Text Comparison Using Soft Cardinality

Sergio Jimenez[1], Fabio Gonzalez[1], and Alexander Gelbukh[2]

[1] National University of Colombia
[2] CIC-IPN Mexico
{sgjimenezv,fagonzalezo}@unal.edu.co
www.gelbukh.com

Abstract. The classical set theory provides a method for comparing objects using cardinality and intersection, in combination with well-known resemblance coefficients such as Dice, Jaccard, and cosine. However, set operations are intrinsically crisp: they do not take into account similarities between elements. We propose a new general-purpose method for comparison of objects using a soft cardinality function that show that the soft cardinality method is superior via an auxiliary affinity (similarity) measure. Our experiments with 12 text matching datasets suggest that the soft cardinality method is superior to known approximate string comparison methods in text comparison task.

1 Introduction

Things are usually not either completely equal or completely different. More often than not we need to decide which objects are more similar than others. For example, in information retrieval task we need to find documents most similar to the user query, while none of these documents is exactly equal to it. While the task of exact comparison is well-defined and the corresponding methods are clear and well understood, approximate comparison is a highly heuristic task for which a great variety of methods have been suggested, each one good for some problems and none good for all—which means that the quest for better and more general approximate comparison paradigms is in full swing.

In this paper we propose a new approximate object comparison method based on soft cardinality. While stemming from fundamental elements from the classic set theory, it uses a new cardinality function that allows for flexibility. Despite the simplicity of our method, preliminary results obtained in a text matching task are encouraging as compared with much more elaborated state-of-the-art approximate string matching techniques.

The paper is organized as follows. Section 2 briefly describes some most closely related approaches and compares our method with them. Section 3 presents the notion of soft cardinality and introduces our method. Section 4 gives the experimental results. Finally, Section 5 concludes the paper.

2 Related Work

Binary similarity measures use different strategies to assess commonalities and differences of objects under comparison. Probably the most popular approximate

E. Chavez and S. Lonardi (Eds.): SPIRE 2010, LNCS 6393, pp. 297–302, 2010.

comparison technique uses the well-known resemblance coefficients [1]. Although resemblance coefficients reflect the degree of similarity, they are based on crisp operations such as set intersection, which do not consider degree of similarity between the elements of the sets. This is a drawback in scenarios with uncertainty. Our proposed method addresses the problem of the inability to manage uncertainty by resemblance coefficients: instead, it exploits the redundancy in sets.

Unlike fuzzy sets [2], our approach does not consider uncertainty in the membership of a particular element. It, however, does handle uncertainty as to the contribution of an element to the cardinality of the set if the element presents redundancy with respect to other elements of the same set.

Classic set theory provides an intuitive mechanism for object comparison using set operations such as intersection $\cap$ and union $\cup$, as well as a function $|*| \in \mathbb{N}$ called cardinality for counting the number of different elements in a set. Using only those components, it is possible to compute similarity between two sets A and B using a family of resemblance coefficients [1] such as:

$$Jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|}, \quad cosine(A,B) = \frac{|A \cap B|}{\sqrt{|A| \times |B|}}, \tag{1}$$

as well as a host of other resemblance coefficients widely used in biosciences, economics, social sciences, engineering and computer science, among other fields. All those expressions return 1 if A and B are equal, 0 if A and B have no elements in common, and otherwise a number between 0 and 1.

3 Soft Cardinality

The classic set cardinality function counts the elements in a set in a crisp manner: if an element is duplicated, it is only considered once. However, if two elements in a set are very similar—nearly duplicates but not exactly—our intuition is that together they should contribute less to the total cardinality than a pair of completely different elements.

Consider the set S of animals in Fig. 1 (a). The classic crisp cardinality function reports 3 different animals in S. Nevertheless, if an affinity function between the elements of the set is used, Fig. 1 (b) shows a better view of the situation. A *soft cardinality* function that reflects this intuition should produce a value less than three but greater than two. Even though the elements of the set shown in Fig. 1 are not sets themselves, the affinity between *tiger* and *lion* induces some type of intersection in terms of the total soft cardinality of the set.

We define a binary affinity function $\alpha(*,*)$ that reflects the similarity between two elements a and b in the set. Obviously,

$$\begin{aligned} &\alpha(a,b) \in [0,1]; \quad &&\alpha(a,b) = \alpha(b,a); \quad &&\alpha(a,a) = 1; \\ &\delta(a,b) = 1 - \alpha(a,b); \quad &&\delta(a,c) \leq \delta(a,b) + \delta(b,c), \end{aligned} \tag{2}$$

where δ is the distance.

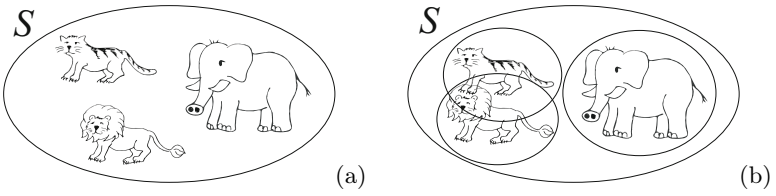

Fig. 1. A set with three elements

Consider Fig. 1 (*b*). If the elements of S' are treated as sets, the affinity function α can be considered as an estimation of the intersection between the elements in the set. Assume an affinity value $\alpha(tiger, lion) = 0.7$, the total soft cardinality of the set can be defined as the crisp cardinality $|S| = 3$ minus the overlap between *tiger* and *lion* (which is 0.7), which gives $|S'|_\alpha = 2.3$.

3.1 Estimating Cardinality of Set Union Using Pairwise Intersections

The affinity binary function α has been proposed as an approximation of the cardinality of the intersection of two elements in a set. Although a and b are not sets themselves (they are just elements), the previous assumption allows us to treat them as sets. If we want to calculate the soft cardinality of a set S with only two elements a and b, it can be calculated using (2) properties and recalling $|A \cup B| = |A| + |B| - |A \cap B|$:

$$|S|_\alpha = \alpha(a, a) + \alpha(b, b) - \alpha(a, b) = 2 - \alpha(a, b),$$

where $|S|_\alpha$ stands for the soft cardinality of the set S given the affinity function α. Similarly, the case of a set S with three elements a, b, c can be treated using the following classic set theory expression:

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |B \cap C| - |A \cap C| + |A \cap B \cap C|\,.$$

However, to estimate $|A \cap B \cap C|$, a ternary affinity function $\alpha(*, *, *)$ is needed. In the general case, the soft cardinality of a set S with elements $s_1, s_2, \ldots, s_n$ requires k-ary affinity functions, $k = 1, \ldots, n$. Nevertheless, the most common affinity (similarity) and distance functions are binary, making the construction of such high-arity functions a problem. We propose to estimate $|S|_\alpha$ using only a binary α function.

Consider a matrix $\mathbf{A}$ of pairwise affinity $\alpha(s_i, s_j)$ between the elements of S. It is symmetric and has all 1's in the diagonal. We will construct a function of this matrix that gives n if $\mathbf{A} = \mathbf{I}_n$ ($n \times n$ identity matrix) and 1 if all entries of $\mathbf{A}$ are 1's. The former case corresponds to a set S in which all its elements are completely different (i.e., classical crisp cardinality), while the later case corresponds to a set S in which all elements are identical. On any other symmetric matrix the function will give a real number in the interval $(1, n)$ that approximates the cardinality of $\bigcup_{i=1}^{n} s_i$.

Consider the following expression that satisfies the previous requirements of reproducibility of the classic crisp cardinality:

$$|S|_\alpha = \sum_{i=1}^{n} \frac{1}{\sum_{j=1}^{n} \alpha(s_i, s_j)^p}, \quad p \geq 0 \tag{3}$$

where p is an adjustable parameter. Using $p = 1$ in this expression with our example, a result of $|S|_\alpha = 2.18$ is obtained. This result is close to the value 2.3 obtained previously out of intuitive considerations.

3.2 Decoupling and Combining Affinity and Importance

Until now, we have assumed an inherent individual cardinality of 1 for each element. This makes sense given that the classic crisp cardinality increases by 1 with each different element in a set. This is equivalent to considering elements as sets represented as circles of area equal to 1 in a Venn diagram with among them. However, often it is reasonable to consider elements represented by circles with different areas. For instance, if the elements of the set are words, their discrimination power can be encoded by the area of the circles: stop words coded by small circles and rare words by large circles. This approach allows us to treat separately (i) the affinity between elements and (ii) the intrinsic importance or weight of the element.

In order to integrate importance or "element cardinality" to the soft cardinality method, consider the role of the inner term $(\sum_{j=1} \alpha(s_i, s_j)^p)^{-1}$ in (3). This term represents the contribution of the element s_i to the total soft cardinality of S and thus can be weighted with a factor w_i to reflect the importance ("cardinality") of the element s_i:

$$|S|_\alpha^p = \sum_{i=1}^{n} \frac{w_i}{\sum_{j=1}^{n} \alpha(s_i, s_j)^p} \tag{4}$$

Vector space model represents documents as vectors whose coordinates are dictionary words. Generally, the values of the coordinates are weights associated with the discrimination power of the words such as *tf-idf* [3]: the importance of the word for the document. However, this model assumes independence between index terms (i.e. dimensions), making it impossible to take into account their relatedness. Unlike vector space model, (4) provides the soft cardinality method a mechanism to keep affinity (correlation) and importance (weight) decoupled but naturally combined.

4 Preliminary Results

The aim of our experiments is to explore the utility and potential of the soft cardinality method in a particular text comparison task. The name-matching task consists in comparing two strings and to decide if the strings refer to the same entity or not. For instance, "King Rail Rallus elegans" and "Rail: King (Rale élégant) Rallus elegans" in the *Birds-Scott2* dataset refer to the same bird. All

possible string pairs are compared with a similarity measure and a threshold θ selects matched pairs.

Cohen *et al.* [4] carried out a comparative study of several similarity measures with twelve name matching data sets.They showed that their SoftTF-IDF measure, on average, outperforms all other measures. We compared our soft cardinality method using the same data sets that they used.

4.1 Experimental Setup

The matching problem between two set of strings can be viewed as a classification problem over the Cartesian product of the sets. As performance measure for each experiment (i.e. pairs dataset-similarity measure) we used *interpolated average precision* (IAP) and F_1-score, which are commonly used in information retrieval [5].

We carried out experiments with more than 148 well-known string similarity measures, see [6] for details. For reference purposes, the results for the best performing character level measure (i.e. bigrams) and the best token level measure (i.e. cosine) are reported. The reported similarity measures are described as follow:

bigrams. Bigrams intersection computed as the quotient of the common bigrams between the two strings and the number of bigrams in the longer string.

cosine. Both strings are tokenized (separated into words) and compared using cosine coefficient.

SC1. Both strings are tokenized. Soft cardinality (3) with $p = 1$ is used to compute the cardinality of each set of tokens and the cardinality of the union of both set of tokens. Cardinalities are combined with cosine coefficient (1). The used auxiliary inter-token affinity function α was bigrams.

SC2. It is the same method as SC1, but using $p = 2$ in (3).

SC.3 It is the same method as SC1, but using $p = 2$ and $w_i = idf_i$ in (4).

STI. SoftTF-IDF proposed by Cohen *et al.* [4] SoftTF-IDF is a fuzzified extension to the well-known *tf-idf* vector space model metric.

4.2 Results

Similarly to Cohen *et al.*, we also noticed that STI (SoftTF-IDF) outperforms on average practically all known text similarity measures. As shown in Table 1, the proposed soft cardinality methods SC2 and SC3 slightly outperformed STI, and SC1 reached a close performance. Both average IAP and F_1-score, reported SC3 and SC2 as the best performing measures for the data sets. Note that unlike STI, SC1 is a basic soft cardinality measure without any parameters to be adjusted. Moreover, SC1 is a static measure, that is, it only uses information in the pair of strings being compared. The SC2 measure is also static, with $p = 2$, see (3), reach the same performance of a STI, which also uses the entire corpus as input. In addition, SC3 performs better than STI using the same *idf* coefficients.

Table 1. IAP results for experiments

Data set	bigrams	cosine	SC1	SC2	SC3	STI
Birds-Scott1	0.848	**0.890**	0.877	0.883	0.883	0.879
Birds-Scott2	0.789	**0.907**	0.895	0.905	0.905	0.897
Birds-Kunkel	0.512	0.857	0.723	0.837	0.868	**0.875**
Birds-Nybird	0.715	0.723	0.734	**0.745**	0.740	0.743
Business	0.529	0.533	0.685	0.732	**0.776**	0.704
Game-Demos	0.722	0.754	0.776	0.784	**0.811**	0.798
Parks	0.881	0.846	0.868	0.859	0.890	**0.897**
Restaurants	0.860	**0.906**	0.899	**0.906**	0.903	0.904
UCD-people	0.797	**0.909**	0.889	**0.909**	**0.909**	**0.909**
Animals	0.066	0.117	0.104	0.116	**0.119**	0.118
Hotels	0.610	**0.722**	0.613	**0.722**	0.697	0.671
Census	0.806	0.412	**0.818**	0.768	0.806	0.726
Average IAP	0.678	0.715	0.740	0.764	**0.776**	0.760
Average F_1-score	0.717	0.779	0.776	0.809	**0.827**	0.808

5 Conclusions

We have proposed a new method for object comparison based on classic set theory, and similarity measure used as a metaphor of elements intersection, which we called soft cardinality method. We have explored the modeling ability of the new method with the case study of text comparison applications.

In particular, soft cardinality allows a nice combination of affinity (similarity) between elements and their importance (weights). This property is useful in text applications where approximate string matching measures can be used as affinity function in combination with practical term weighting schemas such as *tf-idf*. Experimental results showed that our approach gives better results than state-of-the-art approximate string comparison methods in a name matching task.

References

1. De Baets, B., De Meyer, H.: Transitivity-preserving fuzzification schemes for cardinality-based similarity measures. European Journal of Operational Research 160, 726–740 (2005)
2. Zadeh, L.: Fuzzy Sets, Fuzzy Logic and Fuzzy Systems. World Scientific, Singapore (1996)
3. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill Book Co., New York (1983)
4. Cohen, W.W., Ravikumar, P., Fienberg, S.E.: A comparison of string distance metrics for name-matching tasks. In: Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web (2003)
5. Baeza-Yates, R., Ribero-Neto, B.: Modern Information Retrieval. Addison Wesley/ACM Press (1999)
6. Jimenez, S.: A knowledge-based information extraction prototype for data-rich documents in the information technology domain. Master's thesis, National University of Colombia (2008)