CSE 718 – Seminar Report

An Interative Clustering Based Approach To Integrating Source Query Interfaces on the Deep Web

Presenter – Manas Pradhan
pradhan4@buffalo.edu

Overview:

An increasing number of data sources now become available on the Web, but often their contents are only accessible through query interfaces. To integrate the databases, their query interfaces need to be integrated. This can be done in 2 phase out of which the first phase includes the semantic field mapping over the interfaces and then using this field mapping we integrate the interfaces in the second phase. The accuracy of the integration majorly depends on the accuracy of the field matching. This paper mostly concentrates on the field matching phase of the interface integration.

The paper first comments on the limitations of the current approaches: Lack of hierarchical modeling, 1:1 mapping assumption, black box operation and laborious parameter tuning. It then tries to put forward an explanation as in how the clustering based approach overcomes the above drawbacks of the approach. It uses the hierarchical modeling, clustering of fields, handling complex mappings and parameter learning by user interactions for the same. This approach models the interface as an ordered tree of fields before it applies interface matching. In interface matching it handles 2 types of mappings namely 1:1 and 1: m mappings and where 1:m mappings is further decomposed into 'is-a' and 'aggregate' types of mappings. It uses the bridging effect to handle the 1:1 mappings effectively and 1: m mappings are possible to handle because of the hierarchical modeling. After briefly describing the mapping complexities and challenges it then provides the actual implementation of the field matching and clustering algorithm.

This paper also tries to provide the interactive nature to the algorithm by providing a feedback loop wherein the user reactions are used for parameter learning and resolving the uncertainties. The user is asked to answer the matching questions in simple yes/no terms and the answers are used to evaluate the accuracy of the field matching. Also the threshold value is decided not on a trial and error basis but on the basis of the user interactions in this phase. It gives more accuracy to the approach as seen in the experimental results part of the paper. The uncertainties in the algorithm are posed due to the use of homonyms and synonyms for labeling the fields. The approach even tries to reduce these considerably in the feedback phase.

The experiments were conducted by the authors of this paper on 5 different domains: airfare, automobile, real estate, book and job. The accuracy results are provided for each domain distinctly and the results also highlight the importance of each feature introduced in this approach by providing the accuracy results separately for each feature's inclusion into the approach. The accuracy is calculated in terms of precision, recall and F measure which takes a bit of both precision and recall into consideration.

The paper presentation concludes with mentioning some of the related and future works in this area. The related work basically includes study on the bridging effect vs. mapping reusing techniques , user interaction and parameter learning and schema and interface matching. Future work that the paper talks about includes tie resolution methods, automatic interface modeling into the approach and resolving some unresolved uncertainties by user reactions.

Detailed Comments:

The strengths of the paper are its hierarchical modeling and the user learning features. The hierarchical modeling gives the approach the means to handle complex mappings in the interfaces and the user learning phase removes considerable number of uncertainties and reduces some parameter tuning time which becomes a no more trial and error operation.

The weakness of the paper lies in the fact that the paper only considers the importance of the first phase of interface matching and doesn't really provide any solutions towards integrating the interfaces in the approach which has to be assumed to be done manually. The paper could have been more effective by talking a more about the actual implementation and automation of the interface integration based on the accurate field matching achieved by the approach.

The paper does go into the technical details of applying the clustering algorithm and calculating the field similarities by providing the formulae and algorithms and also providing a detailed explanation for the same. It however avoids mentioning some of the information retrieval techniques like normalization with enough depths and providing just a generic explanation about the same.

There were some questions raised during the presentation which doubted if the approach considered the fields just in the query interface or also the ones in the query results. An interesting question included was if the fields that appear only in some of the interfaces are being considered or being excluded while interface matching which remained unanswered due to the paper not dealing with the interface matching in details. However the general comments over the paper suggested that the paper could have been much better had it talked about the interface integration little more than it did.