SC spectra: A linear-time soft cardinality approximation for text comparison

Sergio Jiménez[1] and Alexander Gelbukh[2]

[1] Intelligent Systems Research Laboratory (LISI),
Systems and Industrial Engineering Department
National University of Colombia, Bogota, Colombia
sgjimenezv@unal.edu.co
[2] Center for Computing Research (CIC)
National Polytechnic Institute (IPN), Mexico City, Mexico
www.gelbukh.com

Abstract. Soft cardinality (SC) is a softened version of the classical cardinality of set theory. However, given its prohibitive cost of computing (exponential order), an approximation that is quadratic in the number of terms in the text has been proposed in the past. SC Spectra is a new method of approximation in linear time for text strings, which divides text strings into consecutive substrings (i.e., q-grams) of different sizes. Thus, SC in combination with resemblance coefficients allowed the construction of a family of similarity functions for text comparison. These similarity measures have been used in the past to address a problem of entity resolution (name matching) outperforming SoftTFIDF measure. SC spectra method improves the previous results using less time and obtaining better performance. This allows the new method to be used with relatively large documents such as those included in classic information retrieval collections. SC spectra method exceeded SoftTFIDF and cosine tf-idf baselines with an approach that requires no term weighing.

Keywords: approximate text comparison, soft cardinality, soft cardinality spectra, q-grams, ngrams

1 Introduction

Assessment of similarity is the ability to balance both commonalities and differences between two objects to produce a judgment result. People and most animals have this intrinsic ability, making of this an important requirement for artificial intelligence systems. Those systems rarely interact with objects in real life, but they do with their data representations such as texts, images, signals, etc. The exact comparison of any pair of representations is straightforward, but unlike this crisp approach, the approximate comparison has to deal with noise, ambiguity and implicit information, among other issues. Therefore, a challenge for many artificial intelligence systems is that their assessment of the similarity be, to some degree, in accordance with human judgments.

For instance, names are the text representation–sometimes quite complex, cf. [3,2]–most commonly used to refer to objects in real life. Like humans, intelligent systems when referring to names have to deal with misspellings, homonyms, initialisms, aliases, typos, and other issues. This problem has been studied by different scientific communities under different names, including: record linkage [23], entity resolution [12], object identification [22] and (many) others.

The name matching task [4] consists of finding co-referential names in a pair of lists of names, or to find duplicates in a single list. The methods that use pairs of surface representations are known as static methods and usually tackle the problem using a binary similarity function and a decision threshold. On the other hand, adaptive approaches make use of information throughout the list of names. The adaptability of several of these approaches usually relies on the *tf-idf* weighting or similar methods [20].

Comparison methods can also be classified by the level of granularity in which the texts are divided. For example, the family of methods derived from the edit distance [15] use characters as a unit of comparison. The granularity is increased gradually in the methods based on q-grams of characters [13]. Q-grams are consecutive substrings of length q overlapping $q-1$ characters, also known as *kmers* or *ngrams*. Further, methods such as vector space model (VSM) [20] and coefficients of similarity [21] make use of terms (i.e., words or symbols) as sub-division unit. The methods that have achieved the best performance in the entity resolution task (ER) are those that combine term-level comparisons with comparisons at character or q-gram level. Some examples of these hybrid approaches are Monge-Elkan's measure [17,10], SoftTFIDF [8], fuzzy match similarity (FMS) [5], meta-levenshtein (ML) [18] and soft cardinality (SC) [11].

Soft cardinality is a set-based method for comparing objects that softens the crisp counting of elements that makes the classic set cardinality, considering the similarities among elements. For text comparisons, the texts are represented as sets of terms. The definition of SC requires the calculation of 2^m intersections for a set with m terms. Jimenez *et al.* [11] proposed an approach to SC using only m^2 computations of an auxiliary similarity measure that compares two terms.

In this paper, we propose a new method of approximation for SC that unlike the current approach does not require any auxiliary similarity measure. In addition, the new method allows simultaneous comparison of uni-grams (i.e., characters), bi-grams or tri-grams by combining a range of them. We call these combinations SC spectra (soft cardinality spectra). SC spectra can be computed in linear time allowing the use of soft cardinality with large texts and in other intelligent-text-processing applications such as information retrieval. We tested SC spectra with 12 entity resolution data sets and with 9 classic information retrieval collections overcoming baselines and the previous SC approximation.

The remainder of this paper is organized as follows: Section 2 briefly recapitulates the SC method for text comparison. The proposed method is presented in Section 3. In Section 4, the proposed method is experimentally compared with the previous approximation method and other static and adaptive approaches;

a brief discussion is provided. Related work is presented in Section 5. Finally, in Section 6 conclusions are given and future work is briefly discussed.

2 Soft cardinality for text comparison

The cardinality of a set is defined as the number of different elements in itself. When a text is represented as a bag of words, the cardinality of the bag is the size of its vocabulary of terms. Rational cardinality-based similarity measures are binary functions that compare two sets using only the cardinality of each set and - at least - the cardinality of their union or intersection. Examples of these measures are *Jaccard* ($|A \cap B|/|A \cup B)$, *Dice* ($2|A \cap B|/(|A| + |B|)$) and *cosine* ($|A \cap B|/\sqrt{|A||B|}$) coefficients. The effect of the cardinality function in these measures is to count the number of common elements and compressing repeated elements in a single instance. On the basis of an information theoretical definition of similarity proposed by Lin [16], Cilibrasi and Vitányi [7] proposed a compression distance that takes advantage of this feature explicitly showing its usefulness in text applications.

However, the compression provided by classical cardinality is crisp. That is, two identical elements in a set are counted once, but two nearly identical elements count twice. This problem is usually addressed in text applications using *stemming*, but this approach is clearly not appropriate for name matching. Soft cardinality (SC) addresses this issue taking into account the similarities between elements of the set. SC's intuition is as follows: the elements that have similarities with other elements contribute less to the total cardinality than unique elements.

2.1 Soft cardinality definition

The soft cardinality of a set is the cardinality of the union of its elements treated themselves as sets. Thus, for a set $A = \{a_1, a_2, \ldots, a_n\}$, the soft cardinality of A is $|A|^{'}{=}|\bigcup_{i=1}^{n} a_i|$.

Representing text as bag of words, two names such as "Sergio Gonzalo Jiménez" and "Cergio G. Gimenes" can be divided into terms (tokens) and compared using soft cardinality as it is depicted in Fig. 1. Similarities among terms are represented as intersections. The soft cardinality of each set is represented as the area inside of the resulting cloud-border shape. Similarity measures can be obtained using resemblance coefficients, such as *Jaccard*, obtaining: $sim(A, B) = (|A|^{'} + |B|^{'} - |A \cup B|^{'})/|A \cup B|^{'}|$.

2.2 SC approximation with similarity functions

Computing cardinality of the union of n sets requires the addition of $2^n - 1$ numbers. Besides, each one of those values can be the intersection of n sets. For instance, the cardinality of the union of three sets is $|r \cup s \cup t| = |r| + |s| + |t| -$

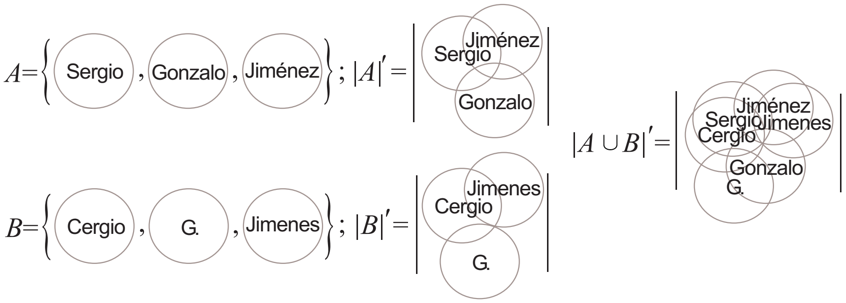

Fig. 1. Example

$|r \cap s| - |s \cap t| - |r \cap t| + |r \cap s \cap t|$. Even for small values of n this computation is not practical.

The soft cardinality can be approximated by using only pairwise comparisons of elements with the following expression:

$$|A|'_{\alpha} \simeq \sum_{i}^{n} \left(\sum_{j}^{n} \alpha(a_i, a_j)^p \right)^{-1} \tag{1}$$

This approximation method makes n^2 calculations of the similarity function $\alpha(*, *)$, which has range $[0, 1]$ and satisfies $\alpha(x, x) = 1$. In our scenario, this function returns the similarity between two terms. In fact, when α is a crisp comparator (i.e., returns 1 when the elements are identical and 0 otherwise) $|A|'_{\alpha}$ becomes $|A|$, i.e., the classical set cardinality. Finally, the exponent p is a tuning parameter investigated by Jimenez *et al.* [11], who obtained good results using $p = 2.0$ in a name-matching task.

3 Computing soft cardinality using sub-strings

The SC approximation shown in (1) is quite general since the function of similarity between the terms α may or may not use the surface representation of both strings. For example, the edit distance is based on a surface representation of characters, in contrast to a semantic relationship function, which can be based on a large corpus or a semantic network. Furthermore, when the surface representation is being used, SC could be calculated by subdividing the text string into substrings and then count the number of different substrings. However, if the unit of the subdivision is q-grams of characters, the resulting similarity measure would ignore the natural subdivision in terms (tokens) of the text string.

Several comparative studies have shown the convenience of the hybrid approaches that first tokenize (split in terms) a text string and then make comparisons between the terms at character or q-gram level [8,4,6,19,11]. Similarly, the definition of SC is based on an initial tokenization and an implicit further subdivision made by the function α to assess similarities and differences between pairs of terms. The intuition behind the new SC approximation is first tokenizing the text. Second, to split each term into a finer-grained substring unit (e.g., bi-grams). Third, to make a list of all the different substrings, and finally, calculate a weighted sum of the sub-strings with weights that depends on the number of substrings in each term.

Consider the following example with the Spanish name "Gonzalo Gonzalez", $A =${"Gonzalo","Gonzalez"}, $a_1 =$"Gonzalo" and $a_2 =$"Gonzalez". Using bi-grams with padding characters[1] as subdivision unit; the pair of terms can be represented as: $a_1^{[2]} =\{\triangleleft G, Go, on, nz, za, al, lo, o\triangleright\}$ and $a_2^{[2]} =\{\triangleleft G, Go, on, nz, za, al, le, ez, z\triangleright\}$. The exponent in square brackets means the size q of the q-gram subdivision. Let $A^{[2]}$ be the set with all different bi-grams $A^{[2]} = a_1^{[2]} \cup a_2^{[2]} =\{\triangleleft G, Go, on, nz, za, al, lo, o\triangleright, le, ez, z\triangleright\}$, $|A^{[2]}| = |a_1^{[2]} \cup a_2^{[2]}| = 11$. Similarly, $|a_1^{[2]} - a_2^{[2]}| = 2$, $|a_2^{[2]} - a_1^{[2]}| = 3$ and $|a_1^{[2]} \cap a_2^{[2]}| = 6$.

Thus, each one of the elements of $A^{[2]}$ adds a contribution to the total soft cardinality of A. The elements of $A^{[2]}$ that also belongs to $a_1^{[2]} - a_2^{[2]}$ or $a_2^{[2]} - a_1^{[2]}$ contributes $1/|a_1^{[2]}| = 0.125$ and $1/|a_2^{[2]}| = 0.11\bar{1}$ respectively; that is the inverse of the number of bi-grams on each term. Common bi-grams between $a_1^{[2]}$ and $a_2^{[2]}$ must contribute with a value in $[0.11\bar{1}, 0.125]$ interval. The most natural choice, given the geometrical metaphor depicted in Fig. 1, is to select the maximum. Finally, soft cardinality for this example is $|A|' \simeq 0.125\times2+0.11\bar{1}\times3+0.125\times6 = 1.33\bar{3}$ in contrast to $|A| = 2$. The soft cardinality of A reflects the fact that a_1 and a_2 are similar.

3.1 Soft cardinality q-spectrum

The SC of a text string can be approximated using a partition $A^{[q]} = \bigcup_{i=1}^{|A|} a_i^{[q]}$ of A in q-grams, where $a_i^{[q]}$ is the partition of i-th term in q-grams. Clearly, each one of the q-grams $A_j^{[q]}$ in $A^{[q]}$ can occur in several terms a_i of A, having indices i satisfying $A_j^{[q]} \in a_i^{[q]}$. The contribution of $A_j^{[q]}$ to the total SC is the maximum of $1/|a_i^{[q]}|$ for each one of its occurrences. The final expression for SC is:

$$|A|'_{[q]} \simeq \sum_{j=1}^{|A^{[q]}|} \max_{i;A_j^{[q]} \in a_i^{[q]}} \left(\frac{1}{|a_i^{[q]}|} \right) . \tag{2}$$

The approximation $|A|'_{[q]}$ obtained with (2) using q-grams is the SC q-spectrum of A.

[1] Padding characters are especial characters padded at the begining and the end of each term before being subdivided in q-grams. These characters allows to distinguish heading and trailing q-grams from those at the middle of the term.

3.2 Soft cardinality spectra

A partition of q-grams allows the construction of similarity measures with its SC q-spectrum associated. The most fine-grained subtring partition is $q = 1$ (i.e., characters) and the coarser is the partition into terms. While partitions such as uni-grams, bi-grams and tri-grams are used in tasks such as entity resolution, the term partition is preferred for information retrieval, text classification and others. Intuitively, finer partitions appear to be suitable for short texts -such as names- and terms seem to be more convenient for documents.

The combination of several contiguous partition granularities can be useful for comparing texts in a particular dataset. Given that each SC q-spectrum provides a measure of the compressed amount of terms in a text, several SC q-spectrum can be averaged or added to get a more meaningful measure. SC spectra is defined as the addition of a range of q-spectrum starting at q_s and ending at q_e, denoted SC spectra $[q_s : q_e]$, having $q_s \leq q_e$. For instance, the SC spectra $[2 : 4]$ uses simultaneously bi-grams, tri-grams and quad-grams to approximate the soft cardinality of a bag of words. Thus, the SC spectra expression is:

$$|A|'_{[q_s:q_e]} = \sum_{i=s}^{e} |A|'_{[q_i]}. \tag{3}$$

4 Experimental Evaluation

The proposed experimental evaluation aims to address the following issues: (i) to determine which of the different substring padding approaches are more suitable for entity resolution (ER) and information retrieval (IR) tasks, (ii) to determine if SC spectra is more convenient than SC q-spectrum, (iii) to compare SC spectra versus the previous SC approximation, (iv) to compare the performance of the proposed similarity measure obtained using SC spectra versus other text measures.

4.1 Experimental Setup

Data sets For experimental evaluation, two groups of data sets were used for entity resolution and information retrieval tasks, respectively. The first group, called ER, consists of twelve data sets for name matching collected from different sources under secondstring framework[2]. The second group, called IR, is composed of nine information retrieval classic collections described by Baeza-Yates and Ribeiro-Neto [1][3]. Each data set is composed of two sets of texts and a gold-standard relation that associates pairs from both sets. The gold-standard in all data sets was obtained from human judgments, excluding *census* and *animal* data sets that were built, respectively, making random edit operations into a list of people names, and using a single list of animal names and considering

[2] http://secondstring.sourceforge.net/

[3] http://people.ischool.berkeley.edu/~hearst/irbook/

as co-referent names pairs who are proper sets at term level. At ER data sets, gold-standard relationship means identity equivalence, and at IR data sets, it means relevance between a query or information need and a document.

Texts in all data sets were divided into terms—i.e., tokenized—with a simple approach using as separator the space character, punctuation, parenthesis and others special characters such as slash, hyphen, currency, tab, etc. Besides, no stop words removal or stemming was used.

Text similarity function The text similarity function used to compare strings was built using a cardinality-based resemblance coefficient replacing classic set cardinality by SC spectra. The used resemblance coefficient was the quotient of the cardinality of intersection divided by the harmonic mean of individual cardinalities:

$$harmonic(A,B) = \frac{|A \cap B| \times (|A| + |B|)}{2 \times |A| \times |B|}. \tag{4}$$

The intersection operation in (4) can be replaced by union using $|A \cap B| = |A| + |B| - |A \cup B|$. Thus, the final text similarity function between two tokenized text strings A and B is given by the following expression:

$$sim(A,B) = 1 + \frac{1}{2}\left(\frac{|A|'_{[q_s:q_e]}}{|B|'_{[q_s:q_e]}} + \frac{|B|'_{[q_s:q_e]}}{|A|'_{[q_s:q_e]}} - \frac{|A \cup B|'_{[q_s:q_e]}}{|A|'_{[q_s:q_e]}} - \frac{|A \cup B|'_{[q_s:q_e]}}{|B|'_{[q_s:q_e]}} \right). \tag{5}$$

Performance Measure The quality of the similarity function proposed in (5) can be quantitatively measured using several performance metrics for ER and IR tasks. We preferred to use interpolated average precision (IAP) because is a performance measure that has been commonly used at both tasks (see [1] for a detailed description). IAP is the area under precision-recall curve interpolated at 11 evenly separated recall points.

Experiments For experiments, 55 similarity functions were constructed with all possible SC spectra using q-spectrum ranging q from 1 to 10 in combination with (5). Each obtained similarity function was evaluated using all text pairs into the entire Cartesian product between both text sets on all 19 data set. Besides, three padding approaches were tested:

single padding to pad one character before and after each token, e.g. the [2:3] spectra sub-division of "sun" is $\{\triangleleft s,\ su,\ un,\ n\triangleright,\ \triangleleft su,\ sun,\ un\triangleright\}$.

full padding to pad $q-1$ characters before and after each token, e.g. the [2:3] spectra sub-division of "sun" is $\{\triangleleft s,\ su,\ un,\ n\triangleright,\ \triangleleft\triangleleft s,\ \triangleleft su,\ sun,\ un\triangleright,\ n\triangleright\triangleright\}$.

no padding e.g.[2:3] spectra for "sun" is $\{su,\ un,\ sun\}$

For each one of the 3135 (55×19×3) experiments carried out interpolated average precision was computed. Fig. 2 shows a results sample for two data sets—*hotels* and *adi*—using *single padding* and *no padding* configurations respectively.

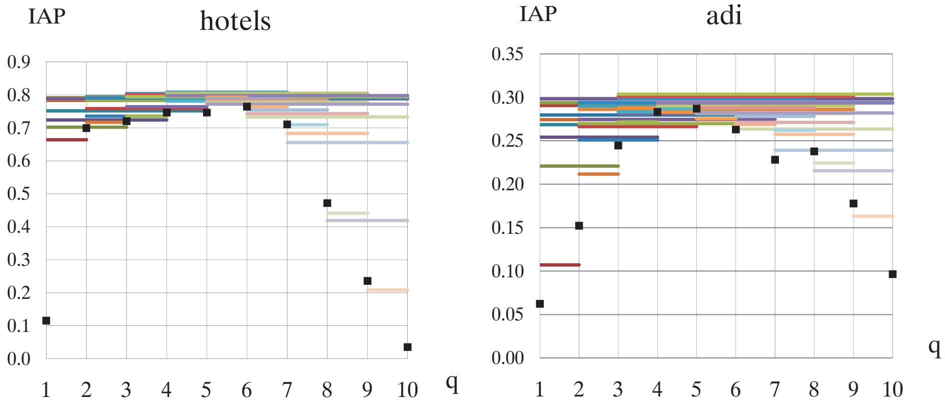

Fig. 2. IAP performance for all SC spectra form $q = 1$ to $q = 10$ for data sets *hotels* and *adi*. Spectra with single q-spectrum are shown as black squares (e.g. [3:3]). Wider spectra are shown as horizontal bars.

4.2 Results

Tables 1 and 2 show the best SC spectra for each data set using the three proposed padding approaches. *Single padding* and *no padding* seems to be more convenient for ER and IR data set groups respectively.

Table 1. Results for best SC spectra using ER data sets

PADDING	full		single		no	
DATA SET	spectra	IAP	spectra	IAP	spectra	IAP
birds-scott1	**[1:2]***	**0.9091**	**[1:2]***	**0.9091**	**[1:2]***	**0.9091**
birds-scott2	[7:8]*	0.9005	**[6:10]**	**0.9027**	[5:9]	0.9007
birds-kunkel	[5:7]*	0.8804	**[6:6]**	**0.8995**	[4:4]	0.8947
birds-nybird	[4:6]	0.7746	**[1:7]**	**0.7850**	[4:5]	0.7528
business	[1:3]	0.7812	**[1:4]**	**0.7879**	[1:4]	0.7846
demos	**[2:2]**	**0.8514**	**[2:2]**	**0.8514**	[1:3]	0.8468
parks	[2:2]	0.8823	[1:9]	0.8879	**[2:4]**	**0.8911**
restaurant	[1:6]	0.9056	**[3:7]**	**0.9074**	**[1:6]**	**0.9074**
ucd-people	**[1:2]***	**0.9091**	**[1:2]***	**0.9091**	**[1:2]***	**0.9091**
animal	[1:10]	0.1186	**[3:8]**	**0.1190**	[3:4]	0.1178
hotels	[3:4]	0.7279	[4:7]	0.8083	**[2:5]**	**0.8147**
census	[2:2]	0.8045	**[1:2]**	**0.8110**	[1:2]	0.7642
best average	[3:3]	0.7801	**[2:3]**	**0.7788**	[1:3]	0.7746
average of best		0.7871		**0.7982**		0.7911

* Asterisks indicate that another wider SC spectra also showed the same IAP performance.

Table 2. Results for best SC spectra using IR collections

PADDING	full		single		no	
DATA SET	spectra	IAP	spectra	IAP	spectra	IAP
cran	**[7:9]**	**0.0070**	[3:4]	0.0064	[3:3]	0.0051
med	[4:5]	0.2939	**[5:7]***	**0.3735**	[4:6]	0.3553
cacm	**[4:5]**	**0.1337**	[2:5]	0.1312	[2:4]	0.1268
cisi	[1:10]	0.1368	[5:8]	0.1544	**[5:5]**	**0.1573**
adi	[3:4]	0.2140	[5:10]	0.2913	**[3:10]**	**0.3037**
lisa	[3:5]	0.1052	[5:8]	0.1244	**[4:6]**	**0.1266**
npl	[7:8]	0.0756	[3:10]	0.1529	**[3:6]**	**0.1547**
time	[1:1]	0.0077	[8:8]	0.0080	**[6:10]**	**0.0091**
cf	[7:9]	0.1574	[5:10]	0.1986	**[4:5]**	**0.2044**
best average	[3:4]	0.1180	**[5:8]**	**0.1563**	[4:5]	0.1542
average of best		0.1257		0.1601		**0.1603**

* Asterisks indicate that another wider SC spectra also showed the same IAP performance.

Fig. 3 shows precision-recall curves for SC spectra in comparison with other measures. The series named *best SC spectra* is the average of the best SC spectra for each data set using *single padding* for ER and *no padding* for IR. *MongeElkan* measure [17] used an internal inter-term similarity function of bi-grams combined with Jaccard coefficient. *SoftTFIDF* used the same configuration proposed by Cohen *et al.* [8] but fixing its normalization problem found by Moreau *et al.* [18]. Soft Cardinality used (1) with $p = 2$ and the same inter-term similarity function used with *MongeElkan* measure.

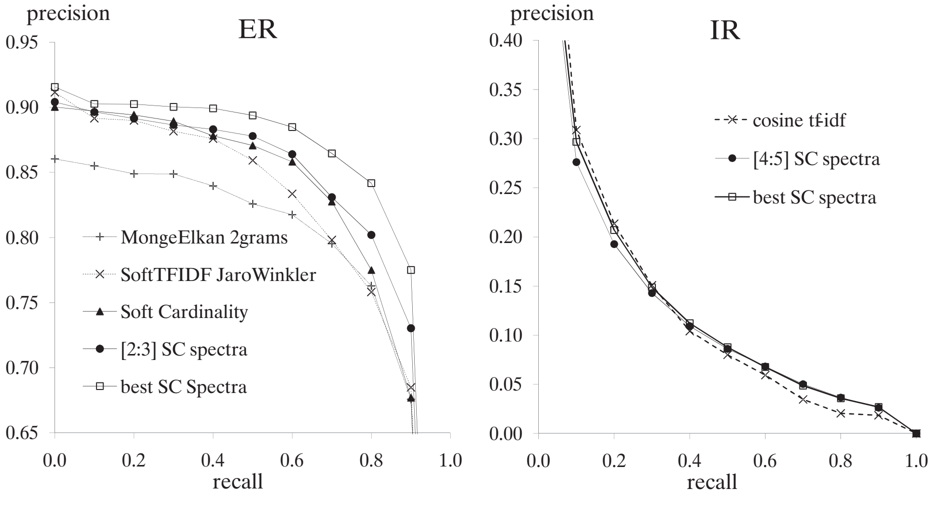

Fig. 3. Precision-recall curves of SC spectra and other measures

4.3 Discussion

Results in Tables 1 and 2 indicate that padding characters seem to be more useful at ER data sets than at IR collections, but using only a single padding character. Apparently, the effect of adding padding characters is important only in collections with relatively short texts such as ER.

Best performing configurations (showed in boldface) were reached—in most of the cases (16 over 19)—using SC spectra instead of single SC q-spectrum. This effect can also be appreciated in Figures 2 (a) and (b), where SC spectra (represented as horizontal bars) tends to outperform SC q-spectrum (represented as small black squares). The relative average improvement of the best SC spectra for each data set versus the best SC q-spectrum was 1.33% for ER data sets and 4.48% for IR collections. Results for best SC q-spectrum were not shown for space limitations. In addition, Fig. 2 qualitatively shows that SC spectra measures tend to perform better than the SC q-spectrum with maximum performance of those that compose a SC spectra. For instance, [7:9] SC spectra at *adi* collection outperforms all SC 7-grams, SC 8-grams and SC 9-grams.

As Fig. 3 clearly shows—for ER data—the similarity measures obtained using the best SC spectra for each data set outperforms the other tested measures. It is important to note that unlike SoftTFIDF, measures obtained using SC spectra are static. That is, they do not use term weighting obtained from term frequencies into the entire data set. Regarding IR, SC spectra reached practically the same performance than *cosine tf-idf*. This result is also remarkable because we are reaching equivalent performance (better at ER data) using considerably less information. Finally, ER results also show that SC spectra is a better soft cardinality approximation than the previous approximation; see (1). Besides, SC spectra require considerably less computational effort than that approximation.

5 Related Work

The proposed weighting schema that gives smaller weights to substrings according to the length in characters of each term is similar to the approach of De La Higuera & Micó, who assigned a variable cost to character edit operations to Levenshtein's edit distance [9]. They obtained improved results in a text classification task using this cost weighting approach. This approach is equivalent to ours because the contribution of each q-gram to the SC depends on the total number of q-grams in the term, which in turn depends on the length in characters of the term.

Leslie *et al.* [14] proposed a k-spectrum kernel for comparing sequences using sub-strings of k-length in a protein classification task. Similarly to them, we use the same metaphor to name our approach.

6 Conclusions and future work

We found that the proposed SC spectra method for text comparison performs particularly well for the entity resolution problem and reach the same results of cosine *tf-idf* similarity using classic information retrieval collections. Unlike several current approaches, SC spectra does not require term weighting. However, as future work, it is interesting to investigate the effect of weighting in SC spectra at term and substring level. Similarly, how to determine the best SC spectra for a particular data set is an open question worth to investigate. Finally, we also found that SC spectra is an approximation for soft cardinality with less computational cost and better performance, allowing the proposed method to be used with longer documents such as those of text information retrieval applications.

Acknowledgements

This research was funded in part by the Systems and Industrial Engineering Department, the Office of Student Wellfare of the National University of Colombia, Bogotá, and throught a grant from the Colombian Department for Science, Technology and Innovation Colciencias, project 110152128465. The second author recognizes the support from Mexican Government (SNI, COFAA-IPN, SIP 20113295, CONACYT 50206-H) and CONACYT–DST India (project "Answer Validation through Textual Entailment").

References

1. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley & ACM Press (1999)
2. Barceló, G., Cendejas, E., Bolshakov, I., Sidorov, G.: Ambigüedad en nombres hispanos. Revista Signos. Estudios de Lingüística 42(70), 153–169 (2009)
3. Barcelo-Alonso, G., Cendejas-Castro, E., Sidorov, G., Bolshakov, I.A.: Formal grammar for hispanic named entities analysis. Lecture Notes in Computer Science 5449, 183–194 (2009)
4. Bilenko, M., Mooney, R., Cohen, W.W., Ravikumar, P., Fienberg, S.: Adaptive name matching in information integration. IEEE Intelligent Systems 18(5), 16–23 (2003), `http://portal.acm.org/citation.cfm?id=1137237.1137369`
5. Chaudhuri, S., Ganjam, K., Ganti, V., Motwani, R.: Robust and efficient fuzzy match for online data cleaning. In: Proceedings of the 2003 ACM SIGMOD international conference on management of data. pp. 313–324. ACM, San Diego, California (2003), `http://portal.acm.org/citation.cfm?id=872757.872796`
6. Christen, P.: A comparison of personal name matching: Techniques and practical issues. In: International Conference on Data Mining Workshops. pp. 290–294. IEEE Computer Society, Los Alamitos, CA, USA (2006)
7. Cilibrasi, R., Vitányi, P.: Clustering by compression. IEEE Transactions on Information Theory pp. 1523–1545 (2005)
8. Cohen, W.W., Ravikumar, P., Fienberg, S.E.: A comparison of string distance metrics for name-matching tasks. In: Proceedings of the IJCAI2003 Workshop on Information Integration on the Web. pp. 73–78 (Aug 2003), `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.15.178`
9. de la Higuera, C., Mico, L.: A contextual normalised edit distance. In: IEEE 24th International Conference on Data Engineering Workshop. pp. 354–361. Cancun, Mexico (2008), `http://portal.acm.org/citation.cfm?id=1547551.1547758`

10. Jimenez, S., Becerra, C., Gelbukh, A., Gonzalez, F.: Generalized mongue-elkan method for approximate text string comparison. In: Computational Linguistics and Intelligent Text Processing, CICLING'09. pp. 559–570. Springer Berlin Heidelberg (2009), http://dx.doi.org/10.1007/978-3-642-00382-0_45
11. Jimenez, S., Gonzalez, F., Gelbukh, A.: Text comparison using soft cardinality. In: String Processing and Information Retrieval, SPIRE'10. vol. 6393, pp. 297–302. Springer Berlin Heidelberg (2010), http://www.springerlink.com/content/x1w783135m36k880/
12. Köpcke, H., Thor, A., Rahm, E.: Evaluation of entity resolution approaches on real-world match problems. In: Proceedings of the 36th International Conference on Very Large Data Bases. Singapore (2010)
13. Kukich, K.: Techniques for automatically correcting words in text. ACM Computing Surveys 24, 377–439 (Dec 1992)
14. Leslie, C., Eskin, E., Noble, W.S.: The spectrum kernel: A string kernel for SVM protein classification. In: Biocomputing 2002 - Proceedings of the Pacific Symposium. pp. 564–575. Kauai, Hawaii, USA (2001), http://eproceedings.worldscinet.com/9789812799623/9789812799623_0053.ht%ml
15. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10(8), 707–710 (1966)
16. Lin, D.: Information-Theoretic definition of similarity. In: Proceedings of the Fifteenth International Conference on Machine Learning. pp. 296–304 (1998), http://portal.acm.org/citation.cfm?id=645527.657297&coll=Portal&dl=GUID%E&CFID=92419400&CFTOKEN=72654004
17. Monge, A.E., Elkan, C.: The field matching problem: Algorithms and applications. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD). pp. 267–270. Portland, OR (Aug 1996)
18. Moreau, E., Yvon, F., Cappé, O.: Robust similarity measures for named entities matching. In: Proceedings of the 22nd International Conference on Computational Linguistics. pp. 593–600 (2008), http://portal.acm.org/citation.cfm?id=1599081.1599156
19. Piskorski, J., Sydow, M.: Usability of string distance metrics for name matching tasks in polish. In: Proceedings of the 3rd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, (LTC-2007), Poznań, Poland, October 5–7, 2007 (2007), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.102.9942
20. Salton, G.: Introduction to modern information retrieval. McGraw-Hill (1983)
21. Sarker, B.R.: The resemblance coefficients in group technology: A survey and comparative study of relational metrics. Computers & Industrial Engineering 30(1), 103–116 (Jan 1996), http://dx.doi.org/10.1016/0360-8352(95)00024-0
22. Tejada, S., Knoblock, C.A.: Learning domain independent string transformation weights for high accuracy object identification. In: Proceedings of International Conference on Knowledge Discovery and Data Mining (SIGKDD) (2002)
23. Winkler, W.E.: The state of record linkage and current research problems. Statistical research divison U.S. Census Bureau (1999), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.39.4336