

Integrating Structured Data on the Web

Thanh Nguyen, School of Computing, University of Utah

May 11, 2011

The explosion of structured data on the Web (e.g., online databases, products, Wikipedia, linked open data) creates many opportunities for integrating and querying these data that go way beyond the simple search capabilities provided by search engines. Although much work has been devoted to data integration in the database community, the Web brings new challenges. There is a large (and growing) volume of data which are heterogeneous and noisy. Before they can be queried, these data need to be organized and cleaned. Because they do not come with an associated schema or ontology, relationships and correspondences among them must be discovered. And because there is so much data, scalable techniques are needed that require little or no manual intervention. The goal of my PhD dissertation is to develop algorithms and tools that enable on-the-fly, automatic integration of large collections of structured Web data. Below, I give a brief overview of some of the problems I have worked on, specifically: Web form schema matching [16, 15], discovery of entity types and relationships for Wikipedia infoboxes [17], and multilingual schema matching for Wikipedia infoboxes [14].

Matching Web Form Schemata. It is estimated there are millions of databases on the Web [11]. The contents of these databases are hidden and are only exposed on demand, as users fill out and submit Web forms. Several applications have emerged which attempt to uncover hidden-Web information and make it more easily accessible, including meta-searchers, hidden-Web crawlers and Web information integration systems. These applications face several challenges, from locating relevant forms [3] and determining their domains [5, 4], to understanding the semantics of the form elements and identifying matches among elements across different forms [12, 9, 22, 8, 18, 20]. Different approaches have been proposed for matching form schemata, but these share an important shortcoming: due to the wide variability in how forms are designed, they require that the forms be pre-processed so as to deal with their heterogeneity. The reliance on pre-processing is problematic for a large heterogeneous data set, since manual pre-processing is expensive and automated approaches are error prone.

To address this problem, we proposed PruSM [16, 15], a new form-matching strategy. PruSM has two notable features: it leverages the availability of a large number of forms to determine *correlations* among attributes, which used as a source of similarity; and to minimize the propagation of matching errors, it prioritizes matches with high confidence by incorporating both syntactic and latent correlation information in an aggregated fashion, i.e., considering *sets* of elements. The set of matches is built incrementally: the initial, high-confidence matches are used to resolve uncertain ones, incrementally grow the set of (certain) matches. We have evaluated PruSM using 2,900 forms in multiple domains. Our experiments show that PruSM obtains high precision and recall without any manual pre-processing; has higher accuracy (from 10% and 68%) than existing holistic approaches; and it is able to find matches for infrequent attributes which are commonplace in form collections.

Discovering Entities and Relationships in Wikipedia. Another rich source of structured data on the Web is Wikipedia. The mass collaboration approach around Wikipedia has led to the creation of a valuable information source. Although the success of Wikipedia can be attributed in

part to the simplicity of adding information, this has also limited its usefulness, notably, when it comes to querying the information. Even though authors are encouraged to provide structure in the pages they create and edit, by selecting appropriate categories and providing infoboxes, often they do not follow the guidelines or follow them loosely. This leads to several problems including data heterogeneity, template duplication and schema drift. Several approaches have been proposed which aim to extract semantic as well as structured information implicit in Wikipedia documents, so as to support more expressive queries [7, 21, 1, 19]. However, a drawback of these approaches is that they rely on categories [19] and template names [1, 21] which can lead to errors due not only to ambiguity present in template names and in categories, but also to inconsistencies, mistakes or irrelevant information added by users.

To address this limitation, instead of relying solely on categories and templates, we leverage the structured information available in Wikipedia infoboxes, and in an attempt to *discover* the correct schema for a given entity, we group together infoboxes (instances) that have similar schemas [17]. But clustering the infobox schemas is a challenging problem. First, we do not have any a priori knowledge of the number of entity types (or clusters), and there is potentially a large number of entity types with a very *skewed distribution*, *i.e.*, a few types being very frequent and many types having low frequency. Second, schema *heterogeneity* is prevalent: even within an entity type, there is a wide variability in the schemas used, with many attributes being *optional* and *rare* [21]. Furthermore, the presence of *ambiguous* attributes can mislead the clustering algorithms.

We take advantage of the large number of available infobox schemas to compute *attribute correlations* for each attribute pair, and use them as a source of similarity (and dissimilarity) information to cluster infobox schemata. In particular, we apply a two-pronged approach: before clustering the infoboxes, we first discover the representative attribute sets that are highly correlated and thus likely candidates for describing an entity—we use these as the basis to group similar infobox schemata together. We also leverage the topological structure of the infoboxes, specifically, the link patterns among the entities to discover meaningful relationships, as well as to reconcile and refine the entity types.

Our experiments using over 100,000 infoboxes extracted from Wikipedia show that our approach outperforms other clustering methods, and that it is effective and able to construct an accurate schema for Wikipedia content, even in the presence of noisy, manually-edited data. The derived entity clusters have high coverage and quality. A comparison against DBpedia data shows that our clusters are meaningful, cohesive and include *new* entity types which are not covered by DBpedia. And since this process is automated, it supports the dynamic nature of Wikipedia. We have also explored the use of the entities and relationships we discover to support structured queries over the infoboxes [17, 13].

Matching Multilingual Schemas in Wikipedia. Recent research has taken advantage of Wikipedia’s multilingualism as a resource for cross-language information retrieval and machine translation, as well as proposed techniques for enriching its cross-language structure. The availability of documents in multiple languages also opens up new opportunities for querying structured Wikipedia content, and in particular, to enable answers that straddle different languages. As a step towards supporting such queries, we proposed **WikiMatch**, a method for identifying mappings between attributes from infoboxes that come from pages in different languages [14]. Our approach leverages *latent semantic analysis* [10] and other kinds of information readily available in Wikipedia to find mappings across multiple infoboxes in a completely automated fashion. Because **WikiMatch** does not rely on learning techniques, it is scalable: not only can it be used to find mappings between many language pairs, but it is also effective for languages that are under-represented and lack sufficient training samples. Another important benefit of our approach is that it does not depend on syntactic similarity between attribute names, and thus, it can be applied to language pairs that have distinct morphologies. We

have performed an extensive experimental evaluation using a corpus consisting of pages in Portuguese, Vietnamese, and English. We also compared WikiMatch against state-of-the-art techniques from data integration [2] and information retrieval [10], as well as to a technique specifically designed to align infobox attributes [6]. The results show that WikiMatch outperforms existing approaches in terms of F-measure, and in particular, it obtains substantially higher recall. We also present a case study where we showed that, through the use of the correspondences derived by WikiMatch, a multilingual querying system is able to derive a higher-quality answers [14].

References

- [1] S. Auer and J. Lehmann. What have innsbruck and leipzig in common? extracting semantics from wiki content. In *ESWC*, pages 503–517, 2007.
- [2] D. Aumüller, H. H. Do, S. Massmann, and E. Rahm. Schema and ontology matching with COMA++. In *SIGMOD*, pages 906–908, 2005.
- [3] L. Barbosa and J. Freire. An adaptive crawler for locating hidden-web entry points. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 441–450, 2007.
- [4] L. Barbosa and J. Freire. Combining classifiers to identify online databases. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 431–440, 2007.
- [5] L. Barbosa, J. Freire, and A. S. da Silva. Organizing hidden-web databases by clustering visible web documents. In *ICDE*, pages 326–335, 2007.
- [6] G. Bouma, S. Duarte, and Z. Islam. Cross-lingual alignment and completion of wikipedia templates. In *CLIAWS3*, pages 21–29, 2009.
- [7] R. Gleim, E. Mehler, and M. Dehmer. Web corpus mining by instance of wikipedia. In *EACL*, 2007.
- [8] B. He and K. C.-C. Chang. Automatic complex schema matching across web query interfaces: A correlation mining approach. *TODS*, 31(1):346–395, 2006.
- [9] H. He and W. Meng. Wise-integrator: An automatic integrator of web search interfaces for e-commerce. In *VLDB*, pages 357–368, 2003.
- [10] M. Littman, S. T. Dumais, and T. K. Landauer. Automatic cross-language information retrieval using latent semantic indexing. In *CLIR*, pages 51–62, 1998.
- [11] J. Madhavan, S. Cohen, X. L. Dong, A. Y. Halevy, S. R. Jeffery, D. Ko, and C. Yu. Web-scale data integration: You can afford to pay as you go. In *CIDR*, pages 342–350, 2007.
- [12] H. Nguyen, T. Nguyen, and J. Freire. Learning to extract form labels. *Proc. VLDB Endow.*, 1(1):684–694, 2008.
- [13] H. Nguyen, T. Nguyen, H. Nguyen, and J. Freire. Querying Wikipedia Documents and Relationships. In *WebDB*, 2010.
- [14] T. Nguyen, V. Moreira, H. Nguyen, H. Nguyen, and J. Freire. Multilingual schema matching for wikipedia infoboxes. Technical report, University of Utah, Salt Lake, UT, 2011.
- [15] T. Nguyen, H. Nguyen, and J. Freire. Prudent schema matching for a large number of web-form interfaces. Technical report, University of Utah, Salt Lake, UT, 2009.
- [16] T. Nguyen, H. Nguyen, and J. Freire. PruSM: A prudent schema matching strategy for web-form interfaces. In *CIKM*, pages 1385–1388. ACM, 2010.
- [17] T. Nguyen, H. Nguyen, and J. Freire. Entity and relationships discovery in wikipedia. Technical report, University of Utah, Salt Lake, UT, 2011.
- [18] W. Su, J. Wang, and F. Lochovsky. Holistic query interface matching using parallel schema matching. In *EDBT*, pages 77–94, 2006.

- [19] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A Core of Semantic Knowledge. In *WWW*, 2007.
- [20] J. Wang, J. Wen, F. Lochovsky, and W. Ma. Instance-based schema matching for web databases by domain-specific query probing. In *VLDB*, pages 408–419, 2004.
- [21] F. Wu and D. S. Weld. Automatically refining the wikipedia infobox ontology. In *WWW*, pages 635–644, 2008.
- [22] W. Wu, C. Yu, A. Doan, and W. Meng. An interactive clustering-based approach to integrating source query interfaces on the deep web. In *SIGMOD*, pages 95–106, 2004.