Putting Things in Context: A Topological Approach to Mapping Contexts to 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 to model a domain, while contexts are system generated models. In this work, we provide a formal mathematical framework that delineates the relationship between contexts and ontologies. We then use the model to handle the uncertainty associated with automatic context extraction from existing documents by providing a ranking method, which ranks ontology concepts according to their suitability to a given context. Throughout this work we motivate our research using QUALEG, a European IST project that aims at providing local governments with an effective tool for bi-directional communication with citizens. We empirically evaluated our model using two real-world data sets, coming from Reuters and news RSS. Our empirical analysis shows that the proposed model can be adopted in practice. The input needed to accurately define a concept by a context is small, and the classification of documents to concepts is accurate.
Keywords: Ontology, Context, Topology mapping

1 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 [35]. 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.

Such an approach was recently adopted in QUALEG, a European Commission project aimed at increasing citizen participation in the democratic process.[1] In QUALEG, contexts are used to classify the input from citizens and map them to the services

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

provided by the local governments. In particular, QUALEG was designed to handle routing of emails to departments, opinion analysis on topics at the forefront of public debates, and identification of new topics on the public agenda.

The two classifications (*i.e.*, global *vs.* local and manual *vs.* automatic) are not necessarily at odds. In the example given above, documents may be email messages from citizens, expressing a local view of a domain. The classification of manual *vs.* automatic modeling of a domain has been the center of attention in the area of data integration in the past few years.

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 to a given context.

Throughout this paper, we motivate our work with examples from the eGovernment domain. However, due to the absence of large scale data sets for this domain, we support our model with an empirical analysis using real-world news syndication traces.

Our contributions are as follows:

- We present a framework for combining contexts and ontologies using topological structures.
- We provide a model for ranking ontology concepts relative to a context that deals with the uncertainty inherent in the context extraction and classification.
- 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. Analyzing traces from Reuters and news RSS data, we analyze several aspects of our model, such as the context size required to define a concept and the accuracy of the classification of documents.

The rest of the paper is organized as follows. We first discuss related work in Section 2. Next, in Section 3 we propose a model for combining contexts and ontologies. In Section 4 we present a ranking model to map contexts to ontologies. Section 5 displays the results and analysis of the model implementation. Finally, Section 6 includes concluding remarks and suggestions for future work.

2 Related Work

This section describes related work in three different research areas, namely context representation and extraction, ontology, and topology.

2.1 Context Representation and Extraction

The context model we use is based on the definition of context as first class objects formulated by McCarthy [25]. McCarthy defines a relation $ist(\mathcal{C}, P)$, asserting that a

proposition P is true in a context $\mathcal{C}$. We use this relation in Section 4.1 when discussing context extraction.

It has been proposed to use a multilevel semantic network to represent knowledge within several levels of contexts [42]. 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 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.

A previous work on contexts [39] uses metadata for semantic reconciliation. They define the semantic domain of an attribute as the set of attributes used to define its semantics. Work by [16] uses 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. In this work, we do not focus on ontology generation, which can be performed in any one of various methods, including those mentioned above. In the eGovernment application that we use as a motivation for this work, the existence of an ontology to which contexts should be mapped is assumed.

The creation of taxonomies from metadata (in XML/RDF) containing descriptions of learning resources was undertaken in [32]. 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 [1] employed a clustering tool that utilizes specialized data analysis functions and 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 presented methods based on information retrieval techniques [43] for extracting contextual descriptions from data and evaluating the quality of the process. In Section 5.2 we compare our experiments with text classification using the Latent Semantic Indexing (LSI) approach presented in the work of [14] and [21]. The approach associates word-based vectors to topics in a taxonomy. The underlying idea of LSI is that the aggregate of all the word contexts in which

a given word does and does not appear provides a set of mutual constraints that largely determines the similarity of meaning of words and sets of words to each other.

Methods which included techniques for analyzing quality of information included Motro and Rakov [30] who 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 [28] is based on estimating loss of information in navigating 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 and we shall use them in our empirical evaluation as well.

To demonstrate our method, we propose in Section 4.1 the use of a fully automatic context recognition algorithm that uses the Internet as a knowledge base and as a basis for clustering [35]. Both the contexts and the ontology concepts are defined as topological sets, for which set distance presents itself as a natural choice for a distance measure.

2.2 Ontology

Ontologies have been 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 [4]. Typically, ontologies are represented using Description Logic [2, 9], where subsumption typifies the semantic relationship between terms, or Frame Logic [18], where a deductive inference system provides access to semi-structured data.

Recent work has focused on ontology creation and evolution and in particular on schema matching. Many heuristics were proposed for the automatic matching of schemata (*e.g.*, Cupid [23], GLUE [8], and OntoBuilder [11]), and several theoretical models were proposed to represent various aspects of the matching process [22, 27, 10].

The realm of information science has produced an extensive body of literature and practice in ontology construction, *e.g.*, [44]. Other undertakings, such as the DOGMA project [41], provide an engineering approach to ontology management. Work has been done in ontology learning, such as Text-To-Onto [24], Thematic Mapping [6], OntoMiner [7], and TexaMiner [15] to name a few. Finally, researchers in the field of knowledge representation have studied ontology interoperability, resulting in systems such as Chimaera [26] and Protègè [31].

Our model of an ontology 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. When experimenting with our model we assume an ontology is given, designed using any of the tools mentioned above.

2.3 Topology

In recent years 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 [5]. Other researchers have implemented

topology in metrical information associated with actions [38, 20]. 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 [40]. Remolina and Kuipers present a general theory of topological maps whereby sensory input, topological and local metrical information are combined to define topological maps explaining such information [33].

Following the success of these works, in this work we use topologies and topology theory as a tool of choice for integrating contexts and ontologies. While the tools we use are inherently different from those of [33] and [40], we follow their basic theme of using topology to integrate local views into a global one.

3 A Model of Context and Ontology

In this section we formally define our model of contexts and ontologies (Section 3.1) and propose a topology-based model to specify the relationships between them (Section 3.2). We conclude in Section 3.3 with a discussion and a few examples from the QUALEG project.

3.1 Contexts and Ontologies

We define a descriptor c_i from domain $\mathcal{D}$ as an index term used to identify a record of information [29]. It can consist of a word, phrase, or alphanumerical term. A weight $w_i \in \Re$ identifies the importance of descriptor c_i in relation to the record of information. For example, we can have a descriptor $c_1 = Musik$, and $w_1 = 6$. A *descriptor set* $\{\langle c_i, w_i\rangle\}_i$ is defined by a set of pairs, descriptors and weights. Each descriptor can define a different point of view of the concept. The descriptor set defines all the different perspectives and their relevant weight, which identifies the importance of each perspective.

By collecting all the different view points delineated by the different descriptors we obtain the *context*. A *context* $\mathcal{C} = \left\{\{\langle c_{ij}, w_{ij}\rangle\}_i\right\}_j$ is a set of finite sets of descriptors. 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 can represent the relevance of a descriptor to Doc. In classic Information Retrieval, $\langle c_{ij}, w_{ij}\rangle$ may represent the fact that the word c_{ij} is repeated w_{ij} times in Doc.

Another example which represents a context from a different perspective can be seen if we take two different descriptors of a published article: first, the publication information of the article, such as article title and author, and second, a set of keywords representing the classification topics of the paper. Both descriptors refer to the same paper, but each descriptor provides a different viewpoint of it. An example of descriptors based on the publication information of a document can be the title and category (press release): $\{\{\langle$Theater im Grenzbereich, $2\rangle\}, \{\langle$Pressemitteilung, $1\rangle\}\}$. In addition, the document can be described by a descriptor set: $\{\{\langle$Musik, $8\rangle\}, \{\langle$Open Air, $1\rangle\}\}$. It is worth noting that the context above has two descriptor sets, each with two pairs of a descriptor and a weight.

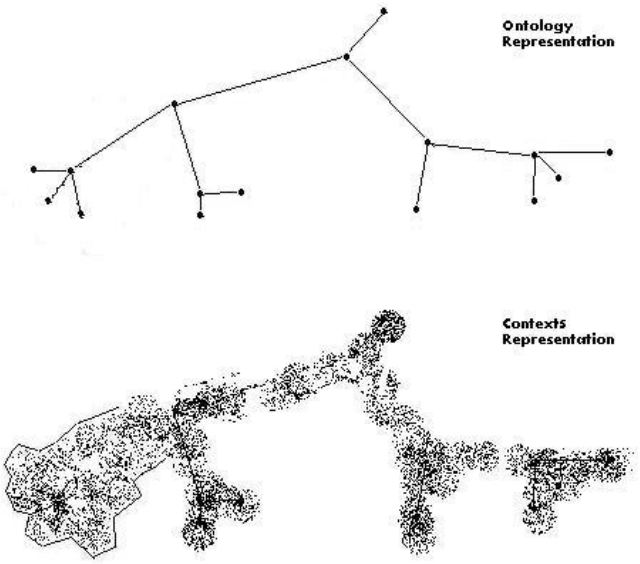

Fig. 1. Contexts and Ontology Concepts

An *ontology* $O = (V, E)$ is a directed graph, with nodes representing concepts (*things* in Bunge's terminology [3, 4]) and edges representing relationships (See Figure 1 (top) for a graphical illustration). A single concept is represented by a name and a context $\mathcal{C}$.

Example 1 (Contexts and ontologies).

To illustrate contexts and ontologies, consider the local government of Saarbrücken. Two ontology concepts in the ontology of Saarbrücken are Perspective du Theatre and Long Day School. The first concept, Perspective du Theatre, is associated with a context that contains descriptors such as:

(Perspectives du Theatre, $\{\{\langle$Öffentlichkeitsarbeit, $2\rangle\}$, $\{\langle$Multimedia, $1\rangle\}$, $\{\langle$Kulturpolitik, $1\rangle\}$, $\{\langle$Musik, $6\rangle\}$, ...$\}$)

and Long Day School is associated with the following context descriptors:

(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 [35] (to be described in Section 4.1) is $\{\{\langle$Musik, $8\rangle\}$, $\{\langle$Open Air, $1\rangle\}\}$. Intuitively, this email may be related to both concepts, possibly with a stronger connection to Perspective du Theatre (due to the higher weight). In this work we demonstrate how such a context can be mapped to ontology concepts.

3.2 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 finite; $\Rightarrow \bigcup_{i \in J} G_i \in \vartheta$
2. $J \subset I \Rightarrow \bigcap_{i \in J} G_i \in \vartheta$
3. $\emptyset \in \vartheta, X \in \vartheta$

The pair (X,ϑ) is called a *topological space* and the sets in ϑ are called *closed sets*.

We now define a context to be a closed set in a topology, representing a family ϑ of all possible contexts in some set X with the subset relation $\subseteq$. X is a set of sets of pairs $\langle c, w \rangle$, 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 real numbers $\Re$).

The topology is defined by the following relation on the context: $\forall \mathcal{C}_a \; \exists \mathcal{C}_b$ such that $\mathcal{C}_a = \left\{ \left\{ \langle c_{ij}, w_{ij} \rangle \right\}_i \right\}_j \subseteq \mathcal{C}_b = \left\{ \left\{ \langle c_{kp}, w_{kp} \rangle \right\}_k \right\}_p$. Identity between contexts is defined as follows: $\mathcal{C}_a = \mathcal{C}_b$ if $c_{kp} = c_{ij}, w_{kp} = w_{ij}, \forall k, p$

The empty set and X are also contexts. Contexts as sets of descriptor sets are closed under intersection and union.

We previously defined contexts as closed sets. Next we define the notion of order of contexts using a *directed set*. A *directed set* is a set ***S*** together with a relation $\geq$, which is both transitive and reflexive, such that for any two elements $a, b \in$ ***S***, there exists another element $c \in$ ***S*** with $c \geq a$ and $c \geq b$. In this case, the relation $\geq$ is said to "direct" the set.

We define a specific directed set using contexts. The definition is illustrated in Figure 2. The different descriptor sets can be viewed as a collection in a bag. We randomly select one descriptor set DS_1. Let Context $\mathcal{C}_1$ define all the descriptor sets that can be created out of one given context - this is only one descriptor set. Let Context $\mathcal{C}_2$ be the sets of descriptors that can be created from two given descriptor sets. Context $\mathcal{C}_2$ contains three descriptor sets: DS_1 from the previous context, DS_2 which is another descriptor set we select, and the union of both descriptor sets, therefore, $\mathcal{C}_1 \leq \mathcal{C}_2$. We can continue and build this directed set by adding another descriptor set to $\mathcal{C}_2$ forming a new Context $\mathcal{C}_3$, where $\mathcal{C}_1 \leq \mathcal{C}_3$ and $\mathcal{C}_2 \leq \mathcal{C}_3$. This process of creating the directed set can continue indefinitely. The contexts directed set is formally defined by:

$\mathcal{C}_0 = \{\emptyset\}$

$\mathcal{C}_n = \{DS_i, DS_i \cup DS_n | \forall DS_i \in \mathcal{C}_{n-1}\}$

This directed set forms a sequence where: $\mathcal{C}_1 \leq \mathcal{C}_2 \leq \mathcal{C}_3 \leq ... \leq \mathcal{C}_n \leq ...$

Whenever a directed set 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.*, $\mathcal{C}_n \rightarrow_{n \rightarrow \infty} v$. In topology theory, such a convergence is termed an *accumulation point*, a point which is the limit of a sequence, also called a limit point. Figure 1 (bottom) and Figure 2 illustrate ontology concepts as points of accumulation. The concept can be viewed as delineating a growing set of descriptors forming the context. The borders outline all of the separate descriptors sets which belong to a specific concept. An overlap between descriptors belonging to different concepts is possible, similarly to dynamic taxonomies [34].

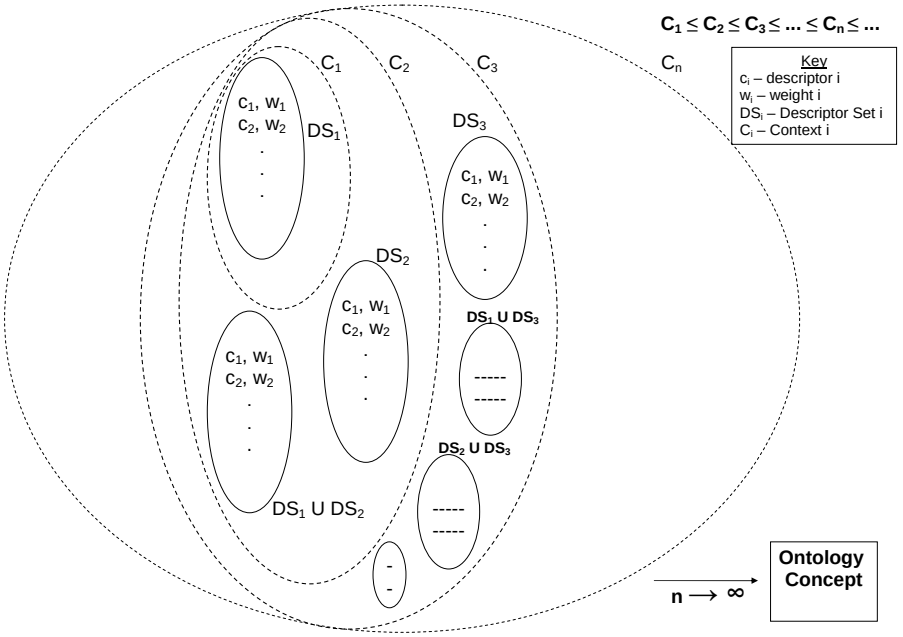

Fig. 2. Contexts Sets Converging to an Ontology Concept

To demonstrate the creation of an ontology concept let a context be a set containing a singleton descriptor set $\{\langle \text{Mathematik}, 2\rangle\}$. If we add another singleton descriptor set of $\{\langle \text{Musik}, 2\rangle\}$ we form a new context which contains three descriptor sets: $\{\{\langle \text{Mathematik}, 2\rangle\}, \{\langle \text{Musik}, 2\rangle\}, \{\langle \text{Mathematik}, 2\rangle, \langle \text{Musik}, 2\rangle\}\}$. As the possible sets of descriptors describing documents increase we advance towards the coverage of the accumulation point. The directed set comprising of these contexts becomes more descriptive. We can converge to an ontology concept, such as *Long Day School*, defined by a set, to which the contexts set belongs. Basically the accumulation point forms the context which includes all the descriptor sets required to define a concept.

With infinite possible contexts, can we ensure the existence of ontology concepts to which these contexts converge? The answer is yes. Looking at the topological definitions, we defined contexts as a subset of a topological space. All of the subsets forming the contexts were defined to be closed sets. According to [17], the following theorem holds in regards to closed sets:

Theorem 1. *A subset of a topological spaces is closed if and only if it contains the set of its accumulation point.*

According to this theorem, any subset of contexts, being closed sets, will necessarily include an accumulation point. If we look at a finite set of descriptor sets, when each time we add another descriptor set, we will obviously reach an accumulation point,

which includes all of the descriptors forming the ontology concepts. However, the above theorem guarantees that even if we have an infinite number of descriptors sets, we will eventually reach an accumulation point, which will also be a context. This context will include all of the descriptor sets defining our concept.

To summarize, prior work [16] has focused on semantic similarity, which is essentially an abstraction / mapping between the domains of the two objects associated with the context of comparison. The work presented here uses points of accumulation to define ontology concepts to be the union of contexts rather than the intersection, as suggested in earlier works. 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 directed set is a relation of contexts that includes all of their possible unions of sets of descriptors. An ontology concept is the accumulation point of the directed set of contexts.

3.3 Discussion and Examples

A context can consist of multiple descriptor sets. Each descriptor set can belong to several ontology concepts simultaneously. For example, a descriptor set $\{\langle \text{Musik}, 2\rangle\}$ can be shared by many ontology concepts that have interest in culture (such as schools, after school institutes, non-profit organizations, *etc.*) although it is not in their main role definition (and hence the low weight assigned to it). Such overlap of contexts in ontology concepts affects, for example, 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 not only judge 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 fits with the theme of Perspective du Theatre will be further analyzed to be correctly classified as "close" to the "for" or "against" category. Opinion analysis can be extended to any number of opinions in the same manner.

4 Ranking Ontology Concepts

Up until now, the proposed model 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 [12] we also argued that

even a well-crafted ontology may vary slightly between organizations within the same domain, such as local governments. Therefore, contexts are bound to vary as well.

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 in the QUALEG project. As a basis for our discussion, we first present the principles of a context recognition algorithm. Details of the algorithm are provided in [35] and the description here is given for the sake of completeness.

4.1 A Context Recognition Algorithm

Several methods were proposed in the literature for extracting context from text. A class of algorithms were proposed in the IR community, based on the principle of counting the number of appearances of each word in a 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. This algorithm was adapted from [35] and is currently part of the QUALEG solution. We use the work in [35] to demonstrate the feasibility of our model. However, other models, such as [21] and [13], can be adopted for context recognition as well.

Let $\mathcal{D} = \{P_1, P_2, ..., P_m\}$ be a set of textual propositions representing a document, where for all P_i there exists a collection of descriptor sets forming the context $\mathcal{C}_i = \{\langle c_{i1}, w_{i1}\rangle, ..., \langle c_{in}, w_{in}\rangle\}$ so that $ist(\mathcal{C}_i, P_i)$ is satisfied. That is, the textual proposition P_i is true for context $\mathcal{C}_i$. The granularity of the textual propositions 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 $\mathcal{C}$ defined by

$$ist(\mathcal{C}, \bigcap_{i=1}^{m} ist(\mathcal{C}_i, P_i)).$$

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 algorithm attempts to reach results similar to those achieved by a human when determining the set of contexts that describe the current scenario.

The context recognition algorithm consists of the following 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, 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 a greater number of relevant documents, less preprocessing (using methods such as Natural Language Processing) is needed in the data collection phase.

4.2 From an Automatically Extracted Context to Ontology Concepts

Given the uncertainty involved in automatically extracting contexts, adhering to 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 that is an accumulation point 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 directed set.
- $\mathcal{C} \nsubseteq \mathcal{C}' \wedge \mathcal{C}' \nsubseteq \mathcal{C}$**:** 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 c_i and c_j with their associated weights w_i and w_j 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,

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

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 between $\langle$Musik, $8\rangle$ and $\langle$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 either Perspective du Theatre or in Long Day School. Therefore, its distance from each of the two points of 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.

Although a normalization step can be used to prevent descriptors with large frequencies from influencing the results - in the previous example this will lead to lowering the value of the weight of Musik - there is an advantage in leaving the higher value descriptors so as to give these descriptors more weight in the process, in order to better represent the weights of the contexts. A higher value means that these descriptors are 'dominant.'

We next show how the proposed ranking mechanism can be utilized for the various tasks of eGovernment, as presented in Section 3.3.

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: The relevant set of ontology concepts is identified, similarly to email routing. Then for each ontology concept, the relative distance of the different opinions of that concept is 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.

5 Experiences and Experiments

In this section we first present the QUALEG project as a test platform for the implementation of the model that maps context to ontology in Section 5.1. We also share our experiences with QUALEG pilots. Next, in Section 5.2 we empirically evaluate various aspects of the proposed model using two real-world data traces, taken from Reuters and news RSS.

5.1 QUALEG Experiences

The section begins with a description of the QUALEG architecture, followed by experiences with e-mail routing and opinion analysis.

QUALEG Architecture The aim of the QUALEG project is to support the electronic interactions between civil servants and citizens. The QUALEG system aims at allowing local governments to maintain a direct connection with citizens through the ongoing adjustment of their policies according to the assessment of citizen needs. This implies that local governments should be able to measure the performance of the services they offer, assess citizen satisfaction, and re-formulate policy orientations on such elements with the participation of citizens.

These tasks are achieved through the implementation of an agent oriented Qualeg architecture, which consists of the following main seven components: (1) Agora - A Web interface to the system through which the citizen interacts via emails, chats and forums with the civil servant. (2) Datamart - The component that stores Qualeg data. (3) Qualeg ontology - A multilingual ontology describing the public and e-Government issues. (4) Knowledge Extractor - The previously described context extraction algorithm activated by the software agents. (5) Qualeg Workflow - The component that handles the flow of processes relevant to the public servants and administrations. (6) A set of agents, which in the backstage handle the main control of the Qualeg system, acting asynchronously and handling the data to be communicated among various modules. (7) A set of Web services offered for seamless data handling to and from the Datamart.

The Knowledge Extraction Agent (KE Agent) has the responsibility to trigger the Knowledge Extraction Module so that the context of the stored information is regularly analyzed. The Knowledge Extraction architecture is illustrated in Figure 3. There are four types of documents that should be analyzed: documents uploaded to AGORA, text in forums, chats, and incoming e-mail messages. In particular, the KE Agent performs periodical searches in the platform's databases for new information to be analyzed. Every transaction with the database is carried out by means of Web services. If new documents are found, the agent triggers the previously described knowledge extraction algorithm on them. Hence, the KE Agent parses all the required information - such as document id, document name, document url - to the KE module. The KE module performs the mapping with reference to an ontology, which defines the set of concepts and their relationships. After the KE process is completed, the context of the document is stored in a database.

Similarly to the KE Agent, the Opinion Analysis Agent (OAAgent) regularly searches in QUALEG's databases to find which documents have to be analyzed by the Opinion Analysis Module (OA Module). Once again, all the agent's database transactions are carried out through Web service calls. If documents requiring analysis are found, the agent triggers the opinion analysis algorithm on them in the same way as the KE agent. Opinion Analysis output is an ontology concept related to an opinion and a list of words representing the context extracted.

The platform for analyzing the information was written in Java running on a standard PC. The processing time was divided into several intervals over a few days to

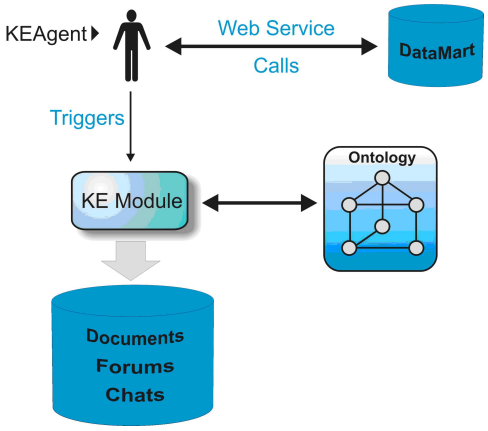

Fig. 3. Knowledge Extraction Architecture

avoid excessive use of the Internet resources employed by the algorithm: Looksmart, Wisenut, Open Directory, Ask, Sponsored Listings, MSN Search, and Vivísimo.

QUALEG Email Routing Our first experience involved the Perspectives du Theatre Festival held during May every year in Saarbrücken, located at the French border of Germany. The festival includes contemporary French theatre, films, street events, music, *etc*. Our challenge was to analyze the festival material and provide a useful set of classifications so that the materials could be rapidly understood and routed to the appropriate civil servants.

The data we received included daily communications (in German) about this event, for a total of 104 emails, primarily emails from citizens to the city hall and press releases and announcements from the city outward. The festival is an annual event and we were given data from 2004 and 2005.

The goal of the topic classification experiment was to identify the topic of an email according to a predefined set of ontology concepts as supplied by Saarbrücken for organizing cultural events. The concepts were Organisation, Veranstalter, Finanzen, Besucher, Informationen, Rahmenprogramm, Spielplan, and Other. Each ontology concept was accompanied by a (manually designed) context that describes it.

To evaluate the proposed model we used a single ontology and two different methods to define and extract contexts. One method was that described in Section 4.1. This method used the technique of mapping contexts to ontology concepts (C2O), as detailed in Section 4.2. The other method was based on conventional Natural Language Processing (NLP) techniques, enhanced by a language domain expert to build a set of rules for identifying relevant words and grammar relevant to the German language. The NLP technique was evaluated with the support of researchers from the University of Southern California's Information Sciences Institute (ISI).

The two techniques are very different. The former is language independent, making it more suitable for multilingual environments at the possible cost of lacking language specific analysis tools used by the latter. The C2O technique uses the Internet as a knowledge base for extracting contexts. These contexts were searched against a list of descriptors that describe concepts in the ontology. The technique was based on a per sentence analysis. For each sentence a classifier was used, automatically trained on keywords and morphological variants (based on the initial list of topics from Saarbrücken). Each sentence in the input was searched against the list of keywords and morphological variants. The NLP technique started from an identical contexts list as the C2O and used the morphological variants of each context.

For both methods the input was parsed at the granularity of sentences. The C2O preprocessing included only partitioning of long sentences according to the search engine requirements. The NLP preprocessing included a Tokenizer, a tool for breaking up compound nouns, and a German Demorpher (Morphy engine), downloaded from the University of Stuttgart (http://www.lezius.de/wolfgang/morphy/). The Demorpher removes case markings, tense markings, *etc*.

Two different experiments were performed. The first experiment was to analyze our model based on the German data. The C2O method achieved a Precision of 85.37%, a Recall of 84.34%, and a total F-Score of 84.85%. This is based on the comparison of the results of the Context Recognition component to that of a human expert. The German input data was classified by two German Language experts and by Saarbrücken local government civil servant employees.

The second experiment analyzed the performance of C2O compared to the NLP technique. In this experiment a subset of 72 emails representing data from a single year was used for comparison. The results showed that C2O achieved an F-Score of 81% while the NLP technique achieved an F-Score of 78%. The results therefore show that the proposed topology-based model of contexts and ontology achieved comparable performance to the NLP technique, with the added value of being language independent.

Opinion Analysis Opinions can be viewed as perspectives expressed in the input information. We modeled opinions to be included in the ontology as concepts, associated with context that provide the local interpretation of each opinion.

There is a difference between the email routing task based on the knowledge extraction and opinion analysis. The knowledge extraction avoids the language specific implementation and bases its analysis techniques on the use of a large corpus of relevant documents taken from the Internet, while the opinion analysis uses techniques from IR and NLP to improve content understanding. As in the knowledge extraction, the results of the opinion analysis are mapped to concepts in the ontology, in this case, opinion concepts. Opinions can be divided into an array of possibilities from extreme positive to very negative. The opinions selected for the experiment were defined in only three categories of concepts - positive, negative, and neutral.

The experiment included 72 emails in German. A set of approximately 6000 opinion verbs and 6000 opinion adjectives taken from ISI [19] were analyzed in English and translated using an online dictionary translation to German. These opinion words are associated with the three opinion concepts. Two possibilities were examined: first, to

translate the emails into English and then analyze the texts for opinion, and second, to translate opinion words. The latter alternative was found to achieve better results, since considerably fewer words are translated, reducing the impact of natural language ambiguity.

The opinion analysis experiments reached a precision of 78.95%, recall of 69.23%, and F-score of 73.77%. The results indicate that it is feasible to use the model to perform opinion analysis, albeit at a lower accuracy than that of routing.

5.2 Experiments

This paper models the relationship between contexts and ontologies as a topology. We experimented with data from Reuters corpus and from news RSS. We start with a description of the two real world data traces and experiment set-up, followed by description of our experiments and an empirical analysis of the results.

The following experiments analyze the model to show how fast the contexts accumulate to a concept, how the quality of the context is attributed to a high number of descriptors from a single context or a multiple contexts, and how concept overlap influences the context representation. The following experiments demonstrate model performance with multiple concepts and analyze how the model performs in a dynamic setting, where a context can be associated with more than a single concept.

Data Sets and Metrics In this paper we present a model of context and ontology relationships. When analyzing our model, we compare it to similar methods of text analysis, which belong to the field of text categorization. This field includes methods of analysis to identify the category to which a given text belongs. Text categorization usually allows text to belong to only one category. Since this classification is rigid and less relevant to our requirements, the model developed in this paper allows a text to belong to more than one category.

The two data traces we used come from Reuters and CNN RSS. In these data traces data are partitioned to topics with no ontological relationships. The experiments focus on the concepts/contexts relationships, for which these data sets serve adequately. Research and experiments on ontological relationships using contexts are reported in [36].

The Reuters data set was taken from a publicly available trace (http://about.reuters.com/researchandstandards/corpus/). We chose 10 news topic categories (referred to hereafter as concepts), for a total of 3,125 data, where a datum is a Reuters news article. The RSS trace was collected during August 2005 from the CNN Web site. Here, we also chose 10 news topic categories including 1,130 data, where a datum is an RSS news header or a news descriptor. The main difference between the Reuters trace and the RSS trace is the datum size. Table 1 describes the two data sets. Concept overlap is explained shortly.

We generated a context for each concept using the C2O algorithm. This context is referred to as context* and the data that was used for this context generation is referred to hereafter as the context* data. The number of data items that were used for generating context* data varies, ranging from one datum to 170 data items for Reuters and from

Table 1. Reuters and RSS Data Set Statistics

Data Set	Reuters	RSS
Size	3,125	1,130
Categories	10	10
Datum per Category	126 - 510	113
Minimum Concept Overlap	21.6%	12.1%
Maximum Concept Overlap	75.5%	86.7%

one datum to 61 for RSS. We also varied the number of context descriptors that were generated from each datum in the context* data, ranging from 1 to 70 descriptors. A varying number of concepts was used, ranging from 1 to 10 concepts depending on the experiment. We use the C2O algorithm, adapted from [35], as an example of a context generator. C2O is known to have generated reasonable contexts in the past (see experiments in [35]).

Given two concepts and their associated contexts, *concept overlap* is defined to be the ratio of the number of common descriptors in both contexts and the minimum context size. Table 1 presents statistics about the minimum and maximum concept overlap found in the data sets.

As a measure of evaluation we use recall and precision metrics. Given a context* $\mathcal{C}$, the recall of a context $\mathcal{C}'$ is defined as the ratio of the number of common descriptors and the size of context* $\mathcal{C}$. A high recall measure means that the C2O algorithm was able to identify correctly a good portion of a context, minimizing false negative. We measure precision with respect to the original classification of data items to categories as given in the data traces. Therefore, the precision of a classification task using contexts is measured as the ratio of the number of correctly classified data items and the number of data items in the experiments. It is worth noting that in most of the experiments the C2O algorithm classifies a datum to a concept whose contexts share the highest number of descriptors with its context, thus setting a lower bound on the algorithm performance. It is also worth noting that QUALEG required using all concepts whose context* shares more than a threshold number of descriptors with a document context. A middle ground may require using a top-K ranked concepts. This leads to a decrease in the precision, yet it increases the recall. Minimizing false positive, in turn, increases the chances of correctly matching data to concepts.

Experiment Results In the first experiment, we evaluated the algorithm ability to generate good representative contexts for concepts (context*). In each experiment we selected a single concept and generated a context using context* data. For each of the remaining data in this category a context was generated and compared with the context*. For each context* data size we repeated the experiment 10 times, each time choosing randomly the context* data. In this setting the average recall and precision over all experiments with the same context* data size is the same.

A graphic illustration of our results is given in Figure 4, which displays the average recall, computed over 10 different ontology concepts in Reuters and news RSS. The

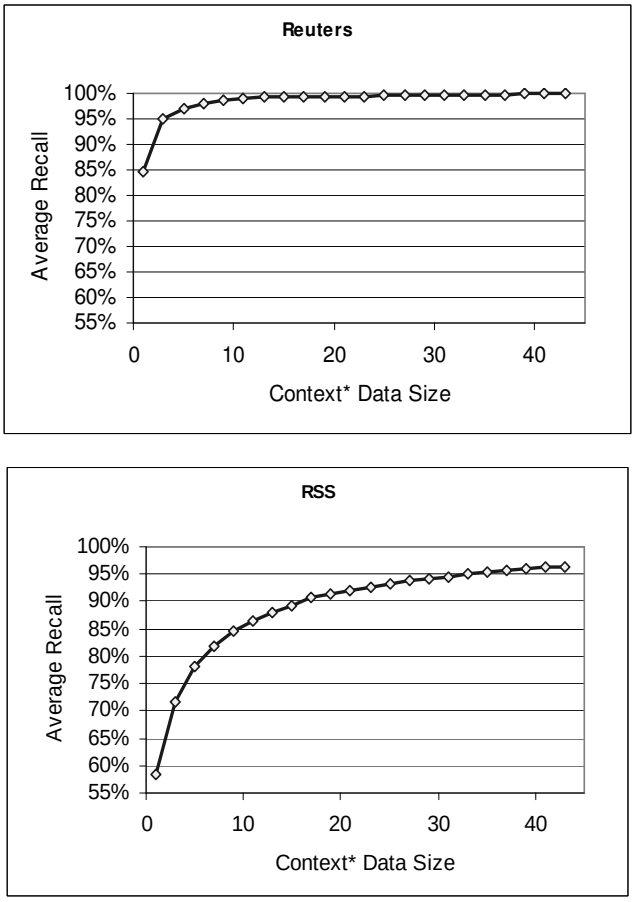

Fig. 4. Number of Contexts for Concept

horizontal axis displays the context* data size. In this experiment each context was limited to 10 descriptors. The vertical axis displays the average recall.

The experiment results indicate that as the number of contexts defining each ontology concept increases, the concept definition improves. While this is to be expected, we also observe that the average recall quickly approaches 100%, although at a different rate for the two data sets. We attribute this different behavior to the sensitivity of the C2O algorithm to different datum length in the two data sets. While Reuters datum is a complete news article, RSS datum contains a header or a short description. Our experience shows that the C2O algorithm we use generates better contexts when given longer texts.

A per concept analysis shows an average recall ranging from 98.86% to 100% in the Reuters data set and from 91.67% to 99.62% in the news RSS data set, when context*

is defined using up to 43 descriptors. Table 2 and Table 3 present a per-concept analysis showing the range of the number of descriptors required to achieve a certain average recall level. It is worth noting that for the Reuters data set some concepts require a training data set of size 5 to achieve 100% recall. There was only one concept that did not reach 100% using a training data set of 43 data items.

Table 2. Average Recall Level vs. Number of Descriptors - Reuters

Average Recall Level	Number of Descriptors
100%	5 - 43
[95% - 100%)	1 - 11
[90% - 95%)	1 - 5
$<$ 90%	1 - 3

Table 3. Average Recall Level vs. Number of Descriptors - RSS

Average Recall Level	Number of Descriptors
[95% - 100%)	15 - 43
[90% - 95%)	9 - 43
$<$ 90%	1 - 35

We next compare two different methods for collecting context descriptors. Method 1 is based on collecting an increasing number of descriptors from a single textual datum. As we increase the number of descriptors, we add to the single context more descriptors with lower relevance value, possibly increasing uncertainty. Method 2 is based on using different descriptors sets, each one based on a different textual datum. In this set of experiments, we chose a single concept for which a context was generated based on a context* and then compared against the remaining data items associated with this category. For Method 1 to achieve a context of size M, the M top descriptors are chosen. For Method 2 the context* was increased, each datum adding up to 5 descriptors to the context*, until the desired number of descriptors was reached.

The results are displayed in Figure 5. As the number of descriptors defining a single context grows, the average recall improves for both methods, yet Method 2 converges faster than Method 1. For the Reuters data, both methods approach 100% recall, while for the RSS data Method 1 performs significantly worse than Method 2. We also observe that recall reaches 100% for 30 descriptors in the Reuters data set while not reaching 100% even for 70 descriptors for the news RSS data set. This is again attributed to the sensitivity of the C2O algorithm to the length of the processed text.

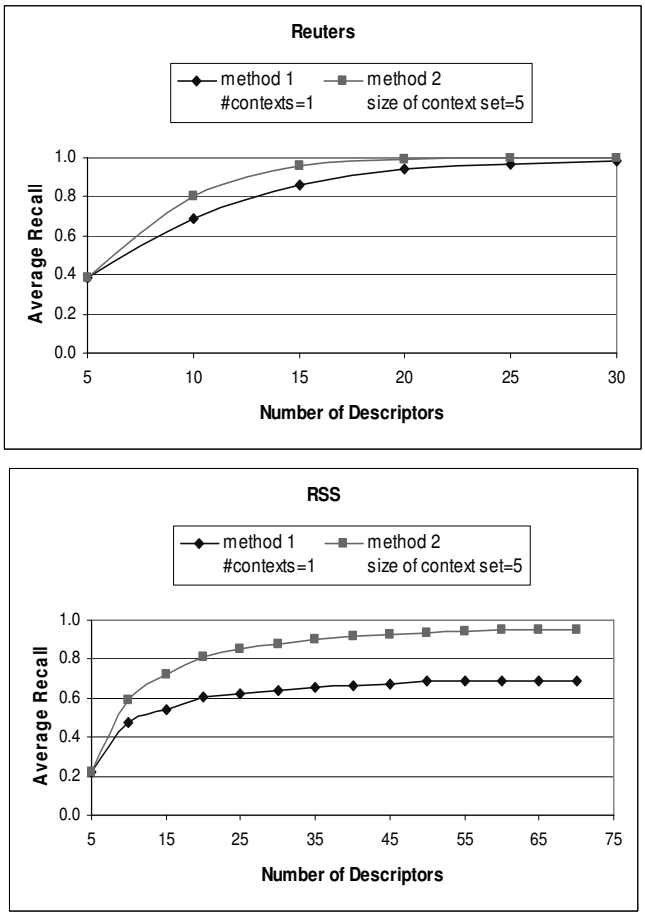

Fig. 5. Comparison of Context Collection Methods Recall

Comparing further the two context generation methods, we took concept pairs and generated contexts* for each concept. We then classified each of the remaining data items to one of the two concepts.

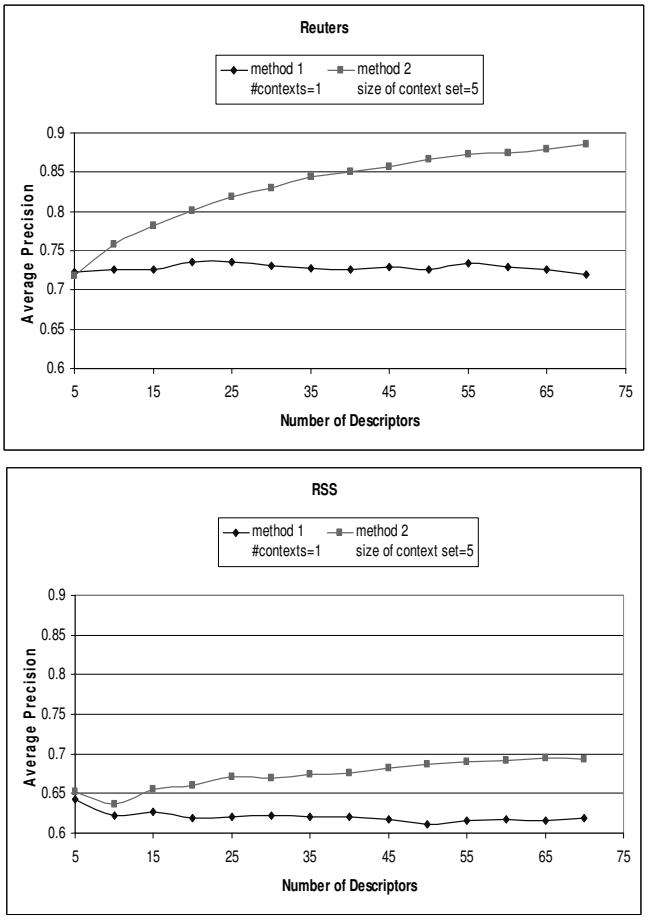

Fig. 6. Comparison of Context Collection Methods Average Precision for Varying Context Sizes

Figure 6 displays the average precision rate for all pairwise combinations of concepts. The horizontal axis displays the number of descriptors in the context* data set and the vertical axis displays the average precision. For both data sets average precision generally increases for Method 2 while remaining unchanged or even decreasing for Method 1. We can conclude that Method 2 performs better and therefore we shall use this method in the remaining experiments.

To evaluate the impact of concept overlap on precision we compared two concept pairs, namely the pair with minimum concept overlap (21.6% for Reuters and 12.1% for RSS) and the pair with maximum concept overlap (75.5% for Reuters and 86.7% for RSS). For each concept in a pair we randomly chose a context* data set and generated a context* for this concept. Then, we classified the remaining data similarly to what was described earlier. The experiments were repeated for various context* data set sizes.

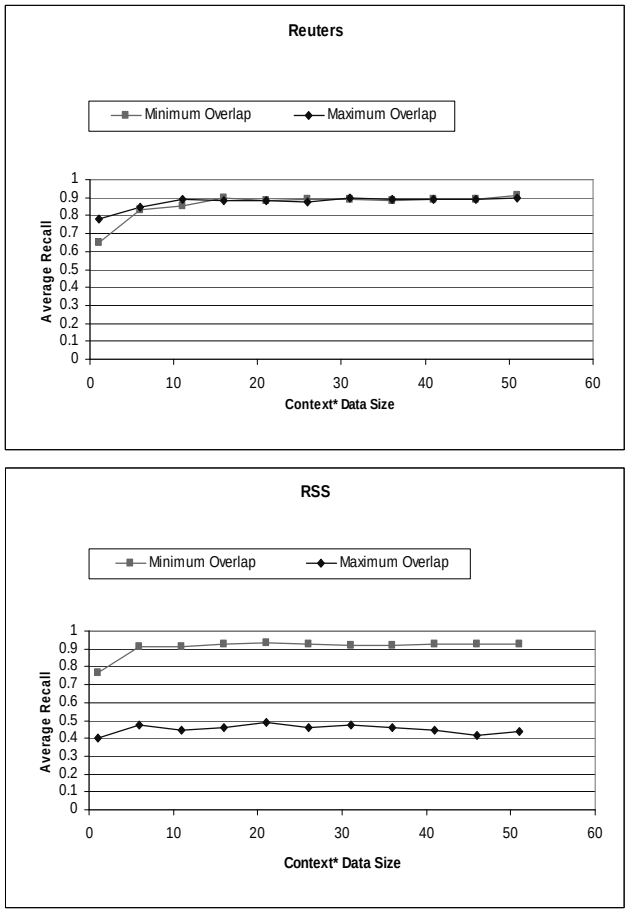

Fig. 7. Concept Overlap and Recall

The results of these experiments are illustrated in Figure 7. We observe different phenomena for each of the data sets. For the Reuters data the average recall converges for a training set of size 30, while for the RSS data the concept pair with a high overlap shows a significantly lower average recall for all tested training sizes. At this time we

are unable to explain these differences. For the news RSS data set there is a wider difference between the minimum overlap pair and maximum overlap pair, which may partially explain this phenomenon.

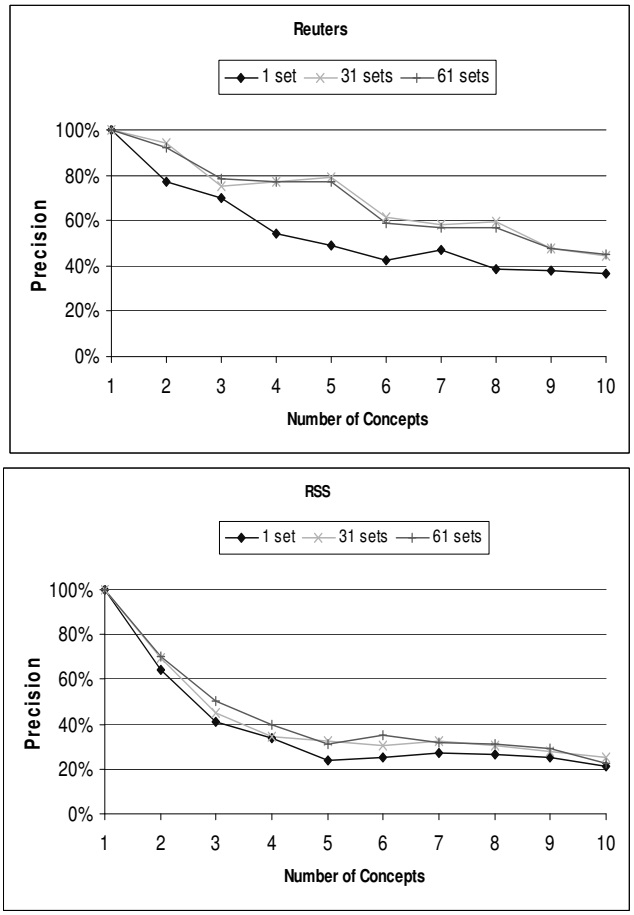

Fig. 8. Number of Concepts

Next, we analyze the impact of the number of concepts on the classification procedure. We repeated the experiment discussed earlier for an increasing number of concepts, varying the size of the context* data set as well. For this set of experiments we have enforced a rigid classification scheme, in which each document is forced to be classified to a single concept. The experiment results are summarized in Figure 8. The horizontal axis displays the number of concepts and the vertical axis presents the classification precision. Each curve represents a different context* data set size. As the

number of concepts increases, the precision declines. It can also be seen that precision improves as we increase the training set size. The marginal effect of increasing the training set size becomes, however, less significant as we increase the training set size. We note again that these precision curves serve as a lower bound on the algorithm efficiency. By using top-K ranked concepts, rather than a single concept, precision decreases toward a lower bound.

To analyze a document belonging to multiple concepts, a random set of concepts and documents was selected as training set. Each descriptor set included ten descriptors. From the ten possible concepts the number of concepts each document can belong to was increased from one to five concepts. The number of concepts used is the X axis. The number of concepts that can be selected is each line (1 to 5) - number of possible concepts. The recall results are presented in Figure 9. For a given K of classified concepts, recall decreases as the number of available concepts increases, as expected. However, an increase from 50% to 91.1% in recall is observed as we increase from one to up to five, out a total of ten, the number of top ranking concepts selected per document. Similarly, for the RSS data set, recall increased from 24.2% to 73.8%.

The precision results for both Reuters and RSS are presented in Figure 10. Both Reuters and RSS results show that the best precision results occur when the number of concepts to which each document can belong is two.

Finally, as a point of reference, we chose a pure vector-space technique [14], [21] of the LSI approach. The two approaches share some similarity in the goal of classifying text into predefined categories. There are also differences. First, the method we used is not based on predefined word-based vectors but rather on a bag of words. This bag of words is determined using words that are not necessarily extracted from the text itself but are associated with a possible context. The size of the word comparison set is not predetermined as in the vector-space method. Another advantage we see to C2O is the use of the Internet as a knowledge base. The Internet allows a set of descriptors to be constantly and automatically updated.

When analyzing our results and the results of the LSI approach, the emphasis on the recall should be taken into consideration. The approach presented here achieved good results in the recall. Although the statistics analysis is not identical in both approaches, the F-score achieved in [21] was 71.1% to 85.3%. The approach presented here with identical weights to the recall and precision, similar to the LSI approach, was from 58% to 100%. The results seem to be especially high when the top five concept categories for recall and top two concepts for precision are used.

6 Discussion and Conclusion

The paper presents a topological framework for combining contexts and ontologies in a model that maps contexts to ontology. Contexts, individual views of a domain of interest, are matched to concepts of an ontology, often considered to be the “golden standard,” for various purposes such as email 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 in automatic

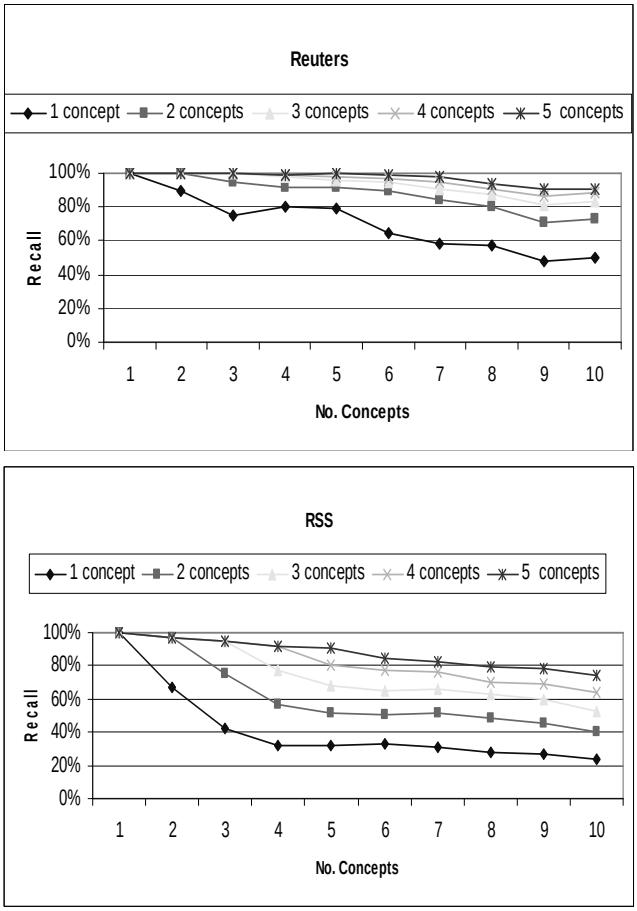

Fig. 9. Recall Number of Possible Concepts

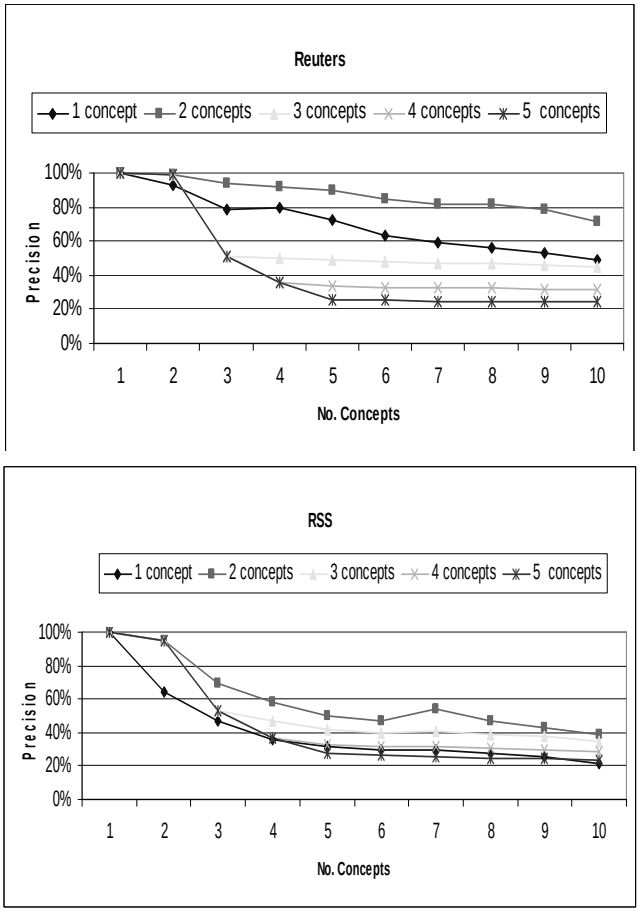

Fig. 10. Precision Number of Possible Concepts

context extraction, is managed through the definition of distance among contexts and a ranking of ontology concepts with respect to a given context.

The proposed model has been implemented as part of QUALEG, an eGovernment project. In this project, information that flows into a local government system is automatically examined, and based on its context, its positioning within the ontology is determined. Since the project involves different countries and different languages, a multilingual ontology system is used. According to the model, different descriptor sets, representing the same concept, can be mapped to the multilingual ontology. To support opinion analysis, 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.

During our experiments with the model, we identified several factors that may contribute to uncertainty. The main reason for errors in ontology concept identification pertains to the preprocessing of the input, which was limited to a minimal and naïve dissection of text. 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 by using different preprocessing methods and utilizing "soft" NLP tools. The ontology definition, which is currently restricted to a small number of words, also contributed to a lower recall rate.

To evaluate empirically the model properties in a controlled environment, we used two real-world data traces, Reuters news reports and RSS news headlines. In these experiments we measured the effectiveness of generating contexts automatically for different concepts. We tested various methods for context extraction and examined the impact of concept overlap and number of concepts on classification quality. We can conclude that the proposed model associating a context with each ontology concept is feasible and the amount of data needed for automatically generating contexts for concepts is relatively small. Context generation can be improved, and we leave such improvements for future research.

In QUALEG the availability of a predefined ontology is assumed. 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. An initial step in this direction is presented in [37].

Acknowledgments

The work of Segev and 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. We thank Amir Teller for his assistance in integrating the Knowledge Extraction component with QUALEG infrastructure and Yulia Turchin for her assistance in experimenting with Reuters and RSS data sets.

References

1. H. Assadi. 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)*, 1998.
2. A. Borgida and R. J. Brachman. Loading data into description reasoners. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 217–226, 1993.
3. M. Bunge. *Treatise on Basic Philosophy: Vol. 3: Ontology I: The Furniture of the World*. D. Reidel Publishing Co., Inc., New York, NY, 1977.
4. M. Bunge. *Treatise on Basic Philosophy: Vol. 4: Ontology II: A World of Systems*. D. Reidel Publishing Co., Inc., New York, NY, 1979.
5. H. Choset and K. Nagatani. Topological simultaneous localization and mapping (slam): Toward exact localization without explicit localization. *IEEE Trans. on Robotics and Automation*, 17(2):125–137, 2001.
6. C. Y. Chung, R. Lieu, J. Liu, A. Luk, J. Mao, and P. Raghavan. Thematic mapping from unstructured documents to taxonomies. In *Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM)*, 2002.
7. H. Davulcu, S. Vadrevu, and S. Nagarajan. Ontominer: Bootstrapping and populating ontologies from domain specific websites. In *Proceedings of the First International Workshop on Semantic Web and Databases*, 2003.
8. A. Doan, J. Madhavan, P. Domingos, and A. Halevy. Learning to map between ontologies on the semantic web. In *Proceedings of the eleventh international conference on World Wide Web*, pages 662–673. ACM Press, 2002.
9. F.M. Donini, M. Lenzerini, D. Nardi, and A. Schaerf. Reasoning in description logic. In G. Brewka, editor, *Principles on Knowledge Representation, Studies in Logic, Languages and Information*, pages 193–238. CSLI Publications, 1996.
10. A. Gal, A. Anaby-Tavor, A. Trombetta, and D. Montesi. A framework for modeling and evaluating automatic semantic reconciliation. *VLDB Journal*, 14(1):50–67, 2005.
11. A. Gal, G. Modica, H.M. Jamil, and A. Eyal. Automatic ontology matching using application semantics. *AI Magazine*, 26(1), 2005.
12. A. Gal and A. Segev. Putting things in context: Dynamic eGovernment re-engineering using ontologies and context. In *Proceedings of the 2006 WWW Workshop on E-Government: Barriers and Opportunities*, 2006.
13. A. Hotho, S. Staab, and A. Maedche. Ontology-based text clustering. In *Proceedings of the IJCAI-2001 Workshop Text Learning: Beyond Supervision*, 2001.
14. V. Kashyap, S. Dalal, and C. Behrens. Professional services automation: A knowledge management approach using lsi and domain specific ontologies. In *Proceedings of the 14th International FLAIRS Conference (Florida AI Research Symposium), Special trach on AI and Knowledge Management*, 2001.
15. V. Kashyap, C. Ramakrishnan, C. Thomas, and A. Sheth. Taxaminer: An experimentation framework for automated taxonomy bootstrapping. *International Journal of Web and Grid Services, Special Issue on Semantic Web and Mining Reasoning*, September 2005.
16. V. Kashyap and A. Sheth. Semantic and schematic similarities between database objects: a context-based approach. *VLDB Journal*, 5:276–304, 1996.
17. J. Kelley. *General Topology*. American Book Company, 1969.
18. M. Kifer, G. Lausen, and J. Wu. Logical foundation of object-oriented and frame-based languages. *Journal of the ACM*, 42, 1995.
19. S. M. Kim, D. Ravichandran, and E. Hovy. ISI novelty track system for trec 2004. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, 2004.

20. S. Koenig and R. Simmons. Passive distance learning for robot navigation. In *Proceedings of the Thirteenth International Conference on Machine Learning (ICML)*, pages 266–274, 1996.
21. T. Liu, Z. Chen, B. Zhang, W.-Y. Ma, and G. Wu. Improving text classification using local latent semantic indexing. In *ICDM*, pages 162–169, 2004.
22. J. Madhavan, P.A. Bernstein, P. Domingos, and A.Y. Halevy. 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)*, pages 80–86, 2002.
23. J. Madhavan, P.A. Bernstein, and E. Rahm. Generic schema matching with Cupid. In *Proceedings of the International conference on very Large Data Bases (VLDB)*, pages 49–58, Rome, Italy, September 2001.
24. A. Maedche and S. Staab. Ontology learning for the semantic web. *IEEE Intelligent Systems*, 16, 2001.
25. J. McCarthy. Notes on formalizing context. *In Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, 1993.
26. D.L. McGuinness, R. Fikes, J. Rice, and S. Wilder. An environment for merging and testing large ontologies. In *Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning (KR2000)*, 2000.
27. S. Melnik, editor. *Generic Model Management: Concepts and Algorithms*. Springer-Verlag, 2004.
28. E. Mena, V. Kashyap, A. Illarramendi, and A. P. Sheth. 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, 2000.
29. C. Mooers. *Encyclopedia of Library and Information Science*, volume 7, chapter Descriptors, pages 31–45. Marcel Dekker, 1972.
30. A. Motro and I. Rakov. Estimating the quality of databases. *Lecture Notes in Computer Science*, 1998.
31. F. N. Noy and M. A. Musen. PROMPT: Algorithm and tool for automated ontology merging and alignment. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000)*, pages 450–455, Austin, TX, 2000.
32. C. Papatheodorou, A. Vassiliou, and B. Simon. Discovery of ontologies for learning resources using word-based clustering. *Proceedings of the World Conference on Educational Multimedia, Hypermedia and Telecommunications (ED-MEDIA 2002)*, pages 1523–1528, 2002.
33. E. Remolina and B. Kuipers. Towards a general theory of topological maps. *Artificial Intelligence*, 152:47–104, 2004.
34. G. Sacco. Dynamic taxonomies: A model for large information bases. *IEEE Trans. Knowl. Data Eng.*, 12(2):468–479, 2000.
35. A. Segev. Identifying the multiple contexts of a situation. In *Proceedings of IJCAI-Workshop Modeling and Retrieval of Context (MRC2005)*, 2005.
36. A. Segev and A. Gal. Putting things in context: A topological approach to mapping contexts and ontologies. In *Proceedings of AAAI-Workshop Workshop on Contexts and Ontologies: Theory, Practice and Applications*, 2005.
37. A. Segev and A. Gal. Ontology verification using contexts. In *Proceedings of ECAI-Workshop on Contexts and Ontologies: Theory, Practice and Applications*, 2006.
38. H. Shatkay and L. Kaelbling. Learning topological maps with weak local odometry information. In *Proc. IJCAI-97*, 1997.
39. M. Siegel and S. E. Madnick. A metadata approach to resolving semantic conflicts. In *Proceedings of the 17th International Conference on Very Large Data Bases*, pages 133–145, 1991.

40. S. Simhon and G. Dudek. A global topological map formed by local metric maps. *In IEEE/RSJ International Conference on Intelligent Robotic Systems*, 3:1708–1714, October 1998.
41. P. Spyns, R. Meersman, and M. Jarrar. Data modelling versus ontology engineering. *ACM SIGMOD Record*, 31(4), 2002.
42. V. Terziyan and S. Puuronen. Reasoning with multilevel contexts in semantic metanetwork. In R. Nossun P. Bonzon, M. Cavalcanti, editor, *Formal Aspects in Context*, pages 107–126. Kluwer Academic Publishers, 2000.
43. C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, second edition, 1979.
44. B.C. Vickery. *Faceted classification schemes*. Graduate School of Library Service, Rutgers, the State University, New Brunswick, N.J., 1966.