SemioSem: A Semiotic-Based Similarity Measure

Xavier Aimé[1,3], Frédéric Furst[2], Pascale Kuntz[1], and Francky Trichet[1]

[1] LINA - Laboratoire d'Informatique de Nantes Atlantique (UMR-CNRS 6241)
University of Nantes - Team "Knowledge and Decision"
2 rue de la Houssinière BP 92208 - 44322 Nantes Cedex 03, France
pascale.kuntz@univ-nantes.fr, francky.trichet@univ-nantes.fr
[2] MIS - Laboratoire Modélisation, Information et Système
University of Amiens
UPJV, 33 rue Saint Leu - 80039 Amiens Cedex 01, France
frederic.furst@u-picardie.fr
[3] Société TENNAXIA
37 rue de Châteaudun - 75009 Paris, France
xaime@tennaxia.com

Abstract. This paper introduces a new similarity measure called SemioSem. The first originality of this measure, which is defined in the context of a semiotic-based approach, is to consider the three dimensions of the conceptualization underlying a domain ontology: the *intension* (*i.e.* the properties used to define the concepts), the *extension* (*i.e.* the instances of the concepts) and the *expression* (*i.e.* the terms used to denote both the concepts and the instances). Thus, SemioSem aims at aggregating and improving existing extensional-based and intensional-based measures, with an original expressional one. The second originality of this measure is to be context-sensitive, and in particular user-sensitive. Indeed, SemioSem is based on multiple informations sources: (1) a textual corpus, validated by the end-user, which must reflect the domain underlying the ontology which is considered, (2) a set of instances known by the end-user, (3) an ontology enriched with the perception of the end-user on how each property associated to a concept c is important for defining c and (4) the emotional state of the end-user. The importance of each source can be modulated according to the context of use and SemioSem remains valid even if one of the source is missing. This makes our measure more flexible, more robust and more close to the end-user's judgment than the other similarity measures which are usually only based on one aspect of a conceptualization and never take the end-user's perceptions and purposes into account.

Keywords: Semantic Measure, Semiotics, Personnalization.

1 Introduction

Currently, the notion of similarity has been highlighted in many activities related to Ontology Engineering such as ontology learning, ontology matching or ontology population. In the last few years, a lot of measures for defining concept (dis-)similarity have been proposed. These measures can be classified according

R. Meersman, P. Herrero, and T. Dillon (Eds.): OTM 2009 Workshops, LNCS 5872, pp. 584–593, 2009.

to two approaches: (i) extensional-based measures such as [14], [9], [7] or [4] and (ii) intensional-based measures such as [12], [8] or [18]. Most of these measures only focus on one aspect of the conceptualization underlying an ontology, mainly the *intension* through the structure of the subsumption hierarchy or the *extension* through the instances of the concepts or the occurences of the concepts in a corpus. Moreover, they are usually sensitive to the structure of the subsumption hierarchy (because of the use of the more specific common subsumer) and, therefore, they are dependent on the modeling choices. Finally, these measures never consider the end-user's perceptions of the domain which is considered [2]. This paper presents SemioSem, a semiotic-based similarity measure, which aims at dealing with these problems. The first originality of SemioSem is to consider the three dimensions of a conceptualization: the *intension* (*i.e.* the properties used to define the concepts), the *extension* (*i.e.* the instances of the concepts) and the *expression* (*i.e.* the terms used to denote both the concepts and the instances). Using the semiotic approach in Knowledge Engineering is not a very new idea, and the 3 aspects of semiotics in ontology has already been underlined [16]. But we propose a practical application of this approach. Thus, SemioSem aims at aggregating three types of measure. The second originality of SemioSem is to be context-sensitive, and in particular user-sensitive. Indeed, SemioSem exploits multiple informations sources: (1) a textual corpus, validated by the end-user, which must reflect the domain underlying the ontology which is considered, (2) a set of instances known by the end-user, (3) an ontology enriched with the perception of the end-user on how each property associated to a concept c is important for defining c and (4) the emotional state of the end-user. The importance of each source can be modulated according to the context of use and SemioSem remains valid even if one of the source is missing. This makes our measure more flexible, more robust and more close to the end-user's judgment than the other similarity measures.

The rest of this paper is structured as follows. Section 2 describes in detail SemioSem: the basic foundations, the formal definitions, the parameters of the end-user and their interactions. Section 3 presents experimental results and compares our measure with related work in the context of a project dedicated to Legal Intelligence within regulatory documents related to the domain "Hygiene, Safety and Environment" (HSE).

2 SemioSem: A Semiotic-Based Similarity Measure

Building an ontology O of a domain D consists in establishing a consensual synthesis of individual knowledge belonging to a specific endogroup (*endogroup* is used here to name the people that agree with the conceptualisation formalized in the ontology, and not only those who have built it). For the same domain, several ontologies can be defined for different endogroups. Then, we call *Vernacular Domain Ontologies* (VDO) this kind of semantical resources[1], because they depends on an endogroup G, with a cultural and educationnal background.

[1] *Vernacular* means native.

Formally speaking, we define a VDO, for a domain D and an endogroup G as follows:

$$O_{(D,G)} = \{\mathcal{C}, \mathcal{P}, \mathcal{I}, \leq^C, \leq^P, dom, codom, \sigma, L\} \text{ where}$$

- $\mathcal{C}$, $\mathcal{P}$ and $\mathcal{I}$ are the sets of concepts, properties and instances of the concepts ;
- $\leq^C$: $\mathcal{C} \times \mathcal{C}$ and $\leq^P$: $\mathcal{P} \times \mathcal{P}$ are partial orders on concept and property hierarchies[2];
- $dom : \mathcal{P} \rightarrow \mathcal{C}$ and $codom : \mathcal{P} \rightarrow (\mathcal{C} \cup Datatypes)$ associates to each property its domain and ventually its co-domain;
- $\sigma : \mathcal{C} \rightarrow \mathcal{P}(\mathcal{I})$ associates to each concept its instances;
- $L = \{L_C \cup L_P \cup L_I, term_c, term_p, term_i\}$ is the lexicon of G relatively to the domain D where (1) L_C, L_P and L_I are the sets of terms associated to $\mathcal{C}$, $\mathcal{P}$ and $\mathcal{I}$, and (2) the fonctions $term_c : \mathcal{C} \rightarrow \mathcal{P}(L_C)$, $term_p : \mathcal{P} \rightarrow \mathcal{P}(L_P)$ and $term_i : \mathcal{I} \rightarrow \mathcal{P}(L_I)$ associate to the conceptual primitives the terms that name them.

An ontology is used in a particular context, which depends on the type of application for which the ontology is used, and the user of the application. This is why E. Rosch considers ontologies as *ecological* [5]. A VDO can be *contextualized*, that is adapted to the context and in particular to the user. Each personnalization of the VDO leads to a different *Personalised Vernacular Domain Ontologie* (PVDO), which respects the formal semantics of the VDO but adds a knowledge layer over the VDO to take into account the particularities of the user. We propose to adapt the VDO to the context by adapting (1) the degrees of truth of the *isa* links defined between concepts and (2) the degrees of expressivity of the terms used to denote the concepts.

We also propose to realize this contextualization and, more precisely, this personnalization, by using additional resources that express some elements of conceptualization specific to the user. Three additional resources are used: (1) a set of **instances** supposed to be representative of the user conceptualization (for instance, in the case of a business information system, these instances are the customers the user deal with, the products he sells to them, etc), (2) a **corpus** given by the user and supposed representative of its conceptualization (for instance, this corpus can be the documents written by the user on a blog or a wiki) and (3) **weighting of properties** of each concept. These weighting express the significance that the user attach to properties in the definition of the concept.

These weightings are fixed by the user as follows: for each property $p \in \mathcal{P}$, the user ordinates, on a 0 to 1 scale, all the concepts having p, in order to reflect its perception on how p is important for defining c. For instance, for the property *has an author*, the concept *Scientific Article* will be put first, secondly the concept *Newspaper Article*, for which the author is less important, thirdly the concept *Technical Manual*.

SemioSem is a semantics similarity measure defined in the context of an PVDO, and uses as resources the PVDO and the three additional resources

[2] $c_1 \leq^C c_2$ means that the concept c_2 is subsuming the concept c_1.

given above. Moreover, SemioSem is based on the three dimensions introduced by Morris and Peirce in their theory of semiotics [11]: (1) the *signified*, *i.e.* the concept defined in intension, (2) the *referent*, *i.e.* the concept defined in extension via its instances, and (3) the *signifier*, *i.e.* the terms used to denote the concept. Thus, SemioSem corresponds to an aggregation of three components: (1) an *intensional* component based on the comparison of the properties of the concepts, (2) an *extensional* component based on the comparison of the instances of the concepts, (3) an *expressional* component based on the comparison of the terms used to denote the concepts and their instances.

Each component of SemioSem is weighted in order to be able to adapt the calculation of the similarities to the way the end-used apprehends the domain which is considered (*e.g.* giving more importance to the intensional component when the end-user better apprehends the domain via an intensional approach rather than an extensional one). These differences of importance are conditioned by the domain, the cognitive universe of the end-user and the context of use (*e.g.* ontology-based information retrieval). For instance, in the domain of the animal species, a zoologist will tend to conceptualize them in intension (via the biological properties), whereas the majority of people use more extensional conceptualizations (based on the animals they have met during their life).

Formally, the function SemioSem: $\mathcal{C} \times \mathcal{C} \rightarrow [0,1]$ is defined as follows:

$$SemioSem(c_1, c_2) = [\alpha * intens(c_1, c_2) + \beta * extens(c_1, c_2) + \gamma * express(c_1, c_2)]^{\frac{1}{\delta}}$$

The following sections present respectively the functions *intens* (cf. section 2.1), *extens* (cf. section 2.2), *express* (cf. section 2.3) and the parameters α, β, γ, δ (cf. section 2.4).

2.1 Intensional Component

From an intensional point of view, our work is inspired by [1] and is based on the representation of the concepts by vectors in the space of the properties. Formally, to each concept $c \in \mathcal{C}$ is associated a vector $\overrightarrow{v_c} = (v_{c1}, v_{c2}, ..., v_{cn})$ where $n = |\mathcal{P}|$ and $v_{ci} \in [0,1], \forall i \in [1,n]$. v_{ci} is the weighting fixed by the end-user which precises how the property i is important for defining the concept c (by default, v_{ci} is equal to 1). Thus, the set of concepts corresponds to a point cloud defined in a space with $|\mathcal{P}|$ dimensions. We calculate a prototype vector of c_p, which was originally introduced in [1] as the average of the vectors of the sub-concepts of c_p. However [1] only considers the direct sub-concepts of c_p, whereas we extend the calculation to all the sub-concepts of c_p, from generation 1 to generation n. Indeed, some properties which can only be associated to indirect sub-concepts can however appear in the prototype of the super-concept, in particular if the intensional aspect is important. Thus, the prototype vector p_{c_p} is a vector in the space properties, where the importance of the property i is the average of the importances of the properties of all the sub-concepts of c_p having i. If for $i \in \mathcal{P}$, $S_i(c) = \{c_j \leq^C c, c_j \in dom(i)\}$, then:

$$\overrightarrow{p_{c_p}}\,[i] = \frac{\sum_{c_j \in S_i(c_p)} \overrightarrow{v_{c_j}}[i]}{|S_i(c_p)|}$$

From an intensional point of view, the more the respective prototype vectors of c_1 and c_2 are close in terms of euclidean distance (*i.e.* the more their properties are close), the more c_1 and c_2 are similar. This evaluation is performed by the function *intens*: $\mathcal{C} \times \mathcal{C} \rightarrow [0, 1]$, which is formally defined as follows:

$$intens(c_1, c_2) = 1 - dist(\overrightarrow{p_{c_1}}, \overrightarrow{p_{c_2}})$$

2.2 Extensional Component

From an extensional point of view, our work is based on the Jaccard's similarity [6]. Formally, the function *extens*: $\mathcal{C} \times \mathcal{C} \rightarrow [0, 1]$ is defined as follows:

$$extens(c_1, c_2) = \frac{|\sigma(c_1) \cap \sigma(c_2)|}{|\sigma(c_1)| + |\sigma(c_2)| - (|\sigma(c_1) \cap \sigma(c_2)|)}$$

This function is defined by the ratio between the number of common instances and the total number of instances minus the number of instances in common. Thus, two concepts are similar when they have a lot of instances in common and few distinct instances.

2.3 Expressional Component

From an expressional point of view, the more the terms used to denote the concepts c_1 and c_2 are present together in the same documents of the corpus, the more c_1 and c_2 are similar. This evaluation is carried out by the function $express : \mathcal{C} \times \mathcal{C} \rightarrow [0, 1]$ which is formally defined as follows:

$$express(c_1, c_2) = \sum_{t_1, t_2} \left(\frac{min(count(t_1), count(t_2))}{N_{occ}} * \frac{count(t_1, t_2)}{N_{doc}} \right)$$

With:

- $t_1 \in words(c_1)$ and $t_2 \in words(c_2)$ where $words(c)$ returns all the terms denoting the concept c or one of its sub-concept (direct or not);
- $count(t_i)$ returns the number of occurrences of the term t_i in the documents of the corpus;
- $count(t_1, t_2)$ returns the number of documents of the corpus where the term t_1 and t_2 appear simultaneously;
- N_{doc} returns the number of documents of the corpus;
- N_{occ} is the sum of the numbers of occurrences of all the terms included in the corpus.

2.4 Parameters of SemioSem

α, β et γ are the (positive or null) weighting coefficients associated to the three components of SemioSem. In a way of standardization, we impose that the components vary between 0 and 1 and that $\alpha + \beta + \gamma = 1$. The values of these three coefficients can be fixed arbitrarily, or evaluated by experiments. We also advocate a method to automatically calculate approximations of these

values. This method is based on the following principle. We consider that the relationship between α, β and γ characterises the cognitive coordinates of the end-user in the semiotic triangle. To fix the values of α, β and γ, we propose to systematically calculate γ/α and γ/β, and then to deduce *alpha* from the constraint $\alpha+\beta+\gamma = 1$. γ/α (resp. γ/β) is approximated by the cover rate of the concepts (resp. the instances) within the corpus. This rate is equal to the number of concepts (resp. instances) for which at least one of the terms appears in the corpus divided by the total number of concepts (resp. instances). The factor $\delta \geq 0$ aims at taking the mental state of the end-user into account. Multiple works have been done in Cognitive Psychology on the relationship beween human emotions and judgments [3]. The conclusion of these works can be summarized as follows: when we are in a negative mental state (*e.g.* fear or nervous breakdown), we tend to centre us on what appears to be the more important from an emotional point of view. Respectively, in a positive mental state (*e.g.* love or joy), we are more open in our judgment and we accept more easily the elements which are not yet be considered as so characteristic. According to [10], a negative mental state supports the reduction in the value of representation, and conversely for a positive mental state. In the context of our measure, this phenomemon is modelised as follows. We characterize (1) a *negative* mental state by a value $\delta \in]1, +\infty[$, (2) a *positive* mental state by a value $\delta \in]0, 1[$, and (3) a *neutral* mental state by a value of 1. Thus, a low value of δ, which characterises a positive mental state, leads to increase the similarity values of concepts which initially would not been considered as so similar. Conversely, a strong value of δ, which characterises a negative mental state, leads to decrease these values.

3 Experimental Results

SemioSem is currently evaluated in the context of a project funded by Tennaxia (http://www.tennaxia.com), an IT Services and Software Engineering company which provides industry-leading software and implementation services dedicated to Legal Intelligence in the areas " Hygiene, Safety and Environment " (HSE). In the context of this project, an ontology of the HSE domain has been built[3]. This ontology, which is particularly dedicated to *dangerous substances*, is composed of 3.776 concepts structured in a lattice-based hierarchy (depth=11; width=1300), and 15 properties. The whole ontology is stored in a database in order to ensure weak calculation time. In order to store the ontology in text-files, we have defined an extension of OWL which includes additional markup for to isa links and labels, dedicated to the representation of prototypicality values.

In order to evaluate our measure and to compare it with related work, we have focused our study on the hierarchy presented in figure 1: the goal is to compute the similarity between the concept *Carbon* and the sub-concepts of *Halogen*; for the expert of Tennaxia, these similarities are evaluated as follows: *Fluorine*=0.6; *Chlorine*=0.6; *Bromine*=0.3; *Iodine*=0.3; *Astatine*=0.1. We have

[3] Tennaxia company - All Rights Reserved. INPI 13th June 2008 N.322.408 – Scam Vlasquez 16th September 2008 N.2008090075.

also elaborated a specific corpus of texts composed of 1200 european regulatory documents related to the HSE domain (mainly laws, regulations, decrees and directives). Table 1 presents the similarity values obtained with three intensional-based measures: Rada, Leacock and Wu. One can note that all the values are equal because these measures only depend on the structure of the hierarchy.

Table 1 depicts the similarity values obtained with three extensional-based measures: Lin, Jiang and Resnik. Table 2 presents the similarity values obtained with SemioSem according to 6 contexts defined by the following parameters: (1) A ($\alpha = 0.7$, $\beta = 0.2$, $\gamma = 0.1$, $\delta = 1$), (2) B ($\alpha = 0.2$, $\beta = 0.7$, $\gamma = 0.1$, $\delta = 1$), (3) C ($\alpha = 0.2$, $\beta = 0.1$, $\gamma = 0.7$, $\delta = 1$), (4) D ($\alpha = 0.33$, $\beta = 0.33$, $\gamma = 0.33$, $\delta = 1$), (5) E ($\alpha = 0.7$, $\beta = 0.2$, $\gamma = 0.1$, $\delta = 0.1$), and (6) F ($\alpha = 0.7$, $\beta = 0.2$, $\gamma = 0.1$, $\delta = 5.0$).

These experimental results lead to the following remarks: (1) in all the contexts, SemioSem provides the same order of similarities as the other measures. In a context which gives priority to the intensional component (cf. context A), SemioSem is better than the other measures. In the context B which gives priority to the extensional component (resp. the context C which gives priority to the expressional component), SemioSem is close to Jiang's measure (resp. Lin's measure). In a context which gives no priority to a specific component (cf. context D), SemioSem is between Lin's measure and Jiang's measure; (2) context E and F clearly show the influence of the emotional factor: a positive mental state

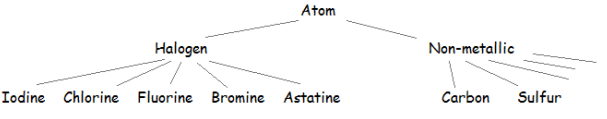

Fig. 1. Extract of the hierarchy of concepts of the HSE ontology

Table 1. Similarity with Carbon

Halogen	Rada	Leacock	Wu	Lin	Jiang	Resnik
Fluorine	0,25	0,097	0,6	0.31	0.14	1.43
Chlorine	0,25	0,097	0,6	0.28	0.12	1.43
Bromine	0,25	0,097	0,6	0.23	0.09	1.43
Iodine	0,25	0,097	0,6	0.22	0.09	1.43
Astatine	0,25	0,097	0,6	0	0	1.43

Table 2. Similarity with Carbon (SemioSem)

Halogen	A	B	C	D	E	F
Fluorine	*0.40*	0.14	0.32	0.27	0.91	0.025
Chlorine	*0.36*	0.12	0.29	0.25	0.90	0.017
Bromine	*0.29*	0.10	0.23	0.20	0.88	0.007
Iodine	*0.28*	0.10	0.23	0.19	0.88	0.006
Astatine	*0.01*	2.10^{-4}	2.10^{-4}	3.10^{-4}	0.63	1.10^{-8}

(cf. context E) clearly increases the similarities values and a negative mental state (cf. context F) clearly decreases similarities values; and (3) the concept *Astatine* is not evocated in the corpus, nor represented by instances. Thus, it is not considered as similar by Lin's and Jiang's measures. SemioSem finds a similarity value thanks to the intensional component.

4 Conclusion

SemioSem is particularly relevant in a context where the perception (by the end-user) of the domain which is considered (and which is both conceptualized within an ontology and expressed by a corpus and instances) can have a large influence on the evaluation of the similarities between concepts (*e.g.* ontology-based information retrieval). We advocate that such an user-sensitive context, which *de facto* includes *subjective knowledge* (whereas ontologies only includes *objective knowledge*), must be integrated in a similarity measure since ontologies co-evolve with their communities of use and human interpretation of context in the use. Formally, SemioSem respects the properties of similarity measure reminded in [4]: *positiveness*[4], *reflexivity*[5] and *symmetry*[6]. But, SemioSem is not a semantic distance since it does not check simultaneously the *strictness property*[7] and the *triangular inequality*[8]. For the extensional component, our first choice was the Amato measure. But one of our goal is to be independent from the modeling structure andthis measure clearly depends on the Most Specific Common Subsumer (MSCS). Moreover, in the case of our experiment, it does not really provide more relevant results. It is why we have adopted the Jaccard measure for its simplicity and its independence from the MSCS but we currently study the use of the Dice measure. For the expressional component, the Latent Semantic Analysis could be adopted but since it is based on the tf-idf approach, it is not really appropriated to our approach: we want to keep the granularity of the corpus in order to give more importance to concepts which perhaps appears less frequently in each document, but in an uniform way in the whole corpus (than concepts which are frequently associated in few documents). Then, as we simply compare the terms used to denote the concepts in the corpus, our approach is clearly limited since it can not deal with expressions such as "t1 and t2 are opposite". To deal with this problem, we plan to study more sophisticated computational linguistic methods. Finally, for the intensional component, our approach can be time-consuming (when the end-user decides to weight the properties of the concepts[9]), but it is really innovative to our knowledge and it

[4] $\forall x, y \in \mathcal{C} : SemioSem(x, y) \geq 0$.
[5] $\forall x, y \in \mathcal{C} : SemioSem(x, y) \leq SemioSem(x, x)$.
[6] $\forall x, y \in \mathcal{C} : SemioSem(x, y) = SemioSem(y, x)$.
[7] $\forall x, y \in \mathcal{C} : SemioSem(x, y) = 0 \Rightarrow x = y$.
[8] $\forall x, y, z \in \mathcal{C} : SemioSem(x, y) + SemioSem(y, z) \geq SemioSem(x, z)$.
[9] Again, by default, all the weightings are equal to 1 and the function *Intens* remains valid. In the case of our experiment, the results obtained in this context for the concept Fluorine are: A - 0.59 ; B - 0.19; C - 0.38; D - 0.37; E - 0.95; F - 0.12.

provides promising results. The parameters alpha, beta, gamma and delta are used to adapt the measure to the context which is related to the end-user perception of the domain according to the intentional, extensional, expressional and emotional dimension. We consider that this is really a new approach and this is why we call our measure SEMIOSEM (Semiotic-based Similarity Measure). Moreover, the aggregation we advocate does not just correspond to a sum: these parameters are used both to modulate the influence of each dimension and/or to adapt the calculus according to the resources with are available. Thus, when no corpus is available, the expressional component can not be used ($\gamma = 0$). A similar approach is adopted for the intentional component (an ontology without properties leads to $\alpha = 0$) and the extensional component (no instances leads to $\beta = 0$). The value of delta (emotional state) can be defined according to a questionary or the analysis of datas given by physical sensors such as the speed of the mouse, the webcam-based facial recognition, etc. To sum-up, SemioSem is more flexible (since it can deal with multiple information sources), more robust (since it performs relevant results under unusual conditions as shown by the case *Astatine* of the experimental results) and more user-centered (since it is based on the domain perception and the emotional state of the end-user) than all the current methods.

References

1. Au Yeung, C.M., Leung, H.F.: Ontology with likeliness and typicality of objects in concepts. In: Embley, D.W., Olivé, A., Ram, S. (eds.) ER 2006. LNCS, vol. 4215, pp. 98–111. Springer, Heidelberg (2006)
2. Blanchard, E., Harzallah, M., Kuntz, P.: A generic framework for comparing semantic similarities on a subsumption hierarchy. In: Proceedings of the 18th European Conference on Artificial Intelligence (ECAI 2008), pp. 20–24. IOS Press, Amsterdam (2008)
3. Bluck, S., Li, K.: Predicting memory completeness and accuracy: Emotion and exposure in repeated autobiographical recall. Applied Cognitive Psychology (15), 145–158 (2001)
4. d'Amato, C., Staab, S., Fanizzi, N.: On the influence of description logics ontologies on conceptual similarity. In: Gangemi, A., Euzenat, J. (eds.) EKAW 2008. LNCS (LNAI), vol. 5268, pp. 48–63. Springer, Heidelberg (2008)
5. Gabora, L.M., Rosch, E., Aerts, D.: Toward an ecological theory of concepts. Ecological Psychology 20(1-2), 84–116 (2008)
6. Jaccard, P.: Distribution de la flore alpine dans le bassin des dranses et dans quelques regions voisines. Bulletin de la Societe Vaudoise de Sciences Naturelles 37, 241–272 (1901) (in French)
7. Jiang, J., Conrath, D.: Semantic similarity based on corpus statistics and lexical taxinomy. In: International Conference on Research in Computationnal Linguistics, pp. 19–33 (1997)
8. Leacock, C., Chodorow, M.: Combining local context and Wordnet similarity for word sense identification. In: WordNet: an electronic lexical database, pp. 265–283. MIT Press, Cambridge (1998)
9. Lin, D.: An information-theoric definition of similarity. In: Proceedings of the 15th international conference on Machine Learning, pp. 296–304 (1998)

10. Mikulincer, M., Kedem, P., Paz, D.: Anxiety and categorization-1, the structure and boundaries of mental categories. Personality and individual differences 11(8), 805–814 (1990)
11. Morris, C.W.: Foundations of the Theory of Signs. Chicago University Press (1938)
12. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Transaction on Systems, Man and Cybernetics 19(1), 17–30 (1989)
13. Resnik, P.: Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research (JAIR) 11, 95–130 (1999)
14. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI 1995), vol. 1, pp. 448–453 (1995)
15. Sanderson, M., Croft, W.B.: Deriving concept hierarchies from text. In: Proceedings of the 22nd International ACM SIGIR Conference, pp. 206–213 (1999)
16. Sowa, J.: Ontology, metadata, and semiotics. In: Ganter, B., Mineau, G.W. (eds.) ICCS 2000. LNCS, vol. 1867, pp. 55–81. Springer, Heidelberg (2000)
17. Tversky, A.: Features of similarity. Psychological Review 84, 327–352 (1977)
18. Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: Proceedings of the 32nd annual meeting of the Association for Computational Linguistics, pp. 133–138 (1994)