Kripke style Dynamic model for Web Annotation with Similarity and Reliability

M. Kopecký[1], M. Vomlelová[2], P. Vojtáš[1]

Faculty of Mathematics and Physics Charles University
Malostranske namesti 25, Prague, Czech Republic

[1]{kopecky|vojtas}@ksi.mff.cuni.cz
[2]marta@ktiml.mff.cuni.cz

Abstract. In this "visionary contribution" (term from conference organizers) we deal with the web semantization as a (semi) automated process of enriching web data in a way understandable for algorithms. It turns out that similarity and dynamic aspects of web data play a role here. We propose a web data extension by a Kripke style dynamic model to describe this process for future extractions.

Keywords: information extraction, semantic annotation, Kripke dynamic logic

1 Motivation

No human can read, understand, and synthesize the whole web information on an everyday basis. So we need automated web data processing. Our dream is to have a multicriterial web search supporting a customer (user) looking for a product (service, resource). Our interest is on web content which is not semantically annotated by owner. To distinguish – we see a difference between schema.org annotation (enabling search engines to improve their job) and our annotation which tries to make (and index upon request) part of the web relevant to user search more like an integrated database.

Closest to our approach is the work of G. Gottlob: Lixto [1](see e.g. [2]) – that is wrapper generation, web data extraction … just to mention a few. On the other side, to our knowledge, Dublin Core Metadata Initiative[2] methods, serve a different purpose – namely for mapping of ontologies to integrate sources annotated by an owner already.

Our motivation example (Fig.1, [4]) shows a web page of a travel agency consisting of several data records (hotels) inside a data region. Our first task is to recognize these areas (in the DOM tree) and make them available for further extraction of e.g. price, quality. The main idea is to use the fact that there is a certain repetition as the

[1] http://www.lixto.com/
[2] http://dublincore.org/

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 105-109.

page was constructed using a template (using Levenshtein similarity and threshold depending on the domain, see [4] with experiments in notebooks, cars and hotels).

So we have the task, algorithm (with preconditions and post conditions), training sets and parameters ensuring best result depending on application domain and possibly a similarity measure. Now we can imagine two scenarios:

- to reuse the algorithm (on a similar page, e.g. created by same template or similar in more general sense) it would be good to have records on previous experiences,
- maybe extracted data are no more available (e.g. many LOD points do not function properly) and we have to extract data again or on demand.

In both cases it would be helpful to have semantic data extended with this information.

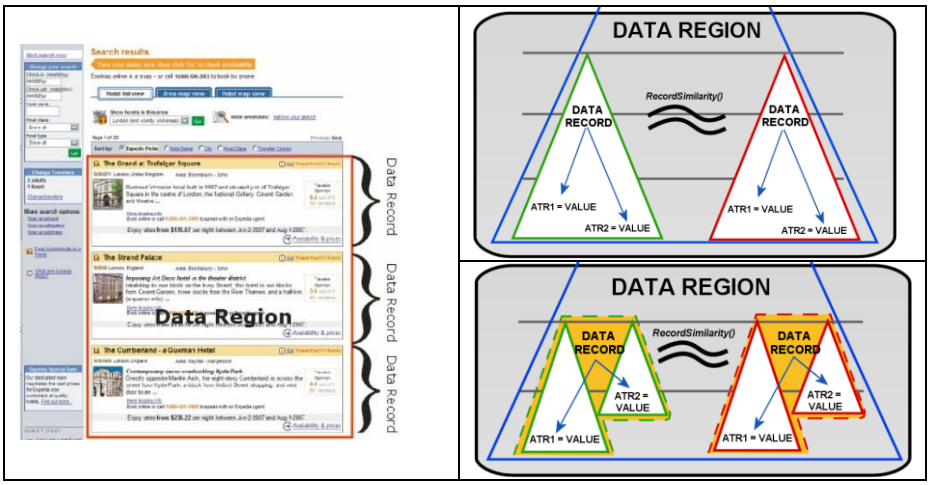

Fig. 1. Web resource, SW extracting data region and data records (in one cell, several cells), [4]

In [5] we presented several methods for mining web information and assisted annotations as we believe this should be the first steps towards the semantic web. Then several methods for processing the gathered data are described. The proposed methods mainly aim on modeling user and his/her preferences and then helping them with reaching their goals. We considered also the connection with a social network (see e.g. SoSIReCR[3]) and friends annotating an initial golden standard.

Our idea her is: we would like to have a formal model to remember origin of web extracted data and the means of this extraction (for possible future reuse and/or re-extraction and comparison of quality of alternative extraction tools).

[3] http://www.sosirecr.cz/index_en.php

2 A model for description of extraction/annotation circumstances

For effective using of changing and/or increasing information we have to evolve tools (e.g. inductive methods) used for creation of specific web service (e.g. recommendation). Our goal is to extend the semantic web foundations to enable describing creation, dynamics and similarities on data. To describe the reliability of extraction algorithms we propose a "half-a-way" extension of dynamic logic [3]. Programs (typically extractors) remain propositional whereas formulas will be more predicate-like (describing properties/attributes of web resources).

Kripke states correspond to different representation of content on the web and results of our extraction and annotation. So, we have two forms of states – input states and output states (with a possible overlap, to be able to describe chaining of extraction algorithms). Programs will usually act on various forms of web content representation, e.g. XML, (X)HTML, tables, DOM, texts, …). Today we can be challenged also by big data downloaded from the web and stored in a cloud. Output of our programs are data in various machine readable forms carrying semantic information, e.g. RDFa enrichment of (X)HTML, relations, FOL, RDF, texts (with PoS, morphology, dependency).

We aim to define a language working with data, hence we need to give our formulas a meaning. To specify formulas of our dynamic logic on each of states the respective semantics is defined using appropriate query language, e.g. XQuery, XPath, FOL, SPARQL, SQL, keyword search, … E.g. an empty SELECT or ASK can give me information on validity of certain statement hidden in FROM, WHERE conditions. New development ([1]) in area of multimodal databases gives extensions of SQL able to handle different forms of data.

Our "half-a-way" extension of dynamic logic has expressions of two sorts (and each sort is/can be typed): Statements about web data: atomic e.g. Φ_0^{RDF}, Φ_0^{FOL}, Φ_0^{RDB}, Φ_0^{XML}, Φ_0^{DOM}, Φ_0^{BoW}, Φ_0^{PoS}, $\Phi_0^{DepTree}$, … and Φ more complex φ^{RDF}, ψ^{FOL}, … with corresponding data model, query language based semantics – all can be subject of uncertainty, probability extensions.

Programs remain propositional: atomic, e.g. Π_0^{σ} for subject extraction, Π_0^{π} for property extraction, Π_0^{ω} for object value extraction in case of html, xhtml, xml data; Π_0^{ner} for named entity extraction in case of text data, and Π more complex $\alpha^{\sigma\pi\omega}$, $\beta^{\sigma\pi\omega}$, $\gamma^{\sigma\pi\omega}$, …

Statements are typically accompanied by information about program creation (data mining tool used for extraction, training data, metric (e.g. precision, recall) …) and there is a lot of reification describing the training and testing data and the metrics of learning. In place of ontologies we assume usage of user created dictionaries (usually very simple). Our model is based on dynamic logic, calculates similarity of Kripke states and describes uncertain/stochastic character of our knowledge.

Hence we are able to express our experience using extraction algorithms in statements like $\{\varphi\}\ \alpha\ \{\psi\}$ or $\varphi \rightarrow [\alpha]_x\ \psi$, where φ is a statement about data D_1 before extraction (preconditions), ψ is a statement about data/knowledge D_2 , K_2 after extrac-

tion (post conditions), α is the program used for extraction. Modality $[\alpha]_x$ can be weighted, describing stochastic aspects of learning. Lot of reification about learning can be helpful.

The main idea of our vision is that if there are some data D_1' similar to D_1 and φ is true in some degree (e.g., because both resources were created using the same template) then after using α we can conclude with high certainty/probability that the statement ψ will be true in some degree on data D_2' (knowledge K_2').
So one has first to train extractors but afterwards check how these extractors are resistant to data changes (with information on specific similarity measure).
We have already provided some experiments with extraction and similarity resistance, but this is out of scope of this "visionary" paper.

In our motivating example from Figure 1, ($\{\varphi\}\ \alpha\ \{\psi\}$ or $\varphi \rightarrow [\alpha]_x\ \psi$) can look like:

α_{Mar} is the software from [4], maybe available free on a URL with training data D_1

φ says: α_{Mar} was trained on data D_1^{ntb}, D_1^{car}, D_1^{hot} in notebooks, cars and hotels domain with parameters of learning (parameters of α_{Mar}, cross validation, metric, …) and precision, recall, … of extraction.

ψ can describe output, e.g. if D_2^{ntb}, D_2^{car}, D_2^{hot} and the similarity resistance i.e. if $D'_1{}^{ntb}$, $D'_1{}^{car}$, $D'_1{}^{hot}$ are similar to D_1^{ntb}, D_1^{car}, D_1^{hot} (in same/close domains) in similarity $\approx^{ntb}$, $\approx^{car}$, $\approx^{hot}$ in degree x^{ntb}, x^{car}, x^{hot} then after running α_{Mar} one can expect P/R …

To conclude, we have presented our vision on how to enable remembering origin of web extracted data and the means of their extraction for future reuse and/or re-extraction. We propose a formal dynamic model for automated web annotation with similarity and reliability (a Kripke style dynamic logic model).

We expect to face several challenges. First is that our doing tries to improve both extractor and resistance against similarity. These are two parameters and it is not clear what the optimum is. The Pareto optimal point – a very good extractor which is highly resistant to similarity changes is probably hard to construct for nontrivial data. Classical data mining optimizes this by cross validation and trying to cover the whole data space. This is difficult to achieve on the web, so we can expect dynamic strategies as web content evolves. To optimize similarity resistance is a new challenge.

Another challenge is as with all semantic web ideas: how to convince people to use them. One way is to concentrate on well-organized communities (medicine, pharmacology) or governmental data sources and publishing regulated by a law. Our vision should act also out of these well-organized communities, so we can expect either positive influence of social network friends or customer creation methods (e.g. Research Gate). We have acquaintance with two sides of the coin working with our social network of IT specialists in the regions of the Czech Republic (https://www.sitit.cz/).

As a future work we will concentrate on experiments in specific domains on bigger and variable data.

Acknowledgement: Research was supported by Czech project Progres Q48.

References

1. M. Aslett. Neither fish nor fowl: the rise of multi-model databases, February 8th, 2013, The 451 Group,
2. R. Baumgartner, G. Gottlob, M. Herzog: Scalable Web Data Extraction for Online Market Intelligence. PVLDB 2(2): 1512-1523 (2009)
3. D. Harel, D. Kozen, J. Tiuryn. Dynamic Logic (Foundations of Computing) The MIT Press, 2000
4. R. Novotny, P. Vojtas, D. Maruscak. Information extraction from web pages. In Proc. 2009 IEEE/WIC/ACM WI&IAT-workshops, 121-124, dl.acm.org id=1632266
5. L. Peska, I. Lasek, A. Eckhardt, J. Dedek, P. Vojtas, D. Fiser: Towards web semantization and user understanding. In EJC 2012, Y. Kiyoki et al Eds. Frontiers in Artificial Intelligence and Applications 251, IOS Press 2013, pp 63-81