Putting Things in Context: A Topological Approach to Mapping Contexts and Ontologies

Aviv Segev, Avigdor Gal
Technion - Israel Institute of Technology
Haifa 32000
Israel
{asegev@tx, avigal@ie}.technion.ac.il

Abstract

Ontologies and contexts are complementary disciplines for modeling views. In the area of information integration, ontologies may be viewed as the outcome of a manual effort of modeling a domain, while contexts are system generated models. In this work, we aim at formalizing the inter-relationships between a manually generated ontology and automatically generated contexts. We provide a formal mathematical framework that delineates the relationship between contexts and ontologies. We then use the model to define the uncertainty associated with automatic context extraction from existing documents and provide a ranking method, which ranks ontology concepts according to their suitability with a given context. Throughout this work we motivate our research using QUALEG, a European IST project that aims at providing local government an effective tool for bi-directional communication with citizens.

Keywords: Ontology, Context, Topology mapping

Introduction

Ontologies and contexts are both used to model views, which are different perspectives of a domain. Some consider ontologies as shared models of a domain and contexts as local views of a domain. In the area of information integration, an orthogonal classification exists, in which ontologies are considered a result of a manual effort of modeling a domain, while contexts are system generated models (Segev, Leshno, & Zviran 2004). As an example, consider an organizational scenario in which an organization (such as a local government) is modeled with a global ontology. A task of document classification, in which new documents are classified upon arrival to relevant departments, can be modeled as an integration of contexts (automatically generated from documents) into an existing ontology. A simple example of a context in this setting would be a set of words, extracted from the document.

This approach was recently taken in QUALEG, a European Commission project aimed at increasing citizen participation in the democratic process.[1] In QUALEG, contexts are used to specify the input from citizens and then to provide services - routing emails to departments, opinion analysis on topics at the forefront of public debates, and the identification of new topics on the public agenda.

The two classifications are not necessarily at odds. In the example given above, documents may be email messages from citizens, expressing a local view of a domain. Yet, the classification of manual vs. automatic modeling of a domain has been the center of attention in the area of data integration and schema matching in the past few years. In particular, many heuristics were proposed for the automatic matching of schemata (*e.g.,* Cupid (Madhavan, Bernstein, & Rahm 2001), GLUE (Doan *et al.* 2002), and OntoBuilder (Gal *et al.* 2005b)), and several theoretical models were proposed to represent various aspects of the matching process (Madhavan *et al.* 2002; Melnik 2004; Gal *et al.* 2005a).

In this work, we aim at formalizing the inter-relationships between an ontology, a manually generated domain model, and contexts, partial and automatically generated local views. We provide a formal mathematical framework that delineates the relationships between contexts and ontologies. Following the motivation given above, we discuss the uncertainty associated with automatic context extraction from existing documents and provide a ranking model, which ranks ontology concepts according to their suitability with a given context. We provide examples from the QUALEG project.

Our contributions are as follows:

- We present a framework for combining contexts and ontologies using topological structures, and model the uncertainty inherent to automatic context extraction.
- We provide a model for ranking ontology concepts relative to a context.
- Using real world scenario, taken from email messages from citizens in a local government, we demonstrate three tasks that involve mapping contexts to ontologies, namely email routing, opinion analysis, and public agenda identification.

The rest of the paper is organized as follows. We first discuss related work on the topic. Next, we propose a model for combining contexts and ontologies and present a ranking model to map contexts to ontologies. The last section includes concluding remarks and suggestions for future work.

[1] http://www.qualeg.eupm.net/

Related Work

This section describes related work in four different research areas, namely context representation, ontologies, context extraction, and topologies.

Context Representation

The context model we use is based on the definition of context as first class objects formulated by McCarthy (McCarthy 1993). McCarthy defines a relation $ist(\mathcal{C}, P)$, asserting that a proposition P is true in a context $\mathcal{C}$. We shall use this relation when discussing context extraction.

It has been proposed to use a multilevel semantic network to represent knowledge within several levels of contexts (Terziyan & Puuronen 2000). The zero level of representation is a semantic network that includes knowledge about basic domain objects and their relations. The first level of representation uses a semantic network to represent contexts and their relationships. The second level presents relationships of metacontexts, the next level describes metametacontext, and so on and so forth. The top level includes knowledge that is considered to be true in all contexts. In this work we do not explicitly limit the number of levels in the sematic network. However, due to the limited capabilities of context extraction tools nowadays (see below), we define context as sets of sets of descriptors at zero level only and the mapping between contexts and ontology concepts is represented at level 1. Generally speaking, our model requires $n + 1$ levels of abstraction, where n represents the abstraction levels needed to represent contexts and their relationships.

Previous work on contexts (Siegel & Madnick 1991) uses metadata for semantic reconciliation. The database metadata dictionary (DMD) defines the semantic and assignment domains for each attribute and the set of rules that define the semantic assignments for each of these attributes. The application semantic view (ASV) contains the applications definition of the semantic and assignment domain and the set of rules defining the applications data semantic requirements. They define the semantic domain of an attribute T as the set of attributes used to define the semantics of T. Work by (Kashyap & Sheth 1996) use contexts that are organized as a meet semi-lattice and associated operations like the greatest lower bound for semantic similarity are defined. The context of comparison and the type of abstractions used to relate the two objects form the basis of a semantic taxonomy. They define ontology as the specification of a representational vocabulary for a shared domain of discourse. Both these approaches use ontological concepts for creating contextual descriptions and serve best when creating new ontologies. The approach proposed herein assumes the existence of an ontology to which contexts should be mapped. Another difference is that in (Kashyap & Sheth 1996), an ontology concept is taken to be the intersection of context sets, while we view ontology concepts as the union of context sets.

Ontology

Ontologies were defined and used in various research areas, including philosophy (where it was coined), artificial intelligence, information sciences, knowledge representation, object modeling, and most recently, eCommerce applications. In his seminal work, Bunge defines Ontology as a world of systems and provides a basic formalism for ontologies (Bunge 1979). Typically, ontologies are represented using a Description Logic (Borgida & Brachman 1993; Donini *et al.* 1996), where subsumption typifies the semantic relationship between terms; or Frame Logic (Kifer, Lausen, & Wu 1995), where a deductive inference system provides access to semi-structured data.

The realm of information science has produced an extensive body of literature and practice in ontology construction, *e.g.*, (Vickery 1966). Other undertakings, such as the DOGMA project (Spyns, Meersman, & Jarrar 2002), provide an engineering approach to ontology management. Work has been done in ontology learning, such as Text-To-Onto (Maedche & Staab 2001), Thematic Mapping (Chung *et al.* 2002), OntoMiner (H. Davulcu & Nagarajan 2003), and TexaMiner (Kashyap *et al.* 2005) to name a few. Finally, researchers in the field of knowledge representation have studied ontology interoperability, resulting in systems such as Chimaera (McGuinness *et al.* 2000) and Protègè (Noy & Musen 2000).

Our model is based on Bunge's terminology. We aim at formalizing the mapping between contexts and ontologies and provide an uncertainty management tool in the form of concept ranking. Therefore, in our model we assume an ontology is given, designed using any of the tools mentioned above.

Context Extraction

The creation of taxonomies from metadata (in XML/RDF) containing descriptions of learning resources was undertaken in (Papatheodorou, Vassiliou, & Simon 2002). Following the application of basic text normalization techniques, an index was built, observed as a graph with learning resources as nodes connected by arcs labeled by the index words common to their metadata files. A cluster mining algorithm is applied to this graph and then the controlled vocabulary is selected statistically. However, a manual effort is necessary to organize the resulting clusters into hierarchies. When dealing with medium-sized corpora (a few hundred thousand words), the terminological network is too vast for manual analysis, and it is necessary to use data analysis tools for processing. Therefore, Assadi (Assadi 1998) has employed a clustering tool that utilizes specialized data analysis functions and has clustered the terms in a terminological network to reduce its complexity. These clusters are then manually processed by a domain expert to either edit them or reject them.

Several distance metrics were proposed in the literature and can be applied to measure the quality of context extraction. Prior work had presented methods based on information retrieval techniques (van Rijsbergen 1979) for extracting contextual descriptions from data and evaluating the quality of the process. Motro and Rakov (Motro & Rakov 1998) proposed a standard for specifying the quality of databases based on the concepts of soundness and completeness. The method allowed the quality of answers to arbitrary

queries to be calculated from overall quality specifications of the database. Another approach (Mena *et al.* 2000) is based on estimating loss of information based on navigation of ontological terms. The measures for loss of information were based on metrics such as precision and recall on extensional information. These measures are used to select results having the desired quality of information.

To demonstrate our method, we propose later in this paper the use of a fully automatic context recognition algorithm that uses the Internet as a knowledge base and as a basis for clustering (Segev, Leshno, & Zviran 2004). Both the contexts and the ontology concepts are defined as topological sets, for which set distance presents iteself as a natural choice for a distance measure.

Topology

In recent years different researchers have applied principles from the mathematical domain of topology in different fields of Artificial Intelligence. One work uses topological localization and mapping for agent problem solving (Choset & Nagatani 2001). Other researchers have implemented topology in metrical information associated with actions (Shatkay & Kaelbling 1997; Koenig & Simmons 1996). In another method of topological mapping, which describes large scale static environments using a hybrid topological metric model, a global map is formed from a set of local maps organized in a topological structure, where each local map contains quantitative environment information using a local reference frame (Simhon & Dudek 1998). Remolina and Kuipers present a general theory of topological maps whereby sensory input, topological and local metrical information are combined to define the topological maps explaining such information (Remolina & Kuipers 2004).

In this work we use topologies as a tool of choice for integrating contexts and ontologies.

A Model of Context and Ontology

In this section we formally define contexts and ontologies and propose a topology-based model to specify the relationships between them.

Contexts and Ontologies

A *context* $\mathcal{C} = \left\{ \{\langle c_{ij}, w_{ij} \rangle\}_j \right\}_i$ is a set of finite set of descriptors c_{ij} from a domain $\mathcal{D}$ with appropriate weights w_{ij}, defining the importance of c_{ij}. For example, a context $\mathcal{C}$ may be a set of words (hence, $\mathcal{D}$ is a set of all possible character combinations) defining a document Doc, and the weights could represent the relevance of a descriptor to Doc. In classical Information Retrieval, $\langle c_{ij}, w_{ij} \rangle$ may represent the fact that the word c_{ij} is repeated w_{ij} times in Doc.

An *ontology* $O = (V, E)$ is a directed graph, with nodes representing concepts (*things* in Bunge's terminology (Bunge 1977; 1979)) and edges representing relationships. A single concept is represented by a name and a context $\mathcal{C}$. Figure 1 (top) displays the graphical representation of an ontology.

Example 1. *To illustrate contexts and ontologies, consider the local government of Saarbrücken. Two ontology concepts in the ontology of Saarbrücken are:*
(*Perspectives du Theatre,* $\{\{\langle$*Öffentlichkeitsarbeit, 2*$\rangle\}$, $\{\langle$*Multimedia, 1*$\rangle\}$, $\{\langle$*Kulturpolitik, 1*$\rangle\}$, $\{\langle$*Musik, 6*$\rangle\}$, ...$\}$)
and
(*Long Day School,* $\{\{\langle$*Förderbedarf, 1*$\rangle\}$, $\{\langle$*Mathematik, 2*$\rangle\}$, $\{\langle$*Musik, 2*$\rangle\}$, $\{\langle$*Interkulturell, 1*$\rangle\}\}$)
A context, which was generated from an email message using the algorithm in (Segev, Leshno, & Zviran 2004) (to be described later) is $\{\{\langle$*Musik,* $8\rangle\}$, $\{\langle$*Open Air,* $1\rangle\}\}$.

Modeling Context-Ontology Relationships

The relationships between ontologies and contexts can be modeled using topologies as follows. A *topological structure (topology)* in a set X is a collective family $\vartheta = (G_i / i \in I)$ of subsets of X satisfying
1. $J \subset I \Rightarrow \bigcup_{i \in J} G_i \in \vartheta$
2. J finite; $J \subset I \Rightarrow \bigcap_{i \in J} G_i \in \vartheta$
3. $\emptyset \in \vartheta, X \in \vartheta$

The pair (X,ϑ) is called a *topological space* and the sets in ϑ are called *open sets*. We now define a context to be an open set in a topology, representing a family ϑ of all possible contexts in some set X. Using the concrete example given above, let X be a set of sets of tuples $\langle c, w \rangle$, where c is a word (or words) in a dictionary and w is a weight. Note that ϑ is infinite since descriptors are not limited in their length and weights are taken from some infinite number set (such as the natural numbers $\mathbb{N}$).

A family $\mathbf{B} = (B_i / i \in I)$ is called a *filter base* (also known as a directed set, indexed set, or a base) if
1. $(\forall i) : B_i \neq \emptyset$
2. $(\forall i)\ (\forall j)\ (\exists k) : B_k \subset B_i \cap B_j$

A *filtered family* is a family of sets $(x_i / i \in I)$ associated with a filter base $\mathbf{B}$ on index I. A filtered family $(x_i) = (x(i) / i \in I, i \in \mathbf{B})$ forms a sequence of sets with the filter base.

We define a specific filtered family based on the concept of a context, as defined above. The definition is illustrated in Figure 2. Let Context Set A_1 define all the context sets that can be created out of one given context - this is only one context. Let Context Set A_2 be the sets of contexts that can be created from two given contexts. Context Set A_2 contains each of the contexts and the union of both contexts. This filtered family can continue expanding indefinitely.

Whenever a filtered family contains contexts that describe a single topic in the real world, such as school or festival, we would like to ensure that this set of contexts converges to one ontology concept v, representing this topic, *i.e.*, $A_n \rightarrow_{n \rightarrow \infty} v$. In topology theory, such a convergence is termed a *point of accumulation*, defined as follows.

Let A be a subset of a topological space X; An element $x \in X$ is a *point of accumulation* of the set A if every neighborhood $V(x)$ meets the set $A - x$, that is, if $x \in \overline{A - x}$. Figure 1(bottom) and Figure 2 illustrates ontology concepts as points of accumulation.

To illustrate the creation of an ontology concept let a context be a set containing a single descriptor

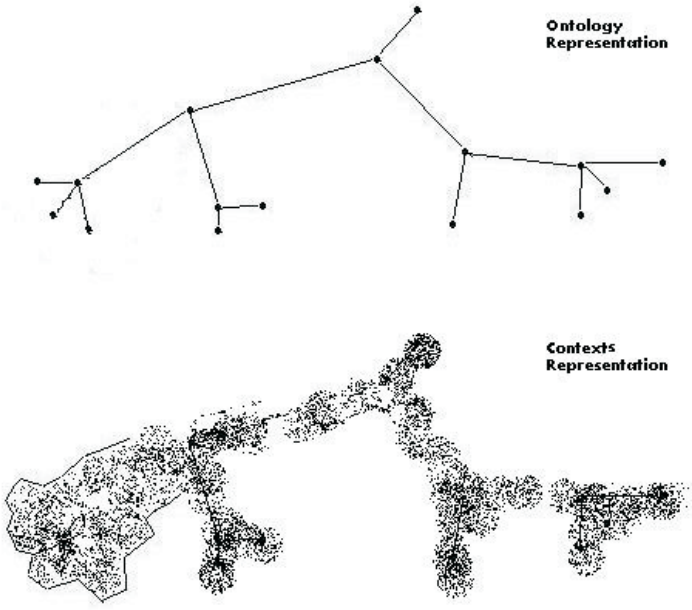

Figure 1: Contexts and Ontology Concepts

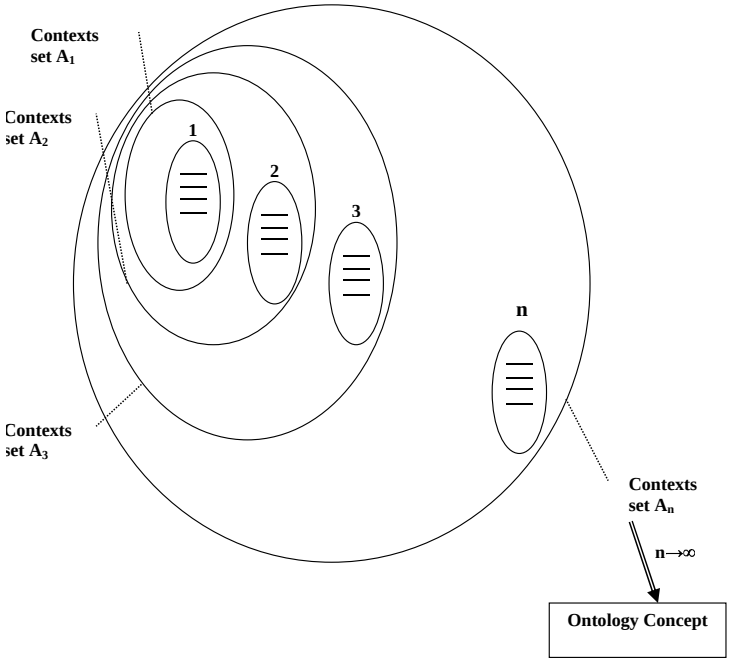

Figure 2: Contexts Sets Converging to Ontology Concept

$\{\langle$Mathematik, 2$\rangle\}$. If we add another context containing a single descriptor of $\{\langle$Musik, 2$\rangle\}$ we form a set of three contexts: $\{\{\langle$Mathematik, 2$\rangle\}$, $\{\langle$Musik, 2$\rangle\}$, $\{\langle$Mathematik, 2$\rangle$, $\langle$Musik, 2$\rangle\}\}$. As the possible sets of descriptors describing documents create an accumulating coverage, we can converge to an ontology concept, such as Long Day School, defined by a set, to which all the contexts set of descriptors belongs.

With infinite possible contexts, can we ensure the existence of a finite number of ontology concepts to which the contexts are mapped? As it turns out, such a guarantee exists in compact topologies. A topological space X is said to be *compact* if every family of open sets $(G_i/i \in I)$ forming a cover of X contains a finite subcover $\{G_{i_1}, G_{i_2}, ..., G_{i_n}\}$. That is, any collection of open sets whose union is the whole space has a finite subcollection whose union is still the whole space. The Bolzano-Weierstrass theorem (Berge 1997) ensures that if X is a compact space, every infinite subset A of X possesses a point of accumulation. Therefore, if the contexts' domain can be covered by a finite cover, such as the number of topics, we can be certain that any infinite set of contexts will accumulate to an ontology concept.

Discussion and Examples

A context can belong to multiple context sets, which in turn can converge to different ontology concepts. Thus, one context can belong to several ontology concepts simultaneously. For example, a context $\langle$Musik, 2$\rangle$ can be shared by many ontology concepts who has interest in culture (such as schools, after school institutes, non-profit organizations, *etc*.) yet it is not in their main role definition. Such overlap of contexts in ontology concepts affects the task of email routing. The appropriate interpretation of a context of an email that is part of several ontology concepts, is that the email is relevant to all such concepts. Therefore, it should be delivered to multiple departments in the local government.

Of particular interest are ontology concepts that are considered "close" under some distance metric. As an example, consider the task of opinion analysis. With opinion analysis, a system should judge not only the relevant area of interest of a given email, but also determine the opinion that is expressed in it. Consider an opinion analysis task, in which opinions are partitioned into two categories (*e.g.*, "for" and "against"). We can model such opinions using a common concept ontology (say, that of Perspectives du Theatre), with the addition of words that describe positive and negative opinions. An email whose context fit with the theme of Perspective du Theatre will be further analyzed to be correctly classified to the "for" or "against" bin. Opinion analysis can be extended to any number of opinions in the same way.

Earlier we have discussed the issue of topological space compactness and its impact on ontology generation. Since there are infinite number of contexts, it may be impossible to suggest a single ontology to which all concepts can be mapped. For local governments, shifting public agenda suggests that a notion of fixed ontology is not at all natural. Nevertheless, we would like to use the Bolzano-Weierstrass theorem to our benefit, and ensure that the contexts domain can be covered by a finite cover, to ensure the existence of points of accumulation.

From the discussion above, it is clear that a fixed ontology cannot serve as a solution. However, when taking a snapshot of a local government, ontology is fixed. Some aspects of the world are beyond the scope of the local government and if we add to the local government ontology a concept that represents all these aspects, we are ensured to have a finite cover of size $n + 1$, with n representing the concepts of current interest. Over time, emails that are beyond the current scope of the local government are accumulated under the $n + 1$ concept, and may be clustered to achieve a new point of accumulation, and thus a new topic of interest in the public agenda.

To summarize, the proposed model employs topological definitions to delineate the relationships between contexts and ontologies. A context is a set of descriptors and their corresponding weights. A filter base is a set of contexts that includes all of their possible unions. If the filter base has a point of accumulation to which the set of contexts converges, then it is defined as an ontology concept. The use of points of accumulations defines ontology concepts to be the union of contexts rather than intersection, as suggested in earlier works. We next turn our attention to the uncertainty inherent in automatic extraction of contexts.

Ranking Ontology Concepts

Up until now, the model we have provided assumed perfect knowledge in the sense that a context is a true representative of a local view and an ontology concept (and its related context) is a true representative of a global view. In the real world, however, this may not be the case. When a context is extracted automatically from some information source (*e.g.*, an email message), it may not be extracted accurately and descriptors may be erroneously added or eliminated. Also, even for manually crafted ontology concepts, a designer may err and provide an inaccurate context for a given concept.

In this section we highlight the uncertainty involved in automatic knowledge extraction and propose a method for managing such uncertainty. In particular, we discuss the impact of uncertainty on the three tasks presented above, namely email routing, opinion analysis, and public agenda.

Context Recognition Algorithms

Several methods were proposed in the literature for extracting context from text. A set of algorithms were proposed in the IR community, based on the principle of counting the number of appearances of each word in the text, assuming that the words with the highest number of appearances serve as the context. Variations on this simple mechanism involve methods for identifying the relevance of words to a domain, using methods such as stop-lists and inverse document frequency. For illustration purposes, we next provide a description of a context recognition algorithm that uses the Internet as a knowledge base to extract multiple contexts of a given situation, based on the streaming in text format of information that represents situations (Segev, Leshno, & Zviran 2004). This algorithm has been used in identifying

context of chat discussions and medical documents, and is currently part of the QUALEG solution.

Let $\mathcal{D} = \{P_1, P_2, ..., P_m\}$ be a series of textual descriptors representing a document, where for all P_i there exists a collection of sets of contexts $\mathcal{C}_{ij}$ so that for each i, $ist(\mathcal{C}_{ij}, P_i)$ for all j. That is, the textual proposition P_i is true in each of the set of contexts $\mathcal{C}_{ij}$. The granularity of the descriptors varies, based on the case at hand, and may be a single sentence, a single paragraph, a statement made by a single participant (in a chat discussion or a Shakespearian play), *etc*. The context recognition algorithm identifies the outer context set $\mathcal{C}$ defined by

$$ist(\mathcal{C}, \bigcap_{i=1}^{m} ist(\mathcal{C}_{ij}, P_i))\forall j.$$

The input to the algorithm is a stream, in text format, of information. The context recognition algorithm output is a set of contexts that attempts to describe the current scenario most accurately. The set of contexts is a list of words or phrases, each describing an aspect of the scenario. The algorithm attempts to reach results similar to those achieved by the human process of determining the set of contexts that describe the current scenario.

The context recognition algorithm consists of four major phases: collecting data, selecting contexts for each text, ranking the contexts, and declaring the current contexts. The phase of data collection includes parsing the text and checking it against a stop-list. To improve this process, the text can be checked against a domain-specific dictionary. The result is a list of keywords obtained from the text. The selection of the current context is based on searching the Internet for relevant documents according to these keywords and on clustering the results into possible contexts. The output of the ranking stage is the current context or a set of highest ranking contexts. The set of preliminary contexts that has the top number of references, both in number of Internet pages and in number of appearances in all the texts, is declared to be the current context. The success of the algorithm depends, to a great extent, on the number of documents retrieved from the Internet. With more relevant documents, less preprocessing (using methods such as Natural Language Processing) is needed in the data collection phase.

From an Automatically Extracted Context to Ontology Concepts

Given the uncertainty involved in automatically extracting contexts, sticking with a strict approach according to which a context belongs to an ontology concept only if it is an element in its associated point of accumulation, may be too restrictive. To illustrate this argument, Let $\mathcal{C}$ be a context in a point of accumulation x and let $\mathcal{C}'$ be an automatically extracted context. The following three scenarios are possible:

$\mathcal{C} \subset \mathcal{C}'$: In this case the context extraction algorithm has identified irrelevant descriptors to be part of the context (false positives). Unless the set of descriptors in $\mathcal{C}'$ that are not in $\mathcal{C}$ is a context in x as well, $\mathcal{C}'$ will not be matched correctly.

$\mathcal{C}' \subset \mathcal{C}$: In this case the context extraction algorithm has failed to identify some descriptors as relevant (false negatives). Therefore, $\mathcal{C}'$ will only be matched correctly if $\mathcal{C}$ is a context in the same filter base.

$\mathcal{C} \not\subseteq \mathcal{C}' \wedge \mathcal{C}' \not\subseteq \mathcal{C}$: This is the case in which both false positives and false negatives exist in $\mathcal{C}'$.

A good algorithm for context extraction generates contexts in which false negatives and false positives are considered to be the exception, rather than the rule. Therefore, we would like to measure some "distance" between an extracted context and various points of accumulation, assuming a "closer" ontology concept to be better matched. To that end, we define a metric function for measuring the distance between a context and ontology concepts, as follows.

We first define distance between two descriptors $\langle c_i, w_i\rangle$ and $\langle c_j, w_j\rangle$ to be:

$$d(c_i, c_j) = \begin{cases} |w_i - w_j| & i = j \\ \max(w_i, w_j) & i \neq j \end{cases}$$

This distance function assigns greater importance to descriptors with larger weights, assuming that weights reflect the importance of a descriptor within a context. To define the best ranking concept in comparison with a given context we use Hausdorff metric. Let A and B be two contexts and a and b be descriptors in A and B, respectively. Then,

$$\begin{aligned} d(a, B) &= \inf\{d(a, b)|b \in B\} \\ d(A, B) &= \max\{\sup\{d(a, B)|a \in A\}, \sup\{d(b, A)|b \in B\}\} \end{aligned}$$

The first equation provides the value of minimal distance of an element from all elements in a set. The second equation identifies the furthest elements when comparing both sets.

Example 2. *Going back to our case study example, the context* $\{\{\langle$*Musik,* $8\rangle\}, \{\langle$*Open Air,* $1\rangle\}\}$ *may be relevant to both Perspective du Theatre and Long Day School, since in both, a descriptor Musik is found, albeit with different weights. The distance between* $\langle$*Musik,* $8\rangle$ *and* $\langle$*Musik,* $6\rangle$ *in Perspective du Theatre is* 2*, and to* $\langle$*Musik,* $2\rangle$ *in Long Day School is* 6*. Assume that* $\{\langle$*Open Air,* $1\rangle\}$ *is a false positive, which does not appear in neither Perspective du Theatre nor in Long Day School. Therefore, its distance from each of the two points accumulation is* 1 *(since* $\inf\{d(a, b)|b \in B\} = 1$*,* e.g.*, when comparing* $\{\langle$*Open Air,* $1\rangle\}$ *with* $\{\langle$*Kulturpolitik, 1*$\rangle\}$*). We can therefore conclude that the distance between the context and Perspective du Theatre is* 2*, which is smaller than its distance from Long Day School (computed to be* 6*). Therefore, Perspective du Theatre will be ranked higher than Long Day School.*

We defer to an extended version of this paper the design of efficient data structures to ensure efficient ranking computation. We now discuss the application of the ranking scheme to the three tasks of email routing, opinion analysis, and public agenda.

Email routing: The user provides QUALEG with a distance threshold t_1. Any ontology concept that matches with a context, automatically generated from an email,

and its distance is lower than the threshold ($d(A, B) < t_1$) will be considered relevant, and the email will be routed accordingly.

Opinion analysis: relevant set of ontology concepts are identified, similarly to email routing. Then for each ontology concept, the relative distance of the different opinions of that concept are evaluated. If the difference in distance is too close to call (given an additional threshold t_2), the system refrains from providing an opinion (and the email is routed accordingly). Otherwise, the email is marked with the opinion with minimal distance.

Public agenda: If all ontology concepts (of the n relevant concepts) satisfy that $d(A, B) \geq t_1$, the email is considered to be part of a new topic on the public agenda, and is added to other emails under this concept. Periodically, such emails are clustered and provided to decision makers to determine the addition of new ontology concepts.

Discussion and Conclusion

The paper presents a topological framework for combining contexts and ontologies in a model that maps contexts to ontologies. Contexts, individual views of a domain of interest, are matched to ontology concepts, often considered to be the "golden standard," for various purposes such as routing and opinion analysis. The model provides a conceptual structure, based on topological definitions, which delineates how and when contexts can be mapped to ontologies. The uncertainty, inherent to automatic context extraction, is managed through the definition of distance among contexts and a ranking of ontology concepts with respect to a given context.

To analyze the context and the mapping of contexts to ontologies, data from a local government, in the form of email messages from citizens, is used. The object of the local government is to analyze the quantities of information flowing in that could not be handled using its human resources. The information is examined to see whether the correct context could be identified and mapped to the right ontology. Since the project involves different countries and different languages, a multilingual ontology system is used. According to the model, different sets of words, representing the same concept, can be mapped to the multilingual ontology.

Each ontology concept was divided into positive and negative citizen opinions about the topics discussed in the email messages. This classification allows the local government to make decisions according to the citizen opinions, which are derived from the information received by email and analyzed only by the algorithm and not by a civil servant.

Initial experiments has yielded reasonable results. The results show that it is possible to automatically perform operations such as information routing and opinion analysis, based on the mapping of contexts and ontologies. We shall briefly provide here a few observations, gathered from the experiments. We defer a complete report on our experiments to an extended version of this paper.

During our experiments with the model we have identified several factors that may contribute to uncertainty. The main reason for errors in ontology concept identification pertains to the preprocessing of the input. The preprocessing was limited to a minimal and naïve dissection of the input. Most of the emails consisted of few sentences only, resulting in a one-shot attempt to determine the correct context. These results could be improved using different preprocessing methods, and the utilization of "soft" NLP tools. The ontology definition, which is currently restricted to a small number of words, also contributed to a low recall rate.

Some problems identified in the mapping of the context to ontology concepts were based on word association. For example, after an email ontology was identified as Perspectives du Theater, an attempt was made to identify its opinion. The number of positive words in the email were counted, and the result was three positive words taken from a predefined list. Therefore, the algorithm identified the opinion as positive. However, a single negative word in the email, not located on the list, transformed the opinion into a negative one. We are currently seeking more advanced techniques to improve opinion analysis. These methods include the analysis of the position of negative and positive words in an email.

As a final comment, we note that the current model assumes the availability of a predefined ontology. Therefore, ontology concepts and their relationships are provided beforehand, and newly extracted contexts are mapped to existing concepts. A possible direction for further research would be to utilize the partial overlapping among contexts to identify ontological relationships, such as generalization-specialization relationships.

Acknowledgments

The work of Gal was partially supported by two European Commission 6^{th} Framework IST projects, QUALEG and TerreGov, and the Fund for the Promotion of Research at the Technion. The authors thank Giora Dula for his useful comments. We thank Amir Taller for his assistance in integrating the Knowledge Extraction component with QUALEG infrastructure.

References

Assadi, H. 1998. Construction of a regional ontology from text and its use within a documentary system. *In Proceedings of the International Conference on Formal Ontology and Information Systems (FOIS-98)*.

Berge, C. 1997. *Topological Spaces*. Dover Publications.

Borgida, A., and Brachman, R. J. 1993. Loading data into description reasoners. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, 217–226.

Bunge, M. 1977. *Treatise on Basic Philosophy: Vol. 3: Ontology I: The Furniture of the World*. New York, NY: D. Reidel Publishing Co., Inc.

Bunge, M. 1979. *Treatise on Basic Philosophy: Vol. 4: Ontology II: A World of Systems*. New York, NY: D. Reidel Publishing Co., Inc.

Choset, H., and Nagatani, K. 2001. Topological simultaneous localization and mapping (slam): Toward exact local-

ization without explicit localization. *IEEE Trans. on Robotics and Automation* 17(2):125–137.

Chung, C. Y.; Lieu, R.; Liu, J.; Luk, A.; Mao, J.; and Raghavan, P. 2002. Thematic mapping from unstructured documents to taxonomies. In *Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM)*.

Doan, A.; Madhavan, J.; Domingos, P.; and Halevy, A. 2002. Learning to map between ontologies on the semantic web. In *Proceedings of the eleventh international conference on World Wide Web*, 662–673. ACM Press.

Donini, F.; Lenzerini, M.; Nardi, D.; and Schaerf, A. 1996. Reasoning in description logic. In Brewka, G., ed., *Principles on Knowledge Representation, Studies in Logic, Languages and Information*. CSLI Publications. 193–238.

Gal, A.; Anaby-Tavor, A.; Trombetta, A.; and Montesi, D. 2005a. A framework for modeling and evaluating automatic semantic reconciliation. *VLDB Journal* 14(1):50–67.

Gal, A.; Modica, G.; Jamil, H.; and Eyal, A. 2005b. Automatic ontology matching using application semantics. *AI Magazine* 26(1).

H. Davulcu, S. V., and Nagarajan, S. 2003. Ontominer: Bootstrapping and populating ontologies from domain specific websites. In *Proceedings of the First International Workshop on Semantic Web and Databases*.

Kashyap, V., and Sheth, A. 1996. Semantic and schematic similarities between database objects: a context-based approach. *VLDB Journal* 5:276–304.

Kashyap, V.; Ramakrishnan, C.; Thomas, C.; and Sheth, A. 2005. Taxaminer: An experimentation framework for automated taxonomy bootstrapping. *International Journal of Web and Grid Services, Special Issue on Semantic Web and Mining Reasoning*. to appear.

Kifer, M.; Lausen, G.; and Wu, J. 1995. Logical foundation of object-oriented and frame-based languages. *Journal of the ACM* 42.

Koenig, S., and Simmons, R. 1996. Passive distance learning for robot navigation. In *Proceedings of the Thirteenth International Conference on Machine Learning (ICML)*, 266–274.

Madhavan, J.; Bernstein, P.; and Rahm, E. 2001. Generic schema matching with Cupid. In *Proceedings of the International conference on very Large Data Bases (VLDB)*, 49–58.

Madhavan, J.; Bernstein, P.; Domingos, P.; and Halevy, A. 2002. Representing and reasoning about mappings between domain models. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence and Fourteenth Conference on Innovative Applications of Artificial Intelligence (AAAI/IAAI)*, 80–86.

Maedche, A., and Staab, S. 2001. Ontology learning for the semantic web. *IEEE Intelligent Systems* 16.

McCarthy, J. 1993. Notes on formalizing context. *In Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*.

McGuinness, D.; Fikes, R.; Rice, J.; and Wilder, S. 2000. An environment for merging and testing large ontologies. In *Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning (KR2000)*.

Melnik, S., ed. 2004. *Generic Model Management: Concepts and Algorithms*. Springer-Verlag.

Mena, E.; Kashyap, V.; Illarramendi, A.; and Sheth, A. P. 2000. Imprecise answers in distributed environments: Estimation of information loss for multi-ontology based query processing. *International Journal of Cooperative Information Systems* 9(4):403–425.

Motro, A., and Rakov, I. 1998. Estimating the quality of databases. *Lecture Notes in Computer Science*.

Noy, F. N., and Musen, M. 2000. PROMPT: Algorithm and tool for automated ontology merging and alignment. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000)*, 450–455.

Papatheodorou, C.; Vassiliou, A.; and Simon, B. 2002. Discovery of ontologies for learning resources using word-based clustering. *Proceedings of the World Conference on Educational Multimedia, Hypermedia and Telecommunications (ED-MEDIA 2002)* 1523–1528.

Remolina, E., and Kuipers, B. 2004. Towards a general theory of topological maps. *Artificial Intelligence* 152:47–104.

Segev, A.; Leshno, M.; and Zviran, M. 2004. Context recognition using internet as a knowledge base. Technical Report TR-04-ISE-1, Technion.

Shatkay, H., and Kaelbling, L. 1997. Learning topological maps with weak local odometry information. In *Proc. IJCAI-97*.

Siegel, M., and Madnick, S. E. 1991. A metadata approach to resolving semantic conflicts. In *Proceedings of the 17th International Conference on Very Large Data Bases*, 133–145.

Simhon, S., and Dudek, G. 1998. A global topological map formed by local metric maps. *In IEEE/RSJ International Conference on Intelligent Robotic Systems* 3:1708–1714.

Spyns, P.; Meersman, R.; and Jarrar, M. 2002. Data modelling versus ontology engineering. *ACM SIGMOD Record* 31(4).

Terziyan, V., and Puuronen, S. 2000. Reasoning with multilevel contexts in semantic metanetwork. In P. Bonzon, M. Cavalcanti, R. N., ed., *Formal Aspects in Context*, 107–126. Kluwer Academic Publishers.

van Rijsbergen, C. J. 1979. *Information Retrieval*. London: Butterworths, second edition edition.

Vickery, B. 1966. *Faceted classification schemes*. New Brunswick, N.J.: Graduate School of Library Service, Rutgers, the State University.