Concept-based query transformation based on semantic centrality in semantic peer-to-peer environment

Jason Jung, Antoine Zimmermann, Jérôme Euzenat

Concept-Based Query Transformation Based on Semantic Centrality in Semantic Peer-to-Peer Environment

Jason J. Jung[1,2], Antoine Zimmerman[2], and Jérôme Euzenat[2]

[1] Inha University, Korea
j2jung@intelligent.pe.kr
[2] INRIA Rhône-Alpes, France
{Antoine.Zimmerman,Jerome.Euzenat}@inrialpes.fr

Abstract. Query transformation is a serious hurdle on semantic peer-to-peer environment. The problem is that the transformed queries might lose some information from the original one, as continuously traveling p2p networks. We mainly consider two factors; i) number of transformations and ii) quality of ontology alignment. In this paper, we propose semantic centrality (SC) measurement meaning the power of semantic bridging on semantic p2p environment. Thereby, we want to build semantically cohesive user subgroups, and find out the best peers for query transformation, i.e., minimizing information loss. We have shown an example for retrieving image resources annotated on p2p environment by using query transformation based on SC.

1 Introduction

Information retrieval process on the p2p networks has been performed by propagating a certain message containing a certain queries to neighbor peers and their neighbors. We assume that the queries for interactions between peers (from source peer to destination peer) are simply represented as a set of concepts derived from the ontology of source peer. For high accessibility, the queries can be transformed into the concepts of destination peer ontology. The concepts in the original query can be replaced to the correspondent concepts resulting from ontology alignment between peer ontologies. More importantly, we propose a novel measurement of semantic centrality (SC), which expresses the power of controlling *semantic* information on semantic p2p network, and show that it is applied to search for the most proper peers for concept-based query transformation.

Thereby, in this study, we introduce a three-layered structure[1] made of superposed networks that are assumed to be strongly linked:

Social layer relating peers (or people) on the basis of common interest;
Ontology layer relating ontologies on the basis of explicit import relationships or implicit similarity;

[1] Please refer to [1] for more description in detail.

G. Dong et al. (Eds.): APWeb/WAIM 2007, LNCS 4505, pp. 622–629, 2007.

Concept layer relating concepts on the basis of explicit ontological relationships or implicit similarity.

We may call this stack of interlinked networks a semantic social space.

Generally, the networks will be characterized here as a set of objects (or nodes) and a set of relations. A network $\langle N, E^1, \ldots E^n \rangle$ is made of a set N of nodes and n sets of object pairs $E^i \subseteq N \times N$ the set of relations between these nodes. These networks can express the relationships between people or many other sort of items. As usual, a path p between node e and e' is a sequence of edges $\langle e_0, e_1 \rangle$, $\langle e_1, e_2 \rangle$, $\ldots$, $\langle e_{k-1}, e_k \rangle$ in which $e_0 = e$ and $e_k = e'$. The length of a path is its number of edges (here k) and the shortest path distance $spd(e, e')$ between two nodes e and e' is the length of the shortest path between them. By convention, $spd(e, e) = 0$.

Definition 1 (Distance network). *A distance network $\langle N, E^1, \ldots, E^n \rangle$ is made of a set N of nodes and n sets of distance functions $E^i : N \times N \longrightarrow [0\ 1]$ defining the distance between nodes (so satisfying symmetry, positiveness, minimality, and triangular inequality).*

It is clear that any network is a weighted network which attributes either 0 or 1 as a weight. The definition of social network analysis can be adapted to distance networks if each time the cardinality of a set of edges if used, it is replaced by the sum of its distances. The distance of a path is obtained by summing the distances of its edges.

In the three-layered model we design to propagate the relational information (e.g., the distance or similarity) not only within a layer but also between layers. We have provided the principles for extracting similarity between concepts in different ontologies and propagating this similarity to a distance and an alignment relation between ontologies. We compute semantic affinities between peers, so that the semantic subgroups can be discovered. By using topological features of the discovered subgroups, two centrality measurements (e.g., local and global centralities) can be obtained. Finally, these centralities are applied to determine the best path on which the queries can travel in p2p network.

2 Inferring Relationships

The numerous relationships that can be found by construction of the concept layer, new relationships can be inferred between the entities. One particularly interesting relationship is similarity: in order to find relationship between concepts from different ontologies, identifying the entities denoting the same concept is a very important feature. As a matter of fact, most of the matching algorithms use some similarity measure or distance in order to match entities.

A distance between two ontologies can be established by finding a maximal matching maximising similarity between the elements of this ontology and computing a global measure which can be further normalised:

Definition 2 (Ontology distance). *Given a set of ontologies N_O, a set of entities N_C provided with a distance function $E_C^{dist} : N_C \times N_C \longrightarrow [0\ 1]$ and a relation $Defines$:*

$N_O \times N_C$, the distance function $E_O^{dist} : N_O \times N_O \longrightarrow [0\ 1]$ is defined as:

$$E_O^{dist}(o, o') = \frac{max(\sum_{\langle c,c' \rangle \in Pairing(Defines(o),Defines(o'))} E_C^{dist}(c, c'))}{max(|Defines(o)|, |Defines(o')|}$$

The resulting measure is minimal ($\forall o \in N_O, E_O^{dist}(o, o) = 0$), but is not guarantee to be a distance unless we apply a closure with the triangular inequality.

This is the measure that is used in the OLA algorithm for deciding which alignment is available between two ontologies [2]. However, other distances can be used such as the well known single, average and multiple linkage distances.

This ontology distance introduces a new relation on the ontology layer which provides a good idea of the distances between ontologies. It is, in turn, a clue of the difficulty to find an alignment between ontologies. It can be used for choosing to match the closest ontologies with regard to this distance. This can help a newcomer in a community to choose the best contact point: the one with whom ease of understanding will be maximised.

It can however happen that people have similar but different ontologies. In order for them to exchange their annotations, they need to know the alignments existing within the ontology network. As the result of applying alignment algorithms, the similarity or distance on the network is the basis for many matching algorithms [2]. Manually extracted alignments can also be added to this relation.

As a result, from concept similarity these algorithms will define a new relation E^{align} at the ontology level.

Definition 3 (Alignment relation). *Given a set of ontologies N_O, a set of entities N_C provided with a relation $E_C^{dist} : N_C \times N_C$, and a matching algorithm $Match$ based on E_C^{dist}, the alignment relation $E^{align} \subseteq N_O \times N_O$ is defined as:*

$$\langle o, o' \rangle \in E^{align} \text{ iff } Match(o, o') \neq \emptyset$$

If one has a measure of the difficulty to use an alignment or of its quality, this network can also be turned into a distance network on which all these measures can be performed. Of course, when an alignment exists between all the ontologies used by two peers, there is at least some chance that they can talk to each others. This can be further used in the social network.

This new relation in the ontology layer allows a new agents to choose the ontology that it will align with first. Indeed, the ontologies with maximal hub centrality and closeness for the alignment network are those for which the benefit to align to will be the highest because they are aligned with more ontologies. In the peer-to-peer sharing application, choosing such an ontology will bring the maximum answers to queries.

This is the occasion to note the difference between the relations in the same network: in the ontology network, the hub ontologies for the import relation are rather complete ontologies that cover many aspects of the domains, while hub ontologies for the E^{align} relation are those which will offer access to more answers.

Once these measures on ontologies are obtained, this distance can be further used on the social layer. As we proposed it is possible to think that people using the same ontologies should be close to each other. The affinity between people can be measured from the similarity between the ontology they use.

Definition 4 (Affinity). *Given a set of people N_S, a set of ontologies N_O provided with a distance $E_O^{dist} : N_O \times N_O \longrightarrow [0\ 1]$ and a relation $Uses : N_S \times N_O$, the affinity is the similarity measure defined as*

$$E^{aff}(p, p') = 1 - \frac{max\left(\sum_{\langle o,o'\rangle \in Pairing(Use(p),Use(p'))} 1 - E_O^{dist}(o, o')\right)}{max(|Use(p)|, |Use(p')|)} \quad (1)$$

Since this measure is normalised, it can be again converted to a distance measure through complementation to 1.

Introducing the distance corresponding to affinity in the social network allows to compute the affinity relationships between people with regard to their knowledge structure. Bottom-up inference from $\mathcal{C}$ allows to find out the semantic relationships between users based on this space.

3 Transformation Path Selection

Affinity measurements between people (in Equ. 1) can play a role of the strength of social tie on a semantic social network. Then, we can apply various social network analysis methods to discover meaningful patterns from the social layer $\mathcal{S}$. In this study, by using cohesive subgroups (communities) identification [3], the linkages on the p2p network should be re-organized to discriminate which peers are more proper to support interoperability among peers.

Basically, the interactions between peers are based on exchanging messages, including either a certain query or answer sets. To make queries understandable on heterogeneous peers, the queries have to be transformed with referring to the corresponding peer ontologies. The peer sending queries should select some other neighbor peers to ask query transformation with their own peer ontologies.

Definition 5 (Query). *A query q can be embedded into a message $\langle p_{src}, p_{dest}, q\rangle$ sent from peer p_{src} to p_{dest}. The ontologies of two peers are denoted as $o_{src} = Use(p_{src})$ and o_{dest}. The query grammar is simply given by $q ::= c|\neg q|q \wedge q|q \vee q$ where $c \in Define(o)$.*

In this study, we are interested in queries consisting of a set of concepts from the peer ontologies, so that the queries can be transformed by *concept replacement* strategy based on correspondences discovered by ontology alignment.

Definition 6 (Correspondence). *A set of correspondences discovered ontology alignment process between two ontologies o_i and o_j is given by*

$$\{\langle c_i, c_j, rel\rangle | E^{align}(o_i, o_j), c_i \in Define(o_i), c_j \in Define(o_j)\} \quad (2)$$

where rel indicates a relation between two classes (e.g., equivalence, subclass, superclass, and so on).

For example, if there exist correspondences $\{\langle c_\alpha^1, c_\beta^3, =\rangle, \langle c_\alpha^2, c_\beta^4, =\rangle\}$ between peer ontologies o_α and o_β, a peer query "$q_\alpha = c_1 \vee c_2$" from α can be transformed to "$q_\beta = c_3 \vee c_4$" for β.

However, we have to deal with the problems;

- what if the correspondences are not enough to transform the queries sent?
- which peers can efficiently help this transformation process?

Thereby, main scheme of our approach is to find out the best transformation path, minimizing information loss from ontology alignment process. In order to reduce information loss caused by ontology mismatching during transforming queries, we can intuitively consider two heuristic criterion; i) minimizing the number of transformations (or length of transformation path), and ii) maximizing the semantic similarities (or correspondences) with neighbors. Instead of meeting these two objectives, we focus on searching for the most *powerful* peer, most likely to help them communicate with each other.

3.1 Measuring Semantic Centrality

When sending a query on semantic p2p network, we need to find out which peer (more exactly, peer ontology) is most useful to transform the query for interoperability between source and destination peer. Thereby, SC of each peer is measured by peer ontology alignment. By mapping peer ontologies, consensual ontology can be built and applied to semantic community identification.

Based on the strengths of social ties E^{aff} between pairs of peers, we can apply a non-parametric approach, e.g., nearest neighborhood method [4]. As extending [3], this task is to maximize "semantic" modularity function $Q^{\diamond}$ on social layer $\mathcal{S}$. With the number of communities k predefined, we find out that the given peer set in a social layer $\mathcal{S}$ can be partitioned into a set of communities (or subgroup) $\mathcal{G} = \{g_1, \ldots, g_k\}$. The users can be involved in more than one community. It means that a certain peer p in g_i can also be taken as one of members of g_j, because the semantics in his ontology is relatively close to both consensus semantics of g_i and g_j. Thus, the modularity function $Q^{\diamond}$ is formulated by $Q^{\diamond}(\mathcal{S}) = \sum_{i=1}^{k} \frac{\sum_{p_a \in g_i, p_b \in g_i} E^{aff}(p_a, p_b)}{|g_i|}$. The only pairs of peers where $E^{aff}(p_a, p_b) \geq \tau_{aff}$ should be considered. Thus, $\mathcal{G}(\mathcal{S})$ can be discovered when $Q^{\diamond}(\mathcal{S})$ is maximized. For computing this, in this paper, we applied an iterative k-nearest neighborhood methods. As changing k, consequently, the social layer is hierarchically re-organized.

Generally, centrality measures of a user are computed by using several features on the social network, and applied to determine the structural power. So far, in order to extract the structural information from a given social network, various measurements such as centrality [5], pair closeness [6], and authoritative [7] have been studied to realize the social relationships among a set of users. Especially, the centrality can be a way of representing the geometrical power of controlling *information flow* among participants on p2p network.

We define two kinds of semantic centralities, with respect to the scope and the topologies of communities;

- *Local* semantic centrality $C_L^{\diamond}$, meaning the power of semantic bridging between the members within the same community, and
- *Global* semantic centrality $C_G^{\diamond}$, implying the power of bridging for a certain target community.

Local SC of peer $p \in g_i$ is easily measured by $C_L^{\diamond}(p, g_i) = \frac{\sum_{p,p' \in g_i, p \neq p'} E^{aff}(p,p')}{|g_i|}$, because we are concerning only $E^{aff}(p_a, p_b) \geq \tau_{aff}$ and regarding them as most potential transformation paths. This is similar to the closeness centrality.

On the other hand, global SC $C_G^{\diamond}$ of peer $p \in g_i$ toward a certain target community g_X is based on three factors; i) the number of available transformation paths (s.t. $E^{aff} \geq \tau_{aff}$), ii) the strength of each path E^{aff}, and iii) the local SC of the peer in target community. Thus, we formulate it as three different ways;

$$C_G^{\diamond}(p, g_X) = \frac{\sum_{p' \in g_X} E^{aff}(p,p') \times C_L^{\diamond}(p', g_X)}{|g_X|} \tag{3}$$

$$= \frac{\left[\max_{p' \in g_X} E^{aff}(p,p')\right] \times C_L^{\diamond}(p', g_X)}{|g_X|} \tag{4}$$

$$= \frac{\max_{p' \in g_X} \left[E^{aff}(p,p') \times C_L^{\diamond}(p', g_X)\right]}{|g_X|} \tag{5}$$

While Equ. 3 can take into account all possible paths to taget community by measuring the average centrality, Equ. 4 and Equ. 5 are focusing on only the maximum affinity path. We empirically evaluated these three different heuristic functions in Sect. 4.

3.2 Query Transformation Strategy

We establish query transformation strategy in accordance with the semantic position of peers in social layer $\mathcal{S}$. Query transformation between heterogeneous peers should be conducted by referring to the following strategies;

- If the peers p and p' are located in a same semantic community, a set of transformation paths $TP_L(p, p')$ between them can be evaluated (or ranked) by $\frac{\sum_{p'' \in TP_L} C_L^{\diamond}(p'')}{\exp(1+|TP_L|)}$ where p'' is on the transformation path TP_L. It means the best transformation path has to be chosen, as the length of the path is shorter and local semantic centralities of the peers on the path are higher.
- If the peers $p_i \in g_i, p_j \in g_j$ are in different semantic communities, a set of transformation paths $TP_G(p_i, p_j)$ between them can be evaluated (or ranked) by $TP_{L_i}(p_i, p_i') + C_G^{\diamond}(p_i', g_j) + TP_{L_j}(p_j', p_j)$, and this can be expanded as $\frac{\sum_{p_i'' \in TP_{L_i}} C_L^{\diamond}(p_i'')}{\exp(1+|TP_{L_i}|)} + C_G^{\diamond}(p_i', g_j) + \frac{\sum_{p_j'' \in TP_{L_j}} C_L^{\diamond}(p_j'')}{\exp(1+|TP_{L_j}|)}$. A global transformation path is decomposed into two local transformation path and a transformation path with best global centrality. Exceptionally, when there is no path between communities, the social layer should be re-organized as decreasing the number of communities k.

Thereby, the best transformation path have to be selected by comparing all candidate ones.

4 Experimental Results

In order to evaluate the proposed approach, we invited seven students and asked them to annotate a given set of images by referring to any other standard ontologies (e.g.,

SUMO, WordNet and ODP). While annotating the images, we could collect peer ontologies for building semantic social space.

4.1 From Peer Ontologies to Social Ties

Here, we want to show the experimental results of building social semantic space by ontology alignment. They are compared with simple co-occurrence patterns between the annotated images by Mika's social centrality C_M [8], which is formulated by $C_M(U_i) = \frac{\sum \frac{\cap_{k=1,k\neq i}^{|U|}(\mathcal{R}_{U_k}, \mathcal{R}_{U_i})}{\mathcal{R}_{U_i}}}{|U|-1}$ where $|U|$ is the total number of peers (or people) on social network. The results are shown in Table 1. We found out that the number

Table 1. Experimental results of a) closeness centrality by co-occurrence patterns, and b) semantic affinity E^{aff} and centrality in semantic social network

(a/b)	AS	AZ	FAK	JE	JJ	JP	SL	C_M	$C_L^{\diamond}$
AS	-	0.98/0.65	0.62/0.33	0.94/0.73	1.00/0.26	0.60/0.32	0.23/0.62	0.73	0.49
AZ	0.98	-	0.62/0.49	**0.94/0.825**	0.98/0.31	0.62/0.3	0.26/0.52	0.73	0.52
FAK	0.78	0.78	-	0.70/0.57	0.78/0.28	0.54/0.22	0.30/0.32	0.65	0.37
JE	0.90	**0.90**	0.53	-	0.90/0.46	0.57/0.49	0.16/0.75	0.66	**0.64**
JJ	1.00	0.98	0.62	0.94	-	0.60/0.72	0.23/0.39	0.73	0.40
JP	0.93	0.97	0.67	0.93	0.93	-	0.13/0.51	**0.76**	0.43
SL	0.44	0.48	0.44	0.32	0.44	0.16	-	0.38	0.52

of annotated resources are barely related to the social centrality. SL annotated the least number of resources, so that his centrality also lowest among people. But, even though JE's annotations were the largest one, JP has shown the most powerful centrality.

4.2 Heterogeneous Query Processing

From the organized three groups $g_A = \{JE, AZ\}$, $g_B = \{JJ, JP\}$, and $g_C = \{AS, FAK, SL\}$ (the number of communities $k = 3$), we compared the image results retrieved by ten concept-based queries generated by every peers, according to the transformation strategies. In Table 2, we show "Precision" performance, because we are emphasizing the information loss effected from query transformation. We found out that Equ. 3 has outperform the others by about 19 % and 11%.

Table 2. Precision performance on query transformation strategies; stp means the simple shortest path on social layer

	g_A				g_B				g_C			
	stp	Equ. 3	Equ. 4	Equ. 5	stp	Equ. 3	Equ. 4	Equ. 5	stp	Equ. 3	Equ. 4	Equ. 5
g_A	0.72	0.75 by Local SC			0.36	0.71	0.57	0.64	0.36	0.74	0.59	0.67
g_B	0.317	0.67	0.54	0.6	0.64	0.69 by Local SC			0.34	0.78	0.62	0.7
g_C	0.425	0.63	0.51	0.57	0.41	0.68	0.54	0.64	0.685	0.67 by Local SC		

5 Discussion and Concluding Remark

Semantic overlay network Various applications (Edutella, Bibster, and Oyster) for sharing resources on p2p network have been released. Most similarly, semantic overlay network [9] concerns query processing for information sharing on p2p network, but it is based on simple keyword matching to estimate the relationships between nodes.

As another important issue, we want to carefully discuss information loss by semantic transformation. While equivalent correspondences (e.g., $\langle c, c', = \rangle$) are acceptable, subsumption correspondences make the transformed queries more specific, and the resources retrieved from peers may (possibly) show higher precision and lower recall results.

As a conclusion, in this paper, we claim a new centrality measurement for providing query-based interactions on p2p network. Especially, we found out very efficient transformation path selection mechanism (e.g., Equ. 3). Moreover, by peer ontology alignment, consensus ontology has been built and applied to identify some semantic communities. We believe that it will play a role of generating semantic geometry to quantify social roles on p2p network.

References

1. Jung, J.J., Euzenat, J.: From personal ontologies to semantic social space. In: Poster of the 4th European Semantic Web Conference (ESWC 2006). (2006)
2. Euzenat, J., Valtchev, P.: Similarity-based ontology alignment in OWL-Lite. In de Mántaras, R.L., Saitta, L., eds.: Proceedings of the 16th European Conference on Artificial Intelligence (ECAI'2004), Valencia, Spain, August 22-27, 2004, IOS Press (2004) 333–337
3. Newman, M.E.J.: Fast algorithm for detecting community structure in networks. Physical Review E **69** (2004) 066133
4. Gowda, K.C., Krishna, G.: Agglomerative clustering using the concept of mutual nearest neighbourhood. Pattern Recognition **10**(2) (1978) 105–112
5. Haga, P., Harary, F.: Eccentricity and centrality in networks. Social Networks **17**(1) (1995) 57–63
6. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proceedings of the National Academy of Sciences **99** (2002) 7821–7826
7. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of ACM **46**(5) (1999) 604–632
8. Mika, P.: Ontologies are us: A unified model of social networks and semantics. In Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A., eds.: Proceedings of the 4th International Semantic Web Conference (ISWC 2005), November 6-10, 2005. Volume 3729 of Lecture Notes in Computer Science., Springer (2005) 522–536
9. Crespo, A., Garcia-Molina, H.: Semantic overlay networks for p2p systems. In Moro, G., Bergamaschi, S., Aberer, K., eds.: Proceedings of the 3rd International Workshop Agents and Peer-to-Peer Computing (AP2PC 2004), July 19, 2004. Volume 3601 of Lecture Notes in Computer Science., Springer (2005) 1–13