data generation process, there is little room for learning what characteristics matched object have.

Record linkage is necessary in many distinct areas. The methods grew up out of the need for linkage in census data [17], but there are significant uses in medical data [11], genealogy [12], astronomical data [13] and citation indexing [5, 15]. Similar methods, albeit with a slightly different history, are used in image registration [2, 18] and feature matching [8]. There are related problems in phylogeny and biological sequence alignment, both of which can be seen in terms of matching [3]. Coupled clustering [9] is a relevant technique which uses linked information from multiple datasets to enable better clustering of data than can be obtained from one.

1.1.1 Learning

One of the key issues in record linkage which is not addressed by the current formalisms is how to learn about the characteristics that linked data has. The Felligi-Sunter model has the facility to adapt the choice of linking or not linking dependent on the level of certainty, but does not have the ability to adapt the grounds for making a link. When the characteristics of linked objects is not fully known a priori, such as in a data mining problem, then the facility for learning the characteristics of linked data is vital in understanding not just the certainty of the link, but the actual grounds for which a link should be made.

This learning-linkage issue is the main focus of this report. The issue will be addressed in a particular simplified circumstance (where link candidates are mutually exclusive), but nevertheless one that covers many uses.

1.2 Context

The methods outlined in this paper can be applied to a general set of problems. However, they are formulated here in terms of a specific problem in order to make the applicability clearer and to provide an obvious grounding to the methods.

The focus of this paper is the issue of finding corresponding optical records to astronomical objects which emit in the far infra-red region of the spectrum. A few astronomical objects emit in the far infra-red, but at the moment little is known about the characteristics of such objects. Furthermore, far infra-red observations have a low spatial resolution. It would be helpful in trying to understand such objects if optical counterparts could be obtained, that is if each infra-red record could be paired with a particular record in an optical database referring to the same underlying object.

The low spatial resolution of the far infra-red measurements makes the task difficult: there are sometimes up to 100 optical records which could potentially be the correct candidate on position alone. In addition not much is known about what characteristics far infra-red emitters have in the optical spectrum. Hence any approach to record linkage will need to be able to learn these characteristics from the data.

2 The Model

We presume that we have an astronomical dataset consisting of a number of objects (stars or galaxies) which emit radiation in the far infra-red region of the electromagnetic spectrum, along with some measurements from such objects, such as position, flux[1] etc. For each infra-red object we also have a list of possible optical records which match. Each optical record will consist of various optical measurements of the object, including position (which is much more accurate than the infra-red position measurement), magnitude, ellipticity shape etc. In addition we could have derived quantities such as the probability that the object is a star or galaxy.

More formally, we presume we have a number M of far infra-red sources, labelled by r, which have known characteristics (eg position, flux). For each of these sources, r, we limit ourselves to considering a limited number of possible optical counterparts: those within a radius R_r from

[1] The flux can be thought of as the magnitude, or brightness of the object in the infra-red.

the measured location of the far-infra object[2]. The number of such optical objects is denoted by N_r. Suppose the relative position of one of those possible optical counterparts, labelled i, to the position of the infra-red source is $\mathbf{x}_i^r$. The magnitude of that optical source is denoted m_i^r. Let z_i^r be an indicator variable that is one if the infra-red source r is associated with the optical object i, and zero otherwise. We also allow the possibility $z_0^r = 1$, representing the fact that the infra-red source has no optical counterpart. Adding the constraint $\sum_{i=0}^{N_r} z_i^r = 1$ ensures that the infra-red source has at most one optical counterpart.

The task we are interested in is twofold. First we wish to be able to learn the characteristics that linked objects have. Second we wish to be able to infer which objects should be linked together. The overall framework we will use to do this is a generative model with the following characteristics.

For the optical records in the region of each infra-red object there are two generative processes. For the *at most one* case where the optical record is a counterpart (i.e. linked to) the far infra-red record, we might expect its position to be local to the measured position of the infra-red source, and have particular characteristics which are different from other optical sources. For the *many* cases where an optical record is not a counterpart, the measurements will be totally independent of the far-infra red object. We can write this model down for each infra-red object r in the following way

$$P(X^r, M^r, Z^r) = P(Z^r) \prod_i P(\mathbf{x}_i^r, m_i^r | z_i^r) \tag{1}$$

where Z^r denotes the family of indicators $(z_i^r)_{i=0,1,\ldots,M}$, M^r the family $(m_i^r)_{i=1,2,\ldots,M}$ and X^r the family $(\mathbf{x}_i^r)_{i=1,2,\ldots,M}$. For $P(x_i^r, m_i^r | z_i^r)$, the two cases $z_i^r = 1$ and $z_i^r = 0$ must be considered, corresponding to the cases where the optical source is or is not linked to the infra red source. It is worth emphasising again that $z_i^r = 1$ for only one value of i and is otherwise zero.

When using this model for inference, the position and magnitude of the optical objects, and the position measurement for the far infra red record are given. Hence by Bayes rule we are interested in finding the posterior values

$$P(z_i^r = 1 | X^r, M^r) = \frac{P(z_i^r = 1) P(\mathbf{x}_i^r, m_i^r | z_i^r = 1) \prod_{j \neq i} P(\mathbf{x}_j^r, m_j^r | z_j^r = 0)}{P_0 + \sum_{i=1}^{N_r} P(z_i^r = 1) P(\mathbf{x}_i^r, m_i^r | z_i^r = 1) \prod_{j=1|j \neq i}^{N_r} P(\mathbf{x}_j^r, m_j^r | z_j^r = 0)} \tag{2}$$

where $P_0 = P(z_0^r = 1) \prod_{j=1}^{N_r} P(\mathbf{x}_j^r, m_j^r | z_j^r = 0)$ is the probability of seeing the data we have in the case that there is no link. The probabilities of (2) are precisely the linkage probabilities, where we have accounted for all the information available.

2.1 Distributions

It can be assumed that many of the parameters given above are known, or failing that, that their distributions are known. For example R_r is chosen to be large enough to definitely include any possible optical counterpart. N_r is determined by that choice of R_r. The measurement error of the far-infra red position $P(\mathbf{x}_i^r | z_i^r = 1)$ can reasonably be assumed to be spherically Gaussian (centred at the true position of the far infra-red object), and for now it is presumed that the variance can be estimated from the characteristics of the instruments.

We assume that, within our radius of interest, the probability of a star or galaxy turning up in one particular position is the same as one appearing in any other position. Hence we set the prior distribution of positions for unlinked optical records using $P(\mathbf{x}_i^r | z_i^r = 0) = \mu_r$, where μ_r is the number of optical objects per unit area. By neglecting the effect of the (at most) one linked object, we can presume $\mu_r = N_r / (\pi (R_r)^2)$ to high enough accuracy. In the event that the spatial distribution of stars and galaxies within the region of interest is not reasonably assumed to be

[2] Theoretically, a priori (before knowing the infra-red location) we should consider every optical record we have to be a possible counterpart. However as we will find later, those optical records at great distances from the infra-red location will have exceedingly small likelihood of being linked, so we can save some computational effort by ignoring them at this stage. Making R_r as big as the whole sky would have negligible effect on the final answers.

homogenous, more localised density measures may need to be used. However in the case here homogeneity is a reasonable assumption, and it would be hard to infer inhomogeneity from the small number (e.g. 50 to 100) of counterparts we are considering.

There are two distributions for optical magnitude. The first is for optical objects which are not counterparts to an infra-red object. This is given by $P(m_i^r|z_i^r = 0)$, and is assumed known, calculated from the overall distribution of optical magnitudes, only a negligible contribution of which will be represented in the far-infra red.

The second magnitude distribution is the critical one. This is the primary unknown which will need to be learnt from the data. This magnitude distribution is that of objects visible in the optical, but which are also far infra-red emitters. As the properties of infra-red emitters is unknown there is no way of setting this distribution a priori. We presume this is represented as a histogram, with B bins, and bin widths of w. Hence

$$P(m_i^r|z_i^r = 1) = \sum_{t=1}^{B} h_t T_t(m_i^r) \tag{3}$$

where T_t is a top hat function that is one if m_i^r lies within the tth bin, and zero otherwise. To ensure correct normalisation to a probability, we have the constraint $w\sum_t h_t = 1$. Hence h_t is the normalised bin height. It is these bin heights that must be learnt. Note that in this derivation we have assumed that the optical magnitude is independent of the flux of the far infra-red object. Including this dependence is theoretically trivial, but will significantly increase the number of parameters as a two dimensional histogram will need to be estimated.

It is also possible that a far-infra red emitter does not emit (or at least does not emit brightly enough to be measured) in the optical. The prior probability that a randomly chosen far infra-red emitter has no optical counterpart is denoted $\gamma = P(z_0^r = 1)$, and is another parameter that must be learnt. If there is a counterpart, then we assume a priori that any of these possible counterparts could be the true one with equal probability, i.e. $P(z_i^r = 1) = (1-\gamma)/N^r$.

2.1.1 Summary

The previous discussion allows a full probabilistic model to be formulated. The likelihood, $P(D|\Theta)$, for the model, where Θ is a compact representation for the parameters we wish to learn, and D is the data $\{X^r, M^r\}$, can be written as

$$P(D|\Theta) = \sum_{Z} \prod_{ri} P(\mathbf{x}_i^r|z_i^r) P(m_i^r|z_i^r) P(z_i^r) \tag{4}$$

where Z denotes the family of all the Z^r for all r. The parameters that need to be learnt are the h_t in $P(m_i^r|z_i^r)$ given in (3), and γ, which is $P(z_0^r = 1)$.

3 The Expectation Maximisation (EM) Algorithm

The previous section provides all the distributions that are needed along with the full likelihood for the model that will be used. In this section, for convenience[3], the basic derivation of the EM algorithm, which allows learning and inference in this model, is provided.

We can show by simple manipulation of probabilities that

$$\log P(D|\Theta) = \langle \log P(D, Z|\Theta)\rangle_{Q(Z)} - \langle \log Q(Z)\rangle_{Q(Z)} + KL(Q(Z)||P(Z|D,\Theta)) \tag{5}$$

where Z is the shorthand for the set of all the z_i^r, and $Q(Z)$ is an arbitrary distribution over Z. $KL(Q||P)$ is the KL divergence and is given by

$$KL(Q(Z)||P(Z|D,\Theta)) = \langle \log Q(Z)\rangle_Q - \langle \log P(Z|D,\Theta)\rangle_Q. \tag{6}$$

[3] For many readers this derivation is obvious and superfluous. However it is included here for readers interested, say, in the astronomical application, who may be unfamiliar with EM.

It is always greater than zero, and is minimised to zero by setting $Q = P$.

Equation (5) is the basis upon which the EM algorithm is developed. We wish to maximise $\log P(D|\Theta)$. As the second term on the right hand side (the entropy) is independent of Θ, whereas the KL divergence is not, we wish to choose Q to minimise the KL divergence. This ensures that changing the parameters by maximising the first term on the right hand side of (5) does not inadvertently reduce the KL divergence at the same time, and thereby *reduce* the log likelihood. This KL divergence is minimised (and is zero) if we set $Q(Z)$ to be the current estimate for the posterior distribution of Z. Hence then maximising

$$\langle \log P(D, Z|\Theta)\rangle_{Q(Z)} \tag{7}$$

can only increase the log likelihood, as the KL divergence can only get larger. Adapting Q to fit the new posterior given the new parameter set does not affect the parameters, and so cannot affect the log probability. Hence this provides an iterative procedure for maximising the log probability.

1. Initialise the parameters.

2. Set Q to be the current estimate of the posterior distribution for Z.

3. Maximise (7) given this Q.

4. Repeat until suitably converged.

4 EM for Record Linkage

In this section the use of EM for the record linkage model is described. As everything is mutually independent between each infra-red source r,

$$\langle \log P(D, Z|\Theta)\rangle_Q = \langle \log \prod_r P(D^r, Z^r|\Theta)\rangle_Q = \sum_r \langle \log P(D^r, Z^r|\Theta)\rangle_Q \tag{8}$$

where again Z^r is shorthand for the family of all z_i^r for a given r, and D^r does the same for the data. Here

$$P(D^r, Z^r|\Theta) = P(Z^r) \prod_{i=1}^{N_r} P(\mathbf{x}_i^r|z_i^r) P(m_i^r|z_i^r), \tag{9}$$

remembering that $z_0^r = 1$ covers the case where no optical counterpart exists. Substituting for $P(z_i^r = 1) = \alpha_i$, where $\alpha_i = (1-\gamma)/N$ for $i > 0$ and $\alpha_i = \gamma$ for $i = 0$, we have

$$\begin{aligned} P(D^r, Z^r|\Theta) = \left(\prod_{i=0}^{N_r} \alpha_i^{z_i^r}\right) \prod_{i=1}^{N_r} \Big[& P(\mathbf{x}_i^r|z_i^r = 1)^{z_i^r} P(m_i^r|z_i^r = 1)^{z_i^r} \\ & \times P(\mathbf{x}_i^r|z_i^r = 0)^{(1-z_i^r)} P(m_i^r|z_i^r = 0)^{(1-z_i^r)} \Big], \end{aligned} \tag{10}$$

where we have used the fact that the z_i^r are indicator variables to pull out the correct components for the linked and unlinked cases. Substituting further we get

$$P(D^r, Z^r|\Theta) = \left(\prod_{i=0}^{N_r} \alpha_i^{z_i^r}\right) \prod_{i=1}^{N_r} \left[P(\mathbf{x}_i^r|z_i^r = 1)^{z_i^r} \left(\sum_t h_t T_t(m_i)\right)^{z_i^r} \mu_r^{(1-z_i^r)} P(m_i^r|z_i^r = 0)^{(1-z_i^r)} \right]. \tag{11}$$

This is the final form of this joint distribution, and given the family of parameters $\Theta \equiv (\gamma, h_t)$, can be calculated directly from the distributions we specified initially. We use equation (9) to calculate the posterior distribution $P(Z^r|D^r, \Theta) = P(Z^r, D^r|\Theta)/\sum_{Z^r} P(Z^r, D^r|\Theta)$, by calculating (9) for all N_r different possibilities that Z^r can take, and then normalising.

Taking logarithms of the above equation, and substituting into (8) we can perform the expectations with respect to Q to get

$$\langle \log P(D,Z|\Theta)\rangle_Q = \sum_r \left[\sum_{i=0}^{n} \langle z_i^r \rangle \log \alpha_i + \sum_{i=1}^{n} \langle z_i^r \rangle \log P(\mathbf{x}_i^r | z_i^r = 1) + \langle z_i^r \rangle \log \left(\sum_t h_t T_t(m_i) \right) \right.$$

$$\left. + (1 - \langle z_i^r \rangle) \log \mu + (1 - \langle z_i^r \rangle) \log P(m_i^r | z_i^r = 0) \right] \quad (12)$$

where all expectations are with respect to $Q(Z)$, taken to be fixed at the current posterior estimate. All the expectations we need are calculable from the marginal posterior distribution $P(z_i^r|D,\Theta)$, obtainable from (11). Note the expectation of an indicator variable is just its probability of being 1.

Only a few parts of Equation (12) are dependent on h^t. Adding a Lagrange multiplier term, $\lambda(w \sum h_t - 1)$, to account for the normalisation constraint on h_t, and taking derivatives we have

$$\frac{\partial}{\partial h_t} \langle \log P(D,Z|\Theta)\rangle_Q = \sum_r \sum_{i=1}^{N_r} \langle z_i^r \rangle \frac{T_t(m_i)}{h_t} + \lambda w$$

Setting the derivative to zero and solving for h_t gives

$$h_t \propto \sum_r \sum_{i=1}^{n} \langle z_i^r \rangle T_t(m_i) \quad (13)$$

where the constant of proportionality is given by the normalisation constraint.

A similar procedure for γ gives

$$\gamma = \frac{\sum_r \langle z_0^r \rangle}{\sum_r \left(\langle z_0^r \rangle + \sum_{i=1}^{N_r} \langle z_i^r \rangle N_r \right)} \quad (14)$$

4.1 Final Procedure

The E-Step involves calculating the marginal posteriors, obtained from (11)

$$\langle z_i^r = 1 \rangle = P(z_i^r = 1 | D^r, \Theta) \quad (15)$$

$$\propto \alpha_i P(\mathbf{x}_i^r | z_i^r = 1) \left(\sum_t h_t T_t(m_i) \right) \prod_{j=1, j \neq i}^{N_r} \left(\mu_r P(m_j^r | z_j^r = 0) \right) \quad (16)$$

for $i > 0$, and

$$\langle z_0^r = 1 \rangle = P(z_0^r = 1 | D^r, \Theta) \propto \alpha_0 \prod_{i=1}^{N_r} (\mu_r P(m_i^r | z_i^r = 0)) \quad (17)$$

and then setting the constant of proportionality to ensure the probabilities sum to one.

The M-Step involves setting h_t and γ according to equations (13) and (14).

5 Extensions

5.1 Mixture-Model Densities

Using histograms to model the conditional magnitude distribution is not ideal. It would be better to use a smoother density model. Two alternatives are either a fixed kernel model or a Gaussian mixture model. For Gaussian kernels we can view the former as a Gaussian mixture model where the kernel widths and positions are fixed. Hence we focus on the Gaussian mixture model. This is a trivial addition to the formalism, as the standard EM learning procedure for mixture models can be used, where the mixture model E step can either be combined with the overall E step, or the mixture model EM can be iterated as a subprocess.

5.2 Star Galaxy Classification

Further variables can be included in the model, such as a flag to label whether an optical object is understood to be a star or galaxy. Adding these variables is straightforward, but will result in additional parameters that need to be estimated. In the case described in this paper this may well provide additional advantages to add a star/galaxy flag as it is expected that far infra-red emitters are more likely to be galaxies.

Other considerations regarding the addition of other variables include whether they need to be modelled jointly with others already introduced. It might be that the variables are not independent from one another, and that dependence must be modelled for result to be accurate.

6 Conclusion

This report outlines an approach for learning the parameters that can be used for linking data from different databases. It has specifically been derived for the astronomical problem of linking far infra-red observations to optical counterparts. Current work includes doing tests of this approach on artificial data, to examine the success and failure modes of the approach, as well as tests on real astronomical data. Additional extensions such as those suggested are being investigated.

Acknowledgments

The work of AJS was partially supported by a fellowship from Microsoft Research Ltd.

References

[1] T.R. Belin and D.B. Rubin. A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, 90(430):694–707, 1995.

[2] L.G. Brown. A survey of image registration techniques. *ACM Computing Surveys*, 24(4):325–376, 1992.

[3] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis.* Cambridge University Press, 1998.

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

[5] C. L. Giles, K. D. Bollacker, and S. Lawrence. Citeseer: An automatic citation indexing system. In *Digital Libraries 98 - The Third ACM Conference on Digital Libraries*, pages 89–98, 1998.

[6] C. W. Kelman, M. A. Kortt, N. G. Becker, Z. Li, J D Mathews, C S Guest, and C D J Holman. Deep vein thrombosis and air travel: Record linkage study. *British Medical Journal*, 372:1072, 2003.

[7] M. D. Larsen and D. B. Rubin. Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, 96(453):32–41, 2001.

[8] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[9] Z. Marx, I. Dagan, J. Buhmann, and E. Shamir. Coupled clustering: A method for detecting structural correspondence. *Journal of Machine Learning Research*, 3(1):747–780, 2002.

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

[11] A. Odn and M. Fahln. Oral anticoagulation and risk of death: a medical record linkage study. *British Medical Journal*, 325:1073–1075, 2002.

[12] D. Quass and P. Starkey. Record linkage for genealogical databases. In *KDD 2003 Workshop on Data Cleaning, Record Linkage and Object Consolidation*, 2003.

[13] R. E. Rutledge, R.J. Brunner, T. A. Prince, and C. Lonsdale. XID: Cross-association of ROSAT/bright source catalog x-ray sources with USNO A2 optical point sources. *Astrophys. J. Suppl.*, 131(1):335–354, 2000.

[14] Y. Thibaudeau. Identifying discriminatory models in record-linkage. In *Proceedings of the Survey Research Methods Section, American Statistical Association (1992)*, pages 835–840, 1992.

[15] B. Wellner, A. McCallum, F. Peng, and M. Hay. An integrated conditional model of information extraction and coreference with application to citation matching. In *Uncertainty in Artificial Intelligence 2004*, 2004.

[16] W. Winkler. Improved decision rules in the Fellegi-Sunter model of record linkage. Technical report, Statistical Research Division, U.S. Bureau of the Census, Washington, DC., 1993.

[17] W. Winkler and Y. Thibaudeau. An application of the Fellegi-Sunter model of record linkage to the 1990 U.S. decennial census. Technical report, Statistical Research Division, U.S. Bureau of the Census, Wachington, DC., September 1999.

[18] B. Zitova and J. Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21:977–1000, 2003.