Ontology Suitability for Uncertain Extraction of Information from Multi-Record Web Documents

David W. Embley[*], Norbert Fuhr, Claus-Peter Klas, Thomas Rölleke

Fachbereich Informatik, Universität Dortmund

June 22, 1999

Abstract

Ontology based data extraction from multi-record Web documents works well [ECLS98, ECJ$^+$98, ECJ$^+$99, EJN99], but only if the ontology is suitable for the Web document. How do we know whether the ontology is suitable? To resolve this question, we present an approach based on three heuristics: density, schema, and grouping. We encode the first heuristic as a density function and use probabilistic models for the second and third. We argue that these heuristics and our computational models for these heuristics correctly determine the suitability of a Web document for a given ontology.

1 Introduction

In the past, information was stored more or less well-structured in databases. Nowadays, a lot of information is presented in semi-structured document languages such as HTML, XML, and others. Automatic extraction of information from these kinds of semi-structured documents is much more difficult than automatic extraction of information from well-structured databases.

One approach to extracting information from semi-structured documents uses so-called ontologies, which provide a kind of semantic description for supporting automatic information extraction. If the ontology is properly set up and properly applied to an HTML document that is suitable for the ontology, then fact extraction works quite correctly and completely [ECLS98, ECJ$^+$98, ECJ$^+$99, EJN99]. This earlier work explains how to properly set up an ontology, but it does not address the problem of determining whether a Web document is suitable for the ontology.

In this earlier work, humans assessed the suitability of a Web document for the ontology and only passed suitable documents to the information-extraction system. However, for a more automated system or for the purpose of adding processing power to make autonomous agents behave more independently, we need to be able to assess automatically the suitability of a document for a given ontology. As one expects, fact extraction works better on "good" pages than on "bad" ones.

We describe in this paper our approach of determining the suitability of ontologies for extracting facts from a page. For a "good" page, the suitability rating should be "high"; for a "bad" page, the suitability rating should be "low". The suitability rating must take into account the inherent uncertainty of information extraction and, at the same time, must provide reliable results to separate "good" pages from "bad" pages.

The approach we take is heuristic and is based on theoretically strong computational models. We propose three heuristics: (1) *density*, (2) *schema*, and (3) *grouping*. We base our density heuristic on the simple observation that for an ontology to be suitable to a page, a reasonably large fraction of the page must contain constants and keywords recognizable by and applicable to the ontology. We base our schema heuristic on the observation that data for each of the attributes in the schema should be

[*] The ideas for this paper were developed while this author was in Dortmund on sabbatical leave from Brigham Young University, Provo, Utah 84602, U.S.A.

present on the page and that the amount of data for each attribute should roughly correspond to the expected cardinalities provided as constraints in the schema. We base our grouping heuristic on the observation that for multi-record pages, the data for the attributes for each record should be grouped together. The computational model for our density heuristic is a ratio of matched to total characters, after appropriate adjustments for overlapping matches. The computational models for our schema and grouping heuristics are probabilistic [CLRC98]. For our schema heuristic we computationally estimate the number of records x and then compute the probability for each attribute a that a page contains x records given that n (uncertain) occurrences of a are recognized by the ontology. For our grouping heuristic we compute the probability that more than an acceptable number of attribute occurrences are missing in a set of groups.

In succeeding sections, we briefly explain the extraction task and then explain each of our suitability heuristics and their computational models. The details provide a justification for our choice of heuristics and their encoding as rules.

2 An Ontology for Extracting Information from Car Ad's

As a task, we deal with the extraction of facts about cars. Consider a semi-structured text as the following which is the result of a text extraction from an HTML document (figure 1).

The document contains free text and record-like data. With an ontology in which we store regular expression for recognizing years, prices, and car models, we can extract information from the document. Of course, the recognition is an uncertain process, and in addition, the uncertainty is strongly influenced by the “suitability” of the ontology, i.e. how does the ontology fit to the content of the document? Depending on the suitability of the ontology, we have high or low trust in the information extracted.

The recognizer produces a list of match facts such as

match(year, “$\{19|20|'\}[0-9][0-9]$”, 78, 80, “9́8”). match(model, “$\{BUICK.*\}$”, 82, 95, “BUICK Century”).

```
NAC/The Salt Lake Tribune Transportation
Classifieds

CLASSIFIEDS | TRANSPORTATION SEARCH
RESULTS

'98 BUICK Century, like brand new, Only
$14,995.  461-8509

'98 BUICK LeSabre, pwr windows, air, V6,
6 passenger.  $15,488.  Barber Bros Super
Center 298-8868

'98 BUICK Century, V-6.  Three to choose
from.  $14,988.  Barber Bros Super Center
298-8868

'98 BUICK Park Ave, White, loaded and
perfect!  Only $23,845.  GARFF VOLVO
297-7108

'98 BUICK Skylark, V6, Blowout!  $9,988.
DL1120 972-8411.

'98 BUICK Century, like new!  $13,977.
DL1120 972-8411.

'97 BUICK Skylark, 2dr, auto, air,
clean as a pin, Low low miles.  Call for
details.  DAN EASTMAN JEEP, 298-3417

¨ 1998 Newspaper Agency Corporation

The classifieds are best viewed with
Internet Explorer 3.02 or above.  Click on
the icon below to download the product.
```

Figure 1: Text extracted from an example Web page

The first parameter shows the attribute name, the second parameter the regular expression which matched, the third and fourth parameter the starting and ending position of the match in the text sequence, and the fifth parameter shows the recognized attribute value. From the sequence of match facts, we want to extract information about cars. However, the pure sequence does not tell us, which year does belong to which model. Also, the quality of the match depends on the ontology that is applied for the recognition. In the following, we develop the suitability of an ontology for obtaining a measure for the trust we have in the extracted facts.

3 Suitability of Ontologies

We propose three criteria for determining the "suitability" of an ontology: *density*, *schema*, and *grouping*.

Density reflects how much of the document is recognized by the ontology. For example, for the sample page shown in figure 1, 80% of the text is recognized as being expressions of the ontology, and thus we have high trust in the extracted facts.

Schema reflects whether the attribute types recognized correspond to the expected attribute types we know from training data. In our sample page, the attribute frequence $year{:}model{:}price = 1{:}1{:}0.9$ is recognized. The closer the attribute frequency of a page to the one learned (expected), the higher is the trust in the extracted facts.

Grouping reflects whether the attributes are grouped in records or spread all over the page. For example, figure 1 shows an attribute sequence such as *year*, *model*, *price*; in contrast, assume a page showing an attribute frequency such as 10 times *year*, then 10 times *model*, then 10 times *price*. Both show an attribute frequency of $year{:}model{:}price = 1{:}1{:}1$. However, since the first page is more record-like than the second, we have a higher trust in the fact extraction of the first.

The extraction works "ok" if the density, the schema, and the grouping of the page with respect to an ontology is "ok." We can express this condition formally in a probabilistic logical program ([Fuh95]) in which we model the three criteria with a weighting function:

```
0.3 criterion_wt(density).
0.4 criterion_wt(schema).
0.3 criterion_wt(grouping).
extraction_ok() :- criterion_wt(X) &
                   criterion_ok(X).
```

For each criterion_ok(X) fact, we compute a probability. The weighted sums of the probabilities results in the probability that the extracted facts are correct. In the following, we take a closer look at each criterion.

3.1 Density

The density criterion tells us how much of a page is recognized. For obtaining a proper fraction between zero and one, we have to decide on a strategy regarding overlaps. For example, "Porsche 911" is matched by model and "911" is matched by price. The whole text consists of 11 characters, however, 11+3 characters were recognized! Overlaps indicate incorrectness and lead to a wrong number of recognized characters. We can handle overlaps as follows: (1) throw overlapping recognitions away, (2) count overlapping characters only once, or (3) select one recognition. We follow (3)—(1) would mean that we throw away a information and (2) would mean that we may count incorrect recognitions. Selecting one recognition raises the question of which one. Ideally, we would select the recognition which maximizes the number of recognized characters. However, to keep it simple in the first phase, we take as an ad hoc solution just the first recognition or we select one of the overlapping recognitions randomly.

Given the number of characters matched and the number of characters on the page, we compute the density ratio ϱ:

$$\varrho \quad := \quad \frac{\text{number of characters matched}}{\text{number of characters on the page}}$$

We choose ϱ as the probability of the density criterion.

ϱ `criterion_ok(density).`

3.2 Schema

3.2.1 A heuristic model

From a set of training pages, we learn the expected record frequency for each attribute. The record frequency tells us, how often an attribute occurs per record. As an example, consider the following values:

attribute	expected record frequency
year	1
model	1
make	1
price	0.8
phone	2

With the expected record frequency, we can determine an expected record length. For example: $1 + 1 + 1 + 0.8 + 2 = 5.8$ attributes per record.

For a current page, let the recognizer produce a page frequency such as 4:4:4:4:7. We assume that this page frequency corresponds to the record frequency, i.e. in the ideal case free text occurring in the page is removed, and only the record part is input for the recognizer.

For comparing the expected record frequency with the current record frequency, we divide the current by the expected record frequency. For example:
4/1 : 4/1 : 4/1 : 4/0.8 : 7/2 = 4 : 4 : 4 : 5 : 3.5.

The more constant the obtained list of numbers is, the closer is the current record frequency to the expected one. A heuristic function could produce a measure for the "constantness" of the obtained list of numbers. As an alternative for the heuristic approach, we develop next a probabilistic model in which the uncertainty of recognition is considered.

3.2.2 A probabilistic model

From a sequence of match facts, we want to derive further knowledge, namely:

1. What is the number of records (ads) in the page?
2. Does the record frequency of attributes (e.g. year:model:price = 1:1:0.9) of the current page correspond to the record frequency of attributes learned from a set of training pages?

First, we investigate the estimation of the number of records in the current page. We assume that all occurrences of an attribute are recognized (recognition is complete), however, not every recognition is correct. Assuming that n_i occurrences of attribute i are recognized and the probability of a correct recognition is p_i, the probability that k of n_i recognitions are correct is expressed via the binomial probability function: (bee model: n bees in a box, each bee with probability p_i in left corner; what is the probability that k bees are in the left corner?)

$$P(k|n_i) = \left(\begin{array}{c} n_i \\ k \end{array} \right) \cdot p_i^k \cdot (1-p_i)^{n_i-k} \tag{1}$$

k	number of correct recognitions
n_i	number of total recognitions
p_i	precision of recognition
$P(k\|n_i)$	probability that k of n_i recognitions are correct

From the set of training pages, we can determine the probability $P(k_i|x)$ that attribute i occurs k_i times if the page has x records (ads).

k_i	number of occurrences of attribute i
x	number of records
$P(k_i\|x)$	probability that attribute i occurs k_i times if page contains x records

Actually, we are interested in the probability that a page has x records if n_i occurrences of attribute i were detected. We compute the probability $P(x|n_i)$ using the theorem of the total probability:

$$P(x|n_i) \quad = \sum_{k_i=0,\ldots,n_i} P(x|n_i \cap k_i) \cdot P(k_i|n_i)$$

Assuming that the number of records x does not depend on number of recognitions n_i given k_i correct recognitions, we obtain:

$$P(x|n_i) \quad = \sum_{k_i=0,\ldots,n_i} P(x|k_i) \cdot P(k_i|n_i)$$

With Bayes, we reach:

$$P(x|n_i) \quad = \sum_{k_i=0,\ldots,n_i} \frac{P(k_i|x) \cdot P(x)}{P(k_i)} \cdot P(k_i|n_i)$$

The probability $P(k_i|x)$ is determined from the training set. For example, we observe $(k_i, x) = (3, 4)$ for one page, and $(k_i, x) = (6, 7)$ for a second page. Assuming a probability distribution, the parameters are estimated with maximum likelihood or other methods. $P(x)$ is the total probability that a page chosen randomly contains x records. $P(k_i)$ is the total probability that attribute i occurs k_i times in the page. With the prior distribution $P(x)$, we can compute $P(k_i)$ via the theorem of the total probability:

$$P(k_i) \quad = \sum_{x} P(k_i|x) \cdot P(x)$$

With equation 1 for $P(k|n_i)$ and the above estimations for $P(k_i|x)$, $P(x)$, and $P(k_i)$, all values for computing the probability $P(x|n_i)$ that a page contains x records given that n_i occurrences of attribute i are recognized are determined. In particular, the recognizer uncertainty (p_i) is considered in this probabilistic approach whereas it was not considered in the heuristic approach.

It remains the aggregation of evidence for the number of records coming from each attribute. Assume we deal with N attributes $t_1, \ldots, t_n$, and we estimate a weighting function $P(t)$ with $\sum_t P(t) = 1$. The weighted sum over the probabilities $P(x|n_i)$ is a reasonable estimate for the aggregation of evidence.

$$P(x|n_1 \wedge \ldots \wedge n_N) \quad = \quad \sum_{t_i} P(t_i) \cdot P(x|n_i)$$

With this theory, we can now produce probabilistic facts such as e. g.

```
0.1 number_of_records(1).
0.2 number_of_records(2).
0.5 number_of_records(3).
0.2 number_of_records(4).
```

Actually, we want to decide whether the schema (the record frequency of attributes) of the current page corresponds to the expected schema. For determining this, we look at the distribution of the probability function of the number of records. The more narrow the distribution, i.e. the more certain the number of records is, the better it fits the ontology (the schema of the set of training pages) to the schema of the current page. A very first estimation of the probability that the schema is ok, is just the maximum of the probabilities for the number of records, i. e. in our example

```
0.5 criterion_ok(schema).
```

3.3 Grouping

3.3.1 A heuristic model

Records (grouping) are easy to recognize for a person: the layout and the structure helps.

For example, consider two pages such as:

99	Buick	$30600
98	Ford	$25000

99 98 Buick Ford $30600 $25000

We see the same information; however, one is grouped into horizontal records. The associations are easier to resolve in the record-like presentation.

How can we recognize automatically, whether a sequence of attributes is record-like or not? Note that we base the record frequency of attributes on the whole sequence of attributes in a page, and thus, the record frequency does not depend on the actual sequence of attributes. The first two criteria, density and schema, yield the same trust for grouped or scattered attributes.

With an estimation of the number of records (see section 3.2) or with information about the grammar of the page, we can split the sequence into groups. For a set of groups, we want to decide whether the split is record-like or not.

A heuristic approach could be to count in each group the number of "selected" attributes, i.e. we select attributes that identify records and check whether they occur constantly over the groups.

3.3.2 A probabilistic model

A more general approach is to base the decision on the probability that more than an acceptable number of attribute occurrences are missing in a set of groups. For a discrete random variable for the number of missing occurrences, the Poisson distribution is appropriate:

$$P(k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda} \tag{2}$$

$P(k)$ is the probability that k occurrences of attributes are missing in a set of groups. The parameter λ of the Poisson distribution corresponds to the average number of missing attribute occurrences.

Assume that the grouping algorithm for the current page produces n groups. From the training set, we know for each attribute the average number missing per record. For example, from a training set assume we have learned a record frequency of attributes of 1:1:0.9. Of course, the training set contains longer and shorter records. From this, we can compute the average number of missing attributes per record (λ_r).

For a current page with n groups, we obtain the number of missing attributes per page:

$$\lambda = \lambda_r \cdot n$$

Since we also know the number of missing occurrences k, we can compute $P(k)$ using eqn (2) and then generate the probabilistic fact

$P(k)$ `criterion_ok(grouping).`

4 Concluding Remarks

We have offered three heuristics for determining whether an ontology is suitable for processing a document, namely a density heuristic, a schema heuristic, and a grouping heuristic. We have argued that these heuristics capture the essence of suitability in terms of high-density applicability, proper value distribution for attributes, and appropriate grouping for records. We have also argued that the computational models we provided for these heuristics accurately reflect the intent of these heuristics. In future work, we must test these hypothesises empirically.

References

[CLRC98] F. Crestani, M. Lalmas, C.J. Van Rijsbergen, and I. Campbell. “is this document relevant?...probably”: A survey of probabilistic models in information retrieval. *ACM Computing Surveys*, 30(4):528–552, December 1998.

[ECJ+98] D.W. Embley, D.M. Campbell, Y.S. Jiang, Y.-K. Ng, R.D. Smith, S.W. Liddle, and D.W. Quass. A conceptual-modeling approach to extracting data from the web. In *Proceedings of the 17th International Conference on Conceptual Modeling (ER'98)*, pages 78–91, Singapore, November 1998.

[ECJ+99] D.W. Embley, D.M. Campbell, Y.S. Jiang, S.W. Liddle, D.W. Lonsdale, Y.-K. Ng, and R.D. Smith. Conceptual-model-based data extraction from multiple-record web pages. *Data & Knowledge Engineering*, 1999. (to appear).

[ECLS98] D.W. Embley, D.M. Campbell, S.W. Liddle, and R.D. Smith. Ontology-based extraction and structuring of information from data-rich unstructured documents. In *Proceedings of the Conference on Information and Knowledge Management (CIKM'98)*, pages 52–59, Washington D.C., November 1998.

[EJN99] D.W. Embley, Y.S. Jiang, and Y.-K. Ng. Record-boundary discovery in web documents. In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data (SIGMOD'99)*, pages 467–478, Philadelphia, Pennsylvania, 31 May - 3 June 1999.

[Fuh95] N. Fuhr. Probabilistic datalog - a logic for powerful retrieval methods. In E.A. Fox, P. Ingwersen, and R. Fidel, editors, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 282–290, New York, 1995. ACM.