Using Ontologies in Semantic Data Mining with SEGS and g-SEGS

Nada Lavrač[1,2], Anže Vavpetič[1], Larisa Soldatova[3],
Igor Trajkovski[4], Petra Kralj Novak[1]

[1] Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia
[2] University of Nova Gorica, Nova Gorica, Slovenia
[3] Aberystwyth University, Wales, United Kingdom
[4] Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University, Skopje, Macedonia
{nada.lavrac,anze.vavpetic.petra.kralj.novak}@ijs.si

Abstract. With the expanding of the Semantic Web and the availability of numerous ontologies which provide domain background knowledge and semantic descriptors to the data, the amount of semantic data is rapidly growing. The data mining community is faced with a paradigm shift: instead of mining the abundance of empirical data supported by the background knowledge, the new challenge is to mine the abundance of knowledge encoded in domain ontologies, constrained by the heuristics computed from the empirical data collection. We address this challenge by an approach, named semantic data mining, where domain ontologies define the hypothesis search space, and the data is used as means of constraining and guiding the process of hypothesis search and evaluation. The use of prototype semantic data mining systems SEGS and g-SEGS is demonstrated in a simple semantic data mining scenario and in two real-life functional genomics scenarios of mining biological ontologies with the support of experimental microarray data.

Keywords. semantic data mining, ontologies, background knowledge, relational data mining

1 Introduction

The most common setting in knowledge discovery is rather simple: given is the empirical data and a data mining task to be solved. The data is first preprocessed, then a data mining algorithm is applied and the ending result is a predictive model or a set of descriptive patterns which can be visualized and interpreted. Data mining algorithms included in the contemporary data mining platforms (e.g., WEKA [20]) provide extensive support for mining empirical data stored in a single table format, usually referred to as propositional data mining.

Data by itself does not carry semantic meaning but needs to be interpreted to convey information. Standard data mining algorithms do not 'understand' the data: data are treated as meaningless numbers (or attribute values) and statistics

are calculated on them to build patterns and models, while the interpretation of the results is left to human experts. An example of an everyday data mining challenge is to use the reference to time when the data was collected. Unless time is the main focus of investigation, as is the case in time series analysis, a data mining algorithm will treat time just like any other attribute. However, as standard data mining algorithms do not have specialized mechanisms to deal with time, it is the role of the domain expert to adequately preprocess the time entry.

It is well known that the performance of data mining methods can be significantly improved if additional relations among the data objects are taken into account: the knowledge discovery process can significantly benefit from the domain (background) knowledge, as successfully exploited in relational data mining and Inductive Logic Programming (ILP) [5]. Additional means of providing more information to the learner is by providing semantic descriptors to the data. Moreover, as discussed in this paper, important additional knowledge to semantic descriptors are also the relations in the underlying ontologies which can be used as an important additional information source for data mining.

Usually, there is abundant empirical data, while the background knowledge is scarce. However, with the expanding of the Semantic Web and the availability of numerous ontologies which provide domain background knowledge and semantic descriptors to the data, the amount of *semantic data* (data which include semantic information, e.g., ontologies and annotated data collections) is rapidly growing. The data mining community is now faced with a paradigm shift: instead of mining the abundance of empirical data supported by the background knowledge, the new challenge is to mine the abundance of knowledge encoded in domain ontologies, constrained by the heuristics computed from the empirical data collection. This paper uses the term *semantic data mining* to denote this new data mining challenge and approaches in which semantic data are mined.

We present g-SEGS, a prototype semantic data mining system implemented in the novel service-oriented data mining environment Orange4WS [16] which supports knowledge discovery workflow construction from distributed data mining services. System g-SEGS is a successor of SEGS, a system for Searching of Enriched Gene Sets [19] designed specifically for functional genomics tasks. While SEGS is a special purpose system for analyzing microarray data with biological ontologies as background knowledge, g-SEGS is a general purpose semantic data mining system. It takes as input (1) domain ontologies in the OWL format, used to construct a version space of hypotheses (patterns) to be mined, and (2) an empirical data collection, annotated by domain ontology terms, used to constrain and guide the top-down search of hierarchically structured space of hypotheses, as well as for hypotheses quality evaluation. The utility of systems g-SEGS and SEGS is demonstrated in three scenarios: a simple hand-crafted scenario, and two functional genomics use cases. In addition to OWL encoded ontologies, used as input to our system g-SEGS, we also use other formats of annotated hierarchically structured data sources, such as the ENTREZ and KEGG hierarchies used in the SEGS real-life functional genomics use case.

The paper is organized as follows. We provide the motivation for this research in Section 2. Section 3 presents the related work. Sections 4 introduces the semantic data mining task and presents the proposed semantic data mining methodology, together with the g-SEGS algorithm implementation. Section 5 presents an illustrative example of using g-SEGS, followed by the presentation of selected results of using SEGS in real-life functional genomics use cases in Section 6. In Section 7, we conclude and give some directions for further work.

2 Motivation

Modern scientifc research is becoming more interdisciplinary, interactive, distributed, knowledge intensive, and data-driven. Semantic Web technologies, such as RDF (Resource Description Framework) and OWL (Web Ontology Language), are becoming popular as technological solutions to many of these challenges to science. The Semantic Web is changing the way how scientific data are collected, deposited, and analysed. Semantic descriptors for data (informational assets) are required to enable automated processing and support of knowledge retrieval, sharing, reuse and discovery.

Ontologies provide logically consistent knowledge models which formally define the semantic descriptors. The RDF data model (triplets *subject-predicate-object*) is simple, yet powerful. Such a representation ensures the flexibility of changing the data structures, and the integration of heterogeneous data sources. Data can be directly represented in RDF as graph data or (semi-)automatically translated from propositional representations. Consequently, more and more data from public relational data bases are now being translated into RDF as *linked data*.[1] In this way, data items from various databases can be easily linked and queried over multiple data repositories through the use of semantic descriptors provided by the supporting ontologies—the domain models or background knowledge.

While contemporary data mining platforms (e.g., WEKA) focus on propositional data, the performance of data mining methods can be significantly improved by providing semantic descriptors to the data and if additional relations among data objects are taken into account, e.g., if the underlying ontologies are used as the main information source for mining.

Semantic data mining has a great potential utility in many applications where ontologies are used as semantic descriptors for the data. For example, in biomedicine, biology, sociology, finance, the number of available ontologies is rapidly growing[2]. In order to support the users, semantic data mining algorithms should be able to import an ontology (or a set of ontologies) in a standard ontology language and output results of data mining in a form which is semantically meaningful to the user. Our system g-SEGS is designed with this goal in mind.

[1] See the Linked Data site `http://linkeddata.org/`

[2] See `http://bioportal.bioontology.org/`

3 Related work

The idea of using hierarchies as background knowledge to generalize terms in inductive rule learning has been proposed already by Michalski [13]. More recent usage of ontologies in data mining includes [6, 2, 18, 3, 12] as well as domain specific systems which use ontologies as background knowledge for data mining [8, 19].

In [6], the use of taxonomies (where the leaves of the taxonomy correspond to attributes of the input data) on paleontological data is studied. The problem was to predict the age of a fossil site on the basis of the taxa that have been found in it – the challenge was to consider taxa at a suitable level of aggregation. Motivated by this application, they studied the problem of selecting an antichain from a taxonomy that improves the prediction accuracy. In [2], background knowledge is in the standard inheritance network notation and the KBRL[3] algorithm performs a general-to-specific heuristic search for a set of conjunctive rules that satisfy user-defined rule evaluation criteria. In [18], ontology-enhanced association mining is discussed and four stages of the (4ft-Miner-based) KDD process are identified that are likely to benefit from ontology application: data understanding, task design, result interpretation and result dissemination over the semantic web. The work of [3] first focuses on pre-processing steps of business and data understanding in order to build an ontology driven information system (ODIS), and then the knowledge base is used for the post-processing step of model interpretation. Liu et al. [12] propose a learning-based semantic search algorithm to suggest appropriate Semantic Web terms and ontologies for the given data.

An ontology driven approach to knowledge discovery in biomedicine is described in [8], where efforts to bridge knowledge discovery in biomedicine and ontology learning for successful data mining in large databases are presented. A domain specific system that uses ontologies and other hierarchies as background knowledge for data mining is SEGS [19]. The SEGS system finds groups of differentially expressed genes, called *enriched gene sets*[4]. Compared to earlier work [17, 9], the novelty of SEGS is that it does not only test existing gene sets (existing ontology terms) for differential expression but it generates also new gene set descriptions that represent novel biological hypotheses.

There has been a large amount of work developing machine learning and data mining methods for graph-based data [1]. However, these methods are not designed to fully exploit the rich logical descriptions of relations provided by the ontologies that support the description of graph-based data. Relational data mining, inductive logic programming (ILP) and statistical relational learning (SRL) methods [5] are more general but they assume the data will be described using Horn clauses or Datalog, rather than RDF and description logics. The most commonly used description logic format for Semantic Web is OWL-DL. OWL-DL

[3] KBRL is based on the RL learning program of [4]

[4] A gene set is enriched if the genes that are members of this gene set are statistically significantly differentially expressed compared to the rest of the genes.

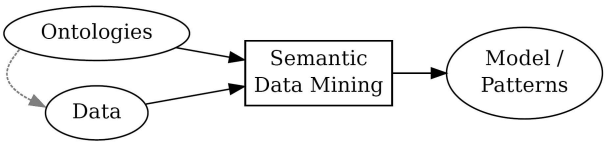

Fig. 1. Schema of a semantic data mining process, with ontologies and annotated data as inputs.

allows to define properties of relations which link entities defined in an ontology as transitive, symmetric, functional, and to assign cardinality to relations. Properties of relations form an important part of the background knowledge model, therefore modifications of existing relational algorithms or even new algorithms are required in order to effectively exploit this knowledge. Lehmann and Haase [11] make the first steps in this direction by defining a refinement operator in the $\mathcal{EL}$ Description Logic; opposed to our work they consider only the construction of consistent and complete hypotheses using an ideal refinement operator.

4 Semantic Data Mining with g-SEGS

This paper uses the term *semantic data mining* to denote a data mining task in which semantic data are mined. This section first introduces this task, followed by the methodology of semantic data mining as implemented in g-SEGS.

4.1 Semantic data mining

A *semantic data mining task*, illustrated in Figure 1, is defined as follows.

Given: a set of domain ontologies, and empirical data annotated by domain ontology terms[5],

Find: a hypothesis (a predictive model or a set of descriptive patterns) by mining the abundance of information in ontologies, constrained by the information in the empirical data collection.

Successfully handling the challenging task of semantic data mining may result in a paradigm shift in which the abundance of domain ontologies will be mined, and the empirical training data will be used mainly to constrain the hypothesis search space by the heuristics computed from the training data collection[6].

The methodology, implemented in the g-SEGS system, assumes that the hypothesis language are logical rules, where rule conditions are conjunctions of

[5] Annotations refer to annotations of instances and of attribute values.

[6] A similar challenge is faced in pattern mining research where the original problem of mining the abundance of data was recently transformed into a problem of mining the abundance of induced patterns, constrained by the heuristics computed from the training data.

ontology terms. While statistical significance of rules could be measured on the fly in the process of rule construction, we have decided to construct all the rules satisfying the support constraint, and to eliminate insignificant rules in rule postprocessing, using a heuristic known from subgroup discovery. As shown in Section 5, semantic data mining results in more general and semantically more meaningful rules, if compared to standard rule learning.

Motivated by the successful applications of SEGS [19, 14], we have decided to generalize SEGS to become domain independent, and developed a new system named g-SEGS (generalized SEGS). From the four main components of SEGS, only the SEGS hypothesis language and the generation and pruning procedure are used unchanged in the new semantic data mining system g-SEGS.

The proposed *semantic data mining methodology*, implemented in g-SEGS, is described below in terms of its four main components: the hypothesis language, the input (domain ontologies and training data), the hypothesis generation procedure and the hypothesis (pattern) evaluation and filtering procedure.

4.2 Hypothesis language

The hypothesis language are descriptive patterns in the form of rules $Class \leftarrow Conditions$, where $Conditions$ is a logical conjunction of ontology terms. For example, a rule whose antecedent is a conjunction of three terms, has the form $Class \leftarrow X \wedge Y \wedge Z$, where X stands for all $x \in X$, Y stands for all $y \in Y$, and Z stands for all $z \in Z$, and where e.g., $X \in Ont1, Y \in Ont2$, and $Z \in Ont3$.

4.3 Input

g-SEGS requires two types of inputs: the ontological background knowledge and the training data.

Background knowledge consists of domain ontologies, typically in the OWL format.[7] Ontologies are used to construct the hypothesis search space.

Training data are class-labeled vectors of attribute values, annotated by the terms in domain ontologies. The data are used to constrain the hypothesis search, and for rule quality evaluation in rule postprocessing.

4.4 Rule construction

Rule construction results in a set of rules satisfying the minimal support criterion. As a rule antecedent is a conjunction of ontology terms, all possible conjunctions of ontology terms can be generated and evaluated for small ontologies. In case of large ontologies, however, the search space needs to be pruned. To do so, we use the subsumption property of a relation which forms the hierarchical

[7] In addition to OWL ontologies, we allow for other formats of annotated hierarchically structured data sources, such as the ENTREZ and KEGG hierarchies, which were used in one of the two real-life functional genomics use cases in Section 6.

backbone of the ontology (e.g. `is-a`). Suppose that rule $C \leftarrow X' \wedge Y' \wedge Z'$ has been constructed by the specialization of rule $C \leftarrow X \wedge Y \wedge Z$, where $X' \preceq X, Y' \preceq Y, Z' \preceq Z$ ($\preceq$ denotes *more or equally specific* relation). If rule $C \leftarrow X' \wedge Y' \wedge Z'$ covers m objects where $m < N$ (m is lower than the support threshold N which determines the minimal number of objects to be covered by each rule), it is pruned and none of its specialization will be constructed. This results in a significant reduction of the hypothesis search space.

In a simplified case, where three ontologies $Ont1$, $Ont2$ and $Ont3$ are given, hypothesis generation consists of creating the conjunctions of individual ontology terms, one from each ontology. Hypothesis construction is performed in a top-down manner, starting from the most general terms in each of the three ontologies, and specializing the rule antecedent as long as the stopping criterion is satisfied (ensuring sufficient coverage of data instances)[8]. If one conjunct does not satisfy the constraint, then its descendents will also not satisfy it, because they cover a subset of instances covered by the conjunction. Therefore, we first construct conjuncts from the top nodes of $Ont1$, $Ont2$ and $Ont3$, and if the conjunction fails to satisfy the given constraint, g-SEGS will not refine the last added term. Note that the efficiency of the algorithm comes from the usage of the hierarchical structure of ontologies.

In addition to `is-a` or `instance-of` subsumption relations there may be other links (relations) among ontology terms, e.g, the `interacts` relation. Consider a simple rule `class(A)` ← `is-a(A,B)`, and suppose that ontology term `B` is linked with term `C` through `interacts(B,C)`. In this case, the rule's antecedent can be refined to form a conjunction `is-a(A,B)` ∧ `interacts(B,C)`. This illustrates a situation which is common to ILP, as one can also make statements about `B` or `C`, not only about term `A` which appears in the rule head `class(A)`. For this reason, as well as due to applying heuristic rule filtering (see the next section), a simple top-down refinement approach to rule construction (e.g. as proposed by Lehmann and Haase [11]) is insufficient.

4.5 Rule filtering and evaluation

As the number of generated rules can be large, uninteresting and overlapping rules have to be filtered. Rule filtering in g-SEGS is done with *wWRAcc* (Weighted Relative Accuray heuristic with example weigths) heuristic [10], using example weights as means for considering different parts of the example space when selecting the best rules. In the *wWRAcc* heuristic defined below, N' denotes the sum of weights of all examples, $n'(C)$ is the sum of weigths of examples of concept C, $n'(Cnd)$ is the sum of weights of all covered examples, and $n'(Cnd \wedge C)$ is the sum of weights of all correctly covered examples of concept C.

$$wWRAcc(C \leftarrow Cnd) = \frac{n'(Cnd)}{N'} \cdot \left(\frac{n'(Cnd \wedge C)}{n'(Cnd)} - \frac{n'(C)}{N'} \right)$$

[8] If the ontology is simply a hierarchy (a tree), with the root of the graph being the most general term, this means that substantial pruning of the search space can be achieved in rule construction.

Rule filtering, using the weighted covering approach, proceeds as follows. It starts with a set of generated rules, a set of examples with weights equal to 1 and parameter k, which denotes how many times an example can be covered before being removed form the example set. In each iteration, we select the rule with the highest *wWRAcc* value, add it to the final rule set, and remove it from the set of generated rules. Then counter m is increased to $m + 1$ and weigths of examples covered by this rule decreased to $\frac{1}{m+1}$, where example weight $\frac{1}{m}$ means that the example has already been covered by $m < k$ rules. These steps are repeated until the algorithm runs out of examples or rules or if no rule has a score above 0. Once the learning process is finished and the rules have been generated and filtered, they are evaluated and sorted using the Fisher's exact test or the original *WRAcc* (Weighted Relative Accuray) measure known from CN2-SD subgroup discovery, which trades-off the generality of a rule and its precision. The *WRAcc* heuristic is defined as

$$WRAcc(C \leftarrow Cnd) = \frac{n(Cnd)}{N} \cdot \left(\frac{n(Cnd \wedge C)}{n(Cnd)} - \frac{n(C)}{N} \right)$$

where N is the number of all examples, $n(C)$ is the number of examples of concept C, $n(Cnd)$ is the number of all covered examples, and $n(Cnd \wedge C)$ is the number of all correctly covered examples of concept C.

4.6 g-SEGS implementation

The g-SEGS system takes as input the ontologies in the OWL format and data in the Orange [15] format, uses the hierarchical structure of the `is-a` relation of ontologies for efficient search and pruning of the rule search space, generates rules by forming conjunctions of terms from different ontologies, and uses the *wWRAcc* (Weighted Relative Accuray heuristic with example weigths) for rule pruning by iteratively selecting the rules and Fischer exact test or *WRAcc* (Weighted Relative Accuray) to sort/rank the selected rules.

g-SEGS is implemented in the Orange4WS [16] environment which upgrades the freely available Orange [15] data mining environment with several additional features: simple creation of new visual programming units (widgets) from distributed web services, composition of workflows from both local and distributed data processing/mining algorithms and data sources, and implementation of a toolkit for creating new web services. By using these tools, we were able to give g-SEGS a user-friendly interface and the ability to be executed remotely as a web service. By mapping the g-SEGS input to the SEGS input we were able to fully reuse the already implemented SEGS system. We defined the g-SEGS web service using WSDL (Web Service Definition Language). Using the web service definition and the set of tools provided by Orange4WS, we created a web service for our system. Finally, also using Orange4WS, we imported the web service into the Orange visual programming environment, thus allowing g-SEGS to be used in various workflows together with other Orange widgets.

A screenshot of an Orange4WS workflow with g-SEGS is shown in Figure 2. The workflow is composed of one widget for loading the dataset (`File`), three

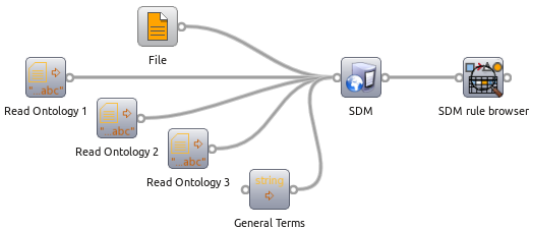

Fig. 2. An Orange4WS workflow with g-SEGS.

widgets for loading the three ontologies (`Read Ontology`), and one widget for specifying top-level ontology terms that are too general to appear in the final rules (`General terms`). These five widgets act as the input to the g-SEGS widget, which generates rules, displayed in the g-SEGS `Rule set browser` widget.

5 An illustrative example

As a proof-of-concept semantic data mining example, consider a bank which has the following data about its customers: place of living, employment, bank services used, which includes the account type, possible credits and insurance policies and so on. The bank also categorized the clients as 'big spenders' or not and wants to find patterns describing big spenders. Table 1 presents the training data.

The application of standard classification rule learning algorithm CN2 (we chose the Orange [15] implementation of CN2) to these data generates the rules presented in the top part of Table 2, and the middle part of this table presents the results obtained by using the CN2-SD subgroup discovery algorithm [10].

While CN2 generates a set of dependent and very specific classification rules, CN2-SD produces rules representing individual subgroup descriptions which are better suited for the comparison with the results obtained with g-SEGS. Note that both sets of rules are rather specific, due to the specificity of the attribute-value data representation. Standard data mining does not provide automated means for rule generalization; if more general rules were desired, the data should have been manually preprocessed and attribute-values generalized to obtain more general rules and therefore more valuable results.

In semantic data mining using g-SEGS, in addition to the data in Table 1, three ontologies shown in Figure 3 are used as input to introduce semantics into the discovery process. The result of applying g-SEGS to these ontologies and the given training data is presented in the bottom part of Table 2.[9]

The result illustrates the following characteristics of semantic data mining by g-SEGS: (a) Conditions of g-SEGS rules are conjunctions of literals, having ontology terms as arguments of predicates bearing the ontology name (and therefore logically defined semantic meaning), while the conditions of CN2 and

[9] The same data and background knowledge could also be used for describing credit holders or clients that have closed their account in a bank.

Table 1. Table of bank customers described by different attributes and class 'big spender'.

id	occupation	location	account	loan	deposit	inv_fund	insur.	big_spender
1	Doctor	Milan	Classic	No	No	TechShare	Family	YES
2	Doctor	Krakow	Gold	Car	ShortTerm	No	No	YES
3	Military	Munich	Gold	No	No	No	Regular	YES
4	Doctor	Catanzaro	Classic	Car	LongTerm	TechShare	Senior	YES
5	Energy	Poznan	Gold	Apart.	LongTerm	No	No	YES
...	...	...	...	...	...	...	...	...
25	Transport	Cosenza	Classic	Car	ShortTerm	No	Family	NO
26	Police	Tarnow	Gold	Apart.	No	No	No	NO
27	Nurse	Radom	Classic	No	No	No	Senior	NO
28	Education	Catanzaro	Classic	Apart.	No	No	No	NO
29	Transport	Warsaw	Gold	Car	ShortTerm	TechShare	Regular	NO
30	Police	Cosenza	Classic	Car	No	No	No	NO

CN2-SD rules are conjunctions of attribute-value pairs, (b) g-SEGS rules are more general compared to rules constructed by CN2, CN2-SD or other non-semantic data mining algorithms, and (c) once the ontologies and the workflows

Table 2. Rules generated by CN2, CN2-SD and g-SEGS from the data in Table 1. Coverage, confidence and WRAcc were computed in postprocessing.

CN2 rules for class big_spender='YES'	**Coverage**	**Confid.**	**WRAcc**
occupation='Doctor'	20.00%	83.33%	0.067
loan='No' ∧ account='Gold'	10.00%	100.00%	0.050
occupation='Health-care'	6.67%	100.00%	0.033
occupation='Education' ∧ account='Gold'	6.67%	100.00%	0.033

CN2-SD rules for class big_spender='YES'	**Coverage**	**Confid.**	**WRAcc**
account='Gold' ∧ investment_fund='No'	33.33%	80.00%	0.100
account='Gold'	46.67%	64.29%	0.067
occupation='Doctor'	20.00%	83.33%	0.067
occupation='Health-care'	6.67%	100.00%	0.033
investment_fund='TechnologyShare' ∧ account='Classic'	13.33%	75.00%	0.033

g-SEGS rules for class big_spender='YES'	**Coverage**	**Confid.**	**WRAcc**
occupation(Public) ∧ bankingService(Gold)	26.67%	87.50%	0.100
bankingService(Gold)	46.67%	64.29%	0.067
occupation(Doctor)	20.00%	83.33%	0.067
occupation(Public) ∧ bankingService(Deposit)	26.67%	75.00%	0.067
occupation(Health)	23.33%	71.43%	0.050
occupation(Doctor) ∧ bankingService(Deposit)	16.67%	80.00%	0.050
location(Bavaria)	16.67%	80.00%	0.050
location(Germany) ∧ occupation(Service) ∧ bankingService(investmentFund)	16.67%	80.00%	0.050

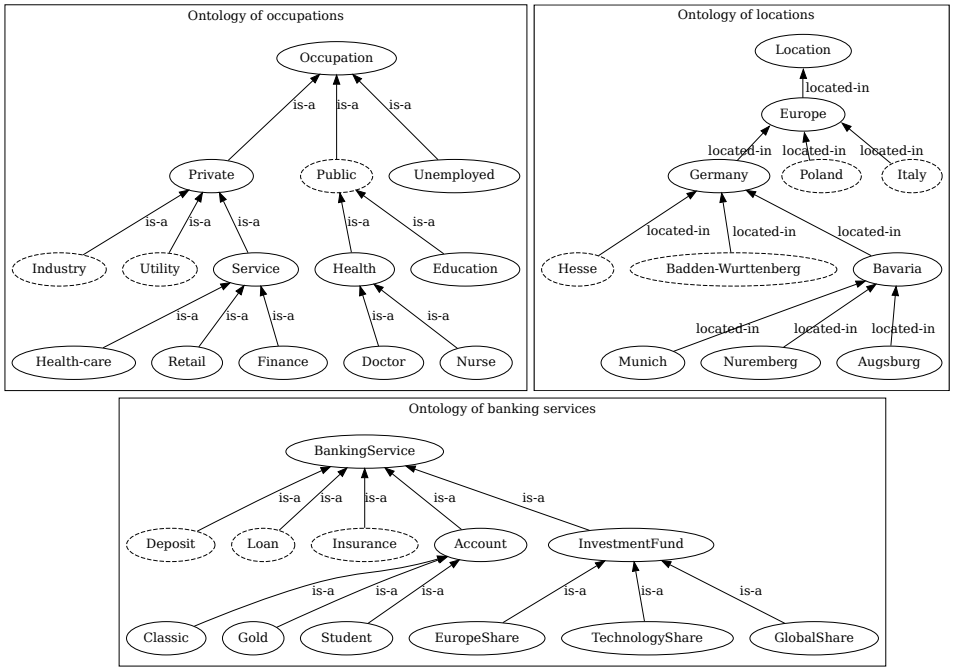

Fig. 3. Ontologies for data in Table 1. Note that these are not the full ontologies, but only the parts needed to interpret the rules presented in this paper. Concepts with omitted subconcepts are drawn with a dashed line.

have been constructed, automated and therefore repeatable data processing and rule construction can be performed, less prone to human processing errors.

6 Functional genomics use cases

This section presents how SEGS was used in two functional genomics use cases, illustrating (1) microarray data analysis by using the Gene Ontology (GO) as background knowledge, and (2) microarray data analysis using three semantic knowledge sources, i.e., GO, KEGG and Entrez, as background knowledge to SEGS.

We first present the results of analzying microarray data with the SEGS algorithm, using the Gene Ontology as background knowledge. The results were obtained in the data analysis task, aimed at distinguishing between samples of acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML), first introduced by Golub et al. [7]. The data contains 73 class-labeled samples of expression vectors, where gene expression profiles (obtained by the Affymetrix HU6800 microarray chip) contain probes for 6,817 genes.

Using GO as background knowledge, our system has generated several gene group describing rules. For space constraints we list a single rule, describing a

group of 18 genes, characterizing the ALL leukemia class.

```
genegroup(all,G) ←
   component(G,nucleus) ∧ {interaction(G,G1) ∧
   process(G1,regulation-of-transcription).
```

The interpretation of this rule is that acute lymphoblastic leukemia (ALL) is characterized by proteins (gene group `G`), which are the products of the genes which are located in the nucleus of the cell, and which interact with the genes (gene group `G1`) which are included in the process of regulation of transcription. Unlike other well known tools that use gene ontologies for analyzing gene expression data PAGE [9] and GSEA [17], which report statistically significant single GO terms and do not use gene interaction data, we are able to find a set of genes described by a conjunction of ontology terms as well as the available gene interaction data to describe features of genes which can not be represented by other approaches.

The second use case in functional genomics presents the results obtained with the SEGS algorithm when analzying microarray data from a different ALL dataset, i.e., a dataset from a clinical trial in acute lymphoblastic leukemia (ALL) (Chiaretti et al., 2004). The ALL dataset was again chosen as it is typical for medical research and has a reference role for such evaluations as it has been a model dataset for other microarray data analysis tools as well. The analysis of differences in gene expression between two lymphocyte subtypes (lymphocyte B and lymphocite T) was performed as follows. Genes were first ranked according to their expression value, and differentially expressed genes were selected by gene filtering according to $logFC$ cut-off value $|0.3|$.

Three semantic knowledge sources were used as background knowledge to SEGS: GO, KEGG and Entrez. As, except for GO, these hierarchies are not available in the OWL format, a dedicated algorithm for merging these three sources was used to form the joint input database format, which can be chosen as a parameter in g-SEGS, in addition to the default OWL format. Space constrains disable us from presenting the set of discovered rules, describing subgroups of differentially expressed genes, formed as conjuctions of terms, e.g., `receptor-binding(G)` ∧ `T-cell-activation(G)` as well as basic information about the rules. Similar to previous research, the results show that one of the main differences between differentially expressed and non-differentially expressed gene groups is the expression of major histocompatibility complex (HLA) related genes.

7 Conclusions

This paper discusses *semantic data mining* as an adequate approach to face a potential paradigm shift in data mining, addressing the new challenge of mining the knowledge in ontologies, constrained by the empirical evidence in the collected data. In our approach, domain ontologies define the hypothesis search

space, and the data is used as means of guiding and constraining the hypothesis search and evaluation.

A prototype semantic data mining system g-SEGS is used to illustrate the approach in a simple semantic data mining scenario, whereas its predecessor SEGS is used to illustrate semantic data mining in two real-life functional genomics scenarios. The g-SEGS system takes ontologies in OWL format and data in a standard attribute-value format as its input, and takes advantage of the hierarchical relationships in ontologies for efficient search and pruning of the hypothesis search sapce. The user friendlly user interface is also one of the key features of the g-SEGS system.

There are many possible fields of application of semantic data mining. It can be directly applied to domains where data are characterized by sparsity and taxonomies are available, like market basket analysis, to give an example. We have demonstrated the usefulness of semantic data mining in two real-life functional genomics scenarios where biological ontologies are mined with the support of experimental microarray data. The prototype semantic data mining system g-SEGS shows major advantages compared to non-semantic systems, as more general rules and automated data preprocessing are performed. There are also advantages compared to ILP and other related approaches since our system uses a standardized encoding of knowledge.

A systematic comparison of g-SEGS to the state of the art relational data mining systems is planned in our further work. The first results of comparing g-SEGS to the state of the art ILP system Aleph indicate that g-SEGS is significantly more efficient, and that using the ontologies in their native format substantially simplifies the system's use in real life scenarios, by reducing the encoding time and ensuring the system's reusability.

Acknowledgments

The research presented in this paper was supported by the Slovenian Ministry of Higher Education, Science and Technology (grant no. P-103) and the EU-FP7 projects e-LICO and BISON.

References

[1] C.C. Aggarwal and H. Wang, editors. Managing and Mining Graph Data. Springer US, 2010.

[2] J.M. Aronis, F.J. Provost, and B.G. Buchanan. Exploiting background knowledge in automated discovery. In *Proc. of the 2nd International Conference on Knowledge Discovery and Data Mining*, pages 355–358, 1996.

[3] L. Brisson and M. Collard. How to semantically enhance a data mining process? In *Proc. of the 10th International Conference on Enterprise Information Systems, ICEIS 2008*, pages 103–116, 2008.

[4] S.H. Clearwater and F.J. Provost. Rl4: A tool for knowledge-based induction. In *Proc. of the 2nd International IEEE Conference on Tools for Artificial Intelligence*, pages 24–30, November 1990.

[5] L. De Raedt. *Logical and Relational Learning*. Springer Berlin Heidelberg, 2008.

[6] G.C. Garriga, A. Ukkonen, and H. Mannila. Feature selection in taxonomies with applications to paleontology. In *Proc. of the 11th International Conference on Discovery Science*, DS '08, pages 112–123. Springer-Verlag, 2008.

[7] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoringt. *Science*, 286:531–537, 1999.

[8] P. Gottgtroy, N. Kasabov, and S. MacDonell. An ontology driven approach for knowledge discovery in biomedicine. In *Proc. of the VIII Pacific Rim International Conferences on Artificial Intelligence (PRICAI)*, 2004.

[9] S.Y. Kim and D.J. Volsky. Page: Parametric analysis of gene set enrichment. *BMC Bioinformatics*, 6(144), 2005.

[10] N. Lavrač, B. Kavšek, P.A. Flach, and L. Todorovski. Subgroup discovery with CN2-SD. *Journal of Machine Learning Research*, 5:153–188, 2004.

[11] J. Lehmann and C. Haase. Ideal Downward Refinement in the $\mathcal{EL}$ Description Logic. In *Proceedings of the 19th International Conference on Inductive Logic Programming* (ILP 2009), LNAI 5989, 73-87, Springer, 2010.

[12] H. Liu. Towards semantic data mining. In *Proc. of the 9th International Semantic Web Conference (ISWC2010)*, November 2010.

[13] R.S. Michalski. A theory and methodology of inductive learning. In R.S. Michalski, J.G. Carbonell, and T.M. Mitchell, editors, *Machine Learning: An artificial intelligence approach*, pages 83–129. Palo Alto: Tioga Publishing Company, 1983.

[14] I. Mozetič, N. Lavrač, V. Podpečan, P. Kralj Novak, et al. Bisociative knowledge discovery for microarray data analysis. in *Proc. of the First Intl. Conf. on Computational Creativity*, 190–199, Springer, 2010.

[15] J. Demšar, B. Zupan, and G. Leban. Orange: From experimental machine learning to interactive data mining, white paper (www.ailab.si/orange). Faculty of Computer and Information Science, University of Ljubljana, 2004.

[16] V. Podpečcan, M. Juršič, M. Žakova, and N. Lavrač. Towards a service-oriented knowledge discovery platform. In *Proc. of the ECML/PKDD Workshop on Third-generation data mining: Towards service-oriented knowledge discovery*, 25-36, 2009.

[17] P. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, and M.A. Gillette. Gene set enrichment analysis: A knowledge based approach for interpreting genome-wide expression profiles. *Proc. of the National Academy of Science, USA*, 102(43):15545–15550, 2005.

[18] V. Svátek, J. Rauch, and M. Ralbovský. Ontology-enhanced association mining. In *Semantics, Web and Mining, Joint International Workshops, EWMF 2005 and KDO 2005*, pages 163–179, 2005.

[19] I. Trajkovski, N. Lavrač, and J. Tolar. SEGS: Search for enriched gene sets in microarray data. *Journal of Biomedical Informatics*, 41(4):588–601, 2008.

[20] I.H. Witten and E. Frank, E. *Data Mining Practical Machine Learning Tools and Techniques* (2nd ed.), 2005. San Francisco: Elsevier.