Towards Web Sites' Analysis Based On Coalgebras*

Daniela Cancila, Furio Honsell, Marina Lenisa

Dipartimento di Matematica e Informatica, Università di Udine
Via delle Scienze 206, 33100 Udine, Italy.
tel. +39 0432 558455, fax: +39 0432 558499
e-mail: `cancila,honsell,lenisa@dimi.uniud.it`.

The social importance of the *World Wide Web* has increased enormously in recent years. *Web usability* issues have therefore become crucial in site development, and the design of a web site is of vital importance from business perspective. In this regard, the Forrester Research report [MMS98] clearly points out that: *"40 % of the users do not return to a site when their first visit resulted in a negative experience"*. In particular, it is worth stating the obvious: if the customer can't find a product, then she will not buy it. The main consequence is a huge financial loss for the site. Moreover, the user has a powerful tool: the click of mouse. In fact, *"she who clicks the mouse gets to decide everything. It is so easy to go elsewhere; all the competitors in the world are but a mouseclick away"* [Nie00].

In this abstract, we present a *coalgebraic interpretation* [JR97] of web interactions, which can formally express various web design notions and properties, such as *user behavior*, *locality*, *reachability*. In [Saz01], Sazonov presented an abstract hyperset model of the web. His aim is to provide an abstract, "static" set-theoretic view of the web in order to build a query language for Web-like Data Bases. Coalgebraically, this amounts to modeling the web as a coalgebra for the (hyper)set functor $\mathcal{P}_f(\mathcal{L} \times \_)$, where $\mathcal{P}_f$ denotes the finite powerset functor and $\mathcal{L}$ is a set of labels for links. In this model, pages are modeled by their labeled links to other pages, and they are equated up-to *bisimilarity*. E.g. all pages with no further links are taken to be equal. This approach appears a little too rigid. Our coalgebraic description is more expressive in various respects. First of all, we give a coalgebraic account of single web sites, whereby pages are modeled by their links together with some extra information. This information can be either related to the contents, or to the architecture or to the presentation of the page, according to which aspect of page/site design we want to focus on. The granularity of this information is not fixed, but, e.g. in the case of contents, it can range from the entire text to a subset of keywords. This provides a formal framework for analysing the design of single sites, and for comparing different sites. We give a coalgebraic account of user *visits* to a web site, in terms of the notion of *relator simulation*. This gives us a formal tool for studying user behaviour, and hence for analysing the usability of a site, which fruitfully complements the more standard, but subjective methods of *heuristic evaluation* and *user testing*. Recently Joseph Goguen has proposed an algebraic method based on semiotics [dS83] for the study of web design issues [Gog99]. Unlike our approach, Goguen analyses other features. Nervertheless it is relevant the use of (co)algebraic tools as a methodology for the analysis of web issues. Actually we are testing our model on web sites. The aim is to study various aspects of a site (under different perspectives, i.e. content, architecture or presentation) for capturing of general properties (e.g. the coherence).

Here is a more detailed description of our model:
• A web page p is represented by a (possibly circular) pair $\langle a, \{l_1 : p_1, \ldots, l_n : p_n\}\rangle$, where $l_i(\in \mathcal{L})$ is a link in p and p_i is the page linked by l_i, and $a \in A$ is a contents (architecture,

* Work partially supported by MIUR Project CoMeta.

presentation) related information on the page p. In this page representation, we allow for more than one link from one page to another. Howevever, a good design principle is that there is at most one link from one page to another. Therefore, in our model, we assume that there is at most one link from one page to another. This allows us simplify the above presentation of a web page p, i.e. $p = \langle a, \{p_1, \ldots, p_n\}\rangle$. Our assumption is justified also by the way in which user visits are recorded in the "log file" (see below), where links are missing. A (possibly dynamic) web site S, i.e. a set of web pages closed under the relation induced by the links, is represented as a *coalgebra* for the (hyper)set functor $F_A = A \times \mathcal{P}_f(X)$. For a set of pages S representing a site, the coalgebra strucure $f : S \to F_A S$ yields the identifier of the current page (state) and the pages reached via the links from the current page (the next states), i.e.: $f(p) = \langle a, \{p_1, \ldots, p_n\}\rangle$. Different choices for the granularity of the information encoded by the set A give rise to different functors related by a natural transformation, $\eta : F_A \dot{\to} F_{A'}$, representing the *abstraction procedure* from one representation to a coarser one. In particular, if (S, f) is the F_A-coalgebra representation of a site, then $(S, \eta_S \circ f)$ is the $F_{A'}$-coalgebra representation given by $F_{A'}$. Correspondingly, the bisimilarity equivalences are η-related.

• The notion of F_A-*coalgebraic bisimilarity* induces an equivalence $\sim_A$, which allows us to highlight the structure of a given web site, according to various features, and detect contents (presentation, architecture) related pages or possibly useless duplications of pages. Formally, the *unique morphism* from a web site (S, f) into the *final* F_A-*coalgebra* gives us the *minimal* (non redundant) web site equivalent to S w.r.t. the chosen page representation.

• In order to model a visit of a site (S, f), we use the information in the "log file" of S, i.e. the file in which the user activity w.r.t. that site is registered. In particular, in the log file, for each page requested by a user, we find the user identifier (client IP), the request time, the address of the required page, and the address of the current page. We consider all the requests with the same user identifier to constitute a single visit. Moreover, we assume each client to have a personal (unlimited) "cash memory". This allows users to avoid to reload from the site a page requested for the second time (either via a normal link or via a "back/forward" link). As a consequence, for each user, in the "log file" only the first visit to each site page is registered. Therefore, by abstracting from possible multiple requests of the same page, a visit can be realistically represented as a *tree* which is a subgraph of the site. Formally: a visit of a site (S, f) is a pair $\langle p_0, (V, g)\rangle$, where $p_0 \in V$ is the *entry page* and (V, g) is an F_A-coalgebra (V, g) such that

– $\forall p \in V$, there exists a *unique* sequence $\langle p_0, \ldots, p_n\rangle$ such that $p_n = p$ and, for all $0 \leq i < n$, $\pi_2 \circ g(p_i) = \{p_{i+1}\}$.

– there is a *functional injective* Γ_A-*simulation* R between (V, g) and (S, f), i.e. R is a functional injective relation on $V \times S$ such that, for all $v \in V$ and $p \in S$, if vRs, then $f(v)\Gamma_A(R)g(s)$, where Γ_A is a suitable *weak relator* (i.e. a functor on *Rel*, up-to identity preservation), extending the functor F_A.

The bisimilarity on visits defines *user behaviours*. The locality of a visit, i.e. the set of pages of the site visited, can be recovered by taking $\pi_2\circ \leq_A$, where $\leq_A$ denotes the greatest Γ_A-similarity relation. Finer models for visits can be obtained by exploiting also the time information in the log file. E.g. the time interval between the visit of successors pages can be taken as an approximation of the time spent by the user for "completely understanding" each single page. The functor used in this case is $F_{A\times\Delta}$, where Δ encodes positive time intervals. A Δ-*enriched visit* is then defind as a pair $\langle p_0, (V, g)\rangle$, where (V, g) is an $F_{A\times\Delta}$-coalgebra, $\langle p_0, (V, \eta_V \circ g)\rangle$ is a standard visit, and $\eta : F_{A\times\Delta} \dot{\to} F_A$ is the abstraction natural transformation. The bisimilarity equivalence on this model allows us study user behaviour also on the base of the time spent for a visit, and hence to judge on the usability of pages/sites.

This approach is promising especially in the direction of applying bisimilarity techniques and coalgebraic logics for studying usability and defining formal properties of web/sites.

We wish to thank Giorgio Brajnik for his constructive comments.

References

[CH02] D. Cancila and F. Honsell. A coalgebraic description of web interactions. Technical Report 0207, Institut für Informatik, LMU München, 2002.

[dS83] F. de Saussure. *Course in General Linguistics*. London: Duckworth, [1916] 1983. trans. Roy Harris.

[Gog99] J. Goguen. An Introduction to Algebraic Semiotics, with Applications to User Interface Design. *Lecture Notes in Artificial Intelligence*, 1562:242–291, 1999.

[JR97] B. Jacobs and J. Rutten. A Tutorial on (Co)Algebras and (Co)Induction. Bulletin, 1997. n.62, p.222-259.

[MMS98] H. Manning, J.C. McCarthy, and R.K. Souza. Why Most Web Sites Fail. *Interactive Technology Series, Forrester Research*, vol. 3(7), September 1998.

[Nie00] J. Nielsen. *Designing Web Usability*. New Riders, 2000.

[Saz01] V. Sazonov. Using Agents for Concurrent Querying of Web-like Databases via a Hyper-Set-Theoretic Approach,. In *PSI'01, Fourth Int. Conf. Perspective of System Informatics*, Novosibirsk, Akademgorodok, Russia, 2001.