Semantic Metrics

Bo Hu, Yannis Kalfoglou, Harith Alani,
David Dupplaw, Paul Lewis, and Nigel Shadbolt

IAM Group, ECS, University of Southampton, SO17 1BJ, UK
{bh, y.kalfoglou, ha, dpd, phl, 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 reduced 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 on the usability of our ideas in potential areas.

1 Introduction

We are currently witnessing a shift of participation in ontology authoring from knowledge engineers to interested practitioners. This change is fueled, partly, by ever growing interest in the Semantic Web and in 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 roots" approach to ontology engineering has put the onus on ontology management and calls for a variety of new tasks, such as ontology ranking, segmentation and evaluation, to name but a few. A common ingredient to accomplish these tasks is the assessment of similarity/dissimilarity between concepts within ontologies or between entire 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 between concepts. The very fact that there are so many options indicates that reaching a consensus on how to capture semantics embedded in ontologies is hard to achieve in the first place. We are particularly interested in building upon all the work from different disciplines and focusing on metrics leveraging the semantics of concepts.

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

S. Staab and V. Svatek (Eds.): EKAW 2006, LNAI 4248, pp. 166–181, 2006.
© Springer-Verlag Berlin Heidelberg 2006

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 of working with and enhancing existing technology and we conclude the paper with several issues that need further investigation.

2 Background

In this section, we review the meanings of distances in different disciplines from which our semantic metrics are drawn. We restrict our focus on ontology languages whose underlying logic satisfies the *Beth property*, e.g. OWL-DL.

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 *distance* has been used in various domains and situations ranging from geometry and physics to information theory. An orthodox distance function must be **non-negative** and **symmetric** 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 the *City Block (Manhattan) Distance*, the *Euclidean Distance*, or the *Chebyshev Distance*. Analogous to the two dimensional space distance, the *Euclidean Distance* is generalised in an m dimensional space to *Minkowski Distance*, $\Delta_{\mathrm{Min}}(p, q) = (\sum_i \mid p_i - q_i \mid^m)^{1/m}$.

The idea of distance, in the broader sense of measuring how far apart two objects are, has been applied to the computation of the discrepancy between documents in Information Retrieval (IR), disagreement between words in a lexical taxonomy in Knowledge Representation, and dissimilarity between strings in Information Theory. The semantic metrics that we propose in this paper stem from the general distance measures that are discussed as follows.

The vector space model (VSM) [19] 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 proceeds 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, [19]) based on which, the documents can be sorted.

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

In 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 [21]. A full account of Shannon's view of the mathematical theory of information, however, is beyond the scope of this paper. We restrict our focus to 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 by *conditional entropy* which measures how much uncertainty a variable Y has, if the knowledge regarding another variable X is completely known. Representing *conditional entropy* as $H(Y \mid X)$, it may be defined as

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

In practice *conditional entropy* can be regarded as a divergence measure between two variables, where $H(X \mid X) = 0$. The larger the conditional entropy, the less information one gains from X with regard to Y and the further apart X and Y are.

Meanwhile, in a discrete domain, the Kullback-Leibler divergence measures the disagreement of two distributions. Let p and q be discrete distributions of a variable, the "distance" between p and q is computed as

$$\Delta_{KL}(p,q) = \sum_i p_{\ i} \log \left(\frac{p_{\ i}}{q_{\ i}} \right)$$

Note that the Kullback-Leibler divergence is not symmetric and is positive definite [5]. It 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 with answers ranging from simple taxonomies to logically sound and coherent constructs whose underlying model supports logic inferences [2]. In order to discuss distances with regard to ontological entities and with regard to ontologies themselves, we need first to clarify our intuitions about ontologies. Instead of giving a full philosophical reflection on the term *ontology*, we take the Artificial Intelligence (AI) approach and restrict an ontology to be "a specification of a conceptualisation" [8]. Although the fact that many models, e.g. database schemata, UML models, and Semantic Network models [22], can be considered ontologies in a broader sense, we normally confine our view of *conceptualisation* to the following formalisation:

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

Description Logics (DLs) are a family of knowledge representation and reasoning formalisms that have attracted substantial research recently, especially after the endorsement of DL-based ontology modelling languages (e.g. OWL [14]) by the Semantic Web initiative [3]. Among the three "sub-species" of OWL, OWL-Lite is based on $\mathcal{SHIF}$ DL and OWL-DL is based on $\mathcal{SHOIN}$ DL [9]. 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. Let CN denote a concept name, C and D be arbitrary concepts, R be a property, n be a non-negative integer, o_i $(1 \leq i \leq n)$ be an instance and $\top$, $\bot$ denote the top and the bottom. A $\mathcal{SHOIN}$ concept is:

$$CN \mid \mathsf{C} \sqcap \mathsf{D} \mid \mathsf{C} \sqcup \mathsf{D} \mid \neg\mathsf{C} \mid \exists\mathsf{R}.\mathsf{C} \mid \forall\mathsf{R}.\mathsf{C} \mid \geq_n \mathsf{R}.\top \mid \leq_n \mathsf{R}.\top \mid \{o_1, \ldots, o_n\}$$

Meanwhile, $\mathcal{SHOIQ}$ extends $\mathcal{SHOIN}$ with qualified number restrictions, $\geq_n \mathsf{R}.\mathsf{C}$ and $\leq_n \mathsf{R}.\mathsf{C}$.

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 mapping each property (role) to a subset of $\mathfrak{D}^{\mathcal{I}} \times \mathfrak{D}^{\mathcal{I}}$. 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) [12], OWL [14], etc. Despite the apparent differences, many of the 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 between concepts is by no means a new idea. It can be approached from 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 [6] and [23]. Intensional approaches exploit features defined directly over the concepts and apply measures such as Tversky's model (e.g. in [4]) and graph-based ones (e.g. in [15]). More specifically, graph-based methods represent ontologies as directed acyclic graphs and count the total number of weighted edges, where the edges could be inheritance relationships and/or properties. Feature-based methods characterise concepts with discrete semantics bearing components, e.g. concept names, property names, domains, etc. and take a weighted average of the similarity/dissimilarity between each pair of components [13]. Both extensional and intensional methods have advantages and disadvantages. On one hand, it may be argued that instance data can best capture the semantics and there are plenty of well studied techniques that can be leveraged. In reality, however, an unbiased population

is not always available, especially for ontologies published on the loosely regulated Web. The applicability of such approaches, therefore, is highly suspect. On the other hand, the intensional approaches would probably not win the battle due to: 1) the ambiguity of converting semantic distinctions—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 logical restrictions of concepts and the distances of concepts are reduced to the distances between different vectors of such semantics bearing signatures.

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

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

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

Unfolding concepts. Semantics of concepts are embedded in DL-based constructs which need to be explicated before computing the distance. Concepts are recursively unfolded till only primitive ones (i.e. concepts that are only defined by names) appear 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 unfolding process. For instance, let $\mathrm{CN} \doteq \mathsf{C}' \in \mathcal{O}$, CN_i and RN_j be concept and property names appearing in C' respectively, and $(\mathrm{CN}_i \doteq \mathsf{C}_i) \in \mathcal{O}$ and $(\mathrm{RN}_j = \mathsf{R}_j) \in \mathcal{O}$. It is possible to thoroughly expand C' by recursively replac-

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

ing 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 $[x/y]$ 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 $k \leq n$.

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

We adopted the tableau construction rules used in many DL-based inferential systems to facilitate the concept unfolding and the signature extraction process. In Figure 3, we present an example of how Book (defined in Figure 1) is unfolded by repetitively applying the transformation rules defined for each and every DL construct (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 completeness.

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

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

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

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

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

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

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

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

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 when a fixed point is reached, i.e. ${}^{n}\mathfrak{C} = {}^{(n-1)}\mathfrak{C}$. As demonstrated in [9], by carefully selecting a subset of admitted conceptual constructs, e.g. the underlying logic models of OWL-Lite and OWL-DL [14], 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 intended meaning of concepts) explicit. As a result, each concept is associated with finite sets of signatures in terms of the primitive concepts and properties. Effectively, concepts are deemed to hold parts of the information of the domain of discourse and thus, in spite of the apparent difference between ontologies and documents in the general sense, techniques for extracting and weighting document surrogates in IR can be applied analogically to concepts.

A straightforward approach to evaluate the influence of semantic signatures is to count the number of their occurrences in each $\mathfrak{C}_i$ of C. A signature is composed by the head (e.g. x and $\langle x, y_0 \rangle$ in Figure 3) and the tail (e.g. Document and hasTitle in Figure 3) separated by a colon. When counting, the heads of the signatures are ignored. The negative construct, $\neg$, states that the target concept

is explicitly excluded and thus value -1 is given to emphasise the restriction. Unexpanded universal quantification, e.g. $\forall \mathsf{R}.\mathsf{B}$, is treated as an atomic signature, as the presence of B is uncertain in the absence of property R. In many ontologies, for many fully defined concepts, the number of primitive concepts and properties is small. Hence, we do not expect to encounter sparse vectors very often. For example, 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 1, where equal weights are assigned to every signature.

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

(a) Definition of thesis concepts

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

(b) Unfolded thesis concepts

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

Table 1. Signature vector space of Book, Phdthesis, and Mastersthesis

	$\mathfrak{C}_1^{\mathsf{Book}}$	$\mathfrak{C}_1^{\mathsf{Phdthesis}}$	$\mathfrak{C}_1^{\mathsf{Mastersthesis}}$
$\top$ (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

Weights of signatures are fine-tuned 1) using the *inverse document frequency weight* (*idf*) [11] scheme from IR with the assumption that signatures appearing in a small number of concepts are more significant for the purpose of discriminat-

ing between concepts 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 an arbitrary ontology $\mathcal{O}$, n_{f_k} the number of concepts that refer to signature k, f_k, and $\mathrm{f}_{f_k,\mathsf{C}_i}$ the frequency of f_k in concept C_i, the *tf-idf* weight, w_{f_k,C_i}, of 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 $\mathcal{O}_m$, signatures such as Document, hasTitle, and hasYear appear in most of the concepts and thus are assigned low weights, whereas humanCreator appears in only one concept and thus is regarded as more important than others. Weights of indirect signatures are adjusted based on the weights of their related properties. For instance, z_0 : Document in Figure 3 is introduced because of humanCreator $\circ$ hasPublication and thus has less influence than x : Document. We decrease the weight of z_0 : Document to $w_{\mathsf{Document}} \cdot w_{\mathsf{humanCreator}} \cdot w_{\mathsf{hasPublication}}$.

Computing distances. By representing concepts as signature vectors, distances between concepts will then equal the distances between vectors in a high dimensional space. When there are more than one resultant $\mathfrak{C}_i$ due to disjunctive constructs (see Section 3.1), the shortest distance is computed.

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

$$\tau(\mathtt{sim}\,(\mathfrak{C}_i, \mathfrak{C}'_j)) = \begin{cases} -\log\left(\mathtt{sim}\,(\mathfrak{C}_i, \mathfrak{C}'_j)\right) & \text{if } \mathtt{sim}\,(\mathfrak{C}_i, \mathfrak{C}'_j) > 0 \\ +\infty & \text{if } \mathtt{sim}\,(\mathfrak{C}_i, \mathfrak{C}'_j) \leq 0 \end{cases} \tag{4}$$

$$\mathtt{sim}\,(\mathfrak{C}, \mathfrak{C}') = \frac{\sum\limits_{w_i \in \mathfrak{C},\ w'_i \in \mathfrak{C}'} w_i \times w'_i}{\sqrt{\sum\limits_{w_i \in \mathfrak{C}} w_i^2}\sqrt{\sum\limits_{w'_i \in \mathfrak{C}'} w_i'^2}} \tag{5}$$

$$\mathtt{sim}\,(\mathsf{C},\mathsf{D}) = \max_{(\mathfrak{C}_i \text{ of } \mathsf{C}, \mathfrak{C}'_j \text{ of } \mathsf{D})} \mathtt{sim}\,(\mathfrak{C}_i, \mathfrak{C}'_j) \tag{6}$$

Due to the introduction of negative numbers for capturing the semantics of $\neg$, there are possibilities for non-positive similarities based on Equation 5. A value of $+\infty$, therefore, represents a pair of totally divergent disjoint concepts.

After taking into account the weighting factors, signature vectors in Table 1 can be refined, and we can 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 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 cardinality restrictions on the hasPublication property and thus there might be fewer instances satisfying all its restrictions. The zero distance between two types of theses also suggests that these two concepts might not be properly defined in that they are identical from the given signatures.

Discussion. We see that 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 semantics bearing signatures that explicitly state the constraints imposed on the instances. We further collapse identical signatures so as to reduce the space complexity. Moreover, despite the apparent similarity, transforming ontologies into graphs cannot preserve the semantics *acoup sur*. Even with labelled edges, graph-based methods always have 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 fact that the introduction of the interim node B increases the length between A and C. 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 scheme to distinguish the contributions from different features, weights of which are normally set up manually by domain experts. We do not intend to undermine the importance of the role of human experts in understanding semantics. We, nevertheless, would like to introduce an automatic weighting mechanism to be complementary to their efforts. The *tf-idf* scheme borrowed from IR proposes a weight for each semantics-bearing (intensional) signature based on the significance of such a signature in introducing semantic discrepancies and thus is inline with the distance metrics. Finally, we consider our metrics as an improvement on techniques from feature-based families. This is evident partially from the fact that when constructing overall similarity/dissimilarity as a 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 based on the argument that the distances between concepts offer a fertile ground from which other metrics—that are effectively aggregations of concept-based distances—can be introduced.

4.1 Distance Between 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 [7,10]. 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-purpose lexicons, e.g. WordNet [16], are sufficient.

The similarity function (Equation 5) 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, C and C' be the concepts from $\mathcal{O}$ and $\mathcal{O}'$ with $\mathfrak{C}$ and $\mathfrak{C}'$ respectively and f'_i be the most similar signature of f_i with $\delta_i = \mathtt{sim}\,(f_i, f'_i)$,

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

Once obtained, the adjusted similarity between signatures can be used in Equation 4 and 3 to compute the similarity between concepts from different ontologies.

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

(a) Definition of Book and related concepts

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

(b) Unfolded concept Book

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

We use the Book concept from $\mathcal{O}_p$ to explain how distance between concepts from different ontologies can be computed. 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 similarly conceptualised. Apparently close concepts 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 from the fact that Book_m requires all books to have a title, a published year, a publisher, etc. while these are not mandatory for Book_p—an instance does not need to have a title, author, date, etc. to be qualified as a Book in ontology $\mathcal{O}_p$.

4.2 Distance Between a Concept and a Set of Concepts

There are occasions where the closeness is sought between a concept on the 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

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

If we emulate $p\left(\mathsf{D} \mid \mathsf{C}\right)$ as $\mathtt{sim}\left(\mathsf{C}, \mathsf{D}\right)$ obtained using Equation 6, we can then approximate the closeness of the ontology $\mathcal{O}$ around C by aggregating the distances between C and every other concept in $\mathcal{O}$. Note that $\mathtt{sim}\left(\right)$ is symmetric while $p\left(\mathsf{D} \mid \mathsf{C}\right)$ does not equal $p\left(\mathsf{C} \mid \mathsf{D}\right)$.

4.3 Distance Between Ontologies

As laid down in Section 2, we view ontologies as organisations of concepts and thus the distance between ontologies is computed out of those between concepts from the respective ontologies. In this paper, several methods are considered in order 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 Equations 3 and 7, we define

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

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 a 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 measure of the extent to which one distribution agrees with another. In order to overcome the asymmetry of KL divergence, Jeffrey-

divergence is proposed. Let $\mathsf{C}_i \in \mathcal{O}$ and $\mathsf{C}'_i \in \mathcal{O}'$ be two concepts from respective ontologies, then the distance between ontologies is computed as:

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

An 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. 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 can then compute the distance from an arbitrary "imperfect" concept C_k to C_0 as $\Delta(\mathsf{C}_k)$. The probability distribution of C_k can, therefore, be approximated as

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

Asymmetric distance measure. Variants of KL divergency are established on the assumption that the ontologies are defined over largely overlapping 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 ought to be obtained from mappings between fundamental semantics 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 6 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 in employing rigorous logics to underpin ontology modelling languages has presented itself as a challenge to several ontology management tasks. In such circumstances, as meaning is emphasised, it is not straightforward to identify the similarity/dissimilarity between concepts, which should be a function of both syntactic and semantic divergences. In this paper, we have demonstrated how concepts can be decomposed into semantics-bearing signatures and how such signatures can yield distance measures among concepts, between a single concept and a group of concepts, and how they may be generalised to compute the distance between ontologies. The proposed semantic measures/metrics can be complementary to other metrics. Compared to tradi-

tional 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 semantically-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 have resulted in many enormous knowledge corpora, especially in medicine, e.g. FMA [18] and GALEN [17]. The sheer size of such ontologies has put a tremendous burden on ontology management tools and have thus become a major obstacle to 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 [20], the authors extracted 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 to the existing segmentation techniques, our distance measures detect the semantic disagreement of different concepts and thus present 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 $\texttt{segmentation}(\mathcal{O}, \mathsf{C}, d) = \{\mathsf{D} \mid \forall\, \mathsf{D} \in \mathcal{O}.\Delta\,(\mathsf{C}, \mathsf{D}) \leq d\}$ where d is an arbitrary real number.

Ontology ranking: Building an 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 engine 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 semantics 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 formulating, 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 Semantic Web applications. The prospective users of such technology are faced with a number of challenges including ambiguity of the meaning of mappings, difficulties in capturing semantics, verification and validation of results and operationalisation in 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 minimises 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. A 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, needs 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 more elegant treatment is preferred. 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 strengths and weaknesses of our approach.

Acknowledgements

This work is supported under the OpenKnowledge and HealthAgents STREP projects funded by EU Framework 6 under Grant numbers IST-FP6-027253 and IST-FP6-027213, and the Advanced Knowledge Technologies (AKT) IRC funded by UK's EPSRC under Grant number GR/N15764/01. The authors are grateful for the input of Srinandan Dasmahapatra in the preparation of this paper.

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. A. Borgida, T. Walsh, and H. Hirsh. Towards measuring similarity in description logics. In *Proceedings of the Description Logics Workshop*, 2005.
5. T.M. Cover and J.A. Thomas. *Elements of Information Theory.* Series in Telecommunications. Wiley, 1991.
6. A.H. Doan, P. Domingos, and A.Y. Halevy. Reconciling schemas of disparate data sources: A machine-learning approach. In *SIGMOD Conference*, 2001.
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. I. Horrocks and U. Sattler. A tableaux decision procedure for $\mathcal{SHOIQ}$. In *Proc. of the 19th Int. Joint Conf. on Artificial Intelligence (IJCAI 2005)*, 2005.
10. Y. Kalfoglou, B. Hu, D. Reynolds, and N. Shadbolt. Semantic integration technologies. 6th month deliverable, University of Southampton and HP Labs, 2005.
11. R. Korfhage. *Information storage and retrieval.* Wiley Computer Publishing, 1997.
12. O. Lassila and R.R. Swick. *Resource Description Framework (RDF) Model and Syntax Specification.* W3C, 1999.
13. A. Maedche and S. Staab. Measuring similarity between ontologies. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management.*, pages 251–263. Springer-Verlag, 2002.
14. D. L. McGuinness and F. van Harmelen. *OWL Web Ontology Language Overview.* W3C, 2003.
15. 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.
16. G. A. Miller. WordNet; a Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995.
17. 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.
18. 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.
19. G. Salton. *Dynamic information and library processing.* Prentice-Hall, Inc., NJ, USA, 1975.
20. J. Seidenberg and A. Rector. Web ontology segmentation: Analysis, classification and use. In *Proceedings of WWW2006*, 2006. to appear.
21. C.E. Shannon. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, 2001.
22. J.F. Sowa. *Knowledge Representation: Logical, Philosophical, and Computational Foundations.* Brooks/Cole, Thomson Learning, 2000. ISBN 0-534-94965-7.
23. F. Wiesman and N. Roos. Domain independent learning of ontology mappings. In *AAMAS*, pages 846–853, 2004.