Emerging Consensus *in-situ*

Bo Hu, Srinandan Dasmahapatra, Paul Lewis

School of Electronic and Computer Science,
University of Southampton, SO17 1BJ, UK
{bh, sd, phl}@ecs.soton.ac.uk

Abstract Traditional ontology mapping techniques are not strictly applicable in a dynamic and distributed environment (e.g. P2P and pervasive computing) in which on-the-fly alignments are sought after. We propose an approach that collaborates the logic formalisms with collaboratively created web repositories. A logic conceptualisation based "signaturing" algorithm is to discover, from concept definitions, the "feature" vectors that uniquely identify concepts; web repositories are used to understand the implications of these features. Such a combination solidifies an on-demand and approximate mechanism that emerges a context-dependent and task-specific consensus among heterogeneous participants of an information exchange task.

1 Introduction

The prevalence of the Internet has made it possible to access a large amount of data. It has been commonly agreed that attaching machine-understandable semantics to web resources gives birth to "smart" applications and thus benefits ordinary web users by partially relieving them from routine tasks [3]. Thus far, the semantics is mainly depicted using ontologies. Due to a lack of universal standards and the diversity of human perspectives, it is inevitable that ontologies describing the same domain of discourse present semantic disagreement to some extent. Therefore, one of the primary tasks to facilitate the envisioned "smart" use of resources becomes establishing a mutual understanding among different ontology-driven, semantics-enhanced systems. This mutual understanding should faithfully reveal the intended meaning of different ontological entities.

Clearly, in order to have a mutual understanding, all the participants involved in an information exchanging task must agree upon a list of words as the semantics carriers and the meanings of these words must "pick out the same individuals in the same context" [14]. In other words, ontological entities pass meanings by not only values but also context-dependent referents. We refer to the structural and naming information attached to a concept the *value* of the concept while classified instances the *referents*. For example, we might use the sentence "*give me a French*" to ask for a French wine or a French movie depending on the conversational context. Current ontology mapping capability performs well in identifying task-independent semantic equivalences. While such techniques are good in generic ontology mapping scenarios, their applicability is

suspect in certain circumstances. Let's take the peer-to-peer (p2p) environment as an example. In such a setting, when trying to establish a consensus, two major obstacles have prevented us from taking the conventional ontology mapping route. On the one hand, data owners would be more incline to well-targeted and task-specific solutions that allow them to share data within the context of a particular conversation instead of large scale and broad sense consensuses offered by some centralised authorities. On the other hand, the quality of a consensus is largely decided by the data that each individual holds. In many cases, such data might be so diverse, ambiguous and incomplete. Any consensuses stemmed therefrom demonstrate a certain degree of imperfectness, which is a function of the data possessed by a data provider. These two characteristics are, of course, not unique to p2p environments. Any applications aiming to provide on-the-fly semantic alignments present such characteristics.

In this paper, the first issue is accommodated by reinterpreting ontology mapping as a task situated in the background knowledge of a particular conversation: concepts are first decomposed into semantics-bearing signatures and are reenforced as feature vectors based on web encyclopedia repositories. When a particular conversation is to be conducted, we generate the corresponding feature vectors so as to reduce the network traffic and the subsequent computational burden (Section 3). During this process, Latent Semantic Analysis (LSA) is leveraged to alleviate the influence of modelling idiosyncrasy. In Section 3 and throughout the rest of this paper, the "Wine Ontology"[1] is used to detail our approach. LSA also leads to a feasible solution to the second topic of this paper. In a p2p or a similar environment, the consensus established w.r.t. a particular conversation should not be pealed off from the data that is held by each individual. The uncertainty of an answer is, therefore, defined as the degree of satisfaction, i.e. to what extent a request can be satisfied based on the local data that the query handling individuals possess. We propose a probabilistic model to quantify such satisfaction and regulate how an appropriate answer should be screened out from other candidates (Section 4). Finally, in Section 5, we conclude the paper with issues worth further discussion and investigation.

2 Preliminaries

Description Logic (DL) is a family of knowledge representation and reasoning formalisms. It has attracted substantial research interest recently, especially after the endorsement of DL-based ontology modelling languages (e.g. OWL [12]) by the Semantic Web initiative [3]. DLs are based on the notions of concepts (i.e. unary predicates) and properties (i.e. binary relations). Using defined constructs, complex concepts can be composed from primitive ones. In the context of DLs, an ontology is normally a 4-tuple $\langle CN, PN, \mathcal{C}, \mathcal{P} \rangle$ where CN is a set of concept names, PN a set of property names, $\mathcal{C}$ a set of concepts and $\mathcal{P}$ a set of properties. Let C and D be arbitrary concepts, P be a property, n be a non-negative integer,

[1] Available from http://www.schemaweb.info/schema/SchemaDetails.aspx?id=62

o_i ($1 \leq i \leq n$) be instances and $\top$, $\bot$ denote the top and the bottom. A $\mathcal{SHOIN}$ DL concept is: ($\mathcal{SHOIN}$ is the underlying logic of OWL-DL)

$$CN \mid \top \mid \bot \mid \mathsf{C} \sqcap \mathsf{D} \mid \mathsf{C} \sqcup \mathsf{D} \mid \neg \mathsf{C} \mid \exists \mathsf{R}.\mathsf{C} \mid \forall \mathsf{R}.\mathsf{C} \mid \geq_n \mathsf{R}.\top \mid \leq_n \mathsf{R}.\top \mid \{o_1, \ldots, o_n\}$$

An interpretation $\mathcal{I}$ is a couple $(\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$ where the nonempty set $\Delta^{\mathcal{I}}$ is the domain of $\mathcal{I}$ and the $\cdot^{\mathcal{I}}$ function maps each concept to a subset of $\Delta^{\mathcal{I}}$ and each property to a subset of $\Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$.

Latent Semantic Analysis (LSA) [6] is an approach to document indexing. For a large corpus of text documents, LSA assumes the existence of an underlying semantic model that can be captured using a term-document matrix with rows corresponding to terms and columns to documents. It then discovers such a model by projecting the original term-document matrix into a lower-dimensional vector space with effectively reduced noise. LSA has been found capable of simulating a variety of human cognitive phenomena and thus emulating the "meaning" discovering process. The advantage of LSA lies in the fact that the resulting correlation between an arbitrary pair of items (terms or documents) is not isolated from the rest of the representation system. The enabling technique of LSA is Singular Value Decomposition (SVD). SVD decomposes an $M \times N$ matrix $\mathfrak{M}$ and represents it as an approximation, $\hat{\mathfrak{M}}$, at a lower dimensionality k:

$$\mathfrak{M} \approx \hat{\mathfrak{M}} = USV^{\mathrm{T}} = (u_1 \cdots u_k) \begin{pmatrix} \delta_1 & & \\ & \ddots & \\ & & \delta_k \end{pmatrix} \begin{pmatrix} v^1 \\ \vdots \\ v^k \end{pmatrix} \tag{1}$$

where S is an $K \times K$ diagonal matrix of singular values, U is an $M \times K$ matrix of eigenvectors derived from the term-term correlation matrix given by $\mathfrak{M}\mathfrak{M}^{\mathrm{T}}$, and V is an $N \times K$ matrix of eigenvectors derived from the document-document correlation matrix given by $\mathfrak{M}^{\mathrm{T}}\mathfrak{M}$. Recently, methods based on LSA have been successfully applied to detect synonyms and acronyms [4].

3 Situating concept interpretation in WIKIPEDIA

Establishing consensus implies aligning different local ontologies. General ontology mapping has been extensively studied recently [2]. In this paper, we take an eclectic approach drawn from both the formal logics realm and Web2.0 applications. More specifically, we i) produce signatures that explicitly and quantitatively characterise the intensional restrictions of concepts, ii) retrieve web documents according to concept signatures as *virtual* instances, which are niched in the context of a particular "topic", and iii) generalise the signatures into individual-independent and task-specific *feature vectors* based on the landscape of their respective *virtual* instances. Similarity among concepts is then reduced to the similarity among their respective feature vectors.

3.1 Signaturing Concepts

In an ontology $\mathcal{O}$, semantics of concepts are concealed in the inter-concept relationships introduced through subsumptions and property references. The first step towards establishing consensus, therefore, becomes making explicit the semantics hidden behind the concept constructs. In order to reveal such "hidden" semantics, we recursively unfold concepts against their constructs till no further actions can be taken. In this paper, we focus on ontologies that can be represented with OWL-DL. More specifically, we restrict ourselves to $\mathcal{SHOIN}(\mathrm{D})$ DL [10]. This restriction is due to both theoretical and practical considerations. On the one hand, $\mathcal{SHOIN}(\mathrm{D})$ is Beth-definable. Although reasoning w.r.t. $\mathcal{SHOIN}(\mathrm{D})$ is NExpTime-complete, it has been demonstrated [10] that deterministic complexity can be achieved by restricting the concept constructs to a carefully selected subset of $\mathcal{SHOIN}(\mathrm{D})$ and/or translating $\mathcal{SHOIN}(\mathrm{D})$ into the less expensive $\mathcal{SHIN}(\mathrm{D})$ whose satisfiability reasoning is ExpTime [1]. On the other hand, after examining the available ones from the Internet, we observe that many ontologies can be or have already been rewritten in RDF(S) or OWL-DL, both of which are recommended by W3C. Methods developed for $\mathcal{SHOIN}(\mathrm{D})$ is, therefore, applicable to those ontologies based on less expressive languages.

If cyclic definitions are not allowed—no primitive concept (property) appears on both sides of an introduction axiom, and all definitions are in their Negation-Normal Form—the negations are applied only to concept names, it is possible to fully unfold the righthand side of all concept introduction axioms and guarantee the termination of such an unfolding process.

$$
\begin{aligned}
\mathsf{WhiteBordeaux} &\doteq \mathsf{Bordeaux} \sqcap \mathsf{WhiteWine} \\
\mathsf{Bordeaux} &\doteq \mathsf{Wine} \sqcap \exists\, \mathsf{locatedIn}.\{\mathsf{BordeauxRegion}\} \\
\mathsf{Wine} &\sqsubseteq\ =_1 \mathsf{hasBody} \sqcap =_1 \mathsf{hasColor} \\
&\quad \sqcap =_1 \mathsf{hasFlavor} \sqcap =_1 \mathsf{hasMaker} \sqcap =_1 \mathsf{hasSugar} \\
&\quad \sqcap \forall\, \mathsf{hasMaker}.\mathsf{Winery} \sqcap \exists\, \mathsf{locatedIn}.\mathsf{Region} \\
&\quad \sqcap \geq_1 \mathsf{madeFromGrape} \\
\mathsf{WhiteWine} &\sqsubseteq \mathsf{Wine} \sqcap \exists \mathsf{hasColor}.\{\mathsf{White}\} \\
\mathsf{Region} &\sqsubseteq \top \quad \mathsf{Winery} \sqsubseteq\ \top
\end{aligned}
$$

Figure 1. The WhiteBordeaux example

We adopt the construct transformation rules [1] to facilitate the concept unfolding. In Fig. 2, we demonstrate how concept WhiteBordeaux (defined in Fig. 1) is unfolded by repetitively applying the transformation rules. There are cases that concepts are only partially defined with necessary conditions (inclusions) instead of fully defined with both necessary and sufficient conditions (equalities). Before unfolding, inclusions (i.e. axioms in the form $\mathsf{C} \sqsubseteq \mathsf{D}$) are rewritten in equalities (i.e. axioms in the form $\mathsf{C} \doteq \mathsf{D}$). This is achieved by introducing a new primitive concept to represent the difference between C and D. For instance, we introduce $\mathsf{C\text{-}spec} \sqsubseteq \top$ and rewrite $\mathsf{C} \sqsubseteq \mathsf{D}$ into $\mathsf{C} \doteq \mathsf{D} \sqcap \mathsf{C\text{-}spec}$. In this

paper, we assume that the set of newly introduced primitive concept names is disjoint with $CN \cup PN$ and bears clues to the original partially defined concepts. The unfolding process stops when no transformation rules are applicable. It has been demonstrated that by carefully selecting a set of admissible constructs, a termination of unfolding is guaranteed w.r.t. acyclic ontologies.

$$\Pi_1^{\mathsf{WB}} = \{\, x : \mathsf{Bordeaux} \sqcap \mathsf{WhiteWine} \,\}$$

$$\ldots$$

$$\Pi_1^{\mathsf{WB}} = \left\{ \begin{array}{l} x : (=_1 \mathsf{hasBody} \sqcap =_1 \mathsf{hasColor} \sqcap =_1 \mathsf{hasFlavor} \\ \quad \sqcap =_1 \mathsf{hasMaker} \sqcap =_1 \mathsf{hasSugar} \\ \quad \sqcap \forall \mathsf{hasMaker.Winery} \sqcap \exists \mathsf{locatedIn.Region} \\ \quad \sqcap \geq_1 \mathsf{madeFromGrape}) \\ \quad \sqcap \exists \mathsf{locatedIn.}\{\mathsf{BordeauxRegion}\} \\ \quad \sqcap \exists \mathsf{hasColor.}\{\mathsf{White}\} \sqcap \mathsf{WhiteWine\text{-}spec} \\ \quad \sqcap \mathsf{Wine\text{-}spec} \end{array} \right\}$$

$$\ldots$$

$$\Pi_1^{\mathsf{WB}} = \left\{ \begin{array}{l} \langle x, y_0 \rangle : \mathsf{hasBody},\ \langle x, y_1 \rangle : \mathsf{hasColor}, \\ \langle x, y_2 \rangle : \mathsf{hasFlavor},\ \langle x, y_3 \rangle : \mathsf{hasMaker}, \\ \langle x, y_4 \rangle : \mathsf{hasSugar},\ y_3 : \mathsf{Winery}, \\ \langle x, y_5 \rangle : \mathsf{locatedIn},\ y_5 : \mathsf{BordeauxRegion\text{-}spec}, \\ \langle x, y_6 \rangle : \mathsf{madeFromGrape},\ y_1 : \mathsf{White\text{-}spec} \end{array} \right\}$$

Figure 2. Unfolding concept WhiteBordeaux

As illustrated in Fig. 2, WhiteBordeaux is completely unfolded into its semantics-bearing signature, Π_1^{WB}. In order to reduce the computational complexity, when unfolding we introduce primitive concepts to substitute nominal individuals. For instance, the fragment "$\ldots \sqcap \exists$locatedIn.{BordeauxRegion}$\sqcap \ldots$" of Bordeaux in Fig. 1 refers to instance BordeauxRegion of concept Region. We introduce primitive concept BordeauxRegion $\sqsubseteq$ Region accordingly and modify the above fragment into a set of equations as:

$$\begin{array}{rl} \mathsf{WhiteBordeaux} \doteq & \ldots \sqcap \exists\, \mathsf{locatedIn.BordeauxRegion} \ldots \\ \mathsf{BordeauxRegion} \doteq & \mathsf{Region} \sqcap \mathsf{BordeauxRegion\text{-}spec} \end{array}$$

Effectively, the resulting signature is composed by DL ABox assertions generated formally according to the conceptualisation. We regard them as semantics-preserving breakdowns of the constraints that are satisfied by any instances belonging to a concept. Fully breaking down into primitive concepts and properties, in some cases, is difficult to achieve. For instance, universal property value restrictions (UPVRs) can only be further expanded when in the same signature, there are elements defined over the quantified property. In Fig. 3(b), "x : $\forall$hasSex.Male" is not unfolded due to the absence of "$\langle x, y \rangle$: hasSex". It is different from "$\forall$hasMaker.Winery" in Fig. 2 because of the presence of "$=_1$ hasMaker" in the latter case. "x : $\forall$hasMaker.Winery" would have been left unexpanded if otherwise.

It is possible that a concept has more than one signature, if it is defined as the disjunction of other concepts. Applying the non-deterministic unfolding rule of disjunction construct ($\sqcup$) results in alternative signatures, each of which captures a part of the intended meaning of the original concept. For instance, in Fig. 3, Human is unfolded into two different signatures.

$$\begin{array}{rl} \mathsf{Human} \doteq & \mathsf{Man} \sqcup \mathsf{Woman} \\ \mathsf{Man} \doteq & \ldots \sqcap \forall\, \mathsf{hasSex.Male} \sqcap \ldots \\ \mathsf{Woman} \doteq & \ldots \sqcap \forall \mathsf{hasSex.Female} \sqcap \ldots \end{array}$$

(a)

$$\begin{array}{l} \Pi_1^{\mathsf{Man}} = \{\ldots, x : \forall \mathsf{hasSex.Male}, \ldots\} \\ \Pi_1^{\mathsf{Human}} = \{\ldots, x : \forall \mathsf{hasSex.Male}, \ldots\} \\ \Pi_2^{\mathsf{Human}} = \{\ldots, x : \forall \mathsf{hasSex.Female}, \ldots\} \end{array}$$

(b)

Figure 3. Rewriting BordeauxRegion

3.2 Weighting signature elements

Signaturing concepts can be seen as a process that gradually makes the semantic restrictions (expressed via concept constructs) explicit. As a result, each concept is associated with finite sets of formulae, being the primitive concepts, primitive properties and unexpanded universal property value restrictions. The initial feature vector is extracted from these formulae.

$\Pi_i^{\times}$ is subject to two "tuning" actions. Firstly, suffixes of X-spec concepts are removed. For instance, "y_1 : White-spec" in our example is rewritten as "y_1 : White" and is considered the same as those featured by "White". Secondly, residual UPVRs are simplified. The unexpanded UPVRs are dissected into properties and concepts. For instance, "$x : \forall \mathsf{hasSex.Male}$" in Π_1^{Human} is expanded into "$\langle x, y\rangle$: *hasSex" and "y : *Male". The new signature elements generated therefrom are marked as optional to be differentiated from the others.

Obviously, different signature elements contribute differently to shaping the final semantics of concepts. In order to evaluate the significance of individual signature elements, a weighting schema is conceived. Basically, we consider concepts that contribute directly to the construction of others more important than those that impinge on others indirectly through properties or chains of properties. This is to emphasis on those elements that are semantically more significant than others. For instance, Bordeaux and WhiteWine are equally important in shaping the meaning of WhiteBordeaux (see Fig. 1) while BordeauxRegion is less significant than Wine w.r.t. Bordeaux as Bordeaux should be narratively interpreted as a special Wine first before narrowing it down to those that are produced in a particular geographic region. Our interests of Bordeaux, therefore, are arguably more in the former than the latter. In order to reflect such a difference in different restrictions and thus different signature elements of a concept, we introduce the weight adjusting coefficient β. β_e of signature element e is estimated as follows: we split an element as the *head* (e.g. "x" and "$\langle x, y_0\rangle$") and the *tail* (e.g. Region and hasBody) separated by a colon.

- if e is a first-class signature element headed by "x", $\beta_e = \omega^c$;
- if e is a first-class signature element headed by "$\langle x, y_i\rangle$", $\beta_e = \omega^p$;
- if e is a non-first-class signature element introduced in Π_i^{C} through a property P or a property chain $\mathsf{P}_1 \cdots \mathsf{P}_n$:

$$\beta_e = \beta_{\mathsf{P}} * \omega^c \qquad \text{or} \qquad \beta_e = \prod_i \beta_{\mathsf{P}_i} * \omega^c;$$

- if e is restricted by Negation, $\beta_e = -\beta_e$;

– if e is an optional element introduced through unfolding residual universal property value restrictions, $\beta_e = 0.5\,\beta_e$.

The weight of an arbitrary e then relies on two initial values ω^c and ω^p corresponding to the first-class element featured by a primitive concept and a primitive property respectively. The exact values of ω^c and ω^p are obtained by either i) assigning manually based on one's domain knowledge and expectation or ii) adopting the *tf-idf* weighting schema used in Information Retrieval (IR) with the assumption that an element appearing in every concept is less significant than those appearing only in a handful of concepts. Weight adjusting coefficients are memorised for each signature element. We then simplify the signatures to a set of terms/phrases. We extract the *bodies* of signature elements and apply Natural Language Processing (NLP) methods to tokenise and stem the *bodies* [11]. Those resulting terms or short phrases that appear more than once in a signature are collapsed into one with an aggregative weight as the sum of those corresponding to every occurrence, i.e. $\beta_e^{\text{total}} = \sum_i \beta_e^i$. When merging multiple appearance, we observe the disjointness between primitive concepts and primitive properties. For convenience, we denote the set of weighted terms/phrases obtained at this stage as γ.

3.3 Generalising Signatures

Terms or short phrases in γ are individual-specific, presenting interindividual variation. Therefore, in order to emerge a consensus among individuals each holding a different local ontology—possibly in different natural languages, it is necessary to situate the interpretation of those terms/phrases stemmed from concept signatures of different ontologies into the same background knowledge. A straightforward approach to drawing such information is treating web repositories as the source of common background knowledge. For instance in order to exploit the feature vectors of WhiteBordeaux, one has to understand all the words (i.e. Color, Flavor, etc.) appearing therein. Such an understanding should not rely on a particular ontology or vocabulary and should reflect the general human cognition of the words. In this paper, inspired by existing studies (e.g. [9]), we juxtapose terms/phrases against the titles of WIKIPEDIA articles and represent each concept as a vector of weighted WIKIPEDIA article titles, referred to as a *wiki-enhanced feature vector*. Note that hereafter we use bold font to denote the wiki-enhanced feature vectors.

WIKIPEDIA is a very appealing and probably the largest source of encyclopaedic knowledge. As a collaboratively edited document repository, it seems reasonable to conjecture that WIKIPEDIA presents most of the modelling (e.g. naming) variation that one will expect in independently developed ontologies. Meanwhile, the great diversity of wikipedians' background ensures that the contents published on WIKIPEDIA generally has better quality than other non-peer-reviewed web resources.

We assume that every conversation or an information exchange task focuses on a particular topic. For instance, when one asks others about "the taste of

white Bordeaux", the topic of this incident is "Wine". If we denote the main WIKIPEDIA article as `t`, we compute the enhanced feature vector as in Fig. 4. In Step 1), when harvesting "virtual instances" from WIKIPEDIA, we utilise three different types of articles to pool a well targeted text corpus: i) the main article (Λ_{Main}) together with other articles that are the m neighbours of `t` (Λ_{Neigbr}), ii) the *List_of_**xxx*** page π and WIKIPEDIA articles directly linked to π and their m neighbours (Λ_{List}), and iii) the corresponding articles in other languages (Λ_{Lang}). WIKIPEDIA articles are retrieved from the following URL patterns:

$$\text{URL for } \tau = \textit{http://}\langle\texttt{ln}\rangle\textit{.wikipedia.org/wiki/}\langle\texttt{t}\rangle$$
$$\text{URL for } \pi = \textit{http://}\langle\texttt{ln}\rangle\textit{.wikipedia.org/wiki/List\_of\_}\langle\textit{xxx}\rangle$$

where $\langle\texttt{ln}\rangle$ is the language code (e.g. "`en`" for English and "`fr`" for French). When a particular topic does not have corresponding WIKIPEDIA entry, one has to manually specify the correct keywords to find the appropriate articles. Outbound links from Λ_{Main} are followed to retrieve articles that are closely related to Λ_{Main}. In many topics, WIKIPEDIA maintains the so-called "*List of*" pages, e.g. the "list of wine producing countries". Articles referred to from such collective pages are normally well-situated. For instance, "France" in "List of wine producing countries" leads to the WIKIPEDIA article titled "French Wine". Links to collective pages might also be available from within the main article. In our approach we gather both types of collective pages and their m-neighbours in Λ_{List}. Pooling all the WIKIPEDIA articles together, we have a well-targeted corpus of text documents, $\Lambda = \Lambda_{\text{Main}} \cup \Lambda_{\text{Neigbr}} \cup \Lambda_{\text{List}}$. Harvested WIKIPEDIA articles are parsed to remove WIKIPEDIA specific tags and commands. In Step 4), the SVD operation helps to reduce modelling variation and discover the latent semantics—unrevealed correlations between terms and articles. Similarly, SVD is performed again in Step 6) to optimise the weights of WIKIPEDIA articles w.r.t. concepts in $\mathcal{O}$.

If multiple natural languages are involved when establishing the consensus, $\mathbf{c}$ (see Step 8 in Fig. 4) needs to be translated into other languages. Although the cross-lingual capability of LSA has been investigated [8], we would rather avoid experimenting with such an approach and opt for a simple solution: when constructing $\mathfrak{M}_{\text{ac}}$ in Step 5), instead of the articles in the same language as the local ontology $\mathcal{O}$, we retrieve those corresponding WIKIPEDIA articles in the target languages as $\Lambda^{x}_{\text{Lang}}$, links to which normally present in the English articles. x will be decided by the context in which the consensus is to be established. For instance, if an individual is expected to communicate with French-speaking groups, $\Lambda^{\text{Fr}}_{\text{Lang}}$ is populated. One example of such articles is the French correspondence of "Wine" available at "*http://fr.wikipedia.org/wiki/Vin*". Subsequently, Step 6) and 7) are carried out based on the new matrix.

After signaturing and generalisation, an arbitrary concept $\mathsf{C} \in \mathcal{O}$ is associated with a WIKIPEDIA-enhanced and semantics-enriched feature vector, $\mathbf{c}$, that represents the context within which C is to be interpreted. Note that the elements in the final enhanced feature vector is not specific to the naming habit of an individual ontology engineer. That is to say that $\mathbf{c}$ contains those terms that

1) harvest relevant WIKIPEDIA articles against $\mathtt{t}$ and pool them into Λ;
2) index every article $a \in \Lambda$ with a list of terms, l_a, and weight $t \in l_a$ based on *tf-idf* schema as $w_{\langle \mathrm{t}_i, \mathrm{a} \rangle}$;
3) construct the term-article matrix $\mathfrak{M}_{\mathrm{ta}}$ with WIKIPEDIA articles as columns and assign cell entries as

$$c_{ij} = \begin{cases} \beta_{kw} \cdot w_{\langle \mathrm{t}_i, \mathrm{a} \rangle}, & \text{if} \quad t_i = t_{\mathrm{kw}} \in \gamma; \\ 0, & \text{if} \quad t_i \notin \gamma. \end{cases}$$

4) perform SVD on $\mathfrak{M}_{\mathrm{ta}}$ and compute the correlation, $\sigma_{\langle \gamma, a \rangle}$, between γ and every indexed WIKIPEDIA article as the cosine of the angle between γ and article term vectors;
5) construct an article-concept matrix $\mathfrak{M}_{\mathrm{ac}}$ with WIKIPEDIA articles as rows, concepts from $\mathcal{O}$ as columns, and cell $c_{ij} = \sigma_{\langle \gamma, a \rangle}$;
6) condense the dimensionality of $\mathfrak{M}_{\mathrm{ac}}$ into $\hat{\mathfrak{M}}_{\mathrm{ac}}$ with SVD;
7) associate every $\mathsf{C} \in \mathcal{O}$ with a vector of WIKIPEDIA articles, denoted as $\mathbf{c}$.
8) (Optional) go to Step 5) and translate $\mathbf{c}$ into other languages based on $\Lambda^{\mathtt{x}}_{\mathrm{Lang}}$.

Figure 4. Algorithm for generating wiki-topic vectors

might not appear in $\mathcal{O}$ but are frequently correlated with parts of the restrictions of concept C. Such extras are pulled out from the referenced web repository, in our case WIKIPEDIA. This is consistent with the observation on inter-individual modelling variability: people tent to use different terms (e.g. synonyms and/or hypernyms) to refer to the same object [6].

4 Answering queries approximately

When trying to establish consensus, a key task is to find the local substitutes of those foreign concepts. In this occasion, we expect that all the individuals have already done their "homework" off-line and trained their article-concept matrix ($\hat{\mathfrak{M}}_{\mathrm{ac}}$) against the same web repository, namely WIKIPEDIA. Acquiring the similarity between a foreign concept C and the local ones requires the query handling peer to incorporate the foreign feature vector $\mathbf{c}$ of C into its local $\hat{\mathfrak{M}}_{\mathrm{ac}}$. Constructing $\hat{\mathfrak{M}}_{\mathrm{ac}}$ from scratch w.r.t. every received foreign feature vector is undesirable due to the high cost in recomputing SVD. A cheap solution is the so-called *folding-in* [6] operation. Let $\hat{\mathfrak{M}}_{\mathrm{ac}} = USV^{\mathrm{T}}$ be the reduced matrix of the query handler, p_{qh}. Every input $\mathbf{c}$ is projected onto the span of U as $(\mathbf{c})^{\mathrm{T}} U S^{-1}$. Once feature vectors of the foreign concepts are put juxtaposed with those local ones, similarity measures such as the *cosine* angle can be employed to compare the foreign concepts against the local ones. Since both the query submiter p_0 and the query handler p_{qh} train and obtain their local *interpreation vectors* based on the same web repository, the folding-in approach is applicable. Alternatively, a more sophisticated and slightly more expensive approach, the SVD-updating method, can be used. Thus far, a few fast updating algorithms have been investigated. Our experience suggested that the one proposed in [15] outperformed many others.

4.1 LSA-based probability

There is a very little opportunity for individuals with independently developed ontologies to find perfect equivalences among themselves. In a majority of cases, we might need to rewrite Q_0 into Q_{qh}:"[...]*the list of French WhiteWine from Boreaux*" and attach to it similarity values. When more than one foreign concept is involved, there is not a plausible model to combine the multiple similarities. In order to tackle this daunting prospect of query rewriting with multiple foreign terms, we adopt the probability model of LSA [7] and a similarity approximation of probability [5]. If C is a concept appearing in Q_0 with $\mathbf{c}$, answers to Q_{qh} can be regarded as a faithful answer to Q_0 with a probability of $p\,(\mathsf{C} \mid \mathsf{C}') = p\,(\mathbf{c} \mid \mathbf{c}',\, \mathbf{u}'_i)$ where C' is the local translation of C in $\mathcal{O}_{\text{qh}}$ and $\mathbf{u}'_i$ are the left singular vectors computed based on $\mathcal{O}_{\text{qh}}$ (see Equation 1).

$p\,(\mathbf{c} \mid \mathbf{c}',\, \mathbf{u}'_i)$ can be derived from two probabilities, $p\,(\mathbf{c}' \mid \mathbf{u}'_i)$ and $p\,(\mathbf{c} \mid \mathbf{u}'_i)$ [5]. $p\,(\mathbf{c} \mid \mathbf{u}'_i)$ is given by a probability variation of LSA model [7]. In this model, the author demonstrated that with some relaxations, an LSA-based similarity can be presented as the probability of a particular document (a column of a term-document matrix $\mathfrak{M}$) in the term-space (rows of $\mathfrak{M}$) based on document-document similarities [7]. This probability is computed as

$$p\,(\mathbf{c}' \mid \mathbf{u}'_i) = e^{(\mathbf{c}' \cdot \mathbf{u}'_1)^2 + \cdots + (\mathbf{c}' \cdot \mathbf{u}'_k)^2} \,/\, Z_k \tag{2}$$

where Z_k is the normalisation constant[2]. Similarly, $p\,(\mathbf{c} \mid \mathbf{u}'_i)$ is computed. Deriving $p\,(\mathbf{c} \mid \mathbf{c}', \mathbf{u}'_i)$ from $p\,(\mathbf{c}' \mid \mathbf{u}'_i)$ and $p\,(\mathbf{c} \mid \mathbf{u}'_i)$, on the other hand, is not straightforward. In this paper we use a similarity emulation of probability [5],

$$p\,(\,\mathbf{c} \mid \mathbf{c}', \mathbf{u}'_i) = p\,(\,\mathbf{c}' \mid \mathbf{u}'_i)^{\epsilon}, \text{ where } \epsilon = \left(\frac{1 - \mathbf{c} \cdot \mathbf{c}'}{1 + \mathbf{c} \cdot \mathbf{c}'} \right)^{1 - p\,(\mathbf{c} | \mathbf{u}'_i)} \tag{3}$$

where $\mathbf{c} \cdot \mathbf{c}'$ gives the similarity value of concepts C and C' computed from their respective feature vectors.

4.2 Fidelity of query rewriting

Since $\mathbf{u}'_i$ can be seen as a representation of p_{qh}'s local data, we define to what extent p_{qh} can understand p_0 as *fidelity* of the query-specific consensus between p_0 and p_{qh} as: $\mathit{fid}\,(\,Q_0 \mid Q^{\text{Tr}}) = p\,(\,\mathbf{c}_1, \ldots, \mathbf{c}_n \mid \mathbf{c}'_1, \ldots, \mathbf{c}'_n, \mathbf{u}'_i)$, where $\mathbf{c}_i$ are the respective feature vectors of concepts in Q_0 while $\mathbf{c}'_i$ are feature vectors of the corresponding concepts in the translation Q^{Tr}. One possible way of continuing the derivation of the fidelity function would be approximating $p\,(\,\mathbf{c}_1, \ldots, \mathbf{c}_n \mid \mathbf{c}'_1, \ldots, \mathbf{c}'_n, \mathbf{u}'_i)$ with Equation 3. We collapse vectors $\mathbf{c}_1, \ldots, \mathbf{c}_n$ to their centroid $\mathbf{c}_c$ while $\mathbf{c}'_1, \ldots, \mathbf{c}'_n$ to $\mathbf{c}'_c$ and apply Equation 3. Hence, we have $\mathit{fid}\,(\,Q_0 \mid Q_i^{\text{Tr}}) \approx p\,(\,\mathbf{c}_c \mid \mathbf{c}'_c, \mathbf{u}'_i)$. The computation of the fidelity requires only limited information from query submitting individuals: the actual queries and those feature vectors

[2] Please refer to [7] for details of how Z_k is defined.

associated with the queries. Meanwhile, computation is localised to each query handler and characterised by its local ontology.

The fidelity, $fid(\, Q_0 \mid Q^{\mathrm{Tr}})$, can be regarded as a criteria for judging the capability of query handlers. Let Q_i^{Tr} be the translation of Q_0 in the local ontology $\mathcal{O}_i$ of p_i, the query rewriting fidelity reflects how good an answer to Q_0 drawn from Q_i^{Tr} is and thus how well p_i can handle Q_0 based on its local knowledge. The higher the fidelity is, the greater chance the original query is satisfactorily answered by the query handler. Such a probabilistic model paves the way for a ranking mechanism of candidate query handlers. A feasible scenario could be that every p_i in a group G estimates its own fidelity based on the local translation of Q_i^{Tr}, the original Q_0 and the incoming foreign feature vectors. A group coordinator can then allocate the query handling task to the one with the highest estimated fidelity value.

5 Concluding Remarks

In a loosely regulated environment, global consensus may not be enforced. It is more likely that each individual or a small group of peers maintains a local ontology. For an outsider to query an established group and retrieve useful information, conventional ontology mapping techniques (e.g. discussed in [13]) are not sufficient. Dynamic and on-the-fly methods for establishing on demand consensuses becomes more desirable. In the meantime, exact equivalences have given way to those less perfect ones. In this paper, we propose a mechanism to emerge and exploit imperfect consensuses among heterogeneous data holders. We situate an alignment task in the background knowledge drawn from public web repositories, e.g. WIKIPEDIA. We combine the strength of both representation formalisms and collaboratively created web repositories when discovering semantics. Our semantic alignment approach also gives birth to a probabilistic model to evaluate approximate query answering by identifying the most appropriate individual to handle the query and ranking candidate answers returned from the chosen one. There are two sources of complexity w.r.t. the proposed approach: i) signaturing concepts using DL transformation rules (NEXPTIME for expressive DLs) and ii) SVD dimensionality reduction. We, however, would like to make the following argument. On the one hand, DL-based reasoning is very expensive in itself. Any methods aiming to manipulate semantics embedded therein, therefore, have to pay the price of high complexity. Meanwhile, high complexity is normally associated with a set of "culprit" DL constructs. Introducing substituting constructs and/or alternative modelling techniques might help us to find a route around the complexity issue. Meanwhile, construct transformation rules are implemented and optimised in many DL systems and the empirical evaluation has confirmed the performance of such systems in tackling real-life ontologies [1]. On the other hand, LSA has been extensively studied in IR. We share the optimistic expectation with those dedicated researchers regarding the practical value of LSA/SVD in latent semantics discovery.

Finalising the implementation and evaluating it against real-life scenarios are our immediate future work. The evaluation will focus on the following aspects: i) the applicability of WIKIPEDIA w.r.t. topics of different popularity, ii) the scalability of the signaturing and weighting algorithms and their applicability in the current landscape of ontologies, and iii) the performance of the proposed probability mechanism for query handling.

Acknowledgements

This work is supported under the OpenKnowledge STREP projects funded by EU Framework 6 under Grant numbers IST-FP6-027253.

References

1. F. Baader, D. Calvanese, D. McGuinness, D. Nardi, and P. Patel-Schneider, editors. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, 2003.
2. R. Benjamins, J. Euzenat, N. Noy, P. Shvaiko, and M. Stuckenschmidt, H. adn Uschold, editors. *International Workshop on Ontology Matching*, 2006.
3. T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, pages 28–37, 2001.
4. V. Bhat, T. Oates, V. Shanbhag, and C. Nicholas. Finding aliases on the web using latent semantic analysis. *Data & Knowledge Engineering*, 49(2):129–143, 2004.
5. S. Blok, D. Medin, and D. Osherson. Probability from similarity. In P. Doherty, J. McCarthy, and M. Williams, editors, *AAAI Spring Symposium on Logical Formalization of Commonsense Reasoning*, pages 36–42. AAAI Press, 2003.
6. S. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.
7. C. Ding. A probabilistic model for latent semantic indexing. *JASIST*, 56(6):597–608, 2005.
8. S. Dumais, T. Landauer, and M. Littman. Automatic cross-linguistic information retrieval using latent semantic indexing. In *AAAI Spring Symposium on Cross Language Information Retrieval*, 1997.
9. E. Gabrilovich and S. Markovitch. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th IJCAI*, pages 1606–1611, 2007.
10. I. Horrocks and U. Sattler. A tableaux decision procedure for $\mathcal{SHOIQ}$. In *Proc. of the 19th Int. Joint Conf. on Artificial Intelligence (IJCAI 2005)*, 2005.
11. R. Korfhage. *Information storage and retrieval*. Wiley Computer Publishing, 1997.
12. D. L. McGuinness and F. van Harmelen. *OWL Web Ontology Language Overview*. W3C, 2003.
13. E. Rahm and P.A. Bernstein. A survey of approaches to automatic schema matching. *The VLDB Journal*, 10:334–350, 2001.
14. L. Steels and F. Kaplan. Bootstrapping grounded word semantics. In T. Briscoe, editor, *Linguistic Evolution through Language Acquisition: Formal and Computational Models*, chapter 3. Cambridge University Press, 1999.
15. H. Zha and H. Simon. On updating problems in latent semantic indexing. *SIAM J. Sci. Comput.*, 21(2):782–791, 1999.