A vector space model for information retrieval with generalized similarity measures. [*]

Biliana Paskaleva[†]
Sandia National Laboratories
P.O. Box 5800, MS 1138
Albuquerque, NM 87185-1138, USA
bspaska@sandia.gov

Pavel Bochev[‡]
Sandia National Laboratories
P.O. Box 5800, MS1320
Albuquerque, NM 87185-1320, USA
pbboche@sandia.gov

ABSTRACT

We develop a new set of similarity functions for a formal vector space model for information retrieval. Our model considers records as multisets of tokens. A token weight maps records into real vectors. Using this vector representation we define a p-norm of a record and pairwise conjunction and disjunction operations on records. With the help of these operations we then develop consistent extensions of set-based similarity functions and new ℓ_p distance-based similarities. We show that with particular classes of token weights and p-values, our definitions recover the standard versions of the similarity functions. The paper concludes with a preliminary study of the new similarities in the context of a model entity matching problem.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Vector space model, similarity measure, information retrieval, set-theoretic operations.

1. INTRODUCTION

Given a generating set of terms, and the associated term weights, the standard vector space model (VSM) [14, 16] for information retrieval encodes documents and queries as vectors of term weights. An integral part of VSM is a similarity function, which measures the closeness between documents. The normalized inner product between vectors defines the cosine similarity, which is a standard choice for a similarity measure 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 have demonstrated significant promise.

In this paper we use set-theoretic and Banach space ideas to develop a class of new similarity functions for the VSM. In particular, we obtain 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 similarity functions. We show that for some values of p and for specific term weighting choices some of the new similarity measures coincide with published set-based similarity functions [8, 9] and the standard VSM cosine similarity. As a result, our approach enables a consistent extension of a wide range of similarities to the VSM context. In a nutshell, we develop extensions of similarity functions, which bridge the standard vector space model with set-based approaches.

We present a preliminary comparative study of the new similarity functions in the context of an entity matching problem. To this end we use the Abt-Buy e-commerce set [1] in conjunction with two different assignment rules - a simple maximal similarity rule and a linear sum assignment rule.

We have organized the paper as follows. The rest of this section introduces some notation. Section 2 reviews the formal variant of the vector space model, which we use in the paper and Section 3 introduces the new similarity functions. In Section 4 we describe the model entity matching problem, the assignment rules and present the results from the study. We summarize our findings in Section 5.

Throughout the paper 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 Hadamar [11], product of $\mathbf{r}, \mathbf{s} \in \mathbf{R}^N$ is the vector $\mathbf{r} \circ \mathbf{s} = (r_1 s_1, \ldots, r_N s_N) \in \mathbf{R}^N$.

2. VECTOR SPACE MODEL

In this section we introduce a Vector Space Model (VSM) for information retrieval, specialized to our needs. Recall that a multiset is a pair (A, μ), where A is an underlying

[*] 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.

[†] High Confidence Systems Environments Department.

[‡] Numerical Analysis and Applications Department.

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.

Given a class of IR problems, let $T = \{t_1, t_2, \ldots, t_N\}$ be a corresponding set of N distinct index terms or keywords, which we call "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 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 $T \times C(T) \times C^2(T) \rightarrow R^+ \cup \{0\}$, 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

$$\omega(t, r, D) = \begin{cases} \alpha > 0 & \text{if } r \in D \text{ and } t \in r \\ 0 & \text{if } r \notin D \text{ or } d \in D \text{ and } t \notin r \end{cases} \qquad (1)$$

We review two examples of token weights. The inverse document frequency [14]

$$\mathsf{idf}(t, D) = \log\left(\frac{|D|}{1 + |D(t)|}\right) , \qquad (2)$$

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) satisfies condition (1):

$$\omega_{\mathsf{idf}}(t, r, D) = \mathsf{idf}(t, D)\nu_r(t) \,. \qquad (3)$$

In (3) $\nu_r(t)$ is the indicator function of r. The normalized term frequency [14]

$$\omega_{\mathsf{tf}}(t, r, D) := \mathsf{tf}(t, r) = \mu_r(t)/|r| \qquad (4)$$

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) and (4) [14, §6.2.2]:

$$\omega_{\mathsf{tf*idf}}(t, r, D) = \omega_{\mathsf{tf}}(t, r, D) \cdot \mathsf{idf}(t, D) \qquad (5)$$

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 [7]. We refer to [14, p.128] for additional variants of tf-idf measures.

Every token weight induces a mapping $\mathbf{w} : C(T) \mapsto \mathbf{R}^N$.

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

which maps records into vectors of token weights.

2.2 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 $\parallel \mathbf{r} \parallel_p = (\sum_{i=1}^N \mathbf{r}_i^p)^{1/p}$ 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 $\parallel \cdot \parallel_p$ and the map ω:

$$\parallel r \parallel_{\omega,p} = \parallel \mathbf{w}(r) \parallel_p \,. \qquad (6)$$

We define the conjunction of two records $s, r \in C(T)$ as

$$\parallel r \wedge s \parallel_{\omega,p} := \Big(\sum_{i=1}^N \mathbf{w}(r)^{p/2} \circ \mathbf{w}(s)^{p/2}\Big)^{1/p} \qquad (7)$$

and their disjunction as

$$\parallel s \vee r \parallel_{\omega,p} := \parallel s \parallel_{\omega,p} + \parallel r \parallel_{\omega,p} - \parallel r \wedge s \parallel_{\omega,p} \,, \qquad (8)$$

respectively. The conjunction (7) and the disjunction (8) 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*

$$\parallel r \wedge r \parallel_{\omega,p} = \parallel r \parallel_{\omega,p} \quad \textit{and} \quad \parallel r \vee r \parallel_{\omega,p} = \parallel r \parallel_{\omega,p} \,. \qquad (9)$$

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

$$\parallel r \wedge s \parallel_{\omega,p} = 0 \quad \textit{and} \quad \parallel r \vee s \parallel_{\omega,p} = \parallel r \parallel_{\omega,p} + \parallel s \parallel_{\omega,p} \,. \qquad (10)$$

PROOF. From (7) it follows that

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

The rest of (9) follows from this identity and (8). The proof of (10) is also straightforward. □

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, $\bar{r}$ is a set of tokens, i.e., a collection of unique elements of T. Note that $\bar{r}$ is a subset of T.

The set-based similarity [8, 12] represents 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, dissociate 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 is. This rules out application of token weights such as the tf measure (4) because the values of $\mu_{\bar{s}}$, $\mu_{\bar{s} \cup \bar{r}}$, and etc. 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 [8], to a general ℓ_p set-norm

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

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 $\omega(t,r,D) := \bar{\omega}(t,D)\cdot\nu_r(t)$. Then,*

$$\begin{gathered} \| r\wedge s \|_{\omega,p} = \| \bar{r}\cap\bar{s} \|_{\bar{\omega},p} \, ; \quad \| r\vee s \|_{\omega,p} = \| r\cup s \|_{\bar{\omega},p} \\ \text{and} \quad \| r \|_{\omega,p} = \| \bar{r} \|_{\bar{\omega},p} \; . \end{gathered} \tag{11}$$

PROOF. The assertion easily follows from definition (6) and by using the fact that $\omega(t,r,D) = \bar{\omega}(t,D)$ whenever $t \in D$. □

This proposition shows that the conjunction and disjunction functions in the vector space model represent consistent extensions of set-based norms of intersections and unions of sets, respectively. In other words, definitions (6), (7), and (8), allow us to bridge key notions in the vector space model and the set-based approach. This makes it possible to obtain consistent extensions of similarity measures from the set theory to the vector space model context.

3. SIMILARITY FUNCTIONS

A similarity function is a mapping $S : C(T)\times C(T) \mapsto [0,1]$. In this paper we restrict attention to similarity functions that are composition of the mapping $\mathbf{w}$ with a vector similarity function $\mathbf{s} : \mathbf{R}^N \times \mathbf{R}^N \mapsto [0,1]$. Succinctly, we assume that

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

where $\mathbf{s} : \mathbf{R}^N \times \mathbf{R}^N \mapsto [0,1]$.

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., [12] or [8] 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 (7), and (8) we extend the set-based definition to

$$J_p(r,s) := \frac{\| r\wedge s \|_{\omega,p}}{||r\vee s||_{\omega,p}} ; \quad p \geq 1 . \tag{12}$$

Extended Normalized Weighted Intersection similarity. This similarity function is related to the Jaccard coefficient but uses different normalization of the set intersection. Using (7), and (8) we obtain the extension

$$N_p(r,s) = \frac{\| r\wedge s \|_{\omega,p}}{\max\{\| r \|_{\omega,p}; \| s \|_{\omega,p}\}} ; \quad p \geq 1 . \tag{13}$$

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 is again in the normalization of the intersection term. The corresponding extension is

$$D_p(r,s) = \frac{2 \| r\wedge s \|_{\omega,p}}{\| r \|_{\omega,p} + \| s \|_{\omega,p}} ; \quad p \geq 1 . \tag{14}$$

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

$$\Delta_p(r,s) = 1 - \frac{\| r - s \|_{\omega,p}}{2\max\{\| r \|_{\omega,p}, \| s \|_{\omega,p}\}} ; \quad p \geq 1 . \tag{15}$$

The ℓ_p-distances corresponding to $p = 1$, $p = 2$ and $p = \infty$ are often called City Block, Euclidean and Chebyshev distance, respectively [15]. 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,*

$$\begin{aligned} J_p(r,s) &= \frac{\| \bar{r}\cap\bar{s} \|_{\bar{\omega},p}}{||\bar{r}\cup\bar{s}||_{\bar{\omega},p}}, \\ N_p(r,s) &= \frac{\| \bar{r}\cap\bar{s} \|_{\bar{\omega},p}}{\max\{\| \bar{r} \|_{\bar{\omega},p}, \| \bar{s} \|_{\bar{\omega},p}\}}, \\ D_p(r,s) &= \frac{2 \| \bar{r}\cap\bar{s} \|_{\bar{\omega},p}}{\| \bar{r} \|_{\bar{\omega},p} + \| \bar{s} \|_{\bar{\omega},p}}. \end{aligned} \tag{16}$$

PROOF. The proof follows directly from Proposition 2. □

This proposition confirms that (12)–(14) 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 [8], while for $p = 2$ the extended Jaccard similarity (12) recovers the Jaccard coefficient used in [9].

4. APPLICATION TO AN ENTITY MATCHING PROBLEM

In this section we report results from a preliminary study of the new similarity functions in the context of a model entity matching (EM) problem. The setting of the model EM problem assumes that there are two different sets of records, denoted by A and B, respectively, which represent the same set of real world entities $\mathcal{E}$ using two different relations. The task is to link the records from A and B corresponding to the same real world entity. We refer to [3] and [7] for comprehensive survey of EM problems.

Our study uses the Abt-Buy e-commerce set [1], which uses two different relations, "Abt" and "Buy", to describe the same set of products (entities). The "Abt" and "Buy" relations include attributes for a name, description, price, identification number and a manufacturer; see Table 1. The Abt-Buy provides the exact matches between the record pairs, which makes it appropriate for entity matching studies.

Table 1: Two records from the Abt-Buy e-commerce set corresponding to the same real world entity.

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	—

We base the entity matching on the "name" attribute, which gives a capsule summary of the product (entity). Our study uses two subsets of the Abt-Buy database, which we label as "Set I" and "Set II", respectively. The sets comprise of record pairs from relation "Abt" and relation "Buy", which correspond to the same entities. Set I has 100 such pairs and Set II - 122 pairs. The set T is the union of all unique terms in the "name" field of "Abt" and "Buy".

We explain the record matching approaches using Set I, the procedures are identical for Set II. Set I has 100 records from "Abt" and 100 records from "Buy". We denote the collections of these records by $A^{(I)}$ and $B^{(I)}$, respectively. Given a token weight ω, the corresponding term-to-document matrices $\mathbf{A}^{(I)}$ and $\mathbf{B}^{(I)}$ have element

$$\mathbf{A}^{(I)}_{ij} = \omega(t_j, a_i, A^{(I)}) \quad \text{and} \quad \mathbf{B}^{(I)}_{ij} = \omega(t_j, b_i, B^{(I)})\,,$$

respectively, where $t_j \in T$, $a_i \in A^{(I)}$, and $b_i \in B^{(I)}$. The rows of $\mathbf{A}^{(I)}$ and $\mathbf{B}^{(I)}$ are the vector space representations of the records in $A^{(I)}$ and $B^{(I)}$. For a given similarity function $S(\cdot,\cdot)$, the similarity matrix $\mathbf{S}$ for Set I has elements

$$\mathbf{S}_{ij} = S(a_i, b_j)\,.$$

This 100×100 matrix gives the pairwise similarity between the records in $A^{(I)}$ and and $B^{(I)}$. To match the records we use two different decision rules.

Maximum similarity assignment. This assignment strategy employes a simple, "greedy-algorithm"-like decision rule. For every record $a_i \in A^{(I)}$ we find the greatest element in row i of the similarity matrix $\mathbf{S}$. The column index of this element gives the record $b_j \in A^{(I)}$, which is linked with a_i. Succinctly,

$$a_i \mapsto b_j \quad \text{where} \quad j = \arg\max_k \mathbf{S}_{ik}.$$

The mapping defined by the maximum similarity assignment is not a bijection because more than one record a_i can be assigned to the same record b_j. For this reason the mapping is also not a surjection because some of the records in $B^{(I)}$ may remain without assignments.

Linear sum assignment. This strategy matches records by solving the following linear program

$$\begin{gathered}\max_{x_{ij}} \sum_{i=1}^{M}\sum_{j=1}^{M} \mathbf{S}_{ij}x_{ij} \quad \text{such that } x_{ij} \in \{0,1\}, \\ \sum_{j=1}^{M} x_{ij} = \sum_{i=1}^{M} x_{ij} = 1\,; \quad i,j = 1,2,\ldots,M.\end{gathered} \tag{17}$$

The unit elements of the solution define the decision rule:

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

Table 2: Error [%] in the solution of the entity matching problem for the extended set-based similarity functions and the tf*idf token weight.

Assignment →		Max. sim. Error [%]		LSAP Error [%]	
ω	$S(\cdot,\cdot)$	Set I	Set II	Set I	Set II
tf*idf	cos	**14**	**13.93**	**15**	**8.20**
tf*idf	J_1	15	13.93	15	6.56
tf*idf	N_1	14	13.93	14	8.20
tf*idf	D_1	15	13.93	15	6.56
tf*idf	J_2	14	13.93	14	6.56
tf*idf	N_2	16	13.93	13	6.56
tf*idf	D_2	14	13.93	14	6.56

Table 3: Error [%] in the solution of the entity matching problem for the extended set-based similarity functions and the idf token weight.

Assignment →		Max. sim. Error [%]		LSAP Error [%]	
ω	$S(\cdot,\cdot)$	Set I	Set II	Set I	Set II
tf*idf	cos	**14**	**13.93**	**15**	**8.20**
idf	J_1	15	13.11	13	6.56
idf	N_1	17	13.11	13	6.56
idf	D_1	15	13.11	13	6.56
idf	J_2	15	13.11	14	4.92
idf	N_2	17	13.11	15	4.92
idf	D_2	15	13.11	14	6.56

The program (17) is Linear Sum Assignment Problem (LSAP) [4, p.74]. The solution of (17) maximizes the "total similarity" of the assignments between the records in $A^{(I)}$ and $B^{(I)}$. The paper [10] is an early example of using (17) for record linkage. For more recent applications to entity matching we refer to [5] or [6]. The Hungarian algorithm [13] is a classical solution method for (17), while the auction algorithm [2] presents a more efficient alternative.

4.1 Discussion of results

We compare the errors in the model entity matching problem when the similarity matrices for Sets I and II are defined using the new similarity functions, and the tf*idf and idf token weights in (5) and (3), respectively. Specifically, we compute $\mathbf{S}$ using the extended Jaccard, Normalized Weighted Intersection and Dice similarities with $p = 1$ and $p = 2$, and the Δ_p similarity with $p = 1, 2, 5$. The standard cosine similarity [14, p.124] with the tf*idf token weight provides the benchmark.

Table 4: Error [%] in the solution of the entity matching problem for Δ_p similarity functions, $p = 1, 2, 5$.

Assignment →		Max. sim. Error [%]		LSAP Error [%]	
ω	$S(\cdot,\cdot)$	Set I	Set II	Set I	Set II
tf*idf	cos	**14**	**13.93**	**15**	**8.20**
tf*idf	Δ_1	17	16.39	14	9.02
tf*idf	Δ_2	30	20.49	14	10.66
tf*idf	Δ_5	33	32.79	26	22.13

Table 2 presents the results with the tf*idf token weight. As expected, the LSAP assignment rule generally performs better than the simple maximum similarity rule. The difference is particularly noticeable for Set II, where the error in the LSAP assignments is reduced by a half, compared with the error in the maximum similarity assignments. Furthermore, with the LSAP assignment the new similarities consistently outperform the cosine similarity for Set II and are slightly better for Set I. With the maximum similarity rule all errors are identical for Set II and the cosine is slightly better for Set I.

Table 3 presents the results with the idf token weight. With this weighting J_1, N_1 and D_1 are equivalent to the set-based similarity functions in [8]. Our first observation is that the idf weighting further decreases the errors of the LSAP assignments for Set II. In addition, the new similarity functions with $p = 2$ perform particularly well in this setting reducing the error to 4.92% for the J_2 and N_2 similarities. Interestingly enough, the errors for the maximal similarity rule are slightly lower for Set II and on the average - slightly higher for Set I.

Finally, Table 4 summarizes the results when the similarity matrix is defined by using the distance based similarities Δ_p and the tf*idf token weight. One important observation is that as p increases, so does the error. One explanation is that as p grows, Δ_p approaches Δ_∞. It is well known that this kind of norm is more sensitive to outliers and so, some degradation of accuracy can be expected.

5. CONCLUSIONS

In this paper we have developed consistent extensions of set-based similarity functions, which bridge set-based models with the Vector Space Model. We also developed a class of ℓ_p distance based similarities. Preliminary studies of the new similarity functions for a model entity matching problem reveal that their performance is comparable to and in some cases exceeds that of the benchmark cosine similarity. These results confirm the utility of the new similarity functions.

6. ACKNOWLEDGMENTS

B. Paskaleva acknowledges support by the Early Career LDRD program at Sandia National Laboratories.

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

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

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

[5] D. Dey, S. Sarkar, and P. De. A probabilistic decision model for entity matching in heterogeneous databases. *Manage. Sci.*, 44(10):1379–1395, Oct. 1998.

[6] 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.

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

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

[9] 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.

[10] M. 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.

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

[12] M.-C. Kim and K.-S. Choi. A comparison of collocation-based similarity measures in query expansion. *Information Processing & Management*, 35(1):19 – 30, 1999.

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

[14] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval.* Cambridge University Press, 2008.

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

[16] S. K. M. Wong and V. 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.