

## Semantic heterogeneity measures of unstructured P2P systems

Thomas Cerqueus  
LINA, University of Nantes  
Nantes, France  
thomas.cerqueus@univ-nantes.fr

Sylvie Cazalens  
LINA, University of Nantes  
Nantes, France  
sylvie.cazalens@univ-nantes.fr

Philippe Lamarre  
LINA, University of Nantes  
Nantes, France  
philippe.lamarre@univ-nantes.fr

**Abstract**—We consider P2P data sharing systems in which each participant uses an ontology to represent information. If all the participants do not use the same ontology, the system is said to be semantically heterogeneous. Several methods have been proposed to reach a degree of interoperability but thorough evaluation of these methods is prevented by a lack of tools to describe the situations in which they have been tested. In this paper we identify components that impact on the semantic heterogeneity, and we define several complementary measures to capture the different facets of heterogeneity. Proposed measures allow to characterize the situation in which a method is evaluated, or to measure the heterogeneity reduction produced by another method.

### I. INTRODUCTION

In this paper, we are interested in data sharing peer-to-peer (P2P) systems where each peer is an individual data source. We focus on systems in which each peer uses an ontology to represent its data and its queries. If all the participants use the same ontology, the system is semantically homogeneous. However when the number of peers is important, it is unlikely that they agree on the use of a single ontology. This leads to a situation of semantic heterogeneity of the system. Several methods have been proposed to reach a degree of interoperability using correspondences between ontologies [1]. Most of them are translation-based solutions. Some proposals use similarity between concepts of a same ontology to better answer queries [2]. These approaches can be classified into two classes (non necessarily distinct): those which try to cope with heterogeneity to obtain interoperability (noted CH-methods), and those which try to decrease heterogeneity (noted DH-methods).

Thorough evaluation of these methods is prevented by a lack of tools to describe the situations in which they have been tested (from a semantic viewpoint). It is generally limited to a specific configuration of a given system. Hence the problem is to define elements that might be useful to describe a given semantic state of the distributed system with respect to heterogeneity, just like the system load describes the global amount of work to be treated by the system.

Our approach consists in identifying components that impact on the semantic heterogeneity of a P2P system, and defining several complementary measures to capture the different facets of heterogeneity. We underline that considering the problem from an evaluation perspective allows us,

like for designers of the experiments, to have a global view of the P2P system. In that case, it is possible to assume knowledge that a given peer wouldn't have itself within the system. Proposed measures should allow to characterize the situation in which a given CH-method is evaluated. They should also enable to measure the heterogeneity reduction of a given DH-method.

The paper is organized as follows. Section II presents scenarios showing that heterogeneity is a multifaceted notion. Section III presents the formal model. Section IV defines several complementary measures of semantic heterogeneity. Section V discusses related work. Section VI concludes.

### II. SEMANTIC HETEROGENEITY IS MULTIFACETED

In this section we present three scenarios that aim to identify several facets of heterogeneity.

*Scenario 1:* Let us consider a P2P system in which participants use different ontologies. Assuming data are uniformly distributed, each participant can potentially answer to queries. In this situation, the possibility for a participant to be understood when he sends queries depends on the number of ontologies used in the system, and the number of participants using the same ontology as he does. This scenario shows that it is crucial to consider the number of ontologies on use, and the number of participants.

*Scenario 2:* We now consider a P2P system in which some participants use an ontology  $o_1$  and others use  $o_2$ . In that case, the capacity to interoperate depends on the disparity between the ontologies  $o_1$  and  $o_2$ . More generally, it depends on the disparities between the participants (disparity between their knowledge and their perceptions). In order to measure heterogeneity of the whole system, it is necessary to consider the disparity between the couples of participants.

*Scenario 3:* Here we consider a P2P system in which participants use different ontologies. We also consider that queries are not sent to the whole system: a query issued by  $p$  is sent to a subset of the participants. The possibility for  $p$  to retrieve relevant documents depends on the capacity of its neighbourhood to understand his queries. So in order to measure the difficulty to interoperate, we have to focus on the neighbourhood of each participant by considering the disparity between them and their neighbourhood.

These three scenarios identify different facets that should be taken into account: the contexts on use, the disparities between participants and the organization of the system.

### III. MODEL

#### A. The basic P2P system

An unstructured P2P system is defined by a graph  $S = \langle \mathcal{P}, \mathcal{N} \rangle$ , where  $\mathcal{P}$  is a set of peers (with  $|\mathcal{P}| > 1$ ) and  $\mathcal{N}$  represents a neighbourhood relation. Each element in  $\mathcal{N}$  is an ordered pair  $(p_i, p_j)$  of  $\mathcal{P}$  such that  $p_j$  is one of  $p_i$ 's neighbours.

*Definition 1:* The neighbourhood of a participant  $p$  within a radius  $n$ , denoted by  $\mathcal{N}_n^p$ , is the set of participants accessible from  $p$  with  $l$  hops, where  $1 \leq l \leq n$ . We consider that  $p$  does not belong to its own neighbourhood. In the system of Figure 1,  $\mathcal{N}_2^{p_1} = \{p_2, p_3, p_4, p_5\}$ .

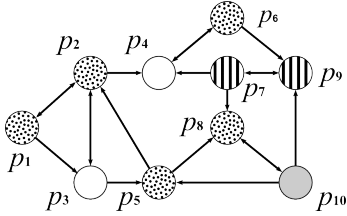


Figure 1. Unstructured P2P system.

#### B. Semantic context of a participant

We consider that an ontology is composed of a set of concepts  $C_o$ , a set of relations  $R_o$  (linking the concepts) and a set of properties  $P_o$  (assigned to the concepts). In practice OWL allows to represent ontologies by defining *classes*, *datatype properties* and *object properties*.

A function measuring the proximity between any two concepts of a same ontology is called an intra-ontology similarity measure:  $sim_o : C_o \times C_o \rightarrow [0, 1]$ . Several measures have been defined in the litterature [3].

*Definition 2:* A semantic context is a couple  $\phi = \langle o, sim_o \rangle$  where  $o$  is an ontology and  $sim_o$  is an intra-ontology similarity measure.

A semantic context enables to express the participant's perception of a domain in a more refined way. Indeed the ontology only reflects the relative organization of the concepts used to model the domain. An intra-ontology similarity brings an additional notion of proximity which expresses how close two concepts are according to the participant.

*Definition 3:* Given a P2P system  $S = \langle \mathcal{P}, \mathcal{N} \rangle$ , a peer-to-context mapping is a function  $\mu : \mathcal{P} \rightarrow \Phi$  mapping each peer to one semantic context.

#### C. Disparity between two semantic contexts

*Definition 4:* A disparity function  $d : \Phi \times \Phi \rightarrow [0, 1]$  is a function that assigns a real value in  $[0, 1]$  to a couple  $\langle \phi, \phi' \rangle$  representing how much  $\phi'$  differs from  $\phi$ . It satisfies the minimality property:  $\forall \phi \in \Phi, d(\phi, \phi) = 0$ .

Some measures presented in the litterature could be used as disparity measure but they are limited to a single component of semantic contexts: the ontology [1] [4]. We think that it may be relevant to consider the differences that come from intra-ontology similarity functions. Indeed some methods of information retrieval use similarity values to extend queries. Thus two participants, using different contexts, could extend a query with different concepts, depending on their intra-ontology similarity. For example, Figure 2 illustrates the case of two participants who rank concepts with respect to their decreasing similarity with the concept *Flower*. Their rankings are different, so they would not extend a query about *Flower* with the same concepts.

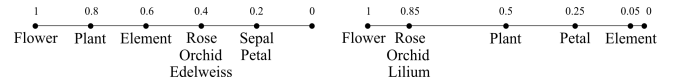


Figure 2. Ranking of the concepts w.r.t. the concept *Flower* where *Rose* is ranked 4<sup>th</sup> with  $sim_{o_1}$  (left) and 2<sup>nd</sup> with  $sim_{o_2}$  (right).

To capture this difference, we introduce the notion of rank of a concept  $c_1$  with regards to another concept  $c$  denoted by  $rank_\phi^{c_1}(c)$  where:  $rank_\phi^{c_1}(c) = |\{s \in S_\phi^c : s \geq sim_o(c_1, c)\}|$ .

For a concept  $c$  having an equivalent,  $S_\phi^c$  is defined by:  $S_\phi^c = \{s \in [0, 1] : \exists c' \in E_o' \text{ such that } sim_o(c, c') = s\}$  where  $E_o'$  designates the set of concepts of  $o$  having equivalents in  $o^1$ . We propose to measure the disorder around each concept. For a concept  $c \in E_o'$ , disorder (denoted by  $dis_{\phi, \phi'}(c)$  and normalized in  $[0, 1]$ ) is defined as:

$$dis_{\phi, \phi'}(c) = \frac{1}{|E_o'|} \sum_{c_0 \in E_o'} \frac{|rank_\phi^c(c_0) - rank_{\phi'}^{eq_{c_0}}(eq_{c_0}^c)|}{\max(|S_\phi^c|, |S_{\phi'}^{eq_{c_0}}|) - 1}$$

Applying  $dis_{\phi, \phi'}$  for each concept of  $o$  having equivalent defines disparity as:

$$d_{disorder}(\phi, \phi') = \frac{1}{|E_o'|} \sum_{c \in E_o'} dis_{\phi, \phi'}(c)$$

#### D. Semantic heterogeneity: definition and typology

*Definition 5:* Let us assume a set  $\mathcal{M}$  of models  $\mathcal{M} = \langle \langle \mathcal{P}, \mathcal{N} \rangle, \langle \Phi, d \rangle, \mu \rangle$  where  $\langle \mathcal{P}, \mathcal{N} \rangle$  is a P2P system,  $\Phi$  is a set of semantic contexts with a disparity function  $d$  and  $\mu$  is a peer-to-context mapping.

A semantic heterogeneity function (or measure) is a function  $\mathcal{H} : \mathcal{M} \rightarrow [0, 1]$  such that:

- $\mathcal{H}(\mathcal{M}) = 0$  if  $|\phi_S| = 1$  (minimality);
- $\mathcal{H}(\mathcal{M}) = 1$  if  $\forall \phi, \phi' \in \phi_S, d(\phi, \phi') = 1$  (maximality).

where  $\phi_S = \{\phi \in \Phi : \exists p \in \mathcal{P} \text{ such that } \mu(p) = \phi\}$ .

Depending on the application domain, several functions might be necessary to capture all the facets of heterogeneity.

<sup>1</sup>We do not formally define the notion of equivalence of two concepts. The interested reader is invited to consult a reference book like [1].

Based on the previous model, we propose a typology of heterogeneity measures. In our view, every measure should consider  $\mathcal{P}$ ,  $\Phi$  and  $\mu$  which are the basic components of the model. Then, we differentiate the measures which are:

- *Structure aware/unaware*: An heterogeneity measure is structure aware if its definition considers the neighbourhood relation  $\mathcal{N}$ . Otherwise it is structure unaware.
- *Disparity aware/unaware*: An heterogeneity measure is disparity aware if its definition considers the disparity function  $d$  on the set of semantic contexts  $\Phi$ . Otherwise, it is disparity unaware.

These two criteria can be combined, leading to four classes of heterogeneity measures. For each class, Table I enumerates the elements of the model that are considered.

Table I  
FOUR CLASSES OF HETEROGENEITY MEASURES.

	Structure unaware	Structure aware
Disparity unaware	$\mathcal{P}, \Phi, \mu$	$\mathcal{P}, \Phi, \mu, \mathcal{N}$
Disparity aware	$\mathcal{P}, \Phi, \mu, d$	$\mathcal{P}, \Phi, \mu, \mathcal{N}, d$

#### IV. MEASURES OF HETEROGENEITY

##### A. Structure unaware measures

1) *Disparity unaware measures*: Notions of richness and evenness are commonly used to measure the heterogeneity of a population (e.g. in biology). Richness is the number of “species” present in a population. Evenness is the relative abundance or proportion of individuals among the “species”. In our context, richness depends on the number of different semantic contexts used in the system. The more contexts there are, the more heterogeneous it is. This idea can be expressed by the following measure:

$$\mathcal{H}_{Rich}(\mathcal{M}) = \frac{|\phi_{\mathcal{S}}| - 1}{|\mathcal{P}| - 1}$$

where  $|\phi_{\mathcal{S}}|$  is the number of different contexts used in the system  $\mathcal{S}$ , and  $|\mathcal{P}|$  is the number of participants.

*Example 1*: In the system presented on Figure 1, four different contexts are used by ten participants:  $\mathcal{H}_{Rich}(\mathcal{M}) = \frac{4-1}{10-1} = 0.33$ . The richness measure does not give any indication on how contexts are distributed. However it is important to differentiate cases where contexts are evenly distributed across the system from cases where several contexts are used only once. To capture this aspect we can adapt the Simpson diversity index [5]:

$$D = \frac{1}{|\mathcal{P}|^2} \sum_{\phi_i \in \phi_{\mathcal{S}}} |\mathcal{P}_{\phi_i}|^2$$

where  $\mathcal{P}_{\phi_i}$  is the set of peers using  $\phi_i$ . We define  $\mathcal{H}_{Even}$  as:

$$\mathcal{H}_{Even}(\mathcal{M}) = \frac{|\mathcal{P}| \cdot (1 - D)}{|\mathcal{P}| - 1}$$

*Example 2*: On the system presented on Figure 1,  $D = 0.34$ . Given  $|\mathcal{P}| = 10$ , we find  $\mathcal{H}_{Even}(\mathcal{M}) = 0.73$ .

If  $\mathcal{H}_{Even}$  is close to 1, we can assert that some participants do not share their semantic context with anyone while others do share it with many others. Measures  $\mathcal{H}_{Rich}$  and  $\mathcal{H}_{Even}$  are complementary because they capture two aspects of the heterogeneity. Indeed a system can be rich (i.e. a lot of different contexts are used) and even (i.e. contexts are used in equal number), or poor and uneven, etc.

2) *Disparity aware measure*: On top of determining diversity, it is interesting to take into account disparity between contexts of the system. Indeed diversity measures do not make any difference between a system  $\mathcal{S}_1$  using  $\eta$  contexts between which disparities are weak, and a system  $\mathcal{S}_2$  using  $\eta$  contexts between which disparities are important. We propose to consider the disparity between participants rather than only consider the contexts they use. If the disparity between participants is globally important, it means that participants have important knowledge differences. As we do not take into account the system topology, we consider the disparity between each pair of participants:

$$\mathcal{H}_{Disp}(\mathcal{M}) = \frac{1}{|\mathcal{P}|^2 - |\mathcal{P}|} \sum_{p_i \neq p_j \in \mathcal{P}} d(\mu(p_i), \mu(p_j))$$

It determines if peers are globally disparate from each other.

##### B. Structure aware measures

In an heterogeneous P2P system, it is interesting to consider the participants’ neighbourhood. If participants are globally far (semantically speaking) from their respective neighbourhoods, the system is highly heterogeneous. Starting from a participant  $p$ ’s neighbourhood  $\mathcal{N}_n^p$ , we propose several measures.

1) *Disparity unaware measure*: First we can number the participants that do not use the same semantic context as  $p$ :

$$\mathcal{H}_{Rap}^n(\mathcal{M}, p) = \frac{|\{p_i \in \mathcal{N}_n^p : \mu(p_i) \neq \mu(p)\}|}{|\mathcal{N}_n^p|}$$

This measure gives basic information about a participant’s neighbourhood, and could eventually be calculated by a participant itself. Indeed, it just requires to be able to determine if another participant uses the same context.

*Example 3*: In Figure 1,  $\mathcal{N}_2^{p_1} = \{p_2, p_3, p_4, p_5\}$ . As  $p_3$  and  $p_4$  do not use the same context as  $p_1$ , we find that  $\mathcal{H}_{Rap}^2(\mathcal{M}, p_1) = \frac{2}{4} = 0.5$ .

We can use  $\mathcal{H}_{Rap}$  to get a global measure:

$$\mathcal{H}_{RapAvg}^n(\mathcal{M}) = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \mathcal{H}_{Rap}^n(\mathcal{M}, p)$$

2) *Disparity aware measure*: The fact that two participants do not use the same context does not induce that they

can not communicate together. So we refine the previous measure by considering a disparity measure:

$$\mathcal{H}_{Dap}^n(\mathcal{M}, p) = \frac{1}{|\mathcal{N}_n^p|} \sum_{p_i \in \mathcal{N}_n^p} d(\mu(p), \mu(p_i))$$

This measure focuses on a particular participant and determines how this latter is understood by its neighbours.

*Example 4:* In Figure 1,  $\mathcal{N}_2^{p_1} = \{p_2, p_3, p_4, p_5\}$ . As  $p_1, p_2$  and  $p_5$  use the semantic context  $\phi$ , and  $p_3$  and  $p_4$  use  $\phi'$ , we find:  $\mathcal{H}_{Dap}^2(\mathcal{M}, p_1) = \frac{3}{5}d(\phi, \phi) + \frac{2}{5}d(\phi, \phi')$ .

As for the disparity unaware measure, a global measure  $\mathcal{H}_{DapAvg}^n$  can be obtained (cf. definition of  $\mathcal{H}_{RapAvg}^n$ ). If  $\mathcal{H}_{DapAvg}^n$ 's value is weak, it means that participants are surrounded by participants able to "understand" them.

*Proposition 1:* All the proposed measures satisfy both properties of minimality and maximality (proofs are trivial).

3) *Using heterogeneity measures to evaluate system organization:* Some measures defined previously enable to determine if a participant is well located in a system with regards to its neighbourhood. Intuitively, the neighbourhood of a participant  $p$  is "favorable" if it is composed of the participants from whom he is close semantically. The neighbourhood of a participant is favorable if considering a bigger neighbourhood increases the heterogeneity (around  $p$ ). Given  $n$  hops, the neighbourhood of  $p$  is favorable if:

$$\forall i \leq n, \forall j > n, \mathcal{H}^i(\mathcal{M}, p) \leq \mathcal{H}^j(\mathcal{M}, p)$$

where  $\mathcal{H}$  is an heterogeneity measure centered on a participant (e.g.  $\mathcal{H}_{Rap}$  or  $\mathcal{H}_{Dap}$ ). Having a condition to determine if a participant's neighbourhood is favorable allows to determine how hard it will be for this participant to be understood: it gives information on its capacity to interoperate.

Given two systems  $\mathcal{S}_1$  and  $\mathcal{S}_2$ , we can say that  $p$ 's neighbourhood is more favorable in  $\mathcal{S}_1$  than in  $\mathcal{S}_2$  if:

$$\forall i \in \llbracket 1, n \rrbracket, \mathcal{H}^i(\mathcal{M}_1, p) \leq \mathcal{H}^i(\mathcal{M}_2, p)$$

We can compare two systems' organization both way: we can use global measures, or we can rely on conditions relative to participants' placement. For instance, we could say that  $\mathcal{S}_1$  is better organized than  $\mathcal{S}_2$  if each participant's neighbourhood is more favorable in  $\mathcal{S}_1$  than in  $\mathcal{S}_2$ .

## V. RELATED WORK

Our work assumes the existence of a disparity measure between two semantic contexts. Distance measures proposed in the field of ontology matching can be adapted and used. [1] and [4] present a number of similarity measures between two ontologies based on terms associated to concepts and on the hierarchical structure of ontologies. In [6] authors propose measures of similarity between ontologies in the alignment space (this latter is defined as a set of ontologies, and a set of alignments between these ontologies). This work can be adapted to define disparity between two participants

of a P2P system, but it does not aim to characterize heterogeneity of the whole system/space. All these measures can be used in our work if we assume that the participants' semantic contexts are only made of ontologies.

In [7] authors define a necessary condition for semantic interoperability in P2P systems, but do not propose any measures. They assume that interoperability is ensured by the presence of correspondences between the different schemas (or ontologies) on use. In this context, two peers are said to be semantically interoperable if translation links exist between each other. Nevertheless this condition does not give information about the translation quality.

## VI. CONCLUSION

In this paper we defined several complementary measures to capture different facets of heterogeneity. They are divided into different classes depending on the fact that they exploit the system topology and/or disparity measures. The proposed measures are meant to be used in an evaluation context, this is why we assumed global knowledge of the system. Depending on the application the designer of the experiments should choose those measures which, together best characterize the system semantic heterogeneity. They can also be used to instantiate P2P systems with specific (semantic) characteristics. Obviously, the proposed measures should be validated through extensive experimentations.

As future work, we plan to propose algorithms that reduce heterogeneity and to evaluate them in different situations of heterogeneity according to the proposed measures.

## REFERENCES

- [1] J. Euzenat and P. Shvaiko, *Ontology matching*. Springer-Verlag, 2007.
- [2] A. Ventresque, S. Cazalens, P. LAMARRE, and P. VALDURIEZ, "Improving interoperability using query interpretation in semantic vector spaces," in *5th European Semantic Web Conference (ESWC)*, 2008, pp. 539–553.
- [3] H. Seddiqui and M. Aono, "Metric of intrinsic information content for measuring semantic similarity in an ontology," in *7th Asia-Pacific Conference on Conceptual Modelling (APCCM)*, 2010, pp. 89–96.
- [4] J. David and J. Euzenat, "Comparison between ontology distances (preliminary results)," in *7th International Conference on The Semantic Web (ISWC)*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 245–260.
- [5] E. Simpson, "Measurement of diversity," *Nature*, vol. 163, p. 688, 1949.
- [6] J. David, J. Euzenat, and O. Šváb Zamazal, "Ontology similarity in the alignment space," in *9th International Semantic Web Conference (ISWC)*, 2010.
- [7] P. Cudré-Mauroux and K. Aberer, "A necessary condition for semantic interoperability in the large," in *3rd International Conference on Ontologies, Databases and Applications of Semantics (ODBASE)*, 2004, pp. 859–872.