Matching uncertain identities against sparse knowledge

Steven Horn[1], Anthony Isenor[2] Moira MacNeil[1], and Adrienne Turnbull[1]

[1] Defence Research and Development Canada, Centre for Operational Research and Analysis
{steven.horn, moira.macneil, adrienne.turnbull}@forces.gc.ca
[2] Defence Research and Development Canada, Atlantic Research Centre
anthony.isenor@drdc-rddc.gc.ca

Abstract. This paper presents a method for fast matching of data attributes contained in a high-volume data stream against an incomplete database of known attribute values. The method is applied to vessel observational data and databases of known vessel characteristics, with emphasis on vessel identity attributes. Due to the large quantity of streaming observations, it is desirable to compute the best matching identity to a sufficient confidence level rather than include all possible identity information in the matching result. The question of which observed attributes to use in the calculation is addressed using information theory and the combination of the information conveyed by each attribute is addressed using evidence theory. An algorithm is developed which matches observations to known identities with a configurable level of desired confidence, represented as a χ^2 value for statistical significance.

Keywords: Entropy, Transferrable Belief Model, Generalized Bayes Theorem, Database, Intelligence, Information, Data Errors

1 Introduction

Data quality is a continual issue when dealing with an automated processing system. Introducing the requirement for real-time processing magnifies the problem by creating an environment where pausing and reflecting on the quality issue impacts the time criticality of the system.

Here we consider data quality as related to both errors (i.e., incorrect values) and inconsistencies (e.g., CA as compared to CAN for Canada; syntax issues; differing vocabularies issues). Inconsistencies are often related to vocabularies issues that sometime require semantic level matching [1]. Such data quality issues influence the system's ability to process the incoming data stream. Some simplistic views of how to deal with data inconsistency have been reported. For example, [2] examined inconsistencies in vessel information reported from a selection of open websites. This investigation illustrated the complexities associated with vocabulary matching when aggregating multiple information sources.

Data inconsistencies are also influence by the data volume. As volume increases, the ability of a user to identify and correct the data drops dramatically.

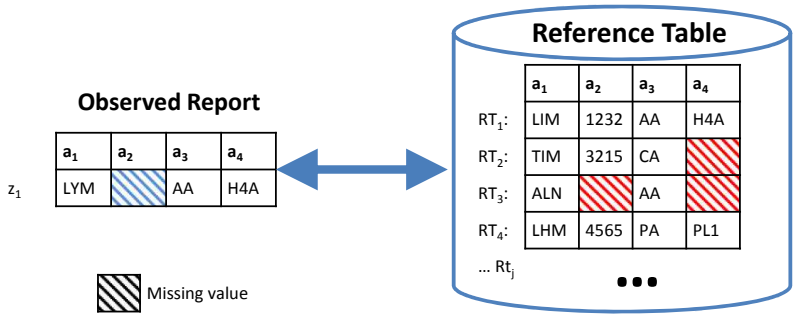

Fig. 1. Illustration of the sparse attribute association problem. a_i are the attribute types and RT_j are the reference table identities. Note that there are missing attribute values in both the report and the reference table rows.

Of course each data stream will be unique and in some sense these differences are related to the data attributes in the stream. A large variation in the content means a more difficult process to identify an error or inconsistency. Also, as fewer restrictions are placed on the data attribute, the automatic identification of quality issues becomes more problematic. An example of this case is a name text field. The observed and reported name of a person, object or thing, has such variation that few restrictions can be placed on the attribute content. A system's ability to learn and correct is influenced by the data attributes and the content permitted in those attributes.

2 Problem formulation

Commercially provided databases of ship identities are available, containing identifying information based on regulation and registration information. This knowledge base is represented as a database table, herein referred to as the Reference Table (RT). An incoming data stream of vessel observations is considered the target data and each incoming target record consists of numerous target attributes. In order to verify the identity of the target, one must associate these target attributes with the values in the RT entries. Fig. 1 illustrates this problem.

Let A be the set of possible attributes, and a_i be the elements of set A. For the results presented here, the set of attribute element labels in the data are: vessel name, Maritime Mobile Service Identifier (MMSI), International Radio Call-sign (IRCS), and International Maritime Organization (IMO) number. To formalize the description of observations, a target observation at time k, denoted by z_k consists of a set of attributes: $z_k = \{a_i\}\ a_i \in A$. The reference table contains the set of known targets, each with a set of attributes $RT_j = \{a_i\}\ a_i \in A$.

An additional challenge arises since the accuracy of the values of a_i in z_k is not ideal due to possible errors either from observation or transmission faults. However, it is assumed that the rate of error is known for a specific source or sensor. For example, an analysis of 561,771 distinct real-world observations resulted

in IMO values that were incorrect in 2.37% of the cases, and MMSI values that were incorrect 6.64% of the cases. These values provide a priori indications on the accuracy of information contained in each attribute. The problem described here is similar to the information retrieval problem with missing data [3] with the additional complexities of an uncertain query and query speed requirement.

3 Cost of attribute search

The entire RT must be searched for each attribute, since one must assume that there are errors in the attribute values and therefore the first result may not be the only, nor correct result. Furthermore, this means that one cannot necessarily use a previous attribute value match to narrow the RT search space. In order to reduce the number of attribute comparisons required to declare a match, the proposed approach is to achieve a significant level of confidence as early as possible in matching against the RT, even if not all observed attributes are used in the association.

Each attribute a_i in z_k carries with it some measure of information when considered against the RT. In this case, the quantification of information gain as the change in information entropy by the attribute is used as the attribute selection criteria. A similar approach has been used by others [4] to quantify database vulnerabilities to deriving hidden fields from a subset of known fields. In this case, the hidden field is the identifier RT_j. Information gain is also used for decision trees in machine learning [5]. The information gain (IG) of an attribute can be calculated as the difference between the entire RT (first sum in Eq. 1), and the conditional entropy from the attribute (second sum in Eq. 1).

$$IG(RT, a_i) = -\sum_v p_t \cdot log_e p_t - \sum_{v \in a_i} \frac{|\{t \in RT | t_i = v\}|}{|RT|} \cdot H(\{t \in RT | t_i = v\}) \quad (1)$$

Where p_t is the probability of a value occurring in the RT, and can be calculated in the frequentist approach from the RT as $p_t = \frac{1}{|RT|}$. The set entropy is defined as $H(X) = -\sum_x p_x \cdot log_e p_x$. Alternatively, the Laplace correction [6] can be used to estimate p_t by assuming p_t as a posterior Bayesian estimate, which better accounts for the possibility that the RT is not complete. The information gain $IG(RT, a_i)$ for each attribute is an indication of the value of that attribute for discriminating the value of RT_j. When an observation z_k is evaluated, those attributes with higher information gain should be searched first as they provide the strongest evidence (for or against) the association. The following section will discuss how each attribute comparison is combined using evidence theory.

4 Uncertainty model

The potential outcomes for each observed attribute are expressed in Table 1 as four possible cases, with a joint likelihood $p_1 \cdot p_2 \cdot p_3$ as a function of the target birth rate λ_b, observation rate λ_o, and error rate λ_e. The probability p_1 relates

to the likelihood that the observed target is or is not described in the RT, p_2 relates to the likelihood that the RT has or has not a non-null value for the attribute, and p_3 relates to the likelihood that the observed attribute is correct or incorrect. Note that the values for λ_e and λ_o are conditional on which source and sensor is providing the observation.

Table 1. Joint likelihood calculation for attribute match cases.

Case	Condition(s)	p_1	p_2	p_3
I:	$z \notin RT$	$\frac{\lambda_b}{\lambda_o}$		
II:	$z \in RT$, $a_i \notin RT_j$	$1 - \frac{\lambda_b}{\lambda_o}$	$1 - \frac{\lvert RT_j \rvert}{\lvert RT \rvert}$	
III:	$z \in RT$, $a_i \in RT_j$, a_i error	$1 - \frac{\lambda_b}{\lambda_o}$	$\frac{\lvert RT_j \rvert}{\lvert RT \rvert}$	λ_e
IV:	$z \in RT$, $a_i \in RT_j$, a_i correct	$1 - \frac{\lambda_b}{\lambda_o}$	$\frac{\lvert RT_j \rvert}{\lvert RT \rvert}$	$1 - \lambda_e$

The generalized Bayesian Theorem (GBT) in the transferrable belief model (TBM) is a standard method for the combination of evidence [7]. The pmf $f(RT_j|z)$ is used in the basic belief assignment (bba), as explained later in the text. The bba is a function on the set of hypotheses (Ω) which consists of elements of the power set for the possible combinations for values of RT_j on the set space 2^{Ω}. The bba function, generates the basic belief mass, which has the property that the sum of the values equals 1 [12], i.e. $\sum_{A \in \Omega} m(A) = 1$.

In our work, the open-wold assumption is adopted, which relaxes this sum such that the hypothesis $m(\emptyset)$ can also have a non-zero mass which represents the evidence (or lack of evidence) that the match is not contained in the RT. This assumption accepts the fact that it is possible the RT is incomplete and that no solution may be possible. The lower bound on the combined evidence is represented by the belief function defined as $bel(A) = \sum_{\emptyset \neq B \subseteq A} m(B), \forall A \subseteq \Omega$, and the upper bound by the plausibility function defined as $pl(A) = \sum_{B:A \cap B \neq \emptyset} m(B)$.

In GBT, the mass function is created assuming that $pl(RT_j|p) = P(RT_j|p)$ and therefore, the mass bba is assigned such that $pl(RT_j) = p(RT_j)$ for the singleton hypotheses [7]. In our application of the GBT, the probability mass function (pmf) $f(RT_j|z)$ used in the bba is calculated using the joint likelihoods ($p_1 \cdot p_2 \cdot p_3$) in Table 1. The null set receives the combined likelihoods from cases I and II, and possible matches receive the combined likelihoods from cases III and IV, which are then normalized. Other techniques exist for generating the bba, such as the use of Akaike information criterion [11], or expert training sets [5].

Each set of evidence generated by an attribute of the observed report z_k, the set of matches, will not necessarily intersect. This is a typical case where each source of evidence is over a non-exhaustive frame of discernment. There are many approaches to deal with this situation [8]. The approach adopted here is to use the disjunctive rule of combination to combine the evidence [9]. This supports the

combination of evidence from multiple attributes which may provide evidence for non-intersecting hypotheses [10].

To declare a confident match, a statistical hypothesis test is set up to determine if there is enough evidence to make a decision. If there is not enough evidence, more attributes must be included. Since the mass functions are represented on the set space 2^{Ω}, the pignistic transform [7] is used to collapse the evidence onto the singular hypothesis set for decision making.

$$BetP(A) = \sum_{X \subseteq \Omega} \frac{|A \cap X|}{|X|} \frac{m(X)}{1 - m(\emptyset)} \tag{2}$$

The pignistic probability for each potential matching RT_j at each stage is calculated. To indicate confidence in the result, a likelihood ratio test is used. Here, we take H_0 as RT_{H0} where $BetP(RT_{H0}) = sup(BetP(RT_j))$ and H_1 as the next highest $RT_{H1} \neq RT_{H0}$. The test statistic is defined as $\Lambda(t) = \mathcal{L}(H_0|z)/\mathcal{L}(H_1|z)$ using $BetP$ as the likelihood, and a threshold η is chosen to reject the null hypothesis according to the desired confidence in the match. The higher the desired confidence, the more information required to achieve the level of confidence and the more computational burden to select and combine the evidence.

The χ^2 distribution percent point function (quantile) is used with significance level α to reject the null hypothesis. H_0 is rejected if $\Lambda(t) < \eta = Q(\alpha, \chi^2(1))$. If H_0 is rejected, then another attribute must be included in the estimate. If there are no more attributes to check, then the observation is declared as not matched.

5 Results and conclusion

The algorithm was implemented in Python 2.7 and used the open source library for Dempster-Shafer theory calculations [13]. Real unclassified production data from the Royal Canadian Navy (RCN), and a commercial RT consisting of almost 600,000 records was used. Fig. 2 shows two examples of observations being associated to the RT. Note that vessel identities have been obfuscated. The first example achieves the desired confidence after considering the IMO, IRCS, and Name of the vessel. The second example is missing an IMO value in z_k but is able to achieve the desired confidence using the reported MMSI and IRCS.

The implementation was evaluated against real data and was able to achieve a processing rate of up to tens of thousands of records per second. Future work involves an in-depth analysis of the Receiver Operating Characteristics and inclusion of fuzzy comparisons such as soundex, edit distance, and Metaphones.

References

1. Graybeal, J., Isenor, A.W., Reuda, C.: Semantic mediation of vocabularies for ocean observing systems. Computers & Geosciences 40, 120–131 (2012)
2. ST-Hilaire, M.-O., Isenor, A.W.: Determining the consistency of information between multiple subsystems used in maritime domain awareness. NATO Science for

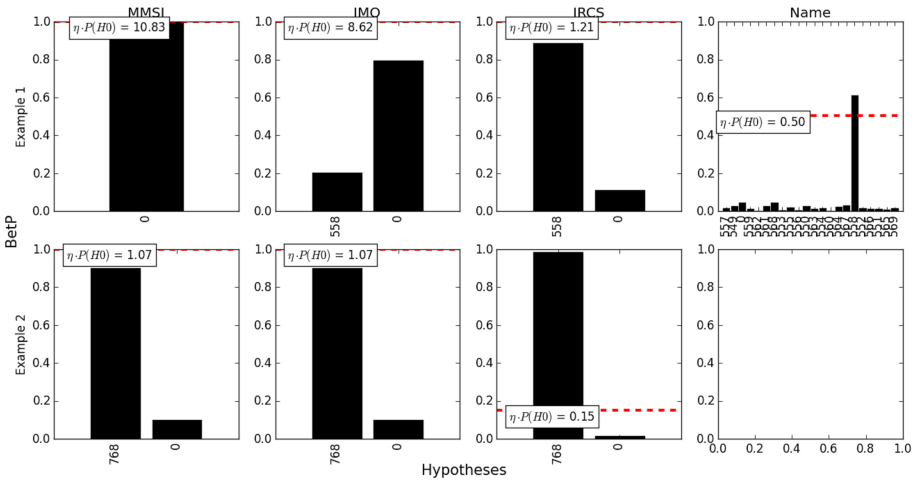

Fig. 2. Two examples of searching the RT. Attributes are searched in order from left to right. The horizontal dashed red line and text box indicate the cut-off probability to declare a match with α=0.001.

Peace and Security Series-E: Human and Societal Dynamics, NATO Advanced Science Institutes Series (2011)

3. Jousselme, A.-L., Maupin, P.: Comparison of uncertainty representations for missing data in information retrieval. Proceedings of the 16th International Conference on Information Fusion, 1902–1909 (2013)
4. Unger, E.A., Harn, L.: Entropy as a measure of database information. Proceedings of the Sixth Annual Computer Security Applications Conference, 80–77 (1990)
5. Elouedi, Z., Mellouli, K., Smets, P.: Belief decision trees: theoretical foundations. Int. J. Approximate Reasoning 28 no. 2-3, 91–124 (2001)
6. Wolpert, D., Wolf, D.: Estimating functions of probability distributions from a finite set of samples; part i: Bayes estimators and the Shannon entropy. Santa Fe Institute 1993-07-046 (1993)
7. Smets, P.: Data fusion in the transferable belief model. Proceedings of the 3rd International Conference on Information Fusion, 21–33 (2000)
8. Janez, F., Appriou, A.: Theory of evidence and non-exhaustive frames of discernment Plausibilities correction methods. Int. J. Approximate Reasoning 18, 1–19 (1998)
9. Smets, P.: Belief functions: the disjunctive rule of combination and the generalized Bayes theorem. Int. J. Approximate Reasoning 9, 1–35 (1993)
10. Denoeux, T.: Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence. J. Artificial Intelligence 172 no. 2-3, 234–264 (2008)
11. Lefevre, E., Vannoorenberghe, P., Colot, O.: Using information criteria in Dempster-Shafer's basic belief assignment. Proceedings of the 1999 IEEE International Fuzzy Systems Conference, 173–178 (1999)
12. Shafer, G.: A mathematical theory of evidence. Princeton University Press (1976)
13. Reineking, T.: A Python library for performing calculations in the Dempster-Shafer theory of evidence. https://github.com/reineking/pyds