

Enhancing a Biological Concept Ontology to Fuzzy Relational Ontology with Relations Mined from Text

Lipika Dey¹ and Muhammad Abulaish²

¹ Department of Mathematics, Indian Institute of Technology Delhi
Hauz Khas, New Delhi – 16, India
lipika@maths.iitd.ernet.in

² Department of Mathematics, Jamia Millia Islamia (A central university)
Jamia Nagar, New Delhi – 25, India
abulaish@computer.org

Abstract. In this paper we investigate the problem of enriching an existing biological concept ontology into a fuzzy relational ontology structure using *generic* biological relations and their strengths mined from tagged biological text documents. Though biological relations in a text are defined between a pair of entities, the entities are usually tagged by their concept names in a tagged corpus. Since the tags themselves are related taxonomically, as given in the ontology, the mined relations have to be properly characterized before entering them into the ontology. We have proposed a mechanism to generalize each relation to be defined at the most appropriate level of specificity, before it can be added to the ontology. Since the mined relations have varying degrees of associations with various biological concepts, an appropriate fuzzy membership generation mechanism is proposed to fuzzify the strengths of the relations. Extensive experimentation has been conducted over the entire GENIA corpus and the results of enhancing the GENIA ontology are presented in the paper.

Keywords: Generic biological relation, Biological ontology enhancement, Fuzzy relational ontology.

1 Introduction

The field of Molecular Biology has witnessed a phenomenal growth in research activities in the recent past. Consequently to aid the process of organizing this large repository of knowledge, there has been a considerable effort towards building structured biological ontologies. Gene Ontology (GO) and GENIA ontology are two of the most popular ones. While the GENIA ontology stores only a set of concepts and the structural semantic relations, GO contains a large collection of biological processes along with biological concepts defined manually. Since manually identification of biological relations and their characterization is a labor-intensive task, several approaches have taken place to automate the process.

Generic biological relations can be characterized based on their occurrence patterns within text. The initial approaches focused on identifying a pre-defined set of verbs representing these relations within text. Thomas *et al.* [7] modified a pre-

existing parser based on cascaded finite state machines to fill templates with information on protein interactions for three verbs – *interact with*, *associate with*, *bind to*. Sekimizu *et al.* [6] have proposed mechanisms for locating a pre-defined collection of seven verbs *activate*, *bind*, *interact*, *regulate*, *encode*, *signal* and *function*. However since it is expensive and labour-intensive to pre-define all such relations exhaustively, Rinaldi *et al.* [5] proposed an automated Literature Based Discovery (LBD) method to characterize these seven relations in terms of the participating entities. Ono *et al.* [4] reports a method for extraction of protein-protein interactions using a dictionary look-up approach. After identifying the dictionary-based proteins within the document to analyze, sentences that contain at least two proteins are selected, which are then parsed with Parts-Of-Speech (POS) matching rules. The rules are triggered by a set of keywords, which are frequently used to name protein interactions like *associate*, *bind* etc. Ciaramita *et al.* [2] have proposed an unsupervised model for learning arbitrary relations between concepts of a molecular biology ontology from the GENIA corpus [3] for the purpose of supporting text-mining and manual ontology building.

In this paper, we present a method for characterizing biological relations mined from a tagged corpus using an ontology-based text-mining approach to extend the underlying ontology into a fuzzy relational ontology. Since biological relations occurring within a text can be directly associated to participating entities, locating only these relations does not provide the true character of the biological relation as an interaction between two biological entities. While it is straightforward to propagate these relations along the ontology tree, consolidating them at the most appropriate level requires significance analysis. For example, analyzing 170 instances out of a total of 219 instances of “expressed in” occurring in the GENIA corpus a break-up reveals that 48 associations are between the concept-pair *<protein_molecule, cell_type>*; 22 instances occur between *<protein_family_or_group, cell_type>*; 21 instances occur between *<protein_molecule, cell_line>*; 10 between *<protein_family_or_group, cell_line>*; 9 between *<DNA_domain_or_region, cell_type>*; 7 between *<RNA_molecule, cell_type>*; 6 between *<DNA_family_or_group, cell_type>*; 5 between *<RNA_molecule, cell_line>*; 4 each between *<RNA_family_or_group, cell_type>* and between *<protein_molecule, tissue>*; 3 each between pairs *<protein_molecule, body_part>* and *<protein_molecule, mono_cell>*; 2 each between pairs *<DNA_domain_or_region, cell_line>*, *<protein_domain_or_region, cell_type>*, *<DNA_domain_or_region, tissue>*, and *<DNA_domain_or_region, body_part>*; 1 instance each between 20 other concept-pairs. While it may not be significant to keep track of the single, dual or triple occurrences, it will also not be appropriate to club all these relations together and state that “expressed in” occurs between concepts *substance* and *source*, which is correct but a case of over-generalization. An appropriate characterization should take into account the proportion of instances reaching at a particular concept-pair against the total occurrences at its parent concept-pair. Thus characterized, the relations can be used to enhance the underlying ontology. We have provided experimental validation of the approach over the GENIA corpus [3].

2 Analyzing Frequently Occurring Biological Relations Extracted from GENIA Corpus

The GENIA ontology is a taxonomy of 47 biologically relevant nominal categories in which the top three concepts are *biological source*, *biological substance* and *other_name*. The *other_name* refers to all biological concepts that are not identified with any other known concept in the ontology. The sub-tree rooted at *source* contains 13 nominal categories and the other rooted at *substance*, contains 34 nominal categories. The GENIA corpus contains 2000 tagged MEDLINE abstracts. Tags are leaf concepts in GENIA ontology. Tags may be nested whereby a tagged Biological entity in conjunction with other entities or processes may be tagged as a different leaf concept. A biological relation is expressed as a binary relation between two biological concepts [4]. Following this definition, while mining for biological relations, we define a relation as an activity co-occurring with a pair of tags within the GENIA corpus. In [1] we had identified a set of 24 root verbs and their 246 variants, which represent biological relations occurring in the GENIA corpus. A complete list of all feasible biological relations and their morphological variants extracted from the GENIA corpus is available on <http://www.geocities.com/mdabulaish/BIEQA/>. We can enhance the GENIA ontology with these relations.

Since the GENIA corpus is tagged with leaf-level concepts, all relations are defined between entities or between leaf-level concept pairs. However keeping track of all instances may not be useful from the aspect of domain knowledge consolidation. This was illustrated through an example in section 1. Hence our aim is to generalize a relation at an appropriate level of specificity before including it in the ontology. This reduces over-specialization and noise.

All molecular biology concepts in the GENIA ontology are classified into two broad categories, *source* and *substance*. Hence the entity pairs occurring with each relation can be broadly classified as belonging to one of the following four categories (i) <source, source> (ii) <source, substance> (iii) <substance, source> and (iv) <substance, substance>. Every instance of a relation belongs to one of these categories and the total number of instances associated to any category can be obtained with appropriate summation. Since a generic concept can represent multiple specific concepts, hence the first step towards characterizing relations is to consolidate the total number of relations belonging to each category, identify the pathways through which they are assigned to a category and then find the most appropriate generalization of the relation in that category.

In order to achieve this, we define a *concept-pair tree* to represent each category. The root node of a concept-pair tree denoted by (L_r, R_r) contains one of the four generic concept-pairs defined earlier. Each node N in a concept-pair tree has two constituent concepts $\langle C_i, C_j \rangle$ denoted as the LEFT and the RIGHT concepts. The LEFT and RIGHT concepts are specializations of L_r and R_r respectively, as obtained from the underlying ontology. Each *concept-pair tree* stores all possible ordered concept-pairs that match the root concept-pair (L_r, R_r) and is generated using a recursive algorithm, described in the next section.

3 Generating Concept-Pair Trees

The concept-pair tree is represented as an AND-OR tree, where each node has links to two sets of children, denoted by L_1 and L_2 . L_1 and L_2 each contain a set of concept-pair nodes. The two sets L_1 and L_2 are themselves connected by the OR operator, while the nodes within each of them are connected with each other through an AND operator. For every node N , the two sets of child nodes L_1 and L_2 are created as follows:

- L_1 consists of concept pairs created by expanding the LEFT concept to consider all its child nodes in the concept ontology, while keeping the RIGHT concept unchanged.
- L_2 is created by keeping the LEFT concept unchanged while considering all children of the RIGHT concept in the concept ontology.
- When any of the concepts LEFT or RIGHT is a leaf-level ontology concept, the corresponding set L_1 or L_2 respectively is NULL.

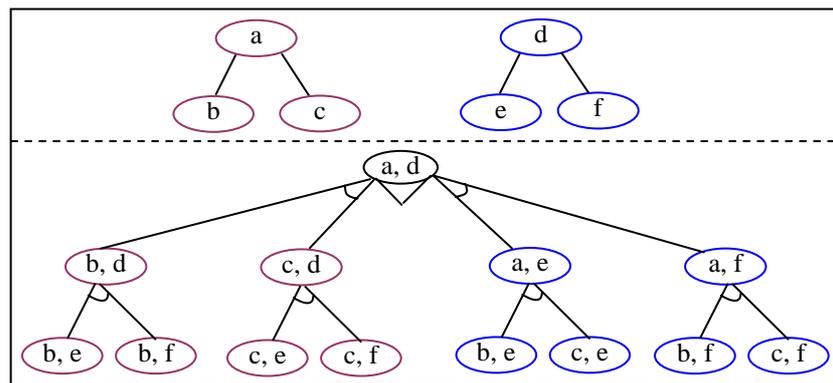


Fig. 1. Sample AND-OR concept-pair tree

Starting from a root concept pair $\langle L_r, R_r \rangle$, the complete concept-pair tree is created recursively as follows:

$$\text{OR} [\text{AND} [\langle \text{children of } L_r, R_r \rangle], \text{AND} [\langle L_r, \text{children of } R_r \rangle]]$$

Let us suppose 'a' and 'd' represent two root concepts in a concept ontology, at each of which an ontology sub-tree is rooted, as shown in Fig. 1. The sets L_1 and L_2 for the root node of the concept pair tree, $\langle a, d \rangle$, are determined as $L_1: \langle b, d \rangle, \langle c, d \rangle$; $L_2: \langle a, e \rangle, \langle a, f \rangle$. Fig. 1 shows the resulting AND-OR tree. AND is represented by '∩', OR is represented using the symbol '∨'. It may be noted that leaf-level pairs occur more than once in the tree. Each occurrence defines a path through which relations between that pair may be propagated up for generalization. Two sets of relations converging at a parent node, could be viewed as alternative models for generalization or could be

viewed as complementing each other to form the total set at the parent level, depending on whether they are coming via the AND path or the OR path. This is further explained in the next section.

3.1 Mapping the Relation Instances over a Concept-Pair Tree

After creating the four different concept-pair trees for the GENIA ontology, the most feasible representation of a relation for each of these categories is obtained using these. Suppose there are N instances of a relation r_g observed over the corpus. Each of these instances is defined for a pair of leaf-level concepts. Based on the generic category of the leaf-level concepts, each relation instance can be mapped to a leaf node in one of the four concept-pair trees.

For each concept-pair tree T^G , all instances that can be mapped to leaf-level nodes of T^G are mapped at the appropriate nodes. These counts are propagated up in the tree exploiting its AND-OR property. Since each leaf-level node has multiple occurrences in a concept-pair tree, each relation instance is mapped to all such leaf-level nodes. For each non-leaf node in the concept-pair tree, the total number of relations is equal to the number of instances propagated up through all its children in either L_1 or L_2 . In order to derive the most appropriate levels for describing a relation, the concept-pair tree is traversed top-down. Starting from the most generic level description at the root level, an information loss function based on set-theoretic approach is applied at each node to determine the appropriateness of defining the relation at that level.

4 Characterizing Relations at Appropriate Levels of Specificity

The process of determining the most specific concept pairs for relations follows a top-down scanning of the AND-OR tree. Starting from the root node, the aim is to determine those branches and thereby those nodes which can account for sufficiently large number of relation instances. When the frequency of a relation drops to an insignificant value at a node the node and all its descendants need not be considered for the relation conceptualization, and may be pruned off without further consideration. The lowest un-pruned node becomes a leaf and is labeled as the most specific concept-pair for defining a relation.

$$Information\ Loss(N) = \frac{|IC_P - IC_N|}{|IC_P + IC_N|} \quad (1)$$

where, IC_N = Count of instances of relation r_g at N , IC_P = count of instances of r_g at parent P of N .

Equation 1 defines a loss-function that is applied at every node N to determine the loss of information incurred if this node is pruned off. The loss function is computed as a symmetric difference between the number of instances that reach the node and the number of relation instances that were defined at its parent. Equation 1 states that if the information loss at a node N is above a threshold, it is obvious that the node N

accounts for a very small percentage of the relation instances that are defined for its parent. Hence any sub-tree rooted at this node may be pruned off from further consideration while deciding the appropriate level of concept pair association for a relation. For our implementation this threshold has been kept at 10%.

Since a parent node has two alternative paths denoted by the expansion of LEFT and RIGHT respectively, along which a relation may be further specialized, the choice of appropriate level is based on the collective significance of the path composed of retained nodes. For each ANDED set of retained nodes, total information loss for the set is computed as the average information loss for each retained child. The decision to prune off a set of nodes rooted at N is taken as follows: Let information loss for nodes retained at L₁ is E₁ and that for nodes retained at L₂ is E₂.

- If E₁ = 0, then L₁ is retained and L₂ is pruned off, otherwise, if E₂ = 0 then L₂ is retained and L₁ is pruned off.
- Otherwise, if E₁ ≈ E₂, i.e., $\text{Min}(E_1, E_2) / \text{Max}(E_1, E_2) \geq 0.995$ then both the subtrees are pruned off, and the node N serves as the appropriate level of specification.
- Otherwise, if E₁ < E₂, then L₁ is retained and L₂ is pruned off. If E₂ < E₁ then L₂ is retained while L₁ is pruned off.

The set of concept-pairs retained are used for conceptualizing the relations.

5 Fuzzification of Relations

Since all relations are not equally frequent in the corpus, hence we associate with each relation a strength S which is computed in terms of relative frequency. Equation 2 computes this strength, where G denotes the category of concept-pairs: *source-substance*, *source-source*, *substance-substance* and *substance-source*. |T^G| denotes the total count of all relations that are defined between ordered concept pairs defined in the tree T^G, and N_{r_g}^G denotes the total number of relation instances of type r_g mapped to T^G.

$$\mu_{(C_i, C_j)}^G(r_g) = \frac{1}{2} \left\{ \frac{|\langle C_i, r_g, C_j \rangle|}{N_{r_g}^G} + \frac{|\langle C_i, r_g, C_j \rangle|}{|T^G|} \right\} \quad (2)$$

Since exact numeric values of strength do not convey much information, hence we choose a fuzzy representation to store the relations. The feasible biological relations are converted into fuzzy relations based on the membership of their strength values to a fuzzy quantifier term set {weak, moderate, strong}. The membership functions for determining the values to each of these categories are derived after analyzing the graphs displaying the distributions of strength. Fig. 2 shows the percentage of feasible relations of each category against the strengths of the relations.

The fuzzy membership functions are derived after analyzing the graphs shown in Fig. 2. Each curve shows only one valley, and this common valley for all trees is

observed at strength 0.4. Hence 0.4 is selected for defining the intermediate class “moderate”. The membership functions for the categories “weak”, and “strong” for each category are obtained through curve-fitting on different sides of the valley, while the membership function for class “moderate” is obtained by using the values surrounding 0.4. The fuzzy membership functions for categories “moderate” and “strong” are always characterized by Gaussian functions, whereas for the category “weak”, different types of functions are derived.

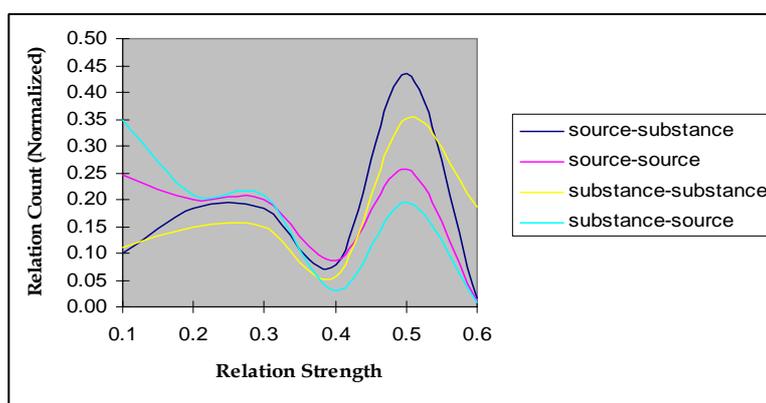


Fig. 2. A plot of relation strengths and their %age counts for all four categories of trees

Table 1. Biological relations and associated generic concept-pairs along with their fuzzy strength

Relation	Generic concept-pairs and their strengths			
	Substance-Source	Substance-Substance	Source-Source	Source-Substance
Induce	(<OC, Nat>, S) (<OC, Art, W>)	(<OC, AA>, S) (<OC, NA>, W)	(<Src, Src>, S)	-----
Inhibits	(<Lip, CT>, W) (<PFG, CT>, W) (<PM, CT>, M) (<DNADR, CT>, W)	(<Sbs, Cmp>, S)	(<CT, Art>, S) (<CT, Nat>, S)	(<Nat, AA>, S) (<Nat, NA>, M)
Activate	(<OC, Nat>, S)	(<Pr, AA>, S) (<Pr, NA>, W)	(<CL, CT>, W) (<CT, CT>, S) (<MC, CT>, W)	(<Src, OC>, S)
Expressed in	(<OC, Src>, S)	(<DNA, OC>, W) (<Pr, AA>, M) (<Pr, NA>, M) (<RNA, OOC>, W)	(<Nat, Org>, W) (<Nat, Tis>, W) (<Nat, CT>, S)	-----

Legend:
OC: Organic compound; **AA:** Amino_acid; **NA:** Nucluc_acid; **OOC:** Other_organic_compound; **Sbs:** Substance; **Nat:** Natural source; **Org:** Organism; **CT:** Cell_type; **Pr:** Protein; **Src:** Source; **Tis:** Tissue; **MC:** Mono_cell; **PFG:** Protein_family_or_group; **Lip:** Lipid; **DNADR:** DNA_domain_or_region; **Art:** Artificial source; **Cmp:** Compound; **PM:** Protein_molecule; **S:** Strong; **M:** Moderate; **W:** Weak

$$\mu_{weak}(x) = a + bx, \text{ where } a = 1.194, b = -2.194 \quad (3)$$

$$\mu_{moderate}(x) = ae^{-\frac{(x-b)^2}{2c^2}}, \text{ where } a = 2.506, b = 0.357, c = 0.032 \quad (4)$$

$$\mu_{strong}(x) = ae^{-\frac{(x-b)^2}{2c^2}}, \text{ where } a = 1.131, b = 0.476, c = 0.049 \quad (5)$$

Sample fuzzy membership functions derived for the category *substance-substance* (shown in yellow color in Fig. 2) is shown below. The membership functions for the fuzzy sets “*weak*”, “*moderate*” and “*strong*” are defined in equations 3, 4 and 5 respectively. Table 1 shows the top 5 relations along with their associated concept pairs and strengths identified for enhancing the GENIA ontology.

6 Enhancing GENIA to a Fuzzy Relational Ontology

We now explain how we propose to extend the GENIA ontology by adding the generic relations to it. Since the relations have variable strengths, hence we propose to maintain a Fuzzy Relational Ontology rather than a crisp ontology structure. In this model there are two categories of relations – *structural* and *generic*. While structural relations are crisp, generic relations have associated fuzzy strengths. We define the Fuzzy Relational Ontology Model as follows:

Definition (Fuzzy Relational Ontology Model) – A Fuzzy Relational Ontology Model Θ_f is a 5-tuple of the form

$\Theta_f = (C, P, \mathfrak{R}_s, \mathfrak{R}_g, S)$, where,

- C is a set of concepts
- P is a set of properties. A property $p \in P$ is defined as a unary relation of the form $p(c)$, where $c \in C$ is the concept associated to the property.
- $\mathfrak{R}_s = \{is-a, kind-of, part-of, has-part\}$ is a set of structural semantic relations between concepts. A structural semantic relation $r_s \in \mathfrak{R}_s$ is defined as a binary relation of the form $r_s(C_i, C_j)$, where $C_i, C_j \in C$ are the concepts related through r_s .
- \mathfrak{R}_g is a set of feasible generic relations between concepts. Like structural semantic relations, a generic relation $r_g \in \mathfrak{R}_g$ can be defined as a binary relation of the form $r_g(C_i, C_j)$, where $C_i, C_j \in C$ are the concepts related through r_g .
- $S = \{weak, moderate, strong\}$, is a term set to represent the strength of the generic biological relations in terms of linguistic qualifiers. A linguistic qualifier $\xi \in S$ is defined as a unary relation of the form $\xi(r_g)$, where $r_g \in \mathfrak{R}_g$ is a feasible generic relation

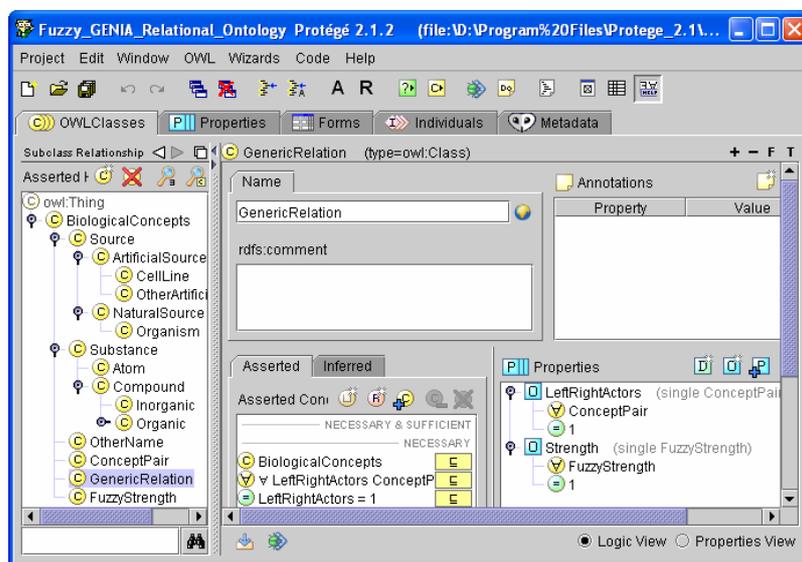


Fig. 3. A snapshot of the Fuzzy Relational GENIA ontology structure

To accommodate generic relations and their strengths, in addition to existing GENIA ontology classes, the fuzzy GENIA relational ontology structure contains three generic classes - a “*ConceptPair*” class, a “*FuzzyStrength*” class and a “*GenericRelation*” class, where the last one multiply inherits from the earlier two classes. The *ConceptPair* class consists of *HasLeftConcept* and *HasRightConcept* properties whose values are the instances of the GENIA *concept* classes. *FuzzyStrength* class has been defined to store the fuzzy quantifiers that can be associated with the generic relations to represent their strength. This class consists of a single property *TermSet* which is defined as a *symbol* and contains the fuzzy quantifiers “weak”, “moderate” and “strong”. The *GenericRelation* class has two properties – *LeftRightActors* and *Strength*. The *LeftRightActors* property is a kind of OWL object property which range is bound to the *ConceptPair* class. The *Strength* property is also a kind of OWL object property for which the range is bound to the *FuzzyStrength* class. All mined generic relations are defined as instances of the class *GenericRelation*. Fig. 3. shows a snapshot of a portion of the enhanced Fuzzy GENIA relational ontology structure. A total of 280 strong, 38 moderate and 576 weak relational links were identified for adding to GENIA. It is observed that each instance of relation has a strong or moderate co-occurrence with a maximum of 4 different pairs. However, the maximum number of weak co-occurrences could go up to 17. For example, Table 1 shows 3 strong and 2 weak instances of the relation “induce”. In our implementation we have restricted the enhancement to include only strong and moderate relations, to keep the ontology comprehensible.

7 Conclusions

In this paper we propose a fuzzy relational ontology model to accommodate generic biological relations into an existing biological ontology. The relations are mined from the GENIA corpus, which contains tagged MEDLINE abstracts. The mined relations which are always defined between a pair of leaf level concepts in the GENIA corpus are generalized using a novel technique. The generalization task is framed as an optimization problem over a AND-OR concept-pair tree. Since the relations occur with varying strengths, the enhanced ontology is modeled as a fuzzy ontology structure. The derivation of the fuzzy membership functions have also been addressed in detail. A glimpse of the experimental results has been provided. Extension of the ontology structure into a rough-fuzzy ontology is being currently studied.

8 Acknowledgement

The first author is presently on leave from IIT Delhi and has taken up a temporary assignment with Webaroo Technologies Limited. She would like to thank Webaroo Technologies Ltd. for providing the support to attend the conference.

References

1. Abulaish, M., Dey, L.: An Ontology-based Pattern Mining System for Extracting Information from Biological Texts. in: Proceedings of the 10th Int. Conf. on RSFDGrC'05, Canada. LNAI 3642, Part II, Springer (2005) 420-429
2. Ciaranita, M., Gangemi, A., Ratsch, E., Saric, J., Rojas, I.: Unsupervised Learning of Semantic Relations between Concepts of a Molecular Biology Ontology. in: Proceedings of the 19th Int. Joint Conf. on Artificial Intelligence (2005) 659-664
3. Kim, J.-D., Ohta, T., Tateisi, Y., Tsujii, J.: GENIA Corpus – A Semantically Annotated Corpus for Bio-Textmining. *Bioinformatics*, Vol. 19, Suppl. 1 (2003) i180-i182
4. Ono, T., Hishigaki, H., Tanigami, A., Takagi, T.: Automated Extraction of Information on Protein-Protein Interactions from the Biological Literature. *Bioinformatics* 17(2) (2001) 155-161
5. Rinaldi, F., Scheider, G., Andronis, C., Persidis, A., Konstani, O.: Mining Relations in the GENIA Corpus. in: Proceedings of the 2nd European Workshop on Data Mining and Text Mining for Bioinformatics, Pisa, Italy (2004)
6. Sekimizu, T., Park, H. S., Tsujii, J.: Identifying the Interaction between Genes and Genes Products Based on Frequently Seen Verbs in Medline Abstract. *Genome Informatics* 9 (1998) 62-71
7. Thomas, J., Milward, D., Ouzounis, C., Pulman, S., Carroll, M.: Automatic Extraction of Protein Interactions from Scientific Abstracts. in: Pacific Symposium on Biocomputing (2000) 538-549