

Acxiom Data Engineering Lab (ADEL)

Working Papers Series 2002



Applied Research in Data Engineering

Publication in Collaboration
with UALR- Donaghey Cyber College

Semantic Data Matching

Russell Deaton, Ph. D. and Ning Zhu
Department of Computer Science and Engineering
The University of Arkansas
Fayetteville, AR 72701
rdeaton@uark.edu

Tom Schweiger
Acxiom Corporation
Fayetteville, AR 72701
Tom.Schweiger@acxiom.com

March 29, 2002

Abstract

Latent semantic analysis (LSA) was applied to the problem of matching business names. It was hypothesized that LSA could address problems, such as terms with multiple meanings and synonomous terms, that pure string matching could not. In a preliminary study, LSA was able to match business names with abbreviations, short record sizes, and typographical errors, as well as match synonyms and higher order structure in the data.

1 Introduction

Automated and real-time management of customer relationships requires robust and intelligent data matching across widespread and diverse data sources. Simple string matching algorithms, such as dynamic programming, can handle typographical errors in the data, but are less able to match records that require contextual and experiential knowledge. Latent Semantic Analysis (LSA) is a machine intelligence technique that can match data based upon higher order structure, and is able to handle difficult problems, such as words that have different meanings but the same spelling, are synonymous, or have multiple meanings. A preliminary study of LSA indicated that it can handle abbreviations, short record sizes, and typographical errors in business name data, as well as match synonyms and reveal higher order structure in the data, such as matching financial institutions though no words are shared. In this paper, the results of the preliminary study are reported.

2 Problem Statement: Data Matching for Customer Data Integration

In order to be successful, companies need to build relationships with customers, and to understand their wants and needs. Technology can support customer relationship matching by creating a real-time, single view of the customer from transaction information that is distributed across diverse and dispersed data sources. The problem of customer data integration is difficult because of errors, both machine and human, incomplete and misleading information, and the size and diversity of the data sources. Systems, such as Acxiom's AbiliTec technology, can enable and ease the task of customer data integration by processing customer records from widespread data sources, merging customer information in a data warehouse, and providing links to the customer information in an appropriate format.

A difficulty in this process is the accurate matching of customer information in different records. This information can include, among other things, business names, addresses, consumer names, and purchases. The problem is the imprecise nature of the customer information across different records. As an example, let's take business names. Business names can include typographical errors, phonetic spelling, homonyms (words that are spelled the same, but have different meanings), synonyms (different words, but different spellings), polysemy (words with multiple meanings), and different combinations of word breaks. Business names can be abbreviated or aliased. Frequently, human intelligence is able to deal with these difficulties by using context and experience, or in other words, the meaning or semantics of the information. The challenge in automated data integration is to reproduce the human capability in machine intelligence.

3 Semantic Data Matching

In this work, a technique called latent semantic analysis will be applied to data matching for customer data integration.

3.1 Background on Latent Semantic Analysis

In approaches such as dynamic programming, information is matched by literally comparing and scoring pairwise string symbols in the data records. Higher scoring pairs of strings are matches, while low scoring pairs are mismatched. While perhaps adequate for overcoming typographical errors in data records, string matching is inadequate for the problems of homonyms, synonyms, and polysemy.

Latent semantic analysis (LSA)[1, 2] attempts to classify records based upon higher order structure (meaning) in the data space that is present, but not obvious, in the patterns of word usage and groups. For a given query, the quality of its classifications are determined by the other words in a record and the size and completeness of the data space, corresponding to context and experience, respectively. It has been applied to a variety of data matching applications, such as intelligent tutoring systems, web-based search and information retrieval, and cross-language matching, and is capable of dealing with the complex problems associated with data matching for customer data integration.

The technique is based on linear algebra, and singular value decomposition, in particular. The data is represented in a term \times document matrix, A , where for example individual terms are the rows, the documents, or records, are the columns, and the entries

$$A = [a_{ij}], \tag{1}$$

are the frequency of occurrence of term i in document j . This is typically a sparse matrix, as every term does not appear in every document. Local or global weights can be applied to increase or decrease the importance of terms among documents. The matrix, A , is factored using singular value decomposition as

$$A = U\Sigma V^T, \tag{2}$$

where U (term) and V (document) contain the left and right singular vectors of A , and Σ is a diagonal matrix containing the singular values. These matrices represent, in a sense, the decomposition of the original information into a linearly independent set of vectors, or factor values. Typically, only the k largest factor values are saved, and thus, the original, and possibly noisy, term-document matrix is approximated by the reduced matrices. These reduced factors are a set of indices with which to represent each term and document as a vector in a k -dimensional space. This truncation of the original dimensionality enables higher order structure to be resolved, while dampening the effects of noise and variability in the data. This means that terms, which might never appear in the same document, can be near each other in the k -dimensional space. Nearness is typically calculated with cosines, Euclidean distance, or dot products. Once derived, the SVD matrices can be used to classify new queries composed of terms or documents. In addition, the entire SVD can be updated with new information by the process of folding-in new terms and documents without redoing the entire SVD.

3.2 Preliminary Analysis

Business name data usually consists of short documents with few terms. LSA is usually applied to much larger documents. In order to verify that LSA was appropriate for Acxiom data and using an evaluation copy of a commercial program from Telecordia, the LSA technique was tested on a sample set of data. The measure of nearness was the dot product, with a score nearer to 1 indicating greater similarity. As a sanity check, documents were used as queries in order to verify that a document would perfectly match itself, which was the case. Then, various key words searched were done to explore the capabilities of the technique. Though not a systematic study, overall, the results were promising. In most cases, terms in a query were matched with appropriate documents. In addition, synonyms for query terms were detected, and higher order relationships emerged. Running times were reasonable. It was discovered that when certain common terms, such as office, center, *etc...*, were removed from the analysis, results improved.

The results for an example query of “richardson” on a set of 132 documents are shown in Table 1. The results for the same query, but with the word “center” removed are shown in Table 2. As is evident, the results for the correct matches were improved, but another

Term	Score
richardson center	0.934
richardson center incorpora	0.933
richardson center adult services	0.935
lumbermart building center	0.910
soderquist center	0.911
soder quist center	0.911
interface computer center	0.908
meeks building center	0.910

Table 1: LSA results for query “richardson.”

Term	Score
richardson center	1.0
richardson center adult services	0.996
allclean	0.996
allclean services	0.995
ibm global services	0.994
vice pres student services	0.994

Table 2: LSA results for query “richardson,” with common term “center” removed.

common word, “services,” introduced relationships between records, thus revealing a higher order structure, *i. e.* all business that are services, in the database.

In another example, the term, “bank,” was used to query a database of over a 100,000 documents. Some of the results are summarized in Table 3. Lines 1 and 2 obviously have a

Number	Term	Score
1	arvest	0.957
2	arvest bank opps	0.997
3	mcilroy bk & tr	0.919
4	federal taxct dls rtl wn	0.976
5	national financial services	0.935

Table 3: LSA results for query “bank.”

semantic relationship, and identify the same entity. In addition, the abbreviation for bank in line 3 was correctly identified. Lines 4 and 5 show entities that are related to ‘banks’ in their function, revealing the synonym “financial” for bank.

4 Conclusion

In conclusion, the preliminary study shows that the LSA technique has promise for addressing some of the problems for data matching. Automated and real-time management of customer relationships requires robust and intelligent data matching across widespread and diverse data sources. Simple string and individual term matching algorithms, such as dynamic programming, can handle typographical errors in the data, but are less able to match efficiently records that require contextual and experiential knowledge. Latent Semantic Analysis (LSA) is a machine intelligence technique that can match data based upon higher order, contextual structure, and is able to handle difficult problems, such as words that have different meanings but the same spelling, that are synonymous, or that have multiple meanings. Therefore, it was hypothesized that LSA could solve similar problems in Acxiom's data matching applications. When applied to a sample of Acxiom data, a simple test of LSA indicated that it can handle abbreviations, short record sizes, and typographical errors in business name data, as well as match synonyms and reveal higher order structure in the data, such as matching financial institutions though no words are shared. This indicated that LSA might be useful in identifying groups of records that represent the same business entity, and in retrieving business entity records in response to incomplete or erroneous user queries.

Several variables that affect LSA performance could be evaluated to improve performance. When the document-term matrix is parsed from the raw data, pre-processing can improve LSA performance, specifically, deletion of common terms, keyword identification by frequency of occurrence, and word order. In the singular value decomposition, global and local weighting of commonly occurring terms ('company'), abbreviations ('Inc.'), symbols ('&'), and numbers are an issue since they frequently occur in business name and address data. Other variables that could receive emphasis include the dimensionality of the LSA space, record sizes, noise in the data, and the effect of different fields on the results.

References

- [1] M. W. Berry, S. T. Dumais, and G. W. O'Brien, "Using linear algebra for intelligent information retrieval," *Siam Review*, vol. 37, pp. 573–595, 1995.
- [2] S. Deerwester, S. T. Dumai, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by latent semantic analysis," *Journal of the Society for Information Science*, vol. 41, pp. 391–407, 1990.