Symbolic Dynamics in Text: Application to Automated Construction of Concept Hierarchies

Fionn Murtagh

Department of Computer Science
Royal Holloway, University of London
Egham, Surrey TW20 0EX, England
fmurtagh@acm.org

Abstract. Following a symbolic encoding of selected terms used in text, we determine symmetries that are furnished by local hierarchical structure. We develop this study so that hierarchical fragments can be used in a concept hierarchy, or ontology. By "letting the data speak" in this way, we avoid the arbitrariness of approximately fitting a model to the data.

1 Introduction

1.1 Symmetry Group and Alternating Permutation Ordinal Encodings in Symbolic Dynamics

In symbolic dynamics, we seek to extract symmetries in the data based on topology alone, before considering metric properties. For example, instead of listing a sequence of iterates, $\{x_i\}$, we may symbolically encode the sequence in terms of up or down, or north, south, east and west moves. This provides a sequence of symbols, and their patterns in a phase space, where the interest of the data analyst lies in a partition of the phase space. Patterns or templates are sought in this topology. Sequence analysis is tantamount to a sort of topological time series analysis.

Thus, in symbolic dynamics, the data values in a stream or sequence are replaced by symbols to facilitate pattern-finding, in the first instance, through topology of the symbol sequence. This can be very helpful for analysis of a range of dyamical systems, including chaotic, stochastic, and deterministic-regular time series. Through measure-theoretic or Kolmogorov-Sinai entropy of the dynmical system, it can be shown that the maximum entropy conditional on past values is consistent with the requirement that the symbol sequence retains as much of the original data information as possible. Alternative approaches to quantifying complexity of the data, expressing the dynamical system, is through Lyapanov exponents and fractal dimensions, and there are close relationships between all of these approaches (Latora and Baranger (1999)).

Later in this work, we will use a "change versus no change" encoding, using a multivariate time series based on the sequence of terms used in a document.

From the viewpoint of practical and real-world data analysis, however, many problems and open issues remain. Firstly (Bandt and Pompe (2002)), noise in the data stream means that reproducibility of results can break down. Secondly, the symbol sequence, and derived partitions that are the basis for the study of the symbolic dynamic topology, are not easy to determine. Hence Bandt and Pompe (2002) enunciate a pragmatic principle, whereby the symbol sequence should come as naturally as possible from the data, with as little as possible by way of further model assumptions. Their approach is to define the symbol sequence through (i) comparison of neighboring data values, and (ii) up-down or down-up movements in the data stream.

Taking into account all up-down and down-up movements in a signal allows a permutation representation.

Examples of such symbol sequences from Bandt and Pompe (2002) follow. They consider the data stream $(x_1, x_2, \ldots, x_7) = (4, 7, 9, 10, 6, 11, 3)$. Take the order as 3, i.e. consider the up-down and down-up properties of successive triplets. $(4, 7, 9) \longrightarrow 012; (7, 9, 10) \longrightarrow 012; (9, 10, 6) \longrightarrow 201; (6, 11, 3) \longrightarrow 201; (10, 6, 11) \longrightarrow 102$. (In the last, for instance, we have $x_{t+1} < x_t < x_{t+2}$, yielding the symbolic sequence 102.) In addition to the order, here 3, we may also consider the delay, here 1. In general, for delay τ, the neighborhood consists of data values indexed by $t, t-\tau, t-2\tau, t-3\tau, \ldots, t-d\tau$ where d is the order. Thus, in the example used here, we have the symbolic representation 012012201201102. The symbol sequence (or "itinerary") defines a partition – a separation of phase space into disjoint regions (here, with three equivalence classes, 012, 201, and 102), which facilitates finding an "organizing template" or set of topological relationships (Weckesser (1997)). The problem is described in Keller and Lauffer (2003) as one of studying the qualitative behavior of the dynamical system, through use of a "very coarse-grained" description, that divides the state space (or phase space) into a small number of regions, and codes each by a different symbol.

Different encodings are feasible and Keller and Sinn (2005a, 2005b) use the following. Again consider the data stream $(x_1, x_2, \ldots, x_7) = (4, 7, 9, 10, 6, 11, 3)$. Now given a delay, $\tau = 1$, we can represent the above by $(x_{6\tau}, x_{5\tau}, x_{4\tau}, x_{3\tau}, x_{2\tau}, x_\tau, x_0)$. Now look at rank order and note that: $x_\tau > x_{3\tau} > x_{4\tau} > x_{5\tau} > x_{2\tau} > x_{6\tau} > x_0$. We read off the final permutation representation as (1345260). There are many ways of defining such a permutation, none of them best, as Keller and Sinn (2005a) acknowledge. We see too that our m-valued input stream is a point in $\mathbb{R}^m$, and our output is a permutation $\pi \in S_m$, i.e. a member of the permutation group.

Keller and Sinn (2005a) explore invariance properties of the permutations expressing the ordinal, symbolic coding. Resolution scale is introduced through the delay, τ. (An alternative approach to incorporating resolution

scale is used in Costa et al. (2005), where consecutive, sliding-window based, binned or averaged versions of the time series are used. This is not entirely satisfactory: it is not robust and is very dependent on data properties such as dynamic range.) Application is to EEG (univariate) signals (with some discussion of magnetic resonance imaging data) (Keller et al. (2005)). Statistical properties of the ordinal transformed data are studied in Bandt and Pompe (2002), in particular through the S_3 symmetry group. We have noted the symbolic dynamics motivation for this work; in Bandt (2005) and other work, motivation is provided in terms of rank order time series analysis, in turn motivated by the need for robustness in time series data analysis.

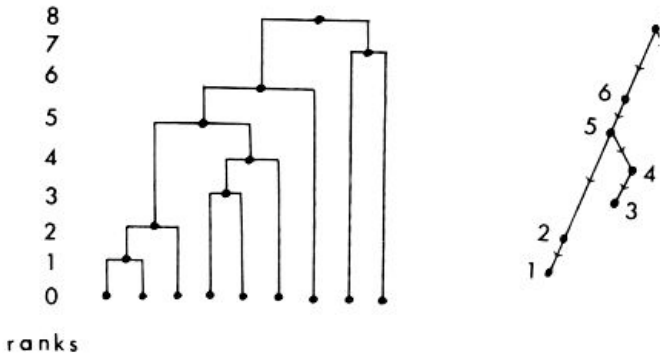

Fig. 1. Left: dendrogram with lower ranked subtree always to the left. Right: oriented binary tree associated with the non-terminal nodes.

Given the permutation representation used, let us note in passing that there is an isomorphism between a class of hierarchic structures, termed unlabeled, ranked, binary, rooted trees, and the class of permutations used in symbolic dynamics. Each non-terminal node in the tree shown in Figure 1 has one or two child nodes. This is a dendrogram, representing a set of $n-1$ agglomerations based on n initial data vectors. A packed representation (Sibson (1980)) or permutation representation of a dendrogram is derived as follows. Put lower ranked subtree always to the left; and read off oriented binary tree on non-terminal nodes (see Figure 1). Then for any terminal node indexed by i, with the exception of the rightmost which will always be n, define $p(i)$ as the rank at which the terminal node is first united with some terminal node to its right. For the dendrogram shown, the packed representation is: (125346879). This is also an inorder traversal of the oriented binary tree. The packed representation is a uniquely defined permutation of $1 \ldots n$. Dendrograms (on n terminals) of the sort shown in Figure 1, referred to as non-labeled, ranked

(NL-R) in Murtagh (1984), are isomorphic to either down-up permutations, or up-down permutations (both on $n - 1$ elements).

1.2 Motivation for an Alternative Ordinal Symbolic Dynamics Encoding

In some respects we follow the work of Keller, Bandt, and their colleagues in using an ordinal coding to provide for an encoding of the data sequence. However in the following areas we need to adopt a different approach.

- We need to handle multivariate time series.
- We need to bypass the two alternative analyses that the ordinal symbolic encoding necessarily leads to, viz. either up-down or down-up.
- Biological verisimilitude is not strong with the ordinal encoding as discussed so far.

We look at each of these in turn.

To handle multivariate time series, Keller and Lauffer (2003), and Keller and Wittfeld (2004) find the best composite time series, using projections on the first factor furnished by correspondence analysis. Correspondence analysis uses a weighted Euclidean distance between profiles (or, using the input data, the χ^2 distance) and for time-varying signals such as EEG signals, it is a superior choice compared to, say, principal components analysis.

In Bandt and Groth (2005), the need for multivariate analysis is established. Among tentative steps towards this are window-based averages of distances.

It is immediate in any inequality, $x_t > x_{t-1}$, that reversing the inequality (e.g. through considering an axial symmetry in the time axis) can lead to a new and different outcome. When we have multivariate data streams, enforcing symmetry is very restrictive. We bypass this difficulty very simply by instead using a change/no change symbolic representation. Financial verisimilitude is lost in doing this (if up = gain, down = loss); but biological verisimilitude, and that of other areas, is aided greatly.

Based on their EEG analysis, Keller and Sinn (2005a) ask: "Does there exist a basic (individual) repertoire of 'ordinal' states of brain activity?". As opposed to this, we target the hierarchy or branching fragment as the pattern that is sought, which suits the dendritic structures of the brain. While rank order alone is a useful property of data, we seek to embed our data (globally or locally) in an ultrametric topology, which also offers scope for p-adic algebraic processing. We move from real data, we take account of ordinal properties, and we end up with a topological and/or algebraic framework. This implies a data analysis perspective which is highly integrated and comprehensive. Furthermore, as an analysis pipeline, it is potentially powerful in bridging observed data with theoretically-supported interpretation.

2 The Topological View: Ultrametric Embedding

1. We seek uncontestable local hierarchical structure in the data. The traditional alternative is to impose hierarchical structure on the data (e.g. through hiearchical clustering, or otherwise inducing a classification tree).
2. We seek to avoid having any notion of hierarchical direction. In practice this would imply that hierarchical "up" (e.g. agglomerative or bottom-up) and hierarchical "down" (e.g. divisive or top-down) should each be considered independently.
3. We may wish to accommodate (i.e., include in our analysis) outliers and random exceptional values in the data. More particularly: we want to handle power law distributions, characterized by independent but not identically distributed values. An example is Zipf's law for text.
4. Therefore, for text we will use the property of linearity of text: words are linearly ordered from start to finish. (Note that a hypertext could be considered as a counter-example.)

The approach to finding local hierarchical structure is described for time series data in Murtagh (2005). We use the same approach here. The algorithm is as follows. The data used is the sequence of frequencies of occurrence of the terms of interest – nouns, noun-substantives – in their text-based order. These terms are found using TreeTagger (Schmid (1994)).

In seeking to use free text, we will also take into consideration the strongest "given" in regard to any classical text: its linearity (or total) order. A text is read from start to finish, and consequently is linearly ordered.

A text endowed with this linear order is analogous to a time series. (This opens up the possibility to generalize the work described here to (i) speech signals, or (ii) music. We will pursue these generalizations in the future.)

3 Quantifying Hierarchical Structure in a Linear Ordered Set: Application

We proceed now to particular engineering aspects of this work. We require a frequency of occurrence matrix which crosses the terms of interest with parts of a free text document. For the latter we could well take documentary segments like paragraphs.

O'Neill (2006) is a 660-word discussion of ubiquitous computing from the perspective of human computing interaction. With this short document we used individual lines (as proxies for the sequence of sentences) as the component parts of the document. There were 65 lines.

Based on a list of nouns and substantives furnished by the part-of-speech tagger (Schmid (1994)) we focused on the following 30 terms:

`support` = { "agents", "algorithms", "aspects", "attempts", "behaviours", "concepts", "criteria", "disciplines", "engineers", "factors", "goals", "interactions", "kinds", "meanings", "methods", "models", "notions", "others",

"parts", "people", "perceptions", "perspectives", "principles", "systems", "techniques", "terms", "theories", "tools", "trusts", "users" }.

This set of 30 terms was used to characterize through presence/absence the 65 successive lines of text, leaading to correspondence analysis of the 65×30 presence/absence matrix. This yielded then the definition of the 30 terms in a factor space. In principle the rank of this space (taking account of the trivial first factor in correspondence analysis, relating to the centering of the cloud of points) is $\min(\ 65 - 1, 30 - 1)$. However through all zero-valued rows and/or columns, the actual rank was 25. Therefore the full rank projection of the terms into the factor space gave rise to 25-dimensional vectors for each term, and these vectors are endowed with the Euclidean metric.

Define this set of 30 terms as the support of the document. Based on their occurrences in the document, we generated the following *reduced* version of the document, defined on this support, which consists of the following ordered set of 52 terms:

`Reduced document` = "goals" "techniques" "goals" "disciplines" "meanings" "terms" "others" "systems" "attempts" "parts" "trusts" "trusts" "people" "concepts" "agents" "notions" "systems" "people" "kinds" "behaviours" "people" "factors" "behaviours" "perspectives" "goals" "perspectives" "principles" "aspects" "engineers" "tools" "goals" "perspectives" "methods" "techniques" "criteria" "criteria" "perspectives" "methods" "techniques" "principles" "concepts" "models" "theories" "goals" "tools" "techniques" "systems" "interactions" "interactions" "users" "perceptions" "algorithms"

This reduced document is now analyzed using the algorithm described earlier. Each term in the sequence of 52 terms is represented by its 25-dimensional factor space vector.

For successive triples, if the triple is to be compatible with the ultrametric inequality, we require the recoded distances to be one of the following patterns: 1,1,1 or 2,2,2 for an equilateral triangle; and 1,2,2 in any order for an isosceles triangle with small base.

The only other pattern is 1,1,2 (in any order) which is not compatible with the ultrametric inequality. (It is seen to represent the case of an isosceles triangle with large base.)

Out of 43 unique triplets, with self-distances removed, we found 31 to respect the ultrametric inequality, i.e. 72%. The ultrametricity of this document, based on the support used, was thus 0.72.

For a concept hierarchy we need an overall fit to the data. Using the Euclidean space perspective on the data, furnished by correspondence analysis, we can easily define a terms $\times$ terms distance matrix; and then hierarchically cluster that. Consistent with our analysis we recode all these distances, using the mapping onto $\{1, 2\}$ for unique pairs of terms.

Note that this is tantamount to having a window encompassing all of the reduced document. It is also interesting to check the ultrametricity coefficient

here. This means therefore the ultrametricity coefficient in the window length n case, versus the ultrametricity coefficient in the window length 3 case. The latter was seen to be (from exhaustive calculation) above, 0.72. For the window length n case, we sampled 2000 triplets, and found the ultrametricity coefficient to be 0.56. Since the linear order is of greater ultrametric (hence, hierarchical) structure, an evident question arises as to whether it should be used as the basis for a retrieved overall or global hierarchy. We do not do this, however, because the greater hierarchical structure comes as the cost of being overly fragmentary. Instead, we adopt the approach now to be described.

Approximating a global ultrametric from below, achieved by the single linkage agglomerative hierarchical clustering method (this best fit from below is optimal), and an approximation from above, achieved by the complete linkage agglomerative hierarchical clustering method (this best fit from above is non-unique and hence is one of a number of best fits from above), will be identical if the data is fully ultrametric-embeddable. If we had an ultrametricity coefficient equal to 1 – we found it to be 0.72 for this data – then it would not matter what agglomerative hierarchical clustering algorithm (among the usual Lance-Williams methods) that we select.

In fact, we found, with an ultrametricity coefficient equal to 0.72, that the single and complete linkage methods gave an identical result. This result is shown in Figure 2.

References

BANDT, C. and POMPE, B. (2002): Permutation Entropy: a Natural Complexity Measure for Time Series, *Physical Review Letters, 88, 174102(4).*

BANDT, C. and SHIHA, F. (2005): Order Patterns in Time Series. Preprint 3/2005, Institute of Mathematics, Greifswald, www.math-inf.uni-greifswald.de/∼bandt/pub.html

BANDT, C. (2005): Ordinal Time Series Analysis. *Ecological Modelling, 182, 229–238.*

BANDT, C. and GROTH, A. (2005): Ordinal Time Series Analysis. Poster Freiburg. www.math-inf.uni-greifswald.de/∼groth

COSTA, M., GOLDBERGER, A.L. and PENG, C.-K. (2005): Multiscale Entropy Analysis of Biological Signals. *Physical Review E, 71, 021906(18).*

DE SOETE, G. (1986): A Least Squares Algorithm for Fitting an Ultrametric Tree to a Dissimilarity Matrix. *Pattern Recognition Letters, 2, 133–137.*

KELLER, K. and LAUFFER. H. (2003): Symbolic Analysis of High-Dimensional Time Series. *International Journal of Bifurcation and Chaos, 13, 2657–2668.*

KELLER. K. and WITTFELD, K. (2004): Distances of Time Series Components by Means of Symbolic Dynamics, *International Journal of Bifurcation and Chaos, 693–704.*

KELLER, K. and SINN, M. (2005): Ordinal Symbolic Dynamics, Technical Report A-05-14, www.math.mu-luebeck.de/publikationen/pub2005.shtml

KELLER. K. and SINN, M. (2005): Ordinal Analysis of Time Series. *Physica A 356, 114–120.*

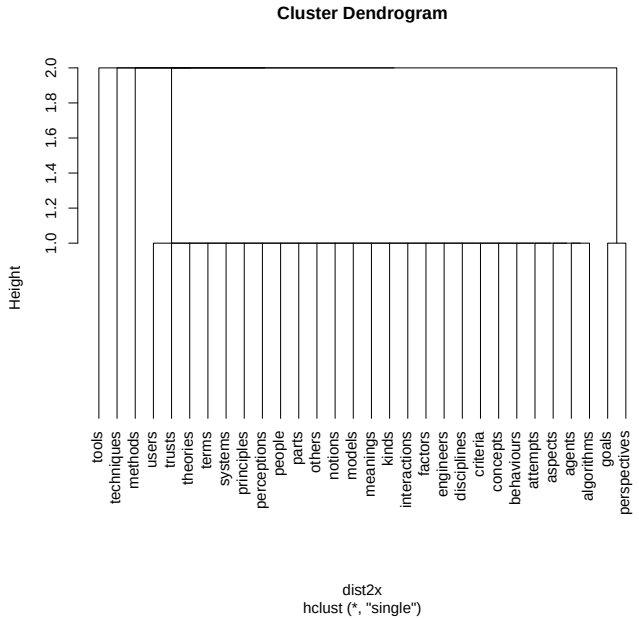

Fig. 2. Single (or identially, complete) linkage hierarchy of 30 terms, comprising the support of the document, based on (i) "no change/change" metric recoded (ii) 25-dimensional Euclidean representation.

KELLER, K., LAUFFER, H. and SINN, M. (2005): Ordinal Analysis of EEG Time Series. *Chaos and Complexity Letters, 2.*

LATORA, V. and BARANGER, M. (1999): Kolmogorov-Sinai Entropy Rate versus Physical Entropy. *Physical Review Letters, 82, 520(4).*

MURTAGH, F. (1984): Counting Dendrograms: a Survey. *Discrete Applied Mathematics, 7, 191–199.*

MURTAGH, F. (2005): Identifying the Ultrametricity of Time Series. *European Physical Journal B, 43, 573–579.*

O'NEILL, E. (2006): Understanding Ubiquitous Computing: a View from HCI, in Discussion following R. Milner, Ubiquitous Computing: How Will We Understand It?", *Computer Journal, 49, 390–399.*

SCHMID, H. (1994): Probabilistic Part-of-Speech Tagging using Decision Trees, Proc. Intl. Conf. New Methods in Language Processing. TreeTagger, www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/ DecisionTreeTagger.html

SIBSON, R. (1980): SLINK: an Optimally Efficient Algorithm for the Single-Link Cluster Method. *Computer Journal, 16, 30–34.*

WECKESSER, W. (1997): Symbolic Dynamics in Mathematics, Physics, and Engineering, based on a talk by N. Tuffilaro, http://www.ima.umn.edu/∼weck/nbt/nbt.ps