# TOWARDS SEMANTIC DATA MINING WITH g-SEGS

*Petra Kralj Novak*[1], *Anže Vavpetič*[1], *Igor Trajkovski*[2], *Nada Lavrač*[1,3]
[1] Department of Knowledge Technologies, Jožef Stefan Institute, Jamova 39, Ljubljana, Slovenia
[2] Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University, 1000 Skopje, Macedonia
[3] University of Nova Gorica, Vipavska 13, 5000 Nova Gorica, Slovenia
e-mail: Petra.Kralj.Novak@ijs.si

## ABSTRACT

**This paper introduces the term semantic data mining to denote a data mining approach where domain ontologies are used as background knowledge for data mining. It is motivated by successful applications of SEGS (search for enriched gene sets), a system that uses biological ontologies as background knowledge to construct descriptions of interesting gene sets in experimental microarray data. We generalized this domain-specific system to perform subgroup discovery on arbitrary data, annotated by ontologies. We present a prototype of the new semantic data mining system named g-SEGS, implemented in the Orange4WS environment, and an illustrative example showing the application potential of semantic data mining.**

## 1 INTRODUCTION

The most common setting in knowledge discovery is that we are given some data and a data mining task. The data is first manually preprocessed, then a data mining algorithm is applied and the ending result is a model or a set of patterns that can be further interpreted and visualized. It is generally recognized that the quality of the end model depends crucially on the quality of the data collection and preparation process.

Data by itself does not carry any meaning; it needs to be interpreted to convey information. Standard data mining algorithms do not 'understand' the data: the data is treated as meaningless numbers and statistics are calculated on them to build models, and the interpretation of the results is left to human domain experts. An example of an everyday data mining challenge is using the reference to time when the data was collected. Unless time is the main interest of investigation (as is the case in time series analysis), time should be treated just like one of the attributes. 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.

This paper introduces the term *semantic data mining* to denote a data mining approach where domain ontologies are used as background knowledge for data mining (schematically presented in Figure 1). Our research is motivated by the SEGS [12] system which successfully uses biological ontologies as semantically annotated background knowledge to find descriptions of differentially expressed gene sets. We realized that some features of SEGS could be useful not only in functional genomics but also in other domains, and decided to generalize SEGS to become domain independent.


Figure 1: The proposed semantic data mining methodology schema.

We present a prototype semantic data mining system g-SEGS, a generalization of the SEGS system. g-SEGS uses as input: (1) data annotated by ontologies and (2) ontologies in the OWL format. The latter are used for efficient search and pruning of the pattern search space to generate patterns in the form of conjunctions of ontology terms, and uses the Fischer exact test and permutation testing to statistically validate the discovered patterns. As such, g-SEGS is a successful proof-of-concept semantic data mining system.

This paper is organized as follows: Section 2 presents the related work. Section 3 presents the new g-SEGS system and Section 4 provides an illustrative example. In Section 5, we conclude and give some directions for further work.

## 2 RELATED WORK

The idea of using hierarchies as background knowledge to generalize terms in knowledge discovery has been proposed already in early machine learning by Michalski [7]. More recent usage of ontologies in knowledge discovery includes [1, 11, 2] as well as domain specific systems that use ontologies as background knowledge for knowledge discovery [4, 12].

In [1], background knowledge is in the form of standard inheritance network notation and the algorithm KBRL—based on the RL learning program (Clearwater & Provost, 1990)—performs a general-to-specific heuristic search for a set of conjunctive rules that satisfy user-defined rule evaluation criteria. In [11], 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 [2] 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.

An ontology driven approach for knowledge discovery in biomedicine is described in [4], 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 as background knowledge for data mining is SEGS [12]. The SEGS system finds groups of genes—the so-called gene sets—that are enriched. A gene set is enriched if the genes that are members of that gene set are statistically significantly differentially expressed compared to the rest of the genes. Compared to earlier work [10, 5], the novelty of the SEGS method proposed by Trajkovski et al. (2008) [12] is that it does not only test existing gene sets for differential expression but it also generates new gene sets that represent novel biological hypotheses. The SEGS method has four main components: the background knowledge, the hypothesis language, the hypothesis generation procedure and the hypothesis evaluation procedure.

## 3 GENERALIZED SEGS: G-SEGS

Motivated by successful applications of SEGS [6, 8], we decided to generalize it to become domain independent and named it g-SEGS. From the four main components of SEGS, only the SEGS hypothesis language and the generation and pruning procedure are general enough to be used unchanged in the new semantic data mining system g-SEGS. System g-SEGS inputs ontologies in the OWL format and data in the Orange [3] format, uses the hierarchical structure of the of the 'is-a' relation in ontologies for efficient search and pruning of the pattern search space, generates patterns in the form of conjunctions of terms from different ontologies, and uses the Fischer exact test and permutation testing to statistically validate the discovered patterns.

Interesting subgroups are constructed by conjunction of terms from the ontologies. All possible descriptions (by making all possible conjunctions) could be generated and evaluated for small ontologies. In case of very large ontologies, however, we need to prune the search space. In this case, we use the hierarchical property of the is-a relations of the ontologies. For example, if we constructed a subgroup with the following description $X \wedge Y \wedge Z$ ($X \in Ont1, Y \in Ont2, Z \in Ont3$), which covers $m$ objects from class A, assuming a threshold of $N$ ($N > m$) as the minimum number of objects that must be covered with the description, then we do not need to construct (and evaluate) all the intersections $x \wedge y \wedge z$, where $x \preceq X, y \preceq Y, z \preceq Z$ ($\preceq$ denotes more specific). This significantly reduces the search space of feasible descriptions.

g-SEGS is implemented in the Orange4WS data mining platform [9], which upgrades the freely available Orange data mining framework with several additional features: simple creation of new 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. We defined the g-SEGS web service using WSDL (Web Service Definition Language). Using the created 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 environment, thus allowing g-SEGS to be used in various workflows together with other Orange widgets.

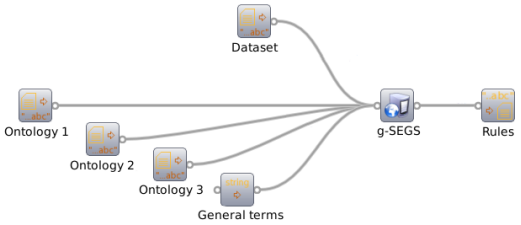


Figure 2: An Orange4WS workflow with g-SEGS.

A screenshot of an Orange4WS workflow with g-SEGS is presented in Figure 2. The workflow is composed of one widget for loading the data (Dataset), three widgets for loading the three ontologies, and one widget for specifying top-level ontology terms that are too general to appear in the final rules. These five widgets act as the input to the g-SEGS widget, which generates rules, displayed in the Rules widget.

g-SEGS inherited some limitations of SEGS, which include the limitation to four input ontologies, using a hierarchical structure (directed acyclic graphs only), which in practice means 'is-a' relations only, and cannot use attributes that are not annotated by ontologies.

## 4 AN ILLUSTRATIVE EXAMPLE

As a proof-of-concept of semantic data mining, we present the following 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 annotated the clients as 'big spenders' or not and wants to find patterns describing big spenders. Table 1 presents the example data.

An application of a 'standard' data mining algorithm (we chose the Orange [3] implementation of CN2) to these data produces the result presented in Table 2. These rules are very specific, due to the specificity of the attribute-values the data is described by. In classical data mining, such data should be manually preprocessed and attribute-values generalized to obtain more general rules and therefore more valuable results.

In addition to the data of Table 1, we propose to use three ontologies (depicted in Figure 3) to bring semantics into the knowledge discovery process. The result of applying g-SEGS to the data from Table 1 and ontologies from Figure 3 is presented in Table 3.

Table 1: A table of bank customers described by different attributes and a class 'big spender'.

| id | occupation | location | account | loan | deposit | investment_fund | insurance | big_spender |
|---|---|---|---|---|---|---|---|---|
| 1 | Doctor | Milan | Classic | No | No | TechnologyShare | 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 | TechnologyShare | Senior | YES |
| 5 | Energy | Poznan | Gold | No | No | No | No | YES |
| 6 | Doctor | Rome | Gold | Apartment | No | No | Regular | YES |
| 7 | Finance | Bavaria | Gold | No | ShortTerm | GlobalShare | No | YES |
| 8 | Health-care | Frankfurt | Classic | Car | No | GlobalShare | Family | YES |
| 9 | Military | Warsaw | Gold | No | ShortTerm | No | Regular | YES |
| 10 | Education | Latina | Gold | Apartment | No | No | Family | YES |
| 11 | Health-care | Karlsruhe | Classic | Apartment | No | EuropeShare | No | YES |
| 12 | Retail | Munich | Classic | Car | LongTerm | TechnologyShare | Regular | YES |
| 13 | Education | Catanzaro | Gold | Car | No | No | No | YES |
| 14 | Doctor | Milan | Classic | No | No | EuropeShare | No | YES |
| 15 | Police | Munich | Gold | Apartment | No | No | No | YES |
| 16 | Retail | Stuttgart | Classic | Car | LongTerm | TechnologyShare | No | NO |
| 17 | Finance | Brescia | Gold | Apartment | No | EuropeShare | Regular | NO |
| 18 | Administration | Tarnow | Classic | Car | No | No | Senior | NO |
| 19 | Materials | Freiburg | Gold | Apartment | ShortTerm | GlobalShare | No | NO |
| 20 | Doctor | Poznan | Classic | Personal | ShortTerm | EuropeShare | Regular | NO |
| 21 | Administration | Cosenza | Classic | Car | No | No | No | NO |
| 22 | Unemployed | Munich | Classic | Car | No | No | No | NO |
| 23 | Military | Kalisz | Classic | Apartment | ShortTerm | EuropeShare | Regular | NO |
| 24 | Manufacturing | Cosenza | Gold | Apartment | LongTerm | No | No | NO |
| 25 | Transportation | Cosenza | Classic | Car | ShortTerm | No | Family | NO |
| 26 | Police | Tarnow | Gold | Apartment | No | No | No | NO |
| 27 | Nurse | Radom | Classic | No | No | No | Senior | NO |
| 28 | Education | Catanzaro | Classic | Apartment | No | No | No | NO |
| 29 | Transportation | Warsaw | Gold | Car | ShortTerm | TechnologyShare | Regular | NO |
| 30 | Police | Cosenza | Classic | Car | No | No | No | NO |

Table 2: Rules generated by CN2 for data from Table 1. Coverage and confidence were computed in postprocessing.

| Rules for class big_spender='YES' | coverage | conf. |
|---|---|---|
| loan='No' & account='Gold' | 13.33% | 100.00% |
| occupation='Doctor' & deposit='No' | 10.00% | 100.00% |
| occupation='Health-care' | 6.67% | 100.00% |
| occupation='Doctor' | 16.67% | 83.33% |
| occupation='Education' & account='Gold' | 6.67% | 100.00% |

Table 3: Rules generated by g-SEGS from Table 1 data and ontologies from Figure 3.

| Rules for class big_spender='YES' | coverage | conf. |
|---|---|---|
| Occupation(Public) & BankingService(Gold) | 26.67% | 87.50% |
| Occupation(Doctor) | 20.00% | 83.33% |
| BankingService(Gold) | 46.67% | 64.29% |
| Location(Germany) & Occupation(Service) & BankingService(InvestmentFund) | 16.67% | 80.00% |
| Location(Bavaria) | 16.67% | 80.00% |

Characteristics of using g-SEGS (semantic data mining) are the following:

- more general rules compared to CN2 or other non-semantic data mining algorithms
- automated and therefore repeatable preprocessing - not prone to errors like human preprocessing
- g-SEGS rules have ontology terms with ontology names as conjuncts, while CN2 rules have attribute-value pairs.

## 5 CONCLUSIONS

This paper introduced the term *semantic data mining* which denotes a data mining approach where domain ontologies are used as background knowledge for data mining. We generalized a domain-specific system SEGS to perform semantic data mining on arbitrary ontology-annotated data.

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. Despite its current limitations, the new semantic data mining system g-SEGS shows major advantages compared to non-semantic systems, which include more general rules and automated data preprocessing. Hence, g-SEGS is a significant step towards practical semantic data mining.


## 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] 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.

[2] L. Brisson and M. Collard. How to semantically enhance a data mining process? In J. Filipe and J. Cordeiro, editors, *ICEIS*, volume 19 of *Lecture Notes in Business Information Processing*, pages 103–116. Springer, 2008.

[3] 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.

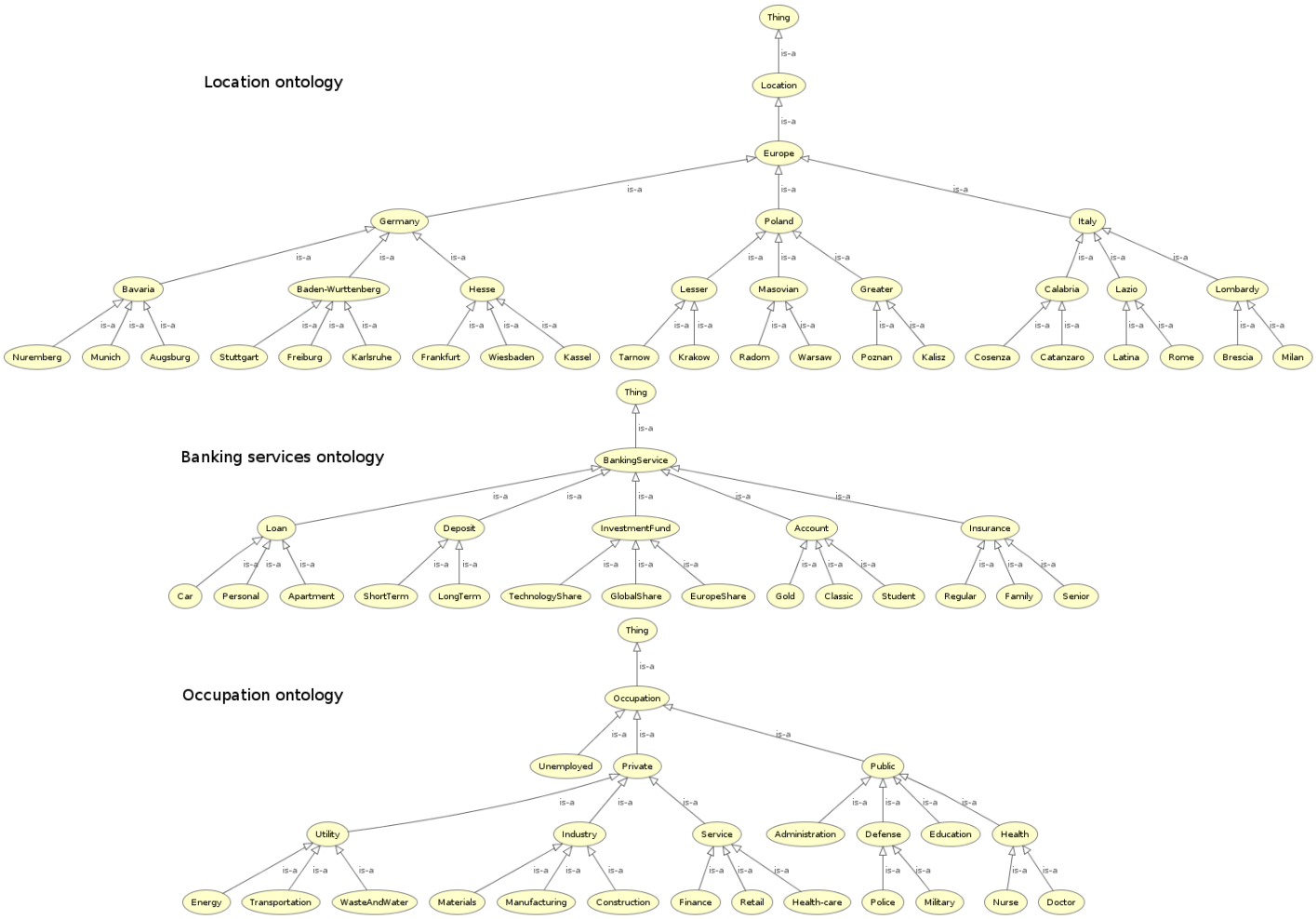


Figure 3: Three ontologies for data in Table 1.


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

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

[6] N. Lavrač, P. Kralj Novak, I. Mozetič, V. Podpečan, H. Motaln, M. Petek, and K. Gruden. Semantic subgroup discovery: Using ontologies in microarray data analysis. In *Proc. 31st Annual Intl. Conf. of the IEEE EMBS*, pages 5613–5616, 2009.

[7] 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.

[8] I. Mozetič, N. Lavrač, V. Podpečan, P. Kralj Novak, H. Motaln, M. Petek, K. Gruden, H. Toivonen, and K. Kulovesi. Bisociative knowledge discovery for microarray data analysis. In *Proc. First Intl. Conf. on Computational Creativity*, pages 190–199, 2010.

[9] V. Podpečan, M. Juršič, M. Žakova, and N. Lavrač. Towards a service-oriented knowledge discovery platform. In V. Podpečan and N. Lavrač, editors, *Third-generation data mining: towards service-oriented knowledge discovery*, pages 25–36, 2009.

[10] 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.

[11] 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.

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