

THE STATE OF RECORD LINKAGE AND CURRENT RESEARCH PROBLEMS

William E. Winkler, U. S. Bureau of the Census¹

ABSTRACT

This paper provides an overview of methods and systems developed for record linkage. Modern record linkage begins with the pioneering work of Newcombe and is especially based on the formal mathematical model of Fellegi and Sunter. In their seminal work, Fellegi and Sunter introduced many powerful ideas for estimating record linkage parameters and other ideas that still influence record linkage today. Record linkage research is characterized by its synergism of statistics, computer science, and operations research. Many difficult algorithms have been developed and put in software systems. Record linkage practice is still very limited. Some limits are due to existing software. Other limits are due to the difficulty in automatically estimating matching parameters and error rates, with current research highlighted by the work of Larsen and Rubin.

KEY WORDS: Computer matching; Modeling; Iterative fitting; String comparison; Optimization.

RÉSUMÉ

Cet article donne une vue d'ensemble des méthodes et des systèmes qui ont été mis en place pour le couplage d'enregistrements. Newcombe qui développa une approche nouvelle, et Fellegi et Sunter avec leur modèle mathématique, nous ont laissé les bases nécessaires pour un traitement moderne de la discipline du couplage d'enregistrements. Dans leur travail fondamental, Fellegi et Sunter ont introduit de puissantes idées pour l'estimation des paramètres sous-jacents, ainsi que d'autres idées qui continuent d'influencer la pratique du couplage d'enregistrement. La recherche sur le couplage d'enregistrements se caractérise par une synergie de la statistique, de l'informatique, et de la recherche opérationnelle. Malgré l'intégration sous formes de logiciels de plusieurs algorithmes, la pratique du couplage d'enregistrements n'en reste pas moins limitée. Cette limitation est due en partie aux défauts des logiciels eux-mêmes, mais aussi aux difficultés à estimer de façon automatique les paramètres sous-jacents ainsi que les taux d'erreurs encourues. Ce problème qui fait présentement l'objet de recherches, a été souligné par Larsen et Rubin.

MOTS CLÉS : Couplage d'enregistrements; modélisation; comparaison de chaînes de caractères; optimisation.

1. INTRODUCTION

Record linkage is the methodology of bringing together corresponding records from two or more files or finding duplicates within files. The term record linkage originated in the public health area when files of individual patients were brought together using name, date-of-birth and other information. In recent years, advances have yielded computer systems that incorporate sophisticated ideas from computer science, statistics, and operations research. Some of the work originated in epidemiological and survey applications.

Very recent work is in the related areas of information retrieval and data mining.

The ideas of modern record linkage originated with geneticist Howard Newcombe (Newcombe et al. 1959, 1962) who introduced odds ratios of frequencies and the decision rules for delineating matches and nonmatches. Newcombe's ideas have been implemented in software that is used in many epidemiological applications and often rely on odds-ratios of frequencies that have been computed a priori using large national health files. Fellegi and Sunter (1969) provided the formal mathematical foundations

¹ William E. Winkler, Statistical Research Division, Room 3000-4, Bureau of the Census, Washington, DC, 20233-9100 USA, bwinkler@census.gov

of record linkage. Their theory demonstrated the optimality of the decision rules used by Newcombe and introduced a variety of ways of estimating crucial matching probabilities (parameters) directly from the files being matched.

The outline of this paper is as follows. The second section provides more details on intuition about and the theoretical model for record linkage. Ideas of Newcombe have had the most important application in the development of national health files of individuals. The more general ideas of Fellegi and Sunter have been instrumental in estimating crucial matching parameters and estimating error rates for wide classes of lists. Methods for overcoming messy-data problems are described systematically in relation to the formal model of Fellegi and Sunter. In the third section, some of the basic research problems are covered. Although some of the problems have been (partially) solved for high quality pairs of lists, the solution methods do not easily extend to most matching situations.

2. BACKGROUND ON RECORD LINKAGE

Howard Newcombe had crucial insights that led to computerized approaches for record linkage. The first was that the relative frequency of the occurrence of a value of a string such as a surname among matches and nonmatches could be used in computing a binit weight (score) associated with the matching of two records. The second was the scores over different fields such as surname, first name, age, etc. could be added to obtain an overall matching score. More specifically, he considered odds ratios

$$\log_2(p_L) - \log_2(p_F) \quad (1)$$

where p_L is the relative frequency among links and p_F is the relative frequency among nonlinks.

Since the true matching status is often not known, he suggested approximating the above odds ratio with the following ratio

$$\log_2(p_R) - \log_2(p_R)^2 \quad (2)$$

where p_R is the frequency of a particular string (first, initial, birthplace, etc). If one matches a large universe file with itself, then the second ratio is a good approximation of the first ratio. Newcombe's ideas have been extended in a variety of ways (e.g., Newcombe et al., 1988, 1992, Gill 1999)

Fellegi and Sunter (1969) introduced a formal mathematical foundation for record linkage. To begin, notation is needed. Two files **A** and **B** are matched. The idea is to classify pairs in a product space $\mathbf{A} \times \mathbf{B}$ from two files **A** and **B** into **M**, the set of true matches, and **U**, the set of true nonmatches. Fellegi and Sunter,

making rigorous concepts introduced by Newcombe (1959), considered ratios of probabilities of the form:

$$R = P(\gamma \in \Gamma | M) / P(\gamma \in \Gamma | U) \quad (3)$$

where γ is an arbitrary agreement pattern in a comparison space Γ . For instance, Γ might consist of eight patterns representing simple agreement or not on the largest name component, street name, and street number. Alternatively, each $\gamma \in \Gamma$ might additionally account for the relative frequency with which specific values of name components such as "Smith", "Zabrinsky", "AAA", and "Capitol" occur. The ratio R or any monotonely increasing function of it such as the natural log is referred to as a matching weight (or score).

The decision rule is given by:

If $R > UPPER$, then designate pair as a match.

If $LOWER \leq R \leq UPPER$, then designate pair as a possible match and hold for clerical review.

If $R < LOWER$, then designate pair as a nonmatch. (4)

The cutoff thresholds $UPPER$ and $LOWER$ are determined by a priori error bounds on false matches and false nonmatches. Rule (4) agrees with intuition. If $\gamma \in \Gamma$ consists primarily of agreements, then it is intuitive that $\gamma \in \Gamma$ would be more likely to occur among matches than nonmatches and ratio (1) would be large. On the other hand, if $\gamma \in \Gamma$ consists primarily of disagreements, then ratio (3) would be small.

Pairs with weights above the upper cut-off are referred to as *designated matches* (or links). Pairs below the lower cut-off are referred to as *designated nonmatches* (or nonlinks). The remaining pairs are referred to as *designated potential matches* (or potential links).

If one considers a situation where there are three matching fields and only simple agree/disagree weights are considered, then a conditional independence assumption can be made to simplify computation.

$$\begin{aligned} &P(\text{agree first, agree last, agree age} | M) \\ &= P(\text{agree first} | M)P(\text{agree last} | M)P(\text{agree age} | M) \quad (5a) \end{aligned}$$

Similarly,

$$\begin{aligned} &P(\text{agree first, agree last, agree age} | U) \\ &= P(\text{agree first} | U)P(\text{agree last} | U)P(\text{agree age} | U) \quad (5b) \end{aligned}$$

This conditional independence assumption must hold on all combinations of fields (variables) that are used in matching. The probabilities $P(\text{agree first} | M)$,

$P(\text{agree last} \mid M)$, $P(\text{agree age} \mid M)$, $P(\text{agree first} \mid U)$, $P(\text{agree last} \mid U)$, and $P(\text{agree age} \mid U)$ are called *marginal probabilities*. $P(\mid M)$ & $P(\mid U)$ are called the *m-* and *u-*probabilities, respectively. The natural logarithm of the ratio R of the probabilities is called the *matching weight or total agreement weight*. The logarithms of the ratios of probabilities associated with individual fields (marginal probabilities) are called the *individual agreement weights*. The *m-* and *u-*probabilities are also referred to as *matching parameters*.

Fellegi and Sunter showed that it is possible to compute the unknown *m-* and *u-* probabilities directly in the 3-variable, conditional independence case. More generally, in the conditional independence situation, the parameters can be computed via a straightforward application of the EM algorithm (Winkler 1988). If the conditional independence assumption does not hold, then the parameters can be computed by generalized EM methods (Winkler 1988, 1989a, 1993b, Armstrong and Mayda 1993, see also Meng and Rubin 1993), by scoring (Thibaudeau 1993), and by Gibbs sampling (Larsen 1996, Larsen and Rubin 1999). The methods of Larsen and Rubin (1999) are the most general. These methods can yield more accurate matching parameters and better decision rules. These parameter-estimation methods do not always yield sufficiently accurate probability estimates for estimating record linkage error rates. An error-rate estimation method that is somewhat supplemental to these is due to Belin and Rubin (1995). Although the method of Belin and Rubin requires calibration data, it is known to work well in a narrow range of situations (Winkler and Thibaudeau, 1991; Scheuren and Winkler, 1993). The situations are those in which there is substantial separation of the curves of log frequency versus matching weight for matches and nonmatches. Generally, such separation occurs with high-quality lists of individuals containing only moderate amounts of typographical error and reasonable amounts of homogeneity in the characteristics respectively used in classifying pairs as matches and nonmatches. With some administrative lists and most agricultural and business lists, such homogeneity does not occur. In some situations, difficulties with business lists can be dealt with via software loops that address list-specific nonhomogeneity. Some of the general methods for dealing with nonhomogeneity of identifying characteristics are described in Winkler (1993). EM methods and ideas for dealing with nonhomogeneity similar to Winkler (1988, 1989a, 1993) have recently been applied to the general problem of text classification in machine learning and data mining by Nigam et al. (1999). The methods of Winkler are

more general because they allow for dependencies of fields and convex constraints on probabilities (either class or marginal) that predispose estimates to subregions of the parameter based on prior knowledge from similar matching situations.

2.1 String Comparators

In many matching situations, it is not possible to compare two strings exactly (character-by-character) because of typographical error. Dealing with typographical error via approximate string comparison has been a major research project in computer science (see e.g., Hall and Dowling, 1980). In record linkage, one needs to have a function that represents approximate agreement, with agreement being represented by 1 and degrees of partial agreement being represented by numbers between 0 and 1. One also needs to adjust the likelihood ratios (3) according to the partial agreement values. Having such methods is crucial to matching. For instance, in a major census application for measuring undercount, more than 25% of matches would not have been found via exact character-by-character matching. Three geographic regions are considered in Table 1. The function Φ_n represents exact agreement when it takes value one and represents partial agreement when it takes values less than one. In the St Louis region, for instance, 25% of first names and 15% of last names did not agree character-by-character among pairs that are matches.

Table 1: Proportional Agreement by String Comparator Values Among Matches Key Fields by Geography

	StL	Col	Wash
	First		
$\Phi_n=1.0$	0.75	0.82	0.75
$\Phi_n \geq 0.6$	0.93	0.94	0.93
	Last		
$\Phi_n=1.0$	0.85	0.88	0.86
$\Phi_n \geq 0.6$	0.95	0.96	0.96

Jaro (1976, see also 1989) introduced a string comparator that accounts for insertions, deletions, and transpositions. The basic Jaro algorithm has three components: (1) compute the string lengths, (2) find the number of common characters in the two strings, and (3) find the number of transpositions. The definition of common is that the agreeing character

must be within half the length of the shorter string. The definition of transposition is that the character from one string is out of order with the corresponding common character from the other string. The string comparator value (rescaled for consistency with the practice in computer science) is:

$$\Phi_j(s1,s2) = 1/3(\#common/str_len1 + \#common/str_len2 + 0.5 \#transpositions/\#common),$$

where s1 and s2 are the strings with lengths str_len1 and str_len2, respectively.

Using truth data sets, Winkler (1990) introduced crude methods for modeling how the different values of the string comparator affect the likelihood in the Fellegi-Sunter decision rule. Winkler also showed how a variant of the Jaro string comparator Φ_n dramatically improves matching efficacy in comparison to situations when string comparators are not used. The Winkler string comparator Φ_n is used in the Generalized Record Linkage System software of Statistics Canada.

2.2 Heuristic Improvement by Forcing 1-1 Matching

Jaro (1989) introduced a linear sum assignment procedure (lsap) to force 1-1 matching because he observed that greedy algorithms often made erroneous assignments. A greedy algorithm is one in which a record is always associated with the corresponding available record having the highest agreement weight. Subsequent records are only compared with available remaining records that have not been assigned. In the following, the two households are assumed to be the same, individuals have substantial identifying information, and the ordering is as shown. A lsap algorithm causes the wife-wife, son-son, and daughter-daughter assignments correctly because it optimizes the set of assignments globally over the household. Other algorithms such as greedy algorithms can make erroneous assignments such as husband-wife, wife-daughter, and daughter-son.

		HOUSEH2		
HOUSEH1		Wife	Daughter	Son
	Husband	c ₁₁	c ₁₂	c ₁₃
	Wife	c ₂₁	c ₂₂	c ₂₃
	Daughter	c ₃₁	c ₃₂	c ₃₃
	Son	c ₄₁	c ₄₂	c ₄₃

c_{ij} is the (total agreement) weight from matching the i th person from the first file with the j th person in the second file. Winkler (1994) introduced a modified assignment algorithm that uses 1/500 as much storage as the original algorithm and is of equivalent speed.

The modified assignment algorithm does not induce a very small proportion of matching error (0.1-0.2%) that is caused by the original assignment algorithm.

2.3 Why the methods do not always work well.

The record linkage methods described above can perform well when there is little typographical variation and other forms of nonhomogeneity in the identifying characteristics of lists. The methods may not work well due to failures of the assumptions used in the models, lack of sufficient variables for matching, sampling or lack of overlap in lists, and extreme variations in the messiness of data. The idiosyncrasies of messy data are most easily described. Each of the following types of errors provides examples of situations where pairs of records will not have homogeneous identifying characteristics.

1. Records that do not address standardize.
2. Records that do not name standardize.
3. Records that have more information or missing matching variables.
4. Records that do not have easily comparable fields.

Name	Ralph Smith	R J Smith
Address	123 Main St	PO Box 9128
Age	54	50

If the PO Box address in the right-most column were replaced by a street address that corresponds almost exactly to the street address given in the second column, then it might be possible to accurately match. If R J Smith is actually Roberta Joan Smith, then the match would be in error. Inconsistencies of name and address information are typically even greater with agriculture and business lists. During name and address standardization, commonly occurring words such as Mister, Road, Post Office Box, etc. are replaced by standardized spellings and the components of the names and addresses are placed in fixed locations. If standardization fails for a record, then automatic matching in software may be impossible. This is due to specific information needed for comparison and computing weights that is missing. If two lists of individuals are small samples, then we may not be able to match on certain commonly occurring names such as John Smith without substantial corroborating information. The difficulty of estimating the overlap of samples has most effectively been dealt with by Deming and Gleser (1959) in situations where there is not matching error. When there is a possibility of matching error, the estimation can be more difficult.

3. BASIC RESEARCH PROBLEMS

The basic research problems have been open since the work of Newcombe et al. (1959) and Fellegi and Sunter (1969). Partial progress in solving the problems has occurred. The major difficulties in all situations have been determining how identifying information can be used and what the relative value of a field is in matching in comparison with other fields.

3.1 When can frequency-based matching improve over simple agree/disagree matching?

The ideas of frequency-based (value-specific) matching were introduced by Newcombe et al. (1959). Fellegi and Sunter (1969) gave two methods for computing frequency-based weights in the context of their formal model that have been extended by Winkler (1988, 1989c). The basic idea is that agreements on specifically rarely occurring values of a field (variable) are better at distinguishing matches than general agreement (non-value-specific) or on commonly occurring fields. For instance,

$$\begin{aligned} &P(\text{agree last name} = \text{'Zabrinsky'}, \text{agree first name} \\ &\text{'Zbigniew'} | M) > \\ &P(\text{agree last name, agree first name} | M) > \\ &P(\text{agree last name} = \text{'Smith'}, \text{agree first name} \\ &\text{'James'} | M). \end{aligned} \quad (6)$$

Reasonably correct frequencies are computed and used in matching. The intuition is that frequency-based weights given by the first and third probabilities in (6) are better able to delineate matches and nonmatches than the simple agree/disagree probabilities given in the second probability in (6). Names by themselves are seldom effectively used in matching. Additional fields such as components of the address, age or full date-of-birth, maiden name, sex, and race are also needed to reduce error rates to acceptable levels. In some early experiments, frequency-based matching often did better than simple agree/disagree matching. With the development of more sophisticated models for estimating agree/disagree matching parameters with the EM algorithm, simple agree/disagree weights sometimes performed better. The reason is due to the fact that, in many files, a moderate number of false matches agree on relatively rarely occurring names. In those situations, pairs that might be in the clerical review region given in (4) might move upward to the designated match region. If there is a substantial number of fields available for matching, then the redundancy provided by the extra fields can reduce matching error. Where substantial redundancy of identifying characteristics are available, it is not clear that the total agreement weights should be moved

upward for pairs associated with less frequently values of a variable.

There are, nevertheless, a number of important situations when it is likely that frequency-based matching may be demonstrated to work at least as well as simple agree/disagree matching. The major situations all involve large national health files that have been significantly cleaned for typographical error and for which accurate probabilities can be computed a priori using true population counts. The research question is "Are there situations for which it can be shown that frequency-based matching improves over simple agree/disagree matching?" It seems that with many business lists, agriculture lists, and general administrative lists that frequency-based matching may not yield improvements because of the large amounts of typographical variation. These lists often have moderate to large proportions of records that fail standardization, have excessively high typographical error rates, and have only moderate overlap. If any one of these three situations occurs, then frequency-based matching may be seriously compromised.

3.2 What is the best method for estimating parameters under conditional independence when non-1-1 (or 1-1) matching is done?

Winkler (1990) showed that parameters estimated under the conditional independence EM sometimes worked better in matching decision rules than probabilities estimated using conventional conditional independence. The conventional methods estimate the marginal probabilities $P(\text{agree field} | M)$ and $P(\text{agree field} | U)$ directly using samples for which truth has been obtained via possibly time-consuming manual review. Winkler (1988) did not need to use the known truth of matches in his examples. The reason that the EM parameters worked better is that they effectively represent the conditional probabilities such as the following

$$\begin{aligned} &P(\text{agree field 1, agree field 2, agree field 3} | M) = \\ &P(\text{agree field 1} | M) P(\text{agree field 2} | \text{field 1, M}) \\ &P(\text{agree field 3} | \text{field 1, field 2, M}). \end{aligned} \quad (7)$$

The EM algorithm decides what ordering of the fields in (7) is optimal in estimating the likelihoods. These probabilities implicitly perform a minor automatic adjustment for the lack of conditional independence. The EM algorithm still makes a homogeneity assumption because it assumes that the same ordering can be applied to all pairs conditional on whether they are a match or nonmatch. Because the EM-parameters are designed to maximize the likelihood, they produce better decision rules than the probabilities estimating under the conventional methods. Winkler (1990) provided an exact comparison of decision rules using

parameters obtained by the two estimation techniques. Caution in the automatic use of the EM-probabilities is needed because the EM may not exactly divide the set of pairs into two classes that correspond exactly to matches and nonmatches. The difficulty of having EM-determined classes that correspond to true matching classes has been addressed by Winkler (1993) and by Nigam et al. (1999). The caution may not apply to conventionally estimated parameters because of the clerical review can assure that estimated parameters are consistent with model assumptions.

The EM probabilities are estimated using all pairs and often used in matching software that forces 1-1 matching. Although the mechanisms for forcing 1-1 matching are not explicitly accounted for, the probabilities are known to work well in those situations. The research question is "When can the EM-probabilities estimated under conditional independence be effectively used in 1-1 matching decision rules?" When can marginal probabilities that are conventionally estimated via samples be effectively used in 1-1 matching?

3.3 When does accounting for dependencies help in matching?

If conditional independence does not hold, then $P(\text{agree first name, agree last name} | M) \neq P(\text{agree first} | M) P(\text{agree last} | M)$.

Decision rules that apply probabilities estimated under the conditional independence assumption may be suboptimal. Smith and Newcombe (1975) gave a modified decision rule that adjusts for the lack of dependence that have been effectively extended and applied by others (Gill, 1999). The modified decision rules are heavily dependent on the assumption that the adjustments based on a sample for which truth is known can be used in a variety of matching situations. The assumption is likely to be reasonable in situations of large national health files for which truth is known on a large subset. Winkler (1989a), Thibaudeau (1993), Armstrong and Mayda (1993), and Larsen and Rubin (1999) have all given formal models for estimating the record linkage parameters (probabilities) under general dependence models. Winkler (1989a) also showed that the values of matching parameters vary significantly from one list to another. The variation occurs even when the lists have the same matching variables and the same amount of overlap but represent different geographic regions. All of the authors have shown that the development of appropriate dependence models takes considerable skill and suitable software. They have also shown that probabilities estimated under dependence are more

accurate. None of the authors has been able to show whether the new parameter-estimation method can be assured to yield appropriately good decision rules in actual record linkage software on a day-to-day basis. A basic research question is "For what types of files and matching situations can general dependence-based probabilities and decision rules improve matching?" There is still considerable empirical evidence that matching under the conditional independence assumption is effective in practice. Winkler (1993, 1994) demonstrated that matching under the conditional independence assumption worked nearly as well as matching under more general dependency models in certain situations. The situations included population files having multiple individuals per household in which 1-1 matching was forced. Winkler (1994) did suggest accounting for dependencies might yield better automatic estimates of error rates.

3.4 What are (suitable) ways of estimating error rates?

The method of Belin and Rubin (1995) is currently the only method for automatically estimating record linkage error rates. Belin and Rubin were able to achieve highly accurate estimates (Winkler and Thibaudeau 1991, Scheuren and Winkler 1993) in a narrow range of situations. The situations generally involved population files where there was good separation between the matching weights associated with nonmatches and matches. If there is not good separation, then methods that use more information from the matching process may ultimately yield suitable estimates in a larger range of situations as suggested by Winkler (1994) and Larsen and Rubin (1999). The estimation methods and the means of evaluating the fits of the latent class models are quite difficult because the usual Chi-square methods do not work (Rubin and Stern, 1993). The basic research question is "How does one automatically estimate error rates?"

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion. A longer version containing advanced research problems in microdata confidentiality, analytic linking, and machine learning is available at <http://www.census.gov/srd/www/byyear.html>. The translation of the abstract to French was facilitated by Dr. Yves Thibaudeau.

REFERENCES

- Belin, T. R., and Rubin, D. B. (1995), "A Method for Calibrating False- Match Rates in Record Linkage,"

- Journal of the American Statistical Association*, **90**, 694-707.
- Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland (1975), *Discrete Multivariate Analysis*, Cambridge, MA: MIT Press.
- Deming, W. E., and G. J. Gleser (1959), "On the Problem of Matching Lists by Samples," *Journal of the American Statistical Association*, **54**, 403-415.
- Dempster, A. P., N. M. Laird, and D. B. Rubin (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, **B**, **39**, 1-38.
- Fellegi, I. P., and A. B. Sunter (1969), "A Theory for Record Linkage," *Journal of the American Statistical Association*, **64**, 1183-1210.
- Gill, L. (1999), "OX-LINK: The Oxford Medical Record Linkage System," in *Record Linkage Techniques 1997*, Washington, DC: National Academy Press, 15-33.
- Jaro, M. A. (1989), "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, **89**, 414-420.
- Larsen, M. D. (1996), "Bayesian Approaches to Finite Mixture Models," Ph.D. Thesis, Harvard University.
- Larsen, M. D., and D. B. Rubin (1999), "Iterative Automated Record Linkage Using Mixture Models," Statistics Department Technical Report, Harvard University.
- Meng, X., and D. B. Rubin (1993), "Maximum Likelihood Via the ECM Algorithm: A General Framework," *Biometrika*, **80**, 267-278.
- Newcombe, H. B. (1988), *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*, Oxford: Oxford University Press (out of print).
- Newcombe, H. B., M. E. Fair, and P. Lalonde (1992), "The Use of Names for Linking Personal Records (with discussion), *Journal of the American Statistical Association*, **87**, 1193-1208.
- Newcombe, H. B., J. M. Kennedy, S. J. Axford, and A. P. James (1959), "Automatic Linkage of Vital Records," *Science*, **130**, 954-959.
- Nigam, K., A. K. McCallum, S. Thrun, and T. Mitchell (1999), "Text Classification from Labeled and Unlabelled Documents using EM," *Machine Learning*, to appear.
- Scheuren, F., and W. E. Winkler (1993), "Regression analysis of data files that are computer matched," *Survey Methodology*, **19**, 39-58.
- Scheuren, F., and W. E. Winkler (1997), "Regression analysis of data files that are computer matched, II," *Survey Methodology*, **23**, 157-165.
- Smith, M. E. and H. B. Newcombe (1975), "Methods for Computer Linkages of Hospital Admission-Separation Records into Cumulative Health Histories," *Meth. Inform. Medicine*, 18-25.
- Thibaudeau, Y. (1993), "The Discrimination Power of Dependency Structures in Record Linkage," *Survey Methodology*, **19**, 31-38.
- Winkler, W. E. (1988), "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 667-671.
- Winkler, W. E. (1989a), "Near Automatic Weight Computation in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Fifth Census Bureau Annual Research Conference*, 145-155.
- Winkler, W. E. (1989b), "Methods for Adjusting for Lack of Independence in an Application of the Fellegi-Sunter Model of Record Linkage," *Survey Methodology*, **15**, 101-117.
- Winkler, W. E. (1989c), "Frequency-based Matching in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 778-783.
- Winkler, W. E. (1990), "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Assn.*, 354-359.
- Winkler, W. E. (1993), "Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 274-279.
- Winkler, W. E. (1994), "Advanced Methods for Record Linkage," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 467-472 (longer version report 94/05 available at <http://www.census.gov/srd/www/byyear.html>).
- Winkler, W. E. (1995), "Matching and Record Linkage," in B. G. Cox *et al.* (ed.) *Business Survey Methods*, New York: J. Wiley, 355-384.
- Winkler, W. E. and Thibaudeau, Y. (1991), "An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Census," U.S. Bureau of the Census, Statistical Research Division Technical report RR91/09.