

From entropy to ontology

Jacques Calmet, Anusch Daemi

Institute for Algorithms and Cognitive Systems (IAKS)

University of Karlsruhe (TH)

Am Fasanengarten 5, D-76133 Karlsruhe

Germany

email: {calmet, daemi}@ira.uka.de

Abstract

The theoretical foundations for a definition of distance on ontologies are laid out in this paper. The distance measure is defined using the well known concepts of *entropy* and *mutual information* from information theory. These formal methods are adapted for practical use in ontologies through the definition of useful discrete random variables. These include centrality measures like degree, closeness or betweenness.

Keywords: Distance, Ontology, Entropy, Mutual Information, Centrality.

1 Ontological structures and distance

It is well known that the concept of ontology refers to some sort of taxonomy of knowledge on a domain together with an advanced structure on this knowledge (see [Gruber, 1993]). In most cases, it is very hard to identify or define such a structure. In this paper we propose a generic approach based on the definition of a concept of distance on ontologies. Once a distance is available, it becomes possible to identify the structure associated to an ontology, which will be the topic of another paper.

Ontologies are related to a model of knowledge, and knowledge in turn to information. Thus it makes sense to introduce the concept of entropy and mutual information, as defined by Shannon for information theory in [Shannon, 1948], on ontologies. Entropy and mutual information in turn enables us to define a distance measure formally. With this distance a sound foundation is given for the capturing of the inherent structure of an ontology.

The paper is structured as follows. In section 2 we introduce the concept of distance from traditional mathematics to computational linguistics, which uses some kind of entropy. In section 3 the history of entropy is briefly described and ends with the definition provided by Shannon for information theory. Section 4 applies the concept of entropy and mutual information to ontologies via a definition of distance. The paper concludes in section 5 with an outlook on further research opportunities related to the application of entropy on ontology.

2 A brief history of distance

The concept of distance (metric) in mathematics is a very old one and ranges back to the times of Aristoteles and Euclid. A distance d between two points a and b is commonly described by three axioms:

1. Positive definiteness: $\forall a, b: d(a, b) \geq 0$ and $d(a, b) = 0$ if and only if $a = b$
2. Symmetry: $d(a, b) = d(b, a)$
3. Triangle inequality: $\forall a, b, c: d(a, b) \leq d(a, c) + d(c, b)$

One well known distance measure for the euclidean space is the Euclidean distance, the sum of the quadratic differences of continuous features in a space, i.e. the real numbers in \mathbb{R}^2 . If this measure is used in a multidimensional space where the dimensions have different semantics, for example height and weight, a standardization of scales is required. One example of this kind of distance is the Mahalanobis distance, which measures the standard deviation and covariance of each feature and expresses the feature values as multiples of the standard deviation and covariance (see [Russel and Norvig, 2003], p. 735). An interesting approach is proposed in [Menger, 1942] as he defines a more probabilistic concept of distance. He introduces a probability distribution F_{pq} whose value $F_{pq}(x)$ is the probability, that the distance between p and q is less than x , for any real number x .

If we have discrete rather than continuous features, the Hamming distance can be used to measure the difference between the features. The Hamming distance is defined as the total number of different characters between two discrete features. One of the many applications of this measure is found in coding theory, where it is used to correct codewords sent via a noisy channel (see [Pierce, 1980]).

To measure the distance of some kind of special discrete features like words, terms, documents and similar corpora of text, computational linguistics has introduced as distance measures *positioning* and *frequency*, among others.

The positioning distance measure is mostly used in Vector Space Models (VSM) where documents and queries are represented in a high dimensional vector space. The basis vectors are composed of index terms,

which are words relevant to the problem. The location of the document vector in the VSM is determined by weights assigned to the index terms for each document (see [Jones and Willett, 1997]). These weights can be calculated in different ways. Usually they are either term frequencies (*tf*) or inverted document frequencies (*tfidf*). The distance between a document and the query is measured using the cosines between the two vectors (also known as normalized correlation coefficient [Sethi, 1999]). The selection of basis vectors for the VSM is critical for successful recognition of similarities between documents, because they can severely warp the distance measure if poorly chosen.

The frequency approach is based on the frequency of terms in a document, term frequency, which is used in the information retrieval measure *tfidf*. Term frequency/inverted document frequency weights the term frequency of a term in a document by a logarithmic factor that discounts the importance of the term in regard to all relevant documents. So terms appearing too rarely or frequently have a lower weight than those holding the balance (see [Hotho *et al.*, 2002]).

A distance measure originating from term frequency and *tfidf* can be calculated by introducing some form of entropy as first presented in [Resnik, 1995]. This method assigns instances of concepts *c* a probability $p(c)$ of encountering them in a document. The information content of such a concept *c* is subsequently defined as $-\log p(c)$. The distance measure is based on the assumption that the more information two concepts have in common, the more similar they are. The information shared by two concepts is indicated by the information content of the concepts that subsume them in a given taxonomy. In practice one needs to measure similarity between terms rather concepts, so calculations have to be done over sets of concepts representing senses of the terms.

It should be noted, that our approach is not using one of these information retrieval measures because they usually require some corpora of text to work on, for example [Miller *et al.*, August 1993]. Instead we try to focus on using solely the information contained in an ontology.

3 The origins of entropy

The concept of entropy originates in physics (the second law of thermodynamics) and statistical mechanics from the works of Maxwell, Boltzmann, Gibbs and others. Entropy has then become increasingly popular in computer science and information theory, particularly through the work of [Shannon, 1948].

In thermodynamics entropy is defined as the irreversible increase of nondisposable energy in the universe, meaning that you can change physical, chemical, and electrical energy completely into heat but the reversal of this process accomplished without outside help or without an inevitable loss of energy in the form of irretrievable heat is impossible [Maxwell, 2001]. This fact is expressed through the second law of thermodynamics: Entropy can only increase in the *universe*. It is quite possible to decrease entropy in a

system, but that must be balanced by an at last equal increase in entropy elsewhere.

Maxwell, Boltzmann and Gibbs have extended these ideas into the domain of statistical mechanics, where *macrostates* and *microstates* play an important role. For example, the temperature of a system defines a macrostate whereas the movement of the molecules, that is the kinetic energy of them, defines the microstates from which the macrostate is composed. The temperature is then recognized as an average of the microstate variables, which represent the average kinetic energy of the system. To put it more formally, the definition of entropy by Boltzmann [Tolman, 1979] is

$$-k \sum_i P_i \log P_i$$

whereas the P_i are the probabilities, that particle *i* will be in a given microstate, and all the P_i are evaluated for the same macrostate. *k* is the famous *Boltzmann constant* in thermodynamics, but otherwise it can be arbitrary.

Entropy, as defined in the work of Shannon, represents the *information content* of a message or, from the point of view of the receiver, the uncertainty about the message the sender produced prior to its reception. It is defined as

$$-\sum_i p(i) \log p(i)$$

whereas $p(i)$ is the possibility of receiving message *i* and Shannon has shown that $\log p(i)$ is the only function that satisfies all requirements to measure information. The unit used is the *bit* invented by John Tukey, Shannon's colleague at Bell Labs to mean binary digit, which is the appropriate unit for entropy because of the conventional use of base-two logs in computing Shannon entropy. The apparent similarity to the Boltzmann equation is evident and Shannon has detailed in his work (see [Shannon, 1948]), that only this equation satisfies the necessary requirements to be a function measuring the uncertainty of a message. The entropy is at maximum value when all events are equiprobable, which is intuitive, because in this case the receiver has maximum uncertainty about which message he may receive next. On the other hand entropy is at minimum, if $p(i) = 1$ or $p(i) = 0$, denoting an event that occurs every time or never, so the receiver knows prior to the receipt of the message what he will get.

4 From entropy to ontology

Entropy is used as a distance measure in many application areas, especially communication theory (coding theory) and statistics. One such definition of distance is the Kullback-Leibler distance or relative entropy (see [Kotz and Johnson, 1981]):

$$D(g \parallel p) = \sum_i g(i) \log \frac{g(i)}{p(i)}$$

It is used for measuring the distance, or error, between a real distribution $p(i)$ in the system and an assumed

distribution $g(i)$. The Kullback-Leibler distance is not a real distance, because it does not satisfy symmetry or the triangle inequality, only positive definiteness. Nonetheless, it is useful to think of the relative entropy as a distance between distributions.

A similar approach will be used to define a distance on ontologies. We'll make use of the relative entropy, mutual information and conditional mutual information (see [Cover and Thomas, 1991]).

4.1 Distance

The distance $d(X, Z; Y)$ between two concepts X and Z in an ontology will be defined as the reduction of uncertainty in X due to knowledge of concept Y , when Z is known. This leads us naturally to the *entropy* $H(X)$ of a discrete random variable. Let X be such a random variable with its associated probability mass function $p(x) = Pr\{X = x\}, x \in \Omega$ and alphabet Ω . The entropy $H(X)$ of X is defined as:

$$H(X) = - \sum_{x \in \Omega} p(x) \log p(x).$$

The conditional entropy $H(X|Z)$ (with $p(z) = Pr\{Z = z\}, z \in \text{alphabet } \Psi$), used in our definition of distance, is defined as:

$$H(X|Z) = - \sum_{x \in \Omega} \sum_{z \in \Psi} p(x, z) \log p(x|z)$$

For the distance we propose the use of the *conditional mutual information*, since we reduce the uncertainty in X given concept Z and further decrease it, if we have more information in the form of concept Y .

$$d(X, Z; Y) = H(X|Z) - H(X|Y, Z)$$

Since there can be more than one concept providing information about X , Y can be seen as a vector of concepts $\mathbf{Y} = Y_1, \dots, Y_n$:

$$d(X, Z; \mathbf{Y}) = d(X, Z; \mathbf{Y}) = H(X|Z) - H(X|\mathbf{Y}, Z)$$

If no additional information is available besides Z and $X \neq Z$, we define $d(X, Z; 0)$ as the mutual information $I(X, Z)$ (the reduction of uncertainty in X due to knowledge of Z) between X and Z :

$$d(X, Z; 0) = I(X, Z) = H(X) - H(X|Z)$$

If we have $X = Z$, we have no information gain, because X provides all information about itself already:

$$d(X, X; 0) = H(X|X)$$

If the concepts are independent of each other (e.g. $p(x|z) = p(x)p(z)$), we have also no mutual information gain as described for example in [Rényi, 1982].

Now if we model our representation of the world with a graph, in the simplest case with a taxonomy or more complex with an ontology, the Y_i are the concepts on a path between X and Z . Of course there is usually more than one path between X and Z , so all the l paths \mathbf{Y}_{XZ} between both concepts can contribute to the reduction of uncertainty in X :

$$\begin{aligned} d(X, Z; \mathbf{Y}_{XZ}) &= d(X, Z; Y_j^l) \\ &= H(X|Z) - \sum_{j=1}^l H(X|Y_1, \dots, Y_{n_j}, Z) \end{aligned}$$

For this modeling it is assumed that the graph representation is cycle free. This sum is in no way optimized and does not account for the interdependencies between the different Y_j^l , and therefore may lead to questionable results. This is an open research question which must be investigated to lead to optimal results.

4.2 Some properties

Positive definiteness

$$\forall X, Z : d(X, Z; \mathbf{Y}) \geq 0$$

Proof

We have $H(X) \geq 0$ and the fact, that conditioning reduces entropy (see [Cover and Thomas, 1991], p. 27): $H(X|Y) \leq H(X)$.

$$\begin{aligned} d(X, Z; \mathbf{Y}) &= d(X, Z; Y_1, \dots, Y_n) \quad (1) \\ &= \underbrace{H(X|Z)}_{\geq H(X|\mathbf{Y}, Z)} - H(X|\mathbf{Y}, Z) \quad (2) \end{aligned}$$

In (2) we use the fact, that conditioning reduces entropy:

$$\begin{aligned} \Rightarrow H(X|Z) - H(X|\mathbf{Y}, Z) &\geq 0 \\ \Rightarrow d(X, Z; \mathbf{Y}) &\geq 0 \end{aligned}$$

$d(X, X; 0) = 0$ follows directly from the definition of conditional entropy ■

Symmetry

$$d(X, Z; \mathbf{Y}) = d(Z, X; \mathbf{Y})$$

Proof

$$\begin{aligned} d(X, Z; \mathbf{Y}) &= H(X|Z) - H(X|\mathbf{Y}, Z) \quad (3) \\ &= I(X; \mathbf{Y}|Z) \quad (4) \\ &= I(X; Y_1, \dots, Y_n|Z) \quad (5) \\ &= I(Y_1, \dots, Y_n|Z; X) \quad (6) \\ &= I(\mathbf{Y}|Z; X) \quad (7) \\ &= d(Z, X; \mathbf{Y}) \quad (8) \end{aligned}$$

(5) used the symmetry of the conditional mutual information.

In other words, X has as much information about the Y_i given Z as the Y_i have about Z given X ■

4.3 Probability Distribution

For selection of a probability mass function $p(x)$ there are several possible approaches. We will propose the use of some centrality measures (Degree, Closeness, Betweenness) described in [Wasserman and Faust, 1994] as a first step.

For simplicity we will use as alphabet Ω the set consisting of one node N with N being a node in the

ontology. With this alphabet we define a discrete random variable $X : \Omega \rightarrow \mathbb{R}$ for each node via:

$$X = \frac{\text{deg}(N)}{2}$$

This definition, with Ω as alphabet and the normalized degree

$$\text{deg}(N) = \frac{\text{degree}(N)}{N}$$

obviously satisfies the axioms for a probability mass distribution.

4.4 Entropy of a node

The degree of a node in an ontology can have the meaning of ambiguousness. For example, a node with a very low degree can be assigned rather easily a unique meaning, because it only has few (if any) connections to others, whereas a node with a high degree has many connections. Depending on the semantics of the connections it can be very difficult to assign a unique semantic meaning to that node.

This is reflected in the entropy assigned to a node. A node with $\text{deg}(N) = 0$ has no ambiguity, so there is no uncertainty about it and therefore (with the assumption $0 \cdot \log(0) = 0$):

$$\begin{aligned} H(N) &= - \sum_{\Omega} \frac{\text{deg}(N)}{2} \log\left(\frac{\text{deg}(N)}{2}\right) \\ &= -0 \log(0) - \left(1 - \frac{0}{2}\right) \log\left(1 - \frac{0}{2}\right) \\ &= 0 \end{aligned}$$

In contrast, a node which is connected to every other node in the network ($\text{deg}(N) = 1$) should have maximum uncertainty:

$$\begin{aligned} H(N) &= - \sum_{\Omega} \frac{\text{deg}(N)}{2} \log\left(\frac{\text{deg}(N)}{2}\right) \\ &= -\frac{1}{2} \log \frac{1}{2} - \left(1 - \frac{1}{2}\right) \log\left(1 - \frac{1}{2}\right) \\ &= -\frac{1}{2}(-1 - 1) \\ &= 1 \end{aligned}$$

Nodes with degree $0 < \text{deg}(N) < 1$ should have a monotonically increasing entropy because more links to other nodes means more uncertainty about that node. This property is ensured with the above defined discrete random variable. It increases monotonically with the degree from 0 to $\frac{1}{2}$, and therefore the entropy increases from 0 to 1.

5 Conclusion

A theoretical approach for a definition of distance on ontologies has been shown. It uses the well known concepts entropy and mutual information, as specified by Shannon, for defining the amount of information some concepts contribute to a specific target concept. To put it another way, the distance measures the reduction of uncertainty of one concept relative

to another, by considering possible target concepts in between them.

Further research has to be done for an optimal selection of probability distributions and discrete random variables. Particularly interesting as discrete random variables may be the betweenness and information centrality measures (see [Stephenson and Zelen, 1989]), since information centrality includes the statistical definition of information on network structures. Another approach would be the use of *algorithmic entropy* or Kolmogorov complexity, as it defines the absolute information content of a string (see [Li and Vitányi, 1993]).

References

- [Cover and Thomas, 1991] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley Series in telecommunications, 1991.
- [Gruber, 1993] T.R. Gruber. Towards principles for the desing of ontologies used for knowledge sharing. *Knowledge Aquisition*, pages 199–220, 1993.
- [Hotho et al., 2002] A. Hotho, A. Maedche, and S. Staab. Ontology-based text document clustering. *Künstliche Intelligenz*, pages 48–54, 04 2002.
- [Jones and Willett, 1997] K. Sparck Jones and P. Willett, editors. *Readings in Information Retrieval*. Morgan Kaufmann, 1997.
- [Kotz and Johnson, 1981] S. Kotz and N.L. Johnson, editors. *Encyclopedia of statistical science*, volume 4, pages 421–425. John Wiley and Sons, 1981.
- [Li and Vitányi, 1993] M. Li and P. Vitányi. *An introduction to kolmogorov complexity and its applications*. Springer Verlag, 1993.
- [Maxwell, 2001] J. C. Maxwell. *Theory of Heat*. Dover; Reprint, 2001.
- [Menger, 1942] K. Menger. Statistical metrics. In *Proceedings of Natural Academic Sciences*, volume 28, page 535, 1942.
- [Miller et al., August 1993] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. *Introduction to WordNet: An On-line Lexical Database*. Cognitive Science Laboratory, Princeton University, August 1993.
- [Pierce, 1980] J.R. Pierce. *An introduction to information theory - Symbols, signals and noise*. Dover, second edition, 1980.
- [Resnik, 1995] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of IJCAI*, pages 448–453, 1995.
- [Russel and Norvig, 2003] S. Russel and P. Norvig. *Artificial Intelligence - A modern approach*. Prentice Hall, second edition, 2003.
- [Rényi, 1982] A. Rényi. *Tagebuch über die Informationstheorie*. Birkhäuser Verlag, 1982.
- [Sethi, 1999] I. K. Sethi. Document representation and IR models. <http://www.cse.secs.oakland.edu/isethi/IR/Coursenotes/Notes1.pdf>, 1999.

- [Shannon, 1948] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, July and October 1948.
- [Stephenson and Zelen, 1989] K. Stephenson and M. Zelen. Rethinking centrality: Methods and examples. *Social Networks*, 11:1–37, 1989.
- [Tolman, 1979] R. C. Tolman. *The principles of statistical mechanics*. Dover, 1979.
- [Wasserman and Faust, 1994] S. Wasserman and K. Faust. *Social network analysis: Methods and applications, Structural analysis in the social sciences*. Cambridge University Press, Cambridge, 1994.