A Fast Algebraic Web Verification Service[⋆]

M. Alpuente[1], D. Ballis[2], M. Falaschi[3], P. Ojeda[1], and D. Romero[1]

[1] DSIC, Universidad Politécnica de Valencia
Camino de Vera s/n, Apdo. 22012, 46071 Valencia, Spain.
{alpuente, pojeda, dromero}@dsic.upv.es
[2] Dip. Matematica e Informatica
Via delle Scienze 206, 33100 Udine, Italy.
demis@dimi.uniud.it
[3] Dip. di Scienze Matematiche e Informatiche
Pian dei Mantellini 44, 53100 Siena, Italy.
moreno.falaschi@unisi.it

Abstract. In this paper, we present the rewriting-based, Web verification service WebVerdi-M, which is able to recognize forbidden/incorrect patterns and incomplete/missing Web pages. WebVerdi-M relies on a powerful Web verification engine that is written in Maude, which automatically derives the error symptoms. Thanks to the AC pattern matching supported by Maude and its metalevel facilities, WebVerdi-M enjoys much better performance and usability than a previous implementation of the verification framework. By using the XML Benchmarking tool xmlgen, we develop some scalable experiments which demonstrate the usefulness of our approach.

1 Introduction

Web-site management is an arduous task today. While it is not so difficult to build a Web site, this task is still somehow a form of art, resulting in an increasing volume of information contained in ever-larger, complex Web sites, which is very difficult to keep up-to-date and correct. Recently, some sophisticated Web-site management tools have been proposed. These provide helpful facilities (see [4, 5]), but unfortunately, they are mostly oriented to Web-site syntactic checking/restructuring so that they can do little by themselves to relieve the problem. A lot of research work has been invested in consistency management and repair of software applications and databases, whereas similar technologies are much less mature for Web systems.

The automated management of data-intensive Web sites is an area to which rule-based technology has a significant potential to contribute. Web sites typically contain and integrate several bodies of data that are linked into a rich navigational structure. It is widely accepted today that declarative representations

---

[⋆] This work has been partially supported by the EU (FEDER) and Spanish MEC TIN-2004-7943-C04-02 project, the Generalitat Valenciana under grant GV06/285, and Integrated Action Hispano-Alemana HA2006-0007. Daniel Romero is also supported by ALFA grant LERNet AML/19.0902/97/0666/II-0472-FA.

are the best way to specify the structural aspects of Web sites as well as many forms of Web-site content. As an additional advantage, rule-based languages such as Maude [14] offer an extremely powerful, rewriting-based "reasoning engine" where the system transitions are represented/derived by rewrite rules indicating how a configuration is *transformed* into another.

In previous work on GVerdi [2, 6], we proposed a rewriting-based approach to Web-site verification and repair. In a nutshell, our methodology w.r.t. a given formal specification is applied to discover two classes of important, semantic flaws in Web sites. The first class consists of correctness errors (forbidden information that occurs in the Web site), while the second class consist of completeness errors (missing and/or incomplete Web pages). The original GVerdi system was developed in Haskell as a stand-alone application with wx-Haskell graphical interface. It allows the user to load a given Web site, together with a Web specification, and is able to recognize and fix the erroneous information appearing in the Web site. This is done by means of a novel rewriting-based technique, called *partial rewriting*, in which the traditional pattern matching mechanism is replaced by a suitable technique based on the *homeomorphic embedding* relation for recognizing patterns inside semistructured documents.

Web services are an emergent paradigm built upon XML as vehicle for exchanging messages across applications. Web services provide a standard means of interoperating between different software architectures, running on a variety of platforms and/or frameworks. Using the WSDL infrastructure for service description allows one to specify the properties, capabilities and behavior/semantics of a Web service. The new WebVerdiService, written in Java, can be seen as a service interface to Verdi-M, that is, an abstract boundary (independent of transmission protocol and data format) that offers any Internet requester entity the capabilities of Verdi-M. The new WebVerdiService relies on a strictly more powerful Web verification engine written in Maude [14] which automatically derives the error symptoms. Thanks to the AC pattern matching supported by Maude and its metalevel features, we have significantly improved both the performance and the usability of the original system. By using SOAP messages and other Web-related standards, a Java Web client that interacts with the Web verification service has also been made publicly available within the WebVerdi-M implementation.

Although there have been other recent efforts to apply formal techniques to Web site management [7, 10, 12, 16, 18, 20, 28], no work addressed the semantic verification of Web sites before. The key idea behind WebVerdi-M is that rule-based techniques can support in a natural way not only intuitive, high level Web site specification, but also efficient Web site verification techniques. As far as we know, rewriting-based techniques have not been explored in the context of Web site verification to date. Previous rewriting-based approaches for Web site processing focus on transformation rather than verification issues, e.g. [24, 8]. Our rule specification language does offer the expressiveness and computational power of functions and is simpler than formalizations of XML schemata based on tree automata often used in the literature such as, e.g. the regular expression types [23].

VeriWeb [28] explores interactive, dynamic Web sites using a special browser that systematically explores all paths up to a specified depth. The user first specifies some properties by means of *SmartProfiles*, and then the verifier traverses the considered Web site to report the errors as sequences of Web operations that lead to a page which violates a property. Navigation errors and page errors can be signaled, but tests are performed only at the http-level. In [20], a declarative verification algorithm is developed which checks a particular class of integrity constraints concerning the Web site's structure, but not the contents of a given instance of the site. In [16], a methodology to verify some semantic constraints concerning the Web site contents is proposed, which consists of using inference rules and axioms of natural semantics. The framework XLINKIT [18, 29] allows one to check the consistency of distributed, heterogeneous documents as well as to fix the (possibly) inconsistent information. The specification language is a restricted form of first order logic combined with Xpath expressions [33] where no functions are allowed.

The paper is organized as follows. Section 2 presents some preliminaries, and in 3 we briefly recall the rewriting-based, Web-site verification technique of [2]. In Section 4, we discuss the efficient implementation in Maude (by means of AC pattern matching) of one of the key ingredients of our verification engine: the *homeomorphic embedding* relation, which we use to recognize patterns within semi-structured documents. Section 5 briefly describes the service-oriented architecture of our verification prototype WebVerdi-M. In Section 6, we present an experimental evaluation of the system on a set of benchmarks. Finally, Section 7 presents our conclusions.

2 Preliminaries

By $\mathcal{V}$ we denote a countably infinite set of variables and Σ denotes a set of *function symbols* (also called *operators*), or *signature*. We consider varyadic signatures as in [15] (i.e., signatures in which symbols do not have a fixed arity, that is, they may be given an arbitrary number of arguments).

$\tau(\Sigma, \mathcal{V})$ and $\tau(\Sigma)$ denote the *non-ground term algebra* and the *term algebra* built on $\Sigma \cup \mathcal{V}$ and Σ, respectively. Terms are viewed as labelled trees in the usual way. Positions are represented by sequences of natural numbers denoting an access path in a term. The empty sequence Λ denotes the root position. Given $S \subseteq \Sigma \cup \mathcal{V}$, $O_S(t)$ denotes the set of positions of a term t that are rooted by symbols in S. $t_{|u}$ is the subterm at the position u of t. $t[r]_u$ is the term t with the subterm rooted at the position u replaced by r. Given a term t, we say that t is *ground*, if no variable occurs in t. A *substitution* $\sigma \equiv \{X_1/t_1, X_2/t_2, \ldots\}$ is a mapping from the set of variables $\mathcal{V}$ into the set of terms $\tau(\Sigma, \mathcal{V})$ satisfying the following conditions: (i) $X_i \neq X_j$, whenever $i \neq j$, (ii) $X_i\sigma = t_i$, $i = 1, ..n$, and (iii) $X\sigma = X$, for all $X \in \mathcal{V} \setminus \{X_1, \ldots, X_n\}$. By ε we denote the *empty* substitution. An *instance* of a term t is defined as $t\sigma$, where σ is a substitution.

By $Var(s)$ we denote the set of variables occurring in the syntactic object s. Syntactic equality between objects is represented by $\equiv$.

3 Rule-based Web site verification

In this section, we briefly recall the formal verification methodology proposed in [2], which allows us to detect forbidden/erroneous information as well as missing information in a Web site. By executing a Web specification on a given Web site, we are able to recognize and exactly locate the source of a possible discrepancy between the Web site and the properties required in the Web specification. An efficient and elegant implementation in Maude of such a methodology is described in Section 4.

3.1 Denotation of Web sites

In this framework, we assume a Web page to be a well-formed *XML document* [32], since there are plenty of programs and online services that are able to validate XML syntax and perform link checking (e.g. [34],[31]). Since XML documents are provided with a tree-like structure, they can be straightforwardly encoded as ground Herbrand terms of a given term algebra $\tau(\mathcal{T}ext \cup \mathcal{T}ag)$, where $\mathcal{T}ext$ represents the set of all the finite strings over a given alphabet and $\mathcal{T}ag$ is a set of tag symbols [2]. Note that XML tag attributes can be considered as common tagged elements, and hence be translated analogously. Therefore, a *Web site* can be seen as a finite set of ground terms belonging to $\tau(\mathcal{T}ext \cup \mathcal{T}ag)$. In the following, we will also consider terms of the non-ground term algebra $\tau(\mathcal{T}ext \cup \mathcal{T}ag, \mathcal{V})$, which may contain variables. An element $s \in \tau(\mathcal{T}ext \cup \mathcal{T}ag, \mathcal{V})$ is called *XML document template*. In our methodology, Web page templates are used to specify properties of Web sites, as described in the following section.

3.2 The Web specification language

A Web specification is a triple (R, I_N, I_M), where R, I_N, and I_M are a finite set of rules. The set R contains the definition of some auxiliary functions which the user would like to provide, such as string processing, arithmetic, boolean operators, etc. R is formalized as a term rewriting system, which is handled by standard rewriting [25]. The rewriting mechanism executes function calls by simply reducing them to their irreducible form (that is, a term that cannot be rewritten further) The set I_N describes constraints for detecting erroneous Web pages (*correctNess rules*). A correctness rule has the following syntax:

$$l \rightharpoonup error \mid C$$

where l is a term, $error$ is a reserved constant, and C is a (possibly empty) finite sequence containing

- membership tests of the form $X \in rexp$ w.r.t. a given regular language $rexp$;[4]
- equations/inequalities over terms.

When C is empty, we simply write $l \rightharpoonup error$.

Informally, the meaning of a correctness rule $l \rightharpoonup error \mid C$ is the following. Whenever (i) a "piece" of a given Web page $p \in \tau(\mathcal{T}ext \cup \mathcal{T}ag)$ can be "recognized" to be an instance $l\sigma$ of l, and (ii) the corresponding instantiated condition $C\sigma$ holds, then Web page p is marked as an incorrect page.

The third set of rules I_M specifes some properties for discovering incomplete/missing Web pages (*coMpleteness rules*). A completeness rule is defined as

$$l \rightharpoonup r \ \langle \mathtt{q} \rangle$$

where l and r are terms and $\mathtt{q} \in \{\mathtt{E}, \mathtt{A}\}$. Completeness rules of a Web specification formalize the requirement that some information must be included in all or some pages of the Web site. We use attributes $\langle \mathtt{A} \rangle$ and $\langle \mathtt{E} \rangle$ to distinguish "universal" from "existential" rules, as explained below. Right-hand sides r of completeness rules can contain functions, which are defined in R. In addiction, some symbols in the right-hand sides of the rules may be marked by means of the symbol $\sharp$. Marking information of a given rule r is used to select the subset of the Web site in order to check the condition formalized by r.

Intuitively, the interpretation of a universal rule $l \rightharpoonup r \ \langle \mathtt{A} \rangle$ (respectively, an existential rule $l \rightharpoonup r \ \langle \mathtt{E} \rangle$) w.r.t. a Web site W is as follows: if (an instance of) l is "recognized" in W, (an instance of) the irreducible form of r must also be "recognized" in *all* (respectively, *some*) of the Web pages that embed (an instance of) the marked structure of r.

Example 1. Let R be a TRS, we define $X \mathbin{+\!\!+} Y$, which concatenates two strings; we also define $X + Y$, which sums two natural numbers. Let (R, I_N, I_M) be a Web specification where I_N and I_M are defined as follows:

$$\begin{aligned}
member(name(X), surname(Y)) &\rightharpoonup \sharp hpage(fullname(X \mathbin{+\!\!+} Y)),\\
&\ status)\ \langle E \rangle\\
hpage(status(professor)) &\rightharpoonup \sharp hpage(\sharp status(\sharp professor),\\
&\ teaching))\ \langle A \rangle\\
project(grant1(X), grant2(Y), total(Z)) &\rightharpoonup error \mid X + Y \neq Z,\\
&\ X\ in\ [0-9]*, Y\ in\ [0-9]*,\\
&\ Z\ in\ [0-9]*
\end{aligned}$$

This Web specification models some required properties for a Web site of a research group. The first rule formalizes the following property: if there is a Web page containing a member list, then for each member, a home page should exist which contains (at least) the full name and the status of this member. The full

---

[4] Regular languages are represented by means of the usual Unix-like regular expression syntax.

name is computed by applying the function $++$ to the name and the surname of the member. The marking information establishes that the property must be checked only on home pages (i.e., pages containing the tag "hpage"). The second rule states that, whenever a home page of a professor is recognized, that page must also include some teaching information. The rule is universal since it must hold for each professor home page (i.e., all the Web pages embedding the marked structure $home(status(professor))$. The third rule states that, for each research project, the total project budget must be equal to the sum of the funds, which have been granted for the first and the second research periods. The membership tests in the conditions ensure that the values assigned to the variables are natural numbers.

Diagnoses are carried out by running Web specifications on Web sites. The operational mechanism is based on a novel, flexible matching technique [2] that is able to "recognize" the partial structure of a term (Web template) within another and select it by computing *homeomorphic embeddings* (cf. [26]) of Web patterns within Web documents.

3.3 Homeomorphic embedding

Homeomorphic embedding relations allow us to verify whether a given XML document template is somehow "enclosed" within another one. Roughly speaking, we consider a simple embedding relation which closely resembles the notion of *simulation* [22], this relation has been widely used in a number of works about querying, transformation, and verification of semistructured data (cf. [11, 10, 1, 21, 9]).

We give a definition of homeomorphic embedding, $\trianglelefteq$, which is an adaptation of the one proposed in [26], where (i) a simpler treatment of the variables is considered, (ii) function symbols with variable arities are allowed, (iii) the relative positions of the arguments of terms are ignored (i.e. $f(a, b)$ is not distinguised from $f(b, a)$), and (iv) we ignore the usual *diving* rule[5] [26].

Definition 1 (homeomorphic embedding). *The* homeomorphic embedding *relation*

$$\trianglelefteq \subseteq \tau(\mathcal{T}ext \cup \mathcal{T}ag, \mathcal{V}) \times \tau(\mathcal{T}ext \cup \mathcal{T}ag, \mathcal{V})$$

on XML documents templates is the least relation satisfying the rules:

1. $X \trianglelefteq t$, *for all* $X \in \mathcal{V}$ *and* $t \in \tau(\mathcal{T}ext \cup \mathcal{T}ag, \mathcal{V})$.
2. $f(t_1, \ldots, t_m) \trianglelefteq g(s_1, \ldots, s_n)$ *iff* $f \equiv g$ *and* $t_i \trianglelefteq s_{\pi(i)}$, *for* $i = 1, \ldots, m$, *and some injective function* $\pi : \{1, \ldots, m\} \to \{1, \ldots, n\}$.

Whenever $s \trianglelefteq t$, we say that t *embeds* s (or s *is embedded* or "*recognized*" in t).

The intuition behind the above definition is that $s \trianglelefteq t$ iff s can be obtained from t by striking out certain parts, in other words, the structure of the template s appears within t, which likely as specific Web data terms.

Let us illustrate Definition 1 by means of a rather intuitive example.

---

[5] The diving rule allows one to "strike out" a part of the term at the right-hand side of the relation $\trianglelefteq$. Formally, $s \trianglelefteq f(t_1, \ldots\ t_n)$, if $s \trianglelefteq t_i$, for some i.

Example 2. Consider the following XML document templates (called s_1 and s_2, respectively):

$$hpage(surname(Y), status(professor), name(X), teaching)$$

$$\begin{aligned} hpage(\ & name(mario), surname(rossi), status(professor), \\ & teaching(course(logic1), course(logic2)) \\ & hobbies(hobby(reading), hobby(gardening))) \end{aligned}$$

Note that $s_1 \trianglelefteq s_2$, since the structure of s_1 can be recognized inside the structure of s_2, while $s_2 \ntrianglelefteq s_1$.

It is important to have an efficient implementation of *homeomorphic embedding* because it is used repeatedly during the verification process as described in the following.

3.4 Web verification methodology

Roughly speaking, the verification methodology works as follows. First, by using the homeomorphic embedding relation of Definition 1, we check whether the left-hand side l of some Web specification rule is embedded into a given page p of the considered Web site. When the embedding test $l \trianglelefteq p$ succeeds, by extending the proof, we construct the biggest substitution[6] σ for the variables in $Var(l)$, such that $l\sigma \trianglelefteq p$. Then, depending on the nature of the Web specification rule (correction or completeness rule), it is as follows:

(Correction rule) evaluating the condition of the rule (instantiated by σ); a correctness error is signalled in the case when the error condition is fulfilled.

(Completeness rule) by a new homeomorphic embedding test, checking whether the right-hand side of the rule (instantiated by σ) is recognized in some page of the considered Web site. Otherwise, a completeness error is signalled. Moreover, from the incompleteness symptom computed so far, a fixpoint computation is started in order to discover further missing information, which may involve the execution of other completeness rules.

4 Verifying Web sites using Maude

Maude is a high-performance reflective language supporting both equational and rewriting logic programming, which is particularly suitable for developing domain-specific applications [30, 17]. As a matter of fact, it has a *clear* and *well-understood* semantics, which eases the development of complex systems. On the other hand, it is *expressive* enough to implement a wide range of applications, ranging from deterministic systems to highly concurrent ones. In addiction, the

---

[6] The substitution σ is easily obtained by composing the bindings X/t, which can be recursively gathered during the *homeomorphic embedding* test $X \trianglelefteq t$, for $X \in l$ and $t \in p$.

Maude language is not only intended for system prototyping, but it has to be considered as a real programming language with competitive performance. In the rest of the section, we recall some of the most important features of the Maude language which we have conveniently exploited for the optimized implementation of our Web site verification engine.

Equational attributes. A Maude program consists of functional *modules*, which define a typed signature Σ, a set of typed variables, and a set of equations implementing the functions in the signature Σ. Equational attributes are a means of declaring certain kinds of equational axioms for a particular binary operator of Σ which Maude uses efficiently in a built-in way. Some of the attributes supported are: `assoc` (associativity), `comm` (commutativity), `id:<identity name>`. Employing equational attributes not only avoids termination problems and leads to much more efficient term evaluation, but it also us allows to define more "flexible" data structures. As an example, let us describe how we model (part of) the internal representation of XML documents in our system.

The chosen representation slightly modifies the data structure provided by the Haskell HXML Library [19] by adding commutativity to the standard XML tree-like data representation. In other words, in our setting, the order of the children of a tree node is not relevant: e.g., $f(a, b)$ is "equivalent" to $f(b, a)$.

```
fmod TREE-XML is
sort XMLNode .
op RTNode : -> XMLNode .                    -- Root (doc) information item
op ELNode _ _ : String AttList -> XMLNode . -- Element information item
op TXNode _ : String -> XMLNode .           -- Seq. of char information items
--- ...  definitions of the other XMLNode types omitted ...
sorts XMLTreeList XMLTreeSeq XMLTree .
op Tree (_) _ : XMLNode XMLTreeList - > XMLTree .
subsort XMLTree < XMLTreeSeq .
op _,_ : XMLTreeSeq XMLTreeSeq -> XMLTreeSeq    [comm assoc id:null] .
op null : -> XMLTreeSeq .
op [_] : XMLTreeSeq -> XMLTreeList .
op [] : -> XMLTreeList .
endfm
```

In the previous module, the XMLTreeSeq constructor `_,_` is given the equational attributes `comm assoc id:null`, which allow us to get rid of parentheses and disregard the ordering among `XML` nodes within the list. The significance of this optimization will be clear when we consider rewriting XML trees with AC pattern matching.

AC pattern matching. The evaluation mechanism of Maude is based on rewriting modulo an equational theory E (i.e. a set of equational axioms), which is accomplished by performing *pattern matching modulo* the equational theory E. More precisely, given an equational theory E, a term t and a term u, we say that t *matches* u *modulo* E (or that t *E-matches* u) if there is a substitution σ

such that $\sigma(t) =_E u$, that is, $\sigma(t)$ and u are equal modulo the equational theory E. When E contains axioms for associativity and commutativity of operators, we talk about *AC pattern matching.* AC pattern matching is a powerful matching mechanism, which we employ to inspect and extract the partial structure of a term. That is, we use it directly to implement the notion of homeomorphic embedding of Definition 1. Let us see an example.

Example 3. Let us define an associative and commutative binary infix operator $\star$ along with the constants a, b, c, d. Then $a \star d$ AC-matches (a part of) the term $t \equiv a \star b \star c \star d$ since t is equivalent — modulo associativity and commutativity— to the term $t' \equiv (a \star d) \star (b \star c)$, and, hence, $a \star d$ trivially matches the first subterm of t'. Thus, we are able to recognize the structure $a \star d$ as a substructure of the term $a \star b \star c \star d$.

Metaprogramming. Maude is based on rewriting logic [27], which is reflective in a precise mathematical way. In other words, there is a finitely presented rewrite theory $\mathcal{U}$ that is universal in the sense that we can represent in $\mathcal{U}$ (as a data) any finitely presented rewrite theory $\mathcal{R}$ (including $\mathcal{U}$ itself), and then mimick in $\mathcal{U}$ the behavior of $\mathcal{R}$. Roughly speaking, there exists a universal Maude program that is able to "reproduce" the computations of any other Maude program (including itself). This leads to novel metaprogramming capabilities that can greatly increase software reusability and adaptability, which have been exploited in our context to implement the semantics of correctness as well as completeness rules (e.g. implementing the homeomorphic embedding algorithm, evaluating conditions of conditional rules, etc.). Namely, during the partial rewriting process, functional modules are dynamically created and run by using the meta-reduction facilities of the language.

Now we are ready to explain how we implemented the homeomorphic embedding relation of Section 3.3, by exploiting the aforementioned Maude high-level features.

Homeomorphic embedding implementation Let us consider two XML document templates l and p. As mentioned above, the critical point of our methodology is to (i) discover whether $l \trianglelefteq p$ (i.e. l is embedded into p); (ii) find the substitution σ such that $l\sigma$ is the instance of l recognized inside p, whenever $l \trianglelefteq p$.

Given l and p, our proposed solution can be summarized as follows. By using Maude metalevel feactures, we first dynamically build a module `M` that contains a single rule of the form

$$\mathtt{eq}\ \mathtt{l} = \mathtt{sub}(\text{"}\mathtt{X_1}\text{"}/\mathtt{X_1}), \ldots, \mathtt{sub}(\text{"}\mathtt{X_n}\text{"}/\mathtt{X_n}), \qquad \mathtt{X_i} \in \mathtt{Var}(\mathtt{l}), \mathtt{i} = \mathtt{1}, \ldots \mathtt{n},$$

where `sub` is an associative operator used to record the substitution σ that we want to compute. Next, we try to reduce the XML template p by using such a rule. Since l and p are internally represented by means of the binary constructor `_,_` that is given the equational attributes `comm assoc id:null` (see

Section 4), the execution of module M on p essentially boils down to computing an AC-matcher between l and p. Moreover, since AC pattern matching directly implements the homeomorphic embedding relation as illustrated in Example 3, the execution of M corresponds to finding all the homeomorphic embeddings of l into p (recall that the set of AC matchers of two compatible terms is not generally a singleton). Additionally, as a side effect of the execution of M, we obtain the computed substitution σ for free as the sequence of bindings for the variables $\mathtt{X_i}$, $\mathtt{i} = \mathtt{1}, \ldots, \mathtt{n}$ which occur in the instantiated rhs

$$\mathtt{sub}(''\mathtt{X_1}''/\mathtt{X_1})\sigma, \ldots, \mathtt{sub}(''\mathtt{X_n}''/\mathtt{X_n})\sigma, \qquad \mathtt{X_i} \in \mathtt{Var(l)}, \mathtt{i} = \mathtt{1}, \ldots \mathtt{n},$$

of the dynamic rule after the partial rewriting step.

Example 4. Consider again the XML document templates s_1 and s_2 of Example 2. We build the dynamic module M containing the rule

$$\mathtt{op\ hpage(surname(Y), status(professor), name(X), teaching)} = \mathtt{sub(''Y''/Y), sub(''X''/X)}\ .$$

Since $s_1 \trianglelefteq s_2$, there exists an AC-match between s_1 and s_2 and, hence, the result of executing M against the (ground) XML document template s_2 is the computed substitution: $\mathtt{sub(''Y''/rossi), sub(''X''/mario)}$.

5 Prototype implementation

The verification methodology presented so far has been implemented in the prototype WebVerdi-M (Web Verification and Rewriting for Debugging Internet sites with Maude). In developing and deploying the system, we fixed the following requirements: 1) define a system architecture as simple as possible, 2) make the Web verification service available to every Internet requestor, and 3) hide the technical details from the user.

In order to fulfill the above requirements, we developed the Web verification system WebVerdi-M as a Web service. A Web service is a software system identified by a URI, whose public interfaces and bindings are defined and described using XML (specifically WSDL). A Web service enhances interoperability among software applications. A Web service is also self-describing and can be published, located, and invoked across the Web. Hence, Web services facilitate the development of distributed applications by adopting a loosely coupled Web programming model. Systems developed in terms of Web services are language independent and platform independent. Additionally, they are easily scalable and extensible by establishing connections to new Web services when necessary.

5.1 WebVerdi-M Architecture

WebVerdi-M is a service-oriented architecture that allows one to access the core verification engine Verdi-M as a reusable entity.

WebVerdi-M can be divided into two layers: front-end and back-end. The back-end layer provides web services to support the front-end layer. This architecture allow clients on the network to invoke the Web service functionality through the available interfaces.

The tool consists of the following components: Web service WebVerdiService, Web client WebVerdiClient, core engine Verdi-M, XML API, and database DB.

Figure 1 illustrates the overall architecture of the system. For the reader interested in more detail, the types of messages and the specific message exchange patterns that are involved in interacting with WebVerdi-M can be found in [3].

Fig. 1. Components of WebVerdi-M

WebVerdiService. Our web service exports six operations that are network-accessible through standardized XML messaging. These operations are: store a Web site, remove a Web site, retrieve a Web site, add Web page to a Web site, check correctness, and check completeness. The Web service acts as a single access point to the core engine Verdi-M which implements our Web verification methodology in in Maude. Following the standards, the architecture is also platform and language independent so as to be accessible via scripting environment as well as via client applications across multiple platforms.

XML API. In order for successful communications to occur, both the WebVerdiService and WebVerdiClient (or any user) must agree to a common format for the messages being delivered so that they can be properly interpreted at each end. The WebVerdiService Web service is developed by defining an API that encompasses the executable library of the core engine. This is achieved by making use of Oracle JDeveloper, including the generation of WSDL for making the API available. The OC4J Server (the web server integrated in Oracle JDeveloper) handles all procedures common to Web service development. Synthesized error symptoms are also encoded as XML documents in order to be transferred from the WebVerdiService Web service to client applications as an XML response by means of the SOAP protocol.

Verdi-M. Verdi-M is the most important part of the tool. Here is where the verification methodology is implemented (see Section 4). This component is implemented in Maude language and is independent of the other system components.

WebVerdiClient. WebVerdiClient is a Web client that interacts with the Web service to use the capabilities of Verdi-M. Our main goal was to provide an *intuitive* and *friendly* interface for the user. WebVerdiClient is provided with a versatile, new graphical interface that offers three complementary views for both the specification rules and the pages of the considered Web site: the first one is based on the typical idea of accessing contents by using folders trees and is particularly useful for beginners; the second one is based on XML, and the third one is based on term algebra syntax. The tool provides all translations among the three views. A snapshot of WebVerdiClient is shown in Figure 2.

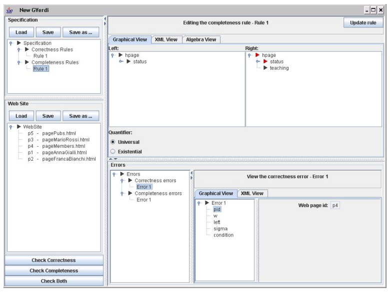

Fig. 2. WebVerdiClient Snapshot

DB. The WebVerdiService Web service needs to transmit abundant XML data over the Web to and from client applications. The common behavior of a user when using the tool is to modify the default rules provided for the Web specification and then verify a particular Web site. After modifying the Web specification, it would be necessary to send back to the service the considered specification as well as the whole Web site to verify. After the application invokes the WebVerdiService Web service with these two elements, synthesized errors are progressively generated and transferred to the client application. The standard Web service architecture requires client applications to wait until all data are received and then errors are sent, which could cause significant time lags in the application. In order to avoid this overhead and to provide better performance to the user, we use a local *MySql* data base where the Web site and Web errors are temporarily stored at the server side.

6 Experimental evaluation

In order to evaluate the usefulness of our approach in a realistic scenario (that is, for sites whose data volume exceeds toy sizes), we have benchmarked our system by using several correctness as well as completeness rules of different complexity for a number of XML documents randomly generated by using the XML documents generator xmlgen available within the XMark project [13]. The tool xmlgen is able to produce a set of XML data, each of which is intended to challenge a particular primitive of XML processors or storage engines by using different scale factors.

Table 1 shows some of the results we obtained for the simulation of three different Web specifications $WS1$, $WS2$ and $WS3$ in five different, randomly generated XML documents. Specifically, we tuned the generator for scaling factors from 0.01 to 0.1 to match an XML document whose size ranges from 1Mb –corresponding to an XML tree of about 31000 nodes– to 10Mb –corresponding to an XML tree of about 302000 nodes– (for an exhaustive evaluation, please refer to `http://www.dsic.upv.es/users/elp/WebVerdi`).

Both Web specifications $WS1$ and $WS2$ aim at checking the verification power of our tool regarding data correctness, and thus include only correctness rules. The specification rules of $WS2$ contain more complex and more demanding constraints than the ones fomalized in $WS1$, with involved error patterns to match, and conditional rules with a number of membership tests and functions evaluation. The Web specification $WS3$ aims at checking the completeness of the randomly generated XML documents. In this case, some critical completeness rules have been formalized which recognize a significant amount of missing information.

The figures shown in Table 1 were obtained on a personal computer equipped with 1Gb of RAM memory, 40Gb hard disk and a Pentium Centrino CPU clocked at 1.75 GHz running Ubuntu Linux 5.10.

Let us briefly comment our results. Regarding the verification of correctness, the implementation is extremely time efficient, with elapsed times scaling lineary. Table 1 shows that the execution times are small even for very large documents (e.g. running the correctness rules of Web specification $WS1$ over a 10Mb XML document with 302000 nodes takes less than 13 seconds). Concerning the completeness verification, the fixpoint computation which is involved in the evaluation of the completeness rules typically burdens the expected performance (see [2]), and we are currently able to process efficiently XML documents whose size is not bigger than 1Mb (running the completeness rules of Web specification $WS3$ over a 1Mb XML document with 31000 nodes takes less than 3 minutes).

Scale factor	**Size**	**Nodes**	**Time**		
			$WS1$	$WS2$	$WS3$
0.01	1 Mb	30,985	0.930 s	0.969 s	165.578 s
0.03	3 Mb	90,528	12.604 s	2.842 s	1768.747 s
0.05	5 Mb	150,528	5.975 s	5.949 s	4712.157 s
0.08	8 Mb	241,824	8.608 s	9.422 s	12503.454 s
0.10	10 Mb	301,656	12.458 s	12.642 s	21208.494 s

Table 1. Verdi-M Benchmarks

Finally, we want to point out that the current Maude implementation of the verification system supersedes and greatly improves our preliminary system, called GVerdi[2, 6], which was not even able to manage correctness for XML document repositories larger than 1Mb within a reasonable time. We are currently working on further improving the performance of our system.

7 Conclusion

In the literature on Web management, Web sites verification has mainly a syntactic focus with a particular concern for the accessibility and usability perspective [4, 5]. This paper can be seen as a step forward towards the formal, semantic verification of Web sites using rule-based technology. First we present an efficient and innovative implementation in Maude –a high-performance reflective functional language– of the rewriting-based, Web verification methodology of [2]. This methodology deals with semantic flaws that are not addressed by classical tools. The framework comes with a language for defining correctness and completeness conditions on Web sites. Then, our rewriting-based verification technique is able to recognize forbidden/incorrect patterns and incomplete/missing Web pages by means of a novel rewriting-based technique, called *partial rewriting*.

In this work, first we exploit Maude's capabilities which are particularly suitable for our implementation, such as associative commutative pattern matching and metaprogramming. We can thus provide WebVerdi-M with a powerful Web verification engine. We have done a comparison of run times of Verdi-M core engine and shown the resulting impressive performance (e.g. less than a second for evaluating a tree of some 30,000 nodes).

Then, we have proposed a service-oriented architecture which makes the Web verification capabilities of the system easily accessible to internet requestors. The resulting prototype WebVerdi-M is publicly available together with a set of examples and its XML API.

In order to make possible technological transfer to industry it is necessary to have tools that are able to give prompt answers on real size examples, as we have shown by our scalable benchmarks. Another important factor, is to reduce the cost of learning to the user. For this reason we have developed a friendly innovative interface for our system.

References

1. S. Abiteboul, P. Buneman, and D. Suciu. *Data on the Web. From Relations to Semistructured Data and XML.* Morgan Kaufmann, 2000.
2. M. Alpuente, D. Ballis, and M. Falaschi. Automated Verification of Web Sites Using Partial Rewriting. *Software Tools for Technology Transfer*, 8:565–585, 2006.
3. M. Alpuente, D. Ballis, M. Falaschi, P. Ojeda, and D. Romero. The Web Verification Service WebVerdi-M. Technical report, DSIC-UPV, 2007.
4. M. Alpuente, S. Escobar, and M. Falaschi (Eds.). *Automated Specification and Verification of Web Sites, 1st Int'l Workshop WWV'05*, volume 157(2). Elsevier, 2006.

5. M. Alpuente, S. Escobar, and M. Falaschi (Eds.). *Automated Specification and Verification of Web Systems, 2nd Int'l Workshop WWV'06*. IEEE Computer Society Press, 2007.
6. D. Ballis and J. García Vivó. A Rule-based System for Web Site Verification. In *Proc. of 1st Int'l Workshop on Automated Specification and Verification of Web Sites (WWV'05)*, volume 157(2). ENTCS, Elsevier, 2005.
7. L. Baresi, P. Fraternali, M. Tisi, and S. Morasca. Towards model-driven testing of a web application generator. In *Proc. of Web Engineering, 5th Int'lernational Conference, ICWE 2005*, volume 3579 of *Lecture Notes in Computer Science*, pages 75–86, 2005.
8. I. D. Baxter, F. Ricca, and P. Tonella. Web Application Transformations based on Rewrite Rules. *Information and Software Technology*, 44(13), 2002.
9. E. Bertino, M. Mesiti, and G. Guerrin. A Matching Algorithm for Measuring the Structural Similarity between an XML Document and a DTD and its Applications. *Information Systems*, 29(1):23–46, 2004.
10. F. Bry and S. Schaffert. Towards a Declarative Query and Transformation Language for XML and Semistructured Data: Simulation Unification. In *Proc. of the Int'l Conference on Logic Programming (ICLP'02)*, volume 2401 of *Lecture Notes in Computer Science*. Springer-Verlag, 2002.
11. F. Bry and S. Schaffert. The XML Query Language Xcerpt: Design Principles, Examples, and Semantics. Technical report, 2002. Available at: `http://www.xcerpt.org`.
12. J. Cavalcanti and W. Vasconcelos. A logic-based approach for automatic synthesis and maintenance of web sites. In *Proc. of the 14th Int'l Conference on Software Engineering and Knowledge Engineering*, pages 619–626. ACM press, 2002.
13. Centrum voor Wiskunde en Informatica. XMark – an XML Benchmark Project, 2001. Available at: `http://monetdb.cwi.nl/xml/`.
14. Manuel Clavel, Francisco Durán, Steven Eker, Patrick Lincoln, Narciso Martí-Oliet, José Meseguer, and Carolyn Talcott. The maude 2.0 system. In Robert Nieuwenhuis, editor, *Rewriting Techniques and Applications (RTA 2003)*, number 2706 in Lecture Notes in Computer Science, pages 76–87. Springer-Verlag, 2003.
15. N. Dershowitz and D. Plaisted. Rewriting. *Handbook of Automated Reasoning*, 1:535–610, 2001.
16. T. Despeyroux and B. Trousse. Semantic Verification of Web Sites Using Natural Semantics. In *Proc. of 6th Conference on Content-Based Multimedia Information Access (RIAO'00)*, 2000.
17. S. Eker, J. Meseguer, and A. Sridharanarayanan. The Maude LTL model checker and its implementation. In *Model Checking Software: Proc. 10 th Intl. SPIN Workshop*, volume 2648 of *LNCS*, pages 230–234. Springer, 2003.
18. E. Ellmer, W. Emmerich, A. Finkelstein, and C. Nentwich. Flexible Consistency Checking. *ACM Transaction on Software Engineering*, 12(1):28–63, 2003.
19. Joe English. The HXML Haskell Library, 2002. Available at: `http://www.flightlab.com/ joe/hxml/`.
20. M. Fernandez, D. Florescu, A. Levy, and D. Suciu. Verifying Integrity Constraints on Web Sites. In *Proc. of Sixteenth International Joint Conference on Artificial Intelligence (IJCAI'99)*, volume 2, pages 614–619. Morgan Kaufmann, 1999.
21. M. F. Fernandez and D. Suciu. Optimizing Regular Path Expressions Using Graph Schemas. In *Proc. of Int'l Conference on Data Engineering (ICDE'98)*, pages 14–23, 1998.

22. M. R. Henzinger, T. A. Henzinger, and P. W. Kopke. Computing Simulations on Finite and Infinite Graphs. In *IEEE Symposium on Foundations of Computer Science*, pages 453–462, 1995.
23. H. Hosoya and B. Pierce. Regular Expressions Pattern Matching for XML. In *Proc. of 25th ACM SIGPLAN-SIGACT Int'l Symp. POPL*, pages 67–80. ACM, 2001.
24. C. Kirchner, Z. Qian, P. K. Singh, and J. Stuber. Xemantics: a Rewriting Calculus-Based Semantics of XSLT. Rapport de recherche A01-R-386, LORIA, 2001.
25. J.W. Klop. Term Rewriting Systems. In S. Abramsky, D. Gabbay, and T. Maibaum, editors, *Handbook of Logic in Computer Science*, volume I, pages 1–112. Oxford University Press, 1992.
26. M. Leuschel. Homeomorphic Embedding for Online Termination of Symbolic Methods. In T. Æ. Mogensen, D. A. Schmidt, and I. H. Sudborough, editors, *The Essence of Computation*, volume 2566 of *Lecture Notes in Computer Science*, pages 379–403. Springer, 2002.
27. N. Martí-Oliet and J. Meseguer. Rewriting Logic: Roadmap and Bibliography. *Theoretical Computer Science*, 285(2):121–154, 2002.
28. B. Michael, F. Juliana, and G. Patrice. Veriweb: automatically testing dynamic web sites. In *Proc. of 11th Int'l WWW Conference*. ENTCS, Elsevier, 2002.
29. C. Nentwich, W. Emmerich, and A. Finkelstein. Consistency Management with Repair Actions. In *Proc. of the 25th International Conference on Software Engineering (ICSE'03)*. IEEE Computer Society, 2003.
30. J. Meseguer S. Escobar, C. Meadows. A Rewriting-Based Inference System for the NRL Protocol Analyzer and its Meta-Logical Properties. *Theoretical Computer Science*, 367(1-2):162–202, 2006.
31. Typke und Wicke GbR. Validate/Check XML. Available at: `http://www.xmlvalidation.com/`.
32. World Wide Web Consortium (W3C). Extensible Markup Language (XML) 1.0, second edition, 1999. Available at: `http://www.w3.org`.
33. World Wide Web Consortium (W3C). XML Path Language (XPath), 1999. Available at: `http://www.w3.org`.
34. World Wide Web Consortium (W3C). Markup Validation Service, 2005. Available at: `http://validator.w3.org/`.