Towards Semantic Summaries over Ontologies

Sebastian Wandelt[1] and Ralf Möller[2]
[1]*WBI, Humboldt Universität zu Berlin, Berlin, Germany*
[2]*STS, Hamburg University of Technology, Hamburg, Germany*

Keywords: Ontologies, Description Logics, Semantic Summaries.

Abstract: Industry is increasingly dependent on the gathering and processing of data to support crucial product development activities. However, support systems for engineers or computer scientists may need to consider terabytes of data, making it very hard to automatically extract useful information. Querying data repositories in order to extract just the right information for the situation at hand remains a challenging problem.
We propose a notion of semantic summaries on top of description logic knowledge bases that supports querying and summarizing information in large (ontological) data repositories. The idea of a semantic summary is to characterize the result set from a broader perspective, instead of describing each domain object. We show that our summarization approach scales for benchmark ontologies up to several million assertional axioms.

1 INTRODUCTION

Industry is increasingly dependent on the gathering and processing of data to support decision making and other activities critical to their business. However, support systems for engineers, including software engineers, need to gather information from data stores that grow up to petabyte size, making efficiency in information retrieval increasingly difficult. Querying data repositories in order to extract just the right information for the situation at hand is a challenging task.

When dealing with huge data sets, it can be helpful to compute any kind of synopses and summaries over the data for two purposes. First, from a query answering system point of view, it might be more efficient to answer (transformed) queries over a summarization, because of reduced complexity of the input. Second, from a user's point of view, it can be easier to explain/comprehend particular relations (e.g. subsumptions, individual relations, etc.) in the ontology. The underlying idea for creating synopses and summarizations is closely related to notions of similarity. First, we discuss similarity in the case of synopses. Technically, synopses can be created in several ways:

- Spatial synopses: Given a particular snapshot (representation of a point of time), a similarity relation/function is computed, which assigns a similarity measure for any two entities (concepts, individuals, etc.) in an ontology. For example, in a clinical setting, two patients can be treated similar, if they share a particular amount of symptoms. In a synopsis, these patients might be merged together and only unmerged/unfolded on further request. The scenario is depicted in Figure 1.

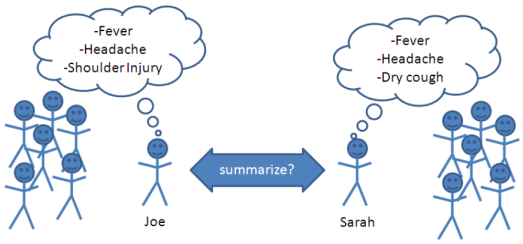

Figure 1: Semantic summary.

Joe and Sarah share symptoms Fever and Headache in our example. For some queries it might sufficient to merge Joe and Sarah into one individual, which then has e.g. only Fever and Headache, or, Fever, Headache, Shoulder injury and Dry cough. The outcome after reasoning over summarizations clearly depends on the chosen strategy. Especially in a clinical setting, for some queries, it is important to retain soundness and completeness in a synopsis, because we do not want to draw wrong conclusions about any of our patients. On the other hand, there might be queries, which do not need to distinguish details about Joe and Sarah, e.g. assume we want to find out all patients with Fever only.

Wandelt S. and Möller R..
Towards Semantic Summaries over Ontologies.
DOI: 10.5220/0004134601960201
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2012), pages 196-201
ISBN: 978-989-8565-30-3

- Temporal synopses: the idea is to use similarity of an individual over time. For example, assume the scenario shown in the Figure 2. There we show one possible progress of a disease for Joe. If we want to query for people with shoulder injuries only, we do not need to distinguish the instances of Joe over time. On the other hand, if we have a query to find patients with rising symptoms for flue, it is inevitable to consider all changes of information on Joe.

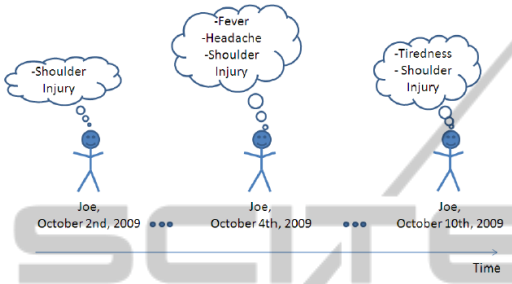

Figure 2: Semantic summary.

- Diagonal synopses: Combine Spatial and Temporal Synopses. This area is widely unexplored, but seems to have high potential to analyze large amounts of fluctuant data.

The challenges about synopses are twofold. First, the choice of the strategy (which entities to treat similar) is crucial to determine its reasoning and explanation capabilities. Second, it is important to identify which synopses should be kept up-to-date for an ontology. Summaries can be seen as an extension of synopses. The idea is to get away from told instance information in ontologies to determine a kind of abstraction of the stored assertional data. Going back to our clinical example, the idea can be visualized as shown in Figure 3.

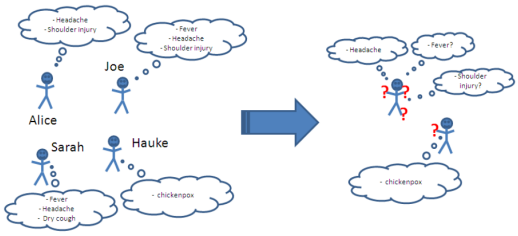

Figure 3: Semantic summary.

For example, in a summary, we only know that there is a patient with Chickenpox. Furthermore, we know that there are three people who have Headache, and possible share Fever and Shoulder Injuries. This summary can be used to talk about the general situation, e.g. allocation of beds, in the hospital. With existing technologies this is not yet possible. Although one can create statistics over ontologies by predefined aggregate queries, these queries do not adapt to new situations (e.g. new diseases).

With summaries, these statistics are just created automatically, without the user having to define any statistics-rules for his ontology. Differential summaries can then be used to determine the recent changes in ontologies, e.g. ”Does the hospital have a similar allocation of beds, as it had 5 years ago?”. We emphasize that it is not only intended to use obvious and directly told information on individuals to create summaries, but to use the locality information to detect all possibly relevant information.

This article discusses the creation of spatial synopses. For each named individual in the ABox, an abstraction is computed, given the told ABox information. The intuition is that the abstraction corresponds to a subset of the assertional knowledge, representing what we know about a given individual and what is sufficient to perform reasoning with respect to the given background knowledge. The situation is depicted in Figure 4.

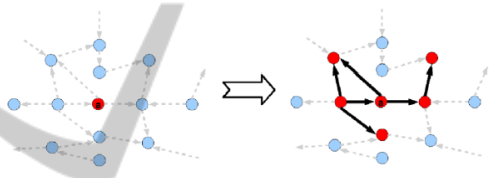

Figure 4: Semantic summary.

The abstractions for each named individual are combined, in order to obtain asynopsis/summary of the whole ontology. In the following, we describe the formal foundations for computing semantic summaries.

2 DESCRIPTION LOGICS

Description logics are a family of languages for knowledge representation. Historically, description logics are descendants of semantic nets (Quillian, 1968) and frame systems (Minsky, 1974). In Artificial Intelligence, description logics are used for formal reasoning about application domains.

In the following, we recapitulate syntax and semantics of the description logic $\mathcal{SHI}$ as far as relevant for this work. Please refer to (Baader, 1999) for further details. We assume a number of disjoint *base sets* as follows: **CN** is a non-empty set of *concept names*, **RN** is a non-empty set of *role names*, **NIN** is a non-empty set of *named individuals*, and **AIN** is a non-empty set of *anonymous individuals*. The set

of *individuals* is $\mathbf{IN} = \mathbf{NIN} \cup \mathbf{AIN}$. The set of $\mathcal{SHI}$ *-concept descriptions* is given by the following grammar:

$$C_1, C_2 ::= \top | \bot | A | \neg C_1 | C_1 \sqcap C_2 | C_1 \sqcup C_2 | \forall R.C_1 | \exists R.C_1$$

where $A \in \mathbf{CN}$ and $R \in \mathbf{Rol}$. With **AtCon** we denote all *atomic concepts*, i.e. concept descriptions which are concept names or negated concept names. For the semantics of concept descriptions please refer to (Baader et al., 2007).

A *TBox* $\mathcal{T}$ is a set of so-called *generalized concept inclusion axioms* $C_1 \sqsubseteq C_2$. A *RBox* $\mathcal{R}$ is a set of so-called *role inclusion axioms* $R_1 \sqsubseteq R_2$. An *ABox* $\mathcal{A}$ is a set of so-called *concept and role assertion axioms* $C(a)$ and $R(a_1, a_2)$. An *ontology* O consists of a 3-tuple $\langle \mathcal{T}, \mathcal{R}, \mathcal{A} \rangle$. We restrict the concept assertion axioms in $\mathcal{A}$ in such a way that each concept description is an atomic concept or a negated atomic concept. This is without loss of generality, since each non-atomic concept description can be given a name in the TBox. The set of TBoxes (RBoxes, ABoxes, ontologies) is denoted with **ST** (**SR**, **SA**, **SO**).

We denote with $clos(C)$ the closure of a concept description C. The closure of a concept description is usually used for syntactical analysis. We assume that a concept description C is usually in negation normal form, i.e. for all $\neg C_1 \in clos(C)$, C_1 is a concept name. Using De Morgan laws, every concept description can be transformed into a concept description in negation normal form. The *negation normal form of a* concept description C is denoted $nnf(C)$. Given a TBox $\mathcal{T}$, the *concept closure of* $\mathcal{T}$, denoted $clos(\mathcal{T})$, is defined as

$$clos(\mathcal{T}) = \bigcup_{C_1 \sqsubseteq C_2 \in \mathcal{T}} (clos(\neg C_1) \cup clos(C_2)).$$

3 INDIVIDUAL ABSTRACTION

In (Wandelt and Möller, 2008), a method is proposed to identify the relevant assertions to reason about an individual. The main motivation is to enable in-memory reasoning over large ontologies, i.e. ontologies with a large ABox, for traditional tableau-based reasoning systems. More formally, given an input individual a, the proposal is to compute a set of ABox assertions $\mathcal{A}_{isl}$ (a subset of the source ABox $\mathcal{A}$), such that for all atomic (!) concept descriptions C, we have $\langle \mathcal{T}, \mathcal{R}, \mathcal{A} \rangle \models C(a)$ iff $\langle \mathcal{T}, \mathcal{R}, \mathcal{A}_{isl} \rangle \models C(a)$.

In order to define subsets of an ABox relevant for reasoning over an individual a, we define an operation which splits up role assertions in such a way that we can apply graph component-based modularization techniques over the outcome of the split.

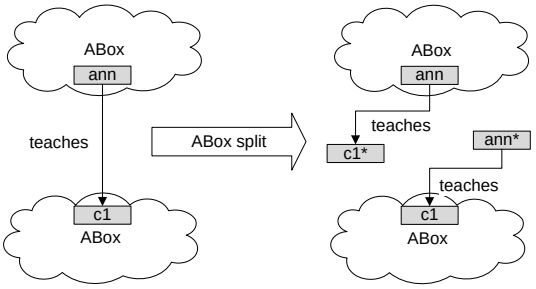

Figure 5: Intuition of an ABox split.

Definition 1 (ABox Split)**.** *Given*

- *a role description R,*
- *two distinct named individuals a and b,*
- *two distinct anonymous individuals c and d, and,*
- *an ABox* $\mathcal{A}$*,*

an ABox split *is a function* $\downarrow_{c,d}^{R(a,b)}: \mathbf{SA} \rightarrow \mathbf{SA}$*, defined as follows:*

- *If* $R(a,b) \in \mathcal{A}$ *and* $\{c,d\} \nsubseteq Ind(\mathcal{A})$*, then*

$$\downarrow_{c,d}^{R(a,b)} (\mathcal{A}) = \mathcal{A} \setminus \{R(a,b)\} \cup \{R(a,d), R(c,b)\} \cup \{C(c) \mid C(a) \in \mathcal{A}\} \cup \{C(d) \mid C(b) \in \mathcal{A}\}$$

- *Else*

$$\downarrow_{c,d}^{R(a,b)} (\mathcal{A}) = \mathcal{A}.$$

The intuition of Definition 1 is depicted in Figure 5. The clouds in Figure 5 indicate a set of ABox assertions. We split up a role assertion and keep the concept assertions for each fresh individual copy. The reason for keeping the asserted concept descriptions is explained below. If the ABox does not contain the role assertion in question, then the split returns an unchanged ABox.

Definition 2 (Extended $\forall$-info Structure)**.** *Given a TBox* $\mathcal{T}$ *in normal form and a RBox* $\mathcal{R}$*, an* extended $\forall$-info structure for $\mathcal{T}$ and $\mathcal{R}$ *is a function* $extinfo_{\mathcal{T},\mathcal{R}}^{\forall}$: $\mathbf{Rol} \rightarrow \wp(\mathbf{Con})$*, such that we have* $C \in extinfo_{\mathcal{T},\mathcal{R}}^{\forall}(R)$ *if and only if there exists a role* $R_2 \in \mathbf{Rol}$*, such that* $\mathcal{R} \models R \sqsubseteq R_2$ *and* $\forall R_2.C \in clos(\mathcal{T})$*.*

Example 1 (Example for an Extended $\forall$-info Structure)**.** *Let*

$$\mathcal{T}_{Ex1} = \{ Chair \sqsubseteq \forall headOf.Department, \exists memberOf.\top \sqsubseteq Person, GraduateStudent \sqsubseteq Student \}$$

Input: Ontology $O = \langle \mathcal{T}, \mathcal{R}, \mathcal{A} \rangle$, individual $a \in NInd(\mathcal{A})$
Output: Individual island $ISL_a = \langle \mathcal{T}, \mathcal{R}, \mathcal{A}^{isl}, a \rangle$
Algorithm:
Let agenda $= a$
Let seen $= \emptyset$
Let $\mathcal{A}^{isl} = \emptyset$
While agenda $\neq \emptyset$ **do**
Remove a_1 from **agenda**
Add a_1 to **seen**
Let $\mathcal{A}^{isl} = \mathcal{A}^{isl} \cup \{C(a_1) \mid C(a_1) \in \mathcal{A}\}$
For each $R(a_1, a_2) \in \mathcal{A}$
$\mathcal{A}^{isl} = \mathcal{A}^{isl} \cup \{R(a_1, a_2) \in \mathcal{A}\}$
If $R(a_1, a_2) \in \mathcal{A}$ is $\mathcal{SHI}$-splittable with respect to O **then**
$\mathcal{A}^{isl} = \mathcal{A}^{isl} \cup \{C(a_2) \mid C(a_2) \in \mathcal{A}\}$
else agenda $=$ **agenda** $\cup (\{a_2\} \setminus$ **seen**$)$
For each $R(a_2, a_1) \in \mathcal{A}$
$\mathcal{A}^{isl} = \mathcal{A}^{isl} \cup \{R(a_2, a_1) \in \mathcal{A}\}$
If $R(a_2, a_1) \in \mathcal{A}$ is $\mathcal{SHI}$-splittable with respect to O **then**
$\mathcal{A}^{isl} = \mathcal{A}^{isl} \cup \{C(a_2) \mid C(a_2) \in \mathcal{A}\}$
else agenda $=$ **agenda** $\cup (\{a_2\} \setminus$ **seen**$)$

Figure 6: Naive algorithm for computation of an individual island.

and

$$\mathcal{R}_{Ex1} = \{headOf \sqsubseteq memberOf\},$$

then the TBox in normal form is

$$\mathcal{T}_{Ex1norm} = \{ \top \sqsubseteq \neg Chair \sqcup \forall headOf.Department, \top \sqsubseteq \forall memberOf.\bot \sqcup Person, \top \sqsubseteq \neg GraduateStudent \sqcup Student \}$$

and the extended $\forall$-info structure for $\mathcal{T}_{Ex1norm}$ and $\mathcal{R}_{Ex1}$ is:

$$extinfo^{\forall}_{\mathcal{T},\mathcal{R}}(R) = \begin{cases} \{Department, \bot\} & \text{if } R = headOf, \\ \{\bot\} & \text{if } R = memberOf, \\ \emptyset & \text{otherwise.} \end{cases}$$

The extended $\forall$-info structure allows us to check which concept descriptions are (worst-case) propagated over role assertions in $\mathcal{SHI}$-ontologies.

Definition 3 ($\mathcal{SHI}$-splittability of Role Assertions)**.** *Given a $\mathcal{SHI}$-ontology $O = \langle \mathcal{T}, \mathcal{R}, \mathcal{A} \rangle$ and a role assertion $R(a_1, a_2)$, we say that $R(a_1, a_2)$ is $\mathcal{SHI}$-*splittable with respect to *O if*

1. *there exists no transitive role R_2 with respect to $\mathcal{R}$, such that $\mathcal{R} \models R \sqsubseteq R_2$,*
2. *for each $C \in extinfo^{\forall}_{\mathcal{T},\mathcal{R}}(R)$*
   - *$C = \bot$ or*
   - *there exists a concept description C_2, such that $C_2(b) \in \mathcal{A}$ and $\mathcal{T} \models C_2 \sqsubseteq C$ or*
   - *there exists a concept description C_2, such that $C_2(b) \in \mathcal{A}$ and $\mathcal{T} \models C \sqcap C_2 \sqsubseteq \bot$*

   and
3. *for each $C \in extinfo^{\forall}_{\mathcal{T},\mathcal{R}}(R^-)$*
   - *$C = \bot$ or*
   - *there exists a concept description C_2, such that $C_2(a) \in \mathcal{A}$ and $\mathcal{T} \models C_2 \sqsubseteq C$ or*
   - *there exists a concept description C_2, such that $C_2(a) \in \mathcal{A}$ and $\mathcal{T} \models C \sqcap C_2 \sqsubseteq \bot$.*

To sum up, for each named individual in the ontology, we use the algorithm from Figure 6, to obtain an abstraction of the individual.

4 SEMANTIC SUMMARIES

Given an individual abstraction for each named individual in an input ontology, it is clear that some (or even many) abstraction are similar to each other. Due to lack of space we do not go into the technical details of computing the similarity of individual abstraction here. However, if one looks at an abstraction as a graph, graph homomorphisms can be used directly to determine similar individual islands.

The key insight is that similar abstraction entail the same set of concept descriptions for there root individual. Therefore these individuals (of similar abstractions) cannot be distinguished with respect to the given background knowledge. This is exactly what we expect from semantic summaries. Thus, for semantic summaries, we propose to look at ontologies as a set of similar individual abstractions.

We performed some first evaluation of this idea with respect to a benchmark ontology. The Lehigh University Benchmark, short LUBM, is a synthetic ontology developed to benchmark knowledge base systems with respect to large OWL applications. The ontology is situated in the university domain. The background knowledge, i.e. the terminology, is described in a schema called Univ-Bench, see (Guo et al., 2005) for an overview over the history, different versions and the predecessor Univ 1.0. The expressivity of the ontology is chosen to be in OWL Lite, which corresponds to the description logic $\mathcal{SHIF}$. However, the de facto expressivity is lower. For instance,

the ontology does not introduce any cardinality/functionality expressions on roles.

In Figure 7, we show the number of individuals in the dataset, for different numbers of universities. It can be seen that the number of individuals increases almost linearly with the number of universities. Around 30 percent of the individuals in the dataset are publications, another 30 percent are undergraduate students, 10 percent are graduate students, 10 percent are courses and graduate courses. The remaining 20 percent of the individuals are for instance professors, assistants and departments. For more details about the data distribution, see (Guo et al., 2005).

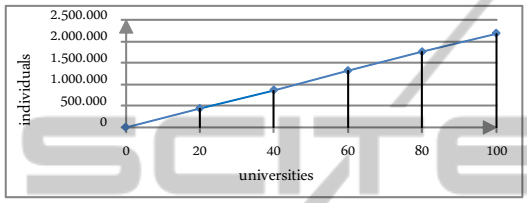

Figure 7: Number of individuals in LUBM.

Next, we evaluated the number of distinct individual abstractions for different number of universities. The result is shown in Figure 8. It can be seen that the number of distinct individual abstactions is constant - compared to the linear number of individuals.

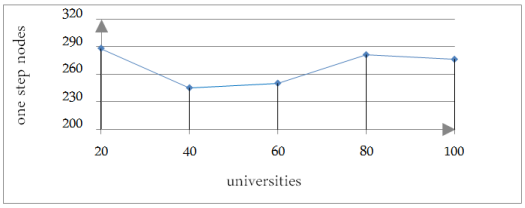

Figure 8: Number of distinct individual abstractions for LUBM.

As a second ontology, we had a look at an ontology from the CASAM project. The CASAM project is focused on computer-aided semantic annotation of multimedia content. The novelty is the aggregation of human and machine knowledge. For a detailed discussion of the research objectives, see (Gries et al., 2010), (Papantoniou et al., 2010), and (Creed et al., 2010). Within the CASAM project, there is a need to define an expressive annotation language which allows for typical-case reasoning systems. The proposed annotation language is defined by the so-called Multimedia Content Ontology, short MCO, introduced in (Gries et al., 2009). Inspired by the MPEG-7 standard, see (ISO/IEC15938-5FCD, 2002), strictly necessary elements describing the structure of multimedia documents are extracted. The intention is to exploit quantitative and qualitative time information in order to relate co-occurring observations about events in videos. Co-occurrences are detected either within the same or between different modalities, i.e. text, audio and speech, regarding the video shots.

For our evaluation with respect to MCO, we have a number of multimedia documents from the CASAM project. The set of test ontologies contains documents with identifiers ranging from 1 to 14. Each document is decomposed into several so-called *delta* files. Each delta represents additional information about the document of concern. We evaluated our summarization techniques with respect to all documents. Here we only show the results for Document 1, since for all the other documents we obtained very similar statistics.

In Figure 9, we show the number of individuals in the dataset, with an increasing delta. It can be seen that most individuals are introduced in the first delta files. The remaining delta files only introduce additional ABox assertions about already known individuals. Please note that the number of individuals is not linear in the number of delta.

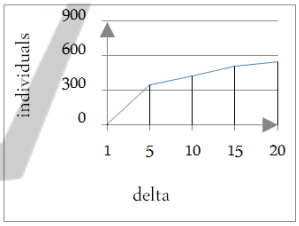

Figure 9: Number of individuals in Document 1.

We have evaluated the number of individual abstractions for different delta. The result is shown in Figure 10. It can be seen that the number of distinct individual abstractions is relatively constant - after most individuals are introduced in the third and fourth delta.

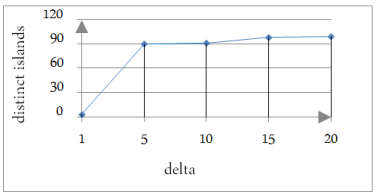

Figure 10: Number of distinct individual abstractions for Document 1.

5 CONCLUSIONS AND FUTURE WORK

We have proposed first ideas for a notion of semantic summaries that supports industrial information

search scenarios by using (domain specific) industry-standard vocabularies to query and summarize information. It has been shown already that summaries can be efficiently managed in a distributed computing setting(Wandelt and Möller, 2010) and can be used for reasoning over the ontology of concern (Wandelt et al., 2010).

For Future Work, we have to evaluate our semantic summary techniques with respect to additional ontologies. Furthermore, we would like to formally implement and evaluate difference operators over ontology summaries, in order to formally capture ontology evolution with temporal synopses.

REFERENCES

Baader, F. (1999). Logic-Based Knowledge Representation. In *Artificial Intelligence Today*, pages 13–41. Springer-Verlag.

Baader, F., Calvanese, D., McGuinness, D. L., Nardi, D., and Patel-Schneider, P. F. (2007). *The Description Logic Handbook*. Cambridge University Press, New York, NY, USA.

Creed, C., Lonsdale, P., Hendley, R., and Beale, R. (2010). Synergistic annotation of multimedia content. In *Proceedings of the 2010 Third International Conference on Advances in Computer-Human Interactions*, ACHI '10, pages 205–208, Washington, DC, USA. IEEE Computer Society.

Gries, O., Möller, R., Nafissi, A., Rosenfeld, M., Sokolski, K., and Wessel, M. (2010). A Probabilistic Abduction Engine for Media Interpretation Based on Ontologies. In Hitzler, P. and Lukasiewicz, T., editors, *RR*, volume 6333 of *Lecture Notes in Computer Science*, pages 182–194. Springer.

Gries, O., Möller, R., Nafissi, A., Sokolski, K., and Rosenfeld, M. (2009). CASAM Domain Ontology. Technical report, Hamburg University of Technology.

Guo, Y., Pan, Z., and Heflin, J. (2005). LUBM: A benchmark for OWL knowledge base systems. *J. Web Sem.*, 3(2-3):158–182.

ISO/IEC15938-5FCD (2002). Multimedia Content Description Interface (MPEG-7). http://mpeg.chiariglione.org/standards/mpeg-7/mpeg-7.htm.

Minsky, M. (1974). A Framework for Representing Knowledge. Technical report, MIT-AI Laboratory, Cambridge, MA, USA.

Papantoniou, K., Tsatsaronis, G., and Paliouras, G. (2010). KDTA: Automated Knowledge-Driven Text Annotation. In Balcázar, J. L., Bonchi, F., Gionis, A., and Sebag, M., editors, *ECML/PKDD (3)*, volume 6323 of *Lecture Notes in Computer Science*, pages 611–614. Springer.

Quillian, R. (1968). Semantic memory. In *Semantic Information Processing*, pages 216–270. MIT Press.

Wandelt, S. and Möller, R. (2008). Island reasoning for ALCHI ontologies. In *Proceedings of the 2008 conference on Formal Ontology in Information Systems*, pages 164–177, Amsterdam, The Netherlands. IOS Press.

Wandelt, S. and Möller, R. (2010). Distributed island-based query answering for expressive ontologies. In Bellavista, P., Chang, R.-S., Chao, H.-C., Lin, S.-F., and Sloot, P. M. A., editors, *GPC*, volume 6104 of *Lecture Notes in Computer Science*, pages 461–470. Springer.

Wandelt, S., Möller, R., and Wessel, M. (2010). Towards scalable instance retrieval over ontologies. *Int. J. Software and Informatics*, 4(3):201–218.