SEMANTIC CLUSTERING BASED ON ONTOLOGIES
An Application to the Study of Visitors in a Natural Reserve

Montserrat Batet, Aïda Valls
Department of Computer Science and Mathematics, Universitat Rovira i Virgili, Av. Països Catalans 26, Tarragona, Spain

Karina Gibert
Department of Statistics and Operations Research, Univ. Politècnica de Catalunya, C. Jordi Girona 1-3, Barcelona, Spain

Keywords: Ontologies, Numerical and categorical data, Semantic similarity, Clustering, Tourism analysis.

Abstract: The development of large ontologies for general and specific domains provides new tools to improve the quality of data mining techniques such as clustering. In this paper we explain how to improve clustering results by exploiting the semantics of categorical data by means of ontologies and how this semantics can be included into a hierarchical clustering method. We want to prove that when the conceptual meaning of the values is taken into account, it is possible to find a better interpretation of the clusters. This is demonstrated with the analysis of real data collected from visitors to of a Natural Reserve. The results of our methodology are compared with the ones obtained with a classical multivariate analysis done in the same database.

1 INTRODUCTION

Nowadays, the extensive use of information and communication technologies provides access to a large amount of data (e.g. Wikipedia, electronic questionnaires). Usually, these resources provide textual data that may be semantically interpreted (e.g. a questionnaire can ask about the "*Main hobby*" of the responder, whose answer could be *dancing* or *trekking*). Classically, in clustering these terms were managed as *categorical* features (nominal or ordinal), whose values are expressed with labels in a predefined set of terms. Categorical values are compared at a syntactic or ordinal level, but without a semantic analysis. Taking into account the conceptual meaning of those terms, more accurate estimations of their degree of similarity (e.g. *trekking* is more similar to *jogging* than *dancing*) should be obtained, improving the quality and interpretability of the clusters obtained.

On one hand, this paper presents a hierarchical clustering method that can deal with numerical, categorical and with variables for which semantic information is available, named *semantic features*. In the core of the clustering algorithm, comparisons between objects are done with a compatibility measure which takes into account the different types of features and permits to make a homogeneous treatment. To provide the semantic knowledge, ontologies will be considered and use used to compare *semantic features*.

On the other hand, we test this clustering method with a dataset obtained from visitors of the Ebre Delta Natural Park. The goal is to find a characterization of the types of visitors in terms of their tourist and trip profiles. The results of our methodology are compared with the ones obtained with a statistical analysis made in with the same database. Results show that the clusters that consider the semantics of the terms give a more valuable classification of the visitors.

The paper is organized as follows: in Section 2, the clustering method for numerical, categorical and semantic features is presented. In Section 3, a dataset is introduced and an analysis of the results of a previous study is done. In section 4, the same data set is tested using the proposed method; the results are studied. The last section gives the conclusions.

2 SEMANTIC CLUSTERING

To include the semantic aspect into the clustering process, *semantic features* are introduced.

Batet M., Valls A. and Gibert K..
SEMANTIC CLUSTERING BASED ON ONTOLOGIES - An Application to the Study of Visitors in a Natural Reserve.
DOI: 10.5220/0003165602830289
In *Proceedings of the 3rd International Conference on Agents and Artificial Intelligence* (ICAART-2011), pages 283-289
ISBN: 978-989-8425-40-9

- A feature X_k is a semantic feature if: X_k takes linguistic values.
- Linguistic values or terms appearing in X_k can be semantically compared exploiting some background, like ontologies.

Since the values of *semantic features* became concepts (i.e. they correspond to labels of concepts in the reference ontology (Studer, Benjamins, and Fensel, 1998)) rather than simple modalities, it is possible to perform comparisons between values using a semantic similarity function.

2.1 Measuring Semantic Similarity

The definition of a measure of distance/similarity between the values of a pair of semantic features is essential for comparing objects in a semantic clustering approach.

This similarity is quantified by determining how concepts are alike based on semantic evidences observed in some knowledge source (*e.g.* a ontology or a corpus). According to the knowledge exploited in order to estimate the similarity between terms these functions can be classified in different families.

In some methods, taxonomies and, more generally, ontologies (Studer et al., 1998) are considered as a graph model in which semantic relations are modelled as is-a links between concepts. Then, the similarity is usually a function of the minimum number of is-a links (i.e. minimum path) between concepts (Rada, Mili, Bichnell, and Blettner, 1989). Similarity can also be a function of other features such as the depth of the concepts in the taxonomy (Leacock and Chodorow, 1998). The main advantage of these *Taxonomy-based* measures is that they only rely in a ontology for assessing the similarity. However, they are affected by their dependency on the degree of completeness, homogeneity and coverage of the ontology.

Other approaches consider not only the ontology but also the distribution of the compared terms in a corpus. These approaches rely on the *Information Content (IC)* of concepts (the inverse to its probability of occurrence in a corpus). Similarity is usually estimated as the IC of the first common ancestor of the compared concepts (Resnik, 1995) (Lin, 1998). In general, IC based measures provide better results than *Taxonomy-based* measures as they exploit a great amount of knowledge. However, they are affected by data sparseness if there is not enough available data to estimate information distribution.

In previous works (Batet, Sanchez, Valls, and Gibert, 2010) we have done an extensive study of the characteristics, performance, advantages and drawbacks of several semantic similarity measures. As a result of this study we propose the use the similarity measure presented in (Batet et al., 2010). This measure is based on the exploitation of all the taxonomical knowledge in a ontology (i.e. the full set of ancestors of a pair of compared concepts). In fact, the measure is based on the ratio between the number of different *is-a* ancestors of terms c_1 and c_2 in a reference ontology and the total number of ancestors of the terms, where $A(c_1)$ and $A(c_2)$ are the is-a ancestors of c_1 and c_2 in the ontology including themselves, respectively.

$$d_{SCD}(c_1, c_2) = \log \frac{\left|A(c_1) \cup A(c_2)\right| - \left|A(c_1) \cap A(c_2)\right|}{\left|A(c_1) \cup A(c_2)\right|} \quad (1)$$

On the contrary to previous approaches, where only a partial view of the modelled knowledge of the ontology is considered (*i.e.* the minimum path between concepts), this measure considers the relationships given by multiple inheritance of the concepts. This approach has been proven to provide a more accurate estimation of the similarity than classical *Taxonomy based* and *IC based* measures (Batet et al., 2010). In addition, as this measure only rely on the ontology, it has a low computationally cost. Therefore, in the semantic clustering proposal that we present, this measure is applied for evaluating semantic features.

2.2 Object Comparison with Numerical, Categorical and Semantic Features

Typically, an object is represented by a multidimensional vector where each dimension represents a feature or variable. Here, the case in which features can indistinctly be either numerical or categorical or semantic is treated. A compatibility measure for comparisons is introduced.

Metrics for mixed numerical and categorical values can be found in the literature (Ahmad and Dey, 2007; Gibert and Cortés, 1997). However, authors are not aware of references including also semantic features. In this sense, our proposal generalizes Gibert's mixed metrics (Gibert and Cortés, 1997) including semantic features.

Data is represented as a matrix, where objects $I = \{1,\ldots,n\}$ are in the rows, while the K features $X_1 \ldots X_K$ are in the columns. Thus, each cell (x_{ik}) contains the value taken by object i for feature X_k.

The distance between a pair of objects i and i', is calculated as the combination of applying a specific

distance for each type of feature X_k. This distance is defined in eq. 2:

$$d^2{}_{(\alpha,\beta,\gamma)}(i,i') = \alpha d^2_\zeta(i,i') + \beta d^2_Q(i,i') + \gamma d^2_S(i,i')$$
$$\text{with } (\alpha,\beta,\gamma) \in [0,1]^3, \alpha+\beta+\gamma = 1 \quad (2)$$

where $\zeta = \{k : X_k \text{ is a numerical feature}, k=1{:}K\}$, $Q = \{k : X_k \text{ is a categorical feature}, k=1{:}K\}$, and $S = \{k : X_k \text{ is a semantic feature}, k=1{:}K\}$. In our proposal $d_\zeta^2(i,i')$ is the normalized Euclidean distance for numerical features, $d_Q^2(i,i')$ is the χ^2 metrics for categorical values and $d_S^2(i,i')$ is the similarity measure introduced in the previous section.

In (2) each component has an associated weight. The weighting constants (α,β,γ) are taken as functions of the features' characteristics. In particular, they depend on the range of distances of each type of feature and how many variables refer.

We have $(\alpha,\beta,\gamma) \in [0,1]^3$ with $\alpha+\beta+\gamma=1$, being $n_\zeta=card(\zeta)$, $n_Q=card(Q)$, $n_S=card(S)$ and $d^2_{\zeta\ \max*}$, $d^2_{Q\ \max*}$, and $d^2_{S\ \max*}$ are the truncated maximums of the different sub-distances.

$$\alpha = \frac{\frac{n_\zeta}{d^2_{\zeta\ \max*}}}{\frac{n_\zeta}{d^2_{\zeta\ \max*}} + \frac{n_Q}{d^2_{Q\ \max*}} + \frac{n_S}{d^2_{S\ \max*}}}$$

$$\beta = \frac{\frac{n_Q}{d^2_{Q\ \max*}}}{\frac{n_\zeta}{d^2_{\zeta\ \max*}} + \frac{n_Q}{d^2_{Q\ \max*}} + \frac{n_S}{d^2_{S\ \max*}}}$$

$$\gamma = \frac{\frac{n_S}{d^2_{S\ \max*}}}{\frac{n_\zeta}{d^2_{\zeta\ \max*}} + \frac{n_Q}{d^2_{Q\ \max*}} + \frac{n_S}{d^2_{S\ \max*}}}$$

Notice that, (α,β,γ) depends on the importance of each type of feature. Firstly, all the components have the same influence in the calculation of $d^2(i,i')$, because they are proportional to the maximum distance for each type of feature; secondly, as truncated maximums are considered they are robust to outliers; and finally (α,β,γ) are proportional to the number of features they represent. So, the complete expression for the compatibility measure is:

$$d^2{}_{(\alpha,\beta,\gamma)}(i,i') = \alpha \sum_{k\in\varsigma} \frac{(x_{ik} - x_{i'k})^2}{s_k^2} + \frac{\beta}{n_Q^2} \sum_{k\in Q} d_k^2(i,i') + \frac{\gamma}{n_S^2} \sum_{k\in S} ds_k^2(i,i') \quad (3)$$

where s_k^2 is the variance of the feature X_k, $d_k^2(i,i')$ is the contribution of a single categorical feature X_k the χ^2 measure, and $d^2_{Sk}(i,i')$ is the contribution of the semantic feature X_k.

3 THE DATASET

In 2004, the *Observatori de la Fundació d'Estudis Turístics Costa Daurada* conducted a study of the visitors of the *Ebre Delta Natural Park*. The data was obtained with a questionnaire made to 975 visitors. The questionnaire was designed in order to determine the main characteristics of the tourism demand and the recreational uses of this natural area.

From these questions, two groups of interest were defined (Anton-Clavé, Nel-lo, and Orellana, 2007): 4 variables that define the tourist profile (origin, age group, accompanying persons and social class) and 6 that model the trip profile (previous planning, first and second reasons for trip, accommodation, length of stay and loyalty).

3.1 Previous Study and Results

In (Anton-Clavé et al., 2007), an statistical dimensionality reduction were used to find visitor's profiles. In particular, a multivariate homogeneity analysis was carried out. Two dimensions were selected for the analysis, keeping a 30% and 26% of variance respectively. In the interpretation phase, it was seen that Dimension 1 can discriminate among the variables relating to type of accommodation, length of stay and reason for the trip. It shows the degree of involvement of the tourist with the nature. The second dimension is determined by the type of group and age and shows the degree of involvement with the services available in the park. However, the total variance represented by the two first dimensions is 56%, which means that 44% of the information contained in the data set is missed, which can seriously affect the interpretation.

From that, five clusters of visitors were identified in (Anton-Clavé et al., 2007), from which the two first groups include a total of 83.9 % of the individuals, concluding that the rest of groups were really small and targeted to a very reduced group of visitors. For this reason, only the two main groups, (*EcoTourism* and *BeachTourism*) were characterised and Chi-square independence test was performed.

In the group *EcoTourism* (44,6%), the main interests are nature, observation of wildlife, culture and sports. People stay mainly in rural

establishments and campgrounds. In the group there are youths (25-24) coming from Catalonia and the Basque Country and it is the first time that visit the Delta. The group *BeachTourism* (39,3%) was characterised by their interest in beach, relaxation, and walking. It is family tourism, staying in rental apartments or second homes. They come from Spain or overseas. The group contains middle-class people, 35-64 years old, who do long and frequent visits.

4 STUDY OF THE VISITORS WITH CLUSTERING

In this section, a hierarchical clustering based on the Ward's criterion (Ward, 1963) is done on the same dataset of visitors to the Ebre Delta. We have taken the same subset of variables (4 that define the tourist profile and the 6 that model the trip profile).

Table 1: Typology of visitors to the Ebre Delta Natural Park with categorical features.

Class	#	Description
864	1	Single outlier visitor
C963	20	Long stage, between 35-early 40s years, 52% stays at second home, 75% are Catalan people, it is not clear the first reason to come (some of them come for walking).
C966	72	Long stage, between 35-early 40s, 68% is at home, half Spanish, half Catalan, have a second residence near the park
C965	37	Long stage, higher fidelity, around 46 years, 65% home, 78% Catalan, their main interest is gastronomy
C921	4	Long stage, more fidelity, between 35-early 40s, 50% goes to the hotel, an important part makes reservation, part of the foreigners concentrated in this group 25% of the class are foreigners, they come for recommendation of other people, 50% Catalan, main interests: relaxation or landscape
C936	16	Shorter stage, between 35-early 40s, almost 60% home, 80% Catalan, main interests: nature or business
C918	8	Youngs, under 30s, 50% stay in camping, 75% makes reservation, 50% Spanish, education tends to be first reason
C964	817	Shorter stage, occasional visit, between 35-early 40ss, mainly hotel, 63% Catalan, main reasons: nature, landscape and sightseeing

In the same way of the previous study, as most of the variables were categorical, only equal or different values are distinguished, leading to a poor estimation of the similarity between responses.

Table 2: Typology of visitors using semantics.

Class	#	Description
C947	110	The 81% comes for nature, but also for relax (35%), they use mainly hotels and rural establishments (79%), they have a reservation (95%)
C966	194	They come for relax (36%), visit the family (14.4%), but the second reason is mainly nature (35%), they have no hotel, they stay at home or at a family house (68,5%), and they have no reservation (99%), this is a group of young people leaving in the area, which repeat the visits more than others.
C968	203	Short stage, around 2 days, they clearly come for nature reasons (91.6%) and second for relax and wildlife (43.6%), they are in hotels or apartments (44.6%) although they have not reservation, mainly Spanish
C955	88	The first reason for coming is heterogeneous (nature, relaxation, beach, landscapes), the second is nature, they stay in a camping (90%), the half have a reservation, mainly Catalan and Spanish but also concentrates a big proportion of foreigners
C944	124	Relax and wildlife (46%) are the first reasons for coming and second is nature (40%), they stay at hotel or cottages (72%), and have reservation (88%). This is a group of slightly older people programming the stay in hotel or apartment, looking for relax or beach
C964	88	Wildlife and the landscape are the first reasons for coming (67%), but also for culture (19.5%) and the second reason is nature, they are mainly in hotel (54%). They are mainly Catalan or Spanish.
C957	84	Stay longer, slightly older that the rest, nature (38%) and beach (16%) are the main interest and second main interest is wildlife, most of them are foreigners with a second home, or that stay in an apartment.
C961	84	They all come for beach, their secondary interests are equally relaxation and nature, they live near the park and their visit is improvised, the stage is longer.

The clustering generates a big and heterogeneous class (with 83.8% of the tourists) which seems to share all type of visitors and other 7 small classes. Therefore, although the interpretation of the small classes is possible (see Table 1), from the point of view of the manager, this partition is useless because

the majority of visitors belong to the same profile as no clear intra-class difference can be identified.

In the knowledge-based approach proposed in this paper, the clustering method is able to manage the meaning of values, relating them to concepts in a given ontology. Therefore, a semantic clustering has been done by considering those categorical variables of the previous experiment as semantic variables and using the metrics proposed in Section 2.2. WordNet (Fellbaum, 1998) ontology is used to estimate the semantic similarity. From the results of this experiment, a cut in 8 classes is recommended for its interpretability (Figure 1).

This partition has clusters of more homogeneous dimension. This is an important fact, since now we can identify typologies of visitors that represent a significant proportion of the total number of visitors.

From the dendogram in Figure 1, it can also be seen that we have obtained clusters with high cohesion, which means that the distances between the members of cluster are quite small in comparison with their distances with objects outside the cluster. Moreover, if the level of partition is increased, then the cohesion of the clusters decreases quickly, which also indicates that the clusters are well defined.

This clustering is coherent with the grouping made by (Anton-Clavé et al., 2007) using multivariate analysis, because the variables about the reasons for visiting the park have a great influence in the formation of the groups. Interests on nature, beach and relax are present in different classes. However, thanks to the semantic interpretation of the concrete textual values provided by the respondents, we have been able to identify that visitors interested in nature are similar to those interested in wildlife. The system has been also able to identify the similarity between hotels and cottages and between second homes and familiar houses. This proofs that the estimation of the relative similarities among objects in terms of the meaning of the values improves the final grouping.

In this way, the two types of visitors identified in statistical analysis as Ecotourism and Beach Tourism have now been refined as follows:

- Ecotourism: visitors that stay in hotels and apartments for relax (C947), visitors with familiars or a second residence (C966), Spanish visitors interested in wildlife (C968) and tourists interested in culture (C955).
- Beach tourism: older people staying in hotels or apartments looking for relax and people that live near the park and go to the beach.

Notice that this is a more rich classification that establishes clear profiles of visitors.

5 CONCLUSIONS

The exploitation of data from a semantic point of view establishes a new setting for data mining methods. In this paper, it has been proposed a method to include semantic variables into an unsupervised clustering algorithm. A combination function that combines numerical, categorical and semantic features has been formally defined. In particular, the contribution of semantic features is obtained by estimating the semantic similarity of textual values from a conceptual point of view exploiting ontologies. Then, a knowledge-based clustering is proposed.

The paper presents an application of this methodology to a dataset obtained from a survey done to the visitors of a Natural Protected Park. The results show that a semantic clustering approach is able to provide a partition of objects that considers the meaning of the textual responses and, thus, the result is more interpretable and permits to discover semantic relations between the objects. The method has produced a more equilibrated grouping and provides useful knowledge about the characteristics of the visitors.

After obtaining these promising results, we will study the effect of using domain ontologies (e.g. medicine) instead of WordNet in the semantic similarity assessment. Moreover, the consideration of a set of ontologies used in an integrated way is also under study.

ACKNOWLEDGEMENTS

This work has been partially supported by the Universitat Rovira i Virgili (2009AIRE-04) and the DAMASK Spanish project (*Data mining algorithms with semantic knowledge*, TIN2009-11005) and the Spanish Government (PlanE, Spanish Economy and Employment Stimulation Plan). M. Batet is supported by an URV research grant.

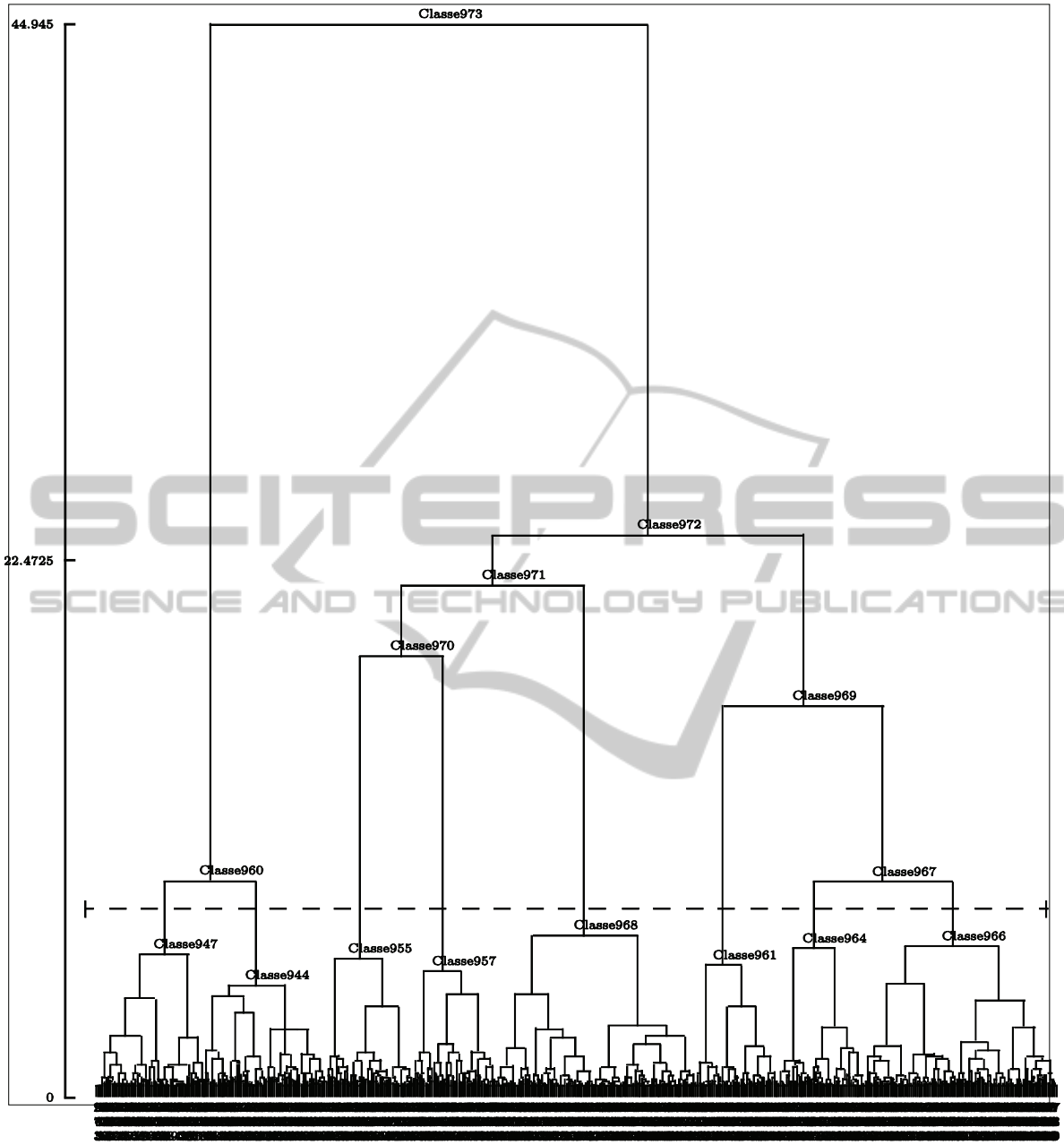

Figure 1: Dendogram with semantic features (8 classes).

REFERENCES

Ahmad, A., and Dey, L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data. *Data Knowledge Engineering, 63*(2), 503-527.

Anton-Clavé, S., Nel-lo, M.-G., and Orellana, A. (2007). Coastal tourism in Natural Parks. An analysis of demand profiles and recreational uses in coastal protected natural areas. *Revista Turismo & Desenvolvimento, 7-8*, 69-81.

Batet, M., Sanchez, D., Valls, A., and Gibert, K. (2010). Exploiting Taxonomical Knowledge to Compute Semantic Similarity: An Evaluation in the Biomedical Domain. In *Trends in Applied Intelligent Systems. 23rd International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2010, LNAI 6096* (pp. 274-283): Springer.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, Massachusetts: MIT Press. More information: http://wordnet.princeton.edu.

Gibert, K., and Cortés, U. (1997). Weighing quantitative and qualitative variables in clustering methods. *Mathware and Soft Computing, 4*(3), 251-266.

Leacock, C., and Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. In *WordNet: An electronic lexical database* (pp. 265-283): MIT Press.

Lin, D. (1998, July 24-27). *An Information-Theoretic Definition of Similarity*. Paper presented at the 15th International Conference on Machine Learning (ICML98), Madison, Wisconsin, USA.

Rada, R., Mili, H., Bichnell, E., and Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics, 9*(1), 17-30.

Resnik, P. (1995, August 20 - 25). *Using Information Content to Evalutate Semantic Similarity in a Taxonomy*. Paper presented at the 14th International Joint Conference on Artificial Intelligence, IJCAI 1995, Montreal, Quebec, Canada.

Studer, R., Benjamins, V. R., and Fensel, D. (1998). Knowledge Engineering: Principles and Methods. *Data and Knowledge Engineering, 25(1-2)*(1-2), 161-197.

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association, 58*, 236-244.