

Automatic Ontology-Based Knowledge Extraction from Web Documents

vs.

Automating the Extraction of Data from HTML Tables with Unknown Structure

Stefan Bischof, Stefan Rümmele

In this report we compare the papers [AKM⁺03] and [ETL03]. We show that the two proposed systems realize different goals with the same or similar underlying technics.

1. Differences of Intentions

- **Source data of interest**

[ETL03] takes web pages containing HTML tables of interest for a given application domain as the input whereas [AKM⁺03] considers unstructured text from webpages for the knowledge extraction process.

- **Resulting data format**

The difference between the output data is similar to the input. While [ETL03] returns structured data fit into a target schema, [AKM⁺03] generates text in a narrative form specified by the user.

- **Interface specification**

[AKM⁺03] is designed as a whole process from knowledge extraction over information management to narrative generation and thus provides an interface for human beings. [ETL03] however, only provides an interface to an information extraction procedure and is not designed as an complete application.

- **Internal data source**

While [AKM⁺03] uses a knowledge base to store the aquired knowledge, [ETL03] doesn't outline a specific internal data storage approach.

- **Project structure**

The approach of [ETL03] is a stand alone project only built up previous research on extraction ontologies. [AKM⁺03] uses several projects for the diverse tasks needed and combines them to a whole process.

2. Implementation Similarities

- **Extraction ontology**

For both papers the central key aspect of the information extraction process is the extraction ontology. Without it, for every new page a wrapper has to be created. But with the ontology the wrapper creation can be automated.

- **Narrow domain of interest**

Since an ontology is a kind of model, it can only represent an restricted field. Hence both approaches do well in a specific application domain but none of them can be expanded to a all-purpose web information extraction tool.

- **Manual onotology creation**

Both papers have the need of manual ontology creation. Out of this fact results, that the adaption to other fields is time-consuming. However [ETL03] state that learning in connection with ontologies is of interest for further research.

- **Need of domain knowledge**

This results out of the previous point. The manual ontology creation can only be achieved with the help of an expert.

3. Conclusion

We want to state that an in-depth comparison of the two papers is difficult. This is based on the fact that [AKM⁺03] is a several times bigger project than the one mentioned in [ETL03]. As a result the authors are not able to give detailed insight into the technical implementation details of every aspect.

Furthermore we cannot say which approach is superior to the other, since [AKM⁺03] deals with natural language processing but [ETL03] is able to use a higher percentage of the found information.

4. References

[AKM⁺03] Harith Alani, Sanghee Kim, David E. Millard, Mark J. Weal, Wendy Hall, Paul H. Lewis, Nigel R. Shadbolt. Automatic Ontology-Based Extraction from Web documents. IEEE Intelligent Systems, Volume 18, Issue 1, January 2003.

[ETL03] David W. Embley, Cui Tao, Stephen W. Liddle. Automating the Extraction of Data from HTML Tables with Unknown Structure. Submitted , May 2003, (source: <http://www.deg.byu.edu/papers/>).