SKEL's Research on Machine Learning for Information Integration on the Web

Georgios Paliouras

Software and Knowledge Engineering Laboratory
Institute of Informatics and Telecommunications,
National Centre for Scientific Research "Demokritos"
153 10 Ag. Paraskevi, Athens, Greece
paliourg@iit.demokritos.gr
http://www.iit.demokritos.gr/skel/

Abstract. This short paper highlights the research activity of the Software and Knowledge Engineering Laboratory (SKEL) on machine learning methods to support information Integration on the Web. SKEL belongs in the Institute of Informatics and Telecommunications of the National Centre for Scientific Research "Demokritos", and has recently coordinated the European research project CROSSMARC, which produced a complete prototype for information integration, employing state-of-the-art techniques for information gathering, information extraction, ontology maintenance and personalization. Following a brief description of the CROSSMARC architecture, this paper presents recent research activity in SKEL towards the automation of information extraction and ontology maintenance. For this purpose, novel machine learning methods have been developed in the areas of grammatical inference, stacked generalization and compression-based clustering.

1. Introduction

Information integration is one of the most promising solutions to the problem of information overload on the Web, aiming to identify the information that is most relevant to the needs and interests of a user (or a group of users) out of the vast quantity of available information. Clearly, this goal is very ambitious and requires the development of truly intelligent systems that are able to capture the users' requirements, while at the same time understanding the intentions of Web content editors. Such intelligent systems typically combine a variety of state-of-the-art technologies for information gathering, information extraction and personalization. For these technologies to be effective, the system needs to also incorporate extensive amounts of domain and task-specific knowledge, in the form of grammars, lexicons, ontologies, etc. The acquisition and maintenance of these knowledge resources is a serious bottleneck, which is worsened by the volatility of online information. Machine learning methods are proposed as a promising solution to this problem, facilitating the automation of the knowledge acquisition process, through the statistical analysis of the available information.

The Software and Knowledge Engineering Laboratory (SKEL) of the Institute of Informatics and Telecommunications in the National Centre for Scientific Research "Demokritos" has set as its main goal for the past decade to advance knowledge technologies that are required for overcoming the obstacle of information overload on the Web. Towards that goal, it has produced innovative research results in the whole chain of technologies employed by intelligent information integration systems: information gathering (retrieval, crawling, spidering), information extraction (named entity recognition and classification, role identification, wrappers), personalization (user communities and stereotypes). The recent emphasis of our research has been on the automation of intelligent system development, customization and maintenance, which involves mainly the employment of machine learning methods for knowledge acquisition, as mentioned above.

This short paper highlights SKEL's most recent research activity in the area of machine learning for information integration, focusing on two of the most demanding tasks: (a) learning for information extraction and (b) maintaining ontologies. The following section presents briefly the CROSSMARC architecture for information integration, which is the main result of the European research project CROSSMARC and provides the framework for our research in this area. Then, sections 3 and 4 present the main elements of our machine learning research for information extraction and ontology maintenance. Finally, section 5 concludes this brief and partial presentation of SKEL research, indicating promising paths for the future.

2. The CROSSMARC approach to information integration on the Web

CROSSMARC (Cross-lingual Multi Agent Retail Comparison)[1] was a European research project that was completed at the end of the year 2003. The main result of CROSSMARC was an open, agent-based architecture for cross-lingual information integration, incorporating the full chain of technologies involved in the process. Initially CROSSMARC was meant to focus on retail comparison systems that collect product information from various suppliers and present it to customers in a localized and personalized manner. In addition to this application however, the CROSSMARC architecture

[1] http://www.iit.demokritos.gr/skel/crossmarc/

has proven equally useful for other information integration tasks, such as employment search engines. Figure 1 presents the CROSSMARC architecture.

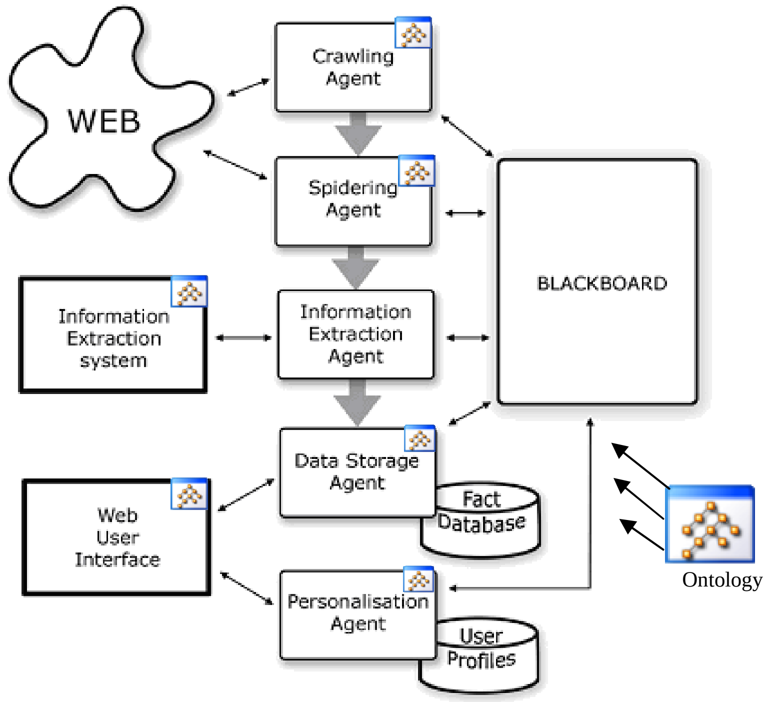

Figure 1. CROSSMARC's agent based architecture.

As mentioned above, CROSSMARC implements the full information integration process, using independent agents that communicate via a blackboard and share the same domain ontology. The information gathering stage is separated into a crawling and a spidering step, collecting interesting sites from the Web and relevant pages from these sites, respectively. Machine learning is used to learn to identify relevant pages and the most promising paths to those pages. The information extraction agent serves as a controller for a number of different information extraction systems, each handling a different language (English, French, Italian and Greek are currently covered). The results of information extraction are stored into the fact database, which is accessed by the end-users through a personalized Web interface. The agent-based design of the CROSSMARC architecture allows it to be open, distributed and customizable. The agents implementing each step of the process can be replaced by any other tool with the same functionality that respects the XML-based communication with the blackboard and the ontology. Furthermore, new information extraction systems, covering different languages can easily be connected to the information extraction agent. The ontology also plays an essential role in all stages, providing terms for modeling the relevance of Web pages, language-independent fact extraction, parameterization of the user models, etc. By collecting domain-specific knowledge in the ontology, the various agents become less dependent on the domain.

More details about the CROSSMARC architecture and the prototype can be found in [1] and [8]. The following sections concentrate on two aspects of the information integration system: the information extraction subsystems and the ontology.

3. Machine learning for information extraction

SKEL has a significant experience in the area of information extraction and the use of learning methods for the customization of information extraction systems to new domains, e.g. [5], [6]. More recently, our research has focused on two problems: learning context-free grammars for information extraction from free text and meta-learning for combining systems that extract information from semi-structured text. The following two subsections highlight our most recent results in the two areas.

3.1. Learning context-free grammars from positive examples

Grammars are an important component of information extraction systems and their construction is a major hurdle in the customization of the system to a new domain. Information extraction from free text usually requires context-free (or even context-sensitive) grammars, which are hard to learn from data. Even harder is the acquisition of these grammars from positive only examples, which is practically the only type of example that a human annotator can provide. This is the reason why there are very few learning methods that deal with this problem, which are usually only applicable to datasets of small size and complexity. In an attempt to overcome this problem we have developed the e-GRIDS algorithm [3], which is based on the same principles as the GRIDS [2] and the SNPR [11] algorithms, but improves them substantially, in order to become applicable to realistic problems.

e-GRIDS performs a beam search in the space of grammars that cover the positive examples, guided by the MDL principle. In other words, it favours simpler grammars, in the sense that the sum of their code length and the code length of the data, assuming knowledge of these grammars, should be small. The starting state for the search is the most specific grammar, which covers only the training data. The search operators compress and generalize the grammars, by merging symbols and creating new ones. The latest version of e-GRIDS, called eg-GRIDS [4], replaces the beam search with a genetic one, within which the grammar-modification operators are treated as mutation operators. Figure 2 depicts graphically the eg-GRIDS architecture.

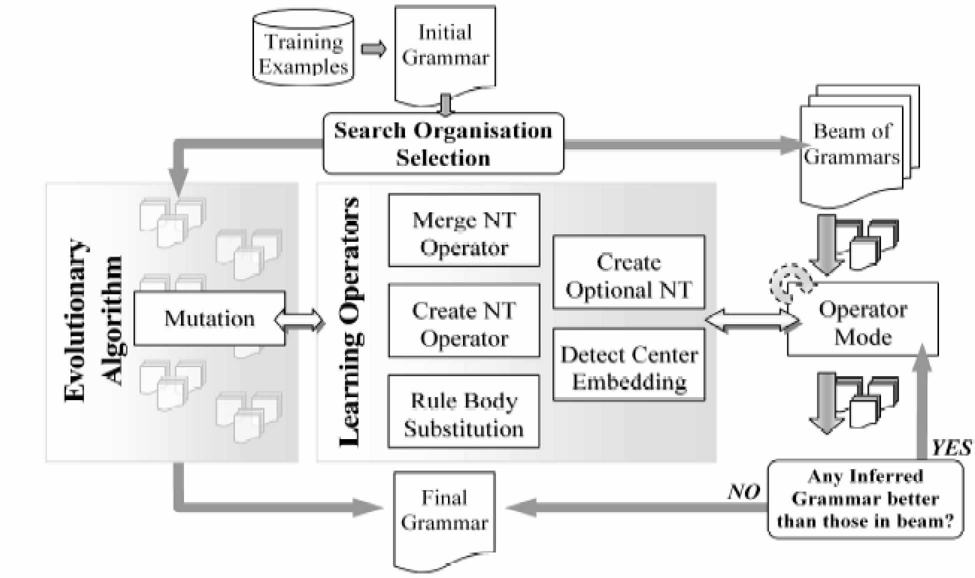

Figure 2. The architecture of the eg-GRIDS algorithm.

The use of genetic search has provided a speed-up of an order of magnitude, facilitating the inclusion of more operators that allow the algorithm to search a larger part of the space and produce much better results. Thus, eg-GRIDS can handle larger datasets, as well as more complex grammars.

3.2. Stacked generalization for information extraction

The extraction of information from semi-structured text is a slightly different task, focusing on the significance of formatting commands in the text and the dynamic construction of many pages from the same template. For this reason, simpler information extraction "grammars" are constructed, which however are usually only applicable within the boundaries of a specific Web site. For this reason, automated acquisition of these "grammars" is also essential.

Our research in this area has focused on the combination of different learning methods in a meta-learning framework, aiming to improve recognition performance by combining their strengths. For this reason, we have developed a stacked generalization framework that is suitable for information extraction, rather than classification which is the typical use of this approach. Figure 3 illustrates ths use of the stacking framework proposed in [7], both at training and at run-time.

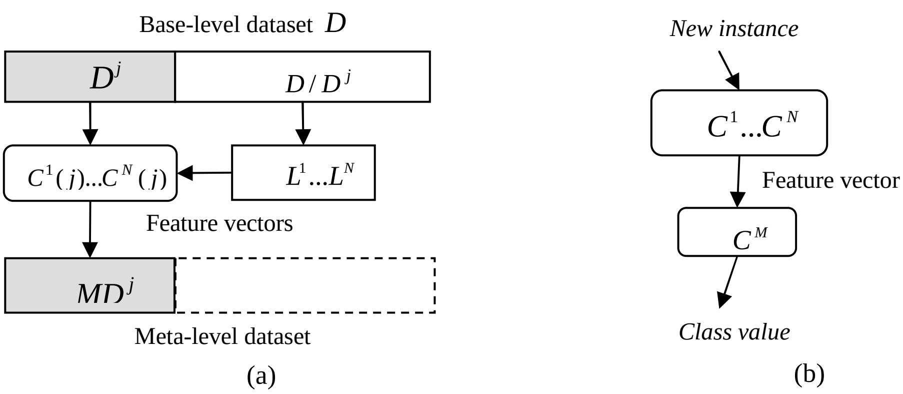

Figure 3. (a) Illustration of the J-fold cross-validation process for creating the meta-level dataset. (b) The stacking framework at runtime.

At training time, the usual cross-validation approach of stacked generalization is followed, which trains all base-level learners ($L^1 ... L^N$) on various subsets of the training dataset (D / D^j) and applies the learned systems ($C^1(j) ... C^N(j)$) on the unseen parts of the dataset (D^j), in order to construct the meta-level dataset (MD^j). However, in the case of information extraction the trained systems may extract different or contradictory information from the same subset of the data. Based on the confidence scores of the information extraction systems, the proposed framework combines their results into a common dataset that is suitable for training a classifier to choose whether to accept or reject an extracted piece of information and if accepted to recognize its type. This approach has led to considerable improvement in recognition performance, which is due to the significant difference in the behavior of the trained base-line systems.

4. Machine learning for ontology maintenance

As mentioned in section 2, the ontology plays an essential role in all stages of the information integration process. However, the relationship of the ontology with the information extraction system is particularly interesting. First, the ontology can be used itself for information extraction, by identifying the occurrence of its instances in the text. In many cases, this may provide an acceptable level of recognition performance. On the other hand, if we have a separate information extraction system, the information extracted by this system can be used to populate the ontology with new instances. We have exploited this two-way interaction between the ontology and the extraction system, in an iterative methodology for maintaining an existing ontology [10]. This methodology is presented in figure 4.

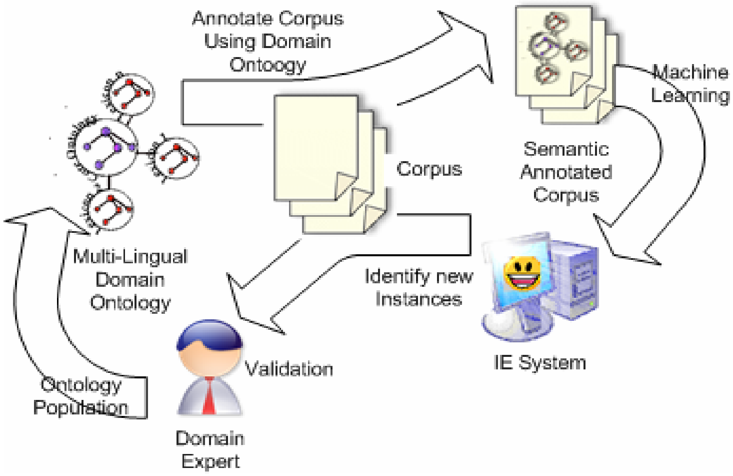

Figure 4. Ontology maintenance methodology.

The proposed process usually commences by an existing domain ontology, which is used to annotate a corpus of raw documents. In this manner, a training corpus for information extraction is formed, without the need for a human annotator. This can lead to a significant speed-up in the development of the information extraction system. The trained system usually generalizes beyond the annotated examples. Therefore, if applied again on the corpus it provides some new instances that do not appear in the initial ontology. These instances are screened by an expert who is responsible for maintaining the ontology. Once the ontology is updated, it can be used again to annotate a new training corpus that will lead to a new information extraction system. This process is repeated until no new instances are added to the ontology. Our initial experiments have shown that this approach works impressively well, even when the initial ontology is very sparsely populated. As a further improvement of this method, we have developed a novel compression-based clustering algorithm, called COCLU [9], which is responsible for identifying lexical variations of existing instances and clustering the lexical variations of new instances. This improvement minimizes the human effort in ontology maintenance.

5. Conclusions

This short paper has presented the recent research activity of the Software and Knowledge Engineering Laboratory (SKEL), focusing on the hardest tasks of learning to extract information and maintaining a domain ontology. For the task information extraction, two recently-published approaches were mentioned: (a) eg-GRIDS for learning context-free grammars that are suitable for free-text extraction and (b) a methodology for stacked generalization of trainable information extraction systems. The former has made learning of context-free grammars applicable to realistic problems, while the latter has been shown to improve considerably the results of existing information extraction systems. Furthermore, a new methodology for the maintenance of domain ontologies was mentioned, which employs an ontology to construct a training dataset for an information extraction system, without the need for human annotation, and then uses this system to suggest instances that are missing from the ontology. This methodology is enhanced by COCLU, a compression-based algorithm for clustering lexical variations of the instances.

We believe that these new methods take us a step closer to the goal of automating the acquisition and maintenance of knowledge resources that are vital elements of the information integration process. We believe that this is a goal worth-pursuing further as this knowledge acquisition process remains a major obstacle in the way of scaling up information integration. In this direction, we are studying the synergies between ontologies, grammars and other structured models for information integration. Furthermore, we are extending our ontology maintenance methodology to cover additional types of relationship, other than lexical variability, and we are investigating the addition of concept discovery.

Acknowledgements

This short paper does not present work solely undertaken by the author, but rather highlights the recent research activity of SKEL on machine learning for information integration. As a result, a number of current and past SKEL members have been involved in the presented work. This research was partially funded by the EC through the IST-FP5 project CROSSMARC (Cross-lingual Multi Agent Retail Comparison), contract number IST-2000-25366. The development of the CROSSMARC architecture has been carried out jointly be the partners of the project: National Centre for Scientific Research "Demokritos" (GR), VeltiNet A.E. (GR), University of Edinburgh (UK), Universita di Roma Tor Vergata (IT), Lingway (FR).

References

1. V. Karkaletsis, C.D. Spyropoulos. "Cross-lingual Information Management from Web pages", *Proceedings of the 9th Panhellenic Conference in Informatics* (PCI), Thessaloniki, 21-23 November, 2003.
2. P. Langley, S. Stromsten. "Learning Context-Free Grammars with a Simplicity Bias", *Proceedings of the Eleventh European Conference on Machine Learning* (ECML), *Lecture Notes in Artificial Intelligence* 1810, pp. 220-228, Barcelona, Spain, Springer-Verlag, 2000.
3. G. Petasis, G. Paliouras, V. Karkaletsis, C. Halatsis and C.D. Spyropoulos, "e-GRIDS: Computationally Efficient Grammatical Inference from Positive Examples", *Grammars*, 2004. Available from http://217.125.102.104/special4.asp
4. G. Petasis, G. Paliouras, C.D. Spyropoulos, C. Halatsis. "eg-GRIDS: Context-Free Grammatical Inference from Positive Examples using Genetic Search," *Proceedings of the 7th International Colloquium on Grammatical Inference* (ICGI), *Lecture Notes in Artificial Intelligence* 3264, pp. 220 - 231, Springer Verlag, 2004.
5. G. Petasis, F. Vichot, F. Wolinski, G. Paliouras, V. Karkaletsis, and C.D. Spyropoulos, " Using Machine Learning to Maintain Rule-based Named-Entity Recognition and Classification Systems", *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (ACL), pp. 426-433, Toulouse, 2001.
6. G. Petasis, A. Cucchiarelli, P. Velardi, G. Paliouras, V. Karkaletsis, and C.D. Spyropoulos, "Automatic adaptation of proper noun dictionaries through co-operation of machine learning and probabilistic methods". *Proceedings of the 23rd ACM SIGIR Conference on R&D in IR* (SIGIR), pp. 128-135, Athens, Greece, 2000.
7. G. Sigletos, G. Paliouras, C. D. Spyropoulos, P. Stamatopoulos. "Stacked generalization for information extraction," *Proceedings of the European Conference in Artificial Intelligence* (ECAI), pp. 549 – 553, Valencia, Spain, 2004.
8. C.D. Spyropoulos, V. Karkaletsis, C. Grover, M.-T. Pazienza, D. Souflis, J. Coch. *Final report of the project CROSSMARC (Cross-lingual Multi Agent Retail Comparison),* October 2003. http://www.iit.demokritos.gr/skel/crossmarc/AnnualReport2003/
9. A. Valarakos, G. Paliouras, V. Karkaletsis, G. Vouros. "A Name-Matching Algorithm for Supporting Ontology Enrichment," *Proceedings of the Panhellenic Conference in Artificial Intelligence* (SETN), *Lecture Notes in Artificial Intelligence* 3025, pp. 381-389, Springer Verlag, 2004.
10. A. Valarakos, G. Paliouras, V. Karkaletsis, G. Vouros. "Enhancing the Ontological Knowledge through Ontology Population and Enrichment," *Proceedings of the International Conference on Knowledge Engineering and Knowledge Management* (EKAW), *Lecture Notes in Artificial Intelligence* 3257, pp. 144-156, Springer Verlag, 2004.
11. G. Wolff. "Grammar Discovery as data compression", *Proceedings of the AISB/GI Conference on Artificial Intelligence*, pp. 375-379, Hamburg, Germany, 1978.