A Method for Semantics-based Conceptual Expansion of Ontology

Liping Zhou
School of Information Sciences
University of Science and Technology
Beijing, China
Zhouliping83_9_20@163.com

Dezheng Zhang
School of Information Sciences
University of Science and Technology
Beijing, China
zdzchina@126.com

Xin Chen
Department of Computer and Information Sciences
University of Alabama at Birmingham, USA
chenxin@cis.uab.edu

Chengcui Zhang
Department of Computer and Information Sciences
University of Alabama at Birmingham, USA
zhang@cis.uab.edu

ABSTRACT

For the past few years, automatic Ontology construction and expansion is one of the most important research subjects in the field of knowledge engineering. Compared with the traditional Term Frequency method, we propose a semantics-based method to extract concepts from a large corpus of text documents and expand the concepts of the known Ontology based on the semantic relations between two terms. The proposed method explores how to identify the candidate concepts, and how to give suggestions to knowledge engineers on where the concepts should be inserted in a given Ontology. The effectiveness of the proposed approach is demonstrated by experiments on a Traditional Chinese Medicine text corpus.

1. INTRODUCTION

Ontology is a data model for semantic representation and conceptualization of a knowledge domain, which uses comprehendible terms to express things existing in our knowledge domain, including entities, objects, relations, processes and so on. Moreover, it employs formal axioms to restrict and regulate the explanation and use of these terms [10]. Ontology provides general expression and construction for domain knowledge and concepts, which enables information sharing and reuse. Through the formal semantic conceptual structure, we can depict domain knowledge and build Ontology with the isomorphic reflection relations of heterogeneous information. Ontology is recognized as one of the most efficient approaches to unifying heterogeneous data, sharing knowledge in different areas, and making natural language to be understood by computers.

A formal definition of Ontology can be described as: *Ontology* is a 5-tuple $\Omega := (\Phi, R, \sigma, W, ins)$, where Φ is the set of known concepts, R is a set of relations on Φ, and σ: $\phi_1 \times \phi_2 \cdots \times \phi_{n-1} \to \phi_n$ is a function, which denotes that the *n*-1 elements on the left uniquely determine the n^{th} element on the right. *W* denotes axioms. For example, a concept belongs to the scope of another concept is an axiom. *ins* represents instances or objects of an concept.

In this paper, we mainly focus on the *is-a* relation of *R*, namely the relation between super-concept and sub-concept. With this relation, we ignore σ, *W* as well as *ins* and reduce The definition of Ontology to $\Omega := (\Phi, is\_a)$. When we mention *Ontology* in this paper, we refer to this reduced definition.

As a computer language for text representation, UNL (Universal Networking Language) was designed and developed by the University of United Nations in 1990s, which can represent documents independent of any kinds of natural languages. The basic structure of UNL is a partially ordered tree. Through UNL, a complex and ambiguous natural language is decoded into a simple and unambiguous computer language. Since Ontology also belongs to the category of natural language, it is inevitably complex and ambiguous. Therefore, we choose to use UNL to represent Ontology as a partially ordered tree structure.

In recent years, the focus of study on Ontology includes several aspects -- automatic construction, automatic reasoning, Ontology expansion, information retrieval by Ontology, text understanding, and so on. The study of thesaurus and automatic Ontology construction can be traced back to three decades ago [7]. At present, there is no standard methodology for the Ontology construction and expansion. Generally speaking, the existing methods can be divided into three categories i.e., manual, semi-automatic, and automatic means. In our on-going effort of studying the effective methods for knowledge transfer from Traditional Chinese Medicine (TCM) doctors, TCM doctors helped to build a TCM Ontology knowledgebase which consists of hundreds of concepts from TCM field. However, the present knowledgebase is not sufficient and still misses quite a lot of valuable domain concepts. How to efficiently expand it to include high-quality TCM Ontology is an urgent problem. Building Ontology manually has been proved to be cumbersome and infeasible. On the other hand, the automatic Ontology construction method is not mature enough for understanding unformatted natural language. Therefore, how to reduce the workloads of knowledge engineers while obtaining high-quality construction results has been the focus of our research. In this paper, we propose a semi-automatic ontology construction approach. Base on the cumulated information and data, we extract important concepts from the text corpora. According to the semantics of the text, we calculate the positions where the

SAC'08, March 16-20, 2008, Fortaleza, Ceará, Brazil.

concepts should be inserted. Hence, the work efficiency of knowledge engineers is greatly improved.

The proposed method has several advantages: 1) It is based on semantic statistics instead of term frequencies. 2) We extend the UNL to fit our specific needs and propose a new approach for calculating similarities among concepts. 3) Candidate concept selection and insertion is converted to a quadratic optimization problem that can be easily solved.

The rest of the paper is organized as follows: Section 2 introduces the concept of UNL and the conceptual semantic matrix. We propose a new Ontology conceptual similarity algorithm in Section 3. Section 4 states the details of Ontology expansion using conceptual semantic matrix by calculating the similarity between two concepts, especially the similarity between a known and an unknown concept. A global concept matrix is then generated for all documents with the candidate concepts extracted. In Section 5, the experiment details and results are illustrated. Section 6 concludes with future work.

2. UNL AND CONCEPTUAL SEMANTIC MATRIX

UNL expresses information or knowledge in the form of a semantic network which is different from natural languages. UNL expressions are unambiguous. Logically it is an ordered semantic network. UNL is composed of 3 parts: concepts, relations, and attributes. In a UNL network, nodes represents concepts, known as Universal Words (UWs); the links between nodes express relation of concepts, known as UNL Relations (URs); the subjective meaning intended by the speaker can be expressed through attributes, and attributes restrict the semantics of a universal word. A text can be expressed by these 3 parts of the UNL. Details can be found in [4].

Traditionally the document is represented by the Vector Space Model (VSM) based on Term Frequency from the documents. In this paper, we use UNL expressions to formalize the characteristics of a document, hence structuralize the Ontology representation. Because a UW is unambiguous in a UNL expression, semantics of a UW are easily differentiated in different UNL expressions. Unlike the approach in [5] in which the weight of a UW is solely determined by its number of URs, we assign different weights to relations (URs) in this study. Instead of considering only the number of URs linked with a UW, the weight of the UW is also determined by the types and weights of relations (URs) on the links. It is assumed that the more URs linked with a UW, the more important the UW is in the sentence. In the proposed method, concepts $C\sigma$ extracted from a document are represented as nodes in a UNL tree. In the rest of the paper, the term "node" and "UW" are used interchangeably. We further categorize URs into three types. The corresponding UW weight calculation algorithms are given below. It is worth mentioning that initially, the weight of each node (UW) is the number of links connected with the node. Then this initial weight is updated according to the category of the URs related to this node.

1. The first category: two UWs linked by a UR have similar rank or 'close' relations in the document, or one UW is frequently used to describe the semantics of the other one. The 'close' relations here include "and", "coo" (co-occurrence), "cnt" (content), "equ" (equivalent), "icl" (included), "iof" (an instance of), "or", "pof" (part-of), "pos" (possessor), "seq" (sequence), and so on [4]. In this case, the two UWs linked by these URs should have the same weight (the greater one of them). And the weight of such kind of UR is set to 3.

2. The second category: for two UWs linked by the same UR, one states the status or the characteristics of the other, or the two UWs are related in terms of actions or events. The URs in this category include "agt" (agent), "aoj" (thing with attribute), "cag" (co-agent), "cob" (effected co-thing), "gol" (goal), "obj" (effected thing), "pur" (purpose), "ptn" (partner), "rsn" (reason), "met" (method) [4], and so on. The weight of the super-node is transferred to its sub-nodes, which implies the transfer of importance of the concept. The weight of the sub-node is thus equal to the maximum weight of its super-nodes plus the number of URs linked to it (its initial weight). The weight of a UR in this category is set to 1.

3. The third category: all the other URs not in the first and the second categories. This involves relations that represent time, frequency, range, quantity, location, restriction, some function words, adverbs, etc. An example in [4] for this type of relation is "mod" (modification). The weight of the super-node is not transferred to its sub-node. The weight of the sub-node is equal to the number of URs linked to it (its initial weight). The weight of a UR in this category is set to 0.

Unlike the approach used in [2], which fills the representation matrix with a finite rule set of linguistic properties and the corpus, we define the conceptual semantic matrix $M(c_\sigma)$ and fill it with the semantic relations among concepts from the corpus. The rows of $M(c_\sigma)$ represent all concepts of the document, known as c_σ, and the columns represent a concept set c_Φ in this document which is a subset of Φ. Note that all the concepts in c_Φ are known since Φ is the set of known concepts. However, c_σ may contain both known concepts that belong to Φ and unknown concepts extracted from the document. The element m_{ij} in $M(c_\sigma)$ represents the semantic relationship between the i^{th} concept in c_σ and the j^{th} concept in c_Φ. We denote the weight of the i^{th} concept in c_σ as λ_i, and the weight of the UR between the i^{th} concept and the j^{th} concept as $\gamma_{i\leftrightarrow j}$. We can obtain the weight of every node. An element in $M(c_\sigma)$ is equal to the sum of the node (the concept in c_σ) weights and the UR weight, which is formalized as $m_{ij} = \lambda_i + \gamma_{i\leftrightarrow j}$. If there is no UR between the two concepts, set $m_{ij} = 0$. In this way, we express the conceptual semantic matrix as $M(c_\sigma)$ based on the weights of conceptual semantic relations, which is different from the classical representation approaches based on term frequency. The proposed method also enables the evaluation of similarity among known and unknown concepts.

Because of the existence of noise in unknown concepts, the indiscriminate use of unknown concepts to calculate the similarity between any two concepts may not be accurate. Therefore, c_Φ instead of c_σ is used to form columns of the matrix. This reduction of matrix scale can also reduce the workload of

computation, hence significantly improve the efficiency.

In Figure 1, an example from a Chinese Medicine document is provided to illustrate the construction of a UNL tree (Figure 1(a)) and the corresponding conceptual semantic matrix (Figure 1(b)). The Chinese sentence is "*中暑，病证名，又名发痧，常见症状有眩晕，口渴，烦躁，脉虚等*", which means "heatstroke is the name of a case, which is also called sunstroke, and its common symptoms include vertigo，thirst，fidgety, feeble-pulse, and so on." This sentence can be represented, by UNL, as a conceptual semantic network and conceptual semantic matrix below. We assume that the concepts "*中暑(heatstroke)*", "*症状(symptom)*", and "*脉虚(feeble-pulse)*" compose c_{Φ} which is a subset of Φ. Take the node *症状(symptom)* for example, since it has the "aoj" (thing with attribute) relation with its super-node "*中暑(heatstroke)*", its node weight is 9 which is the maximum weight of its super-node (i.e., 3) plus the number of links linked to it (i.e., 6). In the conceptual semantic matrix, the corresponding element between "*症状(symptom)*" and "*中暑(heatstroke)*" is then 9+1=10, with 1 being the weight of "aoj".

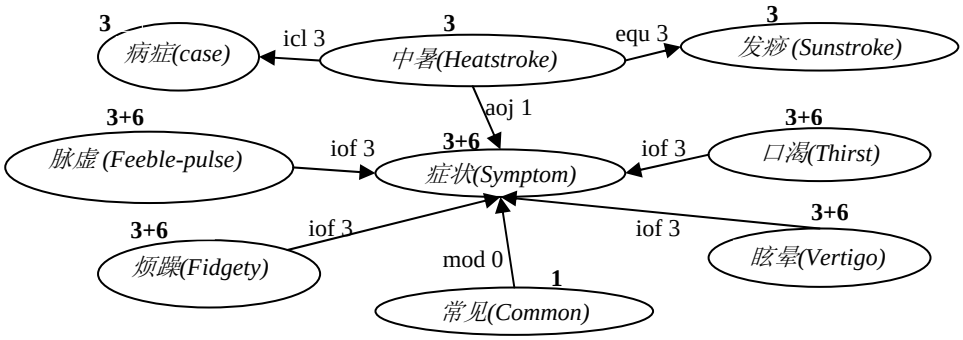

Figure 1(a) An Example UNL Tree (Note: "icl" defines an upper concept or a more general concept. "equ" defines an equivalent concept. "iof" defines a class concept that an instance belongs to. "aoj" defines a thing that is in a state or has an attribute. "mod" defines a thing that restricts a focused thing.)

	中暑 (Heatstroke)	症状 (Symptom)	脉虚 (Feeble-pulse)
中暑(Heatstroke)	3	3+1	0
病症(Case)	3+3	0	0
发痧(Sunstroke)	3+3	0	0
症状(Symptom)	9+1	9	9+3
脉虚(Feeble-pulse)	0	9+3	9
口渴(Thirst)	0	9+3	0
烦躁(Fidgety)	0	9+3	0
眩晕(Vertigo)	0	9+3	0
常见(Common)	0	1	0

Figure 1(b) Conceptual Semantic Matrix of the UNL in 2(a).

3. CALCULATE THE CONCEPTUAL SIMILARITY

In a given ontology, the length of the path between two concepts *x* and *y* indicates how strong their semantic similarity is. The longer the distance between *x* and *y*, the weaker the semantic similarity between them is. Many approaches have been proposed to calculate the semantic similarity between concepts of Ontology. For example, in [6], the authors explore the determination of semantic similarity by a number of information sources, which consist of structural semantic information from a lexical taxonomy and information content from a corpus. However, the implementation of this approach is very complicated. In [3] the authors proposed an algorithm for word similarity calculation in thesauri, which could be applied to distance measures in the hierarchical representation of Ontology. In this paper, we propose a new approach to calculate the conceptual similarity, which follows the following three principles: 1. The greater the distance between two concepts (referred to as the shortest path between them) is, the weaker the similarity between them.

2. The top level concepts close to roots in the structured Ontology are more general (more inclusive) than the lower level concepts close to leaves. Therefore, in the isomorphic paths, with the same distance, the conceptual similarity of two top level concepts is less than that of two lower level concepts.

3. The more siblings a concept has, indicating that its super-node has multiple meanings, the smaller the similarity between this concept and its super-node, and so is the similarity between its siblings and itself.

We define the conceptual similarity as:

$$s(x, y) = \frac{1}{\rho}\left(\frac{\alpha}{\theta\tau}\right)^{1/3} \quad (1)$$

In the formula, α denotes the average depth of two concepts; θ denotes the average number of siblings of the two concepts; ρ denotes the number of steps on the shortest path between the two concepts in the hierarchy; τ is the maximum depth of the ontology. Other definitions include:

1） If $\rho = 0$, then $s(x, y) = 1$.

2） If $\rho \geq k$, then set $s(x, y) = 0$. k is a pre-established threshold, which indicates when the distance between two concepts is too big so that the semantic similarity between them is meaningless, and the value of similarity is set to 0.

3） If $\theta = 0$, then $s(x, y) = \frac{1}{\rho}(\frac{\alpha}{\tau})^{1/3}$.

It can be observed that Formula (1) fulfills the three principles for calculating conceptual similarity as mentioned above, and follows the inequality: $0 \leq s(x, y) \leq 1$.

4. ONTOLOGY CONCEPTUAL EXPANSION

Using the approach mentioned in the previous sections, we can obtain the conceptual semantic matrices of every document. In

this section, we will discuss how to use these matrices to extract candidate concepts, and how to calculate the positions where the candidate concepts should be inserted into a given Ontology.

4.1 Global Concept Matrix

We define C as the corpus, and denote the set of concepts in the documents retrieved according to the concepts from Ontology Ω as $C_\sigma = \{c_1, c_2, ..., c_n\}$. Generally speaking, C_σ contains some concepts which belong to Φ (we call C_Ω) and also some unknown concepts that are not yet included in Φ (we call C_ω). Moreover, C_ω is also the set of candidate concepts from which we select a subset as an expansion of Ontology Ω. We assume that $C_\sigma = C_\omega \cup C_\Omega$, and $C_\omega \cap C_\Omega$ is empty. We propose a new concept named Global Concept Matrix, which represents the semantic accumulation of all important concepts from the corpus. We merge the conceptual semantic matrices $\{M(c_\sigma)_1, M(c_\sigma)_2, ..., M(c_\sigma)_k\}$ generated from C, where *k* is the number of documents in *C*, into a Global Concept Matrix $M(C_\sigma)$. The rows of $M(C_\sigma)$ represent all concepts in C_σ, and every column represents the concept in C_Ω; the element m_{ij} of matrix $M(C_\sigma)$ implies the accumulated semantic relations between the concept c_i in C_σ and the concept c_j in C_Ω. The construction of $M(C_\sigma)$ consists of the following three steps:

Construct UNL networks: Search each sentence in documents in *C*. If the sentence contains the concept(s) in Φ, construct a UNL network for this sentence as demonstrated in Section 2.

Construct conceptual semantic matrices $M(c_\sigma)_i$: According to the algorithm of constructing conceptual semantic matrix introduced in Section 2, convert UNL semantic networks into the corresponding conceptual semantic matrices $M(c_\sigma)_i$. In addition, we ignore the non-content words such as function words, auxiliary words, pronouns, prepositions, exclamations, morpheme words, and other stop words.

Construct Global Concept Matrix $M(C_\sigma)$: For each element m_{ij} of conceptual semantic matrix $M(c_\sigma)_i$, check whether the corresponding concepts in the row and the column already exist in Global Concept Matrix $M(C_\sigma)$; if yes, add it to the current value of m_{ij} in the corresponding position of $M(C_\sigma)$; if not, add a new row (and a new column if the corresponding concept is not yet included in $M(C_\sigma)$) to $M(C_\sigma)$, and insert the value of m_{ij} into the corresponding position. Repeat this process till all the elements in all conceptual semantic matrices $M(c_\sigma)_i$ are checked and merged into $M(C_\sigma)$.

The next step is to select candidate concepts according to $M(C_\sigma)$. We assume that the unknown concepts having similar semantics with the known concepts in Ontology Ω are our target candidate concepts. According to the approach in [8], we assign a weight k_i to each column of $M(C_\sigma)$. Therefore, we have a weight vector $K = \{k_1, k_2, ..., k_m\}$, where *m* is the number of columns of $M(C_\sigma)$. For any two concepts *x* and *y* in rows, the definition of similarity is as follows:

$$S(x, y) = \sum_{t=1}^{m} k_t m_{xt} m_{yt} \quad (2)$$

where, *x* and *y* also denote the subscripts of elements in the matrix, i.e., the x^{th} row refers to the concept *x* and the y^{th} row refers to the concept *y*. This formula is used to calculate the similarity between two row concepts in C_σ. Formula (2) has the same functionality as Formula (1) since both of them are used for calculating the similarity between two row concepts. In theory, the same pair of concepts should generate equal or similar similarities by the two formulas. Therefore, the problem about how to obtain the value of $K = \{k_1, k_2, ..., k_m\}$ is converted to the problem of how to minimize this quadratic formula - $\left|\sum_{i=1}^{m}\sum_{j=1}^{m}(S(x_i, y_j) - s(x_i, y_j))\right|$, where x_i and y_j are known concepts in Ontology Ω. It should be pointed here that in this optimization process, we use the known concepts in Ω only to avoid the unpredictable effects of noise from unknown concepts.

4.2 Concept Extraction and Expansion

By Linear Programming (LP), we obtain the value of $K = \{k_1, k_2, ..., k_m\}$ through minimizing $\left|\sum_{i=1}^{m}\sum_{j=1}^{m}(S(x_i, y_j) - s(x_i, y_j))\right|$. We then use Formula (2) to calculate the similarity of every pair of concepts in C_σ. Furthermore, we focus on calculating the similarity between candidate concepts in C_ω and known concepts in C_Ω. The greater the similarity is, the more likely that this candidate concept can be inserted as the sub-node, super-node or sibling of some known concept in C_Ω.

A similarity threshold value is pre-defined for selecting candidate concepts. More discussions on how the threshold value is selected are included in Section 5. If its similarity with a known concept is greater than this threshold, we will consider the concept as a candidate concept and submit it to the knowledge engineers together with the known concept. The engineers are responsible to judge whether the concept should be inserted into the corresponding position in Ontology.

5. EXPERIMENTS

In on our previous work, we constructed a Traditional Chinese Medicine (TCM) Ontology, based on which we can evaluate the proposed approach. The TCM Ontology has 9 second level nodes such as "处方 (Prescriptions)", "疾病 (Disease)", "病因 (Etiology)". The maximum depth of the TCM Ontology is 8. Since the usable EnConverter for Chinese is not available, we have to convert sentences from texts into UNL semantic networks by hand. We use traditional Chinese medicine cases from medical case database and specialized medical documents from the Internet as the corpus. The process of the experiment contains 5 steps:

1. Search sentences from the corpus, which contain concepts of Ontology Ω, and convert the sentences into UNL networks. Then construct the corresponding conceptual semantic matrix $M(c_\sigma)_i$;
2. Generate the global concept matrix;
3. By Linear Programming, calculate weight vector $K = \{k_1, k_2, ..., k_m\}$. In the experiment, we use the optimizing multinomial software –Ampl /Solver [9] to solve

the quadratic problem;

4. Calculate the similarity between two concepts according to Formula (2), especially the similarities between known concepts and unknown concepts. The average of similarities among known concepts in a given Ontology is used as the similarity threshold. If the similarity between an unknown concept and a known concept is greater than the threshold, the unknown concept is selected as a candidate concept for Ontology expansion;
5. Candidate concepts are submitted to domain engineers together with the contexts where the two concepts emerge. The engineers can decide whether the candidate concepts and their positions in Ontology are valid, and whether these concepts should be inserted into the Ontology.

We selected 20 concepts from TCM Ontology and searched 127 sentences from the corpus to construct conceptual semantic matrices. By applying the proposed algorithm, 210 candidate **(expanded)** concepts are generated. In order to evaluate the results, we choose 9 commonly known concepts and ask the knowledge experts to select all the correct candidate concepts for them. Then we measure the recall, precision, and F-measure [1] of these concepts. The mean values are also presented here. The results are presented in Table 1.

Table 1. Evaluation of the Expanded Concepts

concept	depth	Precision	recall	F-measure
方药 (prescriptions)	2	32.11	85.29	46.66
病因 (etiology)	2	25.00	82.04	38.32
治法 (therapeutic method)	3	34.98	72.35	47.16
药性 (drug properties)	3	30.25	80.59	43.99
虚证 (asthenia syndrome)	4	46.21	62.28	53.05
肝病 (liver disease)	4	51.61	70.11	59.45
中暑 (heatstroke)	5	53.85	58.25	55.96
止血剂 (hemostatic)	5	60.21	60.16	60.18
血厥 (bleeding syncope)	6	78.27	51.27	61.96
Mean		**45.83**	**69.15**	**66.54**

When the depth of concepts increases, i.e. going down the hierarchy of the TCM Ontology, the precision is increasing and the recall is decreasing. F-measure is also increasing, which shows that the less general the concepts are, the more accurate the expanded concepts are. The reason is that the qualities of expanded concepts from general concepts are quite diverse. Some of them are only remotely related to the known concept. In this case, these candidate concepts should not be the neighbor concepts (super-concepts, sub-concepts or siblings) of this known concept in the Ontology. With similarity threshold being the average similarity among known concepts, the number of inserted candidate concepts for top level concepts in Ontology is large. This causes the low precision for these concepts.

In the next step, we vary the threshold for concept similarity from 0.6 times to 1.4 times of the average similarity of all known concepts and observe the changes of the average F-measure. The result is shown in Table 2. This experiment indicates that F-measure is also affected by the similarity threshold. It can be seen that when the threshold increases, the F-measure value increases. After it achieves the max at the similarity threshold of 1.2, it starts to drop with the continued increase in the similarity threshold.

Table 2. F-measure with Different Similarity Thresholds

Threshold	0.6	0.8	1.0	1.2	1.4
F-measure	36.17	52.40	66.54	68.19	47.33

6. CONCLUSION AND FUTURE WORK

In this paper, we propose a new approach of Ontology conceptual expansion based on semantic information statistics. TCM Ontology is used to verify the proposed algorithm, which is different from term-frequency based approaches. Although it identifies some improper candidate concepts or points out wrong positions where a concept should be inserted in Ontology, most of the candidate concepts obtained for specialized concepts (lower level concepts) are considered to be relevant by domain experts. It is also worth mentioning that the proposed approach is not domain-independent and can be applied to many fields.

Our future work includes analyzing the factors that influence the quality of calculation. These factors may include how to select the corpus, how to adjust the threshold of the algorithms, and how to improve similarity formula to make it in accordance with the actual situations. Due to the fact that the UNL networks for TCM documents have to be manually generated, the scale of the corpus is limited. We will need to develop tools to automate this process. In addition, future work also includes adjusting the proposed approach to other research directions such as Ontology mapping, Ontology merging, and automatic Ontology construction.

7. ACKNOWLEDGMENT

The work of Chengcui Zhang was supported in part by NSF DBI-0649894 and the UAB ADVANCE program through the sponsorship of the National Science Foundation.

8. REFERENCES

[1] Tho, O. T., Hui, S. C., Fong, A.C.M., and Cao T. H. 2006. Automatic Fuzzy Ontology Generation for Semantic Web. IEEE Trans. on Knowledge and Data Engineering, 18, 6 (Jun. 2006), 842-856.

[2] Faatz, A., Hoermann, S., Seeberg, C., Steinmetz, R. 2001. Conceptual Enrichment of Ontologies by Means of a Generic and Configurable Approach. In Proceedings of the Workshop on Semantic Knowledge Acquisition and Categorization (Helsinki, Finland, August 2001).

[3] Resnik, P. 1999. Semantic Similarity in a Taxonomy: an Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research, 11.

[4] The Universal Networking Language (UNL) Specifications Version 3 Edition 3 http://www.undl.org/unlsys/unl/UNLSpecs33.pdf

[5] Choudhary, B. and Bhattacharyya P. 2002. Text clustering using semantics. In Proceedings of the eleventh International World Wide Web Conference.

[6] Li, Y., Bandar, Z. A., and McLean, D. 2003. An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. IEEE Transactions on Knowledge and Data Engineering, 15, 4 (July 2003).

[7] Spark-Jones, K. 1997. Readings in Information Retrieval, Morgan Kaufmann.

[8] Bisson, G., Nedellec, C., and Cañamero, L. 2000. Designing Clustering Methods for Ontology Building - The o'K workbench. In Proceedings of the Ontology Learning ECAI-2000 Workshop (August 2000).

[9] Ampl Optimization Software: *http://www.ampl.com*

[10] Borst, W. N. 1997. Construction of Engineering Ontologies for Knowledge Sharing and Reuse. PhD thesis, University of Twente, Enschede.