# On the Comparison of Structured Data

Jyrko Correa-Morris[1] and Noslen Hernández[2]

[1] Institute for Pure and Applied Mathematics (IMPA), RJ - Brazil
[2] Advanced Technologies Application Center (CENATAV) - Cuba
jyrkoc@gmail.com, nhernandez@cenatav.co.cu

**Abstract.** This paper introduces a theoretical framework to characterize measures on structured data. We firstly describe the lattice of structured data. Then, four basic and intuitive properties which any measure on structure data must fulfill are formally introduced. Metrics and kernel functions are studied as particular cases of (dis)similarity measures. In the case of metrics we prove that the well-known edit distances meet all the desirable properties. We also give sufficient conditions for a kernel function to satisfy those properties. Some examples are given for particular kinds of structured data.

**Keywords:** structured data, kernel functions, metrics.

## 1 Introduction

Structured data can be found everywhere. These sophisticated data provide the technical machinery for modeling problems in which, besides observations and measurements, we are interested in representing relations. For this reason, they are very useful in many fields of pattern recognition, artificial intelligence and computer vision. Some examples are the applications they found in biometrics [1–4], image processing and image segmentation [5], clustering [6], object detection, information retrieval, document analysis [7], among others.

In most of the aforecited applications, to compare objects represented by some kind of structured data is a primary task. Many measures have been proposed for comparing these data [8–13]. To decide what measure can be more appropriate for a particular problem is a challenge. A measure is good in as much as it solves the problem at hand. However, there are intuitive properties that any good measure should fulfills independently of the problem being solved. That's why this work does not focus on introducing a new criterion to comparing structured data, but on providing some theoretical elements that allow characterizing what a competent measure is, as well as a better understanding of the various classes of existing measures for structured data.

In order to characterize a good measure for comparing structured data we must first know in detail the space in which such data lie and then to formulate which properties would be desirable for such a measure. With this as a guiding philosophy, this paper introduces four simple intuitive properties which any measure on structured data must meet. These properties are based on the lattice

L. Alvarez et al. (Eds.): CIARP 2012, LNCS 7441, pp. 479–486, 2012.

structure of all spaces of structured data (e.g., the space of strings, the space of graphs, etc.). Recall that a lattice is a pair $(L, \preceq)$ formed by a set $L$ and an order relation $\preceq$ (i.e., a reflexive, antisymmetric, and transitive binary relation) such that any pair of elements in $L$ has a supremum (or joint) and an infimum (or meet), both in $L$. For any space of structured data we count with a natural order relation which is closely connected to the intrinsic notion of (dis)similarity, say "is a substructure of". Several existing measures for comparing structured data are studied on the basis of the introduced properties.

The rest of the paper is organized as follows. Section 2 briefly describes the lattice of structured data. In section 3, the desirable properties for measures on structured data are introduced. Section 4 studies metrics and kernels as special kind of measures. Finally, some conclusions are drawn in Section 5.

## 2 The Lattice of Structured Data

Structured data represents relations between objects in a determined environment. For example, given an alphabet $A$ we have the set $\mathcal{S}_A$ of all possible strings whose characters lie in $A$; given a finite set $X$ we have the set $\mathcal{G}_X$ of all possible graphs with vertices in $X$, and the set $\mathcal{P}_X$ of all possible partitions of $X$. When we refer to a set of structured data without taking into account the explicit form of its element, we write $\mathrm{S}_X$. We frequently use this notation to state results which are valid both for strings, graphs, partitions, and so on.

$\mathrm{S}_X$ is naturally endowed with an order structure: we say that $s \in \mathrm{S}_X$ is less or equal than $s' \in \mathrm{S}_X$ if $s$ is a substructure of $s'$, in notation $s \preceq s'$. This order relation takes a specific form in each of the aforementioned sets: in $\mathcal{S}_A$, $s \preceq s'$ if $s$ is a substring of $s'$, while in $\mathcal{G}_X$ (resp. $\mathcal{P}_X$), $s \preceq s'$ if $s$ is a subgraph of $s'$ (resp. $s$ refines $s'$). Notice that given any two structures $s$ and $s'$, not necessarily one of them is a substructure of the other. If this is the case, we say that $s$ and $s'$ are not *comparable*.

The relation $\preceq$ on $\mathrm{S}_X$ has another important property: given two any structures $s$ and $s'$ it is always possible to find a structure $s \wedge s'$ which is both a substructure of $s$ and $s'$; and if any other structure $s''$ has this property then, $s''$ is a substructure of $s \wedge s'$. This structure $s \wedge s'$ is called the *meet* of $s$ and $s'$. Note that $s \wedge s'$ keeps what is common to both structures.

## 3 Structural Properties

Let $\Gamma$ be a dissimilarity measure on $\mathrm{S}_X$, we will require that $\Gamma$ satisfy the following properties:

*Property 1.* (Symmetry) $\Gamma$ is symmetric.

This property is a standard requirement for any measure responsible for the comparison of the dissimilarity between two objects belonging to a data set. It is based on the simple fact that the likeness among objects does not depend on the order in which they are selected.

*Property 2.* (Collinear monotonicity) If $s \preceq s' \preceq s''$ then, $\Gamma(s, s') \leq \Gamma(s, s'')$.

What is the intuition behind this property? If we have a structure $s$ and we gradually transform $s$ into $\widehat{s}$ by adding more structure on it, then the dissimilarity between $s$ and $\widehat{s}$ increases as the structure $\widehat{s}$ grows. Thinking of $s'$ and $s''$ as being two instances of $\widehat{s}$, $s'$ with less structure than $s''$, the property arises.

*Property 3.* (Dual collinear monotonicity) If $s \preceq s' \preceq s''$ then, $\Gamma(s'', s) \leq \Gamma(s'', s')$.

This property is analogous to the previous. The difference is that in this case we transform the structure $s''$ by removing structure. Then, the dissimilarity with respect to $s''$ increases as the structure declines.

*Property 4.* (Meet predominance) For all $s, s' \in \mathrm{S}_X$, $\Gamma(s, s \wedge s') \leq \Gamma(s, s')$.

This property is based on the principle that the dissimilarity between two objects is determined by the things they have in common. The characteristics that concern only one of them just make the difference. Since $s \wedge s'$ only includes what is common to $s$ and $s'$, the property is natural.

The important point to note here is that these properties furnish some guidelines for the proper performance of a measure to compare structured data. They formalize the fact that the "is a substructure of" relation induces an organization of the elements in $\mathrm{S}_X$ which is compatible with the most elementary notion of dissimilarity. Hereafter, we will refer to these properties as $P1, P2, P3$ and $P4$, respectively.

Another property that could also be included in the above group is collinear additivity: if $s \preceq s' \preceq s''$ then $\Gamma(s, s'') = \Gamma(s, s') + \Gamma(s', s'')$. We instead enunciate $P2$ and $P3$ which capture the essence of collinear additivity, without requiring additivity because this condition can be in general very restrictive. Nevertheless, whenever appropriate, reference to it will be made.

Note also that the property perhaps more intuitive and basic of any dissimilarity measure was not included among the introduced properties. This is because it is a consequence of $P1 - P4$, as it is shown in the following Lemma.

**Lemma 1.** *Let $\Gamma$ be a dissimilarity measure on $\mathrm{S}_X$ that satisfies $P1-P4$. Then, for all $s, s' \in \mathrm{S}_X$, $\Gamma(s, s) \leq \Gamma(s, s')$. That is, no object is more similar to a given object $s$ than the same $s$.*

*Proof.* Indeed, if $s \preceq s'$ then the result follows clearly from $P2$. If $s' \preceq s$ then the result is immediate from $P3$. Now, if $s$ and $s'$ are not comparable, then by applying $P4$ we have that $\Gamma(s, s \wedge s') \leq \Gamma(s, s')$. Since $s \wedge s' \preceq s$, this case is reduced to that previously analyzed. We thus get $\Gamma(s, s) \leq \Gamma(s, s \wedge s') \leq \Gamma(s, s')$, which completes the proof.

Although those properties were only introduced for dissimilarity measures, all of them have an analogous to similarity measures. It is easily seen that if we reverse the order of the inequality in $P2 - P4$, then such analogous properties are obtained.

## 4 Metrics and Kernels for Structured Data

In this section some of the existing measures for structured data are studied in terms of the introduced properties. Given that metrics and kernels are honorable representatives of the classes of dissimilarity and similarity measures, respectively; we studied them separately.

### 4.1 Metrics

Among the most flexible measures for comparing structured data are the edit distances. They have as a tenet to compare structured data by counting the structural distortions (also referred as edit transformations) needed to obtain one structure from another. By structural distortion we mean a change in the structure of the datum (e.g., to insert a character in a string, to insert a vertex in a graph, etc.). Each edit distance has its proper set $\tau$ of edit transformations allowed to be performed. The next theorem shows that an edit distance with all its edit transformation being invertible, meets all of the properties aforeintroduced.

**Theorem 1.** *Let $\tau$ be a set of invertible edit transformations. Let $\Gamma$ be the edit distance so that for all $s, s' \in \mathrm{S}_X$, $\Gamma(s, s')$ is the minimum number of $\tau$-distortions needed to transform $s$ into $s'$. Then, $\Gamma$ satisfies all of the properties $P1 - P4$. In addition, $\Gamma$ is a metric.*

*Proof.* The symmetry of $\Gamma$ follows immediately from the fact that all edit transformations in $\tau$ are invertible. Thus, if for transforming $s$ into $s'$, we need to apply $\tau_i \in \tau$ $n_i$ times; then by applying the inverse transformation $\tau_i^{-1}$ also $n_i$ times, we transform $s'$ into $s$. This gives $P1$.
Consider now $s \preceq s' \preceq s''$. If the edit transformation $\tau_i \in \tau$ is required to transform $s$ into $s'$, then it is also needed in the process of transforming $s$ into $s''$. This is because edit transformations are those changes to be performed with the purpose of transforming one structure into another, so since the structure of $s$ is contained in $s'$ and the structure $s'$ is contained in the structure of $s''$, to pass from $s$ to $s'$ is part of the process of passing from $s$ to $s''$. This proves $P2$. The same argument shows that $P3$ holds.

In order to prove $P4$ it suffices to note that when we transform $s$ into $s'$ we firstly need to remove from $s$ everything that is not in $s''$. Thereafter it only remains in $s$ what is common to $s$ and $s'$. This is just $s \wedge s'$. We thus get P4.

Finally, since $\Gamma$ is a non-negative and symmetric function, what is left is to show that (1) $\Gamma(s, s') = 0$ if and only if $s = s'$, and (2) $\Gamma$ satisfies the triangular inequality. (1) is immediate from the fact that we do not need to perform any edit transformations if and only if $s = s'$. (2) is a consequence of the fact that when we transform $s$ into $s''$ and after $s''$ into $s'$, then we obtain $s'$ from $s$. This process requires $\Gamma(s, s'') + \Gamma(s'', s')$ edit transformations to be done. Because $\Gamma(s, s')$ is the minimum number of edit transformations needed to obtain $s'$ from $s$, we can conclude that $\Gamma(s, s') \leq \Gamma(s, s'') + \Gamma(s'', s')$. This completes the proof.

Note that in this case we can say more: any edit measure satisfying the hypotheses of Theorem 1 meets the collinear additivity property.

Let see now applications of Theorem 1 to some particular examples.

*Example 1.* (An edit distance for strings)
Let $\Gamma_s$ be the edit distance in $\mathcal{S}_A$ whose set of edit transformations is $\tau_s = \{\tau_1, \tau_2\}$, where $\tau_1$ is to insert a character, and $\tau_2$ is to remove a character. Since $\tau_1$ and $\tau_2$ are inverses of each other, Theorem 1 has the following corollary.

**Corollary 1.** *$\Gamma_s$ satisfies $P_1 - P_4$.*

*Example 2.* (An edit distance for graphs)
In the case of graphs, we consider the traditional edit distance $\Gamma_g$ which has edit transformations: $\tau_1$: insert a vertex, $\tau_2$: remove a vertex, $\tau_3$: insert an edge, and $\tau_4$: remove an edge. Because $\tau_1 = \tau_2^{-1}$ and $\tau_3 = \tau_4^{-1}$, we have the following corollary of Theorem 1.

**Corollary 2.** *$\Gamma_g$ fulfills properties $P_1 - P_4$.*

*Example 3.* (An edit distance for partitions)
Mirkin measure $\Gamma_p$ is the most known example of edit distance for partitions. Its edit transformations are: $\tau_1$: insert a pair, and $\tau_2$: remove a pair. Again we have $\tau_1 = \tau_2^{-1}$ and therefore, the following corollary.

**Corollary 3.** *$\Gamma_p$ fulfills properties $P_1 - P_4$.*

To some extent, these results were expected, because it is known that edit distances are fairly good measures and, as was pointed out before, the required properties are very basics for structured data. For the particular case of partitions, an analogous study was done for Variation of Information metric, Dogen metric and the Classification Error metric [14].

### 4.2 Kernels

Let us study in this section under what conditions kernel functions for structured data satisfy properties $P1 - P4$.

**Theorem 2.** *Let $k$ be a kernel function on $\mathrm{S}_X$ and $\varphi$ its feature map (i.e., $k(s, s') = \langle \varphi(s), \varphi(s') \rangle = \sum_i \varphi(s)^i \varphi(s')^i$). If the following conditions hold:*

**C1** *For all $s \in \mathrm{S}_X$, $\varphi(s)$ has all of its components non-negative;*
**C2** *Let $s \prec s'$. If $\varphi(s)^i > 0$ then $\varphi(s)^i \geq \varphi(s')^i > 0$;*
**C3** *If $s \prec s' \prec s''$ then,*

$$\sum_i \varphi(s'')^i \varphi(s)^i - \sum_{i, \varphi(s)^i \neq 0} \varphi(s'')^i \varphi(s')^i < \sum_{j, \varphi(s)^j = 0} \varphi(s'')^j \varphi(s')^j;$$

**C4** *If $\varphi(s)^i$ and $\varphi(s')^i$ are simultaneously no-null, then also is $\varphi(s \wedge s')^i$;*
*then, $k$ satisfies properties $P1 - P4$.*

*Proof.* The proof is a simple verification of the properties $P1 - P4$. Symmetry (i.e., $P1$) follows immediately from the fact that $k$ is kernel.

Now, if $s \preceq s' \preceq s''$ we have that

$$k(s, s') = \sum_i \varphi(s)^i \varphi(s')^i \geq \sum_i \varphi(s)^i \varphi(s'')^i = k(s, s''),$$

because if $\varphi(s'')^i \neq 0$, then $\varphi(s')^i \neq 0$, and condition $C2$ assures that $\varphi(s')^i > \varphi(s'')^i$. This shows that $k$ satisfies $P2$.

Using condition $C3$ we obtain

$$k(s'', s') = \sum_{i, \varphi(s)^i \neq 0} \varphi(s'')^i \varphi(s')^i + \sum_{j, \varphi(s)^j = 0} \varphi(s'')^j \varphi(s')^j > \sum_i \varphi(s'')^i \varphi(s)^i,$$

and as the last term equals $k(s'', s)$, we thus get $P3$.

It remains to prove that $k$ satisfies $P4$. Since $k(s, s \wedge s') = \sum_i \varphi(s)^i \varphi(s \wedge s')^i$ and $k(s, s') = \sum_i \varphi(s)^i \varphi(s')^i$, we can use condition $C4$ for getting that $k(s, s \wedge s')$ has more no-null terms than $k(s, s')$. Moreover, by virtue of condition $C2$, we also have that $\varphi(s \wedge s')^i > \varphi(s')^i$. Therefore, we can conclude $k(s, s \wedge s') > k(s, s')$.

As the reader may have noticed, the previous theorem has an analogous one which is a consequence of simple variations on conditions $C2$ and $C3$.

**Theorem 3.** *If in addition of conditions $C1$ and $C4$, the following statements hold:*

**C'2** *Let $s \prec s'$. If $\varphi(s')^i > 0$ then $\varphi(s')^i \geq \varphi(s)^i > 0$.*
**C'3** *If $s \prec s' \prec s''$ then,*

$$\sum_i \varphi(s)^i \varphi(s'')^i - \sum_{i, \varphi(s'')^i \neq 0} \varphi(s)^i \varphi(s')^i < \sum_{j, \varphi(s'')^j = 0} \varphi(s)^j \varphi(s')^j.$$

*Then $k$ satisfies properties $P1 - P4$.*

*Example 4.* (Kernel for partitions)
Perhaps the simplest example of a kernel that satisfies Theorem 2 is given by the set significance based kernels for partitions (see [12]). The significance $\mu(A/P)$ of subset $A \subseteq X$ with respect to a partition $P$ of $X$ is defined as $\frac{|A|}{|C|}$, provided that there exists a cluster $C$ of $P$ containing $A$, and 0 otherwise. Putting all subsets of $X$ in a determined order $A_1, A_2, \ldots, A_{2^n}$, $n$ the number of objects in $X$, we can assign the vector $V_P$ whose $i^{th}$ component is $\mu(A_i/P)$ to the partition $P$. Thus, $k(P, P') = \langle V_P, V_{P'} \rangle$.

The fulfillment of conditions $C1 - C4$ of Theorem 2 is a consequence of the following fact: if $P \preceq \mathcal{P}'$, then the clusters of $\mathcal{P}$ are smaller than the clusters of $\mathcal{P}'$ and therefore, the component associated to the subset $A_i$ in $V_P$ (if non-null) is greater than its analog in $V_{P'}$. However, as there are subsets $A_i's$ contained in a cluster of $P'$, but not in a cluster of $P$, $V_{P'}$ has more non-null components than $V_P$. We thus get the following corollary.

**Corollary 4.** *The set significance based kernel satisfies P1-P4.*

*Example 5.* (Kernel for graphs. A counterexample)
An example of a kernel that has an arbitrary behavior with respect to properties $P1-P4$ is the selection prototypes based kernels [10]. These kernels use the edit distance (see Section 2) to compare each graph $g$ with a set $T = \{t_1, t_2, \ldots, t_m\}$ of prototypes, and thus, to associate $g$ with the vector $V_g$ whose $i^{th}$ component is $\Gamma_g(g, t_i)$. The similarity between two graphs $g$ and $g'$ is then computed as the inner product between $V_g$ and $V_{g'}$. We shall evince that these kernels, although defined from the edit distance which is a measure as good and flexible as expected, do not meet any of the properties $P2-P4$, whatever the prototypes.

Let $g \prec g'$. If for obtaining $g$ from a prototype $t_i$ we need to add a vertex or an edge to $t_i$, then this vertex or edge is also needed to be added when obtaining $g'$ from $t_i$. We can not make the same claim in the case where a vertex or an edge needs to be removed from $t_i$ because it is possible that such vertex or edge is not in $g$, but does in $g'$. This is the reason why $d_g(t_i, g)$ and $d_g(t_i, g')$ do not always have the same order, and as a consequence the kernel fails to satisfy properties $P2-P4$.

For instance, consider $t_1 = \{V_1 = \{v_1, v_2, v_3\}, E_1 = \{(v_1, v_2), (v_2, v_3)\}\}$ and $t_2 = \{V_2 = \{v_1, v_2, v_3, v_4\}, E_2 = V_2 \times V_2 - Diag(V_2)\}$ as prototypes, and set:

$$
\begin{aligned}
g &= \{V = \{v_1, v_2\}, E = \{(v_1, v_2)\}\},\\
g' &= \{V' = \{v_1, v_2, v_3\}, E' = \{(v_1, v_2), (v_1, v_3)\}\} \cong t_1,\\
g'' &= \{V'' = \{v_1, v_2, v_3, v_4, v_5\}, E'' = E_2 \bigcup \{(v_1, v_5), (v_4, v_5)\}\}.
\end{aligned}
$$

It is easily seen that $V_g = (2, 7)$, $V_{g'} = (0, 5)$, and $V_{g''} = (8, 3)$, and hence, $P2$ and $P3$ fails.

What this example is trying to illustrate is that given any set of prototypes we can always find graphs making the properties fail. This is not difficult to achieve and the reader can easily construct their own examples. It is worthy to mention that these kernels can easily be introduced in the partitions and strings scopes, and even in these scenarios they continue having the same deficiencies.

Although we do not have a rigorous proof, we feel that the marginalized kernels for comparing labeled graphs [9] satisfy the desired properties $P1-P4$.

## 5 Conclusions

Structured data are becoming increasingly important in many applications. Although various measures for comparing structured data have been introduced, there is a lack for theoretical studies analyzing what a good measure is and how we could choose-between different measures. The present paper is an attempt to provide a theoretical framework for characterizing measures on structured data. For this, four basic and intuitive properties that must fulfill such kind of measures were formally introduced. On the basis of these properties, different

existing measures were studied. It was shown how competent measures, as for example the edit distances, fulfilled all the properties. It can be concluded that the introduced theoretical framework is useful not only in the development of new measures but also in the analysis and understanding of the existing ones. For the case of kernel functions, this work establishes sufficient conditions that a kernel function must meets in order to fulfill the introduced properties. A comparative study including the analysis of more measures can be done as future work.

## References

1. Chen, H., Jain, A.K.: Dental biometrics: Alignment and matching of dental radiographs. IEEE Trans. Pattern Anal. Mach. Intell. 27(8), 1319–1326 (2005)
2. Dinu, L.P., Sgarro, A.: A low-complexity distance for dna strings. Fundam. Inform. 73(3), 361–372 (2006)
3. Zhu, E., Hancock, E.R., Ren, P., Yin, J., Zhang, J.: Associating Minutiae between Distorted Fingerprints Using Minimal Spanning Tree. In: Campilho, A., Kamel, M. (eds.) ICIAR 2010, 235–245. LNCS, vol. 6112, Springer, Heidelberg (2010)
4. Dinu, L.P., Ionescu, R.: A genetic approximation of closest string via rank distance. In: SYNASC, pp. 207–214 (2011)
5. González-Díaz, R., Ion, A., Ham, M.I., Kropatsch, W.G.: Invariant representative cocycles of cohomology generators using irregular graph pyramids. Computer Vision and Image Understanding 115(7), 1011–1022 (2011)
6. Correa-Morris, J., Espinosa-Isidrón, D.L., Álvarez-Nadiozhin, D.R.: An incremental nested partition method for data clustering. Pattern Recognition 43(7), 2439–2455 (2010)
7. Bunke, H., Riesen, K.: Recent advances in graph-based pattern recognition with applications in document analysis. Pattern Recognition 44(5), 1057–1067 (2011)
8. Bunke, H., Shearer, K.: A graph distance metric based on the maximal common subgraph. Pattern Recognition Letters 19(3-4), 255–259 (1998)
9. Kashima, H., Tsuda, K., Inokuchi, A.: Marginalized kernels between labeled graphs. In: Proceedings of the Twentieth International Conference on Machine Learning, pp. 321–328. AAAI Press (2003)
10. Bunke, H., Riesen, K.: A Family of Novel Graph Kernels for Structural Pattern Recognition. In: Rueda, L., Mery, D., Kittler, J. (eds.) CIARP 2007. LNCS, vol. 4756, pp. 20–31. Springer, Heidelberg (2007)
11. Riesen, K., Bunke, H.: Approximate graph edit distance computation by means of bipartite graph matching. Image Vision Comput. 27(7), 950–959 (2009)
12. Vega-Pons, S., Correa-Morris, J., Ruiz-Shulcloper, J.: Weighted partition consensus via kernels. Pattern Recognition 43(8), 2712–2724 (2010)
13. Thor, A.: Toward an adaptive string similarity measure for matching product offers. In: GI Jahrestagung (1), 702–710 (2010)
14. Meila, M.: Comparing clusterings: an axiomatic view. In: ICML, pp. 577–584 (2005)