Physica A 442 (2016) 284–299

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Statistical mechanics of ontology based annotations

David C. Hoyle [a,*], Andrew Brass [b]

[a] Thorpe Informatics Ltd., Adamson House, Towers Business Park, Wilmslow Rd., Manchester, M20 2YY, UK
[b] School of Computer Science, University of Manchester, Kilburn Building, Oxford Rd., Manchester, M13 9PL, UK

HIGHLIGHTS

- New model incorporating vocabulary structure to explain term usage.
- Explains patterns seen in real annotation data sets.
- Practical metrics for the assessment of ontologies and annotations are suggested.
- Detailed analysis of regular tree ontology graphs is presented.
- Scaling laws in the growth of the optimal ontology size are derived.

ARTICLE INFO

Article history:
Received 24 March 2015
Received in revised form 16 June 2015
Available online 16 September 2015

Keywords:
Information theory
Ontology
Zipf's law
Scaling law
Annotation

ABSTRACT

We present a statistical mechanical theory of the process of annotating an object with terms selected from an ontology. The term selection process is formulated as an ideal lattice gas model, but in a highly structured inhomogeneous field. The model enables us to explain patterns recently observed in real-world annotation data sets, in terms of the underlying graph structure of the ontology. By relating the external field strengths to the information content of each node in the ontology graph, the statistical mechanical model also allows us to propose a number of practical metrics for assessing the quality of both the ontology, and the annotations that arise from its use. Using the statistical mechanical formalism we also study an ensemble of ontologies of differing size and complexity; an analysis not readily performed using real data alone. Focusing on regular tree ontology graphs we uncover a rich set of scaling laws describing the growth in the optimal ontology size as the number of objects being annotated increases. In doing so we provide a further possible measure for assessment of ontologies.

1. Introduction

With larger and more complex data sets becoming increasingly common, the annotation of data in order to semantically enrich it is a crucial task within data science [1,2]. For example, in molecular biology a gene can be annotated by domain experts with terms, t, from a controlled vocabulary, thereby allowing other researchers to comprehend the function and role of that gene. Similarly, user tagging of information sources such as documents, photographs, or online content, provides additional meta-data and lead to emergent but uncontrolled vocabularies (often called folksonomies [3]). Terms within a vocabulary can be further organized in a hierarchical structure such as a taxonomy [1], in which terms closer to the root of the hierarchy are less specific than those further from the root. Hierarchical organization of vocabulary terms can also be

* Corresponding author.
E-mail addresses: david.hoyle@thorpeinformatics.co.uk (D.C. Hoyle), a.brass@manchester.ac.uk (A. Brass).

http://dx.doi.org/10.1016/j.physa.2015.09.020
0378-4371/

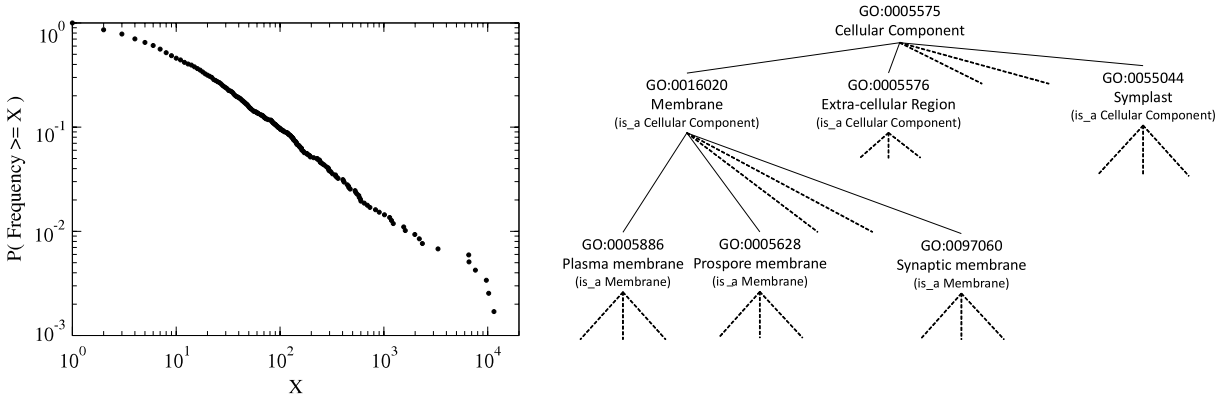

Fig. 1. Left-hand plot shows a Zipf's law plot for GO gene annotations contained within Human GOA. Annotations have been taken from the cellular component sub-ontology. Terms near the root of the cellular component of GO are shown, for illustration, in the right-hand schematic—the dashed lines indicate the presence of further child terms.

used to specify richer semantic relationships between terms—richer than just indicating that one term is a simply a sub-type of another. These richer hierarchical semantic structures are generally called ontologies [4].

The annotations that result from an ontology or taxonomy can exhibit interesting patterns. For example, Kalankesh et al. [5] have shown that distributions of term frequencies, f_t, taken from Gene Ontology (GO) [6] annotations typically follow Zipf's law [7,8]. Fig. 1 shows a Zipf's law plot for annotations taken from the cellular component sub-ontology of GO. The schematic on the right-hand side of Fig. 1 shows, for illustration, part of the ontology that was used to produce the annotation data set plotted on the left-hand side of Fig. 1. Statistical mechanics provides us with a natural tool to understand these annotation patterns, by allowing us to develop a formalism that quantifies both the natural variations in the annotation process, and the ontology structure itself. Although structure-based measures of ontologies already exist [9], within this work we are quantifying the ontology structure from the perspective of the annotations that arise, rather than simply quantifying the ontology structure in isolation. The goals, and ultimately the benefits, of developing a statistical mechanics based formalism are both practical and theoretical.

1.1. Ontologies as information stores

Ontologies and taxonomies, whether formally constructed or emergent, represent a store of information. Organizing a hierarchical store of information requires effort to be expended to create an ordered structure. Work by Ferrer i Cancho et al., within an information theoretic framework, has shown how heavy tailed and essentially hierarchical patterns of term usage can arise simply from a principle of minimizing the communication effort expended when using those terms [10–12]. Within this current paper we also use information theoretic ideas, but it is the process of transferring information from an ontology to an annotated object that we study, *i.e.*, *after* the hierarchical term structure has been determined or prescribed. We do so using an explicit statistical mechanical model that takes into the structure of the ontology. Whilst existing work within the literature has used a specific Hamiltonian to study patterns of word usage, that work has not *per se* been interested in the impact of any underlying prescribed structure in the vocabulary [13]. Similarly, novel work by Palla et al. [14] and Tibély et al. [15] has related tag usage patterns to ontology structure, but focused on an in depth study of observed tag patterns, rather than taking a Hamiltonian model based approach.

The remainder of this paper is organized as follows—in Section 2 we express the annotation process as an ideal lattice gas model in an inhomogeneous field. In Section 3.1 we use the lattice gas model to understand the term frequency patterns seen by Kalankesh et al. [5], and we identify, LD_t, the number of leaf descendants of a node t, as the key quantity controlling the expected term usage frequencies. In Section 4 we derive the most likely natural form for the inhomogeneous field strength, thereby giving rise to a local measure of the ontology. This natural form for the inhomogeneous field also allows us, in Section 5, to construct an ensemble of ontologies of differing complexity. By restricting the ensemble to the class of regular trees we reveal in Section 5.2 a set of transitions in the optimal tree size, and associated scaling laws, as the number of objects being annotated is increased. Finally in Section 6 we discuss a number of possible extensions of the statistical mechanical approach to quantifying ontology structures.

2. Statistical mechanical theory of the annotation process

We consider an ontology to be represented by a rooted Directed Acyclic Graph (DAG) [16,17]. An example DAG, in this case a tree, is shown in Fig. 2. Real-world ontologies are typically not pure trees, and we use a tree structure simply for illustrative purposes. The formalism we develop in this section will be equally applicable to any valid DAG structure. Associated with each node of the DAG is a particular term, and we use node and term interchangeably.