SANDIA REPORT

SAND2012-8069
Unlimited Release
Printed September, 2012

An Extended Vector Space Model for Information Retrieval with Generalized Similarity Measures: Theory and Applications

Biliana S. Paskaleva, Pavel B. Bochev, Arlo L. Ames

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone:	(865) 576-8401
Facsimile:	(865) 576-5728
E-Mail:	reports@adonis.osti.gov
Online ordering:	http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone:	(800) 553-6847
Facsimile:	(703) 605-6900
E-Mail:	orders@ntis.fedworld.gov
Online ordering:	http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2012-8069
Unlimited Release
Printed September, 2012

An Extended Vector Space Model for Information Retrieval with Generalized Similarity Measures: Theory and Applications

Biliana S. Paskaleva
Software Engneering and Qual Envmnt, MS-1138
Sandia National Laboratories, Albuquerque, NM 87185
bspaska@sandia.gov

Pavel B. Bochev
Applied Mathematics and Applications, MS-1320
Sandia National Laboratories, Albuquerque, NM 87185
pbboche@sandia.gov

Arlo L. Ames
Analytics and Cryptography, MS-1027
Sandia National Laboratories, Albuquerque, NM 87185
alames@sandia.gov

Abstract

We present an Extended Vector Space Model (EVSM) for information retrieval, endowed with a new set of similarity functions. Our model considers records as multisets of tokens. A token weight function maps records into a real vectors. Using this vector representation we define a p-norm of a record and pairwise conjunction and disjunction operations on records. These operations prompt consistent extensions of published set-based similarity functions and yield new ℓ_p distance-based similarities. We demonstrate that some well-known similarities form a subset of the new functions resulting from particular choices of token weights and p-values. In so doing,

we establish the equivalence of the corresponding information retrieval models with a properly augmented vector space model. The performance of the extended similarity measures is compared by solving an entity matching (EM) problem for two types of benchmark datasets. Among other things, our results show that the new similarity functions perform particularly well on tasks involving matching of records by keywords.

The EVSM served as foundation for mathematically rigorous definition of EM problem. We developed a supervised EM framework that interprets the EM as the combinatorial optimization problem of finding the maximum weight matching in a weighted bipartite graph connecting records from two databases, also known as Linear Sum Assignment Problem (LSAP). Casting of EM problems into LSAP offers valuable practical and theoretical advantages. There are efficient algorithms that solve LSAP in polynomial time. Availability of such algorithms reduces the task of solving the EM problem to computing weights for the edges of the bipartite graph connecting the records from the databases. This allowed focusing efforts on the development of robust and flexible methodologies for the estimation of the similarity between records and led to the notion of an optimal similarity function (OSF) for MMIR problems. The OSF is sought as a linear combination of similarity functions for the common relation attributes. Solution of a suitably defined quadratic program using training data defines the weights in the linear combination. Computational studies using the Abt-Buy e-commerce set and publication databases comprising of research articles in cloud computing, antennas and information retrieval areas confirm the robustness of our approach.

Acknowledgment

Funding from Laboratory Directed Research and Development program is gratefully acknowledged. The authors would like to acknowledge Diana Frederick for her help collecting data and creating databases, and running data studies. The authors also would like to acknowledge Cynthia Phillips and Joseph Young for answering questions about combinatorial optimization problems.

Contents

Appendix

List of Figures

List of Tables

Nomenclature

EM: entity matching

EVSM: extended vector space model

LSAP: linear sum assignment problem

NWI: normalized weighted intersection

QP: quadratic program

VSM: vector space model

Chapter 1

Introduction

Given a generating set of terms, and the associated term weights, the standard Vector Space Model (VSM) [22, 26] for information retrieval encodes documents and queries as vectors of term weights. A similarity function measuring the closeness between documents is an integral part of the VSM. The normalized inner product between vectors defines the cosine similarity, which is a standard similarity choice in the VSM. Utilization of the inner product by the cosine similarity corresponds to viewing document vectors as elements of a Hilbert space. However, in the broader context of information analysis other non-Hilbertian structures, including set-based approaches [14], have demonstrated significant promise.

In this report we use set-theoretic and Banach space ideas to extend the VSM with several new classes of similarity functions. In particular, we develop extensions of the Jaccard similarity, the Normalized Weighted Intersection (NWI) similarity, and the Dice similarities, as well as a class of ℓ_p norm-based similarities. We show that published set-based similarity functions [14, 15] and the standard VSM cosine similarity form a subset of the new class of similarities, corresponding to specific choices of the term weighting and the p-values.

In so doing we effectively demonstrate the *equivalence* of these approaches with a properly augmented VSM. In other words, a vector representation of documents in conjunction with a suitably defined closeness metric provides an abstraction for a broad class of information retrieval approaches. For example, the extensions of the Jaccard, the NWI, and the Dice similarities developed in this work show that the set-based approach is equivalent to a vector space model in which the closeness between documents is measured by the former. Such an equivalence opens up interesting possibilities to both analyze existing approaches and seek new strategies for information retrieval, which we plan to examine in future work.

Figure 1.1 shows the two dominant paradigms in the IR: VSM and set-based approaches. The purple cloud represents the developed unifying architecture that bridges both approaches.

To further demonstrate the utility of the extended VSM we conduct an entity matching study involving several different benchmark databases. In addition to the Abt-Buy e-commerce set [1] the study utilizes publication databases comprising of research articles in cloud computing, antennas and information retrieval. Among other things, our results show that the new similarity functions perform particularly well on tasks involving matching of records by keywords.

To summarize, the main contributions of this work are as follows:

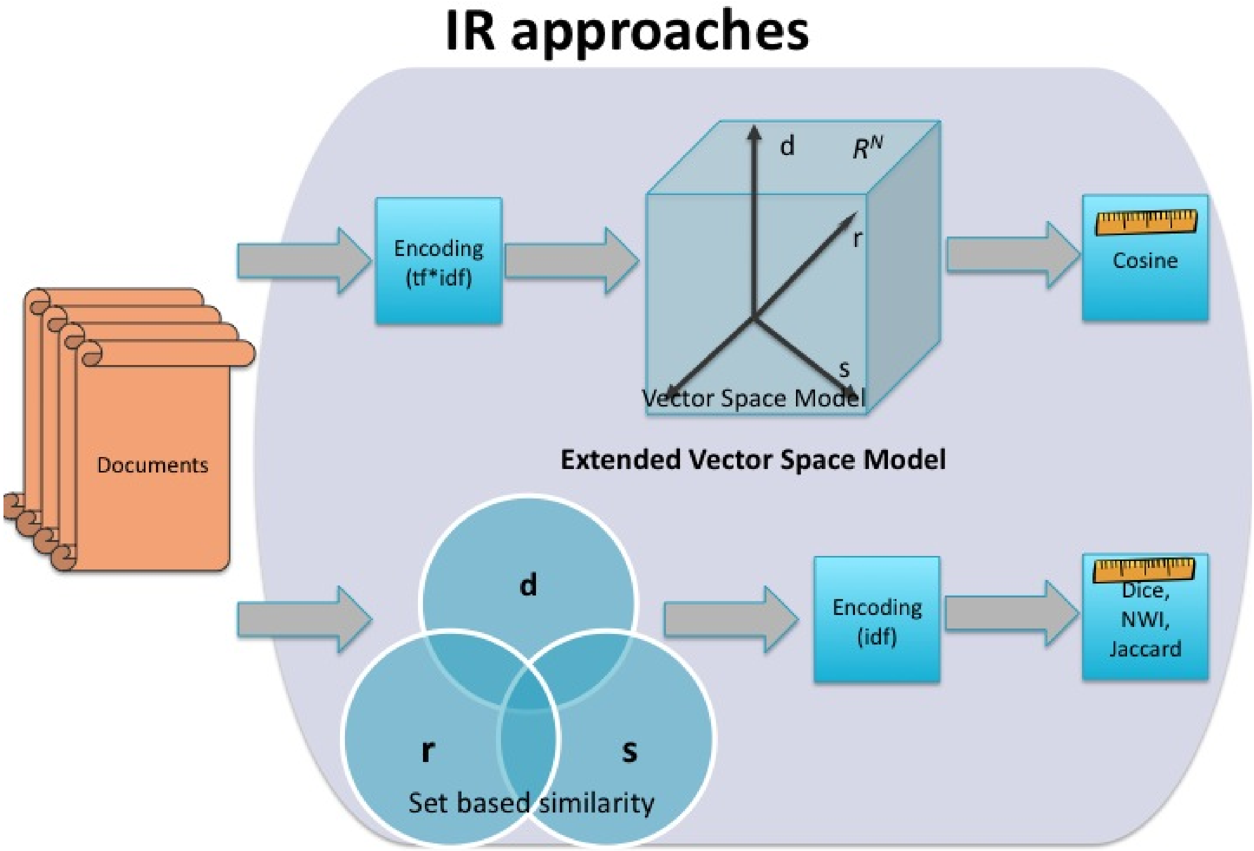

Figure 1.1: Extended Vector Space Model for information retrieval.

- Developed consistent extensions of similarity functions that bridge the standard vector space model with other approaches such as the set-based similarity.

- In so doing, we show that a properly augmented VSM provides an abstraction for a broad class of information retrieval approaches.

- The performance of a representative set of the extended similarity measures is compared and contrasted by solving an entity matching problem for two types of benchmark datasets.

The report is organized as follows. Chapter 2 presents a formal definition of the vector space model used in the work. Section 2.1 discusses token weights and vector representation of documents, Section 2.2 defines pairwise record operation, and Section 2.3 introduces the new similarity functions.

Chapter 3 presents application of the extended VSM to entity matching. The chapter briefly discusses the algorithmic solution of the corresponding mathematical formulation. Section 3.1 describes design of the study, Sections 3.2 and 3.3 present the results of the entity matching study for the scientific publication databases and the Abt-Buy dataset respectively. Our findings are summarized in Section 3.5.

Chapter 4 focuses on the formulation and development of entity matching through optimization-based approximation of a canonical similarity function. Section 4.1 presents a formal statement of the entity matching problem. Section 4.2 describes our approach for derivation of the approximation of an optimal similarity function. Section 4.3 presents results and discussions for case studies of the the optimization-based entity matching approach using the Abt-Buy e-commerce set. Our findings and conclusions are summarized in Section 4.4.

Chapter 2

Extended Vector Space Model (EVSM) for Information Retrieval

This chapter defines a formal VSM for Information Retrieval (IR), which provides the foundation for our approach. Throughout the work lower case bold face symbols denote vectors in Euclidean space $\mathbf{R}^N$, i.e., $\mathbf{r} = (r_1, \ldots, r_N)$. Upper case bold symbols are reserved for matrices in $\mathbf{R}^{N \times M}$. The point-wise q-th power of a vector is the vector $\mathbf{r}^q = (r_1^q, \ldots, r_N^q)$. The point-wise, or Hadamard [17], product of $\mathbf{r}, \mathbf{s} \in \mathbf{R}^N$ is the vector $\mathbf{r} \circ \mathbf{s} = (r_1 s_1, \ldots, r_N s_N) \in \mathbf{R}^N$. Recall that a multiset is a pair (A, μ), where A is an underlying set, and μ is a multiplicity function mapping A to the natural numbers. We will also use the notation $[a_1, a_2, \ldots, a_n]$ with the understanding that the sequence may have repeating elements. The symbol $|\cdot|$ denotes the cardinality of a multiset.

Let $T = \{t_1, t_2, \ldots, t_N\}$ be a dictionary corresponding to a given class of IR problems, i.e., a set of N distinct index terms or keywords, which form the relevant documents. We call the elements of T "tokens." The corpus $C(T)$ of the IR problem is the set of all token multisets $r = [t_1, t_2, \ldots, t_n]$, $n > 0$, and $C^2(T)$ denotes the collection of all multisets of elements of $C(T)$. The elements of $C(T)$ model documents and queries, i.e., a document or a query is a finite multiset of tokens. To simplify the terminology, we do not explicitly differentiate between documents and queries and use the term "record" in reference to both. The elements of $C^2(T)$ model databases (collection of records), i.e., a database is a finite multiset of records. To distinguish the multiplicity functions of different records we write $\mu_r(t)$ for the number of times the token t is encountered in record r. In particular, $\mu_r(t) = 0$ if $t \notin r$. The normalized multiplicity $\nu_r(t) = \mu_r(t)/\max\{1, \mu_r(t)\}$ defines an indicator function with the property that $\nu_r(t) = 1$ if $t \in r$ and $\nu_r(t) = 0$ otherwise.

2.1 Token weights and vector representation

A token weight $\omega(t, r, D)$ is a map $T \times C(T) \times C^2(T) \to R^+ \cup \{0\}$, which ranks the importance of token t in record $r = [t_1, \ldots, t_n]$, relative to a database $D = [r_1, \ldots, r_m]$. We require that

$$\omega(t, r, D) = \begin{cases} \alpha > 0 & \text{if } r \in D \text{ and } t \in r \\ 0 & \text{if } r \notin D \text{ or } d \in D \text{ and } t \notin r \end{cases} \tag{2.1.1}$$

We review two examples of token weights. The inverse document frequency [22]

$$\mathsf{idf}(t,D) = 1 + \log\left(\frac{|D|}{1+|D(t)|}\right), \tag{2.1.2}$$

where $D(t) = \{r \in D \,|\, t \in r\}$ is the multiset of all records in D containing a token $t \in T$, measures whether or not t is common or rare among the records in D. The following variant of (2.1.2) satisfies condition (2.1.1):

$$\omega_{\mathsf{idf}}(t,r,D) = \mathsf{idf}(t,D)\nu_r(t)\,. \tag{2.1.3}$$

In (2.1.3) $\nu_r(t)$ is the indicator function of r. The normalized term frequency [22]

$$\omega_{\mathsf{tf}}(t,r,D) := \mathsf{tf}(t,r) = \mu_r(t)/|r| \tag{2.1.4}$$

is a token weight that depends on t and r but not on D. The normalization by $|r|$ prevents a bias towards longer records (which may have a higher term count regardless of the actual importance of that term in the record). The tf*idf measure is the product of (2.1.2) and (2.1.4) [22, §6.2.2]:

$$\omega_{\mathsf{tf*idf}}(t,r,D) = \omega_{\mathsf{tf}}(t,r,D)\cdot\mathsf{idf}(t,D) \tag{2.1.5}$$

The value of the tf*idf weight is high when t has high frequency in record r, but in overall is not common in the database D [11]. We refer to [22, p.128] for additional variants of tf-idf measures.

In the vector space model every record is represented by an element of the Euclidean space $\mathbf{R}^N$, where $N = |T|$ is the size of the dictionary, i.e., the number of unique tokens in the particular IR context. Formally, the encoding process, which translates records into vectors, is a mapping $\mathbf{w} : C(T) \mapsto \mathbf{R}^N$. It is easy to see that every token weight that satisfies assumption (2.1.1) induces such a mapping viz.

$$C(T) \ni r \mapsto \mathbf{r} \in \mathbf{R}^N; \quad \mathbf{r} = \big(\omega(t_1,r,D),\ldots,\omega(t_N,r,D)\big)\,.$$

In what follows we denote the action of this mapping by $\mathbf{w}(r)$, i.e., $\mathbf{r} = \mathbf{w}(r)$.

2.2 Definitions of pairwise record operations

In this section we introduce and study functions mapping pairs of records into non-negative real numbers. To this end we need the ℓ_p norm

$$\|\,\mathbf{r}\,\|_p := \left(\sum_{i=1}^{N}\mathbf{r}_i^p\right)^{1/p}$$

of a vector $\mathbf{r} \in \mathbf{R}^N$. The p-norm of a record $r \in C(T)$, relative to the token weight ω, is the composition of $\|\cdot\|_p$ and the map ω:

$$\|\,r\,\|_{\omega,p} = \|\,\mathbf{w}(r)\,\|_p\;. \tag{2.2.1}$$

We define the conjunction norm, or simply the conjunction of two records $s, r \in C(T)$ as the Hadamard product of their encodings:

$$\parallel r \wedge s \parallel_{\omega,p} := \Big(\sum_{i=1}^{N} \mathbf{w}(r)^{p/2} \circ \mathbf{w}(s)^{p/2} \Big)^{1/p} . \tag{2.2.2}$$

Using (2.2.2) we define the disjunction norm or simply the disjunction of $s, r \in C(T)$ by

$$\parallel s \vee r \parallel_{\omega,p} := \parallel s \parallel_{\omega,p} + \parallel r \parallel_{\omega,p} - \parallel s \wedge r \parallel_{\omega,p} , \tag{2.2.3}$$

respectively. The conjunction (2.2.2) and the disjunction (2.2.3) are mappings $C(T) \times C(T) \mapsto \mathbf{R}^{+} \cup \{0\}$. The following proposition justifies the choice of names for these operations.

Proposition 1. *For every $r \in C(T)$ there holds*

$$\parallel r \wedge r \parallel_{\omega,p} = \parallel r \parallel_{\omega,p} \quad \textit{and} \quad \parallel r \vee r \parallel_{\omega,p} = \parallel r \parallel_{\omega,p} . \tag{2.2.4}$$

If $r, s \in C(T)$ have no common tokens, then

$$\parallel r \wedge s \parallel_{\omega,p} = 0 \ \ \textit{and} \ \ \parallel r \vee s \parallel_{\omega,p} = \parallel r \parallel_{\omega,p} + \parallel s \parallel_{\omega,p} . \tag{2.2.5}$$

Proof. From (2.2.2) it follows that

$$\parallel r \wedge r \parallel_{\omega,p} = \left(\sum_{i=1}^{N} \mathbf{w}(r)^{p/2} \circ \mathbf{w}(r)^{p/2} \right)^{1/p} = \parallel \mathbf{w}(r) \parallel_{p} = \parallel r \parallel_{\omega,p} .$$

The rest of (2.2.4) follows from this identity and (2.2.3). The proof of (2.2.5) is analogous. □

Our next results establishes connections between the pairwise record operations and some notions of set-based similarity. To avoid confusion we use the bar accent to differentiate between sets and multisets of tokens. Thus, $\bar{r}$ denotes a set of tokens, i.e., a collection of non-repeating elements of T. Clearly, $\bar{r}$ is a subset of T. Note that we can view $\bar{r}$ as a multiset for which $\mu_r(t) = 1$ for all $t \in \bar{r}$.

The set-based similarities [14, 18] represent records as *sets* of tokens and estimates the similarity of records by estimating the similarity of their token sets. Given two token sets $\bar{s}, \bar{r} \subset T$ we can estimate their similarity by assigning values to $\bar{s}$, $\bar{r}$, $\bar{s} \cup \bar{r}$ and $\bar{s} \cap \bar{r}$, and then combining these values into a final similarity score. The set values themselves can be derived from token weights assigned to each token in T, i.e., by using suitable token weights.

However, representation of records as sets of tokens, and the subsequent set operations, *disassociates* the tokens from their parent records. Consequently, the token weights in a set-based similarity cannot depend on a record argument. For instance, the tokens in $\bar{s}$ and $\bar{s} \cap \bar{r}$ are not aware of their multiplicity in the original record(s), whereas the tokens in $\bar{s} \cup \bar{r}$ are not aware of who their parent record was. This rules out application of token weights such as the tf measure (2.1.4) because the values of $\mu_{\bar{s}}$, $\mu_{\bar{s} \cup \bar{r}}$, coincide with the values of the corresponding indicator functions

$v_{\bar{s}}$, $v_{\bar{s}\cup\bar{r}}$, which do not reflect the true frequencies of tokens in their parent records. As a result, the set-based approach typically uses token weights such as idf.

Assuming that $\omega(t,D)$ does not depend on r, we can extend the ℓ_1 and ℓ_2 set-norms of $\bar{r} \subset T$, defined in [14], to a general ℓ_p set-norm

$$\| \bar{r} \|_{\omega,p} = \Big(\sum_{t\in\bar{r}} \omega(t,D)^p \Big)^{1/p} .$$

The following proposition establishes that the conjunction (2.2.2) and the disjunction (2.2.3) of records provide consistent extensions of set-based norms of intersections and unions of sets, respectively, to the vector space model.

Proposition 2. *Given multisets $r,s \in C(T)$, let $\bar{r},\bar{s} \subset T$ denote the sets of unique tokens in r and s, respectively. Assume that $\bar{\omega}$ does not depend on $r \in C(T)$ and define $\omega(t,r,D) := \bar{\omega}(t,D)\cdot v_r(t)$. Then,*

$$\| r \wedge s \|_{\omega,p} = \| \bar{r} \cap \bar{s} \|_{\bar{\omega},p}\,; \quad \| r \vee s \|_{\omega,p} = \| r \cup s \|_{\bar{\omega},p} \qquad \text{and} \quad \| r \|_{\omega,p} = \| \bar{r} \|_{\bar{\omega},p} \ . \tag{2.2.6}$$

Proof. The assertion easily follows from definition (2.2.1) and by using the fact that $\omega(t,r,D) = \bar{\omega}(t,D)$ whenever $t \in D$. □

Proposition 2 enables consistent extension of set-based similarity measures to the vector space model. The significance of this fact is that it allows us to conclude that the set-based similarity is equivalent to a particular instance of the vector space model endowed with the definition of the document ℓ_p norm (2.2.1), the conjunction (2.2.2), the disjunction (2.2.3) and the similarity extensions defined in the next section.

2.3 Generalized similarity functions

A similarity measure is a mapping $S : C(T) \times C(T) \mapsto [0,1]$. In this report we restrict attention to a class of measures defined by the composition of the mapping $\mathbf{w}$, induced by a token measure ω, with a vector similarity $\mathbf{s}$. Succinctly, we consider

$$S(r,s) = \mathbf{s}(\mathbf{w}(r),\mathbf{w}(s)) \quad \forall r,s \in C(T)\,, \tag{2.3.1}$$

where $\mathbf{w} : C(T) \mapsto \mathbf{R}^N$ and $\mathbf{s} : \mathbf{R}^N \times \mathbf{R}^N \mapsto [0,1]$.

In this section we introduce two new classes of similarity functions for the vector space model. The first class exploits the connection between set operations and the conjunction and disjunction functions in Proposition 2 to obtain consistent extensions of set-similarity measures, such as Jaccard or Dice, to the vector space model. We refer to, e.g., [18] or [14] for the set-based definitions of these measures.The second class uses the p-norm of a record to define distance-based similarity functions.

Extended Jaccard similarity. The Jaccard index, also known as the Jaccard similarity coefficient, is a statistic used for comparing the similarity and diversity of sample sets. Using (2.2.2), and (2.2.3) we extend the Jaccard's set-based definition to

$$J_p(r,s) := \frac{\| r \wedge s \|_{\omega,p}}{||r \vee s||_{\omega,p}}; \quad p \geq 1 . \tag{2.3.2}$$

Extended Normalized Weighted Intersection similarity. This similarity function is related to the Jaccard coefficient but uses different normalization of the set intersection. Using (2.2.2), and (2.2.3) we obtain the extension

$$N_p(r,s) = \frac{\| r \wedge s \|_{\omega,p}}{\max\{\| r \|_{\omega,p}; \| s \|_{\omega,p}\}}; \quad p \geq 1 . \tag{2.3.3}$$

Extended Dice similarity. Dice's similarity is named after Lee Raymond Dice, and is also related to the Jaccard coefficient and the normalized weighted similarity. The difference between the two is again in the normalization of the intersection term. The corresponding Dice's extension is

$$D_p(r,s) = \frac{2 \| r \wedge s \|_{\omega,p}}{\| r \|_{\omega,p} + \| s \|_{\omega,p}}; \quad p \geq 1 . \tag{2.3.4}$$

Normalized distance similarity. This similarity is defined using the normalized ℓ_p distance between r and s:

$$\Delta_p(r,s) = 1 - \frac{\| r - s \|_{\omega,p}}{2\max\{\| r \|_{\omega,p}, \| s \|_{\omega,p}\}}; \quad p \geq 1 . \tag{2.3.5}$$

The ℓ_p-distances corresponding to $p = 1$, $p = 2$ and $p = \infty$ are often called City Block, Euclidean and Chebyshev distance, respectively [24]. Thus, we may call $\Delta_1(r,s)$, $\Delta_2(r,s)$, and $\Delta_\infty(r,s)$, City Block, Euclidean and Chebyshev similarity functions, respectively.

Proposition 3. *Assume that $r,s \in C(T)$, $\bar{r},\bar{s} \subset T$, and $\bar{\omega}$ are as in Proposition 2 and let $\omega(t,r,D) := \bar{\omega}(t,D) \cdot \nu_r(t)$. Then,*

$$\begin{aligned} J_p(r,s) &= \frac{\| \bar{r} \cap \bar{s} \|_{\bar{\omega},p}}{||\bar{r} \cup \bar{s}||_{\bar{\omega},p}}, \\ N_p(r,s) &= \frac{\| \bar{r} \cap \bar{s} \|_{\bar{\omega},p}}{\max\{\| \bar{r} \|_{\bar{\omega},p}, \| \bar{s} \|_{\bar{\omega},p}\}}, \\ D_p(r,s) &= \frac{2 \| \bar{r} \cap \bar{s} \|_{\bar{\omega},p}}{\| \bar{r} \|_{\bar{\omega},p} + \| \bar{s} \|_{\bar{\omega},p}}. \end{aligned} \tag{2.3.6}$$

Proof. The proof follows directly from Proposition 2. □

This proposition confirms that (2.3.2)–(2.3.4) are consistent extensions of set-based similarity functions to both a general p and the vector space model context. In particular, for $p = 1$ the extended Jaccard, normalized weighted intersection and Dice similarities recover the functions in [14], while for $p = 2$ the extended Jaccard similarity (2.3.2) recovers the Jaccard coefficient used in [15].

Chapter 3

Application of EVSM to Entity Matching Problems

In this section we examine the performance of the extended VSM using two different types of benchmark databases. In both cases we apply the extended VSM to solve an EM, or a *record linkage* problem for the documents in these databases. We choose this setting because it lends itself to a mathematically precise decision rule based on linear programming [16]. This deterministic setting enables reproducible results, therefore reducing the ambiguities in the assessment of the extended VSM.

Section 3.1 explains the design of the study. There, the adopted formal EM definition and the corresponding solution strategy employed in the study are is provided. Sections 3.2–3.3 define the benchmark databases and present the results from the EM problem for each dataset. The results are discussed in Section 3.4.

3.1 Design of the study

In our study we adopt a formal EM definition and a solution strategy that follow the approach of [16]. It is beyond the scope of this chapter to provide a thorough review of the existing EM literature. Instead, we refer to the comprehensive reviews and studies in [5, 11, 25, 20, 19] and [13] among others. Chapter 4 provides more detailed review and information.

To explain the main ideas consider two databases D^1 and D^2 containing records that describe the same real world entities $\mathscr{E}$ using different attributes. The task is to link the records from D^1 and D^2 corresponding to the same real world entity from $\mathscr{E}$. The set T is the union of all unique terms in D^1 and D^2. For simplicity we assume that $|D_1| = |D_2| = M$, although this is not required for the application of the linear program approach below. Furthermore, we assume that ω is a token weight that fulfills (2.1.1) and that $S(\cdot,\cdot)$ is a similarity function defined as in (2.3.1).

The vector space encoding of the records in D^1 and D^2, induced by the mapping $\mathbf{w}$, is given by the rows of the corresponding term-to-document matrices $\mathbf{D}^1$ and $\mathbf{D}^2$ with elements

$$\mathbf{D}^1_{ij} = \omega(t_j, a_i, D^2) \quad \text{and} \quad \mathbf{D}^2_{ij} = \omega(t_j, b_i, D^1)\,,$$

respectively, where $t_j \in T$, $a_i \in D^1$, and $b_i \in D^2$. Using a similarity function $S(\cdot,\cdot)$ we define the $M \times M$ similarity matrix $\mathbf{S}$ with element

$$\mathbf{S}_{ij} = S(a_i, b_j)\,; \quad a_i \in D^1\,, \quad b_j \in D^2\,. \tag{3.1.1}$$

This matrix gives the pairwise similarity between the records in D^1 and and D^2.

The decision rule Following [16] we match the records using a linear program. Specifically, the records are linked by solving the following optimization problem

$$\begin{gathered}\max_{x_{ij}} \sum_{i=1}^{M} \sum_{j=1}^{M} \mathbf{S}_{ij} x_{ij} \quad \text{such that} \\ x_{ij} \in \{0,1\}, \quad \text{and} \quad \sum_{j=1}^{M} x_{ij} = \sum_{i=1}^{M} x_{ij} = 1\,; \quad i,j = 1,2,\ldots,M.\end{gathered} \tag{3.1.2}$$

The unit elements of the solution define the decision rule:

$$a_i \mapsto b_j \quad \forall x_{ij} = 1\,.$$

The program (3.1.2) is Linear Sum Assignment Problem (LSAP) [6, p.74]. The solution of (3.1.2) maximizes the "total similarity" of the assignments between the records in D^1 and D^2. The paper [16] is the first example of using (3.1.2) for record linkage. For more recent applications to entity matching we refer to [10] or [8]. In our study we solve (3.1.2) using the classical Hungarian algorithm [21]. We refer to [4] for another possible ways to solve the LSAP.

Evaluation of the performance To evaluate the performance of the decision rule we assume that D^1 and D^2 describe the same set of M distinct entities, i.e., there exists a perfect assignment rule $a_i \mapsto b_j$ between the records, which is a bijection $D^1 \mapsto D^2$. Suppose that $\{x_{ij}\}$ is the assignment rule defined by the solution of the LSAP. The unit elements $x_{ij} = 1$ of this rule correspond to two possible outcomes: true positive (TP), or false positive (FP). The precision of the assignment is then defined as

$$P = \frac{|TP|}{|TP| + |FP|}\,,$$

where $|\cdot|$ denotes the number of outcomes of each type. In our study we report the error of the assignment, in percent, defined as

$$E[\%] = (1 - P) \times 100\,. \tag{3.1.3}$$

Note that because we use a setting that admits only two possible outcomes, the precision P is identical with the accuracy

$$A = \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|}\,,$$

where TN and FN stand for true negative and false negative respectively.

Title	Keywords	Abstract	Doc ID
Antennas and propagation			
Simplified Cross-Polarized Multi-Antenna System for Radio Relay Transmission in Wireless Backhaul	“Cross-polarized array antenna; Adaptive antenna control; Radio relay transmission; Wireless backhaul; Directional intermittent packet transmit (lPT) forwarding”	Wireless backhaul systems have been considered as a promising candidate of beyond 3G wireless broadband system for mobile communications. The achievable transmission performance over radio relay channel depends on antenna directivity and radiation patterns of each antenna element. … In this report, we propose a simple method to extend the existing single antenna relay node based on IEEES02.11a to multi-antenna system, where a cross polarized multi-antenna is applied to the existing relay nodes as external equipment. … Simulation results ensure that the proposed multiantenna system with highly efficient packet forwarding protocol, called Intermittent Periodic Transmit (IPT), improves throughput performance of the radio relay transmission in wireless backhaul as compared with conventional omni-directional antenna system.	06206211
Cloud computing			
The Security of Cloud Computing System enabled by Trusted Computing Technology	cloud computing; trusted computing platform; trusted computing; trusted service	Cloud computing provides people the way to share distributed resources and services that belong to different organizations or sites. … We propose a model system in which cloud computing system is combined with trusted computing platform with trusted platform module. In this model, some important security services, including authentication, confidentiality and integrity, are provided in cloud computing system.	05555234
Data mining			
Toward Intelligent Assistance for a Data Mining Process: An Ontology-Based Approach for Cost-Sensitive Classification	cost-sensitive learning, data mining, data mining process, intelligent assistants, knowledge discovery, knowledge discovery process, machine learning, metalearning	A data mining (DM) process involves multiple stages. A simple, but typical, process might include preprocessing data, applying a data mining algorithm, and post processing the mining results. There are many possible choices for each stage, and only some combinations are valid. ....We use the prototype to show that an IDA can indeed provide useful enumerations and effective rankings in the context of simple classification processes. We discuss how an IDA could be an important tool for knowledge sharing among a team of data miners. Finally, we illustrate the claims with a demonstration of cost-sensitive classification using a more complicated process and data from the 1998 KDDCUP competition.	01401890

Table 3.1: Representative records from the antennas, cloud computing and data mining databases.

3.2 Application to scientific publication databases

In the first part of our study we apply the methodology presented in Section 3.1 to solve the EM problem for three different sets of scientific publications. We construct the databases D^1 and D^2 as follows. A set of M randomly selected published papers from a given scientific area defines the set of the “real world” entities $\mathscr{E}$. The database D^1 is defined by extracting the “keywords” attribute from the papers, whereas D^2 contains the “abstract” attribute. Thus, the objective of the EM problem is to match the keywords and the abstract belonging to the same article. Note, that this setting can be also interpreted as a search by keyword and so, our study provides some information about the performance of the VSM in an information retrieval context. In this context D^1 represents a set of queries and D^2 is a set of documents represented by their abstracts.

Using the IEEE Xplore Digital Library we obtained three sets[1] of 100 articles each, in the areas of cloud computing, antennas and propagation and information retrieval. From each article we generated a record comprising of 4 attributes: title, keywords, abstract and document ID. Table 3.1 shows representative records obtained from the first two sets of papers. Because the records identify the papers uniquely, they represent a proxy for the set $\mathscr{E}$ of real world entities. For each set we define $\mathscr{E}$, D^1 and D^2 by first taking a subset of 50 articles and then taking all 100 articles. Thus,

[1] The datasets for this study, except for the full text articles, are available upon request from the authors.

Dataset	Set 1 (M=50)		Set 2 (M=100)	
$S(\cdot,\cdot)$ $\omega \rightarrow$	tf*idf	idf	tf*idf	idf
cos	**6**	**8**	**3**	**4**
J_1	4	*10*	2	*7*
N_1	4	*10*	2	*7*
D_1	4	*10*	2	*7*
J_2	6	8	3	4
N_2	6	8	3	4
D_2	6	8	3	4
Δ_1	8	8	28	4
Δ_2	12	8	6	6
Δ_5	30	14	25	18

Table 3.2: Error in [%] in the solution of the entity matching problem for the extended set-based similarity functions comparing the tf*idf with idf token weights for the antenna database. Errors corresponding to the standard cosine similarity are in boldface. Errors corresponding to standard set-based similarity are in italics.

we obtain two databases per subject area, corresponding to $M = 50$ and $M = 100$, respectively. We refer to these databases as Set 1 and Set 2, respectively.

The study uses a total of 18 different similarities $S(\cdot,\cdot)$ corresponding to the Jaccard (2.3.2), NWI (2.3.3), and Dice (2.3.4) with $p = 1,2$ and the token weight (2.1.3) and (2.1.5), and the Normalized distance (2.3.5) with $p = 1,2,5$ with the same two token weights. For every $S(\cdot,\cdot)$ we compute the similarity matrix (3.1.1) for Set 1 and Set 2 and solve the corresponding LSAP. The errors are estimated according to (3.1.3). The cosine similarity with token measures (2.1.3) and (2.1.5) provides the corresponding reference error. Note that the Jaccard (2.3.2), NWI (2.3.3) and Dice (2.3.4) similarities with $p = 1$ and the token weight (2.1.3) are equivalent to the set based similarities [14], which enables comparison between the extended VSM and the set-based approach.

Tables 3.2, 3.3, and 3.4 summarize the error data for each dataset and similarity. Section 3.4 compares and contrasts these results with the results obtained using the Abt-Buy database, and draws conclusions.

3.3 Application to the Abt-Buy e-commerce database

The second part of the study applies the methodology in Section 3.1 to the Abt-Buy benchmark dataset [1]. This e-commerce dataset uses two different relations, “Abt” and “Buy”, respectively, to describe the “real world” entities $\mathscr{E}$. The entities are specific consumer products.

The “Abt” and “Buy” relations include attributes for a name, description, price, identification

Dataset	Set 1 (M=50)		Set 2 (M=100)	
$S(\cdot,\cdot)$ $\omega \rightarrow$	tf*idf	idf	tf*idf	idf
cos	**24**	**26**	**15**	**24**
J_1	26	*34*	17	*26*
N_1	22	*34*	19	*29*
D_1	26	*30*	17	*26*
J_2	22	30	17	26
N_2	22	24	16	25
D_2	26	30	17	24
Δ_1	46	26	46	26
Δ_2	22	28	19	25
Δ_5	22	32	31	31

Table 3.3: Error in [%] in the solution of the entity matching problem for the extended set-based similarity functions comparing the tf*idf with idf token weights for the cloud computing database. Errors corresponding to the standard cosine similarity are in boldface. Errors corresponding to standard set-based similarity are in italics.

number and a manufacturer; see Table 4.1. The Abt-Buy provides the exact matches between the record pairs, which makes it appropriate for entity matching studies. We base the entity matching on the "name" attribute, which gives a capsule summary of the consumer product (entity). The records from "Abt" define D^1 and the records from "Buy" define D^2. The set T is the union of all unique terms in the "name" field of "Abt" and "Buy".

In the study we use four nested subsets of the Abt-Buy database. These sets, termed Set 1, 2,3 and 4 comprise of 50, 100, 150 and 200 randomly selected records. To ensure that the sets are nested, we define them recursively by first selecting 200 random records for Set 4, then selecting randomly 150 of these records for Set 3 and so on. For each set we proceed to compute the similarity matrix using the same 18 similarities as in Section 3.2, then solve the LSAP problem and compute the error. As before, the cosine similarity with token measures (2.1.3) and (2.1.5) provides the corresponding reference error, while the Jaccard (2.3.2), NWI (2.3.3) and Dice (2.3.4) similarities with $p = 1$ and the token weight (2.1.3) provide for a comparison with the set-based approach. The results from the study are presented and Table 3.6 and discussed in the next section.

3.4 Discussion of results

In this section we present our observations by data sets and then proceed to draw the conclusions.

Scientific publications domain: antenna database The results in Table 3.2 show that the best-performing similarities for this database are Jaccard (2.3.2), NWI (2.3.3), and Dice (2.3.4) with

Dataset	Set 1 (M=50)		Set 2 (M=100)	
$S(\cdot,\cdot)$ $\omega \rightarrow$	tf*idf	idf	tf*idf	idf
cos	**4**	**4**	**10**	**5**
J_1	4	8	4	*4*
N_1	4	8	5	*4*
D_1	4	8	4	*7*
J_2	4	4	8	7
N_2	4	4	8	7
D_2	4	4	8	7
Δ_1	24	0	27	0
Δ_2	4	0	8	4
Δ_5	12	0	14	14

Table 3.4: Error in [%] in the solution of the entity matching problem for the extended set-based similarity functions comparing the tf*idf with idf token weights for the data mining databases. Errors corresponding to the standard cosine similarity are in boldface. Errors corresponding to standard set-based similarity are in italics.

Relation	Name	Description	Price	ID	Manuf.
BUY	Bose Acoustimass 5 Series III Speaker System - 21725	2.1-channel - Black	359.00	202812620	BOSE
ABT	Bose Acoustimass 5 Series III Speaker System - AM53BK	Bose Acoustimass 5 Series III Speaker System - AM53BK/ 2 Dual Cube Speakers With Two 2-1/2’ Wide-range Drivers In Each Speaker/ Powerful Bass Module With Two 5-1/2’ Woofers/ 200 Watts Max Power/ Black Finish	399.00	580	—

Table 3.5: Two records from the Abt-Buy e-commerce set corresponding to the same real world entity.

$p = 1$ and the token weight (2.1.5). Changing the token weight from (2.1.5) to (2.1.3) increases the error in the EM solution by a factor of 2.5 for the first dataset (M=50) and by a factor of 3.5 for the second dataset (M=100). Recall that the combination of J_1, N_1, and D_1 with the token measure (2.1.3) is equivalent to the set-based similarities [14]. Thus, for the antenna database, the extensions of the set-based similarities do outperform significantly their prototypes, and the the standard VSM cosine similarity.

Furthermore, the data for J_2, N_2, and D_2 shows that changing the norm index from $p = 1$ to $p = 2$ reduces the dependence on the token weight but also increases the minimal error in the EM solution. The increase is less pronounced for the larger data set. We also note that the errors of J_2, N_2, and D_2 match the errors of the standard cosine similarity. This could be attributed to the fact that all these measures employ Hilbertian metrics, whereas J_1, N_1, and D_1 rely on a Banach space norm.

Finally, we observe that for this particular database performance of the normalized distance similarities varies widely. Nonetheless, there are two noticeable trends: the errors tend to increase

Dataset	Set 1 (M=50)		Set 2 (M=100)		Set 3 (M=150)		Set 4 (M=200)	
$S(\cdot,\cdot)$ $\omega \rightarrow$	tf*idf	idf	tf*idf	idf	tf*idf	idf	tf*idf	idf
cos	**0**	**0**	**7**	**7**	**7.33**	**8.66**	**7.50**	**8.00**
J_1	0	*0*	9	*9*	8.67	*7.33*	8.5	*8.00*
N_1	0	*0*	7	*9*	7.33	*7.33*	8.50	*8.50*
D_1	0	*0*	9	*9*	8.67	*7.33*	8.00	*8.00*
J_2	0	0	7	7	7.33	7	7.50	8.00
N_2	0	0	6	7	7.33	7	7.50	8.50
D_2	0	0	7	7	7.33	7	7.50	7.00
Δ_1	0	0	10	9	8	10.67	6.50	10.00
Δ_2	0	0	12	10	12.67	11.30	13.00	12.50
Δ_5	12	0	39	24	38	22.66	38	22.50

Table 3.6: Error in [%] in the solution of the entity matching problem for the extended set-based similarity functions comparing the tf*idf with idf token weights for the Abt-Buy database. Errors corresponding to the standard cosine similarity are in boldface. Errors corresponding to standard set-based similarity are in italics.

with the norm index p, and the errors with the (2.1.3) token weight tend to be better than the errors with the (2.1.5) token weight. Note that the latter is opposite to what we observe with the rest of the similarities.

Scientific publications domain: cloud computing database The errors in Table 3.3 reveal that for this database all similarities, except for the normalized distance ones, perform in a similar fashion. As in the antenna database case, an overall improvement in the errors is observed when using the (2.1.5) token weight. While the improvement factors are not as large as in the antenna database, they remain significant. We also observe that for this data set the extended similarities perform comparably to the standard cosine similarity. They yield slightly better results for the first set (M=50) and slightly worse results for the second set (M=100).

The performance of the normalized distance similarities varies quite widely in this case and the trends in their behavior are less obvious. Nonetheless, we see a smaller variation in the error values when these similarities are used with the (2.1.3) token weight. We can view this behavior as a different manifestation of the second trend observed with the antenna database.

Scientific publications domain: data mining database The results in Table 3.4 show that, to a degree, the similarities reprise their performance from the antenna set. In the case $M = 50$ the error with the J_1, N_1, and D_1 doubles when the tf*idf token weight (2.1.5) is replaced by idf (2.1.3). For the same case J_2, N_2, D_2 show no dependence on the token weight.

Moving on to the second case (M=100) we see that both J_1, N_1, and D_1 and J_2, N_2, D_2 are essentially insensitive to the token weight. In contrast, the error with the cosine similarity increases by a factor of 2 when the token weight is switched from idf to tf*idf. Note that this pattern is opposite to what we have so far consistently observed with all but the normalized distance similarities.

The error data for the normalized distance similarities shows the emergence of the trend already observed in the antenna and the cloud computing databases, namely, the improved performance of these functions with the (2.1.3) token weight. For instance, using Δ_1, Δ_2, and Δ_5 with the (2.1.3) token weight allows the LSAP solution to recover the perfect assignment in the case of the first set (M=50). For M=100 the perfect assignment is recovered by Δ_1 with the same token weight. In contrast, using the (2.1.5) token weight, the errors of Δ_1 and Δ_5 significantly exceed those of the rest of the similarities.

E-commerce domain: Abt-Buy database Compared with the scientific publication databases, the results in Table 4.1 reveal a substantially different pattern of behavior for all but the normalized distance similarities. Most notably, for all four sets derived from the Abt-Buy database, the performance of the Jaccard (2.3.2), NWI (2.3.3) and Dice (2.3.4) similarities is essentially independent from the norm index *p and* the choice of token weight. Moreover, the errors of these similarities are close to or identical to the cosine errors.

Insofar as the normalized distance similarities are concerned, the trend for the error to increase with the norm index, observed in the antenna database, emerges here as well with Δ_1 yielding the smallest and Δ_5 - the largest error in all four cases. However, the second trend, namely, better performance of Δ_p with the (2.1.3) token weight, is not strongly present in this study. Below we we evaluate these observations and draw some summary conclusions.

Summary The first conclusion that can be drawn from the study analysis is that the error in the solution of the EM problem for all three scientific publication databases correlates with the choice of token weight, whereas the error in the EM solution for the Abt-Buy e-commerce database does not correlate significantly with this choice.

In particular, for the Jaccard (2.3.2), NWI (2.3.3), and Dice (2.3.4) similarities the trend is for the error to increase when (2.1.5) is replaced by (2.1.3). This trend can be explained by comparing the structure of the records in the scientific publication databases on the one hand and the Abt-Buy database on the other hand. In particular, the "keyword" attribute in the former tends to include a large number of records with repeating terms. For example, the record *cloud computing, trusted computing platform; trusted computing; trusted service*; see Table 3.1, contains multiple repetitions of the terms "computing," and "trusted," which appear in stable combinations such as "trusted computing" and "trusted service." The abstract record of the article will likely contain the same combinations of terms, and so, the term frequency becomes an important characteristic of the record, which is not accounted for by the idf token weight.

Remark 1. *We note that this reasoning does not apply to the Δ_p similarities, which generally exhibit the opposite trend of decreasing errors with the* idf *token weight for the three scientific publication databases. While at this stage of our study an explanation of this behavior remains an* *open question, it is clear that the larger errors and the more erratic behavior of these similarities requires extra caution in their practical application.*

Consider now the Abt-Buy database. The "name" attribute in the Abt-Buy contains records whose structure resembles that of natural language sentences. As a rule, such records contain much fewer repeating words and their lengths do not vary widely. For such records the term frequency tends to have less discriminating power because the tf values are close. As a result, for such records the token weight (2.1.5) differs from (2.1.3) by an almost constant factor. Consequently, both token weights will generate essentially equivalent similarity matrices, which in turn would lead to almost identical solutions of the EM problem. The results in Table 3.6 corroborate this conclusion.

3.5 Conclusions

In this chapter we formulated a systematic approach for extension of the VSM for information retrieval by new classes of similarity measures. The resulting EVSM provides an abstraction that includes other approaches such as set-based similarity.

We investigated the EVSM in the context of an entity matching problem using two different types of databases. Comparative error analysis of a representative set of similarity measures reveals that the extensions of the set-based similarities, obtained by using the tf*idf token weight, generally outperform their prototypes for records that exhibit the characteristics of "keywords," i.e., records with multiple repeating terms. Our analysis shows that the extended similarities and their prototypes are essentially equivalent for records whose structure resembles that of natural language sentences.

Our studies show that the performance of a given similarity measure varies with the structure of the records, the token weight and the underlying metric structure of the vector space. Availability of a systematic approach to the construction of similarities opens up a possibility to seek a similarity that is optimized, in a suitable sense, for a given dataset. Next chapter of the report focusses on approaches for defining optimal superposition similarity function.

Chapter 4

Entity Matching through an Optimization-based Approximation of a Canonical Similarity Function

In this chapter we focus on EM instances that map into a Linear Sum Assignment Problem (LSAP). Transformation of such EM problems into LSAP requires computation of pairwise weights between relations. Typically, conditional probabilities for match between relations attributes define these weights. Training-based EM frameworks use the training data to estimate the probabilities. However, because construction of training sets often requires manual record linking, their size may not be large enough for accurate probability estimates. To mitigate this problem, we use the training data to approximate an optimal similarity function for the given relation pair. We seek this function as a linear combination of similarity functions for the common relation attributes. Solution of a suitably defined quadratic program (QP) defines the weights in the linear combination. Computational studies using the ABT-Buy database confirm the robustness of our approach.

Databases represent real-world entities as records with flat or hierarchical data structure. The record generating process maps the characteristics of a real-world entity into attributes that can be represented and stored on digital computers. The mapping of entities into records can introduce errors, depends on the application context and can differ from database to database. As a result, the same real-world entity can generate non-identical records.

Deciding whether or not non-identical records refer to the same real-world entity is an essential task for information management in a wide spectrum of applications ranging from health care to law enforcement. It is telling that across disciplines the task itself is known under a multiplicity of names such as entity matching, entity resolution, entity reconciliation, record linkage, and record de-duplication, to name a few. In this work we adopt the term *Entity Matching* (EM) to refer both to the broader class of problems in this field as well as to the specific instance that we study.

Generally, there's a consensus that [23] is the first study of EM in a probabilistic setting, whereas [12] abstracted that setting to a formal probabilistic model for record linkage. The abundance of the current EM literature reflects the diversity of viewpoints and approaches coming from different fields of study. On the one end of the spectrum are the probabilistic approaches, which solve the EM problem by estimating the probability of a match between two records see, e.g., [23, 16, 12, 13, 10]. On the other end are approaches that abstract EM to a generic computation problem on a set of records in terms of abstract match, merge and link functions [2, 5]. For instance, given "black-box" match and merge operations along with a partial order on records, [2] defines EM as the task of finding the largest subset D' of the merge closure $\bar{D}$ of a document set D that also dominates $\bar{D}$.

However, a comprehensive review of all extant approaches is beyond the scope of this work. Instead, we limit ourselves to a brief summary of the work relevant to this report and refer to the excellent surveys [25, 19, 20, 13, 5, 11, 13], and the references therein for more information about past and current EM research.

In what follows we focus on EM for relational records with flat data structures, comprised of multiple attributes, i.e., the records are tuples of attribute values. Our main goal is the formulation of robust and flexible algorithms for EM which can adapt to varied application contexts. To this end, we adopt a mathematical abstraction of the EM problem which facilitates this objective by narrowing down the algorithmic design space. Specifically, we interpret EM as the combinatorial optimization problem of finding the maximum weight matching in a weighted bipartite graph connecting records from two databases[1]. In other words, we map EM to a Linear Sum Assignment Problem (LSAP) [6]. This choice reflects our focus on *matching* rather than *merging* records. The latter is an equally important and complex task, which is beyond the scope of this work.

Casting EM problems into LSAP offers valuable practical and theoretical advantages. There are efficient algorithms that solve LSAP in polynomial time [7], such as the Auction algorithm [3, 4] and the classical Hungarian algorithm [21]. Availability of such algorithms allows us to reduce the task of solving the EM problem to the task of computing weights for the edges of the bipartite graph connecting the records from the databases. This in turn allows us to focus efforts on the development of robust and flexible methodologies for the estimation of the similarity between records that work across multiple application domains. Last but not least, LSAP tends to perform better than matching schemes based on greedy-type algorithms because it optimizes the assignment globally over the complete set of records [25].

To the best of our knowledge, the first application of LSAP in the context of EM appears in [16]. This work considers a two-stage matching algorithm which combines LSAP with the matching techniques from [12]. Solution of an LSAP at the first stage provides an optimal assignment between the two sets of records. The second stage uses the optimal decision procedure from [12] to decide whether an assignment is a match. The paper defines the weights for the LSAP problem by summing up individual weights for agreement or disagreement of the record's attributes. Computation of the latter follows along the lines in [12].

The probabilistic decision model in [10] is another example of EM solution by LSAP. The weights in this model account for the cost (to a decision maker) of a false negative (type-I error) or a false positive (type-II error) result. This is different from [16] where the LSAP weights do not include such costs. Another difference is that the model in [10] is a single stage procedure which performs the match solely based on the LSAP solution. The distance-based approach in [8] uses the same LSAP setting but defines the weights as a linear combination of the expected distances

[1] A practical example of such EM is the problem of record linkage between different versions of the same customer or product lists, maintained by different departments of an organization.

between the attribute values. Expert ranking of the predictive power of the attributes is converted to coefficients for these expected distances. A further development of this approach appears in [9], which uses logistic regression to estimate the probabilities required to compute the LSAP weights.

Accuracy of probabilistic EM models necessarily depends on the accuracy of the conditional probability estimates involved in the model and by extension - on the size of the training sets available for the estimation of these probabilities. Because construction of training sets typically requires manual record linking, their size may not be large enough to achieve satisfactory results. This observation is the primary motivation for the distance-based approach in [8]. However, estimation of the expected distance in [8] relies on expert knowledge to asses the predictive power of various attributes and is prone to subjective bias. In either case the robustness of the EM solution can suffer.

The chapter is organized as follows. Section 4.1 presents a formal statement of the entity matching problem. Section 4.2 describes our approach for derivation of the approximation of an optimal similarity function. Section 4.3 presents results and discussions for case studies of the the optimization-based entity matching approach using the Abt-Buy e-commerce set. Our findings are summarized in Section 4.4.

4.1 Statement of the entity matching problem

A comprehensive entity matching process involves multiple steps, such as data preparation, data blocking, and matching and merging of records [11]. This works focuses solely on the task of matching records from two different relations, assuming that all necessary data preparation steps have already been performed. In so doing we obtain simple, yet sufficiently representative mathematical formalization of the EM problem which enables effective algorithm development.

We assume that there is a countable set E of real world entities e_l and a finite collection $A = \{A_1,\dots,A_K\}$ of attribute spaces. The attribute values $a_{i,l} \in A_i$ encode distinct characteristics of the entities $e_l \in E$. A relation is a pair $\{\rho,R\}$, where $R = A_{i_1} \times A_{i_2} \times \cdots \times A_{i_M}$ is a tensor product of attribute spaces and $\rho : E \mapsto R$ is a mapping that represents the record-generating process. The records are tuples of attribute values, i.e.,

$$R \ni r_l = \rho(e_l) = (a_{i_1,l},\dots,a_{i_M,l}) \quad \forall e_l \in E\,.$$

In what follows $\{\rho_k,R_k\}$, $k = 1,2$ are two relations that share a non-empty set of attributes $\{A_{j_1},\dots,A_{j_N}\}$, $N \le K$. Our approach uses only the common attributes between the relations. Thus, without loss of generality $\{A_{j_1},\dots,A_{j_N}\} = \{A_1,\dots,A_N\}$ and $R_1 = R_2 = A_1 \times \dots \times A_N$. The record of $e_l \in E$ in relation $\{\rho_k,R_k\}$ is the tuple

$$\rho_k(e_l) = (a^k_{1,l},\dots,a^k_{N,l}) = r_{k,l} \in R_k \quad \forall e_l \in E\,.$$

We assume that $r_{k,l}$ contains all the information about e_l in the relation $\{\rho_k,R_k\}$. In other words, there are no separate classes of records that describe relationships between entities [2].

A *similarity* function $S : R_1 \times R_2 \mapsto [0,1]$ compares the records. The value of $S(l,m) := S(r_{1,l}, r_{2,m})$ gives the level of "similarity" between $r_{1,l} \in R_1$ and $r_{2,m} \in R_2$.

We formulate the EM problem for a subset $D \subset E$ with a finite dimension $L = |D|$. The order in which ρ_k maps the entities from D into records in R_k is a permutation of the index set $I_L = \{1,2,\ldots,L\}$. Without loss of generality this permutation is the identity for R_1 and some non-trivial σ for R_2. Succinctly,

$$\begin{aligned} R_1(D) &= \{r_{1,l} = \rho_1(e_l) \,|\, e_l \in D\} \\ R_2(D) &= \{r_{2,l} = \rho_2(e_{\sigma(l)}) \,|\, e_{\sigma(l)} \in D\} \end{aligned} \tag{4.1.1}$$

The subset $R_k(D) \subset R_k$ contains the records corresponding to D in the relation $\{\rho_k, R_k\}$. An assignment function is bijection $\beta : R_1(D) \mapsto R_2(D)$. The map $\beta(r_{1,l}) = r_{2,m}$ defines a permutation $\beta(l)$ of L, i.e., $\beta(r_{1,l}) = r_{2,\beta(l)}$. Given a similarity function S, the total similarity between $R_1(D)$ and $R_2(D)$ relative to β is

$$S(\beta, D) = \sum_{l=1}^{L} S(r_{1,l}, \beta(r_{1,l})) = \sum_{l=1}^{L} S(l, \beta(l)) \,.$$

Definition 1. *(Entity Matching Problem) We are given a finite set of entities $D \subset E$ and (4.1.1) defines the corresponding records $R_k(D) \subset R_k$, $k = 1,2$. Given a similarity function S, the solution of the entity matching problem is an assignment function $\beta : R_1(D) \mapsto R_2(D)$, which maximizes the total similarity $S(\beta, D)$ between the records.*

Dependence of the EM solution on a similarity function and the requirement to maximize the total similarity between the records are the two key aspects of this definition. The former accounts for the fact that the quality of the entity matching process depends on the quality of the measures, which quantify the likeliness of the records. The latter reflects the interpretation of EM in this work as a combinatorial optimization problem. Indeed, let $\mathbf{S}$ be the $L \times L$ matrix with element

$$\mathbf{S}_{ij} = S(i,j);\ r_{1,i} \in R_1(D);\ r_{2,j} \in R_2(D) \,. \tag{4.1.2}$$

Then, Definition 1 is equivalent to the linear program

$$\begin{aligned} &\max_{x_{ij}} \sum_{i=1}^{L} \sum_{j=1}^{L} S_{ij} x_{ij} \ \text{ such that } x_{ij} \in \{0,1\} \\ &\sum_{j=1}^{L} x_{ij} = \sum_{i=1}^{L} x_{ij} = 1 ; \quad i,j = 1,2,\ldots,L \end{aligned} \tag{4.1.3}$$

The program (4.1.3) is Linear Sum Assignment Problem (LSAP) [6, p.74]. The unit elements $x_{i_l j_l} = 1$ of an optimal solution $\{x_{ij}\}$ give the assignments $r_{1,i_l} \to r_{2,j_l}$ that maximize the total similarity $S(\beta, D)$. The corresponding permutation $\beta(i_l) = j_l$ defines an assignment function $\beta(r_{1,i_l}) = r_{2,j_l}$, which solves the EM problem.

The record $r_{2,l} = \rho_2(e_{\sigma(l)})$, which implies $\beta(r_{1,l}) = r_{2,\beta(l)} = \rho_2(e_{\sigma\circ\beta(l)})$. From $r_{1,l} = \rho_1(e_l)$, it follows that $r_{1,l}$ and $r_{2,\beta(l)}$ correspond to the same entity iff $\beta = \sigma^{-1}$. We call σ^{-1} the *true assignment* function. Because the EM solution β depends on the similarity function S in general $\beta \neq \sigma^{-1}$. This prompts the following concept.

Definition 2. *A similarity function $S : R_1 \times R_2 \mapsto [0,1]$ is* optimal *for an entity set D, $|D| = L$ if for any permutation σ of the index set I_L, the LSAP (3.1.2) has a unique solution $\beta = \sigma^{-1}$.*

The precision, recall and accuracy of an EM solution β are

$$P(\beta) = \frac{t_p}{t_p + f_p}; \quad R(\beta) = \frac{t_p}{t_p + f_n}$$

and

$$A(\beta) = \frac{t_p + t_n}{t_p + f_p + t_n + f_n},$$

respectively. Where t_p, t_n, f_p and f_n are the numbers of true positive, true negative, false positive and false negative links between the records.

Proposition 4.1.1. *If S is an optimal similarity function for D, then $P(\beta) = R(\beta) = A(\beta) = 1$.*

This results justifies the term ”optimal” similarity function, and the next one provides a simple sufficient condition for optimality of *S*.

Proposition 4.1.2. *The condition*

$$S(l, \sigma^{-1}(l)) > S(l,k); \;\; k \neq \sigma^{-1}(l) \tag{4.1.4}$$

is sufficient for $S : R_1 \times R_2 \mapsto [0,1]$ to be an optimal similarity function for D.

Proof. A similarity function S is optimal if and only if the objective of (3.1.2) has a strict maximum at σ^{-1}:

$$S(\sigma^{-1}, D) > S(\alpha, D) \quad \forall \alpha \neq \sigma^{-1}. \tag{4.1.5}$$

We prove that (4.1.4) implies (4.1.5). Let $\alpha \neq \sigma^{-1}$ be arbitrary permutation of I_L. There exists at least one pair of indices $p, q \in I_L$ such that $p \neq q$, $\alpha(p) = \sigma^{-1}(q)$, $\alpha(q) = \sigma^{-1}(p)$, and

$$\begin{aligned} S(\alpha, D) &= \sum_{i=1, i\neq p,q}^{L} S(i, \alpha(i)) + S(p, \alpha(p)) + S(q, \alpha(q)) \\ &< \sum_{i=1}^{L} S(i, \sigma^{-1}(i)) = S(\sigma^{-1}, D). \end{aligned}$$

This proves the proposition. □

4.2 Approximation of an optimal similarity function

While in theory an optimal similarity function allows (3.1.2) to recover the true assignment $\beta = \sigma^{-1}$, it is unlikely that such a function will be readily available in practice. Accordingly, we propose to develop an approximation using the *canonical* similarity function

$$S(\rho_1(e_p), \rho_2(e_q)) = S(r_{1,p}, r_{2,\sigma^{-1}(q)}) = \delta_{p,q} \tag{4.2.1}$$

where $\delta_{p,q}$ is the Kronecker delta, as a template.

To this end we assume that there is a finite set of entities $\bar{E} \subset E$, $|\bar{E}| = \bar{L}$, with records $R_k(\bar{E})$, $k = 1,2$ for which we know the true assignment function $\bar{\sigma}^{-1}$. We refer to $R_k(\bar{E})$ as the *training set*. Furthermore, we assume that the common attribute spaces A_i have collections of similarity functions

$$s_{ij} : A_i \times A_i \mapsto [0,1];\ i = 1,...,N;\ j = 1,...,K_i.$$

which extend to maps $R_1 \times R_2 \mapsto [0,1]$ viz.

$$S_{ij}(r_{1,k}, r_{2,l}) := s_{ij}(a^1_{i,k}, a^2_{i,l})$$

We approximate (4.2.1) by a convex combination of these similarity functions:

$$\bar{S} = \sum_{i=1}^{N} \sum_{j=1}^{K_i} y_{ij} S_{ij};\ \sum_{i,j} y_{ij} = 1;\ 0 \le y_{ij} \le 1. \tag{4.2.2}$$

We use the training set $\{R_1(\bar{E}), R_2(\bar{E})\}$ to determine the coefficients y_{ij}. The true assignment for this set is $\bar{\sigma}^{-1}$. Consequently, to approximate (4.2.1) by (4.2.2) we require

$$\bar{S}(p,q) = \begin{cases} 1 & \text{for } q = \bar{\sigma}^{-1}(p) \\ 0 & \text{for } q \neq \bar{\sigma}^{-1}(p) \end{cases} \tag{4.2.3}$$

for $p,q = 1,\dots,\bar{L}$. Let $\bar{K} = \sum_{i=1}^{N} K_i$, the number of attribute similarity functions. Conditions (4.2.3) define a $\bar{L}^2 \times \bar{K}$ system of algebraic equations $\bar{\mathbf{S}}\mathbf{y} = \mathbf{d}$ where $\mathbf{y} \in \mathbf{R}^{\bar{K}}$ is a vector of unknown coefficients, $\mathbf{d} \in \mathbf{R}^{\bar{L}^2}$ is the vector

$$d_i = \begin{cases} 1 & \text{if } i \le \bar{L} \\ 0 & \text{if } i > \bar{L} \end{cases}$$

and $\bar{\mathbf{S}} \in \mathbf{R}^{\bar{L}^2 \times \bar{K}}$ is a matrix of coefficients. We ask that $\bar{L}^2 > \bar{K}$, which usually holds in most practical cases. To find the coefficients we solve the constrained optimization problem

$$\begin{cases} \min\limits_{\mathbf{d}} \|\bar{\mathbf{S}}\mathbf{y} - \mathbf{d}\|_p \quad \text{subject to} \\ \sum\limits_{i=1}^{\bar{K}} y_i = 1; \quad 0 \le y_i \le 1. \end{cases} \tag{4.2.4}$$

In summary, the optimization-based approach for entity matching comprises of three steps. At the first “training” step, we approximate the canonical similarity function using a given set of training records. At the second step we use the approximate similarity function to compute the coefficients S_{ij} of the LSAP formulation (3.1.2). Solution of this combinatorial optimization problem at the third step provides the solution of the entity matching problem.

4.3 Application to Abt-Buy e-commerce set

We test and study the optimization-based entity matching approach using the Abt-Buy e-commerce set [1]. The relations in the Abt-Buy involve five attributes $\{A_1, A_2, A_3, A_4, A_5\}$, where A_1 is a name,

Table 4.1: Two records from the Abt-Buy e-commerce set corresponding to the same real world entity.

Relation	A_1	A_2	A_3	A_4	A_5
BUY	Bose Acoustimass 5 Series III Speaker System - 21725	2.1-channel - Black	359.00	202812620	BOSE
ABT	Bose Acoustimass 5 Series III Speaker System - AM53BK	Bose Acoustimass 5 Series III Speaker System - AM53BK/ 2 Dual Cube Speakers With Two 2-1/2' Wide-range Drivers In Each Speaker/ Powerful Bass Module With Two 5-1/2' Woofers/ 200 Watts Max Power/ Black Finish	399.00	580	—

A_2 is a description, A_3 is a price. A_4 is an identification number, and A_5 is a manufacturer. The attributes of the "Buy" and "Abt" relations are $A_1 \times A_2 \times A_3 \times A_4 \times A_5$ and $A_1 \times A_2 \times A_3 \times A_4$, respectively. Table 4.1 shows an example of "Buy" and "Abt" records corresponding to the same real world entity.

For the application of the optimization-based EM approach we use only the first three attributes, i.e., we set $R_1 = R_2 = A_1 \times A_2 \times A_3$. Let C_1 and C_2 be the corpora of the name and description fields in the Abt-Buy set and D_1, D_2 – the corresponding dictionaries, i.e, the sets of distinct words occurring in each corpus. We identify the attribute spaces for the name and description fields in Abt-Buy with the respective corpora, i.e., $A_1 = C_1$, and $A_2 = C_2$. An obvious choice for A_4 is the set of non-negative real numbers. Using the notation from Section 4.1 the records in R_k have the form $r_{k,l} = \{a_{1,l}, a_{2,l}, a_{3,l}\}$, where $a_{1,l}$ and $a_{2,l}$ are "bags of words" and $a_{3,l}$ is a non-negative real number.

Our approach consists of two sages: training and testing stages. In the training stage, we use training sets to estimate the weights for the optimal superposition similarity measure. In the testing stage the weights are applied to the same set of metrics used to estimate the weights and the resulting optimal similarity superposition measure is applied to a testing set, and errors are calculated. In the study we use the four nested subsets of the Abt-Buy database described in Section 3.3 for training sets. These sets, termed Set 1, 2, 3 and 4 comprise of 50, 100, 150 and 200 randomly selected records respectively. To ensure that the sets are nested, we define them recursively by first selecting 200 random records for Set 4, then selecting randomly 150 of these records for Set 3 and so on.

For each set we proceed to compute the weights for the optimal similarity matrix using different subsets of the generalized similarities metrics as defined in Section 3.2 . Then, we solve the LSAP problem and compute the training error. In the testing stage, we apply the weights to define the optimal superposition similarity measure and apply the measure to a testing set consisting of 500 randomly selected pairs of records. The solution the LSAP problem is then used to calculate the accuracy with the EM performed for the testing set. The records in the testing set are selected randomly but in a way such that they do not include elements from the training sets.

As in the previous studies, the cosine similarity with token measures (2.1.3) and (2.1.5) provides the corresponding reference error, while the Jaccard (2.3.2), NWI (2.3.3) and Dice (2.3.4) similarities with $p = 1$ and the token weight (2.1.3) provide for a comparison with the set-based approach.

Dataset	Set 1 (M=50)		Set 2 (M=100)		Set 3 (M=150)		Set 4 (M=200)	
$S(\cdot,\cdot)$ $\omega \rightarrow$	tf*idf	idf	tf*idf	idf	tf*idf	idf	tf*idf	idf
cos	**44**	**52**	**52**	**54**	**59.33**	**60.67**	**59**	**62**
J_1	52	*48*	51	*54*	60.00	*60.67*	61.00	*61.50*
N_1	48	*50*	50	*54*	59.33	*62.67*	60.00	*64*
D_1	52	*48*	51	*54*	58.67	*60.67*	60.00	*62*
J_2	46	48	52	56	58.67	60.67	59.50	62
N_2	48	48	52	55	60.00	62	60	64.50
D_2	46	52	52	56	60.00	60.67	60.00	63
Δ_1	64	52	64	56	70.67	64.67	71.50	64
Δ_2	48	52	52	56	62.00	64	62.50	64
Δ_5	48	50	53	56	64.00	62.67	62.50	62

Table 4.2: Error in [%] in the solution of the entity matching problem for the descriptor attribute, using extended set-based similarity functions and comparing the tf*idf with idf token weights for the Abt-Buy database. Errors corresponding to the standard cosine similarity are in boldface. Errors corresponding to standard set-based similarity are in italics.

4.3.1 Approximation of a canonical similarity function

Because assignment of identification numbers is an ad hoc process, the attribute values in A_4 and A_5 have limited value for comparing records. For this reason we do not consider A_{4-5} in the approximation of the optimal similarity function. For the "price" attribute we use a single similarity function defined as relative numerical error

$$s_{31}(a_{3,k}, a_{3,l}) = \frac{|a_{3,k} - a_{3,l}|}{\max\{|a_{3,k}|, |a_{3,l}|\}}.$$

To measure similarity of A_1 and A_2 we use the generalized similarity metrics from the extended vector space model developed in the previous chapter.

Table 3.6 compares the the errors from application of individual similarity metrics to the name attribute only in the Abt-Buy database. Table 4.2 compares the errors from application of individual similarity metrics to the descriptor attribute only in the Abt-Buy database. In both cases, the similarity metrics are applied to the four sets and and errors are calculated after solving the LSAP. The tables also compares the errors between the similarity metrics using tf*idf with idf token weights. The first observation is that the errors values for the descriptor attribute are much higher compared to the error values for the name attribute as described in Table 3.6. For the set of fifty documents, the error values given by some of the similarity metrics for the descriptor attribute are lower than 50 [%]; however, as the size of the set increase to 200 the error values for all of the similarity metrics vary in the range of 60–70 [%]. These results indicate that the descriptor attribute has less discriminative power compared to the name attribute.

For the training sets 1–4, the EM with only price attribute performs poorly and gives errors of

Testing Dataset (M=500)	Name		Descriptor	
$S(\cdot,\cdot)$ $\omega \rightarrow$	tf*idf	idf	tf*idf	idf
cos	**13**	**12.80**	**70.20**	**70.00**
J_1	11.60	*12.00*	72.40	*71.40*
N_1	12.20	*14.80*	71.80	*73.00*
D_1	11.40	*12.40*	72.20	*70.80*
J_2	12.80	12.00	71	72
N_2	13.20	15.20	71.40	73.20
D_2	13.20	13.20	71.20	71.20
Δ_1	14.60	15.80	78.20	70.60
Δ_2	19.00	18.80	73.80	70
Δ_5	30.20	25.80	75.60	70.80

Table 4.3: Error in [%] in the solution of the entity matching problem for the testing set for name and descriptor attributes extended set-based similarity functions comparing the tf*idf with idf token weights for the Abt-Buy database.

96 [%], 97 [%], 97.30 [%] and 97 [%] for the four sets respectively.

This observation is confirmed by the results in table 4.3 that compares the errors for the name and the descriptor attributes for the data set containing 500 pairs of documents. The results in Table 4.3 show that on average the error values for the descriptor attribute are three to four times higher compared to the name attribute. The lower error value for the name attribute is given by the $J_{1,2}$ similarity metrics with tf*idf weighting. The lower error value for the descriptor attribute is given by the cosine and Δ_2 similarity metrics with idf weighting; however, the error of 70 [%] is very high and approximately six time higher than the lower error for the name attribute. The EM with only price attribute gives again very high error of 98 [%] using similarity function defined as relative numerical error.

4.3.2 Selection of optimal superposition similarity function from subset of two similarities

Tables 4.2– 4.4 show the results for the weights selection that approximates the optimal superposition similarity function when two similarities are used in the selection process. The similarities are selected in such way as to represent the three main classes of similarities: set-based, distance-based, and the benchmark cosine. The tables compare the results for the four training sets (Sets 1–4) and the test set (Set 5), and when using ℓ_1, ℓ_2 and ℓ_∞ norms for the optimizations functional.

Table 4.4 shows the classification results and the weights selection for the optimal similarity metric based on superposition of $N_1 - \mathsf{idf}$ and $\Delta_5 - \mathsf{tf} * \mathsf{idf}$ similarities. $N_1 - \mathsf{idf}$ is selected as

Weight	S_{ij}	Attribute	50	100	150	200
		ℓ_1 optimization				
y_{11}	$N_1 - \mathsf{idf}$	A_1	1	1	1	1
y_{12}	$\Delta_5 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
		Training Error [%]	0	9	7.33	8.50
		Testing Error [%]	14.80	14.80	14.80	14.80
		ℓ_2 optimization				
y_{11}	$N_1 - \mathsf{idf}$	A_1	0.67	0.67	0.71	0.72
y_{12}	$\Delta_5 - \mathsf{tf} * \mathsf{idf}$	A_1	0.33	0.33	0.29	0.28
		Training Error [%]	0.00	9.00	8.66	8.50
		Testing Error [%]	12.80	12.80	13.40	13.60
		ℓ_∞ optimization				
y_{11}	$N_1 - \mathsf{idf}$	A_1	0	0	0	0
y_{12}	$\Delta_5 - \mathsf{tf} * \mathsf{idf}$	A_1	1	1	1	1
		Training Error [%]	6.00	22.00	24.00	27.00
		Testing Error [%]	30.20	30.20	30.20	30.20

Table 4.4: Weight selection and training (1–4) and testing sets errors for ℓ_1, ℓ_2 and ℓ_∞ optimizations using $N_1 - \mathsf{idf}$ and Δ_5-tf*idf similarities for the name attribute in the Abt-Buy database.

representative of set-based similarity measures and $\Delta_5 - \mathsf{tf} * \mathsf{idf}$ as a representative of the distance-based similarity measures. The individual errors for these two similarities for the training and testing sets can be found in Tables: 3.6 and 4.3 respectively. The results in Table 4.4 show that the ℓ_1 norm of the optimization functional selects the better performing similarities of the two: $N_1 - \mathsf{idf}$ for all four training sets. Conversely, the ℓ_∞ norm of the optimization functional always selects the worst performing of the two: $\Delta_5 - \mathsf{tf} * \mathsf{idf}$ resulting in a higher classification error for the testing set compare to the ℓ_1 optimization. The results is consistent with the higher sensitivity of the ℓ_1 norm to outliers.

It is important to note that these two norms of the optimization functional: ℓ_1 and ℓ_∞ are not sensitive to the increase of the training set size and perform consistently even with limited number of training samples. This fact indicates that a small numbers of labeled data may be sufficient for good estimation of the weights.

The ℓ_2 norm of the optimization functional, however, estimates nonzero weights for the both similarity metrics, assigning higher weight to the better performing metric $N_1 - \mathsf{idf}$. Moreover, for the training sets size of 50 and 100 samples, the optimal superposition metric gives better results compared to the individual metric performance and the same as the benchmark cosine metric. This result demonstrates that (1) a superposition of two similarities can achieve improved performance compared to individual similarities performances and (2) superposition of similarities can approximate results achieved by top performing similarities or could perform better.

Table 4.5 shows the results for the weights selection and the corresponding errors when $D_2 - \mathsf{idf}$ and cosine-$\mathsf{tf} * \mathsf{idf}$ similarities are used in the selection process. $D_2 - \mathsf{idf}$ is selected as representative of the set-based similarities and cosine-$\mathsf{tf} * \mathsf{idf}$ represent the cosine class. The individual errors

Weight	S_{ij}	Attribute	50	100	150	200
		ℓ_1 optimization				
y_{11}	$D_2 - \mathsf{idf}$	A_1	1	1	1	1
y_{12}	cosine$-\mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
		Training Error [%]	0	7	7.33	7.00
		Testing Error [%]	13.2	13.20	13.20	13.20
		ℓ_2 optimization				
y_{11}	$D_2 - \mathsf{idf}$	A_1	1	1	1	1
y_{12}	cosine-$\mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
		Training Error [%]	0	7	7.33	7.00
		Testing Error [%]	13.20	13.20	13.20	13.20
		ℓ_∞ optimization				
y_{11}	$D_2 - \mathsf{idf}$	A_1	1	1	1	1
y_{12}	cosine-$\mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
		Training Error [%]	0	7	7.33	7
		Testing Error [%]	13.2	13.20	13.20	13.20

Table 4.5: Training (1–4) and testing sets errors for ℓ_1, ℓ_2 and ℓ_∞ optimizations using $D_2 - \mathsf{idf}$ and cosine-$\mathsf{tf} * \mathsf{idf}$ similarities for the name attribute in Abt-Buy database.

for these two similarities for the training and testing sets can be found in Tables: 3.6 and 4.3. Note that these two similarities perform equally well for the four training sets. The results in Table 4.5 show that the ℓ_1, ℓ_2 and ℓ_∞ optimizations select always one of the the two similarities: $D_2 - \mathsf{idf}$. These results indicate that the cosine and the set-based similarities can be considered to belong to the same equivalence class of similarities.

As in the case above, the norms of the optimization functional: $\ell_{1,2}$ and ℓ_∞ are not sensitive to the increase of the training set size and perform consistently even with limited number of training samples. This fact demonstrate again that a small numbers of labeled data could provide good estimation of the weights.

Table 4.6 shows the results for the third configuration of set of two similarities for selection of weights that approximate the optimal superposition similarity. In this case $\Delta_5 - \mathsf{tf} * \mathsf{idf}$ and cosine $\mathsf{tf} * \mathsf{idf}$ are selected as representative of the distance-based and the cosine similarity measures respectively. The individual errors for these two similarities for the training and testing sets can be found in Tables: 3.6 and 4.3 . The results in Table 4.6 show that the ℓ_1 optimization select the better performing similarities of the two: cosine$-\mathsf{tf} * \mathsf{idf}$. Conversely, the ℓ_∞ optimization selects the worst performing of the two: $\Delta_5 - \mathsf{tf} * \mathsf{idf}$ resulting in a higher classification error for the testing set compere to the ℓ_1 optimization. The results is consistent with the results shown in Table 4.6and the higher sensitivity of the ℓ_1 norm to outliers.

It is important to note again, that these two norms of the optimization functional: ℓ_1 and ℓ_∞ are not sensitive to the increase of the size of the training set and perform consistently even with limited number of training samples. This fact indicates that a small numbers of labeled data may be sufficient for good estimation of the weights.

Weight	S_{ij}	Attribute	50	100	150	200
		ℓ_1 optimization				
y_{11}	cosine-$\mathsf{tf} * \mathsf{idf}$	A_1	1	1	1	1
y_{12}	$\Delta_5 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
		Training Error [%]	0	7	7.33	7.50
		Testing Error [%]	13	13	13	13
		ℓ_2 optimization				
y_{11}	cosine-$\mathsf{tf} * \mathsf{idf}$	A_1	0.76	0.71	0.73	0.75
y_{12}	$\Delta_5 - \mathsf{tf} * \mathsf{idf}$	A_1	0.24	0.29	0.27	0.25
		Training Error [%]	0	7	7.33	7
		Testing Error [%]	12.60	12.80	12.59	12.80
		ℓ_∞ optimization				
y_{11}	cosine-$\mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
y_{12}	$\Delta_5 - \mathsf{tf} * \mathsf{idf}$	A_1	1	1	1	1
		Training Error [%]	6	22	24	27
		Testing Error [%]	30.20	13.20	30.20	30.20

Table 4.6: Training (1–4) and testing sets errors for ℓ_1, ℓ_2 and ℓ_∞ optimizations using cosine-$\mathsf{tf} * \mathsf{idf}$ and $\Delta_5 - \mathsf{tf} * \mathsf{idf}$ similarities for the name attribute in Abt-Buy database.

The ℓ_2 norm of the optimization functional, however, estimates nonzero weights for the both similarity metrics, assigning higher weight to the better performing metric cosine$-\mathsf{tf} * \mathsf{idf}$. Moreover, for all the training sets, the optimal superposition metric gives better results compared to the best performer of the two individual metric and thus performs better than the benchmark cosine metric. This results are consistent with the results presented in Table 4.4, where the optimization includes also the distance-based class but in this case in conjunction with the set-based class of similarities. The fact that the ℓ_2 norm of the optimization functional distributes the weights between the cosine/set-based and the distance-based similarities suggests the the distance-based class of similarities does not belong to the same equivalence class as the cosine and set-based similarities and thus they provide independent measure similarity between documents.

4.3.3 Selection of the optimal similarity superposition function based on the set-based similarities class

In this section we present the results for the optimal superposition similarity function, defined and optimized over the set-based similarities class. More specifically, the search for the weights is defined over the set of similarities that includes $J_{1,2} - \mathsf{tf} * \mathsf{idf}, \mathsf{idf}$, $N_{1,2} - \mathsf{tf} * \mathsf{idf}, \mathsf{idf}$, $D_{1,2} - \mathsf{tf} * \mathsf{idf}, \mathsf{idf}$.

Tables 4.10–4.9 show the results for the weights selection and optimal superposition similarity function for ℓ_1, ℓ_2 and ℓ_∞ optimization respectively. As observed in the other case studies, the ℓ_1 norm of the functional as shown in Table 4.10 tends to select the optimally performing metric,

Weight	S_{ij}	Attribute	50	100	150	200
$y_{1,1}$	$J_1 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,2}$	$N_1 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,3}$	$D_1 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,4}$	$J_2 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,5}$	$N_2 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,6}$	$D_2 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,7}$	$J_1 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,8}$	$N_1 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,9}$	$D_1 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,10}$	$J_2 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,11}$	$N_2 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,12}$	$D_2 - \mathsf{idf}$	A_1	1	1	1	1
		Training Error [%]	0	7	7.33	7
		Testing Error [%]	13.20	13.20	13.20	13.20

Table 4.7: Training (1–4) and testing sets errors for the EM problem, for ℓ_1 optimizations using set-based similarities for the name attribute in Abt-Buy database.

Weight	S_{ij}	Attribute	50	100	150	200
$y_{1,1}$	$J_1 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,2}$	$N_1 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,3}$	$D_1 - t\mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,4}$	$J_2 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,5}$	$N_2 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,6}$	$D_2 - \mathsf{tf} * \mathsf{idf}$	A_1	0.9770	0	0	0
$y_{1,7}$	$J_1 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,8}$	$N_1 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,9}$	$D_1 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,10}$	$J_2 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,11}$	$N_2 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,12}$	$D_2 - \mathsf{idf}$	A_1	0.0229	1	1	1
		Training Error [%]	0.00	7.00	7.33	7.00
		Testing Error [%]	13.20	13.2	13.20	13.20

Table 4.8: Training (1–4) and testing sets errors for the EM problem, for ℓ_2 optimizations using set-based similarities for the name attribute in Abt-Buy database.

which in this case is $D_2 - \mathsf{idf}$. The $D_2 - \mathsf{idf}$ perform the same as the benchmark cosine $-\mathsf{tf} * \mathsf{idf}$ for the four training sets. The ℓ_2 norm of the functional and for a training size of 50 samples as shown in Table 4.8, selects superposition of two similarities: $D_2 - \mathsf{idf}$ and $D_2 - \mathsf{tf} * \mathsf{idf}$ assigning a very high weight of 0.98 to $D_2 - \mathsf{tf} * \mathsf{idf}$ and a small weight of 0.02 to $D_2 - \mathsf{idf}$. As the training size increase to size of 100 and 200 records, the optimization selects only one of the twelve similarities $D_2 - \mathsf{idf}$, assigning weight of 1 to it and 0 to the rest of the similarities. The testing error for the 50

Weight	S_{ij}	Attribute	50	100	150	200
$y_{1,1}$	$J_1 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,2}$	$N_1 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,3}$	$D_1 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,4}$	$J_2 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,5}$	$N_2 - \mathsf{tf} * \mathsf{idf}$	A_1	0.51	0.29	0.37	0
$y_{1,6}$	$D_2 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0.22	0.21	0.48
$y_{1,7}$	$J_1 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,8}$	$N_1 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,9}$	$D_1 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,10}$	$J_2 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,11}$	$N_2 - \mathsf{idf}$	A_1	0	0.27	0.22	0
$y_{1,12}$	$D_2 - \mathsf{idf}$	A_1	0.49	0.21	0.20	0.51
Training Error [%]			0	7	7.33	7
Testing Error [%]			13.20	13.20	13.40	12.60

Table 4.9: Training (1–4) and testing sets errors for the EM problem, for ℓ_∞ optimizations using set-based similarities for the name attribute in Abt-Buy database.

training samples is equal to the cases with bigger size of the training sets.

In contrast with earlier observations, when the selection was performed over two classes of similarities and the ℓ_∞ optimization tended to select one of the classes, usually the outlier, the results presented in Table 4.9 show that in this case the ℓ_∞ optimization selects more than one metric of the same class for the creation of the optimal superposition similarity function. For example, for training set size of 50 samples, the ℓ_∞ optimization selects 2 similarities: $N_2 - \mathsf{tf} * \mathsf{idf}$ and $D_2 - \mathsf{idf}$ with weights of 0.51 and 0.49 respectively. When the size of the training set increases to 100, the weights become almost uniformly distributed between 4 similarities: $N_2 - \mathsf{tf} * \mathsf{idf}$ and $D_2 - \mathsf{idf}$ and $D_2 - \mathsf{tf} * \mathsf{idf}$ and $D_2 - \mathsf{idf}$. Similar observations holds for the case of 150 training samples, however, the highest weight value of 0.37 is assigned to $N_2 - \mathsf{tf} * \mathsf{idf}$. When the size of the training set however increases to 200, only two metric again participate in the optimal superposition similarity metric: this time $D_2 - \mathsf{tf} * \mathsf{idf}$ and $D_2 - \mathsf{idf}$ with weights of 0.48 and 0.51 respectively and providing the lowest testing set classification error of 12.60 [%].

The results from this study and in particular the ℓ_∞ optimization demonstrate that multiple combinations of similarities from the set-based class can lead to similar performance of the EM, suggesting that the similarities within the set-based class of similarities exhibit very similar characteristics. More studies need to be conducted with increased training data sets size to investigate further sensitivity and convergence of the performance.

Weight	S_{ij}	Attribute	50	100	150	200
$y_{1,1}$	$J_1 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,2}$	$N_1 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,3}$	$D_1 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,4}$	$J_2 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,5}$	$N_2 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,6}$	$D_2 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,7}$	$J_1 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,8}$	$N_1 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,9}$	$D_1 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,10}$	$J_2 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,11}$	$N_2 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,12}$	$D_2 - \mathsf{idf}$	A_1	1	1	1	1
$y_{1,13}$	cosine $-\mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,14}$	$\Delta_1 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,15}$	$\Delta_2 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,16}$	$\Delta_\infty - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,17}$	cosine $-\mathsf{idf}$	A_1	0	0	0	0
$y_{1,18}$	$\Delta_1 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,19}$	$\Delta_2 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,20}$	$\Delta_\infty - \mathsf{idf}$	A_1	0	0	0	0
		Training Error [%]	0	7	7.33	7
		Testing Error [%]	13.20	13.2	13.20	13.20

Table 4.10: Training (1–4) and testing sets errors for the EM problem, for ℓ_1 optimizations using the set-based distance, and cosine similarities classes for the name attribute in Abt-Buy database.

4.3.4 Selection of the optimal superposition similarity function based on the set-based, distance and cosine similarities classes

In this subsection we present results for the optimal selection of the superposition similarity function overt the full space of the generalized similarity metrics. In particular, we include all similarities presented earlier that cover the three main classes: generalized set-based, generalized distance-based and the benchmark cosine with tf*idf and idf encodings.

Tables 4.10–4.12 show the results for the weights selection and optimal superposition similarity function for ℓ_1, ℓ_2 and ℓ_∞ optimization respectively. As observed in the case study presented in Section 4.3.3 , the ℓ_1 norm of the functional tends to select the optimally performing metric, which in this case is also $D_2 - \mathsf{idf}$. The $D_2 - \mathsf{idf}$ perform the same as the benchmark cosine $-\mathsf{tf} * \mathsf{idf}$ for the four training sets. The ℓ_2 norm of the functional as shown in Table 4.11 and for a training size of 50 samples, selects superposition of two similarities: $D_2 - \mathsf{idf}$ and $D_2 - \mathsf{tf} * \mathsf{idf}$ assigning a very high eight of 0.9768 to $D_2 - \mathsf{tf} * \mathsf{idf}$ and a small weight of 0.0231 to $D_2 - \mathsf{idf}$. As the training size increase to 100 and 200, the optimization selects only $D_2 - \mathsf{idf}$, assigning weight of 1 to it and 0 to the rest of the similarities. The testing error for the 50 training samples is equal to the cases

Weight	S_{ij}	Attribute	50	100	150	200
$y_{1,1}$	$J_1 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,2}$	$N_1 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,3}$	$D_1 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,4}$	$J_2 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,5}$	$N_2 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,6}$	$D_2 - \mathsf{tf} * \mathsf{idf}$	A_1	0.98	0	0	0
$y_{1,7}$	$J_1 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,8}$	$N_1 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,9}$	$D_1 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,10}$	$J_2 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,11}$	$N_2 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,12}$	$D_2 - \mathsf{idf}$	A_1	0.02	1	1	1
$y_{1,13}$	cosine $-\mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,14}$	$\Delta_1 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,15}$	$\Delta_2 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,16}$	$\Delta_\infty - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,17}$	cosine $-\mathsf{idf}$	A_1	0	0	0	0
$y_{1,18}$	$\Delta_1 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,19}$	$\Delta_2 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,20}$	$\Delta_\infty - \mathsf{idf}$	A_1	0	0	0	0
Training Error [%]			0	7	7.33	7
Testing Error [%]			13.20	13.2	13.20	13.20

Table 4.11: Training (1–4) and testing sets errors for the EM problem, for ℓ_2 optimizations using the set-based distance, and cosine similarities classes for the name attribute in Abt-Buy database.

with bigger size of the training sets, showing equivalency in the performance of the $D_2 - \mathsf{idf}$ and $D_2 - \mathsf{tf} * \mathsf{idf}$ and the ability of the optimization to select it.

Consistently with the earlier observations, the ℓ_∞ optimization, as shown in Table 4.12 tends to select more than one metric for the creation of the optimal superposition similarity function. For example, for training set size of 50 samples, the ℓ_∞ optimization selects 3 similarities: $D_2 - \mathsf{tf} * \mathsf{idf}$, $\Delta_\infty - \mathsf{idf}$ and $\Delta_\infty - \mathsf{tf} * \mathsf{idf}$ with the highest weight assigned to the $\Delta_\infty - \mathsf{tf} * \mathsf{idf}$. As the training size increase to 100 and above, the selection of the weights converges to two similarities: $\Delta_\infty - \mathsf{tf} * \mathsf{idf}$ and $\Delta_\infty - \mathsf{idf}$ with almost uniform weight distribution between the two. The classification error for the testing set is higher in this cases (27 [%])compared to the weight selection based on 50 records training set size (24 [%]). These results confirm again that the set-based and distance-based similarities do not belong to the same equivalence classes. In general, the set -based class of similarities performs better than the distance-based for the EM problem. Including set-based similarities in conjunction with similarities from the distance-based class could enhance the performance of the EM problem with respect to the case when only the similarities from the distance-based class are considered.

Weight	S_{ij}	Attribute	50	100	150	200
$y_{1,1}$	$J_1 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,2}$	$N_1 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,3}$	$D_1 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,4}$	$J_2 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,5}$	$N_2 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,6}$	$D_2 - \mathsf{tf} * \mathsf{idf}$	A_1	0.02	0	0	0
$y_{1,7}$	$J_1 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,8}$	$N_1 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,9}$	$D_1 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,10}$	$J_2 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,11}$	$N_2 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,12}$	$D_2 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,13}$	cosine $-\mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,14}$	$\Delta_1 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,15}$	$\Delta_2 - \mathsf{tf} * \mathsf{idf}$	A_1	0	0	0	0
$y_{1,16}$	$\Delta_\infty - \mathsf{tf} * \mathsf{idf}$	A_1	0.83	0.46	0.48	0.51
$y_{1,17}$	cosine $-\mathsf{idf}$	A_1	0	0	0	0
$y_{1,18}$	$\Delta_1 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,19}$	$\Delta_2 - \mathsf{idf}$	A_1	0	0	0	0
$y_{1,20}$	$\Delta_\infty - \mathsf{idf}$	A_1	0.15	0.54	0.52	0.49
		Training Error [%]	0	20	22.67	24
		Testing Error [%]	24	27.20	27.40	27.80

Table 4.12: Training (1–4) and testing sets errors for the EM problem, for ℓ_∞ optimizations using the set-based, distance, and cosine similarities classes for the name attribute in Abt-Buy database.

4.3.5 Selection of the optimal superposition similarity function over different attributes

In this subsection we present the results for the selection of optimal similarity metric when the search is performed over different similarities for different attributes such as name, and descriptor. Table 4.13 shows the results for selection of optimal similarity function when two attributes are considered: name and descriptor. The cosine–$\mathsf{tf} * \mathsf{idf}$ similarity metric is applied to both. In this case, all three optimization norms and for almost all training set sizes the optimization selects the name attribute with the cosine–$\mathsf{tf} * \mathsf{idf}$ metric and assigns 0 weight to the descriptor attribute. This results in 13 [%] error for the testing set. The only exception is the ℓ_∞ optimization for the training set size of 50, where the weights are assigned between name and descriptor attributes. This results in the lowest observed classification error of 9 [%].

Table 4.14 shows the results for selection of optimal similarity function when again the same two attributes are considered: name and descriptor, and $\Delta_5 - \mathsf{tf} * \mathsf{idf}$ metric from the distance-based similarities class is applied to each attribute. In this case, similar to the results presented in Tables 4.13, $\ell_{1,2}$ optimizations norms, for all training set sizes selects the name attribute and

Weight	S_{ij}	Attribute	50	100	150	200
		ℓ_1 optimization				
y_{11}	cosine–$\mathsf{tf} * \mathsf{idf}$	A_1	1	1	1	1
y_{12}	cosine–$\mathsf{tf} * \mathsf{idf}$	A_2	0	0	0	0
		Training Error [%]	0	7	7.33	7.50
		Testing Error [%]	13	13	13	13
		ℓ_2 optimization				
y_{11}	cosine–$\mathsf{tf} * \mathsf{idf}$	A_1	1	1	1	1
y_{12}	cosine–$\mathsf{tf} * \mathsf{idf}$	A_2	0	0	0	0
		Training Error [%]	0	7	7.33	7.50
		Testing Error [%]	13	13	13	13
		ℓ_∞ optimization				
y_{11}	cosine–$\mathsf{tf} * \mathsf{idf}$	A_1	0.61	1	1	1
y_{12}	cosine–$\mathsf{tf} * \mathsf{idf}$	A_2	0.39	0	0	0
		Training Error [%]	0	7	7.33	7.50
		Testing Error [%]	9	13	13	13

Table 4.13: Training (sets 1–4) and testing sets errors for ℓ_1, ℓ_2 and ℓ_∞ norms of the optimization functional using cosine similarity metric with tf*idf encoding.

Weight	S_{ij}	Attribute	50	100	150	200
		ℓ_1 optimization				
y_{11}	$\Delta_5 - \mathsf{tf} * \mathsf{idf}$	A_1	1	1	1	1
y_{12}	$\Delta_5 - \mathsf{tf} * \mathsf{idf}$	A_2	0	0	0	0
		Training Error	6	22	22.67	27
		Testing Error	30.20	30.20	30.40	30.20
		ℓ_2 optimization				
y_{11}	$\Delta_5 - \mathsf{tf} * \mathsf{idf}$	A_1	1	1	1	1
y_{12}	$\Delta_5 - \mathsf{tf} * \mathsf{idf}$	A_2	0	0	0	0
		Training Error	6	22	24	27
		Testing Error	30.20	30.20	30.20	30.20
		ℓ_∞ optimization				
y_{11}	$\Delta_5 - \mathsf{tf} * \mathsf{idf}$	A_1	0.39	0.03	0.19	0.13
y_{12}	$\Delta_5 - \mathsf{tf} * \mathsf{idf}$	A_2	0.61	0.97	0.81	0.87
		Training Error	4	25	28.77	33
		Testing Error	28.40	52.20	59.39	38.39

Table 4.14: Training (sets 1–4) and testing sets errors for ℓ_1, ℓ_2 and ℓ_∞ norms of the optimization functional, and using Minkowski similarity measure with tf*idf weighting scheme.

assigns 0 weight to the descriptor attribute. This results in 30.20 [%] error for the testing set. The ℓ_∞ optimization however, distributes the weight between the name and descriptor attributes. For training set with size 50, the optimization assigns weight of 0.61 to the descriptor and 0.39 to name attribute. This results in the lowest observed error of 28.40 [%]for this particular configuration. As the size of the training set increases, the wight for the descriptor attribute increases, which results in higher classification error for the testing set of 38.39 [%] for the training set size of 200.

Weight	S_{ij}	Attribute	50	100	150	200
		ℓ_1 optimization				
y_{11}	$\Delta_2-\mathsf{tf}*\mathsf{idf}$	A_1	1	1	1	1
y_{12}	cosine–$\mathsf{tf}*\mathsf{idf}$	A_2	0	0	0	0
		Training Error	0	12	12.67	13
		Testing Error	19	19	19	19
		ℓ_2 optimization				
y_{11}	$\Delta_2-\mathsf{tf}*\mathsf{idf}$	A_1	1	1	1	1
y_{12}	cosine–$\mathsf{tf}*\mathsf{idf}$	A_2	0	0	0	0
		Training Error	0	12	12.67	13
		Testing Error	19	19	19	19
		ℓ_∞ optimization				
y_{11}	$\Delta_2-\mathsf{tf}*\mathsf{idf}$	A_1	0.9963	1	1	1
y_{12}	cosine–$\mathsf{tf}*\mathsf{idf}$	A_2	0.0037	0	0	0
		Training Error	0	12	12.67	13
		Testing Error	18	19	19	19

Table 4.15: Training (sets 1–4) and testing sets errors for ℓ_1, ℓ_2 and ℓ_∞ optimizations using $\Delta_2-\mathsf{tf}*\mathsf{idf}$ metric for the name attribute and cosine-$\mathsf{tf}*\mathsf{idf}$ for the Descriptor attribute.

Table 4.15 shows the results for selection of optimal similarity function when two attributes are considered: name and descriptor and two different type of similarities are applied to each attribute. In particular, $\Delta_2-\mathsf{tf}*\mathsf{idf}$ is applied to the name attribute and the cosine–$\mathsf{tf}*\mathsf{idf}$ similarity metric is applied to the descriptor attribute. In this case, similar to the results presented in Table 4.13, all three optimization norms and for almost all training set sizes the optimization selects the name attribute with the $\Delta_2-\mathsf{tf}*\mathsf{idf}$ metric and assigns 0 weight to the descriptor attribute. This results in 19 [%] error for the testing set. Again, the only exception is the ℓ_∞ optimization for the training size of 50, were the weights are assigned between name and descriptor attribute, with very high weight assigned to name and very low to the descriptor. This however, results in a slightly lower error for the test set of 18 [%] .

Table 4.16 shows the results for selection of optimal similarity function when two attributes are considered: name and descriptor for ℓ_∞ norm for the optimization of the functional. In contrast to the cases considered above, 12 similarities are applied to the attribute descriptor (all set based with $\mathsf{tf}*\mathsf{idf}$ and idf encoding) and cosine $-\mathsf{tf}*\mathsf{idf}$ to the name attribute. It is again interesting to note that for training size of 50 records, the optimization gives weight 0.77 to the name attribute with the cosine $-\mathsf{tf}*\mathsf{idf}$ metric and 0.23 to the descriptor with $D_2-\mathsf{idf}$ metric, leading to testing error of 8.8 [%]. This error value is lower compared to the error associated with the stand alone name attribute with the cosine $-\mathsf{tf}*\mathsf{idf}$ metric. As the size of the training data set increases to 200, the optimization selection converges to selection of only name attribute with the cosine $-\mathsf{tf}*\mathsf{idf}$ metric; however, the testing error increases with approximately 4 [%]. For the $\ell_{1,2}$ norms for the optimization of the functional, the optimization selects always only name attribute with the cosine

Weight	S_{ij}	Attribute	50	100	150	200
$y_{1,1}$	$J_1 - \mathsf{tf} * \mathsf{idf}$	A_2	0	0	0	0
$y_{1,2}$	$N_1 - \mathsf{tf} * \mathsf{idf}$	A_2	0	0	0	0
$y_{1,3}$	$D_1 - \mathsf{tf} * \mathsf{idf}$	A_2	0	0	0	0
$y_{1,4}$	$J_2 - \mathsf{tf} * \mathsf{idf}$	A_2	0	0	0	0
$y_{1,5}$	$N_2 - \mathsf{tf} * \mathsf{idf}$	A_2	0	0	0	0
$y_{1,6}$	$D_2 - \mathsf{tf} * \mathsf{idf}$	A_2	0	0.09	0.09	0
$y_{1,7}$	$J_1 - \mathsf{idf}$	A_2	0	0	0	0
$y_{1,8}$	$N_1 - \mathsf{idf}$	A_2	0	0	0	0
$y_{1,9}$	$D_1 - \mathsf{idf}$	A_2	0	0	0	0
$y_{1,10}$	$J_2 - \mathsf{idf}$	A_2	0	0	0	0
$y_{1,11}$	$N_2 - \mathsf{idf}$	A_2	0	0	0	0
$y_{1,12}$	$D_2 - \mathsf{idf}$	A_2	0.23	0	0	0
$y_{1,13}$	cosine $-\mathsf{tf} * \mathsf{idf}$	A_1	0.77	0.91	0.91	1
		Training Error [%]	0	4.67	4.67	7.50
		Testing Error [%]	8.8	9.80	9.80	13.00

Table 4.16: Training and testing errors for two attributes: descriptor and name, and training sets with varying sizes; ℓ_∞ norm for the optimization.

$-\mathsf{tf} * \mathsf{idf}$ metric for all the four sizes of the training set.

Weight	S_{ij}	Attribute	50	100	150	200
$y_{1,1}$	$J_1 - \mathsf{tf} * \mathsf{idf}$	A_2	0	0	0	0
$y_{1,2}$	$N_1 - \mathsf{tf} * \mathsf{idf}$	A_2	0	0	0	0
$y_{1,3}$	$D_1 - \mathsf{tf} * \mathsf{idf}$	A_2	0	0	0	0
$y_{1,4}$	$J_2 - \mathsf{tf} * \mathsf{idf}$	A_2	0	0	0	0
$y_{1,5}$	$N_2 - \mathsf{tf} * \mathsf{idf}$	A_2	0	0	0	0
$y_{1,6}$	$D_2 - \mathsf{tf} * \mathsf{idf}$	A_2	0.05	0.20	0	0.22
$y_{1,7}$	$J_1 - \mathsf{idf}$	A_2	0	0	0	0
$y_{1,8}$	$N_1 - \mathsf{idf}$	A_2	0	0	0	0
$y_{1,9}$	$D_1 - \mathsf{idf}$	A_2	0	0	0	0
$y_{1,10}$	$J_2 - \mathsf{idf}$	A_2	0	0	0	0
$y_{1,11}$	$N_2 - \mathsf{idf}$	A_2	0	0	0	0
$y_{1,12}$	$D_2 - \mathsf{idf}$	A_2	0.22	0.04	0.22	0
$y_{1,13}$	$\Delta_5 - \mathsf{tf} * \mathsf{idf}$	A_1	0.73	0.76	0.78	0.78
		Training Error [%]	8	17	23.33	19.5
		Testing Error [%]	24.40	23.60	22.20	21.40

Table 4.17: Training and testing errors for two attributes: descriptor and name, and training sets with varying sizes; ℓ_2 norm for the optimization.

Table 4.17 shows the results for selection of optimal similarity function when two attributes are considered: name and descriptor for ℓ_2 norm for the optimization of the functional. Similar to the case shown in Table 4.16,12 similarities are applied to the attribute descriptor (all set-based with $\mathsf{tf} * \mathsf{idf}$ and idf encoding) but the distance-based metric $\Delta_5 - \mathsf{tf} * \mathsf{idf}$ is applied to the name attribute. This case contrasts the case shown in Table 4.16, since the $\Delta_5 - \mathsf{tf} * \mathsf{idf}$ metric is lower performer for the name attribute compared to the benchmark cosine $-\mathsf{tf} * \mathsf{idf}$ metric.

For this case, the ℓ_2 norm for the optimization functional selects the $\Delta_5 - \mathsf{tf} * \mathsf{idf}$ metric applied to the name attribute along with the $D_2 - \mathsf{tf} * \mathsf{idf}$ metric applied to the descriptor attribute. Higher value is given to the $\Delta_5 - \mathsf{tf} * \mathsf{idf}$ metric in conjunction with name attribute; however, the error for the testing set achieved with the superposition of the two attributes and the corresponding similarities is better compared to the case when the attribute-metric are applied stand-alone.

4.4 Conclusions

In this chapter we formulated and developed a robust and flexible algorithm for EM that can adapt to various application contexts. More specifically, a supervised EM frameworks is developed, that interprets the EM as the combinatorial optimization problem of finding the maximum weight matching in a weighted bipartite graph connecting records from two databases. The casting of EM problems into LSAP offers valuable practical and theoretical advantages. There are efficient algorithms that solve LSAP in polynomial time. Availability of such algorithms allows us to reduce the task of solving the EM problem to the task of computing weights for the edges of the bipartite graph connecting the records from the databases. This in turn allows us to focus efforts on the development of robust and flexible methodologies for the estimation of the similarity between records that work across multiple application domains.

Our approach uses training data to approximate an optimal similarity superposition function for a given relation pair. This function is seek as linear combination of the generalized similarity functions for the common relation attributes. Solution of a suitably defined Quadratic Program (QP) defines the weights in the linear combination. Last but not least, LSAP tends to perform better than matching schemes based on greedy-type algorithms because it optimizes the assignment globally over the complete set of records.

Computational studies using the ABT-Buy database confirm the utility, flexibility and the robustness of our approach. Two distinct pattern in the behavior of the optimal similarity superposition function were observed when the search for the optimal function is restricted over two classes at time. The first pattern is that the optimization selects the best performing metric from the search space irregardless of the norm of the optimization functional. This is usually the case when the search for the optimal superposition similarity function is defined over the space of the set-based and cosine similarities. The second pattern observed is that ℓ_1 norm of the optimization functional selects the best performing metric, the ℓ_2 selects combination of similarities and ℓ_∞ selects the worst performing. This observation holds usually in the cases when the search for the optimal superposition similarity function is defined over the space of the set-based, cosine similarity , and the distance-based similarities. These results suggest that the optimal similarity superposition function forms two equivalence classes. The first class is the class for which the first pattern holds and this equivalence class comprises of set-based and cosine similarity metrics. The second equivalence class is the class for which the second patter holds and comprises of distance-based similarities.

The computational results for the various case studies have also demonstrate that superposition of similarities from different classes can achieve improved performance compared to individual similarities performances. Such example results are shown in Tables 4.4 and 4.6. Moreover, superposition of similarities that are not necessary top performers can approximate results achieved by top performing similarities or in some cases even better. These results facilitate the utility of the optimization approach. The analyst will not know in advance in many cases which similarity function or combinations thereof will turn out to give the best performance for a specific EM task. Therefore, it pays to include a range of different similarities so that the optimization can pick the best possible selection.

The computation results also demonstrate the ability of the optimization to rank correctly the importance and contribution of the individual attributes to the performance of the EM. In particular, the optimization properly recovers that descriptor and price attributes has less discriminative power compared to the name attribute. Last but not least, the results show that the optimization approach, especially for ℓ_1 and ℓ_2 norms of the optimization functional performs reliably even with small size of the labeled data. This indicates the utility of our method over probabilistic EM models and other models, that exhibit higher sensitivity to the size of the training sets and require larger training size to achieve satisfactory results.

References

[1] Benchmark datasets for entity resolution.
http://dbs.uni-leipzig.de/en/research/
projects/object_matching/fever
/benchmark_datasets_for_entity_resolution, Database Group Leipzig.

[2] Omar Benjelloun, Hector Garcia-Molina, David Menestrina, Qi Su, Steven Euijong Whang, and Jennifer Widom. Swoosh: a generic approach to entity resolution. *The VLDB Journal*, 18(1):255–276, January 2009.

[3] Dimitri P. Bertsekas. A new algorithm for the assignment problem. *Mathematical Programming*, 21:152–171, 1981. 10.1007/BF01584237.

[4] Dimitri P. Bertsekas. The auction algorithm for assignment and other network flow problems: A tutorial. *Interfaces*, 20(4):133–149, July/August 1990.

[5] David G. Brizan and Abdullah U. Tansel. A Survey of Entity Resolution and Record Linkage Methodologies. *Communications of the IIMA*, 6(3):41–50, 2006.

[6] Rainer Burkard. *Assignment problems*. SIAM, Society for Industrial and Applied Mathematics, Philadelphia, 2009.

[7] Rainer E. Burkard and Eranda Cela. Linear Assignment Problems and Extensions. In P.M. Pardalos and D.-Z. Du, editors, *Handbook of Combinatorial Optimization*, volume Supplement Volume A, pages 75–149. Kluwer Academic Publishers, 1999.

[8] D. Dey, S. Sarkar, and P. De. A distance-based approach to entity reconciliation in heterogeneous databases. *Knowledge and Data Engineering, IEEE Transactions on*, 14(3):567 –582, may/jun 2002.

[9] Debabrata Dey. Entity matching in heterogeneous databases: A logistic regression approach. *Decision Support Systems*, 44(3):740 – 747, 2008.

[10] Debabrata Dey, Sumit Sarkar, and Prabuddha De. A probabilistic decision model for entity matching in heterogeneous databases. *Manage. Sci.*, 44(10):1379–1395, October 1998.

[11] A.K. Elmagarmid, P.G. Ipeirotis, and V.S. Verykios. Duplicate record detection: A survey. *Knowledge and Data Engineering, IEEE Transactions on*, 19(1):1 –16, jan. 2007.

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

[13] L. Gu, R. Baxter, D. Vickers, and C Rainsford. Record linkage: Current practice and future directions. CMIS Technical Report 03/83, CSIRO Mathematical and Information Sciences, 2003.

[14] Marios Hadjieleftheriou and Divesh Srivastava. Weighted set-based string similarity. *IEEE Data Eng. Bull.*, 33(1):25–36, March 2010.

[15] A. Huang. Similarity measures for text document clustering. In *Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand*, pages 49–56, 2008.

[16] Matthew A. Jaro. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):pp. 414–420, 1989.

[17] Charles R. Johnson. *Matrix Theory and Applications.*, volume 40 of *Proceedings of symposia in applied mathematics*. American Mathematical Society, Providence, R.I., 1990.

[18] Myoung-Cheol Kim and Key-Sun Choi. A comparison of collocation-based similarity measures in query expansion. *Information Processing & Management*, 35(1):19 – 30, 1999.

[19] Hanna Köpcke and Erhard Rahm. Frameworks for entity matching: A comparison. *Data & Knowledge Engineering*, 69(2):197 – 210, 2010.

[20] Hanna Köpcke, Andreas Thor, and Erhard Rahm. Evaluation of entity resolution approaches on real-world match problems. *Proc. VLDB Endow.*, 3(1-2):484–493, September 2010.

[21] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

[22] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[23] H. B. Newcombe, J. M. Kennedy, S. J. Axford, and A. P. James. Automatic linkage of vital records. *Science*, 130(3381):954–959, 1959.

[24] Shraddha Pandit and Suchita Gupta. A comparative study on distance measuring approaches for clustering. *International Journal of Research in Computer Science*, 2(1):29–31, 2011.

[25] William E. Winkler. Overview of record linkage and current research directions. RESEARCH REPORT SERIES Statistics #2006-2, U.S. Census Bureau, Washington, DC 20233, 2003.

[26] S. K. M. Wong and Vijay V. Raghavan. Vector space model of information retrieval: a reevaluation. In *Proceedings of the 7th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '84, pages 167–185, Swinton, UK, UK, 1984. British Computer Society.

Appendix A

Definitions

Attribute: in a database management system an attribute may describe a component of the database, such as a table or a field, or may be used itself as another term for a field.

Database: an organized collection of data; collection of documents, images, database records, and combination of these can be viewed as database. The internet might be a database.

Information Retrieval: the activity of obtaining information resources relevant to an information need from a collection of information resources.

Metric: an abstraction of the notion of distance in a metric space; a real-valued function used to compare a single attribute of a record against a single attribute of another record or a query. Multiple metrics might address an individual mode.

Mode (information): different types of information: image, video, text, graphics, etc.; can be a mean of expression of an attribute. Given a set of data bits that represent information, the mode describes how to meaningfully interpret those bits.

Record (database): a set of fields in a database related to one entity.

Query: a question, composed of any combination of modes, against which records of the database will be compared to find best matches. Query is used as the way of retrieving the information from database.

DISTRIBUTION:

1	MS 9004	Howard Hirano, 8100
1	MS 1027	Arlo Ames, 5635
1	MS 1320	Pavel Bochev, 1414
1	MS 1081	Biliana Paskaleva, 6923
1	MS 1081	Marjorie McCornack, 6923
1	MS 0782	Paul Smith, 6633
1	MS 1027	Travis Bauer, 5635
1	MS 1138	Daniel Horschel, 6925
1	MS 0612	Review & Approval Desk, 4916
1	MS 0899	Technical Library, 9536 (electronic copy)
1	MS 0359	D. Chavez, LDRD Office, 1911

Sandia National Laboratories