Entity Identification in Documents Expressing Shared Relationships

JOHN R. TALBURT, NINGNING WU, ELIZABETH PIERCE, CHIA-CHU CHIANG*
CHRIS HEIEN, EBONY GULLEY**, JAMIA MOORE
Department of Information Science
*Department of Computer Science
** Department of Applied Science
University of Arkansas at Little Rock
2801 S. University Ave, Little Rock, AR 72204
USA
http://technologize.ualr.edu/eriq/

Abstract: This paper addresses the problem of entity identification in documents in which key identity attributes are missing. The most common approach is to take a single entity reference and determine the "best match" of its attributes to a set of candidate identities selected from an appropriate entity catalog. This paper describes a new technique of multiple-reference, shared-relationship identity resolution that can be employed when a document references several entities that share a specific relationship, a situation that often occurs in published documents. It also describes the results obtained from a recent test of the multiple-reference, shared-relationship identity resolution technique applied to obituary notices. The preliminary results show that the multiple-reference technique can provide higher quality identification results than single-reference matching in cases where a shared relationship is asserted.

Key-Words: Entity Identification, Entity Resolution, Identity Management, Feature Extraction, Text Mining, Obituary Notices

1 Introduction

The Entity-Relationship Model introduced by Peter Chen (1976) augmented the earlier Relational Model introduced in 1970 (Codd, 1970), which gave birth to the modern Relational Database Management Systems (RDBMS). When a specific instance of a RDBMS is created, the entity-relationships are described by a formal database schema that posits the existence of primary keys, which unambiguously identify all of the principal entities (people, places, and things) in the system. However in many real-world contexts, transactions must be processed that do not contain a unique key that identifies the record as belonging to a particular entity.

Consequently, almost every organization must deal to a greater or lesser extent with the problem of Entity Resolution, the process in which records that are determined to represent the same real-world entity are successively located and merged (Benjelloun, 2005). In the case that a record resolves to a previously known entity (a master record), then the process is also known as Entity Identification. A more formal algebraic treatment can be formulated by considering that an entity resolution process defines an equivalence relation on the underlying set of transaction records (Talburt, 2007).

In the case of documents there are two issues. The first is locating and extracting entity references from unstructured or semi-structured text documents. There are numerous text mining techniques that can be used to facilitate this process of "feature extraction." The approach to feature extraction taken for this project was relatively straightforward ontology-based, pattern matching as described by Arlotta (2003), Embley (1999), Hammer (1997), Hashemi (2003), Laender (2002), Wessman (2005), and many others.

The second problem is that in most cases, the entity references that appear in a document are not sufficient to identify a specific entity because key identification attributes are not present. For example a reference to person with a common name in a large city would be insufficient to identify the particular individual because there could many different people of the same name in that city. These incomplete references are referred to as "identity fragments."

2 Problem Formulation

The traditional approach to entity identification is to take a single identity fragment, compare its attributes to the same attributes from a set of possible candidate identities, and select the "best match" as the identification (Hashemi, 2003). The candidate identities used for the comparison are usually taken from an entity catalog. Figure 1 illustrates the single-reference, attribute-matching technique. However, this technique generally suffers from one of two serious drawbacks that can be difficult to deal with - either the catalog is replete with many identity choices that are very similar, or the catalog is incomplete and the correct identity is not among the choices. For this reason, the best-match algorithm usually takes the form of a belief function that gives a level of confidence in the choice that is made.

2.1 Single Reference Technique

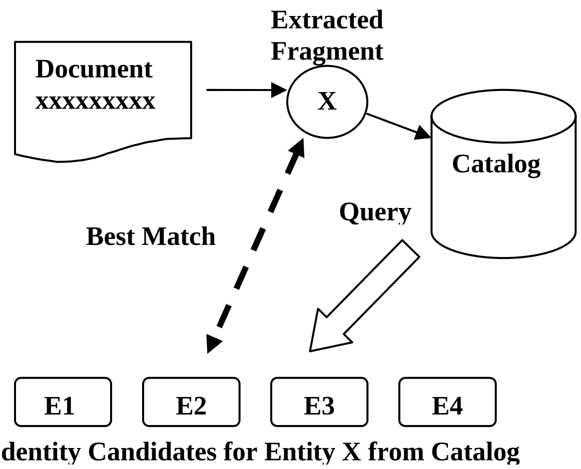

Figure 1: Single-Reference, Attribute Matching

An entity catalog is simply a repository for storing the identifying attributes for a collection of known entities. For example, the United State Postal Service (USPS) maintains a current and complete file of postal delivery points in the United States and it territories. Each delivery point is uniquely identified through a combination of attributes such as street number, pre-directional, street name, street suffix, post-directional, secondary type, secondary number, city, state, and zip code.

There are also many entity reference catalogs that maintain identification information for individuals and organizations. For example, zabasearch.com is a public website that maintains name, address, and telephone information for millions of individuals in the US, and is indexed by name and state. *Yahoo! People Search* and *PeopleFinders.com* are other examples of indexed catalogs available through the Internet.

Entity catalogs can be employed for entity identification in a number of ways, but only two will be discussed here: single-reference, attribute matching and multiple reference, shared relationship. Before discussing each technique in detail, first consider a published obituary that will be used as an example of an unstructured source document to which these techniques can be applied.

2.2 Obituary Example

A typical obituary is free-format text document. It usually contains the following items of information: decedent's name, age at death, and in many cases, names of the decedent's relatives such as parents, spouse, children, and siblings. However, it rarely provides any detailed information such as street addresses that could be used to establish the actual identity of any of the individuals referenced in the document. A typical example from Texas newspaper might read as follows:

> "William Doe, age 87, of Dallas
> died on July 4, 2006. He was
> survived by his wife, Mary Doe,
> and…"

In a metropolitan area like Dallas, Texas, there could be hundreds of people with the name William Doe. Thus additional information is needed in order to establish the actual identity of the decedent.

After the identity fragments are extracted, the catalog is searched using the partial set of attributes from the reference. Using the obituary example just described, this might amount to a query of the catalog for all entities with a name similar to William Doe with addresses in Dallas, Texas. The query produces a number of possible candidates for the identity of the extracted fragment. The last step is to evaluate the likelihood that each candidate represents the actual identity of the fragment.

Although this technique is straightforward in concept, it can be difficult to implement for the two reasons cited earlier, either too much data or too little data. Ambiguity can be introduced when there are several possible candidates, especially when some of the candidates have incomplete attributes. The second problem is the underlying assumption about the completeness of the catalog.

Name	Age	Address
Bill Doe	85	123 Pine St
William T. Doe	90	456 Oak St
Bill Doe, Jr	not given	123 Pine St
Bill Doe, Sr	not given	789 Elm St

Table 1: Candidates for William Doe

Table 1 illustrates the problem when a catalog query that produces many candidates. The first candidate has a similar name (Bill as a nickname for William) and similar age. The second candidate has the exact first and last name, but has a middle initial, and also has a similar age. The third has a similar name, but introduces the possibility of a name qualifier (Jr). This is further complicated by the fourth candidate with similar name and a different name qualifier (Sr). Both the third and fourth candidates lack any age information that could be used for disambiguation. Any believe function that attempted to weight the degree of match would likely yield at best score of no more than 50% because of the lack of discrimination among equally likely candidates.

If the query had only returned one candidate, the candidate Bill Doe, 85, in the first row, then the confidence of the match would be much higher, perhaps on the order of 95%. However, this must be tempered by some knowledge about the completeness of the catalog, information that may not be available. In this case, the single candidate is clearly the best and only match to reference, but it begs the question of whether there other viable candidates that simply have not been included in the catalog.

Previous work to identify decedents in public obituary notices using a large commercial catalog found that only about 20% of the decedents could be reliably identified (95% match confidence or higher) using this technique (Hashemi, 2003).

3. Problem Solution

A second technique for identification can be use when the source asserts a relationship among more than one identity fragment, and the reference catalog has attributes that can confirm the relationship. Again, the obituary notice can be used as an example.

In many instances, an obituary notice will assert the decedent's relationship to one or more other individuals, locations, or organizations. In this case, the technique is based on the fact that related person are likely to have shared the same residential address at some point in time. Thus, it may be possible to resolve identities by finding pairs of candidate identities generated from different fragments that share a common address.

The basic premise of the process is that if a published notice asserts that two identity fragments are of related individuals, and if the two identity candidate lists generated from these fragments share a common residential address, then there is a high probability that the identities that share the common address represent the correct identities of the original fragments. This concept is illustrated in Figure 3.

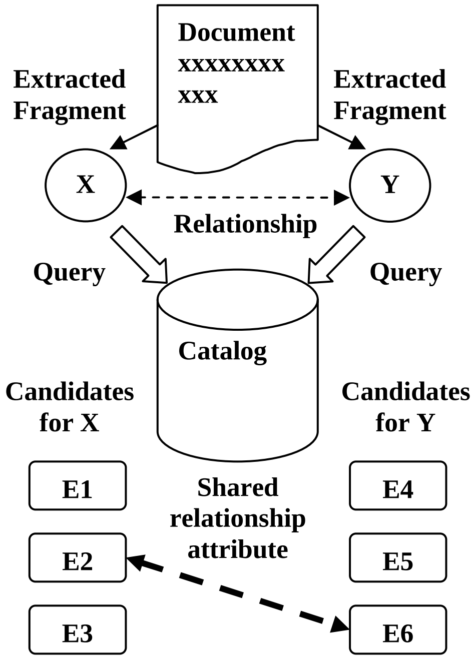

Figure 2: Multiple-Reference, Shared Relationship

Here the catalog is used to generate multiple candidate lists, one for each fragment extracted from the document. Figure 3 shows a simple example where there are two fragments X and Y. In this case, 3 candidate identities for X (E1, E2, and E3) are found in the catalog, and 3 candidate identities (E4, E5, and E6) are found for Y. The resolution comes from finding an intersection between the two lists where a common set of attributes confirm the relationship, in this case between identities E2 and E6. In the case of related individuals, this confirmation could be a common residential address. However, the same methodology could be applied to other entity types and relationships, such as, business entities that share officers or employees.

William Doe Candidates	Mary Doe Candidates
Bill Doe, 123 Oak	Mary Ellen Doe, 678 Willow
Bill S. Doe, 789 Hickory	Mary Doe, 654 Elm
William Q. Doe, 456 Pine	***Mary Doe, 789 Hickory***

Table 2: Candidates for William and Mary Doe

Table 2 shows how this resolution might look for the previous obituary example that asserts that William Doe of Dallas, TX, was survived by this spouse, Mary Doe.

In Table 2, the most likely conclusion is that the correct identities are the Bill S. Doe and Mary Doe who resided at the shared address “789 Hickory Street.” Although it is possible that an unrelated Bill Doe and Mary Doe occupied this same address at different times (a false positive), it is more likely that the identity has been resolved. A question for further research is how often false positives occur and under what circumstances. For example, one could reasonably expect more false positives to occur in an apartment complex than a single family residence, and for more commonly occurring names. An important feature of this technique is that variations in name and other attributes do not dramatically increase the uncertainty in identification as it does in the single-reference technique.

4. Conclusion

This section presents the results of a recent proof-of-concept test to apply the multiple-reference, shared relationship technique to online obituary notices. For these trials, an automated process was written in Java to extract fragments from the notices appearing on the public website *seacoastonline.com*. This site was selected because many of the notices on this site provide a complete or partial street address of the decedent. These cases of open identity allow for the validation of the process by comparing the process outcome to the known identity.

4.1 Automated Trial Results

Two catalogs, the website zabasearch.com (C1 in Table 3), and a commercial Customer Data Integration product (C2 in Table 3), were used in the resolution process. The results of two trials are shown, the first trial was conducted in November 2006, and the second trial in February 2007.

The difference in results between the November and February trials is primarily due to the implementation of several process improvements that addressed quality problems discovered after the November trial. Most of these improvements were in the area of feature extraction. For example in the second trial, improvements to the web crawler resulted in the recovery of information from almost 200 more obituary notices than in the first trial.

Refinements in the fragment extraction routine also resulted in more obituary notices being classified as “short obits.” A short obit is one where only a decedent fragment is found, i.e. no relatives are named. Because the multiple-reference technique cannot be applied to short obits, they were subtracted from the total number of obits to calculate the “net resolution rate” for each catalog (Rows F and N in Table 3).

	Item Description	Nov-06	Feb-07
A	Obits Extracted	1,781	1,971
B	Short Obits	269	755
Catalog 1 Result			
C	Candidates from C1	141,612	141,549
D	Resolutions by C1	111	271
E	Gross Res Rate (D/A)	6.2%	13.7%
F	Net Res Rate D/(A-B)	7.3%	22.3%
G	Res Obits w/Street	59	155
H	Res where All-Match	30	64
I	Res where Some-Match	7	64
J	Est. Acc. (H+.43*I)/G	55.9%	59.0%
Catalog 2 Results			
K	Candidates from C2	52,729	168,507
L	Resolutions by C2	143	386
M	Gross Res Rate (L/A)	8.0%	19.6%
N	Net Res Rate (L/(A-B)	9.5%	31.7%
O	Res Obits w/Street	67	227
P	Res where All-Match	30	56
R	Res where Some-Match	9	93
S	Est. Acc. (P+.43*R)/O	50.6%	42.3%

Table 3: Summary of Trial Results

In general, the statistics from the two trials show that the refinements to the feature extraction program resulted in significant quantitative improvements, but provided little or no qualitative improvements. For example, Rows F and N of Table 3 show that the extraction refinements effectively tripled the number of obits that could be resolved for both catalogs.

However, finding a resolution (in this case, finding a common address) does not necessarily mean that the correct identity has been found. Fortunately, a large portion of the obits listed on the seacoastonline website used for these trials explicitly stated a street address or street name for the decedent. For example, in Trial 1, 59 of the 111 obituary notices resolved using Catalog 1 had a stated street or street address (Rows G and D of Table 3).

By comparing the stated street address (when present) to the street address derived from the resolution process, it is possible to estimate the accuracy of the resolution process. However, the comparison is not simply match or no-match. Most resolutions from this process result in more than one address intersection between the two candidate lists. For this reason, there were actually three validation outcomes:

1. The stated address matched the resolved address (All-Match),
2. The stated address matched at least one of the resolved addresses (Some-Match), or
3. The stated address did not match any of resolved addresses (No-Match).

The problem is that most individuals reside at a number of different addresses over a relative short period of time. According to the USPS, about 14% of the US population move each year (USPS, 2007). Therefore any address reference catalog that maintains historical as well as current addresses for individuals (as do the two catalogs used for these trials) is likely to have more than one address for the same person.

Consequently the Some-Match Cases probably represent one of two situations; either the non-matching addresses are alternate addresses for the decedent, or the non-matching addresses are for another individual. In the former case, these resolutions should be counted as correct in estimating the accuracy of the resolution process. In the latter case, the actual identity of the decedent is unresolved.

As it turns out, the commercial catalog (Catalog 2) used its own business logic to maintain a linkage among various names and addresses combinations that were believed to belong to the same individual. An inspection of the links from a sample of 100 Some-Match Catalog 2 resolutions found that 43% of the resolutions all had the same link, i.e. all addresses were for the same individual. From this, the estimated accuracy of the resolution process was calculated as:

(#All-Match + 43% of #Some-Match)/Total Res

Using this formula, Rows J and S in Table 3 show that there was no significant increase in the accuracy of resolutions from Trial 1 to Trial 2 for either catalog. In the case of Catalog 1, the accuracy is slightly over 50%, and for Catalog 2, accuracy actually fell below 50% for Trial 2.

The 43% estimate is probably low for two reasons. The first is that the business logic used to maintain the links in Catalog 2 are probably not 100% comprehensive. It is likely that some of the addresses not linked to the same individual, such as Post Office boxes, did actually belong to that individual. The second reason is that it is also likely that some of the No-Match resolutions are actually correct. If the address stated in the obituary is a very recent address for the decedent, it is possible that the address has not had enough time to be cataloged.

In any case, the results are significantly better than the 20% accuracy of the single-reference, best match technique of the prior study (Hashemi, 2003). Because of the short obits, both the single- and multiple-reference techniques would be needed in any attempt to resolve obituaries notices on a large scale.

4.2 FUTURE WORK

Future work will focus on four areas.

1. Continue to improve the quality of the fragment extraction process. Many notices in the study were not able to be resolved because the extractor was unable to correctly identify the decedent and relative fragments in the document.
2. Extension of the current study to include other obituary sources.
3. Derive more reliable values for false positive and false negative identification rates.
4. Try to improve the results by combining catalog resolution with other soft computing techniques such as fuzzy sets, rough set analysis, neural networks, and evolutionary/genetic computing.
5. Test the resolution process using other document types, entity types, and catalogs.

REFERENCES

Arlotta, L., Crescenzi, V., Mecca, G., & Merialdo, P. (2003) Automatic annotation of data extracted from large websites. *Sixth International Workshop on the Web and Databases* (pp. 7-12). San Diego, CA.

Benjelloun, O., Garcia-Molina, H., Su, Q., & Widom, J. (March 3, 2005). Swoosh: A Generic Approach to Entity Resolution. *Technical Report*, Stanford InfoLab.

Chen, P. (1976). The entity relationship model - Towards a unified view of data. *ACM Transactions on Database Systems, 1*(1), 9-36.

Codd, E.F. (1970). A relational model of data for large shared data banks. *Communications of the ACM, 13*(6), 377-387.

Embley, D., Campbell, D., Jiang, Y., Liddle, S., Lonsdale, D., Ng, Y., & Smith, R. (1999). Conceptual-model-based data extraction from multiple-record web pages. *Data & Knowledge Engineering, 31*(3), 227-251.

Hammer, J., Garcia-Molina, H., Cho, J., Aranha, R., & Crespo, A. (1997). Extracting semistructured information from the web. *Workshop on Management of Semistructured Data.*

Hashemi, R., Ford, C., Bansal, A, Sieloff, S., & Talburt, J. (2003). Building semantic-rich patterns for extracting features from online announcements. In P. Isaias (Ed.), *International Association for Development of Information Society (IADIS) International Conference on WWW/Internet 2003*. Algarve, Portugal.

Laender, A., Ribeiro-Neto, B., da Silva, A., & Teixeira. J. (2002). A brief survey of web data extraction tools. *SIGMOD Record, 31*(2), 84-93.

Talburt, J., Wang, R., Hess, K., & Kuo, E. (2007). An algebraic approach to data quality metrics for entity resolution over large datasets. In L. Al-Hakim (Ed.), *Information Quality Management: Theory and Application* (pp. 1-22). Hershey, PA: Idea Group Publishing Group.

USPS Website. (2007). USPS Postal Facts, http://www.usps.com/

Wessman, A., Liddle, S., Embley, D. (2005). A generalized framework for an ontology-based data-extraction system. *Information Systems Technology and its Applications*, 4th International Conference ISTA'2005 (239-253). Palmerston North, New Zealand.