Semantic Metrics

Bo Hu, Srinandan Dasmahapatra, Yannis Kalfoglou,
Harith Alani, Nigel Shadbolt

IAM Group, ECS, University of Southampton, SO17 1BJ, UK
{bh, sd, y.kalfoglou, ha, nrs}@ecs.soton.ac.uk

Abstract In the context of the Semantic Web, many ontology-related operations, e.g. ontology ranking, segmentation, alignment, articulation, reuse, evaluation, can be boiled down to one fundamental operation: computing the similarity and/or dissimilarity among ontological entities, and in some cases among ontologies themselves. In this paper, we review standard metrics for computing distance measures and we propose a series of semantic metrics. We give a formal account of semantic metrics drawn from a variety of research disciplines, and enrich them with semantics based on standard Description Logic constructs. We argue that concept-based metrics can be aggregated to produce numeric distances at ontology-level and we speculate for the usability of our ideas through application example areas.

1 Introduction

We witness a shift in participation of ontology authoring from knowledge engineers to interested practitioners. This is fueled, partly, by the ever growing interest in the Semantic Web and semantic technologies in general. It is causing an unprecedented influx of ontologies in the public domain. For instance, as of March 2006 we encountered at least 100 Wine related ontologies in various formats (e.g. OWL, RDF(S), DAML, etc.) and some 200 ontologies with definitions of the omnipresent concept *person*. This emerging "grass root" approach to ontology engineering has put the onus on ontology management and calls for a variety of new tasks, such as ontology ranking, segmentation, evaluation, just to name a few. We observe a common root for accomplishing these tasks: engineer the similarity and/or dissimilarity assessment with respect to ontology concepts or even among ontologies themselves.

We see several areas as relevant: knowledge representation, statistical clustering, data mining, information retrieval, all of which have contributed to the problem of computing similarity/dissimilarity of concepts. We are particularly interested in building upon all this work and focusing on metrics leveraging the semantics of concepts. This is a key requirement in the ontology heterogeneity problem especially when we consider that reaching consensus on how to capture semantics embedded in those ontologies is hard to achieve in the first place.

In this paper, we narrow our focus to the description logic (DL) based OWL language. We investigate a series of distance measures that our semantic metrics

draw upon. These are discussed in Section 2. We then explore how different metrics can be semantically enriched and applied to the computation of distances between concepts in Section 3, and how can they be extended to ontologies themselves (Section 4). Finally, in Section 5, we present three major applications in which our metrics can be used as a complementary means to work with and enhance existing technology and we conclude this paper with several points that need further investigation.

2 Background

2.1 Distance measures

In mathematics, the concrete idea of *distance* between two spatial points has been abstracted as a metric or distance function over a set $\mathfrak{S}$ so that $\Delta : \mathfrak{S} \times \mathfrak{S} \rightarrow \mathfrak{R}$ where $\mathfrak{R}$, the set of real numbers, is the numeric representation of *distance*. Stemming from the spatial distance between two points, the term of *distance* has been used in versatile domains and situations ranging from geometry, physics to information theory. An orthodox distance function must be **non-negative** and **symmetry** and satisfy the **triangle inequality**.

In two dimensional euclidean space, the distance between points, $\{p_1, p_2\}$ and $\{q_1, q_2\}$, can be computed as *City Block (Manhattan) Distance*, *Euclidian Distance*, or *Chebyshev Distance*. Analogous to two dimensional space distance, *Euclidian Distance* is generalised in m dimensional space as *Minkowski Distance* in the form of $\Delta_{\text{Min}}(p, q) = (\sum_i \mid p_i - q_i \mid^m)^{1/m}$.

The idea of distance, with a broader sense of measuring how far apart that two objects are, has been applied to compute the discrepancy of documents in Information Retrieval (IR), disagreement of words in a lexical taxonomy in Knowledge Representation, and dissimilarity of strings in Information Theory. The semantic metric that we proposed in this paper stems from the generally sensed distance measures that are discussed as follows.

The vector space model (VSM) has been widely used in traditional IR to compute the similarity of documents. VSM creates a space in which both the candidate documents and the queries are represented as vectors. Normally, VSM is proceeded in three steps: 1) document indexing: by extracting content bearing terms from the document text, a document can be reduced to a vector of indexing *key-words*; 2) index weighting: the *key-words* are weighted to enhance the relevance between documents and the query; and 3) document ranking: the numeric similarity values between vectors of *key-words* are obtained (see Equation 1) based on which documents can be sorted.

$$\Delta_{VSM}(p, q) = -\log \mathtt{sim}_{VSM}(p, q) = -\log \frac{\sum\limits_i p_i \times q_i}{\sqrt{\sum\limits_i p_i^2}\sqrt{\sum\limits_i q_i^2}} \tag{1}$$

In the information theory, *entropy* (denoted as $H(X)$) is borrowed from thermodynamics to measure the information content of a message or uncertainty of

a message from the receiver's perspective [18]. A full account of Shannon's view of mathematical theory of information, however, is beyond the scope of this paper. We restrict our focus to the information gain with respect to one variable based on the observation of another and use such a measure as distance between arbitrary objects. This is captured in *conditional entropy* (or equivocation) that measures how much uncertainty a variable Y has, if the complete knowledge regarding another variable X is completely known. Representing using $H(X \mid Y)$, *conditional entropy* is defined as

$$H(X \mid Y) = -\sum_{x,y} p\,(x,\, y) \log \frac{p\,(x)}{p\,(x,\, y)} \tag{2}$$

In practice, although it is not as restrict as metric, *conditional entropy* can be regarded as a divergency measure between two variables, where $H(X \mid X) = 0$. The more conditional entropy is, the less information gains from Y regarding X, the further apart X and Y are.

The *entropy* theory leads to another distance. In a discrete domain, the Kullback-Leibler divergence measures the disagreement of two distributions. Let p and q be the discrete distributions of a variable, the "distance" between p and q is computed as

$$\Delta_{KL}\,(p, q) = \sum_i p_k \log \left(\frac{p_k}{q_k} \right)$$

Note that Kullback-Leibler divergence is not symmetry and is positive definiteness [4]. Kullback-Leibler divergence has several symmetrised variants that fit better as distance metrics.

2.2 Ontology and ontology languages

"What counts as an ontology?" is still a highly debated question answers to which range from simple taxonomies to logically sound and coherent constructs whose underlying model supports description logic-based inferences [2]. In order to discuss the distance with respect to ontological entities and ontologies themselves, we need first clarify our intuitions about ontologies. In stead of giving full philosophical reflection on the term *ontology*, we take a rather opportunistic Artificial Intelligence (AI) approach and restrict ontology to be "a specification of a conceptualisation" [8]. Albeit the fact that many models, e.g. database schemata, UML models, and Semantic Network models [19], can be considered ontologies in a broader sense, we normally confine our view of *conceptualisation* as the following formalisation:

an ontology is a four-tuple $\langle \mathcal{C}, \mathcal{P}, \tau_c, \tau_p \rangle$, where $\mathcal{C}$ is a set of unary predicates as concepts, $\mathcal{P} \subseteq \mathcal{C} \times \mathcal{C}$ a set of binary relations as properties and τ_c and τ_p introduction axioms of concepts and properties respectively.

Description logics Description Logics (DLs) are a family of knowledge representation and reasoning formalisms that have attracted substantial research recently, especially, after the DL-based ontology modelling languages (e.g. OWL [12])

are considered to be of crucial importance for the Semantic Web initiative [3]. DLs are based on the notions of concepts (i.e. unary predicates) and properties (i.e. binary relations). Using different constructs, complex concepts can be built up from primitive ones.

Constructor	**Syntax**	**Semantics** (*Interpretation*)
Top (Universe)	$\top$	$\mathfrak{D}^{\mathcal{I}}$
Bottom (Nothing)	$\bot$	$\emptyset$
Primitive Concept	A	$\mathsf{A}^{\mathcal{I}} \subseteq \mathfrak{D}^{\mathcal{I}}$
Conjunction	$\mathsf{C} \sqcap \mathsf{D}$	$\mathsf{C}^{\mathcal{I}} \cap \mathsf{D}^{\mathcal{I}}$
Disjunction	$\mathsf{C} \sqcup \mathsf{D}$	$\mathsf{C}^{\mathcal{I}} \cup \mathsf{D}^{\mathcal{I}}$
Negation	$\neg\, \mathsf{C}$	$\Delta^{\mathcal{I}} \setminus \mathsf{C}^{\mathcal{I}}$
Universal quantification	$\forall\, \mathsf{R}\,.\,\mathsf{C}$	$\{\ c \in \Delta^{\mathcal{I}} \mid \forall\ d \in \Delta^{\mathcal{I}} : \langle\ c, d\ \rangle \in \mathsf{R}^{\mathcal{I}} \rightarrow d \in \mathsf{C}^{\mathcal{I}}\ \}$
Existential quantification	$\exists\, \mathsf{R}\,.\,\mathsf{C}$	$\{\ c \in \Delta^{\mathcal{I}} \mid \exists\ d \in \Delta^{\mathcal{I}} : \langle\ c, d\ \rangle \in \mathsf{R}^{\mathcal{I}} \wedge d \in \mathsf{C}^{\mathcal{I}}\ \}$
Number Restrictions	$\geq n\, \mathsf{R}\,.\,\top$	$\{\ c \in \Delta^{\mathcal{I}} \mid \sharp\{\ d \in \Delta^{\mathcal{I}} : \langle\ c, d\ \rangle \in \mathsf{R}^{\mathcal{I}}\ \} \geq n\ \}$
	$\leq n\, \mathsf{R}\,.\,\top$	$\{\ c \in \Delta^{\mathcal{I}} \mid \sharp\{\ d \in \Delta^{\mathcal{I}} : \langle\ c, d\ \rangle \in \mathsf{R}^{\mathcal{I}}\ \} \leq n\ \}$
Primitive Property	R	$\mathsf{R}^{\mathcal{I}} \subseteq \mathfrak{D}^{\mathcal{I}} \times \mathfrak{D}^{\mathcal{I}}$

Table 1. Syntax and semantics of DL-based constructs

Let *CN* denote a concept name, C and D be arbitrary concepts, R be a property, n be a non-negative integer and $\top$, $\bot$ denote the top and the bottom. A *concept introduction axiom* in DLs is either $CN \stackrel{\cdot}{\sqsubseteq} \mathsf{C}$ (partial definition) or $CN \doteq \mathsf{C}$ (full definition). An interpretation $\mathcal{I}$ is a couple $(\ \mathfrak{D}^{\mathcal{I}},\ \cdot^{\mathcal{I}})$ where the nonempty set $\mathfrak{D}^{\mathcal{I}}$ is the domain of $\mathcal{I}$ and the $\cdot^{\mathcal{I}}$ function maps each concept to a subset of $\mathfrak{D}^{\mathcal{I}}$ while each role to a subset of $\mathfrak{D}^{\mathcal{I}} \times \mathfrak{D}^{\mathcal{I}}$. The interpretation of some DL-based constructs are illustrated in Table 1. The uniform syntax and unambiguous semantics of DLs lend themselves to powerful reasoning algorithms that can automatically classify the domain knowledge in hierarchical structures.

Thus far, many ontology languages have been proposed and standardised, e.g. RDF(S) [11], OWL [12], etc. Despite the apparent differences, many of current ontology languages aiming at facilitating semantic web applications can be regarded as tractable and decidable subsets of description logics.

3 Semantic metric of concepts

Distance of concepts is, by no mean, a new idea. It can be approached in two directions, extensional and intensional. Extensional approaches normally assume an unbiased population of instance data from which a numeric similarity/dissimilarity can be obtained by applying probability distributions, concept co-occurrences and cosine measures of vectors, e.g. in [5] and [20]. Intensional

approaches exploit features defined directly over the concepts and apply measures such as Tversky's model and graph-based ones, e.g. in [13]. More specifically, graph-based methods represent ontologies as directed acyclic graphs and count the total number of weighted edges, being them inheritance relationships and/or properties; feature-based ones characterise concepts with discrete semantic bearing components, e.g. concept names, property names, domains, etc. and aggregate, as a weighted average, the similarity/dissimilarity of each pair of components. Both extensional and intensional methods have advantages and disadvantages. On one hand, although instance data have been argued that can best capture the semantics and there are plenty of well studied techniques that can be leveraged, in reality an unbiased population is not always available especially for ontologies published on the loosely regulated Web. The applicability of such approaches, therefore, is highly sceptical. On the other hand, the intensional approaches would probably not win the battle due to: 1) the ambiguity of converting semantic discrepancy—e.g. *equivalent*, *more general than*, etc.—into numeric values, 2) the computational complexity demonstrated by both graph-matching and SAT problems, and 3) their reliance on good modelling habits of those people constructing the ontologies. Intensional ones might also require more involvement from human observers, e.g. weighting different types of edges in graph-based algorithms. In this paper, we adopt an eclectic approach: we produce signatures characterising the logic restrictions of concepts and the distance of concepts is reduced to the distances between different vectors of such semantic bearing signatures.

In this section and throughout the rest of the paper, two ontologies are used as examples and the test-bed for the proposed metrics. They are bibliography ontologies revised and simplified from publicly available ones and are denoted as at $\mathcal{O}_m$[1] and $\mathcal{O}_p$[2] respectively.

3.1 Concept as a set of signatures

Each concept in an ontology encapsulates a subset of instance data from the domain of discourse. In a broader sense, concepts are effectively constraint systems against which instance data are evaluated. For instance, concept Book (defined as in Figure 1 using DL-based constructs) specifies that a book is a Document that has at least one title, at least one publisher, etc.

Unfolding concepts Semantics of concepts are embedded in the DL-based constructs which need to be explicated before computing the distance. Concepts are recursively unfolded till only primitive one—concepts that only defined by names—appears on the righthand side of the concept introduction axioms. If cyclic definitions are not allowed, i.e. such that no primitive concepts appear on both sides of a concept introduction axiom, it is possible to unfold the righthand side of all concept introduction axioms and guarantee the termination of such an

[1] http://visus.mit.edu/bibtex/0.01/bibtex.owl.
[2] http://www.aktors.org/ontology/portal.

$$\begin{aligned}
\mathsf{Book} &\doteq \mathsf{Document} \sqcap \geq_1 \mathsf{hasTitle} \sqcap \geq_1 \mathsf{hasYear} \\
&\quad \sqcap \geq_1 \mathsf{hasPublisher} \sqcap \geq_1 \mathsf{humanCreator.Author} \\
\mathsf{Author} &\doteq \mathsf{Human} \sqcap \geq_2 \mathsf{hasPublication.Document} \\
&\mathsf{Document} \sqsubseteq \top \quad \mathsf{Human} \sqsubseteq \top
\end{aligned}$$

Figure 1. Book in $\mathcal{O}_p$ and related concepts

unfolding process. For instance, let $\mathrm{CN} \doteq \mathsf{C}' \in \mathcal{O}$, CN_i and RN_j be concept and role names appearing in C' respectively, and $(\mathrm{CN}_i \doteq \mathsf{C}_i) \in \mathcal{O}$ and $(\mathrm{RN}_j \doteq \mathsf{R}_j) \in \mathcal{O}$. It is possible to thoroughly expand C' by recursively replacing defined concept names appearing on the righthand side of $\mathrm{CN} \doteq \mathsf{C}'$ with the concept definitions in $\mathcal{O}$, i.e. $\mathsf{C}\,[\,\mathrm{CN}_i/\mathsf{C}_i,\, \mathrm{RN}_j/\mathsf{R}_j\,]$ where $[\frac{x}{y}]$ defines the process of replacing all occurrences of x with y. Such a process terminates due to the acyclic nature of $\mathcal{O}$ and results in a finite set of logic formulae. Subsequently, semantic signatures are extracted from the unfolded concepts.

$\mathcal{S}$: a non-empty set of instances; $\mathcal{L}$: associating each $a \in \mathcal{S}$ with a set of concepts; $\mathcal{R}$: mapping each property to a subset of $\mathcal{S} \times \mathcal{S}$. For all a, $b \in \mathcal{S}$, if C, C_1, C_2 are concepts and R is property:

$\mathrm{r}_{\sqcap}$: $C_1 \sqcap C_2 \in \mathcal{L}(a)$, then $C_1 \in \mathcal{L}(a)$ and $C_2 \in \mathcal{L}(a)$.
$\mathrm{r}_{\sqcup}$: $C_1 \sqcup C_2 \in \mathcal{L}(a)$, then $C_1 \in \mathcal{L}(a)$ or $C_2 \in \mathcal{L}(a)$.
$\mathrm{r}_{\forall}$: $\forall R.C \in \mathcal{L}(a)$ and $\langle a, b\rangle \in \mathcal{R}(R)$, then $C \in \mathcal{L}(a)$.
$\mathrm{r}_{\exists}$: $\exists R.C \in \mathcal{L}(a)$, then $\exists b.b \neq a$ and $\langle a, b\rangle \in \mathcal{R}(R)$ and $C \in \mathcal{L}(b)$.
$\mathrm{r}_{\geq}$: $\geq_n R.C \in \mathcal{L}(a)$, then $\exists b_1, \ldots b_k.b_i \neq b_j$ and $\langle a, b_i\rangle \in \mathcal{R}(R)$ and $C \in \mathcal{L}(b_i)$ and $k \geq n$.
$\mathrm{r}_{\leq}$: $\leq_n R.C \in \mathcal{L}(a)$, then $\exists b_1, \ldots b_k.b_i \neq b_j$ and $\langle a, b_i\rangle \in \mathcal{R}(R)$ and $C \in \mathcal{L}(b_i)$ and $k \leq n$.

Figure 2. Transformation rules of some DL constructs [2]

We adopted the tableaux algorithms used in many DL-based inferential systems to facilitate the concept unfolding and signature extracting process. In Figure 3, we present an example of how concept Book (defined in Figure 1) is unfolded by repetitively applying the transformation rules defined for each and every DL constructs (see Figure 2 for the rules of some DL constructs)—a detailed description of such rules can be found in [2]. The unfolding process for Book stops when only primitive concepts and properties, namely Document and Human, remain. $\top$ is included for semantic completeness.

As illustrated in Figure 3, concept Book is associated with one set of semantic bearing signatures that fully capture the meaning of Book by means of primitive concepts and properties. There are two points to be further addressed. Firstly, there might be cases that concepts are defined with the union

$$
{}^{0}\mathfrak{C}_1^{\mathsf{Book}} = \left\{ \begin{array}{l} x : \mathsf{Document} \sqcap \geq_1 \mathsf{hasTitle} \sqcap \geq_1 \mathsf{hasYear} \sqcap \\ \quad \geq_1 \mathsf{hasPublisher} \sqcap \geq_1 \mathsf{humanCreator.Author} \end{array} \right\}
$$

$$
{}^{1}\mathfrak{C}_1^{\mathsf{Book}} = \left\{ \begin{array}{l} x : \mathsf{Document} \sqcap \geq_1 \mathsf{hasTitle} \sqcap \geq_1 \mathsf{hasYear} \sqcap \\ \quad \geq_1 \mathsf{hasPublisher} \sqcap \\ \quad \geq_1 \mathsf{humanCreator.}(\mathsf{Human} \sqcap \geq_2 \mathsf{hasPublication.Document}) \end{array} \right\}
$$

$$
{}^{2}\mathfrak{C}_1^{\mathsf{Book}} = \left\{ \begin{array}{l} x : \mathsf{Document},\ x :\geq_1 \mathsf{hasTitle},\ x :\geq_1 \mathsf{hasYear}, \\ x :\geq_1 \mathsf{hasPublisher}, \\ x :\geq_1 \mathsf{humanCreator.}(\mathsf{Human} \sqcap \geq_2 \mathsf{hasPublication.Document}) \end{array} \right\}
$$

$$
{}^{3}\mathfrak{C}_1^{\mathsf{Book}} = \left\{ \begin{array}{l} x : \mathsf{Document},\ \langle x, y_0 \rangle : \mathsf{hasTitle},\ \langle x, y_1 \rangle : \mathsf{hasYear}, \\ \langle x, y_2 \rangle : \mathsf{hasPublisher},\ \langle x, y_4 \rangle : \mathsf{humanCreator}, \\ y_4 : \mathsf{Human} \sqcap \geq_2 \mathsf{hasPublication.Document} \end{array} \right\}
$$

$$
{}^{4}\mathfrak{C}_1^{\mathsf{Book}} = \left\{ \begin{array}{l} x : \mathsf{Document},\ \langle x, y_0 \rangle : \mathsf{hasTitle},\ \langle x, y_1 \rangle : \mathsf{hasYear}, \\ \langle x, y_2 \rangle : \mathsf{hasPublisher},\ \langle x, y_4 \rangle : \mathsf{humanCreator}, \\ y_4 : \mathsf{Human},\ \langle y_4, z_0 \rangle : \mathsf{hasPublication.Document} \\ \langle y_4, z_1 \rangle : \mathsf{hasPublication.Document} \end{array} \right\}
$$

$$
{}^{5}\mathfrak{C}_1^{\mathsf{Book}} = \left\{ \begin{array}{l} x : \mathsf{Document},\ \langle x, y_0 \rangle : \mathsf{hasTitle},\ \langle x, y_1 \rangle : \mathsf{hasYear}, \\ \langle x, y_2 \rangle : \mathsf{hasPublisher},\ \langle x, y_4 \rangle : \mathsf{humanCreator}, \\ y_4 : \mathsf{Human},\ \langle y_4, z_0 \rangle : \mathsf{hasPublication},\ z_0 : \mathsf{Document} \\ \langle y_4, z_1 \rangle : \mathsf{hasPublication},\ z_1 : \mathsf{Document},\ x : \top \end{array} \right\}
$$

Figure 3. Unfolding concept Book in $\mathcal{O}_m$

of other concepts that are either fully defined elsewhere in the same ontology or introduced as anonymous ones. Applying indeterminate $\sqcup$ unfolding rules (see Figure 2) results in alternatives sets of formulae each of which captures part of the intended meaning of the original concept. For instance, if we have "$\mathsf{Human} \doteq \mathsf{Man} \sqcup \mathsf{Woman}$" and Man and Woman as "$\ldots \sqcap \forall \mathsf{hasGenderMale} \sqcap \ldots$" and "$\ldots \sqcap \forall \mathsf{hasGenderFemale} \sqcap \ldots$" respectively. After unfolding, we have two separate sets of signatures.

$$
{}^{i}\mathfrak{C}_1^{\mathsf{Human}} = \{\ldots,\ x : \forall \mathsf{hasGender.Male},\ \ldots\} \text{ or}
$$
$$
{}^{i}\mathfrak{C}_2^{\mathsf{Human}} = \{\ldots,\ x : \forall \mathsf{hasGender.Female},\ \ldots\}
$$

Secondly, property universal quantifications can only be further expanded when there are instances defined over the property, i.e. $y : \mathsf{Male}$ is included, in the above example, if and only if there are $x : \forall \mathsf{hasGender.Male}$ and $\langle x, y \rangle : \mathsf{hasGender}$. They are left unexpanded, otherwise.

The unfolding process stops till a fix point is reached, i.e. ${}^{n}\mathfrak{C} = {}^{(n-1)}\mathfrak{C}$. As demonstrated by Donini and colleagues [6], by carefully selecting a subset of admitted conceptual constructs, e.g. the underlying logic models of OWL-Lite and OWL-DL [12], unfolding can be performed, in the worst case, as ExpTime-complete problems and a termination is guaranteed with respect to acyclic ontologies.

Weighting signatures Unfolding concepts can be seen as a process that gradually makes the semantics (the intend meanings of concepts) explicit. As a result, each concept is associated with finite sets of signatures, being them the primitive concepts and properties. Effectively, concepts are deemed to hold parts of the information of the domain of discourse and thus, albeit the apparent difference between ontologies and documents in general sense, techniques extracting and weighting document surrogates in IR can be analogised to concepts.

A straightforward approach to evaluate the influence of semantic signatures is to count the number of their occurrence in each $\mathfrak{C}_i$ of C. Negative construct, $\neg$, states that the target concept is explicitly excluded and thus value -1 is given to emphasis the semantics. Unexpanded universal quantification, e.g. $\forall \mathsf{R}.\mathsf{B}$, is treated as atomic signatures, as the presence of B is uncertain due to the absence of property R. Comparing to fully defined concepts, in many ontologies, the number of primitive concepts and properties are small. Hence, we do not expect to encounter sparse vectors very often.

As examples, concepts Phdthesis and Mastersthesis (see Figure 4(a)) from $\mathcal{O}_m$ are unfolded as illustrated in Figure 4(b). Their signature vectors and that of concept Book are presented in Table 2, where equal weights are assigned to every signature.

$$\begin{aligned}\mathsf{Phdthesis} \doteq\ & \mathsf{Document} \sqcap \geq_1 \mathsf{hasAuthor} \sqcap \geq_1 \mathsf{hasTitle} \sqcap \\ & \geq_1 \mathsf{hasSchool} \sqcap \geq_1 \mathsf{hasYear} \\ \mathsf{Mastersthesis} \doteq\ & \mathsf{Document} \sqcap \geq_1 \mathsf{hasAuthor} \sqcap \geq_1 \mathsf{hasTitle} \sqcap \\ & \geq_1 \mathsf{hasSchool} \sqcap \geq_1 \mathsf{hasYear}\end{aligned}$$

(a) Definition of thesis concepts

$$\begin{aligned}{}^{n}\mathfrak{C}_1^{\mathsf{Phdthesis}} = \quad & x : \mathsf{Document},\ \langle x, y_0\rangle : \mathsf{hasAuthor},\ \langle x, y_1\rangle : \mathsf{hasTitle}, \\ & \langle x, y_2\rangle : \mathsf{hasSchool},\ \langle x, y_3\rangle : \mathsf{hasYear},\ x : \top \\ {}^{n}\mathfrak{C}_1^{\mathsf{Mastersthesis}} = \quad & x : \mathsf{Document},\ \langle x, y_0\rangle : \mathsf{hasAuthor},\ \langle x, y_1\rangle : \mathsf{hasTitle}, \\ & \langle x, y_2\rangle : \mathsf{hasSchool},\ \langle x, y_3\rangle : \mathsf{hasYear},\ x : \top\end{aligned}$$

(b) Unfolded thesis concepts

Figure 4. Thesis concepts in $\mathcal{O}_m$

Weights of signatures are fine-tuned 1) using the *inverse document frequency weight* (*idf*) [10] scheme from IR with the assumption that signatures appearing in a small number of concepts are more significant than those that are frequently referred to by many concepts and 2) by reducing the weights of signatures referred to indirectly through properties.

Let N be the number of concepts in ontology $\mathcal{O}$, n_{f_k} the number of concepts that refer to signature f_k, and $\mathrm{f}_{f_k,\mathsf{C}_i}$ the frequency of signature f_k in concept C_i,

	$\mathfrak{C}_1^{\mathsf{Book}}$	$\mathfrak{C}_1^{\mathsf{Phdthesis}}$	$\mathfrak{C}_1^{\mathsf{Mastersthesis}}$
$\top$ (top)	1	1	1
Document	3	1	1
Human	1	0	0
hasAuthor	0	1	1
hasPublisher	1	0	0
hasPublication	2	0	0
hasTitle	1	1	1
humanCreator	1	0	0
hasSchool	0	1	1
hasYear	1	1	1

Table 2. Signature vector space of Book, Phdthesis, and Mastersthesis

the *tf-idf* weight, w_{f_k,C_i}, of signature f_k in concept C_i is computed as

$$w_{f_k,\mathsf{C}_i} = \mathrm{f}_{f_k,\mathsf{C}_i} \times (\log_2 N/n_{f_k} + 1)\,, \text{ where } n_{f_k} \neq 0.$$

In $\mathcal{O}_m$, such signatures as Document, hasTitle, and hasYear appear in most of the concepts and thus are assigned low weights whereas humanCreator appears in only one concepts and thus is regarded more important than others. Weights of indirect signatures are adjusted based on the weights of their related properties. For instance, $z_0 : \mathsf{Document}$ in Figure 3 is introduced because of $\mathsf{humanCreator} \circ \mathsf{hasPublication}$ and thus has less influence than $x : \mathsf{Document}$. We decrease the weight of $z_0 : \mathsf{Document}$ to $w_{\mathsf{Document}} \cdot w_{\mathsf{humanCreator}} \cdot w_{\mathsf{hasPublication}}$.

Computing distances By representing concepts as signature vectors, distances between concepts then equals to the distances between vectors in a high dimensional space. Distance based on the vector space model can be regarded as metric as it satisfies the three axioms discussed in Section 2.1. When there are more than one resultant $\mathfrak{C}_i$, the shortest distance is computed.

$$\Delta\,(\mathsf{C},\mathsf{D}) = \min_{(\mathfrak{C}_i \text{ of } \mathsf{C}, \mathfrak{C}'_j \text{ of } \mathsf{D})} \tau(\mathtt{sim}\,(\mathfrak{C}_i, \mathfrak{C}'_j)) \tag{3}$$

$$\mathtt{sim}\,(\mathsf{C},\mathsf{D}) = \max_{(\mathfrak{C}_i \text{ of } \mathsf{C}, \mathfrak{C}'_j \text{ of } \mathsf{D})} \mathtt{sim}\,(\mathfrak{C}_i, \mathfrak{C}'_j) \tag{4}$$

$$\tau(\mathtt{sim}\,(\mathfrak{C}_i, \mathfrak{C}'_j)) = \begin{cases} -\log\left(\mathtt{sim}\,(\mathfrak{C}_i, \mathfrak{C}'_j)\right) & \text{if } \mathtt{sim}\,(\mathfrak{C}_i, \mathfrak{C}'_j) > 0 \\ +\infty & \text{if } \mathtt{sim}\,(\mathfrak{C}_i, \mathfrak{C}'_j) \leq 0 \end{cases} \tag{5}$$

$$\mathtt{sim}\,(\mathfrak{C}, \mathfrak{C}') = \frac{\sum\limits_{w_i \in \mathfrak{C},\ w'_i \in \mathfrak{C}'} w_i \times w'_i}{\sqrt{\sum\limits_{w_i \in \mathfrak{C}} w_i^2}\sqrt{\sum\limits_{w'_i \in \mathfrak{C}'} w_i'^2}} \tag{6}$$

Due to the introduction of negative numbers for capturing the semantics of $\neg$, there are possibilities of non-positive similarities based on Equation 6. $+\infty$, therefore, represents a pair of totally divergent concepts.

After taking into account the weighting factors, signature vectors in Table 2 are refined and we approximate the distances among concepts as

$$\Delta(\mathsf{Book}, \mathsf{Phdthesis}) = -\log(\mathtt{sim}(\mathsf{Book}, \mathsf{Phdthesis})) \approx 2.101$$
$$\Delta(\mathsf{Book}, \mathsf{Mastersthesis}) = -\log(\mathtt{sim}(\mathsf{Book}, \mathsf{Mastersthesis})) \approx 2.101$$
$$\Delta(\mathsf{Phdthesis}, \mathsf{Mastersthesis}) = -\log(\mathtt{sim}(\mathsf{Phdthesis}, \mathsf{Mastersthesis})) \approx 0$$

It demonstrates that, by definition, the distance between the two types of theses is "shorter" than that between theses and book. Such a conclusion is evident if we consider properties as restrictions defined over concepts that screen out unqualified instances from the domain of discourse. Book requires at least two hasPublication. Intuitively, it presents a stronger constraint than those that do not have number restrictions on hasPublication property and thus there might be fewer instances satisfying all its restrictions. The zero distance between two types of thesis also suggests that these two concepts might not be properly defined in that they are identical from the given signatures.

Discussion We see our distance metrics have the following advantages. Anonymous concepts, also known as restrictions, have always been the "trouble maker" in graph-based and feature-based approaches. When unfolding concepts, we expand restrictions together with other defined concepts, e.g. $x : \exists \mathsf{R}.\mathsf{C}$ is replaced by $\langle x, y\rangle : \mathsf{R}$ and $y : \mathsf{C}$. Anonymous concepts are, therefore, replaced by semantic bearing signatures that explicitly state the constraints imposed on the instances. We further collapse identical signatures so as to reduce the space complexity. Meanwhile, albeit the apparent similarity, transforming ontology into graphs cannot preserve the semantics *a coup sur*. Even with labelled edges, graph-based methods always have the difficulty in justifying the semantic significance of transitive properties. For instance, it takes the distance between A and C in $A \rightarrow B \rightarrow C$ to be greater than that in $A \rightarrow C$ due to the introduction of the interim node B. This is intuitively incorrect and can be avoided if we fully unfold the interim concept B to the most basic signatures as well. Furthermore, many feature-based approaches adopt a weighting schema to distinguish the contributions from different features, weights of which are normally set up manually by domain experts. We are not to disparage the importance of human experts' role in understanding semantics. We, nevertheless, would like to introduce an automatic weighting mechanism to be complementary to their efforts. The *tf-idf* schema borrowed from IR proposes a weight for each semantic bearing signature based on the significance of such a signature in introducing semantic discrepancies and thus fit perfectly with the distance metrics. Finally, we consider our metric as an improvement of techniques from feature-based families with well-founded mathematic models. This is evident partially from the fact that when constructing overall similarity/dissimilarity as weighted average, feature-based

approaches assume the semantic homogeneity of different features, which is not necessarily true.

4 Extending semantic metrics of concepts

In this section, we demonstrate how to generalise the semantic metric discussed in previous sections to other ontology related measurements. Our work is laid on the argument that the distances between concepts offers a fertile ground from which other metrics—that are effectively aggregations of concept-based distances—can stem.

4.1 Distance of concepts from different ontologies

Computation of $\Delta(\mathsf{C}, \mathsf{C}')$, where C and C' belong to different ontologies, needs to be bootstrapped by the similarity between primitive concepts and properties from respective ontologies. Ontology Mapping/Alignment techniques have been extensively studied recently and many tools have been developed to automatically or semi-automatically map ontological entities [7,9]. When bootstrapping $\Delta(\mathsf{C}, \mathsf{C}')$, we require only the similarities between primitive concepts and properties and thus simple string distance algorithms and/or those enhanced by external general-purposed lexicons, e.g. WordNet [14], are sufficient.

The similarity function (Equation 6) is adjusted to reflect the similarities computed by ontology mapping algorithms. Let w_i and w_i' be the weights of signatures f_i and f_i' from $\mathcal{O}$ and $\mathcal{O}'$ respectively, being them the primitive concepts, properties, and universal quantifications, C and C' concepts from $\mathcal{O}$ and $\mathcal{O}'$, and f_i' the most similar signature of f_i with $\delta_i = \mathtt{sim}(f_i, f_i')$,

$$\mathtt{sim}(\mathfrak{C}, \mathfrak{C}') = \frac{\sum\limits_{w_i \in \mathfrak{C},\ w_i' \in \mathfrak{C}'} \delta_i w_i \times \delta_i w_i'}{\sqrt{\sum\limits_{w_i \in \mathfrak{C}} (\delta_i w_i)^2} \sqrt{\sum\limits_{w_i' \in \mathfrak{C}'} (\delta_i w_i')^2}} \tag{7}$$

Note that if a signature does not have correspondence, its weight is set to 0.

In our example, concept Book (see Figure 5(a)) from $\mathcal{O}_p$ is unfolded as illustrated in Figure 5(b). With the initial correspondences between primitive concepts (e.g. Reference versus Document) and properties (e.g. hasPublisher versus published-by) from respective ontologies, which might be provided by an automatic mapping system or hand-crafted by human experts, we computed the distance between the two book concepts to be approximately 7.15. By definition, apparent close concepts $\mathsf{Book} \in \mathcal{O}_m$ (denoted as Book_m) and $\mathsf{Book} \in \mathcal{O}_p$ (denoted as Book_p) are effectively semantically different. The absolute positive distance value between these two concepts indicates a semantic divergence which is evident in the fact that Book_m requires all books to have title, published year, publisher, etc. while such are not mandatory for Book_p—an instance does not need to have title, author, date, etc. to be qualified as a Book in ontology $\mathcal{O}_p$.

$$
\begin{aligned}
\mathsf{Book} &\doteq \mathsf{Publication} \sqcap \forall \mathsf{published\text{-}by.Organization} \\
\mathsf{Publication} &\doteq \mathsf{Reference} \sqcap \forall \mathsf{has\text{-}author.Person} \sqcap \forall \mathsf{has\text{-}date.Calendar\text{-}Date} \sqcap \\
&\quad \forall \mathsf{has\text{-}place\text{-}of\text{-}pub.Location} \\
\mathsf{Reference} &\sqsubseteq \top \quad \mathsf{Location} \sqsubseteq \top \quad \mathsf{Calendar\text{-}Date} \sqsubseteq \top \\
&\quad \mathsf{Organization} \sqsubseteq \top
\end{aligned}
$$

(a) Definition of Book and related concepts

$$
{}^{n}\mathfrak{C}_1^{\mathsf{Book}} = \quad x : \mathsf{Reference},\ x : \forall \mathsf{has\text{-}author.Person},\ x : \forall \mathsf{has\text{-}date.Calendar\text{-}Date},\ x : \forall \mathsf{has\text{-}place\text{-}of\text{-}pub.Location},\ x : \forall \mathsf{published\text{-}by.Organization},\ x : \top
$$

(b) Unfolded concept Book

Figure 5. Book and related concepts in $\mathcal{O}_p$

4.2 Distance between a concept and a set of concepts

There are occasions that the closeness is sought after between a concept on one hand and a set of interrelated concepts as a group on the other hand. For instance, one might need a measurement to represent how dense an ontology is with regard to an arbitrary concept. Let $\mathsf{C} \in \mathcal{O}$ be the target concept, $\mathsf{D} \in \mathcal{O}$ a concept from $\mathcal{O}$ that does not equal to C, Equation 2 can be rewritten as

$$
\Delta(\mathsf{C}, \mathcal{O}) = -\sum_{\mathsf{D} \in \mathcal{O},\ \mathsf{D} \neq \mathsf{C}} p(\mathsf{D} \mid \mathsf{C}) \log p(\mathsf{D} \mid \mathsf{C}) \tag{8}
$$

If we take $p(\mathsf{D} \mid \mathsf{C})$ as $\mathtt{sim}(\mathsf{C}, \mathsf{D})$ obtained using Equation 4, we can then approximate the closeness of ontology $\mathcal{O}$ around C by aggregating the distances between C and every other concept in $\mathcal{O}$.

4.3 Distance between ontologies

As laid down in Section 2, we reckon ontologies as a construction of concepts and thus the distance of ontologies is a function of those between concepts from respective ontologies. In the paper, several methods are considered to aggregate individual distances.

Summation of feature distances The *city block distance*—the sum of the distances between individual signatures—is the simplest aggregation function. Based on Equation 3 and 7, we define

$$
\Delta(\mathcal{O}, \mathcal{O}') = \left(\sum_{\mathsf{C}_i \in \mathcal{O}} \left(\min_{\mathsf{C}'_j \in \mathcal{O}'} \Delta(\mathsf{C}_i, \mathsf{C}'_j) \right)^{\lambda} \right)^{1/\lambda} \tag{9}
$$

where λ might take the value of the number of concepts in $\mathcal{O}$ in which case the distance measure is not symmetric.

The disadvantage of Minkowski style distance function is that if the distance between an arbitrary pair of signatures is significantly larger or smaller than that of others, the aggregated result might be falsely amplified or diminished.

Kullback-Leibler (KL) model Also known as *relative entropy*, KL divergence is a natural quasi-distance of the extent to which one distribution agrees with another. In order to overcome the asymmetric characteristic intrinsic to KL divergence, Jeffrey-divergence was proposed. Let $\mathsf{C} \in \mathcal{O}$ and $\mathsf{C}' \in \mathcal{O}'$ be two concepts from respective ontologies, distance between ontologies is computed as

$$\Delta_J(\mathcal{O}, \mathcal{O}') = \sum_i p(\mathsf{C}_i) \log \frac{p(\mathsf{C}_i)}{p(\mathsf{C}'_i)} + \sum_i p(\mathsf{C}'_i) \log \frac{p(\mathsf{C}'_i)}{p(\mathsf{C}_i)}$$

Ontology is effectively a constraint system specifying how instances should be distributed among different concepts. In an arbitrary domain of discourse, the more rigorous the restrictions are, the fewer instances are qualified to instantiate a particular concept. If we define an imaginary "perfect" concept, C_0, as one imposed with no restrictions except the domain top, e.g. $\langle$`owl:Thing`$\rangle$. Assume, the rigorousness of C_0 is 0. We than can compute the distance from an arbitrary "imperfect" concept C_i to C_0 as $\Delta(\mathsf{C}_i)$. The probability distribution of C_i can, therefore, be approximated as

$$p(\mathsf{C}_i) = 1 - \frac{\Delta(\mathsf{C}_i)}{\sum_{j=0}^{n} \Delta(\mathsf{C}_j)} \tag{10}$$

Asymmetric distance measure Variants of KL divergency are established on the assumption that the ontologies are defined over largely overlapped domains and thus distances can be estimated by examining the distributions of "imaginary" instances. When such a prerequisite cannot be assumed, i.e. one does not have *a priori* knowledge of the interpretation domains of ontologies, distance is ought to be obtained from mappings between fundamental semantic bearing signatures and is deemed an aggregation of those computed using Equation 8:

$$\Delta_A(\mathcal{O}, \mathcal{O}') = -\sum_{\mathsf{C} \in \mathcal{O}} p(\mathsf{C}) \sum_{\mathsf{D} \in \mathcal{O}'} p(\mathsf{C} \mid \mathsf{D}) \log p(\mathsf{C} \mid \mathsf{D})$$

where $p(\mathsf{C} \mid \mathsf{D})$ is the similarity based on Equation 4 and Equation 7 and $p(\mathsf{C})$ as in Equation 10. Note that Δ_A is asymmetric, i.e. $\Delta_A(\mathcal{O}, \mathcal{O}') \neq \Delta_A(\mathcal{O}', \mathcal{O})$.

5 Discussion and Conclusions

The increasing interest on employing rigorous logics to underpin ontology modelling languages has presented itself as a challenge on several ontology management tasks. In such circumstances, as "meanings" are emphasised, it is not straightforward to identify the similarity/dissimilarity between concepts, which

should be a function of both syntactical and semantical divergences. In this paper, we demonstrated how the tableau-based algorithm of DLs can benefit distance measures among concepts, between a single concept and a group of concepts, and how to generalise it so as to compute distance between ontologies. The proposed semantic measures/metrics can be complementary to other metrics. Compared to traditional approaches, however, a DL-based one is capable of conveying not only the syntactic but also semantic information.

We envisage several applications of our distance measures/metrics in the context of semantic-enriched applications:

Ontology segmentation: An obvious application of the distance measures is ontology segmentation. With the growing interest in tackling interoperability issues, ontologies have quickly become a convenient vehicle for domain knowledge. Extensive efforts from different communities results in many enormous knowledge corpora, especially in medicine, e.g. FMA [16] and GALEN [15]. The sheer size of such ontologies has put a tremendous burden on ontology management tools and thus becomes the major obstacle hesitating people who seek only a small part of the knowledge encapsulated in such ontologies. Ontology segmentation is envisaged as a neat solution to cope with the size issue. In a recent paper [17], the authors extract a semantically complete part of an ontology by traversing upwards and downwards along *links*—concept inheritance relationships and properties—with the guidance of heuristic rules. Other approaches include graph-based clustering, query-based partitioning, etc. It is our contention that fragmenting an ontology is tantamount to computing semantic distance between concepts. The success of a segmentation strategy, therefore, depends directly on a good metric. As a complementary method of the existing segmentation techniques, our distance measures detect the semantic disagreement of different concepts and thus presents a criteria against which concepts can be filtered in/out. For instance, if one would like to extract a set of concepts around C, the segmentation can be formalised as $\texttt{segmentation}(\mathcal{O}, \mathsf{C}) = \{\mathsf{D} \mid \forall\, \mathsf{D} \in \mathcal{O}.\Delta(\mathsf{C}, \mathsf{D}) \leq d\}$ where d is an arbitrary real number.

Ontology ranking: Building ontology is a time-consuming, error-prone process that requires trained eyes and minds. The Web has made such a task easier by offering search-and-access functionality to various on-line ontology repositories [1]. A search engines normally returns a list of candidates ranked according to a predefined ordering schema. Ranking resultant ontologies of a search query is effectively finding the closeness of a group of concepts w.r.t. those specified in the query. From discussions in Section 4.3, we have

$$\Delta(Q, \mathcal{O}) = -\sum_{\mathsf{C} \in Q} \left(p(\mathsf{C}) \sum_{\mathsf{D} \in \mathcal{O}} \texttt{sim}(\mathsf{D}, \mathsf{C}) \log(\texttt{sim}(\mathsf{D}, \mathsf{C})) \right)$$

Note that queries might be fragments of ontologies and thus cannot be fully unfolded. $\Delta(Q, \mathcal{O})$, therefore, might vary depending on the semantic completeness of queries and the initial similarities of respective semantic bearing signatures. $p(\mathsf{C})$ can be assigned manually by people submitting queries. As a default behaviour of querying, we assume people have some knowledge of the queries that they

are asking, are able to justify the relative significance of different parts of the queries, and can express the relative significance using numeric values. Having obtained the distances between Q and $\mathcal{O}_i$ from the candidate list, $\mathcal{O}_1, \ldots, \mathcal{O}_n$, one can then rank the resultant ontologies by comparing their numeric distance values, e.g. ranking ontologies with smaller $\Delta(Q, \mathcal{O})$ closer to the top of the list.

Ontology mapping: Ontology mapping is a complex and necessary task for most of Semantic Web's applications. The perspective users of such technology are faced with a number of challenges including ambiguity of the meaning of mappings, difficulties of capturing semantics, verification and validation of results and operationalisation in the beneficiary Semantic Web application. The approach proposed in Section 4.1 provides a clear and straightforward metric for measuring the semantic discrepancy between concepts from different ontologies. An intuitive method is to nominate for a concept C from $\mathcal{O}_1$ a concept D_i from $\mathcal{O}_2$ that minimise the distance $\Delta(\mathsf{C}, \mathsf{D}_i)$.

Semantic metrics can be further improved. Firstly, universal quantification, thus far, is regarded as an atomic signature. Although it is semantically coherent, this approach might increase the size of signature corpus in practice. Possible solution could be to consider $\forall \mathsf{R}.\mathsf{C}$ as a complex signature whose weight is the product of w_{R} and w_{C}. The appropriateness of such a weighting scheme, nevertheless, need further evaluation. Secondly, the complement (negation) construct results in a -1 count of the corresponding signature to differentiate it from missing signatures. It increases the possibility of similarities with negative numeric values. Currently, we equally assume that a pair of concepts having negative similarity do not overlap and thus are far apart from each other. We, however, do not distinguish cases with smaller negative similarity values from those with larger ones. The subtle differences between negative similarities might be necessary to answer such questions as "*are the distance of* $\mathsf{C} \sqcap \mathsf{D}$ *and* $\mathsf{C} \sqcap \neg\mathsf{D}$ *and the distance between* $\mathsf{C} \sqcap \mathsf{D} \sqcap \mathsf{E}$ *and* $\mathsf{C} \sqcap \neg\mathsf{D} \sqcap \neg\mathsf{E}$ *the same?*" Although an answer can be found indirectly by comparing similarities, a neat treatment is necessary. Finally, the use of two bibliography ontologies is only to demonstrate the applicability of semantic metrics. More empirical evaluation and a comprehensive comparative study against other approaches will further reveal the strength and weak points of our approach.

Acknowledgements

This work is supported under the OpenKnowledge and HealthAgents funded by EU Framework 6 and the Advanced Knowledge Technologies (AKT) IRC funded by EPSRC, UK.

References

1. H. Alani and C. Brewster. Ontology ranking based on the analysis of concept structures. In *K-CAP '05: Proceedings of the 3rd international conference on Knowledge capture*, pages 51–58. ACM Press, 2005.

2. F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, 2003.
3. T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, pages 28–37, 2001.
4. T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Series in Telecommunications. Wiley, 1991.
5. A.H. Doan, P. Domingos, and A.Y. Halevy. Reconciling schemas of disparate data sources: A machine-learning approach. In *SIGMOD Conference*, 2001.
6. F. M. Donini and Fabio Massacci. EXPTIME tableaux for $\mathcal{ALC}$. *Journal of Artificial Intelligence*, 124:87–138, 2000.
7. M. Ehrig, J. de Bruijn, D. Manov, and F. Martin-Recuerda. State-of-the-art survey on Ontology Merging and Aligning V1. Technical Report Deliverable 4.2.1, Institut AIFB, Universität Karlsruhe, July 2004.
8. T. Gruber. A translation Approach to Portable Ontology Specification. *Knowledge Acquisition*, 5(2):199–221, 1993.
9. Y. Kalfoglou, B. Hu, D. Reynolds, and N. Shadbolt. Semantic integration technologies. 6th month deliverable, University of Southampton and HP Labs, April 2005.
10. R. Korfhage. *Information storage and retrieval*. Wiley Computer Publishing, 1997.
11. O. Lassila and R.R. Swick. *Resource Description Framework (RDF) Model and Syntax Specification*. W3C, 1999.
12. D. L. McGuinness and F. van Harmelen. *OWL Web Ontology Language Overview*. W3C, 2003.
13. S. Melnik, H. Garcia-Molina, and E. Rahm. Similarity Flooding: A Versatile Graph Matching Algorithm and its Application to Schema Matching. In *Porceedings of the 18th International Conference on Data Engineering (ICDE)*, pages 117–128, 2002.
14. G. A. Miller. WordNet; a Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995.
15. A. Rector and J. Rogers. Ontological Issues in using a Description Logic to Represent Medical Concepts: Experience from GALEN. In *Proceedings of IMIA WG6 Workshop*, 1999.
16. C. Rosse and José L. V.. Jr. Mejino. A reference ontology for biomedical informatics: the foundational model of anatomy. *J. of Biomedical Informatics*, 36(6):478–500, 2003.
17. J. Seidenberg and A. Rector. Web ontology segmentation: Analysis, classification and use. In *Proceedings of WWW2006*, 2006. to appear.
18. C.E. Shannon. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, 2001.
19. J.F. Sowa. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks/Cole, Thomson Learning, 2000. ISBN 0-534-94965-7.
20. F. Wiesman and N. Roos. Domain independent learning of ontology mappings. In *AAMAS*, pages 846–853, 2004.