Preliminary Ideas on Concept to Query Closeness Metrics

Alejandra Segura N.
Depto. Sistemas de Información
Universidad del Bío-Bío
Concepción, Chile
e-mail: asegura@ubiobio.cl

Christian Vidal-Castro.
Depto. Sistemas de Información
Universidad del Bío-Bío
Concepción, Chile
e-mail: cvidal@ubiobio.cl

Claudia Martinez A.
Depto. Ingeniería Informática
Universidad Católica de la Santísima Concepción
Concepción, Chile
e-mail: cmartinez@ucsc.cl

Salvador Sánchez-Alonso.
Depto. Ciencias de la Computación
Universidad de Alcalá
Madrid, España
e-mail: salvador.sanchez@uah.cl

***Abstract*—The usefulness of knowledge models for information retrieval tasks such as digital resource tagging, query expansion, and recommending, among others, requires that the query concept be present in the model, i.e., exists matching. If the query concept is absent in the ontology, exists matching problem. In this case, it can be identified in the model other sematic and syntactically closeness concepts to the query concept. Once identified the closest concept is possible to extract relevant knowledge from the ontology. The goal of this work is to propose a solution to the problem mentioned, by identifying those variables that can affect the closeness between a query and concepts in a domain ontology. Using these variables as a starting point, we propose 6 indexes for measuring the degree of closeness. We present the results of implementing a search-selection algorithm using indexes based on exact words, contained words, coincidences in descriptive fields, new words and approximate depth. These indices are validated via a case study, and, from these results, we recommend adjustments needed for building a global concept closeness index in future works.**

Keywords-component; Ontology; Semantic Web; Web Information Retrieval.

I. INTRODUCTION

Information retrieval (IR) involves several processes, among which we can distinguish indexing, query, search and relevance assessment [1].

Knowledge models, mainly thesauri, terminologies and ontologies, provide external knowledge that can semantically enrich, either directly or indirectly, several tasks related to information retrieval, such as digital resource tagging, indexing, querying and recommending. For example, indexing can either build an index by extracting information for resource tagging or it can use the knowledge model itself as an indexing system [2, 3]. Knowledge models are used in the formulation and refinement processes to navigate among the modeled concepts, and also in query expansion to disambiguate or further specify the initial user query by adding new information to the query [4, 5]. In relevance assessment, knowledge models have been used to rank results according to relevance, in what is called “score of results” [6].

Knowledge models can be used either manually, i.e. the user defines the query through model navigation, or automatically, through the use of algorithms that extract relevant information. Despite the fact that manual extraction can yield more precise results, its application is limited due to the large size and structure of many models and because, in most cases, users must have previous knowledge of the model [7]. On the other hand, automatic knowledge extraction allows using large knowledge models and makes their structure transparent to the users. However, their use is generally restricted to those cases where the query is exactly represented in the model.

According to [8], an ontology is “an explicit specification of a conceptualization”. Ontologies involve two parts: syntax and semantics. The first considers symbols and the set of rules for combining them, and the second refers to the meaning of expressions. Ontologies rigorously specify a conceptual framework in a domain, with the goal of facilitating communications, interaction, exchange and information sharing between different computational systems.

Knowledge representation, therefore, requires domain knowledge, representation languages and mechanisms for inferring new knowledge. As indicated in [9], ontologies are the tool of choice for formal knowledge representation oriented to computer-assisted semantic analysis.

A problem associated to the use of knowledge models as a basis for automatic knowledge extraction occurs when an exact match to the query cannot be found. Then, the knowledge model can be examined looking for concepts that are closely related to the query. Accessing the concept closest to the query in turn makes it possible to access other semantically related concepts. That is, new knowledge can be extracted and then utilized in any other information retrieval processes, such as digital resource tagging, indexing, recommending or query expansion.

This article describes an algorithm that, given a query, extracts those concepts that are closest to the query from a given ontology. It also presents an analysis of the proposed algorithm's initial assessment.

The remainder of the article is organized as follows. Section II formalizes the problem considering the non-exact correspondence between the query and the knowledge modeled in the ontology. Section III analyzes previous works related to syntactic and semantic similarity metrics, and also to ontology-based query expansion algorithms. Section IV describes the research methodology, while Section V describes the evaluation process, which includes validation by experts, and the design and application of a questionnaire. Section VI analyzes and discusses the results. Finally, Section VII presents our conclusions and future research directions.

II. THE QUERY MATCHING PROBLEM

In this section, we define the basic elements of the matching problem, which are the query concept and the domain ontology.

Query Concept: The user's query is the query concept (QC), which is formed by a set of words w, that is, QC = $\{w_1, . . ., w_n\}$. Let QC' be the same query concept after linguistic processing, that is, after removing morphological variations (stemming) and ignoring stopwords. Common, frequently-used words generally do not provide information and thus are considered stopwords. The set of stopwords includes prepositions, articles, adverbs, conjunctions, possessive and demonstrative pronouns, and some verbs and nouns. Stopword lists are generally language dependent, but some domain-dependent stop-word lists have also been built [10]. Stemming is the process by which morphological variations of the terms are extracted, e.g., conjugations as well as prefix and suffix derivational morphemes. A derivational morpheme is appended or prepended to a lexical base to form a new derived word. Eliminating these morphemes leaves only the root. Therefore, the lemma represents the variations of the derived terms [11].

Domain ontology: in this work, based on [12], we define a domain ontology as a triplet O={C, R, I}, where C is the set of classes, R is the set of relationships between classes and instances, and I is the set of class instances. Any concept modeled in the ontology is represented either in the classes or in the instances. Every ontology modeled concept (OC) is a set of words such that OC= $\{v_1, . . ., v_y\}$. The ontologies have relationships related to concept taxonomies such as *is-a* or *part-of,* even though they can also include domain-specific relationships to take into account the modeling requirements of the knowledge domain.

Considering the above definitions, the matching problem between a query concept and the concepts modeled in a domain ontology exists when:

$$\nexists OC \in O : QC' = OC$$

Most of the work in information retrieval that makes use of knowledge models assumes that there is a matching between the query concept and at least one concept in the ontology. Although the query concept is absent from the ontology, can be identified in the model other closeness concepts to the query concept (QC). The degree of closeness to the query concept might be determined according to syntactic and semantic variables. It should be noted that there is little information about the query concept context to determine the closeness between the query concepts and the concepts in the ontology. Specifically, we only know the concept (and set of words) and the domain of knowledge where it is immersed.

Our proposal presents an algorithm aimed at extracting those concepts in the ontology which are closest to a query concept for which no exact match exists.

III. RELATED WORK

The problem of matching a query to a domain ontology has been studied in relatively few ontology-based query expansion algorithms, most of which perform the query expansion only if the query concept exists exactly in the model, that is, there is a concept which contains the same words keeping the same order [3, 4, 13-17]. Moreover, our study of related work also reviews relevant syntactic and semantic similarity measures, as they indirectly affect the problem at hand. The similarity has been managed both syntactically and semantically. The syntactic aspect is based on the comparison of two strings and the semantic aspect through the comparison between two concepts present in the model. In the latter case exists two approaches: one based on the structure and another based on the information content.

The edit distance measure (e) proposed by [18] is used to determine the degree of syntactic similarity between 2 strings A and B. It is defined as the number of removal, replacement or append operations needed to convert string A into string B.

These semantic similarity measures consider that both concepts are represented in the model. Other structure-based measures use as a basis the number of nodes separating both concepts. Proposed measures utilize variables such as the depth of the lowest common ancestor (LCA), the local density of the sub-tree containing both concepts, the distance between the concepts and the types of relationships among them. For example, the measure proposed by [19] is calculated as the shortest route between the concepts. The measure proposed by [20] is calculated as a function of the depth of the LCA and the number of links between the concept and said ancestor. The similarity measure proposed by [21] is a function of the concepts' depth, the depth of the LCA and the shortest distance between the concepts. The same authors propose another similarity measure that also includes the local specificity [22]. Li, et al. [23] propose a similarity measure that takes into account the shortest route between the concepts, the depth of the LCA and empirical information.

Information-based semantic similarity measures consider the information content of the model's derived nodes and corpus statistics, such as the concept's frequency in the corpus and the corresponding inverse frequency. The more information two concepts share, the higher their similarity. Some similarity measures in this category are the ones proposed by Resnik [24], which take the information content of the LCA into account. Jiang and Conrath [25] and Lin

[26] suggest improvements to Resnik's measure which also consider the information content of each concept.

The above mentioned semantic and syntactic similarity measures are not directly applicable to the correspondence problem, as they can be used only if both concepts are present in the model. In our case, however, the query concept is absent from the model. Nevertheless, these works are relevant to formulating our proposed solution.

Previous work proposes a query expansion algorithm based on domain ontologies [5]. The same work also defines an algorithm for finding the concept closest to a query concept not present in the ontology. The closest concept is defined as the concept that contains the largest number of words in common with the query concept, and that contains the smallest number of words that do not belong to the query concept.

IV. METHODOLOGY

Figure 1 shows the framework that describes how the matching problem is addressed. In any IR process, when the query concept is absent in the model, it is processed linguistically. Then, the concepts that share words with the query concept are extracted from the ontology. These concepts are also processed linguistically to calculate the indexes of closeness. Finally, based on the indexes, the global concept closeness index is estimated.

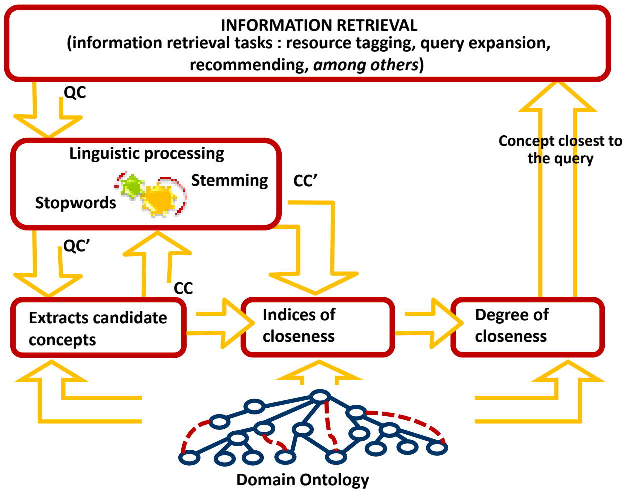
INFORMATION RETRIEVAL
(information retrieval tasks : resource tagging, query expansion, recommending, *among others*)
QC
Linguistic processing
Stemming
Stopwords
CC'
QC'
CC
Concept closest to the query
Extracts candidate concepts
Indices of closeness
Degree of closeness
Domain Ontology

Figure 1. Matching problem framework.

We define the candidate concepts (CC) for a given query concept (QC) as all those concepts present in the ontology that share words with the linguistically pre-processed query QC'. Linguistic processing includes stopword elimination and plural extraction: in other words, a full stemming process is not performed.

Let QC$_i$ be a query concept composed of a set of words w, and QC'$_i$ is the linguistically preprocessed query. Also, let OC be any concept modeled in a domain ontology, which in turn is defined as a set of words v_{nt} such that OC$_1$ = {v$_{11}$,. . ., v$_{1p}$}, . . ., OC$_n$ = {v$_{n1}$, . . ., v$_{nt}$}. Then, if QC$_i$ is not present in the ontology, we can say that OC$_n$ is a candidate concept for QC'$_1$ if and only if

$$\exists v_{nt} \in OC_n \ \land \ \exists w'_{ip} \in \ QC'_i \ / \ v_{nt} = w'_{ip}$$
$$\land \ v_{nt} \neq stopword \land$$
$$\land \ \ v_{nt} = stemm(v_{nt})$$

The closeness from the query concept QC' and the candidate concepts CC is a function of the following

1.-The number of words that match the query concept words. Two types of coincidences, exact and contained, are considered.

1.1.- A contained coincidence occurs when the query concept word is contained within a candidate concept's word.

1.2.- An exact coincidence occurs when the query concept word is syntactically identical to a word in the candidate concept. The greater the number of coincident words, the closer the query concept and the candidate concept.

2.- Word positions. In addition to the coincidences among the query concept's words and the candidate concepts' words, the coincident word's position is also considered. Analysis of this parameter can vary according to each language's grammatical rules. A candidate concept's closeness to the query concept increases if word positions also coincide.

3.- Number of new or mismatching words. This criterion counts the number of CC words that do not coincide either exactly or approximately with any query concept word. Stop-words are ignored. The fewer the new or mismatched words, the higher the candidate concept's closeness to the query concept.

4.- Concept depth. The depth is defined as the longest path from the candidate concept to the model's root class, considering hierarchical is-a relationships. In a domain ontology, the deeper the concept the more specific it is.

5.- Parent relevance. This item considers the parents of the candidate concepts and quantifies the number of its descendants that are also candidate concepts. Closeness increases if a candidate concept belongs to a sub-tree with a greater candidate concept density.

6.- Descriptive fields representation. This item considers the occurrence of the query concept in any descriptive field associated to the candidate concept, such as the <definition> or <description> fields. Concept closeness increases if the candidate concept's descriptive fields contain the query concept.

The 7 variables just mentioned are considered relevant for determining the closeness between a query concept and those concepts modeled in an ontology. Next, we show the 6 indices to be calculated by the algorithm for each candidate concept. Each index takes values between 0 and 1.

Normalized exact word index

$$ind_{coin_{ex}} = \left(\frac{c_{word_{ex}}}{t_{word_{qc}}} \right) \qquad \textbf{(1)}$$

where:

c_word_ex : Number of query words that are also present, exactly, in the candidate concept (variable 1.1).
t_word_qc : Total number of words in the query concept QC, ignoring stopwords.

ISBN: 978-1-61208-200-4

Normalized contained word index

$$ind_{coin_{co}} = \left(\frac{c_{word_{co}}}{t_{word_{qc}}} \right) \qquad \textbf{(2)}$$

where:

c_word_co : Number of query words that are contained in a word present in the candidate concept (variable 1.2).
t_word_qc : Total number of words in the query concept QC, ignoring stopwords.

Normalized new word index

$$ind_{nue} = \left[1 - \frac{t_{word_{cc}} - (c_{word_{co}} + c_{word_{ex}})}{t_{word_{cc}}} \right] \qquad \textbf{(3)}$$

where:

c_word_ex : Number of words in the query concept that are also present in the candidate concept (variable 1.1).
c_word_cc : Number of words in the query concept that are contained in a word present in the candidate concept (variable 1.2).
t_word_cc : Total number of words in the candidate concept CC, ignoring stopwords
t word cc –(c_word_cc + c_word_ex): Number of new or mismatching words (variable 3).

Descriptive fields coincidence index

$$ind_{coin_{des}} = \left(\frac{c_{word_{des}}}{t_{word_{qc}}} \right) \qquad \textbf{(4)}$$

where:

c_word_des : Number of words in the query concept that are an exact or partial match to words in the candidate concept's descriptive fields (variable 6).
t_word_qc : Total number of words in the query concept QC, ignoring stopwords.

Normalized aproximate depth index.

Given the computational complexity inherent to the problem of exactly calculating a concept's maximum depth in a formal domain ontology, this index is defined as an approximation to the candidate concept's maximum depth (variable 4). A formal domain ontology usually includes a large number of modeled concepts, each of which can have several parent nodes, therefore many routes to the root node can exist.

To calculate the semantic distance each line of inheritance has value 1. Therefore we assume that any inheritance relationship with a class that belongs to another ontology is assigned half this value (t_subclassof_o=0.5). This avoids calculating the depth in external ontologies.

$$ind_{prof_{app}} = \left(\frac{ind_{prof}}{\max\limits_{cc\ \ of\ \ qc} (ind_{prof})} \right)$$

$$ind_{prof} = \left(\frac{t_{antcc}}{p_{subclassof_O}} \right) \qquad \textbf{(5)}$$

$$p_{subclassof_O} = \left(\frac{t_{subclassof_O}}{t_{class_O}} \right)$$

where:

p_subclassof_o : parent relationship average for all concepts in the ontology.
t_subclassof_o : total number of subclass relationships in the ontology. Subclass relationships have weight 1, while references to classes in other ontologies or sub-ontologies have weight 0.5.
t_class_o : total number of classes modeled in the ontology that have at least one parent (except for root nodes).
t_ant_cc : total number of parents of a candidate concept.
max(cc of qc): maximum approximate depth of all candidate concepts for a given query.

Candidate density in sibling index (variable 5).

$$ind_{den} = \max_{parent\ \ CC} \left(\frac{c_{sibling_{CC}}}{t_{sibling_{CC}}} \right) \qquad \textbf{(6)}$$

where:

max parent cc : maximum value among all the candidate concept's parents.
parent_cc : number of concept candidate's parents.
c_sibling_cc : number of candidate concept's siblings that are also candidate concepts.
t_sibling_cc : total number of candidate concept's siblings.

V. EVALUATION

We wanted to evaluate the algorithm by examining the concepts it retrieves and determining their closeness to the query concept. The algorithm is evaluated using the Subcellular Anatomy for the Nervous System (SAO) ontology, which is available in the OWL language. This ontology provides a method for describing sub-, supra- and macro-cellular structures. SAO "describes the parts of neurons and glia and how these parts come together to define supracellular structures such as synapses and neuropil", and was developed by the Open Biological and Biomedical Ontology Foundry (http://www.obofoundry.org/crit.shtml) with the stated aim of providing updated domain ontologies in several knowledge areas for the scientific community [27].

For evaluation purposes, we utilize test queries extracted from the syllabi of four central nervous system Anatomy courses. Query concepts are extracted from the contents list for each course. Details can be found in Table I.

TABLE I. SYLLABI USED FOR ALGORITHM EVALUATION

Course details
Learning and Memory: Activity-Controlled Gene Expression in the Nervous System. Fall 2009 . http://ocw.mit.edu/courses/biology/7-340-learning-and-memory-activity-controlled-gene-expression-in-the-nervous-system-fall-2009/Syllabus/
Psychology 202 Biopsychology. Fall 2009. http://courses.washington.edu/psy222/Syllabi/Psy%20202%20Fall%2009%20syl.pdf
Neurophysiology 1012 and 2012. Spring 2009. http://www.neuroscience.pitt.edu
Neuro 405- Neurophysiology .Fall 2010. http://webpub.allegheny.edu/employee/l/lfrench/Neurophys%20syllabus%20F06.htm

73 initial query concepts were identified. For each initial query concept, the algorithm generated a list of candidate concepts sorted by closeness, according to the scores of the 6 indices mentioned in Section 4.

Faculty from the Universidad Católica de la Santísima Concepción, with professional experience in medicine, specifically in anatomy and cellular biology, participated as experts in this study. At first, these experts were consulted to filter the initial concepts and to select those that were coherent with their research lines. Three experts agreed in the selection of 7 initial query concepts. Table II details algorithm results for these 7 queries.

TABLE II. LIST OF INITIAL QUERY CONCEPTS USED IN THE FIRST PHASE OF THE EVALUATION WITH THE RESULTS OF THE ALGORITHM

	Indices												
		coinex		**coinco**		**nue**		**coindes**		**prof_app**		**den**	
Queries	**t**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**
Activation of the NMDA receptor	30	0	0.67	0	0.67	0	1	0.53	1	0	0.67	0.04	1
AMPA receptor endocytosis	31	0	0.67	0	0.67	0	1	0.41	1	0	0.67	0.04	1
APs-Ca channels	24	0	0.33	0	0.33	0	0.5	0.47	1	0	1.33	0.11	1
Brain	17	0	0	0	1	0	0.33	0.13	1	1	2	0.03	1
cerebro spinal fluid	11	0	0.33	0	0.33	0	0.33	0.13	1	0.33	1	0	1
Electrical principles of neuronal function	49	0	0.25	0	0.25	0	0.5	0.27	1	0	0.5	0	1
Neurons	75	0	1	0	1	0	1	0.12	1	0	2	0	1

t: total number of candidate concepts
coin_ex: normalized exact word index.
coin_co: normalized contained word index.
coin_des: descriptive fields coincidence index.
nue: normalized new word index.
prof_app: normalized approximate depth index.
den:candidate density in sibling index.
A: minimun.
B: maximum.

Each expert evaluated the first 10 candidate concepts, randomly sorted, for each of the 7 initial queries through a questionnaire named “Concept closeness evaluation in a domain ontology”. This instrument was designed to gather expert opinions regarding:

- the conceptual closeness of the initial query concept to the candidate concept,
- the closeness rank of the 10 candidate concepts, a number from 1 to 10, where 10 denotes greatest closeness.

In the first phase of the evaluation, a single expert was chosen so as to do an exploratory case study. This was done to find out “the relationship between the closeness ranking determined by the expert and the indices computed for each candidate concept by our algorithm”.

The expert evaluated the closeness of 10 candidates for 7 initial queries. Of these 70 measurements, 64 candidate concepts (91%) were considered close and only 6 (9%) were found to be not close or unrelated to the initial query. The expert indicated that the strategy he applied was, after determining closeness, to perform a top-down revision based upon his experience with the candidate concepts in terms of their composition relationships.

VI. Results and Discussion

These results were analyzed using Pearson correlation analysis [28]. This Pearson analysis was performed to find correlations between the closeness rank specified by the expert (on a scale of 1 to 10, 10 denoting greatest closeness) and each of the indices proposed in Section 4. Analysis results are shown in Table III. Five of the correlation indices were found to be positive relationships, that is, a higher expert ranking yields a larger estimated index. Pearson coefficients concentrated in the (0.46, 0.07) range.

The best correlation values for the closeness rank were obtained for the new word index (0.46), the exact word index (0.40) and the contained word index (0.40). On the other hand, the least relevant correlation was obtained for the descriptive fields coincidence index (0.07). This low correlation can be explained by considering the poor structure and flexibility allowed when filling these descriptive fields. Additionally, it must be considered that the goal of these fields is mainly to provide information to other users.

TABLE III. THE CLOSENESS RANKING DETERMINED BY THE EXPERT AND THE RANKING GIVEN BY THE ALGORITHMS AND THE CALCULATED INDICES.

ordex-coinex	ordex-app	ordex-nue	ordex-def	ordex-prof	ordex-dens
0.40	0.40	0.46	0.07	-0.02	0.19

ordex: closeness ranking determined by the expert
coin_ex: normalized exact word index.
coin_co: normalized contained word index.
coin_des: descriptive fields coincidence index.
nue: normalized new word index.
prof_app: normalized approximate depth index.
den: candidate index in sibling index.

The only index that showed a negative low correlation was the approximate depth index, with a value of -0.02. Beforehand, we expected a higher positive correlation, under the premise that candidate concepts that are deeper in the ontology are more specific, which would in turn yield a higher closeness rank (closer to 10) with respect to the query concept. However, the data shows that the deeper the depth index the lower the closeness rank is, i.e. the candidate concept has a ranking closer to 1. This can be explained by noting that the query concepts (content lists of a course) and the concepts in the ontology have differences in the granularity/specialization. The query concept was assumed to be very specific, so then a deep candidate concept would be very close. However, not all query concepts are specific. Then, if the query concept is of a general nature, its closest candidate concepts will also be of a general nature. Therefore, the relevance of the depth index depends on whether the query concept is of a general or a specific nature. Unfortunately, as the query concept is not present in the ontology, we do not have information about its depth.

In general terms, our results show that the indices proposed in this work are useful as a measure of the closeness between a query concept and concepts modeled in an ontology. As such, they can be used as a starting point for the development of a global closeness index that can be used to rank those concepts that are closest to the query concept.

VII. CONCLUSIONS AND OUTLOOK

The usefulness of knowledge models for information retrieval tasks such as digital resource tagging, query expansion, and recommending, among others, requires that the query concept be present in the model. This work addresses the matching problem that occurs when the query concept is not present exactly in the model. We postulate that it is possible to find concepts that are syntactically and/or semantically close to the query concept, even if the query is not represented in the ontology, and that the closeness between the query concept and a candidate concept can be determined as a function of 7 variables. Based on these 7 variables, we define 6 normalized indices for estimating concept closeness, which are the exact word index, the descriptive fields coincidence index, the contained word index, the new word index, the approximate depth index, and the candidate index in siblings index. After a first evaluation phase, we conclude that 5 of the 6 indices are positively correlated with the closeness rank perceived by domain experts. Moreover, one of the proposed indices warrants further research as its incidence on closeness rank depends on the generality or specificity of the query concept. This, in turn, leads us to envision a mechanism that allows knowing a priori a query concept’s depth so as to be able to calibrate the candidate concepts’ closeness rank.

As future work, we must determine the degree of incidence define in the closeness rank estimation. In order to do this, we will perform a new evaluation with a larger number of experts, and also we will consider changing the knowledge domain area so as to generalize the results obtained to date.

ACKNOWLEDGMENTS

This work was partially supported by DIUBB 115215 3/RS Chile; Bío-Bío University, Chile. Some research in ontologies is also funded by Organic.Lingua CIP-ICT-PSP.2010.6.2 project, reference number 270999.

REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*: Addison-Wesley-Longman, 1999.

[2] C. D. Nguyen, K. J. Gardiner, and K. J. Cios, "Protein annotation from protein interaction networks and Gene Ontology (In press)," *Journal of Biomedical Informatics,* p. 6, 2011.

[3] G. Zou, B. Zhang, Y. Gan, and J. Zhang, "An Ontology-Based Methodology for Semantic Expansion Search," in *FSKD '08: Proceedings of the 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, 2008, pp. 453-457.

[4] Y.-F. Huang and C.-H. Hsu, "PubMed smarter: Query expansion with implicit words based on gene ontology," *Knowledge-Based Systems.,* vol. 21, pp. 927-933, 2008.

[5] A. Segura N., S. Sánchez, E. García-Barriocanal, and M. Prieto, "An empirical analysis of ontology-based query expansion for learning resource searches using MERLOT and the Gene ontology," *Knowledge-Based Systems,* vol. 24, p. 15, 2011.

[6] F. Farfan, V. Hristidis, A. Ranganathan, and M. Weiner, "XOntoRank: Ontology-Aware Search of Electronic Medical Records," in *Proceedings of the 2009 IEEE International Conference on Data Engineering*: IEEE Computer Society, 2009.

[7] J. Bhogal, A. Macfarlane, and P. Smith, "A review of ontology based query expansion," *Information Processing and Management: an International Journal,* vol. 43, pp. 866-886, 2007.

[8] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowledge Acquisition,* vol. 5, pp. 199-220, 1993.

[9] M.-Á. Sicilia, E. García-Barriocanal, C. Pages, J. Martinez, and J. Gutierrez, "Complete metadata records in learning object repositories some evidence and requirements," *International Journal of Learning Technology,* vol. 1, pp. 411-424, 2005.

[10] Á. F. Zazo, C. G. Figuerola, J. L. Alonso Berrocal, and R. Emilio, "Reformulation of queries using similarity thesauri," *Information Processing and Management: an International Journal,* vol. 41, pp. 1163-1173, 2005.

[11] M. F. Porter and K. W. P. Sparck-Jones, *An algorithm for suffix stripping*: Morgan Kaufmann Publishers Inc., 1997.

[12] K. Todorov, P. Geibel, and K.-U. Khnberger, "Mining concept similarities for heterogeneous ontologies," in *Proceedings of the 10th industrial conference on Advances in data mining: applications and theoretical aspects* Berlin, Germany: Springer-Verlag, 2010.

[13] A. Abdelali, J. Cowie, and H. S. Soliman, "Improving query precision using semantic expansion," *Information Processing and Management: an International Journal,* vol. 43, pp. 705-716, 2007.

[14] M. C. Díaz-Galiano, M. T. Martín-Valdivia, and L. A. Ureña-López, "Query expansion with a medical ontology to improve a multimodal information retrieval system," *Comput. Biol. Med.,* vol. 39, pp. 396-403, 2009.

[15] M.-C. Lee, K. H. Tsai, and T. I. Wang, "A practical ontology query expansion algorithm for semantic-aware learning objects retrieval," *Computers &amp; Education,* vol. 50, pp. 1240-1257, 2008.

[16] L. Ma, L. Chen, Y. Gao, and Y. Yang, "Ontology Based Query Expansion in Vertical Search Engine," in *FSKD '09: Proceedings of the 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery*, 2009, pp. 285-289.

[17] J. Tuominen, T. Kauppinen, K. Viljanen, and E. Hyonen, "Ontology-Based Query Expansion Widget for Information Retrieval," in *Proceedings of the 5th Workshop on Scripting and Development for the Semantic Web (SFSW 2009), 6th European Semantic Web Conference (ESWC 2009)*, 2009.

[18] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady,* vol. 10, p. 3, 1996.

[19] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," *IEEE Transactions on Systems, Man and Cybernetics,* vol. v, pp. 17-30, 1989.

[20] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics* Las Cruces, New Mexico: Association for Computational Linguistics, 1994.

[21] H. Al-mubaid and H. A. Nguyen, "A Cluster-Based Approach for Semantic Similarity in the Biomedical Domain," in *The 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, San Francisco, CA, 2006.

[22] H. A. Nguyen and H. Al-Mubaid, "New ontology-based semantic similarity measure for the biomedical domain," in *IEEE International Conference on Granular Computing,* 2006, pp. 623-628.

[23] Y. Li, Z. A. Bandar, and D. McLean, "An approach for measuring semantic similarity between words using multiple information sources," *Knowledge and Data Engineering, IEEE Transactions on,* vol. 15, pp. 871-882, 2003.

[24] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proceedings of the 14th international joint conference on Artificial intelligence - Volume*

1 Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc., 1995.

[25] J. J. Jiang and D. W. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy," in *International Conference Research on Computational Linguistics (ROCLING X)*, Taiwan., 1997, p. 15.

[26] D. Lin, "An Information-Theoretic Definition of Similarity," in *Proceedings of the Fifteenth International Conference on Machine Learning*: Morgan Kaufmann Publishers Inc., 1998.

[27] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis, "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration," *Nat Biotech,* vol. 25, pp. 1251-1255, 2007.

[28] J. Rodgers and A. Nicewander, "Thirteen Ways to Look at the Correlation Coefficient," *The American Statistician,* vol. 42, pp. 59-66, 1988.