# Selecting a Relevant Set of Examples to Learn IE-Rules

J. Turmo and H. Rodríguez

Dept. Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya. Spain
{turmo,horacio}@lsi.upc.es

**Abstract.** The growing availability of online text has lead to an increase in the use of automatic knowledge acquisition approaches from textual data, as in Information Extraction (IE). Some IE systems use knowledge learned by single-concept learning systems, as sets of IE rules. Most of such systems need both sets of positive and negative examples. However, the manual selection of positive examples can be a very hard task for experts, while automatic methods for selecting negative examples can generate extremely large example sets, in spite of the fact that only a small subset of them is relevant to learn. This paper briefly describes a more portable multi-concept learning system and presents a methodology to select a relevant set of training examples.

## 1 Introduction

The growing availability of on-line text has lead to an increase in the use of automatic knowledge acquisition approaches from textual data as in Information Extraction (IE). The aim of an IE system consists in automatically extracting pieces of information relevant for a set of prescribed concepts (scenario).

One of the main drawbacks of applying IE systems is the high cost involved in manually adapting them to new domains and text styles. In recent years, the portability of IE systems to new domains has been improved by the use of a variety of Machine Learning techniques. In fact, learning systems like SRV [1], RAPIER [2], CRYSTAL [3] and WHISK [4], among others, have been used into IE systems to learn rules (IE-rules) for each concept in the scenario from a collection of training documents. However, some drawbacks remain in the portability:

- *Text style problem*: existing IE-rule learning systems depend on the supported text style (structured texts, semi-structured texts or free texts).
- *Combination problem*: IE-rule learning systems are mostly single-concept learning systems; consequently, an extractor (e.g. a set of IE-rules) is learned for each concept within the scenario in an independent manner. Moreover, the order of execution of the learners is set manually and so are the scheduling and way of combining the resulting extractors.

- *Training set size problem*: The size of the set of positive examples[1], $\mathcal{E}^+$, can be small to accurately learn an extractor for a concept within the scenario. This could be the case when dealing with some combinations of text style and domain.
- *Training set relevance problem*: The size of the negative example set[2], $\mathcal{E}^-$, if needed, can be extremely large to be tractable, while only a small subset is relevant to learn.

EVIUS is a multi-concept learning system to learn IE-rules from free text, that deals with such drawbacks. It has been incorporated to a multilingual IE system, M-TURBIO [5]. EVIUS strategies for the first two drawbacks are described in [7]. This paper shortly describes EVIUS and presents a method to deal with the *training set relevance* problem.

## 2 Test-Domain Model

Our methodology has been tested on the domain of mushrooms[3]. The reasons for choosing this domain are, basically, the difficulty of the documents, their lexical and grammatical richness, the frequent use of ellipsis and anaphora and the variety of information to be extracted. Furthermore, we can find fuzzy features, as in *olor algo dulce* (a rather sweet smell), multivaluated features, as in *se encuentran en prados o zonas soleadas* (found in pastures or sunny places) , features whose values can be expressed as intervals, as in *su color varía de rojo sangre a marrón ligeramente claro* (its colour ranges from blood red to slightly pale brown) and features that change values throughout a mushroom's life, as in *blanco constante que pasa a amarillo huevo con la edad* (permanent white changing to egg yellow with time). Figure 1 partially shows the scenario of extraction represented as hierarchy of frames.

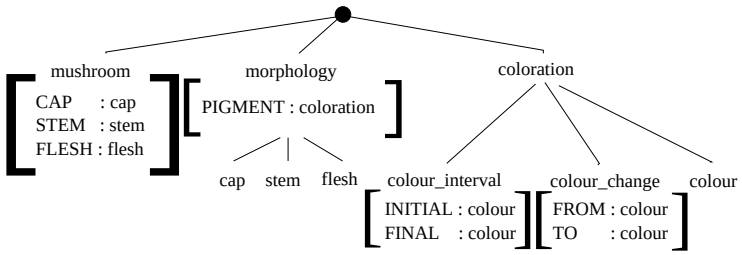


**Fig. 1.** Representation of part of the mycological scenario

[1] This set is, in general, manually selected from the training corpus.
[2] This set is, in general, automatically selected from the training corpus.
[3] A set of documents (Catalan/Spanish) describing mycological specimens has been used for training and testing

In this paper, we will focus on the extraction of the feature PIGMENT in the scenario, since it presents the largest richness to express values. Generalization of results can be easily accomplished.

## 3 EVIUS Description

The input of EVIUS consists of 1) a set of preprocessed texts (POS-tagged, partially-parsed and semantically tagged[4]), as training corpus and, 2) a scenario of extraction $\mathcal{S}$. Using such information as background knowledge, EVIUS learns a set of extractors for the whole set of concepts within the scenario. In this process two kind of such concepts are distinguished: *fundamental concepts*, which are terminal concepts in the scenario hierarchy, concepts having neither by-default features nor inherited ones; and *complex concepts*, owning features, inherited or by default. Such concepts are learned as sets of rules in the form of conjunction of first-order predicates, which are predefined within a predicate model.

**Definition of the predicate model.** As a preprocess of EVIUS, the training corpus is translated into the following predicate model:

- Attributive predicates: pos_X(A), word_X(A), lemma_X(A), isa_X(A) and has_hypernym_X(A), where X is instantiated with categories related to node A of the parsed corpus.
- Relational meta-predicates: distance_le_X(A,B), stating that there are X terminal nodes, at most, between A and B.
- Relational predicates: ancestor(A,B), where B is the syntactic ancestor of A, and brother(A,B), where B is the right brother node of A sharing the syntactic ancestor.

Examples used to learn a concept $c$ can be modeled as tuples $< a_1, \ldots, a_n >$, where each $a_i$ is a terminal node within the input parse. On the one hand, if $c$ is a fundamental concept, then $< a_1, a_2 >$ is defined as the pair of terminal nodes being delimiters of a textual instance of $c$. For instance, example *rojo algo claro* (slightly pale red) for concept *colour* is represented as $< a, b >$, where $a$ and $b$ are, respectively, the nodes corresponding to *rojo* and *claro*. On the other hand, if $c$ is a complex concept, then $a_i$ is the value of the i-th feature of an example. For instance, assuming the concept *colour* has been learned and added to the background knowledge, then example *vira de rojo algo claro a marrón* (changes from slightly pale red to brown) for concept *colour_change*, will be represented as $< a, b >$, where $a$ and $b$ are, respectively, terminal nodes corresponding to the value for the feature FROM (the first *colour*) and the value for the feature TO (the second *colour*).

[4] With EuroWordNet (EWN - http://www.hum.uva.nl/~ewn) synsets. No attempt has been made to sense disambiguate such tags.

The resulting rules can take the pattern $c(A_1, \ldots, A_n)$ :- <*conjunction of predicates*>, where each $A_i$ is a variable representing either the limits of possible textual instances for a fundamental concept, or values of i-th feature of possible instances for a complex one. For instance, one of the rules learned for the concept *colour* in our experiments is:

```
colour(A,B) :- has_hypernym_03464624n(A), ancestor(A,C)
               has_hypernym_03464624n(B), ancestor(B,C).
```

which represents those expressions delimited by words that are achromatic colours (hyponyms of 03464624n) and share the same syntactic ancestor in the parsed sentence.

**Learning a Concept.** An initial extractor is learned with FOIL (First-order Induction Learning) [6] by manually selecting the set of positive examples, $\mathcal{E}^+$, and generating automatically the set of negative ones, $\mathcal{E}^-$. The task of selecting $\mathcal{E}^+$ can be very hard due to the complexity of some concepts in the scenario. In this sense, we can proceed by using a supervised *bootstrapping approach*. Initially, a small set of positive examples is selected by hand and an extractor is learned from them. The resulting set of rules can be applied to the training documents and extracted values can be manually classified as positive and negative examples. These new examples are added to the initial ones and the learning process is iterated until no further improvement is achieved. This approach is simpler for the human experts than selecting positive examples from the whole set of documents but, independently of whether it is used or not, $|\mathcal{E}^+|$ can be inadequate to accurately learn an extractor.

In order to improve the accuracy of such an initial extractor, EVIUS uses an active learning approach by 1) incrementally adding artificial examples to $\mathcal{E}^+$, 2) learning a new extractor from the new set of examples and 3) merging both extractors by appending unrepeated and non empirically subsumed[5] rules from the new extractor. Artificial examples are created by combining features from both the examples covered and uncovered by the current rules, as explained in [7]. Experiments demonstrate that using this technique the F value[6] improves in four points with only two iterations.

**Learning the Whole Scenario.** As a whole learning process, EVIUS learns a set of extractors for the set of concepts within in the scenario. This is done by using a multistrategy constructive learning (MCL) approach, which integrates: a) *closed-loop learning*, in which concepts to be learned at each step (as explained before) are determined, b) *deductive restructuring* [9], by adding a new set of

[5] A rule $r_1$ is empirically subsumed by another one, $r_2$, if examples covered by $r_1$ are also covered by $r_2$.

[6] This measure is defined in the MUC conferences (http://www.muc.saic.com) as a consensus between recall and precision.

examples being instances of the learned concept $c$ and c) *constructive learning*, by adding new attributes isa_$c$(X) for each new generated example X. The set of new examples are used for further learning.

One of the drawbacks that appears when using FOIL to learn, as in EVIUS, is the training set relevance problem. In the next sections some methods for selecting a relevant set of negative examples, $\hat{\mathcal{E}}^-$, from the whole, $\mathcal{E}^-$, are compared.

## 4 The Training Set Relevance Problem

A common assumption in Machine Learning is to consider all non-positive examples as negative ones. This closed-world assumption can be used by FOIL only when dealing with small learning spaces. When the learning space consists of hundreds of predicates, as in our case, the resulting $\mathcal{E}^-$ becomes untractable in practice. Moreover, only a small subset is effectively relevant to learn the concept. Three strategies can be applied to select a set of relevant negatives, $\hat{\mathcal{E}}^-$:

- *Use of intuitive observations.* Sometimes, all positive examples share a set of properties. $\hat{\mathcal{E}}^-$ can be generated by selecting those non positive examples sharing these properties. For instance, taking strings of words as examples, Freitag [1] computes the minimum and the maximum lengths in words of the positive examples and takes as negative ones those other examples whose length is between these values. In [7], some experiments were done by using Freitag's strategy to learn an initial extractor for the concept *colour* in the mycological domain. In order to test them, 58 documents of the mycological domain were used (45 for training and 13 for testing), taking five different corpora with different sizes from the 45 training documents (5, 15, 25, 35 and 45 documents). Recall (R), precission (P) and F measure for our five different sizes of training corpus are presented, as baseline, in the last line, in table 3.
- *Use of a distance measure.* A more general approach consists in defining a distance measure between examples and selecting the closest negative ones to each positive. However, as a consequence of the extremely large size of $\mathcal{E}^-$, the bigger $\mathcal{E}^+$ is, the higher the cost to compute all distance values between positive and negative examples is, until becoming prohibitive.
- *Use of clustering techniques.* In order to reduce the cost in practice of the second strategy, we can take into account that some positive examples can be very similar to each others. Then, applying clustering techniques to $\mathcal{E}^+$, we can take the *medoids*[7] as the most representative positive examples to compute the distances explained in the second strategy.

The subsections above describe the distance measure, the clustering method used in the third strategy, and the method applied to select the negative examples that are closest to the positive ones.

[7] A medoid is an actual data point representing a cluster in a similar way to a *centroid*.

### 4.1 Distance Measure Definition

Bearing in mind the input partially-parsed semantically-tagged training corpus and the fact that examples in $\mathcal{E}^+ \cup \mathcal{E}^-$ follow pattern $e_i =< a_{i,1}, \ldots, a_{i,n} >$, we define a function $\mathcal{P}$ between the example space and an $n+4$ dimensional space, $\mathcal{P}(e_i) = (n_i, \delta_i, \mathcal{L}_i, \mathcal{W}_i, \mathcal{S}_{i,1}, \ldots, \mathcal{S}_{i,n})$, where $n_i$ is the number of words between $a_{i,1}$ and $a_{i,n}$; $\delta_i$ is a number that codes the syntactic paths between nodes $a_{i,1}, \ldots, a_{i,n}$, as described in [5]; $\mathcal{L}_i$ is the set of lemmas involved between $a_{i,1}$ and $a_{i,n}$; $\mathcal{W}_i$ is the set of words between $a_{i,1}$ and $a_{i,n}$; and, $\mathcal{S}_{i,j}$ is the set of all possible senses for the lemma occurring in $a_{i,j}$, preserving ambiguities[8], or as names of concepts in the scenario[9].

Due to the different nature of these dimensions, an Heterogeneous Overlap-Euclidean Metric (HOEM) [10] has been used with the aim of measuring the difference between examples. It has been defined as the euclidean distance among the following $n+4$ distance values:

$$d_1(\mathcal{P}(e_i), \mathcal{P}(e_j)) = |n_i - n_j| \qquad d_2(\mathcal{P}(e_i), \mathcal{P}(e_j)) = \frac{max(\delta_i, \delta_j)}{min(\delta_i, \delta_j)}$$

$$d_3(\mathcal{P}(e_i), \mathcal{P}(e_j)) = |\mathcal{L}_i \cup \mathcal{L}_j| - \frac{|\mathcal{L}_i \cap \mathcal{L}_j|}{|\mathcal{L}_i \cup \mathcal{L}_j|}$$

$d_4$ is the same formula as $d_3$, but substituting $\mathcal{L}$ with $\mathcal{W}$ [10]. And finally,

$$\forall k \ \in [5, n+4] : \ d_k(\mathcal{P}(e_i), \mathcal{P}(e_j)) = min_{s \in \mathcal{S}_{k,i}, \ r \in \mathcal{S}_{k,j}} \{dc(s, r)\}$$

where $dc(s, r)$ is the conceptual distance between two synsets $s$ and $r$ in EWN as defined in [11].

### 4.2 Clustering $\mathcal{E}^+$

Set $\mathcal{E}^+$ can be partitioned into a set of clusters, $\mathrm{E} = \{< \mathcal{E}_i^+, m_i >\}$, being $m_i$ the medoid of cluster $\mathcal{E}_i^+$. This can be done by adopting some clustering technique [12]. We have applied an agglomerative clustering technique based on medoids. The closest example to the others in a cluster, according to the distance average, will be selected as medoid. However, as the desired number of clusters is unknown, agglomerative techniques generate dendrograms and, as a consequence, a dendrite method has to be used in order to select the best set of clusters within the dendrogram. We propose the following simplification of the dendrite method explained in [13]. Being $\mathrm{E}_g$ the set of clusters in the level

[8] This is why $\mathcal{S}_{i,1}, \ldots, \mathcal{S}_{i,n})$ are not collapsed into a single set

[9] In our MCL approach, explained in section 3, new semantics are added as predicates isa_$c$ and new labels $c$ are linked as virtual synsets to EWN.

[10] These distances is defined from Jaccards' coefficient of similarity between sets, being the fraction between the cardinality of the intersection and the cardinality of the union

$g$ of the dendrogram, and $m$ the associated general medoid among individual clusters, we define,

$$\mathrm{B}_g = \sum_{i=1}^{|\mathrm{E}_g|} \Big[ |\mathcal{E}_i^+| \sum_{k=1}^{n+4} d_k(\mathcal{P}(m), \mathcal{P}(m_i))^2 \Big]$$

$$\mathrm{W}_g = \sum_{i=1}^{|\mathrm{E}_g|} \Big[ \sum_{j=1}^{|\mathcal{E}_i^+|} \Big( \sum_{k=1}^{n+4} d_k(\mathcal{P}(e_j), \mathcal{P}(m_i))^2 \Big) \Big]$$

and we redefine the Calinski value, which measures how different are both the clusters between themselves ($\mathrm{B}_g$ value) and the examples between themselves and within each cluster ($\mathrm{W}_g$), as $c_g = \frac{B_g\,(n-g)}{W_g\,(g-1)}$. The $\mathrm{E}_g$ having the first local maximum value $c_g^{max}$ is selected as the best set of clusters, E. As a consequence, the set of the most representative positive examples will be $\hat{\mathcal{E}}^+ = \{m_i \in \mathrm{E}\}$.

This strategy has been applied to the different corpus sizes described in section 3. We have obtained different $\hat{\mathcal{E}}^+$ sets, for each one of them. An average of 14.4% and a maximum of 21.9% of reduction, from $\mathcal{E}^+$ to $\hat{\mathcal{E}}^+$, was achieved.

Once obtained E, we classify all negative examples into such a set, generating sets $\mathcal{E}_i^-$ for each cluster $\mathcal{E}_i^+$, as follows: a negative example, $e^-$, belongs to $\mathcal{E}_i^-$ if $d(m_i, e^-)$ is the minimal one with respect to other clusters[11]. As a final step, set $\hat{\mathcal{E}}^-$ of relevant negative examples to learn the concept, will be selected from $\mathcal{E}_i^-$ sets. The subsection bellow describes a study of different approaches.

### 4.3 Selecting a Relevant Set of Negative Examples

At least two hypotheses could be applied to select $\hat{\mathcal{E}}^-$, when using $\mathcal{E}_i^-$ sets: a) the larger the size of a cluster $\mathcal{E}_i^+$ is, the larger is the number of negative examples from $\mathcal{E}_i^-$ to be selected, and b) the more similar the medoids are, the fewer negative examples have to be selected. From the first hypothesis, $E_i^-(\alpha) = \{e^- \in \mathcal{E}_i^- \mid d(e^-, m_i) \leq \alpha\}$ can be taken as relevant enough for the $i$-th cluster, where $\alpha$ is a distance value, being computed by taking into account the second hypothesis. The following six formulas have been experimented for such a computation, from which only 2 and 3 are dependent on the cluster:

$$\frac{\sum_{i,j \leq |\mathrm{E}|} d(m_i, m_j)}{|\mathrm{E}|} \tag{1}$$

$$max_{j \leq |\mathrm{E}|}\{d(m_i, m_j)\} \tag{2}$$

$$\frac{\sum_{j \leq |\mathrm{E}|} d(m_i, m_j)}{|\mathrm{E}| - 1} \tag{3}$$

$$min_{i \leq |\mathrm{E}|}\big\{max_{j \leq |\mathrm{E}|}\{d(m_i, m_j)\}\big\} \tag{4}$$

[11] A negative example can belong to more than one cluster.

$$max_{i \leq |\mathrm{E}|} \left\{ \frac{\sum_{j \leq |\mathrm{E}|} d(m_i, m_j)}{|\mathrm{E}| - 1} \right\} \quad (5)$$

$$\frac{\sum_{i \leq |\mathrm{E}|} \frac{\sum_{j < |\mathrm{E}|} d(m_i, m_j)}{|\mathrm{E}|-1}}{|\mathrm{E}|} \quad (6)$$

Taking into account these hypotheses, two approaches have been studied.

*First approach.* $\hat{\mathcal{E}}^-$ is generated as the union of sets $E_i^-(\alpha)$. Two different corpus sizes have been used here (5 and 25 documents). Table 1 shows that, for the 5-documents corpus and for all six $\alpha$ formulas, the recall and the F values outperform those using Freitag's method[12]. However, the second formula generates a bigger set $\hat{\mathcal{E}}^-$ (6088) than Freitag's method (2790) does. Moreover, the sizes of $\hat{\mathcal{E}}^-$ generated by the rest of the formulas when using 25 documents, always exceed the Freitag's method (13116). As a conclusion, this alternative cannot be used because extremely large sets $\hat{\mathcal{E}}^-$ are generated.

**Table 1.** Results from applying $\alpha$ formulas.

| Formula | 5 docs | | | | | 25 docs | |
|---|---|---|---|---|---|---|---|
| | $\alpha$ | $\lvert\hat{\mathcal{E}}^-\rvert$ | R | P | F | $\alpha$ | $\lvert\hat{\mathcal{E}}^-\rvert$ |
| 1 & 6 | 0.759 | 658 | 50.00 | 98.08 | 66.23 | 0.679 | **37867** |
| 2 | - | **6088** | 58.82 | 96.77 | 76.17 | | |
| 3 | - | 1259 | 78.43 | 93.02 | 85.10 | - | **38536** |
| 4 | 1.362 | 2465 | 55.88 | 91.93 | 69.51 | 1.352 | **66749** |
| 5 | 1.415 | 2534 | 55.88 | 91.93 | 69.51 | 1.404 | **67049** |
| Freitag's baseline | | 2790 | 43.14 | 97.78 | 59.87 | | 13116 |

*Second approach.* For each cluster, a number of negatives $N_i$ is computed according to the dimension of $\mathcal{E}_i^+$, as follows:

$$N_i = \beta \, |\mathcal{E}_i^+| \qquad \beta = \frac{\sum_{i \leq |\mathrm{E}|} |E_i^-(\alpha)|}{|\mathrm{E}|} \quad (7)$$

Then, $\hat{\mathcal{E}}^-$ is generated as the union of the $N_i$ negative examples within each $\mathcal{E}_i^-$ being closest to the associated $m_i$. We have tested formula 7 combined with formulas 1, 4, 5 and 6 as $\alpha$ distance values[13] and using 5 documents. The results of our experiments are shown in table 2. We can see that using values from

[12] The - mark means that every cluster has a different $\alpha$ value, so they cannot be shown in practice.

[13] Formulas 2 and 3 cannot be used within $\beta$ formula because they depend on the clusters.

formulas 4 and 5 as $\alpha$, the results outperform those using Freitag's method taking only about half the number of negative examples. The use of formula 4 seems to generate a more restricted set $\hat{\mathcal{E}}^-$ without loosing accuracy. As a conclusion, we have adopted formula 7 combined with formula 4 as $\alpha$ value, in order to select set $\hat{\mathcal{E}}^-$.

**Table 2.** Results from applying different $\alpha$ values to $\beta$.

| Formula | $\alpha$ | $\beta$ | $\lvert\hat{\mathcal{E}}^-\rvert$ | R | P | F |
|---|---|---|---|---|---|---|
| 1 & 6 | 0.759 | 21 | 380 | 29.41 | 100 | 45.45 |
| 4 | 1.362 | 79 | **1000** | **43.14** | **100** | **60.28** |
| 5 | 1.415 | 81 | 1021 | 43.14 | 100 | 60.28 |
| Freitag's baseline | | | 2790 | 43.14 | 97.78 | 59.87 |

## 5 Final Experiments

We have applied the method for selecting $\hat{\mathcal{E}}^-$ (explained before) to the 5 different corpus sizes. The results are presented in table 3. Comparing them to those in table 3, a much smaller set $\hat{\mathcal{E}}^-$ is generated by using our method than using Freitag's one. Moreover, the resulting F values show that better extractors are learned when using small sizes of traning corpus. However, for bigger sizes (45 documents), F values resulting from applying the Freitag's method tend to be slightly better (0.15 points over), but a much bigger set of examples is taken. Finally, comparing the number of rules, the extractors learned by using our approach are slightly more compact than those learned by using the baseline method.

**Table 3.** Test results from both Freitag's baseline and 7.4 method.

| Corpus Size | Freitag's baseline | | | | | 7.4 method | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\lvert\hat{\mathcal{E}}^-\rvert$ | R | P | F | Rules | $\lvert\hat{\mathcal{E}}^-\rvert$ | R | P | F | Rules |
| 5 | 2790 | 43.14 | 97.78 | 59.87 | 11 | 1021 | 43.14 | 100 | 60.28 | 9 |
| 15 | 7553 | 56.86 | 100 | 72.50 | 15 | 2784 | 59.80 | 100 | 74.84 | 14 |
| 25 | 13116 | 62.74 | 98.45 | 76.64 | 33 | 7621 | 72.55 | 97.37 | 83.15 | 30 |
| 35 | 18036 | 73.53 | 97.40 | 83.80 | 37 | 10640 | 73.53 | 97.40 | 83.80 | 35 |
| 45 | 29523 | 75.49 | 98.72 | 85.56 | 39 | 17479 | 74.51 | 100 | 85.39 | 37 |

## 6 Conclusions

This paper describes some of the remaining drawbacks of the existing IE-rule learning systems. EVIUS has been briefly described as a possible methodology to deal with the *combination problem*, *training set size problem* and the *training set relevance problem*.

We have presented a new method to deal with the later one. This method consists in selecting a calculated number of negative examples, being these the closest to the positive ones. Some experiments have been done in order to compare the new method to a baseline one. These comparisons prove that our method generates a much smaller set of relevant examples that is good enough to learn good extractors.

## References


1. D. Freitag: Machine Learning for Information Extraction in Informal Domains. Ph.D. thesis, Computer Science Department. Carnegie Mellon University. (1998)
2. M.E. Califf and R. Mooney.: Relational learning of pattern-match rules for information extraction. In Workshop on Natural Language Learning, ACL (1997) 9–15.
3. S. Soderland, D. Fisher, J. Aseltine, and W. Lehnert.: Crystal: Inducing a conceptual dictionary. In proceedings of IJCAI (1995) 1314–1321.
4. S. Soderland.: Learning information extraction rules for semi-structured and free text. Machine Learning, 34, (1999) 233–272.
5. J. Turmo, N. Català, and H. Rodríguez.: An adaptable IE system to new domains. Applied Intelligence, 10(2/3) (1999) 225–246.
6. J. R. Quinlan.: Learning Logical Definitions from Relations. Machine Learning, 5(3) (1990) 239–266.
7. J. Turmo and H. Rodríguez.: Learning IE-rules for a Set of Related Concepts. In proceedings of CoNLL (2000) 115–118.
8. R.S. Michalski.: Towards a unified theory of learning: Multistrategy task-adaptive learning. In B.G. Buchanan and D. Wilkins, editors, Readings in Knowledge Acquisition and Learning. Morgan Kauffman. (1993)
9. H. Ko.: Empirical assembly sequence planning: A multistrategy constructive learning approach. In I. Bratko R. S. Michalsky and M. Kubat, editors, Machine Learning and Data Mining. John Wiley & Sons LTD. (1998)
10. D. Randall and T. Martínez.: Improved heterogeneous distance functions. Journal of Artificial Intelligence, 1. (1997)
11. Eneko Agirre and German Rigau.: A Proposal for Word Sense Disambiguation using Conceptual Distance. In Proceedings of RANLP, Tzigov Chark, Bulgaria (1997)
12. B. Everitt.: Cluster analysis. Edward Arnold, cop, 3. (1993)
13. T. Calinski and J. Harabasz.: A dendrite method for cluster analysis. Communications in Statistics, 3, (1974) 1–27.