STUDY OF CHALLENGES AND TECHNIQUES IN LARGE SCALE MATCHING

Sana Sellami, Aicha-Nabila Benharkat, Youssef Amghar
LIRIS-INSA de Lyon, National Institute of Applied Sciences of Lyon, 69621 Villeurbanne, France

Rami Rifaieh
San Diego Supercomputer Center, University of California, La jolla - 92093-0505, California, U.S.A.

Keywords: Matching, Quality of Matching (QoM), Large Scale, Optimization techniques.

Abstract: Matching Techniques are becoming a very attractive research topic. With the development and the use of a large variety of data (e.g. DB schemas, ontologies, taxonomies), in many domains (e.g. libraries, life science, etc), Matching Techniques are called to overcome the challenge of aligning and reconciling these different interrelated representations. In this paper, we are interested in studying large scale matching approaches. We define a quality of Matching (QoM) that can be used to evaluate large scale Matching systems. We survey the techniques of large scale matching, when a large number of schemas/ontologies and attributes are involved. We attempt to cover a variety of techniques for schema matching called Pair-wise and Holistic, as well as a set of useful optimization techniques. One can acknowledge that this domain is on top of effervescence and large scale matching need much more advances. So, we propose a contribution that deals with the creation of a hybrid approach that combines these techniques.

1 INTRODUCTION

Actually, we are witnessing an explosive growth in the amount of data being collected in the business and scientific area. Databases in these domains are filling up with huge amounts of data information with different representations. These data are heterogeneous, frequently changing, distributed, and their number is increasing rapidly. The presence of vast heterogeneous collections of data causes one of the greatest challenges in the data integration field.
Hence, Matching techniques attempt to develop automatic procedures that search the correspondences between these data in order to obtain useful information. In fact, Matching is an operation that takes data as input and returns the semantic similarity values of their elements/attributes. In our paper, we describe new research work of large scale matching, which differs from the existing research papers (Rahm and Bernstein, 2001), (Shvaiko and Euzenat, 2005) in terms of large scale necessities. In fact, traditional schema Matching works are developed for small scale and static integration scenarios, in which automatic Matching technique is often an option to reduce human labour. In contrast, in large-scale data integration scenarios (Madhavan et al., 2007), the Matching needs to be as automatic as possible and scalable to large quantity of data. Furthermore, current matching algorithms have been performed with simple data holding a small number of components, whereas in practice, real world data are voluminous. The size of data can impact match accuracy because it determines the search space for match candidates. In consequence, the quality of Matching will be decreased. We introduce, then, the major criteria of an ideal Matching system at large scale. We define a quality of Matching (QoM) in terms of factors and metrics that can be used to evaluate large scale matching systems. This analysis of state of the art allows us to make some conclusions and observations about the existing matching works. Depending on these observations, we suggest the creation and the elaboration of a hybrid approach that combines these known techniques to deal with a large scale Matching.
This paper is organized as follows. In section 2, we define and describe a quality of Matching (QoM) to evaluate large scale matching systems. Section 3 presents a review of state of the art matching at

Sellami S., Benharkat A., Amghar Y. and Rifaieh R. (2008).
STUDY OF CHALLENGES AND TECHNIQUES IN LARGE SCALE MATCHING.
In *Proceedings of the Tenth International Conference on Enterprise Information Systems - DISI*, pages 355-361
DOI: 10.5220/0001721903550361

large scale. In section 4, we describe our vision for a large scale matching. Finally, we conclude and discuss future work.

2 LARGE SCALE MATCHING SYSTEMS EVALUATION

Evaluations of schema matching systems have been deeply studied in (Do et al., 2002) discussing various aspects (input, output, match quality measures, effort) that contribute to the match quality obtained as the result of an evaluation. In the large scale context, we define and propose a Quality of Matching (QoM) which is an evaluation of large scale matching systems. The quality concept has been used in several domains as an important phase of evaluation in the current information systems. However, there exists little of work (Bernstein et al., 2004), (Duchateau et al., 2007) which tackles the aspect of quality in the matching process at large scale. Therefore, we estimate that is important and interesting to relate the aspect of quality to the scalable matching techniques. In fact, the quality assessment brings to the users an optimal solution to accomplish their needs. Quality of matching (QoM) means for us an optimization of large scale matching system. We firstly need to identify which quality factors to be evaluated. The selection of the appropriate quality factors implies the selection of metrics and the implementation of evaluation algorithms that measure and estimate such quality factors. We distinguish between two aspects: the factors that influence the quality and the metrics to evaluate and measure the quality of matching process.

2.1 Quality Factors in Large Scale Matching

The factors that have an influence on large scale are essentially related to the context (input data and domain) and matching systems or algorithms. We summarize these quality factors in the following paragraph.

2.1.1 Factors Related to the Context

Input Data. Quality depends on the internal quality of the sources (their coherence, their completeness, their freshness, etc.), on the confidence about producers of these sources. Moreover, we should determine the type, representation and structure of data that have been used (schemas, ontologies, taxonomies, query interfaces etc). These characteristics influence the quality of matching.

Domain. Data reside at different sources and consequently they are extracted from different domains. Data managed by different sources are typically heterogeneous, and data can be incorrect, incomplete, and noisy, thus it may be data of poor quality. Therefore, it is important to determine if the data source result from different or the same domains, the characteristic of domains, etc.

2.1.2 Factors Related to Matching Systems/ Algorithms

Techniques. In a context where the information is produced by sophisticated algorithms, the quality measurement requires a fine knowledge of the computing process of this information. Moreover, the use of these algorithms and techniques (i.e. the type of the matchers implemented (schema vs. instance level, element vs. structural level, language vs. constraint based, etc), auxiliary information, optimization techniques, etc.) could be very expensive.

Needs in Runtime Performance. The quality of matching solutions is measured in terms of how long applications take to be run to completion when tasks of applications are allocated to nodes based on decisions of matching algorithms. This duration is called execution time. Efficient matching algorithms must keep times to a minimum.

Complexity. The matching problem is an extreme case in terms of size and complexity. In fact, the schema matching problem is a combinatorial problem with an exponential complexity. This complexity is due to the large number and size of data (number of schemas/components), the expensive computation of semantic similarity (e.g using the auxiliary resources). Consequently, this makes the naïve matching algorithms for large schemas prohibitively inefficient. Therefore, the complexity is a property that affects the quality of matching algorithms.

Human Interaction (Wang et al., 2007). Matching operation cannot be entirely automated; it is still largely conducted by hand, in a labor-intensive and error-prone process. The manual matching has now become a key bottleneck in building large-scale information management systems. Therefore, user or designer input is necessary to generate correct matchings.

2.2 Quality Metrics in Large Scale Matching

In this section, we define the metrics that are involved individually in existing large scale matching systems evaluations. Our classification could be a support to QoM (Quality of Matching):

Performance. The performance is measured in terms of efficiency and pertinence: Efficiency is the time needed by the system to solve a matching problem. Pertinence evaluates the relevance of matching results. This metric can be calculated by precision and recall values (Do et al., 2002).

Accuracy. Called also Overall has been proposed specifically in schema matching context. This measure considers the post-match effort needed for adding false negative and removing false positives. Accuracy depends on both Recall and Precision measures.

Manual Effort (Wang et al., 2007). It's very important to specify the kind of manual effort during the pre-match process and the post-match process (correction and improvement of the match output).

Scalability. It is a property of systems to keep functioning correctly even with the adding new elements. A system, whose performance improves after adding hardware, proportionally to the capacity added, is said to be a scalable system. An algorithm, design, program, or other system is said to scale if it is suitably efficient and practical when applied to large situations (e.g. large input data set or large number of participating nodes in the case of a distributed system).

Adaptability (Bharadwaj et al., 2004). Refers to the degree to which adjustments in practices, processes, or structures of systems are possible to projected or actual changes of their environment. This criterion could measure the degree of change that a system can support.

Extensibility. Means that the system has been so architected that the design includes all of the hooks and mechanisms for expanding/enhancing the system with new capabilities without having to make major changes to the system infrastructure. Therefore, matching systems should be extended by adding matching techniques, algorithms or customized data structures and operators.

3 REVIEW OF EXISTING MATCHING APPROACHES

We are interested in our work in Matching techniques that aim at identifying semantic correspondences between schemas, ontologies, query interfaces, etc. In the literature, we can distinguish between two matching approaches: Pair-Wise matching and holistic matching. We discuss in this section the research works related to these approaches and we underline the most employed optimization techniques.

3.1 Pair-wise Matching

Matching has been approached mainly by finding pair-wise attribute correspondences, to construct an integrated schema for two sources. Several pair-wise matching approaches over schemas and ontologies have been developed.

3.1.1 Schema Matching

Being a central process for several research topics like data integration, data transformation, schema evolution, etc., schema matching has attracted much attention (Avesani et al., 2007), (Bernstein et al., 2004), (Lu et al., 2005), (Smiljanic et al., 2006), (Do and Rahm, 2007) by researchers community. We are more interested to the approaches that integrate the clustering and fragmentation techniques. In fact, these techniques aim at reducing the dimension of the matching problem and improving the quality of the matching (QoM). In (Do and Rahm, 2007), the authors have developed the fragment-based match approach, i.e., a divide and conquer strategy which decomposes a large matching problem into smaller sub-problems by matching at the level of schema fragments. This approach is done «*a priori*» before the matcher's execution. The fragment-based approach represents an effective solution to treat large schemas. However, only few static fragment types are supported and matching large fragments lead to long execution time. The authors in (Smiljanic et al., 2006) propose a clustered schema matching technique which is a technique for improving the efficiency of schema matching by means of clustering. The clustering is introduced *«a posteriori*» after the generation of matching elements. Clustering is then used to quickly identify regions in the schema repository which are likely to include good matchings for the smaller schema. The clustered schema matching is achieved by an adaptation of the clustering algorithm K-means (Xu and Wunsch, 2005). Moreover, Clustering was

combined with B&B (Branch and Bound) algorithm to find highly ranked matchings. Using this optimization algorithm allows to discover efficiently the best solutions in the whole search space. Though, the improved efficiency comes at the cost of the loss of some matchings. The loss mostly occurs among the matchings which rank low. However, there is no measure of cluster's quality that can be used to decide which clusters have better chances to produce good matchings. In addition, the proposed approach is restricted to 1:1matchings.

3.1.2 Ontology Matching

Ontology matching is a promising solution to the semantic heterogeneity problem. It finds correspondences between semantically related entities of the ontologies. The increasing awareness of the benefits of ontologies for information processing has lead to the creation of a number of large ontologies about real world domains. The size of these ontologies causes serious problems in managing them. Actually, many approaches (Hu and Qu, 2006), (Hu et al., 2006), (Qu et al., 2006), (Stuckenschmidt and Klein, 2004), (Wang et al., 2006) have been proposed in literature to study the large ontology matching problem. For instance, to cope with the large ontologies matching, (Hu and Qu, 2006) propose a partitioning-based approach to address the block matching problem. The authors consider both linguistic and structural characteristics of domain entities based on virtual documents for the relatedness measure. Partitioning ontologies is achieved by a hierarchical bisection algorithm to provide block mappings. Another approach has been proposed (Wang et al., 2006) to deal with large and complex ontologies. This is a divide-and-conquer strategy which decomposes a large matching problem into smaller sub-problems by matching at the level of ontology modules. This method uses the E-connection (Grau et al., 2005) to transform the input ontology into an E-connection with the largest possible number of connected knowledge bases. However, this approach does not discover the complex mappings and does not realize the matching between several voluminous ontologies.

3.2 Holistic Matching

Traditional schema matching research has been determined by pair-wise approach. Recently, holistic schema matching has received much attention due to its efficiency in exploring the contextual information and scalability. Holistic matching matches multiple schemas at the same time to find attribute correspondences among all the schemas at once. These schemas are usually extracted from web query interfaces in the deep Web. Several current approaches to holistic schema matching (Chang et al., 2005), (He and Chang, 2006), (He et al., 2004), (He and Chang, 2003), (Madhavan et al., 2005), (Pei et al., 2006), (Su et al., 2006b), (Su et al., 2006a) rely on a large amount of data to discover semantic correspondences between attributes. Holistic approach has been introduced in (He and Chang, 2003). The authors propose MGS framework which is an approach for global evaluation, building upon the hypothesis of the existence of a hidden schema model that probabilistically generates the schemas that we had observed. The authors propose to apply X^2 hypothesis testing to quantify how consistent the schema model is with the data. Nevertheless, this approach does not take into consideration complex mappings. DCM framework has been proposed in (He et al., 2004) for local evaluation, lying on the observation that co-occurrence patterns across schemas often reveal the complex relationships of attributes. However, these approaches suffer from noisy data. The works suggested in (Chang et al., 2005),(He and Chang, 2006) outperform these approaches by adding sampling («*a priori*») and voting («*a posteriori*») techniques, which is inspired by bagging predictors. HSM (Holistic Schema Matching) (Su et al., 2006b) and PSM (Parallel Schema Matching) (Su et al., 2006a) have been proposed to find matching attributes across a set of Web database schemas of the same domain. HSM and PSM are purely based on the occurrence patterns of attributes and requires neither domain-knowledge nor user interaction. The approach presented in (Pei et al., 2006) proposes a novel clustering-based approach to schema matching. However, this approach focused only on 1:1 matchings.

3.3 Summary and Classification of Matching Approaches

In this section, we propose a classification of the previous described approaches in (Figure1) according to the optimization techniques. We categorize these techniques in four classes: machine learning techniques, description logics, heuristic algorithms and statistical algorithms. In fact, most of the proposed approaches at large scale integrate these techniques to improve and optimize the quality of Matching (QoM). (Figure1) can be read from two points of view: In top down view, we present different input data occuring in both holistic and pair-wise approaches. In bottom up view, we

can base the classification on methods related to the optimization techniques (e.g clustering, modularization, etc). This classification is inspired from the one presented in (Shvaiko and Euzenat, 2005) by taking into consideration only large scale matching techniques.

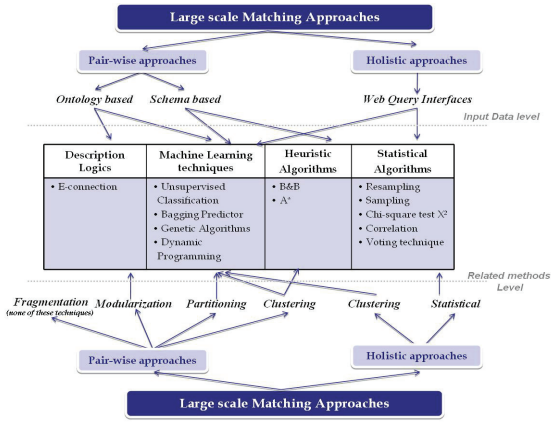

Figure 1: Classification of large scale Matching Approaches.

We can outline from our study on matching the following observations and some open issues that require further research:

- In pair-wise approach, matching is only achieved between two data sources (schemas/ontologies). However, scalable matching system must be able to realize matching among great number of data sources in order to satisfy the needs of real applications. Therefore, pair-wise approaches do not satisfy the scalability criterion.
- Holistic matching is a statistical approach. This approach focuses on observations of the co-occurrence information of attributes across many web query interfaces which involve small number of components in the Deep Web. Then, Holistic approaches have not been applied to ontologies or taxonomies.
- In the majority of existing matching works, the complex mappings are not determined. Most of the existing approaches are focused on the simple matching (1:1). However, discovering complex mappings is a critical semantic operation in the matching problem. Since, the ultimate goal of schema Matching is to derive a Mapping from multiple sources to target (Bernstein et al., 2008), (Melnik et al., 2007).
- Holistic or pair-wise approaches integrate optimization techniques, which are usually performed either in *a priori* matching or *in a posteriori* matching.
- Few works have proposed quality factors and criteria. In the majority of existing works, quality has been defined in terms of precision and recall measures. Therefore, this is insufficient to evaluate the real quality of matching (QoM) system at large scale.
- The majority of Pair-wise matching approaches find attribute correspondences with using auxiliary information. Several works have been proposed for this purpose. For instance, approaches proposed in (Bernstein et al., 2004),(Do and Rahm, 2007) describe the utility to use several matchers. The main idea is to combine the similarities predicted by multiple matchers to determine correspondences. Holistic matching, on the other hand, does not employ any semantic resource for the determination of the correspondences.

4 A NEW VISION FOR LARGE SCALE MATCHING

Based on these observations, we illustrate our vision about a large scale matching system that must include the following points: First, we assume that is interesting to combine the holistic and pair-wise approaches. In fact, Matching in pair-wise systems is usually achieved between only two voluminous data sources. In contrast to this approach, holistic matching is performed between a set of query interfaces. The combination of holistic and pair-wise matchers analyzes schemas/elements under different aspects, resulting in more stable and accurate similarity for heterogeneous schemas. Therefore, their combination can effectively improve the quality of matching. Second, we note the importance of optimization techniques, specially clustering and fragmentation approaches. The main purpose is to deal with large data. With the reduced problem size, we aim to optimize and improve the quality of the matching (QoM). We also underline that the approaches including optimization techniques have a better quality match. Moreover, we notice that these techniques have been integrated either before matching operation (e.g splitting *a priori*) or after matching operation (e.g grouping *a posteriori*). We estimate that is interesting to have a matching system including these techniques in *a priori* and *posteriori* steps. In fact, splitting *a priori* represents an efficient alternative to deal with very large data representations and to reduce the size of large matching problem into small sub-problems. Moreover, grouping *a posteriori* allow us to select and preserve the highly ranked correspondences

result. This step improves the efficiency of schema matching. The combination of these techniques increases the feasibility of large scale matching system. Third, we consider that is important to integrate a quality evaluation in every step of a matching process. Quality evaluation is essential to guarantee the reliability of data representation in order to avoid noisy data. It ensures the consistency of using algorithms and techniques. Moreover, it is necessary to evaluate the matching results and to estimate if the matching system satisfies the quality criteria. Precisely, this quality evaluation allows us to test the performance, accuracy, scalability, adaptability and extensibility of matching system at large scale. Finally, we assess that is essential to employ some auxiliary semantic information to identify finer matching and to deal with the lack of background knowledge in matching tasks. It's also the way to obtain semantic mappings between different input data. Following these ideas, we describe here an instance of our vision for a large scale matching system. (Figure 2) outlines a general procedure for matching at large scale.

Let a set of voluminous (size and number) data, we are going to split up all these sources. This dividing step includes several quality constraints: splitting criteria, reliability of the fragments obtained characteristics of data (structure, format), etc… This phase can be either automatic or manual. Thereafter, we apply a holistic matcher to find similar fragments with a statistical manner. For data in the same domain, those are about a specific kind of topic, usually share common characteristics. The matching resulted can be saved for reusing in the next operations. After determining the similar fragments, we use a pair-wise matcher to find the more complex relations between components. We can employ an auxiliary semantic resource to find these correspondences (e.g determining mapping expressions). Afterwards, we group a posteriori the matching results to select the highly ranked matchings that represent the most pertinent results. We test then the quality of these results to satisfy the accuracy criterion. These results will be saved for a forthcoming use.

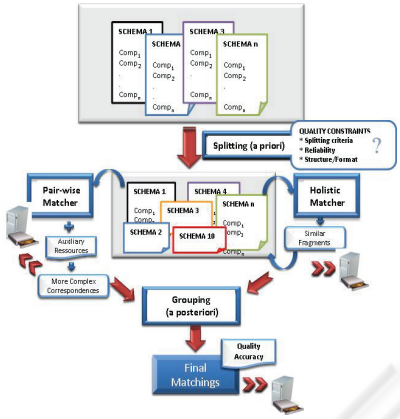

Figure 2: A general Procedure for large scale Matching system.

5 CONCLUSIONS AND FUTURE WORKS

This paper presented a broad scope of matching at large scale categories and characteristics, and surveyed related work. We have presented our motivation to study the solutions for matching at large scale. Since quality is very important to evaluate matching systems, we have described metrics to measure the quality of Matching (QoM) and defined the different factors that influence the quality. We have achieved a state of the art study covering existing approaches: Pair-wise and holistic Matching. We have summarized this survey with listing some important issues and research trends for Matching techniques at large scale. To resume, matching at large scale requires deep domain knowledge: characteristics and representations of data, user's needs, time performance, etc. There is no matching system that can tackle completely all the problems mentioned in this study. We intend in the future to design a matching system that provides all the features described in the previous sections: formalizing quality metrics, splitting, and grouping (e.g clustering) techniques (in *a priori* and *posteriori* phases). The finality of this work is to conceive a complete matching system able to realize matching at large scale between several schemas, ontologies, taxonomies to be applied in various fields such as biology, phylogeny, etc.

REFERENCES

Avesani, P., Yatskevich, M., and Giunchiglia, F. (2007). A large scale dataset for the evaluation of matching

systems. In 4rd European Semantic Web Conference, ESWC'07.

Bernstein, P. A., Green, T. J., Melnik, S., and Nash, A. (2008). Implementing mapping composition. VLDB J., accepted for publication.

Bernstein, P. A., Melnik, S., Petropoulos, M., and Qui, C. (2004). Industrial-strength schema matching. SIGMOD Record, 33(4):38–43.

Bharadwaj, V., Reddy, Y. V. R., Srinivas, K., Reddy, S., Selliah, S., and Yu, J. (2004). Evaluating adaptability in frameworks that support morphing collaboration patterns. In 13th IEEE International Workshops on Enabling Technologies (WETICE 2004), Infrastructure for Collaborative Enterprises, pages 186–191, Modena, Italy.

Chang, K. C.-C., He, B., and Zhang, Z. (2005). Toward large scale integration: Building a metaquerier over databases on the web. In CIDR, pages 44–55.

Do, H. H., Melnik, S., and Rahm, E. (2002). Comparison of schema matching evaluations. In Web, Web-Services, and Database Systems, pages 221–237.

Do, H. H. and Rahm, E. (2007). Matching large schemas: Approaches and evaluation. Inf. Syst., 32(6):857–885.

Duchateau, F., Bellahsene, Z., and Hunt, E. (2007). Xbenchmatch: a benchmark for xml schema matching tools. In VLDB, pages 1318–1321.

Grau, B. C., Parsia, B., Sirin, E., and Kalyanpur, A. (2005). Automatic partitioning of owl ontologies using E-connections. In Proceedings of the 2005 International Workshop on Description Logics (DL2005), volume 147, Edinburgh, Scotland, UK,.

He, B. and Chang, K. C.-C. (2003). Statistical schema matching across web query interfaces. In Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, pages 217–228, San Diego, California, USA.

He, B. and Chang, K. C.-C. (2006). Automatic complex schema matching across web query interfaces: A correlation mining approach. ACM Trans. Database Syst., 31(1):346–395.

He, B., Chang, K. C.-C., and Han, J. (2004). Discovering complex matchings across web query interfaces: a correlation mining approach. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 148–157, Seattle, Washington,USA.

Hu, W. and Qu, Y. (2006). Block matching for ontologies. In The Semantic Web - ISWC 2006, 5th International Semantic Web Conference, volume 4273, pages 300–313, Athens, GA, USA.

Hu, W., Zhao, Y., and Qu, Y. (2006). Partition-based block matching of large class hierarchies. In The Semantic Web - ASWC 2006, First Asian Semantic Web Conference, volume 4185, pages 72–83, Beijing, China.

Lu, J., Wang, S., and Wang, J. (2005). An experiment on the matching and reuse of xml schemas. In 5th International Conference, ICWE 2005, pages 273–284, Sydney, Australia.

Madhavan, J., Bernstein, P. A., Doan, A., and Halevy, A. Y. (2005). Corpus-based schema matching. In Proceedings of the 21st International Conference on Data Engineering, ICDE 2005, pages 57–68, Tokyo, Japan.

Madhavan, J., Cohen, S., Dong, X. L., Halevy, A. Y., Jeffery, S. R., Ko, D., and Yu, C. (2007). Web-scale data integration: You can afford to pay as you go. In Proc. Third Biennial Conference on Innovative Data Systems Research(CIDR 2007), pages 342–350, Asilomar, CA, USA.

Melnik, S., Adya, A., and Bernstein, P. A. (2007). Compiling mappings to bridge applications and databases. In Proceedings of the ACM SIGMOD International Conference on Management of Data, pages 461–472, Beijing, China.

Pei, J., Hong, J., and Bell, D. A. (2006a). A novel clustering-based approach to schema matching. In Advances in Information Systems, 4th International Conference, ADVIS 2006, volume 4243, pages 60–69, Izmir, Turkey.

Qu, Y., Hu, W., and Cheng, G. (2006). Constructing virtual documents for ontology matching. In WWW, pages 23–31.

Rahm, E. and Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. VLDB J., 10(4):334–350.

Shvaiko, P. and Euzenat, J. (2005). A survey of schema-based matching approaches. Journal on Data Semantics IV, 3730:146–171.

Smiljanic, M., van Keulen, M., and Jonker, W. (2006). Using element clustering to increase the efficiency of xml schema matching. In Proceedings of the 22nd International Conference on Data Engineering Workshops, ICDE 2006, page 45.

Stuckenschmidt, H. and Klein, M. C. A. (2004). Structure-based partitioning of large concept hierarchies. In The Semantic Web - ISWC 2004: Third International Semantic Web Conference, volume 3298, pages 289–303, Hiroshima, Japan.

Su, W., Wang, J., and Lochovsky, F. H. (2006a). Holistic query interface matching using parallel schema matching. In Proceedings of the 22nd International Conference on Data Engineering, ICDE 2006, page 122.

Su, W., Wang, J., and Lochovsky, F. H. (2006b). Holistic schema matching for web query interfaces. In Advances in Database Technology - EDBT 2006, 10th International Conference on Extending Database Technology, pages 77–94.

Wang, G., Rifaieh, R., Goguen, J., Zavesov, V., Rajasekar, A., and Miller, M. (2007). Towards user centric schema mapping platform. In International Workshop on Semantic Data and Service Integration, Vienna, Austria.

Wang, Z., Wang, Y., Zhang, S., Shen, G., and Du, T. (2006). Matching large scale ontology effectively. In The Semantic Web - ASWC 2006, First Asian Semantic Web Conference, volume 4185, pages 99–105, Beijing, China.

Xu, R. and Wunsch, D. (2005). Survey of clustering algorithms. Neural Networks, IEEE Transactions on, 16:645– 678.