STUDY OF CHALLENGES AND TECHNIQUES IN LARGE SCALE MATCHING

Sana Sellami, Aicha-Nabila Benharkat
LIRIS-INSA de Lyon, National Institute of Applied Sciences of Lyon, 69621 Villeurbanne, France
sana.sellami@insa-lyon.fr, nabila.benharkat@insa-lyon.fr

Rami Rifaieh
San Diego Supercomputer Center, University of California,La Jolla, California 92093-0505
rrifaieh@sdsc.edu

Youssef Amghar
LIRIS-INSA de Lyon, National Institute of Applied Sciences of Lyon, 69621 Villeurbanne, France
youssef.amghar@insa-lyon.fr

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

Abstract: Matching Techniques are becoming 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 theses 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. In fact, there are many databases and information sources available through the web covering different dynamic domains: Web, Deep Web[1], biology,etc. 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. For instance, GO (Gene Ontology) spans tenth of thousands of concepts arranged in a taxonomic form in (geneontology Website, 2007). UMLS (Unified Medical Language System) covers 476,313 concepts (Bodenreider et al., 1998). Any phylogenetic taxonomy can easily cover thousands of species and descendants.

[1]The deep Web (Bergman, 2001) is qualitatively different from the surface Web. Deep Web sources store their content in searchable databases that only produce results dynamically in response to a direct request.

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 (e.g XML schemas, ontologies, taxonomies, relational database schemas) and returns the semantic similarity values of their elements/attributes. Recently, Matching attracts more attention by researches community. For that, several surveys (Do et al., 2002),(Rahm and Bernstein, 2001),(Shvaiko and Euzenat, 2005)have been proposed covering many of the existing approaches. The survey of (Rahm and Bernstein, 2001) is devoted to a classification of schema Matching approaches and a comparative review of matching systems. As well, the survey in (Shvaiko and Euzenat, 2005) presents a new classification taking into account some novel schema/ontology matching approaches. In our paper, we describe new research works of large scale matching, that differs from (Rahm and Bernstein, 2001) and (Shvaiko and Euzenat, 2005), and we propose survey 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 labor. In contrast, in large-scale data integration scenarios (Madhavan et al., 2007), the Matching process 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. The bigger the input data are, the greater the search space for match candidates will be. In consequence, the quality of Matching (execution time, relevance, etc) will be decreased.

The main motivation of this paper is that current schema Matching techniques don't scale. We survey the existing matching approaches at large scale called holistic and Pair-wise. Moreover, we introduce 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 matching systems and to ensure high outcome for Matching large-scale data. 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 large scale. In section 4, we describe our vision for 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. There are a variety of approaches to study the quality of data in information integration and data search (Kahn et al., 2002),(Gertz et al., 2004),(Peralta et al., 2004),(Burgess et al., 2007). However, there exists little of work which tackles the aspect of quality in the matching process at large scale. In (Bernstein et al., 2004), the authors test their system taking into account the scalability and extensibility criteria. Practically, all matchers are evaluated using precision and recall measures. For example, in (Smiljanic, 2006), the author proposes the evaluation of quality in terms of performance. The performance of a schema matching system consists of efficiency (which expresses how much one system performs faster than the other) and effectiveness (expressed through precision and recall). In (Duchateau et al., 2007), the authors propose quality measures of matching using a number of scoring functions. More specially, the quality of Matching (QoM) is based on the use of quality measures to evaluate the matching system. 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. Therefore, 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 (Peralta et al., 2004) implies the selection of metrics and the implementation of evaluation algorithms that measure and estimate such quality factors. In this respect, a metric is a specific instrument that can be used to measure a given quality factor. We distinguish between two aspects (Fig. 1): the factors that influence the quality and the metrics to evaluate and measure the quality of matching process. We propose (in section 2.1) the factors that mainly depend on the quality of the context (input data and the characteristics of the domain) and the features of matching systems and algorithms. On the other hand, we define the metrics (performance, accuracy, scalability, etc) in term of characteristics of the matching process that builds the resulting data from sources.

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 com-

pleteness, 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 nave 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 (fig.1) could be a support to QoM (Quality of Matching):

- **Performance:** The performance is measured in terms of efficiency and pertinence: Efficiency:It 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 in (Melnik et al., 2002) 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.

Figure 1: Quality of Matching (QoM): factors and metrics

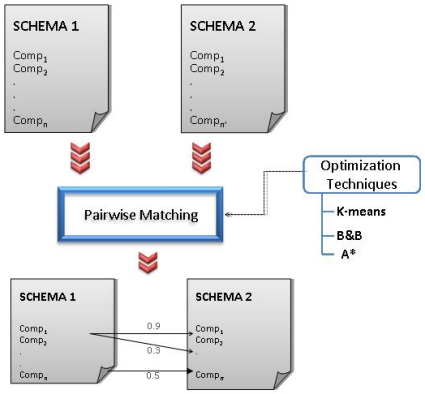

Figure 2: Pair-wise Schema Matching

3 REVIEW OF EXISTING MATCHING APPROACHES

We are interested in our work in Matching techniques that aim at identifying semantic correspondences between schemas, e.g database schemas, ontologies, XML message format, query interfaces, taxonomies, 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 emplyed 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 (Fig. 2)has attracted much attention (Avesani et al., 2005), (Avesani et al., 2007),(Bernstein et al., 2004),(Lu et al., 2005),(Rahm et al., 2004),(Saleem and Z.Bellahsene, 2007),(Smiljanic et al., 2006),(Do and Rahm, 2007) by researches community. We are more interested to the approaches that integrate the clustering and fragmentation techniques (Do and Rahm, 2007),(Rahm et al., 2004),(Smiljanic et al., 2006). In fact, these techniques aim at reducing the dimension of the matching problem. The main purpose is to optimize and improve the quality of the matching (QoM) process. In (Rahm et al., 2004), 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. The main criterion is that a fragment is a rooted sub-graph down to the leaf level in the schema graph. In this way, a fragment is always determined with a node in the schema graph. This approach is done "a priori" before the matcher's execution. The approach is achieved in two matching steps: The first step is the fragments identification of the two schemas that are sufficiently similar and the second step is to match similar fragments. For instance, COMA++ (Aumueller et al., 2005) implements this approach. 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. Moreover, no complexity study exists about fragmentation approach. 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. Then the schema matcher then looks for matchings only within these regions, i.e., clusters. The clustered schema matching is achieved by the clustering algorithm K-means (Xu and Wunsch, 2005). The authors choose an adaptation of the k-means clustering algorithm. Indeed, this choice is based on the simplicity and the non-exponential complexity of the algorithm. Moreover, Clustering was combined with B&B (Branch and Bound) algorithm in (Clausen and Zilinskas, 2002)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 in (Smiljanic et al., 2006) is restricted to 1:1matchings.

3.1.2 Ontology Matching

Ontology matching (Fig. 3)is a promising solution to the semantic heterogeneity problem. It finds correspondences between semantically related entities of the ontologies. These correspondences can be used for various tasks, such as ontology merging, query answering, data translation, or for navigation on the semantic web. Thus, matching ontologies enables interoperation between the knowledge and data expressed in the matched 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 (Ehrig and Staab, 2004),(Hovy, 1998),(Hu and Qu, 2006),(Hu et al., 2006),(Qu et al., 2006),(Stuckenschmidt and Klein, 2004),(Wang et al., 2006a),(Wang et al., 2006b) have been proposed in literature to study the large ontology matching problem. For instance, in (Hu et al., 2006), the authors propose a method for partition-based block matching that is appropriate to large class hierarchies. Large class hierarchies are one of the most common kinds of large-scale ontologies. The two large class hierarchies are partitioned, based on both structural affinities and linguistic similarities, a priori into small blocks respectively. The matching process is then achieved between blocks by combining the two kinds of relatedness found via predefined anchors and virtual documents between them. The partitioning process is realized based on ROCK (Robust Clustering Using Links) algorithm (Guha et al., 1999). ROCK is a robust hierarchical clustering algorithm ROCK that employs links and not distances when merging clusters. However, this approach is not completely applicable to large ontologies and it partitions two large class hierarchies separately without considering the correspondences between them. In addition, it only assumes matchings between classes, thus it is not a general solution for ontology matching. To cope with the large ontologies matching, (Hu and Qu, 2006)propose then 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 (Qu et al., 2006). Partitioning ontologies is achieved by a hierarchical bisection algorithm to provide block mappings. However, classical hierarchical clustering algorithms are not appropriate for large scale data sets due to the quadratic computational complexities in both execution time and store space (Xu and Wunsch, 2005). Another approach has been proposed (Wang et al., 2006a),(Wang et al., 2006b) to deal with large and complex ontologies. The authors propose a Modularization-based Ontology Matching approach (MOM). 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 approach includes sub-steps for large ontology partitioning, finding similar modules, module matching and result combination. 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 and keep the semantics of the original ontology in a specific way E-connection is modeling the union of all domains. However, this approach doesn't discover the complex mappings and doesn't realize the matching between several voluminous ontologies.

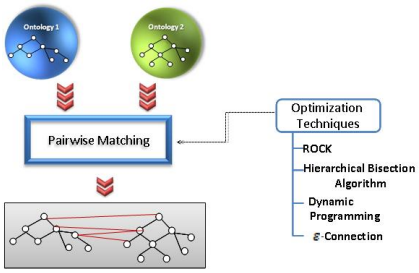

Figure 3: Pair-wise Ontology Matching

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 (Fig. 4) 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. The deep Web refers to World Wide Web content not part of the surface Web indexed by search engines. The data sources in the deep Web are structured and accessible only via dynamic queries instead of static URL links. Several current approaches to holistic schema matching (Chang et al., 2005),(He and Chang, 2006),(He and Chang, 2005),(He et al., 2004),(He and Chang,

2003),(He et al., 2005),(Madhavan et al., 2005),(Pei et al., 2006a),(Pei et al., 2006b),(Su et al., 2006b),(Su et al., 2006a),(Wu et al., 2004) rely on a large amount of data to discover semantic correspondences between attributes. Holistic approach has been introduced in (He et al., 2004),(He and Chang, 2003). The authors in (He and Chang, 2003) propose statistical approaches MGS (for hypothesis modeling, generation, and selection) and DCM (He et al., 2004) (Dual Correlation Mining) framework. The MGS framework 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. This evaluation estimates all possible "models," where a model expresses all attributes matchings. The authors propose also to apply X^2 hypothesis testing to quantify how consistent the schema model is with the data. Nevertheless, this approach doesn't take into consideration complex mappings. DCM framework has been proposed 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 (He et al., 2004),(He and Chang, 2003) by adding sampling and voting techniques, which is inspired by bagging predictors. Specifically, this approach creates a set of matchers, by randomizing input schema data into many independently downsampled trials, executing the same matcher on each trial and then aggregating their ranked results by taking majority voting. The sampling step is achieved a priori and the voting technique is integrated a "posteriori". 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 integrates several steps: matching score calculation that measures the probability of two attributes being synonym, grouping score calculation that estimates whether two attributes are grouping attributes. PSM form parallel schemas by comparing two schemas and deleting their common attributes. HSM and PSM are purely based on the occurrence patterns of attributes and requires neither domain-knowledge nor user interaction. The approaches presented in (Pei et al., 2006a),(Pei et al., 2006b)propose a novel clustering-based approach to schema matching. First, schemas are clustered based on their contextual similarity. Second, attributes of the schemas that are in the same schema cluster are clustered to find attribute correspondences between these schemas. Third, attributes are clustered across different schema clusters using statistical information gleaned from the existing attribute clusters to find attribute correspondences between more schemas. The K-means algorithm has been used to these three clustering tasks and a resampling method (Monti et al., 2003) has been proposed to extract stable attributes from a collection of data. These approaches focused only on 1:1 matchings.

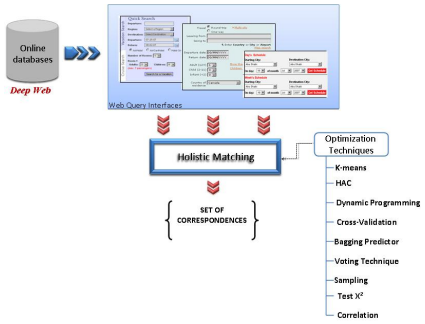

Figure 4: Holistic Matching

3.3 Summary and classification of Matching approaches

In this section, we propose a classification of the previous described approaches in (Fig. 5) 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). (Fig. 5) can be read from two point of views: 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.

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

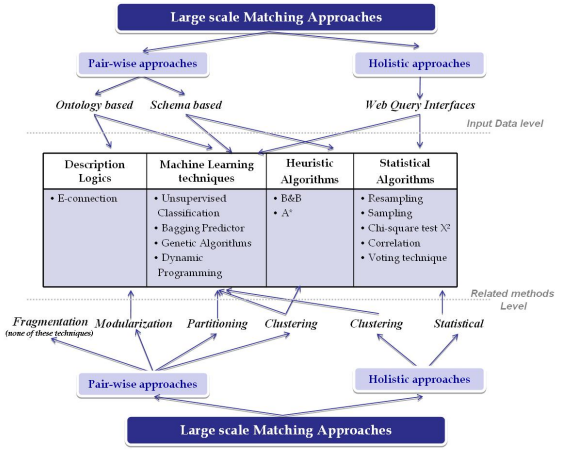

Figure 5: Classification of large scale Matching Approaches

web query interfaces which involve small number of components in the Deep Web (e.g TEL-8, BAMM, etc from UIUC Web Integration repository(datasets, 2003)). Then, Holistic approaches are not 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 realized either in a priori matching or in a posteriori matching. According to figure 5, we can notice that machine learning techniques (e.g unsupervised classification) are the most used.
- 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 (Bernstein et al., 2004),(Aumueller et al., 2005),(Thor 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 it is interesting to combine the holistic and pair-wise approaches. In fact, pair-wise matching is usually achieved between only two voluminous data sources. In contrast to this approach, holistic matching is performed between a set of query interfaces (few components) from the deep web. Using a single pair-wise matcher may be imprecise. 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 representations (schemas, ontologies, taxonomies). With the reduced problem size, we aim to optimize and improve the quality of the matching process (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, regarding few works on quality factors for matching, we consider that is important to integrate a quality evaluation in every step of 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. We have presented in section 2 our definitions for quality metrics relating to 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 large scale matching system. (Fig. 6) outlines a general procedure for matching at large scale. Let a

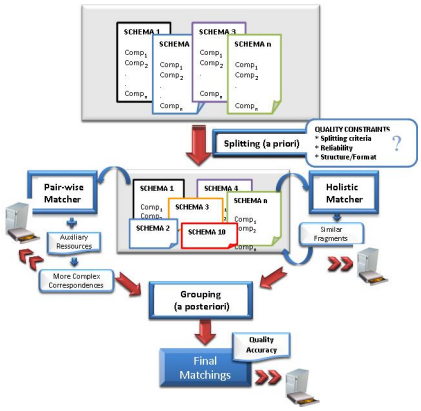

Figure 6: A general Procedure for large scale Matching system

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.

5 CONSLUSION 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. Existing approaches resort to different techniques: machine learning techniques, heuristic algorithms, statistical algorithms, etc. Except, they don't convey the matching in large scale applied for different fields and applications. 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

Aumueller, D., Do, H. H., Massmann, S., and Rahm, E. (2005). Schema and ontology matching with coma++. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 906–908, Baltimore, Maryland, USA.

Avesani, P., Giunchiglia, F., and Yatskevich, M. (2005). A large scale taxonomy mapping evaluation. In *The Semantic Web - ISWC 2005, 4th International Semantic Web Conference*, volume 3729, pages 67–81, Galway, Ireland. Springer.

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*.

Bergman, M. K. (2001). The deep web: Surfacing hidden value.

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.

Bodenreider, O., Nelson, S. J., Hole, W. T., and Chang, H. F. (1998). Beyond synonymy: exploiting the umls semantics in mapping vocabularies. *Proc AMIA Symp*, pages 815–819.

Burgess, M. S., Gray, W. A., and Fiddian, N. J. (2007). Using quality criteria to assist in information searching. *International Journal of Information Quality*, 1(1):83–99.

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.

Clausen, J. and Zilinskas, A. (2002). Subdivision, sampling, and initialization strategies for simplical branch and bound in global optimization. *Computers and Mathematics with Applications*, 44:943–955.

datasets, U. I. (2003). Uiuc.icq datasets.

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.

Ehrig, M. and Staab, S. (2004). Qom - quick ontology mapping. In *The Semantic Web - ISWC 2004: Third International Semantic Web Conference*, pages 683–697, Hiroshima, Japan.

geneontology Website, T. (2007). The geneontology website.

Gertz, M., Özsu, M. T., Saake, G., and Sattler, K.-U. (2004). Report on the dagstuhl seminar: data quality on the web. *SIGMOD Record*, 33(1):127–132.

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

Guha, S., Rastogi, R., and Shim, K. (1999). Rock: A robust clustering algorithm for categorical attributes. In *Proceedings of the 15th International Conference on Data Engineering (ICDE 1999)*, pages 512–521, Sydney, Austrialia.

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. (2005). Making holistic schema matching robust: an ensemble approach. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 429–438, Chicago, Illinois,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.

He, H., Meng, W., Yu, C. T., and Wu, Z. (2005). Wise-integrator: A system for extracting and integrating complex web search interfaces of the deep web. In *VLDB*, pages 1314–1317.

Hovy, E. (1998). Combining and standardizing large-scale, practical ontologies for machine translation and other uses. In Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC). Granada, Spain.

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.

Kahn, B. K., Strong, D. M., and Wang, R. Y. (2002). Information quality benchmarks: product and service performance. *Commun. ACM*, 45(4):184–192.

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.

Melnik, S., Garcia-Molina, H., and Rahm, E. (2002). Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In *Proceedings of the 18th International Conference on Data Engineering (ICDE 2002)*, pages 117–128, San Jose, CA.

Monti, S., Tamayo, P., Mesirov, J. P., and Golub, T. R. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52(1-2):91–118.

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.

Pei, J., Hong, J., and Bell, D. A. (2006b). A robust approach to schema matching overweb query interfaces. In *Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDE Workshops)*, page 46, Atlanta,GA, USA.

Peralta, V., Ruggia, R., Kedad, Z., and Bouzeghoub, M. (2004). A framework for data quality evaluation in a data integration system. In *SBBD*, pages 134–147.

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.

Rahm, E., Do, H. H., and Massmann, S. (2004). Matching large xml schemas. *SIGMOD Record*, 33(4):26–31.

Saleem, K. and Z.Bellahsene (2007). A scalable approach for large-scale schema mediation. pages 26–31.

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

Smiljanic, M. (2006). *XML schema matching: balancing efficiency and effectiveness by means of clustering*. PhD thesis, University of Twente, Zutphen, The Netherlands.

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.

Thor, A. and Rahm, E. (2007). Moma - a mapping-based object matching system. In *CIDR*, pages 247–258.

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. (2006a). Effective large scale ontology mapping. In *Knowledge Science, Engineering and Management, First International Conference, KSEM 2006*, volume 4092, pages 454–465, Guilin, China.

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

Wu, W., Yu, C. T., Doan, A., and Meng, W. (2004). An interactive clustering-based approach to integrating source query interfaces on the deep web. In *SIGMOD Conference*, pages 95–106.

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