Formal Model of Meta–Information Acquisition from Information Resources

Vojtěch Svátek[1] and Václav Snášel[2]

[1] Department of Information and Knowledge Engineering,
University of Economics, Prague, W. Churchill Sq. 4, 130 67 Praha 3, Czech Republic
svatek@vse.cz
[2] Department of Computer Science, Technical University of Ostrava,
17. listopadu 15, 708 33 Ostrava-Poruba, Czech Republic
vaclav.snasel@vsb.cz

Abstract. An outline of formal model describing the acquisition of 'meta–information' on information resources is being proposed, which should enable to compare the quality of different analysis procedures. It is illustrated an example from website analysis.

1 Introduction

While the amounts of information resources available on the internet as well as in company intranets are exponentially growing, their retrieval is still predominantly based on brute-force methods such as full-text indexing. It is however clear that the presence of reliable *meta-information* can bring significant added value to resource discovery. Since it is not feasible to assign meta-information to each resource by hand (as in the traditional library indexing), various knowledge-based, statistical, algebraic and other methods have recently been designed for automatic acquisition of meta-information from the full content of the resources. Examples are *text classification* or *information extraction*, which are frequently applied to either plain text (e.g. articles) or websites. The research on (often quite similar) meta-information acquisition methods is actually fragmented among many different communities, such as that of information retrieval, (semi-structured) databases, automated semantic annotation (for the semantic web), library subject categorisation and the like. It is highly desirable to find a model that would enable to capture (at least partially) the essence of different approaches, to *compare* different methods and tools, and to determine their *complementarity* and/or *supplementarity* with respect to prototype tasks.

Currently, to the knowledge of the authors, all existing models of meta-information acquisition belong to two extreme categories:

1. Models restricted to a *single type* of meta-information. Typical approaches are *classification* of information resource to a particular, predefined category (e.g. subject topic), and *retrieval* of resource that fulfils a predefined role with respect to the given resource (e.g. the bibliography page within a personal website).

2. *Open* models enabling to specify an almost arbitrary structure of meta-information. A typical example is the apparatus of *semantic web ontology languages* such as OWL [4].

It seems highly desirable to fill in the blank space between the two extremes. Our recent work on *problem-solving models* for deductive web mining, in particular, the TODD framework [8], successfully unified apparently dissimilar information resource analysis methods. It however addresses the problem of (mostly) *vertical co-operation* between different analysis tools, while we need *horizontal alignment* of such tools.

In this paper, we propose a model that associates a resource with a collection of properties that belong to three types (closed-domain properties, object properties and content properties) distinguished by the range of their values. The evaluation of 'correct' value assignment differs according to property type, the results can however be aggregated into a single table. We hypothesise that such a model is particularly useful when multiple complementary/supplemantary procedures can be applied on different data structures related to same underlying entities. This is typical for websites (hence the choice of illustrative example), which are known to exhibit a high degree of redundancy in presentation, but the model can presumably be generalised to other types of resources, too.

In section 2 we explain the model itself. Section 3 presents an in-breadth example demonstrating the whole approach. Section 4 discusses some limitations of the approach and compares it with the web-ontology approach. Finally, section 5 wraps up the paper.

2 Resources, Properties, Values and Meta–Information

In this section we formally define the important notions and explain their role in the whole model.

Definition 1. *Let* $\mathcal{R}$ *be the* universe of information resources *(further only resources). Let* $R_t \subseteq \mathcal{R}$ *be a* set of resources *of same* type t*.*

An information resource can be any unique entity that carries some information content. In the case of web, it can be for example a physical web page, a hyperlink, or a whole website. 'Physical web page', 'hyperlink', and 'website' can be considered as types of resources.

Definition 2. *Let* $P = P_{CD} \cup P_O \cup P_{Cn}$ *be a set of* properties *relevant for a set of resources* R*;* P_{CD}*,* P_O *and* P_{Cn} *are pairwise disjoint.* P_{CD} *will be called* closed–domain properties*,* P_O object properties *and* P_{Cn} content properties*.*

Every p *from* $P_{CD} \cup P_{Cn}$ *has an associated* value set*,* V_p*. Every* p *from* P_O *has an associated resource type* t_p*.*

We assume that a fixed set of properties can characterise the resources (of a given type) from different aspects, e.g. to indicate the *author* of a given page,

the *direction* ('upward', 'downward' etc.) of a given hyperlink, or the *startup page* of a website. The first is an example of a content property, the second of a closed–domain property, and the third of an object property. While the value set of closed–domain properties is assumed as enumerated, the value set of content properties is derived from some standard open data type. For simplicity, let us assume that content properties have character strings for values—this is the typical type of target information in information extraction. The model can thus be understood as unifying, from a very particular viewpoint, the paradigms of *information retrieval* (values of object properties refer to retrieved resources), *text classification* (values of closed-domain properties can be understood as classes of resources) and *information extraction* (values of closed-domain properties correspond to semantically labelled text extracted from the content of resources).

Note that the model is not limited to meta–information in the usual sense, i.e. 'information about the resource itself', but extend it to any information that could be acquired from the resource and is 'somehow' related to it. Typically, it is the case of information about the entity that 'owns' and/or 'created' the resource, e.g. about the company a website is devoted to. This is not a conceptual problem, since such 'second–order' meta–information can be directly captured by *composed* properties, e.g. 'location–of–company–owning–the–site'.

Definition 3. *For each* $r \in R, p \in P$, *the* reference value *of* p *for* r *will be denoted as* $Ref(r,p)$.

- *If* $p \in P_{CD} \cup P_{Cn}$ *then* $Ref(r,p) \in V_p$.
- *If* $p \in P_O$ *then* $Ref(r,p) = r_k \in R_t$, *where* t *is the resource type associated with* p.

Note that we do not introduce the reference value in the sense of 'ontologically true' value, but just as the value to which other values would be compared. The reference value could be even 'N/A' (i.e., 'not available') if the real value cannot be derived from the resource content in principle.

Now, we will introduce the notion of *meta–information set*, which corresponds to a (possibly partially) filled 'template' over the set of properties.

Definition 4. *A* meta–information set *of resource* $r \in R$, $\mathcal{M}_r$, *is a set* $\{(p_1, v_1), (p_2, v_2), \ldots, (p_k, v_k)\}$, *where* $\{(p_1, p_2, \ldots, p_k)\}$ *are all properties relevant for* R. *A* reference meta–information set *of resource* $r \in R$, $\mathcal{M}_r^{Ref}$, *is a set* $\{(p_1, Ref(r,p_1), (p_2, Ref(r,p_2)), \ldots, (p_n, Ref(r,p_n))\}$, *where* $\{(p_1, p_2, \ldots, p_n)\}$ *are all properties relevant for* R. *Every resource has* exactly one *reference meta–information set.*

For simplicity, we assume that a meta–information set always covers *all* properties, some of them may however have the value 'N/A'.

Definition 5. *In a given* context, *every* p *from* P_{Cn} *has an associated* value–acceptability relation $Acc_p \subseteq V_p \times V_p$.

The value–acceptability relation may, depending on the nature of the property, amount e.g. to:

– strict equality: $Acc_p(v, Ref(r,p))$ iff $v = Ref(r,p)$ (e.g for a company registration code)
– term permutation (e.g. for keyword lists)
– term superset with length restriction (e.g. for person names—if the page author is 'John Smith', we could possibly accept the value 'Dr. John Smith', sometimes even 'page written by John Smith' but not a long sentence).

The acceptability relation may not be fixed for the given property: it may vary according to 'context'. The notion of context may, in practice, correspond e.g. to an *evaluation session* for different meta–information acquisition procedures. We may thus bias the evaluation toward 'strictness' or 'sloppiness', and obtain coarser or finer distinctions of the procedures' quality. Alternatively, the context may be the *syntactical standard* for true meta–information.

Definition 6. *A meta–information set of resource $r \in R$, $\mathcal{M}_r$, is* correct *for r in property p if it contains a pair (p, v) such that:*

1. *if $p \in P_{CD} \cup P_O$ then $v = Ref(r,p)$*
2. *if $p \in P_{Cn}$ then $Acc_p(v, Ref(r,p))$.*

Note: We chose to require strict identity for properties from $P_{CD} \cup P_O$. In principle, we could introduce some 'acceptability relation' even for these. For example, relaxed criteria for object properties might require, in a certain context, merely sub/super–object (part–of) relation to hold (instead of identity), and similarly, for closed–domain properties, the sub/superclass relationship in a hierarchy could fulfil such role.

Definition 7. *Given two meta–information sets $\mathcal{M}_{r_i}$, $\mathcal{M}'_{r_i}$ of the same resource r_i, $\mathcal{M}_{r_i}$ is* superior *(or equal) to $\mathcal{M}'_{r_i}$ with respect to r_i, $\mathcal{M}_{r_i} \geq \mathcal{M}'_{r_i}$ if $\mathcal{M}_{r_i}$ is correct for r_i in every property in which $\mathcal{M}'_{r_i}$ is correct for r_i.*

This enables to compare the performance of two meta–information acquisition procedures on the same resource (e.g. web document).

3 Example

Let us demonstrate our scheme on a 'toy' example: a hypothetical web page of a small toy producer (Fig. 1).

Let us consider this page as our 'current resource' on which the procedures will be evaluated. Let the *reference meta-information set* wrt. this resource be:

$\mathcal{M}^{Ref} = \{$ (p_1 = *page_type*, 'Main information page')
(p_2 = *page_author*, "John Black")
(p_3 = *name_of_company_referenced_by_page*, "BlackWood Ltd.")
(p_4 = *domain_of_competence_of_company_referenced_by_page*, "toys")
(p_5 = *pricelist_location*, '`/html/body/ul`')
(p_6 = *contact_address_location*, 'N/A') $\}$

```
<html>
<head>
<meta author="John Black">
<meta description="Production and sale of wooden and textile toys">
</head>
<body>
<h1>BlackWood Ltd.</h1>
<p>An opportunity for lovers of original hand-carved and hand-knitted toys.
BlackWood specialises in toys for children aged over 6 years,
and in collectioners' items.</p>
<p>In our shop you can find:</p>
<ul>
<li>wooden animals from 4,-
<li>knitted dolls from 8,-
<li>toy furniture collection, special offer for just 120,-
</ul>
<hr>
<it>Page maintained by J. Black, last modification Feb 28, 2002.</it>
</body>
</html>
```

Fig. 1. Example resource: page of a toy producer

p_1 is a closed–domain property, p_2, p_3 and p_4 are content properties, and p_5 and p_6 are object properties. The acceptability relations for content properties will be defined as follows:

- $Acc_2(x, y)$ holds (for sequences of terms) if x contains the last term from y and no more than three other terms.
- $Acc_3(x, y)$ holds (for sequences of terms) if x contains the first term from y and no more than one other term.
- $Acc_4(x, y)$ holds (for sequences of terms) if x contains all terms from y and less than $|y|$ other terms.

We will consider four *meta–information acquisition procedures* Pr_1, Pr_2, Pr_3 and Pr_4. For instructiveness, we will assume that[3]

- Pr_1 is a Naive Bayes page categoriser operating on unigram representation.
- Pr_2 is an extractor of META tag content, equipped with heuristics mapping META attributes on target meta–information properties.
- Pr_3 is a linguistic, parser–based information extractor equipped with a database of domain–neutral lexical indicators.
- Pr_3 is an HTML–and–punctuation–based information extractor equipped with a database of domain–neutral lexical indicators.

[3] These hypothetical procedures roughly correspond to some of the tools developed within the *Rainbow* project [7].

The meta–information sets produced by the procedures will be ($\mathcal{M}(k)$ denoting the output of the $k-$th procedure, for the given resource):

$\mathcal{M}(1) = \{$ ($p_1 =$ *page_type*, 'Main information page')
($p_2 =$ *page_author*, 'N/A')
($p_3 =$ *name_of_company...*, 'N/A')
($p_4 =$ *domain_of_competence...*, 'N/A')
($p_5 =$ *pricelist_location*, 'N/A')
($p_6 =$ *contact_address_location*, 'N/A') $\}$

(the keyword–based categoriser is clearly designed for classification only, and does not care about structural patterns; assignment to 'Main information page' might be due to the presence of several 'promotion' keywords such as 'opportunity', 'lovers' or 'original')

$\mathcal{M}(2) = \{$ ($p_1 =$ *page_type*, 'N/A')
($p_2 =$ *page_author*, "John Black")
($p_3 =$ *name_of_company...*, 'N/A')
($p_4 =$ *domain_of_competence...*,
"Production and sale of wooden and metallic toys")
($p_5 =$ *pricelist_location*, 'N/A')
($p_6 =$ *contact_address_location*, 'N/A') $\}$

(assuming that the meta–attribute 'description' is tentatively mapped on the property *domain_of_competence...*)

$\mathcal{M}(3) = \{$ ($p_1 =$ *page_type*, 'N/A')
($p_2 =$ *page_author*, "J. Black")
($p_3 =$ *name_of_company...*, "BlackWood")
($p_4 =$ *domain_of_competence...*,
"toys for children aged over 6 years, collectioners' items")
($p_5 =$ *pricelist_location*, '`/html/body/ul`')
($p_6 =$ *contact_address_location*, 'N/A') $\}$

(the value of *page_author* is identified with the subject that 'maintains' the object 'page'; the value of *name_of_company...* is identified with the subject that 'specialises' in *something*—which is, in turn, identified with the value of *domain_of_competence...*; finally, the *pricelist_location* is identified with the HTML element immediately following that with the phrase '...you can find')

$\mathcal{M}(4) = \{$ ($p_1 =$ *page_type*, 'Main information page')
($p_2 =$ *page_author*, "Page maintained by J. Black")
($p_3 =$ *name_of_company...*, "BlackWood Ltd.")
($p_4 =$ *domain_of_competence...*, 'N/A')
($p_5 =$ *pricelist_location*, '`/html/body/ul`')
($p_6 =$ *contact_address_location*, 'N/A') $\}$

(the *page_type* is recognised by presence of free–text paragraphs as well as a list—a mixture presumably typical for 'Main information page'; the value of *page_author* is identified with the slanted text in the page footer; the value of *name_of_company...* is identified with the text in the topmost header, containing

the pattern 'Ltd.'; finally, the *pricelist_location* is identified with the unordered HTML list containing estimated 'price–patterns' in each item).

Taking into account the acceptability relations for p_2, p_3 and p_4, the 'correctness scores' of the meta–information sets for the individual properties are as follows:

Property	Pr_1	Pr_2	Pr_3	Pr_4
page_type	1	0	0	1
page_author	0	1	1	0
name_of_company...	0	0	1	1
domain_of_competence...	0	0	0	0
pricelist...	0	0	1	1
contact...	0	0	0	0

From the table we can deduce that e.g.:

- $\mathcal{M}(4)$ is superior to $\mathcal{M}(1)$
- $\mathcal{M}(3)$ is superior to $\mathcal{M}(2)$

It is however clear that *complementarity* of procedures is more important than *superiority*. We can easily see that:

- no combination of procedures can correctly acquire all meta–information
- if we removed the 'inaccessible' $p4$ and 'inavailable' $p6$, the minimal combinations of procedures needed for correct acquisition would be $\{Pr_1, Pr_3\}$, $\{Pr_2, Pr_4\}$ and $\{Pr_3, Pr_4\}$.

Furthermore, it is more reasonable to remove the *closed–world assumption*, since ignorance should not be treated the same as error. The original table will then look as follows:

Property	Pr_1	Pr_2	Pr_3	Pr_4
page_type	1	?	?	1
page_author	?	1	1	0
name_of_company...	?	?	1	1
domain_of_competence...	?	0	0	?
pricelist...	?	?	1	1
contact...	?	?	?	?

We could also construct the table for *pairs* of procedures. Here, in the case where both procedures return a value, we can either demand, for a correct result, that *both* are correct, or just that *at least one* is correct.

With the first interpretation, the table looks as follows:

Property	Pr_1,Pr_2	Pr_1,Pr_3	Pr_1,Pr_4	Pr_2,Pr_3	Pr_2,Pr_4	Pr_3,Pr_4
page_type	1	1	1	?	1	1
page_author	1	1	0	1	0	0
name_of_company...	?	1	1	1	1	1
domain_of_competence...	0	0	?	0	0	0
pricelist...	?	1	1	1	1	1
contact...	?	?	?	?	?	?

We can see that, in the first (more natural?) interpretation, the combinations involving Pr_4 now become inferior, since its incorrect claim of knowing the value or *page_author* invalidates the correct suggestion of Pr_2 or Pr_3, respectively. The combination $\{Pr_1, Pr_3\}$ then appears as clear 'winner'.

With the second interpretation, the table looks as follows (changes in boldface):

Property	Pr_1,Pr_2	Pr_1,Pr_3	Pr_1,Pr_4	Pr_2,Pr_3	Pr_2,Pr_4	Pr_3,Pr_4
page_type	1	1	1	?	1	1
page_author	1	1	0	1	**1**	**1**
name_of_company...	?	1	1	1	1	1
domain_of_competence...	0	0	?	0	0	0
pricelist...	?	1	1	1	1	1
contact...	?	?	?	?	?	?

The three candidate combinations would then become 'equally correct' again.

4 Discussion

The present model strikes for certain balance between simplicity and coverage. However, due to stress on the former, many important aspects of real-world meta-information acquisition tasks remain uncovered. Among the most important limitations (and topics for future research) are probably the following:

- The model does not treat different methods of combining the results of multiple procedures for the same property (such as voting)
- No uncertainty is allowed in the results of the procedures.
- The properties are single-valued.
- The model is static, it does not anticipate possible change of property values over time.
- The model does not introduce homomorphism (similarity, deformation) for the results of multiple procedures, it is main problem for modelling meta-information

It might also be interesting to compare the typology of properties in our model with the inventory of *web ontology languages* such as OWL [4]. In OWL, content properties would correspond to *datatype properties with embedded data type*, while closed–domain properties would correspond to (object/datatype) *properties with enumerated class / data type*, respectively, since enumeration is possible, in principle, for both types of properties in OWL. Unlike web ontology languages, we however treat *class membership* simply as (closed–domain) properties, to keep the model general enough. Mapping (of semantically relevant elements) from OWL to our model seems to be rather straightforward: a structured ontology could be flattened to our model by replacing pairs of chained properties with new, composed, properties. Having done such mapping, we could

then e.g. 'tap' on existing tools producing meta-information in the form of RDF triples [6].

In Kalfoglou [2] was used *CHU space* for Ontology Mapping. When we want model structure of meta-information (infomorphisms and homomorphisms) by Chu spaces see [5].

Chu space is a matrix over a set Σ. It is formalized in as follows:

Definition 8. *A Chu space $A = (A, r, X)$ over a set Σ, called the alphabet, consist of a set of points constituting the carrier A, a set X of states constituting cocarrier, and a function $r : A \times X \to \Sigma$ constituting the matrix.*

In our case it is matrix with rows corresponding to resources and columns corresponding to properties. Chu space entries are drowning from $\{0, ?, 1\}$. Where 0 indicates resources has not properties, ? indicates activity in progress and 1 indicates finished activity. Formally, carrier A is set of resources, cocarrier X is set of properties and alphabet is $\Sigma = \{0, ?, 1\}$.

5 Conclusions

We presented a model of meta-information acquisition from information resources, and illustrated it on an example from website analysis. Although the web is probably the most characteristic area of application for this kind of model, it could probably be applied to other types of resources.

Future work will address some of the limitations discussed in section 4. Since the model is currently merely descriptive, we examine some existing algebraic formalisms that could extend it with more rigorous semantics. We would also like to elaborate on the problem of transformation/mapping from ontology languages suggested in section 4; an adequate method could possibly be adapted from state-of-the-art research on ontology transformation [3] and alignment [1]. Finally, we plan to implement a simple software tool and test it on the top of a collection of *Rainbow* web services [7].

Acknowledgments

The research is partially supported by grant no. 201/03/1318 of the Grant Agency of the Czech Republic, "Intelligent analysis of the WWW content and structure".

References

1. Doan, A., Halevy, A., Noy, N.: Proceedings of the Semantic Integration Workshop collocated with the Second International Semantic Web Conference (ISWC-03), online at http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS//Vol-82/.

2. Kalfoglou Y., Schorlemmer W. M.: IF-Map: An Ontology-Mapping Method Based on Information-Flow Theory. J. Data Semantics I 2003: 98-127.
3. Omelayenko, B., Klein, M. C. A. (Eds.): Knowledge Transformation for the Semantic Web. Frontiers in Artificial Intelligence and Applications Vol. 95, IOS Press 2003.
4. OWL Web Ontology Language Reference, W3C Recommendation, 10 February 2004, http://www.w3.org/TR/owl-ref/
5. Pratt, V. R.: Chu Spaces. Course notes for the School in Category Theory and Applications. Coimbra, Portugal, 1999.
6. Resource Description Framework (RDF), W3C, http://www.w3.org/RDF/
7. Svátek, V., Kosek, J., Labský, M., Bráza, J., Kavalec, M., Vacura, M., Vávra, V., Snášel, V.: Rainbow - Multiway Semantic Analysis of Websites. In: 2nd International DEXA Workshop on Web Semantics (WebS03), Prague, IEEE 2003.
8. Svátek, V., Labský, M., Vacura, M.: Knowledge Modelling for Deductive Web Mining. In: 14th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2004), Whittlebury Hall, Northamptonshire, UK. Springer Verlag, LNCS, 2004.