Two Tier Ontology Alignment

First A. Hongzhe Liu, Second B. Hong Bao, and Third C. Junkang Feng

Abstract—It is desirable to distribute data and knowledge of a huge amount of cultural heritage and to make them available readily to people, but the distributed, heterogeneous and autonomous nature of the databases of digital museums gives rise to the challenge of achieving the best retrieval results in cross-system searching. To make this difficult task tractable, we analyze semantic heterogeneities among these data sources and describe an two tier (concept and instance) approach to accomplish ontology alignment by using the information flow theory.

Keywords—Digital museum, information flow, ontology alignment, precision, recall.

I. INTRODUCTION

In the open environment of the Internet and the Web, information resources are heterogeneous and are indexed with different vocabularies and organized according to different schemes. How to achieve the best retrieval results in cross-domain searching has presented a particular challenge to the information profession. In information retrieval, users typically are neither, nor should they need to be, aware of the behind-the-scenes mechanisms for matching their query terms to the vocabularies employed by various systems. The ideal approach would be to provide a "one-stop" seamless searching instead of requiring the user to search individual databases or collections separately. To enable such an approach, it is important to render the different knowledge organization systems interoperable.

Chinese civilization has been around for thousands of years, and as a result a huge amount of cultural heritage and antiques are scattered all over the vast territory of China. All kinds of digital museums have been developed for them. Each of these museums maintains large digital archives of their collections. They represent an extremely valuable cultural heritage resource, and yet access and exploitation of the data is constrained due to the distributed and heterogeneous nature of the resource. Therefore it is highly desirable to find an avenue to integrating these distributed and heterogeneous systems.

The remainder of this paper is organized as follows: Section 2 describes a motivating example in digital museum domain. Section 3 defines the ontology mapping. Section 4 lists the related works. Section 5 proposes our solution. Finally section 6 summarizes our contribution of the work and gives a conclusion.

II. MOTIVATING EXAMPLE

Given two digital museums 1 and 2 with local ontology alignment as the follows:

![Two Tier Ontology Alignment](image)

Table 1 Instances of digital museum 1

<table>
<thead>
<tr>
<th>Bronze Ware title</th>
<th>Dynasty (instance values)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NiShouWen NiBao Bo首纹立鸟镈</td>
<td>XianShan夏商</td>
</tr>
<tr>
<td>JuanTunWen Nao卷纹钮鈒</td>
<td>XianShan夏商</td>
</tr>
<tr>
<td>ShenRenWenShuangNao G神人纹双鸟鼓</td>
<td>XianShan夏商</td>
</tr>
<tr>
<td>JiaYiBingDing Bo豎甲乙丙丁</td>
<td>Xzhou西周</td>
</tr>
<tr>
<td>HuWenZheng Nao虎纹征（徵）</td>
<td>Xzhou西周</td>
</tr>
<tr>
<td>YinHouJianZhou应侯见钟</td>
<td>Xzhou西周</td>
</tr>
<tr>
<td>YinWenNao云纹铙</td>
<td>ChunLuChinGuo春秋战国</td>
</tr>
<tr>
<td>ZengHouYu BianZhong保侯乙编钟</td>
<td>ChunLuGaoGao春秋战国</td>
</tr>
<tr>
<td>HuNiYe ChunZhou祖镈于</td>
<td>ChunLuGaoGao春秋战国</td>
</tr>
<tr>
<td>ZhanGou G春秋战国</td>
<td>ChunLuGaoGao春秋战国</td>
</tr>
<tr>
<td>QiuGuanHuLuo Sheng曲管葫芦笙</td>
<td>ChunLuGaoGao春秋战国</td>
</tr>
<tr>
<td>QinGong Bo秦公镈</td>
<td>QinHun秦汉</td>
</tr>
<tr>
<td>GongZong公钟</td>
<td>QinHun秦汉</td>
</tr>
<tr>
<td>TiejinGong G鈛金铜鼓</td>
<td>QinHun秦汉</td>
</tr>
</tbody>
</table>

Table 2 Instances of digital museum 2

<table>
<thead>
<tr>
<th>Bronze Ware title</th>
<th>Dynasty (instance values)</th>
</tr>
</thead>
<tbody>
<tr>
<td>JuanYi TunWen Nao卽云纹钮鈒</td>
<td>Shang商</td>
</tr>
<tr>
<td>BianXingShouWenNao定义兽面纹鈒</td>
<td>ChunQu春秋</td>
</tr>
<tr>
<td>JinHong Ou Zhong晋侯木仲</td>
<td>XiZhou西周</td>
</tr>
<tr>
<td>WangSunHaoYong Zhong王孙诰甬钟</td>
<td>Late ChunQu春秋晚期</td>
</tr>
<tr>
<td>GongZong公钟</td>
<td>QinHun秦汉</td>
</tr>
<tr>
<td>XinHouGong Q秦国公</td>
<td>XiZhou西周</td>
</tr>
<tr>
<td>NiShouWen NiBao Bo首纹立鸟镈</td>
<td>Shang商</td>
</tr>
<tr>
<td>PanSheWenBian Bo蟠蛇纹编镈</td>
<td>Late ChunQu春秋晚期</td>
</tr>
<tr>
<td>GongBao公镈</td>
<td>QinHun秦汉</td>
</tr>
<tr>
<td>JiaYiBingDing Bo豎甲乙丙丁</td>
<td>XiZhou西周</td>
</tr>
</tbody>
</table>
Two standards for classifying Bronze Ware are as follows:
III. PROBLEM DEFINITIONS

In order define our problem, we give the following definition:

**Definition 1 (Ontology)** An ontology is a pair \( O = \langle C, R \rangle \) where \( C \) is a finite set of *concept* connected by \( R \) which is a finite set of relations.

**Definition 2 (conceptual alignment)** A conceptual alignment \( B = \langle C, B \rangle \), consists of a collection of concepts \( C = \{C_i \mid i \in I\} \) from ontology \( O = \{O_i \mid i \in I\} \) and a collection of alignment bridge \( B = \{B_{ij} \mid i \neq j \in I\} \) between them.

**Definition 3 (ontology alignment)** Ontology alignment is the process that not only include the conceptual alignment bridge, also the ontology instances can be transformed between each other according to their conceptual alignment.

As shown with the motivating example in Section 2 when we looked at mappings between local ontologies, normally we can identify approximate conceptual alignment between two ontology based on their literal meaning. But the source ontology instances cannot transformed into the target ontology entities directly due to their take different instance values. Information flow theory proposes a principle where the two tiers involved in a classification, namely the types and the particulars both contribute to the flow of information. Drawing on this fundamental idea and that semantic inter-operability is essentially a problem of information flow, we propose an two tier ontology alignment approach to distributed digital museums.

IV. RELATED WORKS

A project similar to ours is the SCULPTEUR[1] project, which is an ontological-based solution for navigating, searching and retrieving digital cultural heritage information from multiple distributed digital museums. The nature of the work would seem an ontology based semantic integration for multiple heterogeneous sources such as the work of RDFT: mapping meta-ontology for integration task[2], MAFRA: mapping framework for distributed ontologies[3], OntoMerge: Ontology Translation by merging and reasoning[4], Observer: interoperation across preexisting ontologies[5], OSIF: framework for ontology integration system[6], somewhere: querying distributed ontologies[7].

Marco Schorlemmer and Yannis Kalfoglou propose an information-theoretic foundation for progressive ontology alignment in [8]. Ontology based integration can overcome semantic heterogeneities, but there is not a universal ontology mapping method for all solutions especially for some specific domain. For example, the classification heterogeneity, which is a commonplace in Chinese antique classifications, would not seem solvable by currently available ontology based methods. In order to find the internal correspondences between different classifications, we propose an information flow based solution, which classifies as *types* the hierarchy of the classifications and as tokens the instances that satisfy the corresponding types, and an agreed classification with an indexed family of *infomorphisms* act as core connections of them.

V. USING INFORMATION FLOW FOR ONTOLOGY ALIGNMENT

**A. Information Flow Theory**
Information flow is a rigorous, mathematical theory of knowledge distribution proposed by Barwise and Seligman in 1997. It helps solve problems of how information about one or more components of a distributed system carries information about other components.

**Classification.** A classification is a structure \( A = < \text{tok}(A), \text{typ}(A), \models \text{F} A > \), where \( \text{tok}(A) \) is a set of objects to be classified, called the tokens of \( A \), \( \text{typ}(A) \) is a set of objects used to classify the tokens, called the types of \( A \), and \( \models \text{F} A \) is a binary relation between \( \text{tok}(A) \) and \( \text{typ}(A) \) that determines which tokens are classified by which types. If \( a \models \text{F} A \) then we say that \( a \) is of type \( \alpha \) in classification \( A \).

**Theory.** Given a classification \( A \), a sequent is a pair \( (\Gamma, \Delta) \) of sets of types of \( A \).

A token \( a \) of \( A \) is said to satisfy the sequent \( (\Gamma, \Delta) \) if,
\[
(\forall \alpha \in \Gamma) [a =_\alpha \alpha] \Rightarrow (\exists \alpha \in \Delta) [a =_\alpha \alpha]
\]

We say that \( \Gamma \) entails \( A \) in \( A \), written \( \Gamma \models \Delta \), if every token of \( A \) satisfies \( (\Gamma, \Delta) \).

If \( \Gamma \models \Delta \), then the pair \( (\Gamma, \Delta) \) is said to be a constraint supported by the classification \( A \).

The set of all constraints supported by \( A \) is called the complete theory of \( A \), denoted by \( \text{Th}(A) \). The complete theory of \( A \) represents all the regularities supported by the system being modeled by \( A \).

- **Entailment:** a constraint of the form \( \alpha \models \beta \) represent the claim that \( \alpha \) entails \( \beta \).
- **Necessity:** a constraint of the form \( \alpha \models \beta \) represent the claim that the type \( \alpha \) is necessarily the case, without any preconditions.
- **Exhaustive:** a constraint of the form \( \alpha \models \beta \) represent the claim that every token is of one of the two types \( \alpha \) and \( \beta \), again without any preconditions.
- **Incompatible types:** a constraint of the form \( \alpha \models \beta \) represents the claim that no token is of both types \( \alpha \) and \( \beta \).
- **Incoherent types:** a constraint of the form \( \alpha \models \beta \) represents the claim that no token is of type \( \alpha \).

**Infomorphism.** Let \( A = < \text{tok}(A), \text{typ}(A), \models \text{F} A > \), and \( C = < \text{tok}(C), \text{typ}(C), \models \text{F} C > \), be two classifications. An infomorphism between \( A \) and \( C \) is a contravariant pair of function \( f = (f^\alpha, f^\gamma) \) that satisfies the following Fundamental Property of Infomorphism:
\[
f^\gamma(c) =_\alpha \alpha \text{ iff } c =_\alpha f^\gamma(\alpha)
\]

For all tokens \( c \) of \( C \) and all types \( \alpha \) of \( A \). We refer to \( f^\gamma \) as "f-up" and \( f^\alpha \) as "f-down". We take account of the fact that the functions \( f^\gamma \) and \( f^\alpha \) act in opposite directions by writing \( f : A \rightarrow C \).

**Local logics** A local logic \( L = < A, L, N_L > \) consists of a classification \( A \), a set \( L \) of sequents (satisfying certain structural rules) involving the types of \( A \), called the constraints of \( L \), and a subset \( N_L \subseteq A \), called the normal tokens of \( L \), which satisfy all the constraints of \( L \). A local logic \( L \) is sound if every token is normal; it is complete if every sequent that holds of all normal tokens is in the consequence relation \( \models_L \). A logic is natural if it is generated by some classification, and a natural logic is sound and complete. Using infomorphisms, we can move local logics around from one classification to another even though this normally preserve neither soundness nor completeness.

If \( \Gamma \) is a set of types of \( A \), we denote by \( f^\Gamma \) the set of translations of types in \( \Gamma \). If \( \Gamma \) is a set of types of \( B \), we denote by \( f^\Gamma \) the set of types of \( A \) whose translations are in \( \Gamma \). The following two inference rules allow us to pass from one classification to another:

- **f-Intro:**
  \[
  f^\Gamma \models A \models \Delta \quad \models A \models \Delta \quad \models A \models \Delta
  \]

  The first rule allows us to go from a sequent of \( A \) to a sequent of \( B \); the second rule allows us to go the other way round. It can be proven (which is omitted in this paper and interested user is referred to Barwise and Seligman 1997) that \( f^\text{-Intro} \) preserves soundness, whereas \( f^\text{-Elim} \) preserves completeness.

Given an infomorphism \( f : A \leftrightarrow B \) and a logic \( L \) on one of these classifications, we obtain a natural logic on the other. If \( L \) is a logic on \( A \), then \( f^{-1}[L] \) denotes the logic on \( B \) obtained from \( L \) by \( f^\text{-Intro} \). If \( L \) is a logic on \( B \), then \( f^\Gamma[L] \) denotes the logic on \( A \) obtained from \( L \) by \( f^\text{-Elim} \). For any binary channel \( C \) as Fig.4, we define the local logic \( \text{Log}_C(D) \) on \( D \) induced by that channel as \( \text{Log}_C(D) = d^{-1}[\rho[\text{Log}(P)]] \)

![Fig. 4 Binary logic channel C](image)

As we have observed, if \( \text{Log}(P) \) is sound and complete, and \( p \), \( d \) is token surjective, then \( \text{Log}_C(D) \) is sound and complete.

**B. Proposed Solution**

Followed by above, \( C \) connect \( P \) and \( D \) by means of two pairs of contra-variant functions, it captures an existing duality between concepts and instances: Each pair consists of a map of concepts on the so called type level and map of instances on the so called token level, and pointing in the opposite direction. We model the two Bronze Ware categories shown in figure 2 and figure 3 as \( P(C) \) and \( D(C) \), and they are connected by an agreed understanding \( C(C) \), the type, token and generated theory details of the two local logics are illustrated bellow. The diagram is a general model of the coordinated channel between two local ontology classifications, and it faithfully captures the semantic alignment between them, according to the Barwise-Seligman theory of information flow. With new tokens are shared, new types participate in the meaning coordination, and yielding the newly coordinated channel. The distributed IF logic of the natural logic determined by the core of each new channel captures the semantic integration achieved so far. Because usually the token information is not as complete as needed or its computational cost is too expensive, in practice complete semantic integration will seldom be achieved, the
ontology coordination process will usually yield only a partial semantic integration involving a fraction of communicated types and shared tokens.

Fig. 5 shows how both categories base their concepts upon their agreed understanding. For example, the agreed understanding is materialized by two maps that form the alignment. It requires the classification of particular instances of both categories according with the agreed understanding, since it is this agreed way of classification which will determine how the both categories are going to be related to each other. The following table 3 and 4 shows connections classified into the types which determines a theory of how these concepts are related.

Table 3 Connections classified according to types from one point of view

<table>
<thead>
<tr>
<th>Instance connection</th>
<th>XiangShang</th>
<th>XiZhou</th>
<th>ChunQiuZhanGuo</th>
<th>Cymbals</th>
<th>Bell</th>
<th>Bo</th>
<th>Drum</th>
<th>Chun</th>
<th>Sheng</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We can conclude with the following constraints according to above table, which is the theory about relations between the two antique categories:

XiangShang \( \equiv \) Cymbals, Bo, Drum; XiZhou \( \equiv \) Cymbals, Bell, Bo; ChunQiuZhanGuo \( \equiv \) Cymbals, Bell, Bo, Drum, Chun, Sheng; QinHan \( \equiv \) Bell, Bo, Drum

Table 4 Connections classified according to types from another point of view

<table>
<thead>
<tr>
<th>Instance connection</th>
<th>Cymbals</th>
<th>Bell</th>
<th>Bo</th>
<th>Drum</th>
<th>Chun</th>
<th>Sheng</th>
<th>XiangShang</th>
<th>XiZhou</th>
<th>ChunQiuZhanGuo</th>
<th>QinHan</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Constraints concluded:

Cymbals \( \equiv \) XiangShang, XiZhou, ChunQiuZhanGuo; Bell \( \equiv \) XiangShang, ChunQiuZhanGuo, QinHan; Bo \( \equiv \) XiangShang.
A gain, we model two ‘dynasty’ concepts’ instance values and their instances classified according to the values as two classifications, and their generated theories as $P(D)$ and $D(D)$ separately, with $C(C)$ as their core connection. The alignment between two ‘dynasty’ concepts is shown in Fig.6:

1. $C_i$ chooses one instance value $c_i$ from its instance values set $\{c_1, c_2, \ldots, c_k\}$, and then chooses an instance $o_{k_i}$ from instance set $\{o_{k_1}, o_{k_2}, \ldots, o_{k_n}\}$ which has an instance value $c_i$, denoted as $o_{k_i} \in c_i$.

2. $C'_j$ receives the instance $o_{k_i}$ and finds which instance value(s) $c'_j$ from $\{c'_j\}$ it belongs to, if it is failed to find any instance value, then end this round;

3. For every locating instance value $c'_j \in C'_j$ such that $o_{k_i} \in c'_j$; update it in $C_i$ & $C'_j$ association matrix, then concept $C'_j$ select every object $o'_u$ from its instance value set $\{o'_u\}$ which is different from $o_{k_i}$, and sends them back to $C_i$ to indicate that $o_{k_i}$ and $o'_u$ all belong to the same instance value in $C'_j$.

4. $C_i$ receive $o'_u$ and judge it whether match any its instance values of it, if not, do nothing. If yes, update this in $C_i$ & $C'_j$ association matrix;

**Association matrix**

In our model, we assume that equivalent concepts keeps an association matrix, which is used to capture relationships among objects. After each round of the alignment, the association matrix might be updated to reflect the new relationships among objects. Following are the concept association matrix $S[m][n]$ of equivalent concept $C_i$ from $O$ and $C'_j$ from $O'$, concept $C_i$ has instance values $c_1, c_2, c_3, \ldots, c_n$, whereas concept $C'_j$ has instance values $c'_1, c'_2, \ldots, c'_m$;

$s_{ij}$ is value of the association matrix between two concept instance values:

$s_{ij} = \{ o \}$ collection of instances that belongs to instance value $c_i$ and $c'_j$.

**Association strength**

Denote $AS(c_i, c'_j)$ as association strength between two instance value $c_i$ and $c'_j$.

$AS(c_i, c'_j) = |s_{ij}|$, number of instances of $s_{ij}$

Given local Ontology $O$ and $O'$ have a set of concepts $\{C_i\}$ and $\{C'_j\}$ separately, the association strength of concept $C_i$ and $C'_j$ denote as $AS(C_i, C'_j)$:

$AS(C_i, C'_j) = \sum_{i=1}^{m} \sum_{j=1}^{n} |s_{ij}|$

The association strength of $O$ and $O'$ denote as $AS(O, O')$.
AS(O,O') = \sum_{k=1}^{s} \omega_k \cdot AS(C_i,C'_j)

s is number of equivalent concepts exist in ontology O and O'. \omega_k is the weight of the association strength of concept C_i and C'_j, it can be defined by the need of application.

We can set threshold value for the association strength AS(c_i,c'_j), AS(C_i,C'_j), AS(O,O'), if their value are greater than the threshold, then the involved instance values, concepts or ontology can be treated as ‘relevant’.

**Two tier ontology alignment**

The two tier local ontology alignment is as following fig. 7. It does the ontology alignment between concepts and instance values.

![Two tier ontology alignment](image)

**D. Peer to peer query application**

*Peer to peer query architecture*

We have constructed a prototype system that involves querying distributed heterogeneous ontologies, the system consists of a collection of peers that maintain local ontologies, including repositories of data instantiating certain concepts of local ontologies. Peers are acquainted with neighbors via peer alignment algorithm, that is a set of mappings rules, the mapping rules are formulated according to association strength computed of the local ontology peers, there is a collection of association strength values between peer ontology concepts and concept instance values. If the association strength value of one ontology peer is higher the threshold value, then a query is passed to its neighbor and relevant concepts and instance values are mapped according to association strengths between them. Queries are posed to a given peer using one local ontology of the peer. The answers that are expected are not only instances of given peer classes but also instances of classes of foreign peers obtained via specified mappings.

![Distributed peer-to-peer query architecture](image)

**Precision/recall and the association strength threshold value**

*Precision* and *Recall* have been very widely used in Information Retrieval literature to measure loss of information incurred when the answer to a query issued to the information retrieval system contains some proportion of irrelevant data.

\[
\text{Precision} = \frac{\text{Probability(Relevant|Retrieved)}}{\text{RetrievedSet}}
\]

\[
\text{Recall} = \frac{\text{Probability(Retrieved|Relevant)}}{\text{RelevantSet}}
\]

Learn from practice, we know that the higher(lower) the threshold values for the association strength AS(c_i,c'_j), AS(C_i,C'_j) and AS(O,O') we set, the higher(lower) of the *Precision*, but the lower(higher) of the *Recall*. We can adjust the threshold value according to the application to achieve optimum result.

**VI. CONCLUSION AND CONTRIBUTION**

*Information flow* emphasises that, since information is carried by particular tokens, *information flow* crucially involves both types and tokens. Barwise and Seligman realised the fundamental duality between types and tokens, which is central to all channel theoretic constructions. Thus, although meaning coordination is usually thought of as a process during which concepts of separate ontologies are being aligned at the type-level, the logical relationship between concepts arises when tokens are being connected by means of an IF channel. Knowing what these connections at the token-level are is therefore fundamental for determining the semantic alignment of ontologies at the type-level.

In this paper, we proposed an two tier ontology alignment algorithm for heterogeneous digital museums and show why conceptual level alignment is insufficient in semantic
 interoperability. We construct our distributed peer-to-peer query architecture based on that.

The problem of decentralized querying distributed data through distributed ontologies is that the **Precision** and **Recall** is hard to measure and control, experiment proves that our information flow based two tier ontology alignment solution can optimize the **Precision** and **Recall** rate as well as make them measurable by adjusting the threshold value of association strength.

**Reference**


