# Back to Gold's Age: Bridging the Gap Between Traditional Grammar Inference and Web Information Extraction

Valter Crescenzi, Giansalvatore Mecca, Paolo Merialdo

Università di Roma Tre Università della Basilicata Università di Roma Tre
crescenz@dia.uniroma3.it mecca@unibas.it merialdo@dia.uniroma3.it

## 1 Introduction

Since Gold's Theorem (1967), grammar inference for regular languages has been a thoroughly studied topic, with an elegant theoretical background and well established techniques. One of the main contributions of these works is the study of properties of those languages for which the inference process can be performed in a completely automatic way, and of the relative algorithms.

Recently, information extraction from Web sites has imposed itself as a relevant research field for the so called "Semantic Web". Since extraction is performed by *wrappers*, which are essentially grammar parsers for the HTML code of Web pages, grammar inference could in principle play a fundamental role in this field. However, despite 30 years of research, due to some limitations of the traditional framework, essentially none of the recent approaches to Web information extraction reuse theories and techniques from the grammar inference community. As a consequence, most of the techniques that have been proposed do not have a formal background from the grammatical viewpoint. Moreover, they are essentially non-automatic, in the sense that the inference process heavily relies on the intervention of a human trainer.

Is there a chance to reconcile these two research communities?

To answer this question, in this paper we discuss what are the main limitations of traditional grammar inference techniques when used for automatic wrapper generation. Our considerations directly derive from recent advances in the RoadRunner [CMM01] project, an ongoing research effort for fully automatic wrapper generation. Based on this, we discuss what we believe is a relevant contribution of the project, by showing how it extends and reconciles traditional grammar inference with modern information extraction.

## 2 The Grammar Inference Inheritance

Grammar inference is a well known and extensively studied problem (for a survey of the literature see, for example [Pit89]).

Gold gave a simple and widely accepted model of inductive learning also called "learning from text" which goes on as follows [Gol67]: ($a$) consider a target

language $L$ we want to infer, for example a regular expression like `a(bc)+(d)?`[1]; ($b$) assume a learner is given a (possibly infinite) sequence of positive samples of $L$, that is, a sequence of strings belonging to $L$, like, for example `abcbc, ad, abcd` ...; ($c$) after each sample, the learner produces as output a new guess on the target language.

Intuitively, each new sample adds new knowledge about the language, and therefore can help in identifying the solution. We say that a class of languages is *identifiable in the limit* if, for each language $L$ in the class, the learner converges towards the right hypothesis after a finite number of positive examples.

Unfortunately, not all languages are inferrable in the limit. Gold himself produced the first negative results on the inductive inference of grammars (Gold's Theorem [Gol67]). To give an example, from his theorem it follows that even regular languages cannot be identified in the limit. As a consequence, the large body of research on inductive inference that originated from Gold's seminal works has concentrated on the problem of finding restricted classes of regular grammars for which learning from positive data is possible.

In the early '80s Angluin has posed several milestones on this subject by finding necessary and sufficient conditions for an indexed class of languages to be inferable from positive examples [Ang80]. Based on her work, subsequent works introduced several classes; a prominent example is the class of $k$-reversible grammars [Ang82], which were proven to be identifiable in the limit, and for which algorithms were developed and formally proven to be correct.[2]

Due to their property of being inferrable in the limit, these classes represent a natural candidate for automatic wrapper generation on the Web. However, there is a number of serious shortcomings associated with this approach that seriously limit its practical applicability. Let's consider for example $k$-reversible grammars. This is a family of classes of languages (based on the value of $k$, we have respectively 0-reversible languages, 1-reversible languages, 2-reversible languages etc.); each of these classes is a subset of the regular languages.[3]

The main limitations of $k$-reversible grammars are as follows:

- first, the inference algorithm for $k$-reversible grammars assumes that the value of $k$ is known to the inference system; each value of $k$ has its own algorithm, so that the algorithm for 2-reversible languages does not return correct results on 1-reversible or 3-reversible samples;
- second, to produce a correct result, the algorithm assumes that the inference system is given a *characteristic sample* of the language; intuitively, the characteristic sample of a language $L$ is a sort of finite *fingerprint* that discriminates $L$ from any other language of the class; in automata-theoretic

[1] As usual, in the following, + means "one or more occurrence", ? means "zero or one occurrence", i.e, "optional", | means disjunction

[2] Recently, Fernau has generalized Angluin's work about reversible languages [Fer02] and other works, including those on terminal distinguishable languages [RN87]. A discussion about the relationship between Fernau's generalization and our work is outside the scope of this paper, and will be discussed in a forthcoming work [CM02].

[3] We omit the formal definition for space reasons.

| *a. John Doe's web page* | *b. Frank Smith's Web page* |
|---|---|
| <html><body><br>  <h1>John Doe</h1><br>  DB Group <br>  <ul><br>  <li>Distrib. Systems</li><br>  <li>Programming 101</li><br>  <li>Advanced Prog.</li><br>  </ul><br></body></html> | <html><body><br>  <h1>Frank Smith</h1><br>  <br>  <ul><br>  <li>Operating Systems</li><br>  <li>Databases</li><br>  </ul><br></body></html> |

**Fig. 1.** Two sample Web pages

terms, it is a set of samples with two main characteristics: (*a*) informally speaking, it is a set of samples that has the property of "covering" the whole automaton, i.e., touching all states and traversing all transitions of the language automaton; (*b*) among all the sample sets that have this property, it is the one with the minimal length.

To give an example, suppose we have a database storing data about professors and their courses: assume that each professor teaches one or more courses and s/he is member of at most one research group. Suppose a Web sites publishes pages generated from the database, like those in Figure 1.

It is easy to verify that these pages belong to the following grammar (here, $\Delta^+$ denotes a placeholder for strings to extract):

<html><body><br>  <h1>$\Delta^+_{name}$</h1><br>   ( $\Delta^+_{resGroup}$ )? <br>  <ul><br>  (<li>$\Delta^+_{course}$</li>)+<br>  </ul><br></body></html>

Note also that the grammar can easily be used as a wrapper for extraction purposes, i.e., to parse the pages and extract the original dataset.

Unfortunately, even though the grammar can be shown to be 1-reversible, it is not possible to apply algorithms for $k$-reversible grammars to this example, for two reasons. First, we need to know the value of $k$ (1 in this case) before applying the algorithm. This means that, before being able to actually apply the algorithm, we need to find a guess for $k$; finding such a guess is definitely not a simple task [Chi00].

Second, even if we know by some means the correct value for $k$, still we need an input characteristic sample, otherwise the output will not be correct. Requiring that the input is a characteristic sample is a serious drawback. In fact, when wrapping a Web site it is not reasonable to assume that input pages have the minimal length property. For example, the two pages shown below do

not represent a characteristic sample for the grammar; in fact, even though they have the properties of fully "covering" the language automaton, still they do not have minimal lengths (this is due to the fact that there are respectively two and three courses taught by the two professors, whereas in a sample of minimal length only one or two course may appear). As a consequence, if we run the algorithm for 1-reversible grammars on the above pages, the output would be the following:

```
<html><body>
 <h1>Δ+name</h1><p>
 (</p><ul><li>Distrib. Systems</li><li> | <b>DBGroup</b></p><ul><li>)
  Δ+course</li>
  <li>Δ+course</li>
 </ul>
</body></html>
```

It can be easily seen that such a grammar is clearly useless for extraction purposes (for example, the list of courses has not been recognized).

## 3 Information Extraction for the Web

The limitations discussed in the previous section have made traditional grammar inference techniques little appealing to researchers in the information extraction field. As a consequence, most of the recent approaches to information extraction depart from Gold's model of inductive inference.

In essence, they give up the attempt to find classes of grammars that are inferrable in the limit, for which automatic algorithms exist; on the contrary, they consider the full class of regular grammars (or subclasses for which the property of inferrability is not known) and study algorithms for these classes. The study of these algorithms has evolved along two major direction: some researchers have proposed solutions that work in presence of additional information (typically a set of labelled examples or a knowledgeable teacher's responses to queries posed by the learner); others have investigated approaches that allow for a certain imprecision in the inference process [KY97], for instance within the *probably approximately correct* (PAC) model proposed by Valiant [Val84].

Many recent proposals [Ade98,Fre98,Sod99,KWD97,MMK99,EJY99] move along these lines. These works have studied the problem of (semi-)automatically generating the wrappers for extracting data from fairly structured HTML pages. The main limitation of these systems is that they are not completely automatic; in fact:

- they need a labor intensive *training phase*, in which the system is fed with a number of labelled examples (i.e., pages that have been labelled by a human expert to mark the relevant pieces of information);
- the algorithms assume a a-priori knowledge about the organization of data in the target pages; in particular, most of these approaches consider that pages contain a list of flat records, and nesting is not allowed;

The fact that the process is not completely automatic is somehow unsatisfactory for a Web information extraction system. In fact, after a wrapper has been inferred for a bunch of target pages, a small change in the HTML code can lead to the wrapper failure and, as a consequence, the labelling and training phase has to be repeated (involving the a human intervention again).

## 4 The ROADRUNNER Inference Engine

Compared to these works, the ROADRUNNER [CMM01] project was conceived as an effort to overcome the limitations of semi-automatic information extraction proposals, by making the information extraction process fully automatic. In this sense, the system neither relies on user-specified examples, nor on any a priori knowledge about how the page contents are organized, i.e., the schema they follow; moreover, the system is not restricted to flat records, but can handle arbitrarily nested structures.

At the same time, a priority in the project has been that of developing a formal theoretical framework for studying the computational properties of the algorithms that were developed. In this sense, one of our goals was to establish a clear connection with Gold's inductive inference, in order to reuse known theoretical results.

A key feature of ROADRUNNER is that it is focused on *data-intensive web sites*, i.e., large collections of Web pages generated from an underlying database. We consider a nested relational data model [AB95], that is, typed objects with nested sets and tuples. Tuples have attributes (possibly optional), which in turn may be either atomic attributes or sets of tuples. Figures 2.a and b show a convenient way to represent types and instances by means of trees. Their serialization in HTML may be modelled by means of *Mark-Up Encoding Functions* which associate to every node of the schema tree (and therefore to all corresponding nodes which are its instances in the instance tree) a pair of strings (see Figure 2.c). The serialization of an instance is reduced to an in-order visit of the corresponding tagged tree. For example, the instance of Figure 2.c would be serialized into the HTML code of Figure 2.d.

We call *mark-up languages* the class of languages that correspond to encodings of instances of nested types. This is a subclass of regular languages. Each language of the class corresponds to a certain type and encoding function. In this framework the wrapper generation problem can be described as the problem of inferring the language associated to a type given the encodings of a collection of nested relational instances of that type.

A key contribution of the project stands in the definition of a subclass of mark-up encodings, called *Prefix Mark-Up Encodings* [CM02]. The class of languages corresponding to Prefix Mark-Up Encodings has a number of nice features. First, it is possible to show that it is identifiable in the limit; an important consequence is that it makes sense to study fully automatic algorithms for grammars in this class; in particular, it is possible to prove that the algorithms we have developed in [CMM01] correctly infer these languages in the limit. Second,

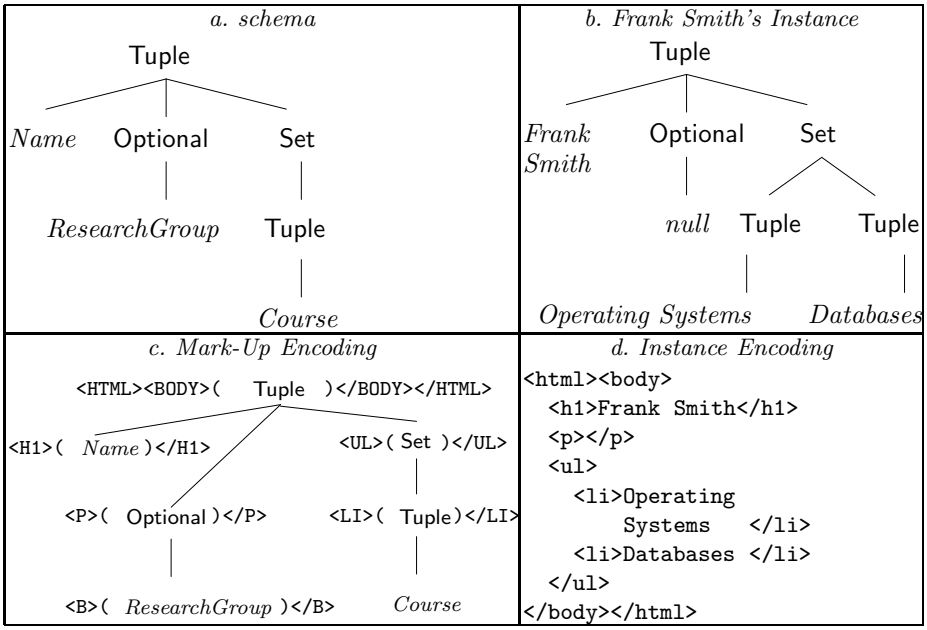


**Fig. 2.** A Schema about Professors, their Research Group and Teaching Activities

the characteristic samples we have found seem to be particularly interesting because they have a nice and clean interpretation in term of database types. Consider a collection of database instances sharing the same type: instances can sometimes underuse their type. This happens, for example, when set types are used to model objects that always have the same cardinality (and would have been more accurately typed with a tuple); as another example consider when an attribute has always the same value (and could have been omitted); a final example occurs when an optional attribute is either never or always *null* (it either could have been considered non-optional or omitted). The concept of *rich* collection of instances [GM99, Definition 1] summarized this consideration as follows: A collection of instances $I$ is *rich* with respect to the common type if it satisfies the following three properties:

- *set richness*: each set node is instantiated in $I$ with at least two distinct cardinalities;
- *optional richness*: each optional node is instantiated in $I$ with at least one *null* and one non-*null* object;
- *basic richness*: each leaf node is instantiated in $I$ with at least two distinct constants.

Intuitively, a collection of instances is rich with respect to a type if it makes full use of this type, and it then contains enough information to recover the type. It can be shown that the characteristic samples for prefix mark-up languages

are exactly sets of encodings of rich collections of instances. Intuitively, this definition requires that the set of samples covers all states and all transitions in the language automaton, but there is no requirement of minimality in the length of samples.

This new notion of characteristic sample represents the most significative difference with respect to the traditional setting, since it makes the input needed for grammar inference more "natural" and practically available in our case. To show this more practically, consider again the example about professors and research group introduced in Section 2. The two pages shown in Figure 1 are a rich set of samples for the target grammar.

## 5 Contributions

We believe that the theory of inductive inference provides an elegant theoretical setting that can lead to significantly better and formally understandable algorithms for wrapper generation. This inheritance has been essentially ignored by works in Web information extraction because of the limitations of the algorithms known so far, and in particular:

1. in order to start the wrapper generation phase, it is necessary to choose a specific class of languages (i.e., a value of $k$ in the case of $k$-reversible languages) before the wrapper inference can start, but usually the right class is not known a priori;
2. the assumption that the input to the wrapper generator is a characteristic sample is quite unrealistic in the Web framework, were samples are chosen in a random way.

RoadRunner solves these problems by identifying a natural class of languages, with a direct counterpart in terms of database structures and HTML tagging; in some sense this class represents a trade-off between expressiveness of the languages and effectiveness of the inference algorithm, since it allows to do inference using a natural notion of input.

Our future work consists in trying to reach an optimal trade-off between these two aspects, augmenting the expressiveness of prefix mark-up language without compromising the naturalness of the corresponding characteristic sample. In this sense, the addition of an union type to our database-like data model seems to be a quite promising direction.

## References


[AB95] S. Abiteboul and C. Beeri. On the power of languages for the manipulation of complex objects. *VLDB Journal*, 4(4):117–138, 1995.

[Ade98] B. Adelberg. NoDoSE – a tool for semi-automatically extracting structured and semistructured data from text documents. In *ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'98), Seattle, Washington*, 1998.

[Ang80] D. Angluin. Inductive inference of formal languages from positive data. *Information and Control*, (45):117–135, 1980.

[Ang82] D. Angluin. Inference of reversible languages. *Journal of the Association for Computing Machinery*, 29(3):741–765, 1982.

[Chi00] B. Chidlovskii. Wrapper Generation by $k$-reversible Grammar Induction. *Proc. Int. Workshop on Machine Learning and Information Extraction (ECAI'00)*, 61–72, 2000.

[CM02] V. Crescenzi and G. Mecca. In preparation. 2002.

[CMM01] V. Crescenzi, G. Mecca, and P. Merialdo. RoadRunner: Towards automatic data extraction from large web sites. In *Int. Conf. on Very Large Data Bases (VLDB'2001), Roma, Italy, September 11-14*, pages 109–118, 2001.

[EJY99] D. W. Embley, Y. S. Jiang, and Ng Y. Record-boundary discovery in web documents. In *ACM SIGMOD International Conf. on Management of Data*, pages 467–478, 1999.

[Fer00] H. Fernau. On learning function distinguishable languages. Technical Report WSI-2000-13, Wilhem-Schickard-Institut für Informatik, 2000.

[Fer02] H. Fernau. Identification of function distinguishable languages. *Theoretical Computer Science*, 2002. To Appear.

[Fre98] D. Freitag. Information extraction from html: Application of a general learning approach. In *Proceedings of the Fifteenth Conference on Artificial Intelligence AAAI-98*, pages 517–523, 1998.

[GM99] S. Grumbach and G. Mecca. In search of the lost schema. In *Seventh International Conference on Data Base Theory, (ICDT'99), Jerusalem (Israel), Lecture Notes in Computer Science, Springer-Verlag*, pages 314–331, 1999.

[Gol67] E. M. Gold. Language identification in the limit. *Information and Control*, 10(5):447–474, 1967.

[KWD97] N. Kushmerick, D. S. Weld, and R. Doorenbos. Wrapper induction for information extraction. In *International Joint Conference on Artificial Intelligence (IJCAI'97)*, 1997.

[KY97] S. Kobayashi and T. Yokomori. Learning approximately regular languages with reversible languages. *Theoretical Computer Science*, 174:251–257, 1997.

[MMK99] I. Muslea, S. Minton, and C. A. Knoblock. A hierarchical approach to wrapper induction. In *Proceedings of the Third Annual Conference on Autonomous Agents*, pages 190–197, 1999.

[Pit89] L. Pitt. Inductive inference, DFAs and computational complexity. In K. P. Jantke, editor, *Analogical and Inductive Inference, Lecture Notes in AI 397*, pages 18–44. Springer-Verlag, Berlin, 1989.

[RN87] V. Radhakrishnan and G. Nagaraja. Inference of regular grammars via skeletons. *IEEE Transactions on Systems, Man and Cybernetics*, 17(6):982–992, 1987.

[Sod99] S. Soderland. Learning information extraction rules for semistructured and free text. *Machine Learning*, 34(1–3):233–272, 1999.

[Val84] L. G. Valiant. A theory of learnable. *Communication of the Association for Computing Machinery*, 27:1134–1142, 1984.