Statistical Classification for Wrapper Induction

Rémi Gilleron, Patrick Marty, Marc Tommasi, Fabien Torre

Mostrare Project , INRIA Futurs Lille
and University of Lille 3 - France

Abstract

The standard document formats of the Web today, `HTML` and `XML`, rely on tree structures that encompass textual information. The main goal of the Mostrare project is to incorporate novel approaches for modeling tree structure and emerging techniques for machine learning into adaptive information extraction systems for the Web. Semantic Web metadata could contribute to improve information extraction tools.

In this paper we present an approach based on statistical classification for wrapper induction. We have developed the CaFeIn platform. It is parameterized by a document representation model and a supervised classification algorithm that can operate on that model. Currently CaFeIn includes a number of feature-sets for textual and structured documents, that can be easily customized or extended. In the future, CaFeIn might also have to account for semantic information.

1 Information Extraction from Semistructured Documents

During the last decade, the World Wide Web has evolved to the most important public data store on world. The web community is highly interested in adequate information representation so that information on the web can be accessed more easily. A major challenge in that perspective is adaptive information extraction that can fully exploit the tree structure of web documents. Tree structure is present in the recent web formats, `HTML` and `XML`, in order to encompass textual information.

Information extraction means to populate a database with values extracted from a collection of documents. Programs that extract information are called wrappers. Writing wrappers by hand is inappropriate, extremely laborious and error prone. Therefore we consider the automatic design of information extraction systems from examples using machine learning techniques. Previous approaches are studied in the field of *wrapper induction*. Data formats on the web are diverse, but tree structured data will become prominent.

The objective of the Mostare project is to develop adaptive information extraction systems for semi-structured documents, that can fully exploit available tree structure. We approach this goal in two research lines:

- *Modeling Tree Structure for Information Extraction*: define and investigate models of tree structures as needed by information extraction; develop corresponding algorithms and software components ([2, 1]).

- *Machine Learning for Information Extraction:* develop learning algorithms that induce models of tree structures and apply them to information extraction ([3]). Combine learning algorithms for tree and string models so that they apply to diverse data formats, and possibly to heterogeneous data ([8]).

Our approach should be sufficiently flexible so that we can integrate semantic information on the web once available. Tree based wrapper could be use to enrich semistructured data with semantic metadata annotations. Those annotations can be done by adding semantic tags into the tree structure, or by replacing initial tags by much more informative ones.

2 Statistical Classification for Wrapper Induction

When considering heterogeneous HTML and XML documents as inputs of information extraction tasks – that may be either scattered into pieces or hidden in large parts of purely textual data – different tree wrappers need to be combined with textual wrappers. For classification tasks, some machine learning algorithms realize such combinations. We have started to examine how such methods perform for information extraction. In a first step, we reformulate information extraction as classification tasks. Second we apply known techniques and combine them with structural wrapper induction.

The CaFeIn framework ([7, 8]) improve previous approaches based on supervised classification for information extraction ([6, 4, 5]). CaFeIn is a single-slot wrapper induction framework for text data or semi-structured data. It is parameterized by a document representation model and a supervised classification algorithm that can operate on that model. Any supervised classification algorithm can be used in the CaFeIn framework.

For text data, the classification task consists in deciding whether a tuple of two positions in a text are respectively the beginning and the end of data to extract. The document representation use only textual features. For semi-structured data, like `XML` or `HTML` trees, data to extract are contained in leaves. Thus the classification task consists in deciding whether a leaf is to be extracted or not. In that case, the document representation can use textual features, tree structure based features or combinations of both textual and tree view of document.

In CaFeIn, the document representation model is an adaptive attribute-value representation. Currently CaFeIn includes a number of feature-sets for textual and structured documents, that can be easily customized or extended. This allows the integration of domain knowledge easily or of semantic web ressources such as ontologies.

We have developed a CaFeIn prototype and examined its performance with different learning algorithms for classification (C4.5 [9], GloBoost [10]) and with different models of data representation. CaFeIn is competitive with the others wrapper induction systems based on specialized learning algorithms. It will be continued with multi-slots wrappers induction, and with a deeper integration of semantic information and semantic web ressources.

References

[1] Y. André, A.-C. Caron, D. Debarbieux, Y. Roos, and S. Tison. Extraction and implication of path constraints. In *Proceedings of the 29th Symposium on Mathematical Foundations of Computer Science*, volume 3153 of *Lecture Notes in Computer Science*, pages 863–875, Prague (Czech Republic), august 2004. Springer Verlag.

[2] I. Boneva, J.-M. Talbot, and S. Tison. Expressiveness of Spatial Logic for Trees, 2005. submitted.

[3] J. Carme, A. Lemay, and J. Niehren. Learning node selecting tree transducer from completely annotated examples. In *7th International Colloquium on Grammatical Inference*, volume 3264 of *Lecture Notes in Artificial Intelligence*, pages 91–102. Springer Verlag, 2004.

[4] H. L. Chieu and H. T. Ng. A maximum entropy approach to information extraction from semi-structured and free text. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI 2002).*, pages 786–791, 2002.

[5] A. Finn and N. Kushmerick. Multi-level boundary classification for information extraction. In *In Proceedings of the European Conference on Machine Learning, Pisa, 2004.*, 2004.

[6] D. Freitag and N. Kushmerick. Boosted wrapper induction. In *AAAI/IAAI*, pages 577–583, 2000.

[7] P. Marty and F. Torre. Classer pour extraire : représentations et méthodes. Technical Report Grappa report 0103, GRAPPA, december 2003.

[8] P. Marty and F. Torre. Codages et connaissances en extraction d'information. In M. L. et Marc Sebban, editor, *Actes de la Sixième Conférence Apprentissage CAp'2004*, pages 207–222. Presses Universitaires de Grenoble, 2004.

[9] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.

[10] F. Torre. Boosting correct least general generalizations. Technical Report 0104, GRAPPA, april 2004.