A TWO LAYERED MODEL FOR EVOLVING WEB RESOURCES

Alfredo Milani and Silvia Suriani

Abstract: In this paper the key features of a two-layered model for describing the semantic of dynamical web resources are introduced.

In the current Semantic Web proposal [Berners-Lee et al., 2001] web resources are classified into static ontologies which describes the semantic network of their inter-relationships [Kalianpur, 2001][Handschuh & Staab, 2002] and complex constraints described by logical quantified formula [Boley et al., 2001][McGuinnes & van Harmelen, 2004][McGuinnes et al., 2004], the basic idea is that software agents can use techniques of automatic reasoning in order to relate resources and to support sophisticated web application.

On the other hand, web resources are also characterized by their dynamical aspects, which are not adequately addressed by current web models.

Resources on the web are dynamical since, in the minimal case, they can appear or disappear from the web and their content is upgraded. In addition, resources can traverse different states, which characterized the resource life-cycle, each resource state corresponding to different possible uses of the resource. Finally most resources are timed, i.e. they information they provide make sense only if contextualised with respect to time, and their validity and accuracy is greatly bounded by time.

Temporal projection and deduction based on dynamical and time constraints of the resources can be made and exploited by software agents [Hendler, 2001] in order to make previsions about the availability and the state of a resource, for deciding when consulting the resource itself or in order to deliberately induce a resource state change for reaching some agent goal, such as in the automated planning framework [Fikes & Nilsson, 1971][Bacchus & Kabanza,1998].

Keywords: Temporal Resources, Dynamic Web, evolutionary resources

Introduction

The basic notion of resource in the semantic web [Berners-Lee et al., 2001] is characterised by a unity of structure, content, and location, i.e. a resource has a structure, which is defined in the ontology, a content, i.e. the actual values of their properties, and a unique location, i.e. an URI [Berners-Lee & Fielding,1998], which uniquely identified it in term of its web location.

In our model a resource is still representing a single entity, but entities can evolve over time with respect to the current value of their contents, and also in their structural and semantic description, in other words, the notion of a resource can be intuitively intended as *the invariant aspects with respect to time of a given URI*. For example our department web page is the same resource, despite of the fact that it is continuously updated, in the content and in the structure since our web server was established in 1995.

Resources can be dated and resources can be updated. For many type of resources it is possible to specify when the information will be update, moreover the resource timestamp also provide a relevant information about its validity.

Consider for example:

a) a web page about the history of the independence war,
b) a personal CV,
c) the news of an online newspaper, and
d) stock exchange prices,

they are all web entities with a different rate of upgrade.

The advantages in explicitly defining the date/update features of a resource are apparent with respect to the trust/validity of the information provided by the resource.

Moreover consider for example a) with respect to b), and assume that these two web resources are not updated since the two years ago. It is clear that the info in a) can be used in any moment (assumed that the source is trustworthy), since we do not expect big new facts about the Independence War, on the other hand, discovering that the personal CV was not updated since two years ago, make this information not valid, thus an hypothetical software agent looking for employee information can decide to look for another CV of the same person or to ask the person to provide an upgraded copy.

Update rate can be estimated for c) and d), online newspapers and stock exchange prices are update at different pace, in the first case the content and structure can completely change after some hours, while in the latter the actual value of the price is the only thing which is likely to change, very rapidly when the stock market is open, and to remain still until next opening, during stock market closing hours. A software agent can exploit this information for its cognitive purposes by browsing the online newspaper by the hours or by the week (e.g. sport news about football matches) and the stock prices by the minutes.

Resource States

The state of a resource is an abstract characterisation of structural properties and actual values, which significantly characterise the resource. Associated with the state, there is the possibility that the resource evolves over time by moving from one state to another, in a transition path, which describes the dynamic evolution of the resource.

In the first instance a resource state is an ontological category which is simply characterised by logical constraints about the values of structural properties provided by the ontology, i.e. different ontological concept which share the same schema but not the same actual values. For example FatMan and SlimMan are instances of the concept of Man, they can be defined by constraints over the values of the properties Weight and Height, the interesting aspect is that the FatMan and SlimMan has a dynamical relationship since an individual can move from one state of another by upgrading its weight, (and less probably its height).

In the most general case, resources can allow structural properties to change, i.e. while moving from one state to another the resource evolve its schema.

It is straightforward to represent the admissible states, and the admissible state transitions of a given resource by a labelled transition diagram in which the label represent conditions or event over web resources or time which trigger the state transition.

Def. LTD for Web resources. A labelled transition diagram for a web resource it is a pair $\{N, \delta\}$

Where N is a set of nodes representing the states of the web resource, and δ represents the labelled arcs of the diagram, i.e. the state transition function which defines for every pair of nodes $n_1,n_2 \in N$ a condition L, which labels the arc (n_1,n_2), condition L is a condition over web resources (static and dynamic items, operations, web services, conditions over property values etc.) and time conditions (i.e. current date or general date/time functions).

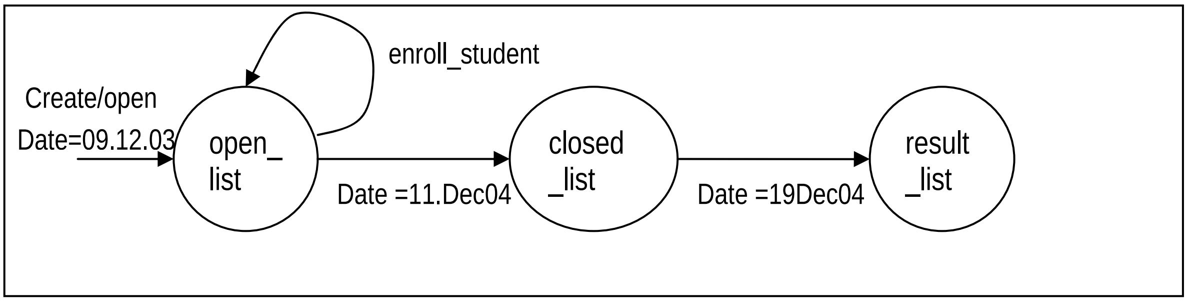

Fig.1 Students List State Transition Diagram

The labelled arc (L, n_1,n_2) denotes the fact that a resource can move from state n_1 to state n_2 when condition L is met, i.e. when the guard is verified, the logical conditions are true or the specified events take place.

By convention (false, n_1,n_2) denotes the fact that no transition is possible between state n_1 and state n_2.

A self-reference loop will represent a resource update; i.e. the resource is remaining in the same state while possibly changing its informative content.

Consider for example the online student enrolment list, exam_list, for the exam code 503 Programming Languages Course which will be held on 12th December 2004, this web resource it is continuously updated since its opening time 3 days before the exam and it is closed the day before and finally it is updated one more time with the list of candidates grades one week after the exam.

The exam_list it is an individual entity despite of the upgrade operations, which are operated on it.

In term of state transitions the evolution of the list can be represented by the state transition network in the figure 1. It is worth noticing that the self-reference loop labelled enroll_student represent the fact that after an enroll_student event.

In this framework the dynamics of resources are represented by appropriate state transition diagrams, which model the resource lifecycle.

A Two Layer Model for Dynamical Web

In order to give an account of the static and dynamic relationships of web resources a two-layer model architecture is proposed in which state transition diagrams are defined over a given ontological network.

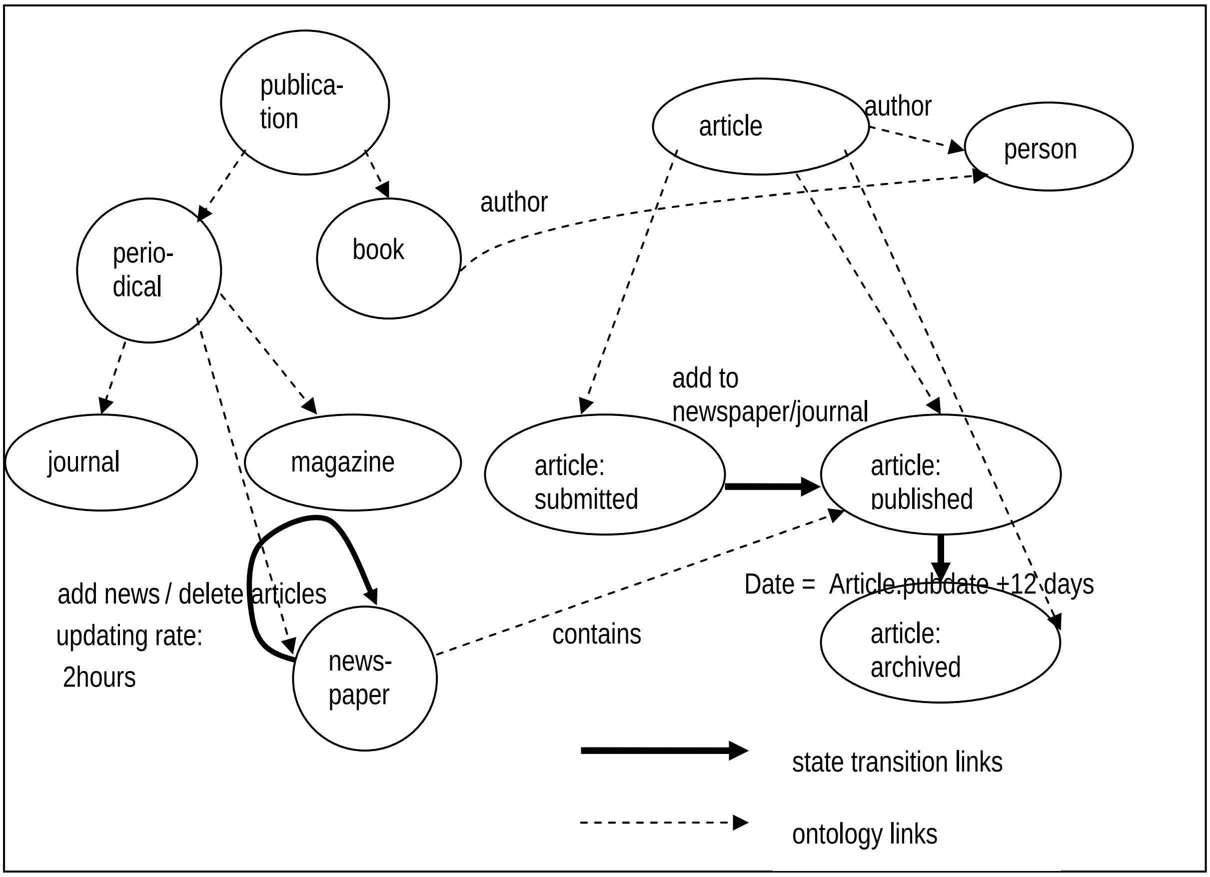

Fig.2 Students List, newspaper and articles state diagrams

For example in figure 2 is represented an online newspaper represented where news change, say, every two hours. In this case the continuous updating of the newspaper resource is represented by a single self reference labelled by operation, i.e. events add/delete news and the conditions that two hours are passed by the previous updated, the newspaper, is related by other ontological concepts (ontological link are represented by dashed lines) such as relationships of type *subclass* with respect to *periodical* and relation *contains* with respect to *published article*.

Articles inherit general concepts such as *author* relationship and can also have a more complex life cycle expressed by a LTD (submitted, published, archived) partially independent from newspaper.

Resource state transitions can be enabled by conditions over other web resources or not web resources.

Reasoning and Acting on Dynamic Resources

State transition diagrams and labelled transition diagrams are formalism popular in the area of process modelling and concurrent system modelling [Gogolla & Parisi-Presicce,1998][Cardell-Oliver et al.,1992].

The labelled transition diagram of a given resource can also be seen as an equivalent deterministic finite state automata DFA, whose transition guards (i.e. the logical conditions) denotes set of symbols in the alphabet defined by the possible binary combination of all atomic conditions. A given guard denotes the set of symbols, which correspond to atomic conditions, which makes the guard itself true.

It would be interesting to investigate the possibility of applying techniques of linear temporal logics [Manna & Pnueli,1991][Vardi,1991] used in circuit testing in order to evaluate LTL queries over a particular state of the dynamical resources.

For example an agent which has found the resource can reason about the truth of LTL modal formula such as Possibly(S) where S is a desired state of the resource, taking appropriate measure, such as *abandoning* the resource if the desired state is unreachable (e.g. the conference submitting deadline is over then transition to state "submitted" is impossible), *waiting* if the state transition is a matter of time or of exogenous events (e.g. wait until tomorrow for the President elections results), inducing the state transition by agent *deliberative action* [Milani & Ghallab,1991](e.g. reserve a ticket in order to buy it), *maintain conditions* which avoid unlikely transitions, consult the resource if the update rate or the type of time validity requires it (i.e. refresh the stock exchange prices, check again the weather forecast service).

Conclusion

The presented preliminary model extends the ontology-based approach to the semantic web, in order to represent the dynamical aspects of web resources, which evolves over time. The classical semantic web hypothesis of web resources as identified by URI is no more valid when web resources evolve; i.e. they assume different states. States of resources are represented by ontological concepts, while labeled transition diagrams are used for description admissible states of resources. The transitions are labeled by the conditions which trigger the state transition, i.e. operation, events or conditions over other resources. The resulting semantic description of web resources consist of a two layered graph which represents both static (i.e. the ontological concepts) and dynamic (i.e. the labeled transitions) relationships among concepts.

The finite state machine model allows to employ powerful reasoning technique in the semantic web graph, as for example LTL (linear temporal logic) in order to make prevision about state of web resources over time. Moreover, this rich knowledge description network can be exploited by web agents who use planning techniques for triggering the desired state transitions.

Bibliography

[Bacchus & Kabanza,1998] Bacchus,F. Kabanza,F., 1998 Planning for temporal extended goals, Annals of Mathematics and Artificial Intelligence, 22:5-27, 1998

[Berners-Lee & Fielding,1998] T. Berners-Lee, R.Fielding, U.C. Irvine, L.Masinter, Uniform Resource Identifiers (URI): Generic Syntax, Request for Comments: 2396, the Internet Society, August 1998

[Berners-Lee et al., 2001] L.Tim Berners Lee, J. Hendler, O. Lassila: The Semantic Web, Scientific American, May 2001

[Boley et al., 2001] Harold Boley, Said Tabet, and Gerd Wagner. Design Rationale of RuleML: A Markup Language for Semantic Web Rules. In International Semantic Web Working Symposium (SWWS), 2001.

[Cardell-Oliver et al.,1992] Rachel Cardell-Oliver, Roger Hale, and John Herbert. An embedding of timed transition systems in HOL. In Higher Order Logic Theorem Proving and its Applications, pages 263--278, Leuven, Belgium, Sept 1992

[Fikes & Nilsson, 1971] R.E.Fikes, N.J.Nilsson. STRIPS: A new approach to the application of theorem proving to problem solving, Artificial Intelligence, 2(3/4), 1971

[Gogolla & Parisi-Presicce,1998] Gogolla, M. and Parisi-Presicce, F., 1998, "State Diagrams in UML - A Formal Semantics using Graph Transformation", Proceedings ICSE'98 Workshop on Precise Semantics of Modeling Techniques (PSMT'98),

[Handschuh & Staab, 2002] S. Handschuh and S. Staab. Authoring and annotation of web pages in CREAM. In The Eleventh International World Wide Web Conference (WWW2002), Honolulu, Hawaii, USA 7-11 May, 2002

[Hendler, 2001] J. Hendler, Agents and the Semantic Web, IEEE Intelligent Systems, vol.16, no.2, Mar./Apr. 2001, pp.30-37.

[Kalianpur, 2001] SMORE - Semantic Markup, Ontology, and RDF, Editor Aditya Kalyanpur A. V. Williams Building College Park, Maryland 20742

[Manna & Pnueli,1991] Manna, Z., Pnueli, A. The temporal logic of reactive and concurrent systems: Specification. Springer Verlag, 1991

[McGuinnes & van Harmelen, 2004] D.L. McGuinness F.van Harmelen, OWL Web Ontology Language Overview W3C Recommendation 10 February 2004

[McGuinnes et al., 2004] D.L. McGuinness, R. Fikes, J. Hendler, and L.A. Stein. ``DAML+OIL: An Ontology Language for the Semantic Web''. In IEEE Intelligent Systems, Vol. 17, No. 5, pages 72-80, September/October 2002

[Milani & Ghallab,1991] A.Milani, M.Ghallab eds. "New Direction in AI Planning", IOS Press 1996

[Vardi,1991] M. Vardi. An automata-theoretic approach to linear temporal logic. In F. Moller and G. Birtwistle, editors, Logics for Concurrency, pages 238-266. Springer Verlag, 1996

Authors' Information

Alfredo Milani – Department of Mathematics and Informatics, University of Perugia, Via Vanvitelli, 1, 06100 Perugia, Italy; e-mail: suriani@dipmat.unipg.it

Silvia Suriani – Department of Mathematics and Informatics, University of Perugia, Via Vanvitelli, 1, 06100 Perugia, Italy; e-mail: suriani@dipmat.unipg.it

EFFECTIVENESS OF TITLE-SEARCH VS. FULL-TEXT SEARCH IN THE WEB

Peretz Shoval and Tsvi Kuflik

Abstract: Search engines sometimes apply the search on the full text of documents or web-pages; but sometimes they can apply the search on selected parts of the documents only, e.g. their titles. Full-text search may consume a lot of computing resources and time. It may be possible to save resources by applying the search on the titles of documents only, assuming that a title of a document provides a concise representation of its content. We tested this assumption using Google search engine. We ran search queries that have been defined by users, distinguishing between two types of queries/users: queries of users who are familiar with the area of the search, and queries of users who are not familiar with the area of the search. We found that searches which use titles provide similar and sometimes even (slightly) better results compared to searches which use the full-text. These results hold for both types of queries/users. Moreover, we found an advantage in title-search when searching in unfamiliar areas because the general terms used in queries in unfamiliar areas match better with general terms which tend to be used in document titles.

Keywords: Indexing, Information retrieval, Precision of search results, Search engines, Title search, Web search.