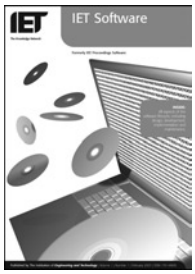
Published in IET Software
Received on 1st June 2007
Revised on 3rd December 2007
doi: 10.1049/iet-sen:20070062

In Special Issue on Language Engineering




ISSN 1751-8806

# Ontological approach for the semantic recovery of traceability links between software artefacts

*Y. Zhang R. Witte J. Rilling V. Haarslev*

*Department of Computer Science and Software Engineering, Concordia University, Montreal, Canada*
*E-mail: rilling@cse.concordia.ca*

**Abstract:** Traceability links provide support for software engineers in understanding relations and dependencies among software artefacts created during the software development process. The authors focus on re-establishing traceability links between existing source code and documentation to support software maintenance. They present a novel approach that addresses this issue by creating formal ontological representations for both documentation and source code artefacts. Their approach recovers traceability links at the semantic level, utilising structural and semantic information found in various software artefacts. These linked ontologies are supported by ontology reasoners to allow the inference of implicit relations among these software artefacts.

## 1 Introduction

Software maintenance, often also referred to as software evolution, constitutes a major part of the total cost occurring during the life span of a software system [1, 2]. As software ages, the task of maintaining it becomes more complex and more expensive. Maintainers often face the challenge to identify and comprehend the different representations and interrelationships that exist among software artefacts and knowledge resources involved in software maintenance [3, 4]. Therefore from a maintainer's perspective, exploring [5] and linking these artefacts and knowledge resources becomes a key challenge [6]. However, it is a well-known fact that even in organisations and projects with mature software development processes, software artefacts created as part of well-defined processes end up to be disconnected from each other [6, 7]. This lack of traceability among software artefacts is caused by several factors, including: (1) the fact that these artefacts are written in different languages, including natural language, formal specifications, modelling and programming languages; (2) they describe a software system at various abstraction levels (design against implementation); (3) processes applied within an organisation do not enforce maintenance of existing traceability links and (4) a lack of adequate tool support to establish and maintain traceability among various artefacts.

As a result, maintainers tend to spend a large amount of effort on synthesising and integrating information from various sources in these systems to establish links among artefacts. Existing research in software traceability focuses on reducing the cost associated with this manual effort by developing automatic assistance in establishing and maintaining traceability links among software artefacts [7].

Software documentation, like requirements and design documents, contains a large amount of information in the form of description and text written in natural language. These documents combined with source code represent two of the main software artefact types utilised in software comprehension [6, 7]. Existing source code/document traceability research [6, 8] mainly focuses on connecting documents and source code using information retrieval (IR) techniques. However, these IR approaches ignore structural and semantic information that can be found in both documents and source code, therefore limiting both their precision and applicability.

In this research, we introduce a novel formal ontological representation that integrates two of the major software artefacts, source code and software documentation. The goal of using this common representation is to reduce the conceptual gap caused by the type of abstractions and





© The Institution of Engineering and Technology 2008

languages found in these artefacts. Having this common representation allows maintainers not only to explore knowledge relevant to their given task across the different modelled artefacts, but also to enrich their current understanding of a system. The use of ontologies for managing traceability between software and other artefacts therefore becomes an important aspect of software language engineering. It provides non-trivial relationships between artefacts and helps with consistency management in a language-based manner.

Unlike existing approaches using IR, we developed a text mining system to semantically analyse software documents. The discovered concept instances from both source code and documents are used to establish traceability links between these software artefacts. In addition, the formal ontological representation also allows us to take advantage of automated reasoning services provided by ontology reasoners to infer implicit relations (links) between these two types of artefacts.

An overview of our approach is shown in Fig. 1. In the first step, ontologies for source code and document artefacts are created and automatically populated from existing source code and documentation artefacts. In the second step, traceability links among the modelled artefacts are established to allow for querying and reasoning upon this knowledge base to infer implicit relations among the modelled software artefacts [9].

Our research is significant for several reasons. First, software artefacts other than source code, such as documentation, contain rich semantic information that is not used by existing reverse engineering tools. Introducing an ontological representation for software documentation enables us to utilise natural language processing (NLP) techniques [10] to 'understand' parts of the semantics conveyed by these artefacts and to establish additional traceability links among these artefacts.

Secondly, the automatic ontology population is an important issue in ontology engineering. In particular for artefacts containing natural language constructs, ontology population with semantically rich information is an ongoing challenge. Our approach demonstrates a feasible solution within the software domain that goes beyond current techniques.

Furthermore, the uniform ontological representation for both source code and documentation allows us to share common concepts between different resources, easing the integration of information by allowing for the automatic recovery and establishment of traceability links among documentation and source code artefacts.

Finally, representing software artefacts in a formal ontology allows programmers to reason about various implicit relations between software artefacts. Taking advantage of these established traceability links and existing ontology-based knowledge representation techniques such as description logics (DL) [11] and ontology reasoners [12], users can define new concepts and roles (types of relations). This newly modelled knowledge can then be used to support specific maintenance activities (e.g. program comprehension) and consistency checking of artefacts through ontology queries.

The remainder of the paper is organised as follows: in Section 2, we provide relevant background of our research, including formal ontologies and text mining techniques used for ontology population. Section 3 presents our approach to traceability recovery, including a high-level description of our methodology and the requirements that have to be satisfied by the ontology design presented in this section.

Section 4 describes the general system architecture of the Semantic Web-Enabled Software Maintenance Environment we implemented to support our traceability recovery approach. Section 5 covers in detail the automatic population of the software ontology from both source code and documents. Section 6 provides an evaluation and application example of our approach. Related work is discussed in Section 7 and conclusions and future work are presented in Section 8.

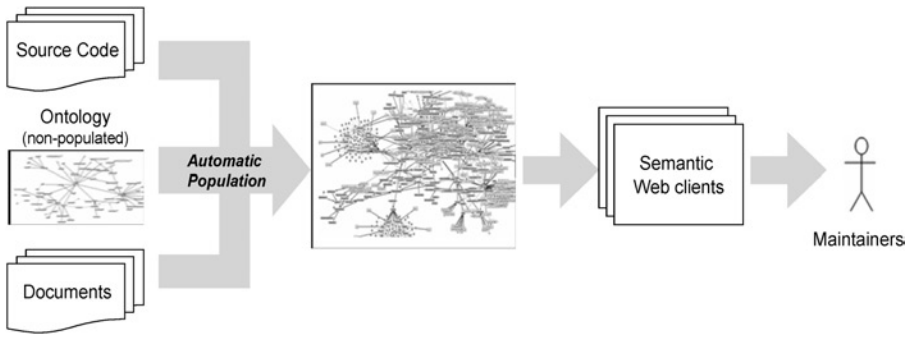



**Figure 1** *Ontology-based software maintenance overview*


© The Institution of Engineering and Technology 2008

doi: 10.1049/iet-sen:20070062

## 2 Background

In this section, we introduce background relevant to our research, including ontologies and their formalisms, DL, as well as text mining techniques used for ontology population. Readers familiar with these concepts can safely skip this section.

### 2.1 Ontology and description logics

The term 'ontology' originates from philosophy, where it denotes the study of existence. In computer science, the most common definition has been provided by Gruber [13]: 'An ontology is an explicit specification of a conceptualisation'. Here, ontologies are typically used as a formal and explicit way of specifying the concepts and relationships in a domain of discourse. DLs [11], a family of knowledge representation formalisms, have long been regarded as a standard ontology language. DL is also a major foundation of the recently introduced Web Ontology Language (OWL) recommended by the W3C [14]. DL represents the knowledge of a domain by first defining the relevant concepts of the domain in a taxonomy and then utilising these concepts to specify properties of individuals occurring in the domain. Furthermore, ontologies allow for the extension of a resulting model to reflect more closely an 'open' world assumption [11], by allowing for the integration of incomplete knowledge and extendibility of the ontological model. The use of DL allows for a formal characterisation of subsumption relationships between concepts: A concept `C` is considered a sub-concept of `D` if all instances of `C` are also instances of `D`.

Basic elements of DL are atomic concepts and atomic roles, which correspond to unary predicates and binary predicates in first-order logic. Complex concepts are then defined by combining basic elements with several concept constructors. For example, in the domain of software design technique and documentation structure, given atomic concepts such as `DesignPattern` and `Paragraph`, as well as an atomic role `contains` that describes a relation between these two concepts, a new concept `DesignPatternDoc` can be defined by a conjunction constructor and existential qualifier

$$\texttt{DesignPatternDoc} \equiv \texttt{Paragraph} \sqcap \exists \texttt{contains.DesignPattern}$$

Individuals existing in the domain and their relations can be specified as instances of their corresponding concepts and roles. For example, the following DL expressions define $p$ as a paragraph instance, `abstract_factory_pattern` as a design pattern instance and the relation contains between the paragraph body $p$ and the `abstract_factory_pattern`

$$\texttt{p: Paragraph}, \texttt{abstract\_factory\_pattern: DesignPattern}, (\texttt{p}, \texttt{abstract\_factory\_pattern}): \texttt{contains}$$

Having DL as the specification language for a formal ontology enables also the use of reasoning services provided by DL-based knowledge representation systems. Unlike many logic programming approaches that cannot guarantee completeness, DL reasoning services are proven to be sound, complete and terminating. Moreover, DL reasoning is automatic and does not require the development of logic programs to extract the desired inferences. DL reasoning is usually performed on demand, triggered by relevant queries to the knowledge base. The Racer system [12] is an ontology reasoner that has been highly optimised to support very expressive DLs [11, 12, 15], that is, DLs that support at least the DL $\mathcal{ALC}$ [11, Chapter 2] extended by transitive roles, inverse roles, role hierarchies and number restrictions on roles (this logic is denoted as $\mathcal{SHIN}$). An example for an even more expressive DL is $\mathcal{SROIQ}$ [16] that extends $\mathcal{SHIN}$ by qualified number restrictions, nominals and a restricted form of role composition. Typical services provided by Racer include terminology inferences (e.g. concept consistency, subsumption, classification and ontology consistency) and instance reasoning (e.g. instance checking, instance retrieval, tuple retrieval and instance realisation). For example, given the above concept definition of `DesignPatternDoc`, as well as the assertions about instance $p$ and `abstract_factory_pattern`, the ontology reasoner can automatically infer that $p$ is also an instance of `DesignPatternDoc`.

For a more complete coverage of DLs and Racer, we refer the reader to [11, 12].

### 2.2 Text mining and ontology population

Text Mining is commonly known as a knowledge discovery process that aims to extract non-trivial information or knowledge from unstructured text [17]. Unlike IR systems [18], text mining does not simply return documents pertaining to a query, but rather attempts to obtain semantic information from the documents themselves, using techniques from NLP [10] and artificial intelligence. In the software domain, for example, a text mining system can be employed to obtain information about individual software entities mentioned in documents, like the system architecture, its components and their relationships with packages or classes. These mentions, called named entities, can then be exported in a structured format for further (automated) analyses or browsing by a user. Commonly used export formats are XML or relational database tuples. When text mining results are exported in the form of instances (individuals) for an existing ontology, this process is also called ontology population [19], which is again different from ontology learning, where the concepts themselves (and their relations) are (semi-)automatically acquired from natural language texts.

Text Mining systems are often implemented using component-based frameworks, such as GATE (General Architecture for Text Engineering) [20] or IBM's UIMA



© The Institution of Engineering and Technology 2008

(Unstructured Information Management Architecture). Within the text mining process, a number of standard NLP techniques are commonly performed. These include first dividing the textual input stream into individual tokens with a (Unicode) Tokeniser, using a Sentence Splitter to detect sentence boundaries and running a statistical Part-of-Speech (POS) tagger that assigns labels (e.g. noun, verb and adjective) to each word. Larger grammatical structures, such as Noun Phrases (NPs) and Verb Groups (VGs), can then be created based on these tags using chunker modules. Based on these foundational analysis steps, more semantically-oriented analyses can be performed, which typically require domain- and language-specific algorithms and resources.

# 3 Ontology design for traceability

We now present our approach to traceability recovery, starting with a high-level description of our methodology. Based on the described approach, we derive a number of requirements (Section 3.2) that have to be satisfied by the ontologies. Our ontology design fulfilling these requirements is then presented in Section 3.3.

## 3.1 Methodology

As stated above, the goal of our work is the automatic establishment of traceability links between source code and natural language documents. In particular, we are concerned with 'deep links' between individual entities on the source code side (e.g. a class or method) and the corresponding mentions of these entities in a document on the level of individual words or phrases. That is, in contrast to IR approaches, we are not interested in creating links on the level of complete documents or paragraphs, but rather much more fine-grained links that precisely show the connections between individual words and code entities. Obviously, these links can only be created between entities that appear on both the source code side and documentation side: for example, class names, methods or variables. But documents may also contain much more higher-level concepts that can not be directly mapped to a single code entity, like descriptions of algorithms, architectures or requirements. It is here that our ontology-based method provides a strong advantage compared with existing, semantic-poor approaches, as it allows us to explicitly encode knowledge of the software domain in a formal language that can be automatically evaluated, for example, through queries and reasoning. For example, we can model knowledge about design patterns (e.g. names, types) and their relations with other entities (e.g. classes, methods) in an ontology. An end user, like a software maintainer, can then navigate a single, unified, formal representation that covers both code and documentation, which allows for a number of novel use cases, like the automatic establishment of traceability links, the generation of concept-based document views and the automatic summarisation (slicing) of documents based on an explicit query by a user.

*3.1.1 Traceability link recovery through ontology alignment:* Before any more advanced tasks can be performed on our ontological code/document representation, we first have to establish the deep traceability links. This involves a number of steps:

1. Building ontologies modelling the domains of source code and software documents.

2. Creating a knowledge base by automatically populating these ontologies through code analysis and text mining.

3. Establishing traceability links between code and documents through ontology alignment.

Step (1) requires an ontology design that is rich enough to capture the concepts and their relations to support all further analysis steps. These requirements are listed explicitly in Section 3.2 and our resulting design is covered in Section 3.3.

Concepts alone cannot support analysis tasks concerned with a concrete system and its documentation: This requires the creation of a knowledge base containing instances (individuals) for the modelled concepts. The automatic population stipulated by Step (2) is described in Section 4.

After populating the ontologies, in Step (3) traceability links are established through ontology alignment [21, 22]. Our approach is illustrated in Fig. 2. The upper part shows concepts of the 'document and source code domain. Instances of these concepts are detected through the analysis of concrete systems, shown in the bottom half. The traceability links between code and documentation are created through ontology alignment, based on class and instance information. Since our documentation ontology and source code ontology share many concepts from the programming language domain, such as `Class` or `Method`, we focus in this research on matching instances, based on their names and properties.

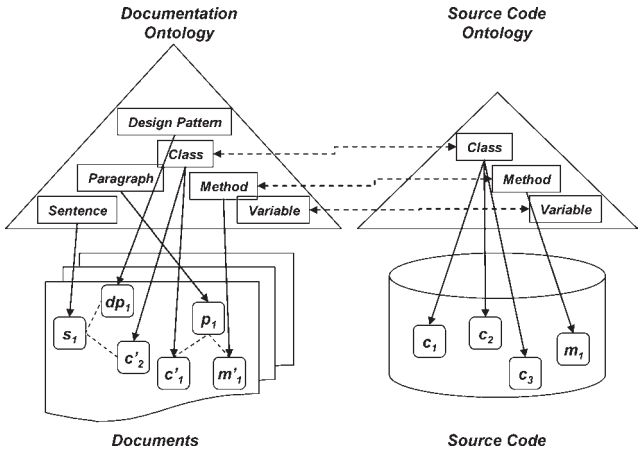



**Figure 2** *Linking instances from source code and documentation*


© The Institution of Engineering and Technology 2008

doi: 10.1049/iet-sen:20070062

The result is a single formal, ontology language-based representation of both source code and natural language that can now be utilised by numerous higher-level tasks.

*3.1.2 Querying code and documentation in a single representation:* The linked source code and documentation ontologies enable us to combine knowledge from both software implementation and documentation. Using the established links between the source code and documentation ontology, a user can perform ontological queries on either documents or source code using any of the contained concepts, entities or relations. Note that this is a much more powerful paradigm than IR-based analysis [6, 8], which only returns a (ranked) list of documents without further structure.

Questions that might be asked by a software maintainer and encoded into an ontology query, are:

- What are the document passages (sentences/paragraphs) describing the source code entity (class, method etc.) I am currently visiting in my editor?

- What design patterns are mentioned in the code's documentation within the same paragraph as the class I am currently working on?

- What architectural descriptions are contained in the documentation? What are its parts, and how do they relate to the source code base?

Note that these queries cross the conceptual boundary between code and natural language documents, which is not possible with other approaches that lack the ability to model both kinds of artefacts in the same formal representation. In fact, queries can also be asked from the viewpoint of a documentation writer, asking for all code segments related to a text currently being written or edited.

*Query Example:* As an example, assume that the populated document ontology contains the class AST as part of a visitor pattern. In order to retrieve all documented information related to the detected pattern, the query shown in Fig. 3 can be used to retrieve all text paragraphs that describe the sub-classes of AST.

```
var query = new Query();                        // define a new query
query.declare("P", "C");                        // declare two query variables
query.restrict("P", "Paragraph");               // P is a paragraph
query.restrict("C", "Class");                   // C is a class
query.restrict("C", "hasSuper",
 "net.refractions.udig.catalog.util.AST");      // C is a sub-class of AST
query.restrict("P", "contains", "C");           // P contains C
query.retrieve("P");                            // this query only retrieve P
var result = ontology.query(query);             // perform the query
```

**Figure 3** *Querying the ontology for paragraphs in the documentation describing AST sub-classes in the source code*

This query utilises both the programming language semantics, such as the inheritance relation between query variable *C* and the class AST (direct and indirect), and the structural information of documentation, such as the `contains` relation between *P* and *C*. The result of this query therefore contains all text paragraphs that describe the sub-classes of AST, that is, the visitor pattern [23]. It has to be noted that the role `contains` is a transitive relation to describe the document structure. The ontology reasoner can automatically resolve the transitivity from `Paragraph` to `Sentence` and from `Sentence` to `Class`.

*3.1.3 Documentation slicing and focused summarisation:* The potentially large amount of documentation associated with a legacy system is the motivation for delivering more advanced, semantically-oriented navigation features to software development environments. Often, information pertaining to a certain entity, like a class, is distributed over many different documents, which makes it difficult for a developer to quickly gain a comprehensive overview on all the available knowledge. With our ontology model, we can deliver a virtual document containing all the information pertaining to a specific entity (or a set of entities) of interest to a software engineer. This is also known as (focused) automatic summarisation in the area of NLP and can be compared with slicing of documents, instead of source code.

Documents can not only be re-structured based on their links to source code, but also using the design-level concepts that relate to particular maintenance tasks. For example, using the classification service of the ontology reasoner, one can classify document pieces that relate to a specific concept or a set of concepts (Fig. 4).

*Document Slicing Example:* For example, a visitor pattern [23] within a document can be considered as all the text paragraphs that describe or contain information related to the concept `Visitor`. The following new concept `VisitorPatternDoc` can be used to retrieve paragraphs that relate to such a visitor pattern. Similarly, a new concept `HighlevelDoc` can be also defined to retrieve all documents that contain the high-level design concept `Architecture` or `DesignPattern`. The ontology reasoner can then automatically classify documents according to these concept definitions.

$$\texttt{VisitorPatternDoc} \equiv \texttt{Paragraph} \sqcap \exists \texttt{contains.Visitor}$$

$$\texttt{HighlevelDoc} \equiv \texttt{DocumentFile} \sqcap \exists \texttt{contains.}(\texttt{Architecture} \sqcup \texttt{DesignPattern})$$



© The Institution of Engineering and Technology 2008

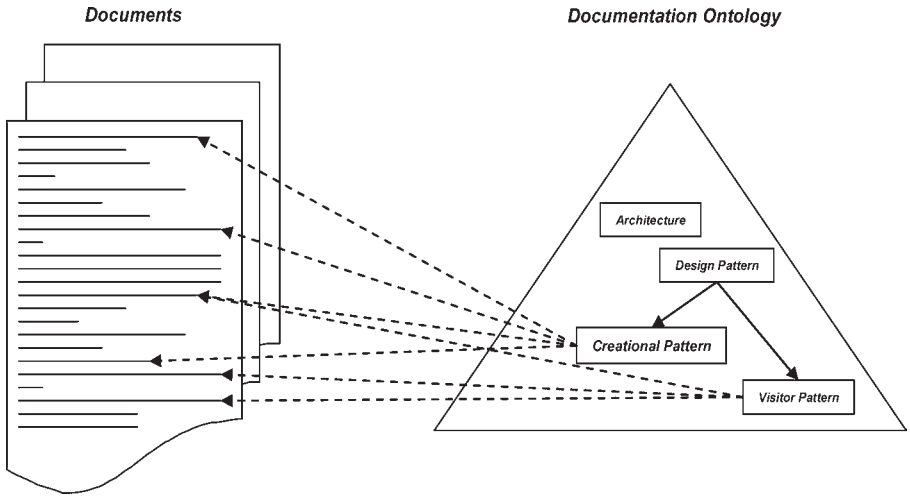



**Figure 4** *Classification of documentation through ontological reasoning*

## 3.2 Detected requirements

Based on the stated goal and approach, we can now define the requirements for traceability recovery systems using the ontology language OWL-DL more precisely.

*3.2.1 Requirement #1: ontology design for traceability:* Our approach relies on the existence of a single formal language model for both source code and documentation. Both types of artefacts must be modelled in enough detail to allow the establishment of the 'deep links' stipulated above. In particular, the documentation and source code sub-hierarchies must be modelled such that they share a number of concepts: otherwise, cross-artefact queries and reasoning as exemplified above would not be possible. This requirement is addressed in Section 3.3.

*3.2.2 Requirement #2: ontology implementation:* To illustrate the feasibility of this approach, it is necessary to implement the ontologies and create a program comprehension environment that manages the various software entities and the ontologies. As stated in the background section, this implementation should be based on established standards, such as the W3C standard OWL. We discuss implementation issues in Section 4. Furthermore, the implementation must be based on an ontology language that supports queries and formal reasoning, which are required for the more advanced traceability use cases, such as document summarisation.

*3.2.3 Requirement #3: automated ontology population:* Creating a knowledge base, that is, populating the ontology with instances, is a necessary prerequisite for an automated environment. Thus, automatic population methods must be developed that can take existing source code and its accompanying documentation and create instances in the modelled ontology. Our ontology population subsystems are described in detail in Section 5.

*3.2.4 Requirement #4: automatic traceability recovery:* After populating the software ontology, traceability links must be identified and created, based on the ontology alignment strategy outlined above.

In addition to these fundamental requirements, this work also requires a formal evaluation, based on clearly defined tasks, metrics and data, as well as a demonstration of its usefulness within a real-world scenario. We address these points in our evaluation and application section (Section 6).

## 3.3 Ontological representation of software artefacts

We start addressing the stated Requirement #1 by designing our software ontology. The goal of our design is to cover the rich structural and semantical knowledge contained in software artefacts, in particular for source code and documentation. This underlying ontological model is one of the major differences between our semantic approach and common 'bag-of-words' approaches used in IR [6, 8].

To satisfy our second requirement, we employ the formal ontology standard OWL-DL. These web ontologies are an emerging technology for capturing the semantics in a domain of discourse. Based on available semantic web technologies, they also allow us to deliver further semantic support to end users [24] that goes beyond traceability recovery.

Our software ontology contains two major sub-ontologies for representing source code and documents within the software domain, described in detail in the following two subsections.

*3.3.1 Source code ontology:* The source code sub-ontology contains more than 100 common concepts from the programming language domain, as well as some additional concepts that are Java specific. Fig. 5 shows a


© The Institution of Engineering and Technology 2008

doi: 10.1049/iet-sen:20070062

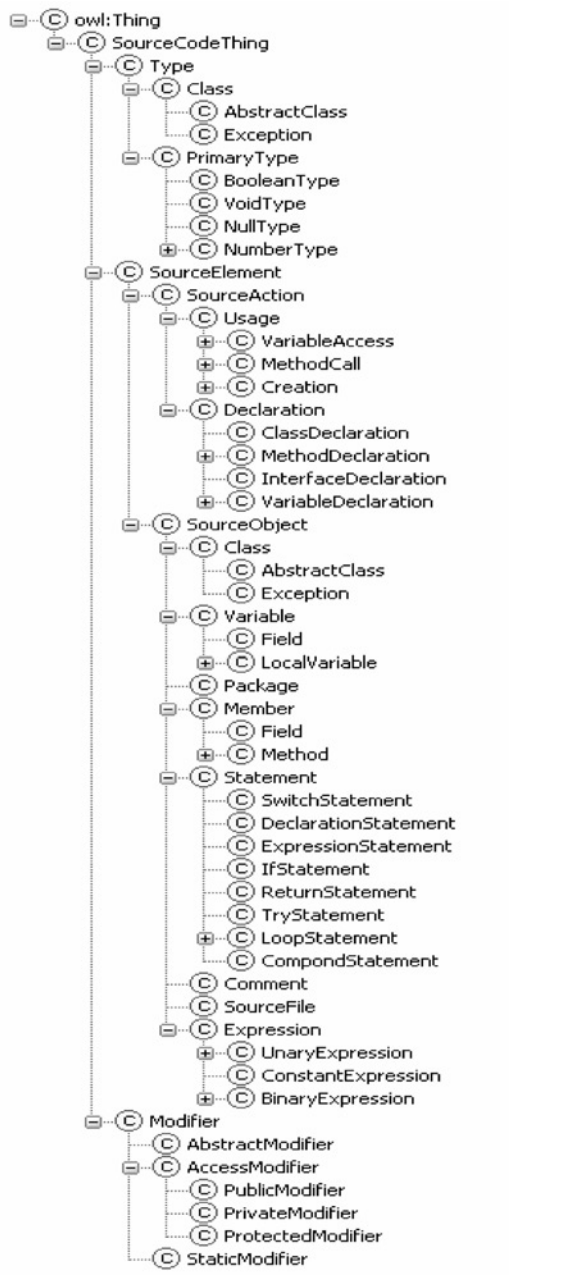



**Figure 5** *Concept hierarchy in the source code sub-ontology*

partial view of the Java taxonomy modelled in the source code ontology.

As suggested in [25], concepts in a domain can be classified into two categories, objects and actions. Objects typically refer to entities that may appear in the domain, and actions refer to their associations.

The `SourceObject` concept is introduced to denote all entities that may appear in the source code. Sub-concepts of SourceObject include `SourceFile`, `Package`, `Class`, `Method`, `Variable`, `Statement` and `Expression` etc. Each of these concepts typically corresponds directly to some programming language constructs in the source code, and therefore instances of these concepts can be automatically identified by source code parsing. This is discussed in detail in Section 5.1.

A `SourceAction` typically denotes a semantic association between two or more `SourceObjects`. For example, a `MethodCall` action refers to a method invocation expression within the body of a method that invokes another method. The properties of `MethodCall` therefore include the caller method, the called method, the types of parameters sent by this method call and the return type of the called method.

The object part of the ontology (concept `SourceObject` and its sub-concepts) is designed to contain complete syntactical information of the source code. The action part (`SourceAction` and its sub-concepts) consists of selected semantic information that is of particular interest to software maintainers.

Several utility concepts are defined to specify properties of other source code entities. For example, the concept `Type` represents an abstract type of source code entities and a `Modifier` concept is used to specify the accessibilities of entities, such as public, private etc.

The use of DL allows the formal characterisation of subsumption relationships between concepts. A concept `C` is considered as a sub-concept of `D` if all instances of `C` are also instances of `D`. As a result, if an individual is specified as a `Method` in our ontology, it will be automatically recognised as a `Member`, and further, as a `SourceObject`.

In addition, we provide necessary conditions for all the concepts in the ontology using different concept constructors. For example

$$\texttt{Method} \sqsubseteq \texttt{SourceObject} \sqcap \exists \texttt{definedIn.Class}$$

describes a situation in which a method is a type of source object, and it has to be defined in a class, and

$$\texttt{Variable} \sqsubseteq \forall \texttt{hasType.Type}$$

means that a variable must have a type. Some concepts are defined by providing necessary and sufficient conditions. For example

$$\texttt{Member} \equiv \texttt{Method} \sqcup \texttt{Field}$$

can be interpreted as a class member that is either a method or a field, and

$$\texttt{PublicField} \equiv \texttt{Field} \sqcap \exists \texttt{hasModifier.PublicModifier}$$

defines that a public file is a field with public modifiers.

*Relationships:* Within the source code ontology, many roles are defined to specify relationships among concepts. A



© The Institution of Engineering and Technology 2008

**Table 1** Role names in source code ontology

| Role name | Description |
|---|---|
| definedIn | SourceObject A is defined in SourceObject B |
| hasSuper | Class A has super Class B |
| hasSub | Class A has sub Class B |
| Call | Method A calls Method B |
| indirectCall | transitive relation of method call |
| access | Method A read/write Variable B |
| read | Method A read Variable B |
| write | Method A write Variable B |
| hasType | Variable A has Type B |
| typeOf | Variable A is of Type B |
| create | Class A creates instance of Class B |

partial view of the role names and their descriptions found in our source code ontology is shown in Table 1.

Additionally, we also provide for each of these roles their corresponding inverse role. For example, the inverse role of `hasSuper` is `hasSub`, that is, if class $C_1$ `hasSuper` class $C_2$, then $C_2$ also `hasSub` class $C_1$. Similarly, if a method $M$ `write` a variable $V$, then $V$ is `writtenBy` $M$.

One of the advantages of using DLs is its ability to define transitive roles. If a role `R` is defined as a transitive role, and if `(a,b)` $\in$ `R` and `(b,c)` $\in$ `R` are specified, then `(a,c)` $\in$ `R` also holds. Transitive roles are especially useful for specifying part-of relationships between source code elements (through `definedIn` role), inheritance relationships between classes (through `hasSuper` role) and indirect calling relationships (through `indirectCall` role).

The definition of subsumption relationships between roles is also supported. More formally, if role `R` is a sub-role of `S` and `(a, b)` $\in$ `R` holds, then `(a, b)` $\in$ `S` will also hold. For example, in our ontology, the `read` role is a sub-role of `access`. Let $M$ and $V$ be instances of `Method` and `Variable`, respectively, if a relation `(M, V)` $\in$ `read` is discovered, then `(M, V)` $\in$ `access` is also implied.

More expressive DLs allow defining a role as the composition of two or more roles. If both `(a, b)` $\in$ `R` and `(b, c)` $\in$ `S` hold, and a role `T` is defined as `T` $\equiv$ `R` $\circ$ `S`, then `(a, c)` $\in$ `T` also holds. For example, in our ontology, a role `classReadVariable` can be defined as the composition of `contain` and `read`

$$\texttt{classReadVariable} \equiv \texttt{contain} \circ \texttt{read}$$

In such a case, if class C contains a method M, and the method M reads a variable V, that is, `(C, M)` $\in$ `contain`, and `(M, V)` $\in$ `read`, then `(C, V)` $\in$ `classReadVariable` is also implied.

Instances of roles (i.e. relationships between source code entities) are often implicit, but can still be recognised through static source code analysis. For example, if within the body of a method, a variable is found in the left hand side of an assignment expression, such as

```
public int aMethod(){
    ……
    aField = 1;
    ……
}
```

then an instance of the `write` role will be created, indicating that the method writes the variable, like: `(aMethod, aField)` $\in$ `write`.

*Characterising object-oriented programming*: We designed our ontology to support three key features specific to the object-oriented programming (OOP) paradigm, namely encapsulation, inheritance and polymorphism.

(1) Encapsulation: A `hasModifier` role is defined to specify whether a `SourceObject` (e.g. `Class`, `Method` or `Field`) is public, private or protected.

(2) Inheritance: We define `hasSuper` and `hasSub` to represent class inheritance. The `hasSuper` and `hasSub` roles are transitive.

(3) Polymorphism: As described earlier, we use static source code analysis techniques to populate the source code ontology. Polymorphism is often used to specify a particular dynamic aspect of OOP languages, which can only be captured partially by static analysis approaches.


© The Institution of Engineering and Technology 2008

doi: 10.1049/iet-sen:20070062

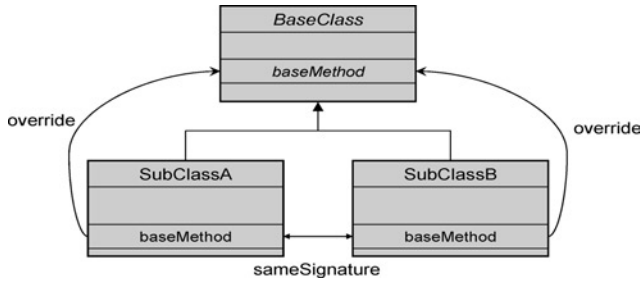



**Figure 6** *Static representation of polymorphism in OOP*

Within our ontology, we provide an `override` role to denote that a method overrides another method defined in the superclass. A `sameSignature` role is also defined to represent the relationship between two methods that are defined in different subclasses and share the same signature, that is, they both override the same method defined in the superclass. If two methods are in a `sameSignature` relationship, both of them may potentially be invoked when the method in the superclass is invoked, for example, through dynamic binding (polymorphism).

For example, as shown in the UML class diagram in Fig. 6, both SubClassA and SubClassB inherit from BaseClass. Therefore the following relations can be automatically identified

$$(\texttt{SubClassA.baseMethod}, \texttt{BaseClass.baseMethod}) \in \texttt{override}$$

$$(\texttt{SubClassB.baseMethod}, \texttt{BaseClass.baseMethod}) \in \texttt{override}$$

and

$$(\texttt{SubClassA.baseMethod}, \texttt{SubClassB.baseMethod}) \in \texttt{sameSignature}$$

Both the `override` and the `sameSignature` role are transitive. Override is a sub-role of the symmetric `sameSignature` role.

*3.3.2 Documentation ontology:* The documentation ontology further extends the source code ontology with a large body of concepts that can be discovered in software documents. At the current stage of our research, this ontology is mainly based on concepts found in the software domain, including concepts such as programming languages, algorithms, data structures and design decisions, especially design patterns and software architectures. Fig. 7 shows the top-level concepts of the documentation ontology.

The software documentation ontology has been designed to support automatic ontology population through a text mining system by adapting the ontology design requirements outlined in [26] for the software engineering domain. Specifically, we included

A *text model* to represent the structure of documents, for example, classes for sentences, paragraphs and text positions, as well as NLP-related concepts that are discovered during the analysis process, like noun phrases and co-reference chains. These are required for anchoring detected entities (populated instances) in their originating documents.

*Lexical information* facilitating the detection of entities in documents, like the names of common design patterns, programming language-specific keywords or architectural styles; and lexical normalisation rules for entity normalisation.

*Relationships* between the classes, including the ones modelled in the source code ontology. These relationships allow us to automatically restrict NLP-detected relations to semantically valid ones (e.g. a relationship like `Variable` implements `interface`, which can result from parsing a grammatically ambiguous sentence, can be filtered out since it is not supported by the ontology).

Finally, *source code entities* that have been automatically populated through source code analysis can also be utilised for detecting corresponding entities in documents [27].

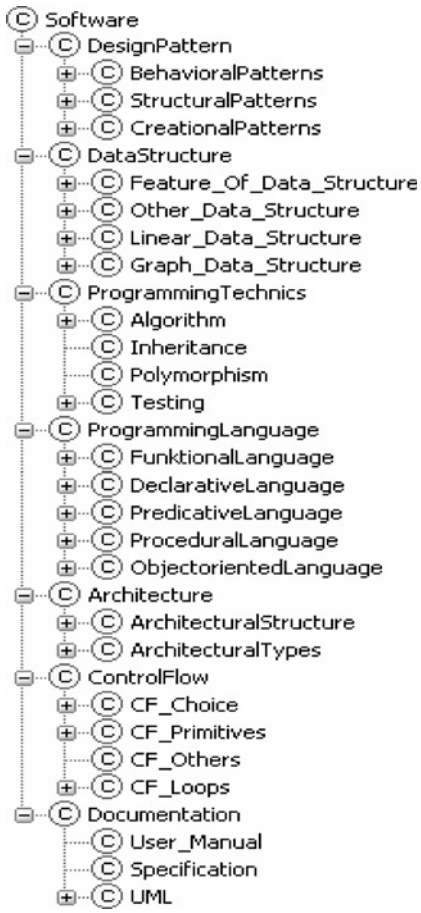



**Figure 7** *Top-level concept hierarchy in the documentation ontology*



© The Institution of Engineering and Technology 2008

# 4 System architecture and implementation

In this section, we describe the general system architecture of our Semantic Web-Enabled Software Maintenance Environment, and illustrate how the detected Requirement #2 of such a system are fulfilled by our implementation. In particular, we describe the overall architecture of our approach in Section 4.1 and some implementation details in Section 4.2. A particular feature of our implementation, an Eclipse plug-in for querying the ontology, is further described in Section 4.3.

## 4.1 Architecture overview

Our system is based on common Semantic Web infrastructures, including an ontology reasoner, a knowledge management system and public RDF/OWL APIs. In order to utilise the structural and semantic information found in various software artefacts, we have developed source code analysis and text mining sub-systems that can automatically extract concept instances and their relations from source code and documents, respectively. An Eclipse integrated user interface is provided to support the querying and exploring of the populated software ontology. Our system has been designed to enable software maintainers in both discovering concepts and relations within a software system, as well as automatically inferring implicit relations among different artefacts. The automatic population of the software ontology stipulated by Requirement #3 is handled by two sub-systems for source code analysis and text mining, described in more detail in the next section. The discovered instances can then be linked through ontology alignment [21, 22] techniques. Based on the software ontology, users can also define new concepts or instances for particular reverse engineering tasks through an ontology management interface.

## 4.2 System implementation

The right side of Fig. 8 shows the software ontology as described previously in Section 3.3. The two sub-ontologies are modelled in OWL-DL [14, 28] and were created using the Protégé-OWL extension of Protégé [29], a free ontology editor.

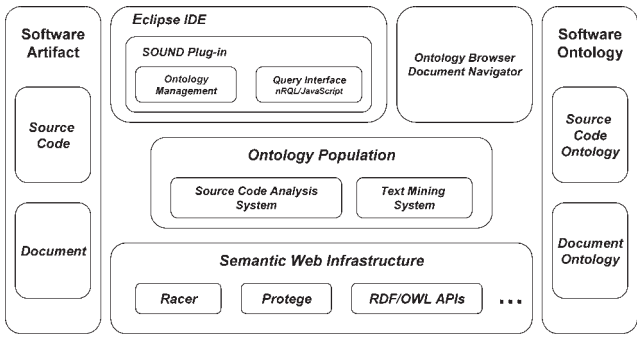



**Figure 8** *Semantic web-enabled software maintenance architecture*

Racer [12], an ontology inference engine, is adopted to provide reasoning services. The Racer system is a highly optimised DL system that supports reasoning about instances, which is particularly useful for the software maintenance domain, where one has to process a large amount of instances efficiently.

Automatic ontology population is handled by two subsystems: The source code analysis and the document analysis. The source code analysis subsystem is based on JDT [30], which is a Java parser provided by Eclipse [31]. The system analyses the abstract syntax tree (AST) of Java source code and identifies entities and their relations to populate the source code ontology.

The text mining subsystem has been implemented based on the GATE (General Architecture for Text Engineering) framework [20], one of the most widely used NLP tools. Starting with version 3.0, GATE has been featuring built-in ontology support in form of an abstraction layer between the components of an NLP system and the various ontology representations. This layer is built on Jena [32] as RDF-Store, which enables the use of OWL ontologies from within GATE.

## 4.3 Query interface

Within our system, a software engineer can use the Racer query language nRQL [15] to retrieve instances of concepts and roles in the ontology. An nRQL query uses arbitrary concept names and role names in the ontology to specify properties of the result. In a query, variables can be used to bind to instances that satisfy the query.

However, the use of nRQL queries is still largely restricted to users with a high mathematical/logical background due to nRQL's syntax, which, although comparatively straightforward, is still difficult for programmers to understand and therefore may be difficult to apply. To bridge this gap between practitioners and Racer, we have introduced an additional scripting language, based on JavaScript, as a query language. We introduced a set of built-in functions and classes in the JavaScript interpreter Rhino [33] to simplify querying the ontology for users.

Within the JavaScript interpreter, we provide a set of logic functions for formulating complex concepts. Using these logic functions, users can construct their own concepts. For example, the concept `ClassMember` discussed in Section 3.3.1 can be specified using the built-in functions as

```
ontology.define_concept
                (″DesignPatternDoc},
                      AND
          (″Paragraph, EXIST(″contains},
              ″DesignPattern})))
```


© The Institution of Engineering and Technology 2008

doi: 10.1049/iet-sen:20070062

```
var design_pattern_doc = new Query();
                // create an new query
design_pattern_doc.declare("P", "DP");
                // declare two query variables in the query
design_pattern_doc.restrict("P", "Paragraph");
                // restrict P to be bound to a paragraph instance
design_pattern_doc.restrict("DP", "DesignPattern");
                // restrict DP to be bound to a design pattern
                 instance
design_pattern_doc.restrict("P", "contains", "DP");
                // restrict P contains DP
design_pattern_doc.retrieve("P");
                // the query will only retrive instances of P
var result = ontology.query(design_pattern_doc);
                // perform the query
```

**Figure 9** *Retrieve paragraphs describing design pattern instances*

Two classes, Query and Result, are provided to assist users in composing queries and manipulating the results. Users can arbitrarily use the vocabulary in the ontology to retrieve instances with specified properties. The typical procedure of composing a query is as follows: (1) query variables are declared; (2) restrictions that apply to the variables are specified using concepts, roles and instances in the ontology and (3) the query is submitted to the built-in JavaScript object called 'ontology'.

The result of a query is a set of tuples that satisfy the specified restrictions. For example, the query script (Fig. 9) retrieves all paragraphs that contain design pattern instances from the documentation ontology.

The query first declares two variables *P* and DP, and then specifies that *P* shall be bound to an instance of `Paragraph` and DP to an instance of `DesignPattern`. The third restriction specifies that *P* and DP shall have a `contains` relation. The next statement states that this query only retrieves instances bound to *P*.

The scriptable query language allows users to benefit from both the declarative semantics of DLs as well as the fine-grained control abilities of procedural languages.

The query interface of our system is a plug-in that provides OWL integration for the Eclipse software development platform. Table 2 compares our JavaScript query language with the nRQL and SPARQL [34] languages through a simple query.

# 5 Automatic ontology population

One of the major challenges for software maintainers is the large amount of information that has to be explored and analysed as part of typical maintenance activities. Therefore support for automatic ontology population is essential for the successful adoption of Semantic Web technology in software maintenance (Requirement #3). In this section, we describe in detail the automatic population of our ontologies from existing artefacts: source code (Section 5.1) and documents (Section 5.2).

## *5.1 Source code ontology population*

Concepts in the source code ontology typically have a direct mapping to source code entities, allowing instances of these concepts to be automatically recognised by our source code ontology population subsystem. Within our implementation, we utilise the Eclipse JDT compiler in order to read the source code, performing common tokenisation and syntax analysis to produce an AST. Our population subsystem traverses the AST created by the JDT compiler to identify concept instances and their relations, which are then passed to an OWL generator for ontology population (Fig. 10). Furthermore, it can also identify instances of roles (i.e. relations between source code entities) by statically analysing the source code.

As an example, consider a single line of Java source code: `public int sort()`, which declares a method called `sort`. A simplified AST corresponding to this line of source code is shown in Fig. 10. We traverse this tree by first visiting the root node `Method Declaration`. At this step, the system understands that a `Method` instance shall be created. Next, the `Name Node` is visited to create the instance of the `Method` class, in this case `sort`. Then, the `Modifier Node` and `Type Node` are also visited, in order to establish the relations with the identified instance. As a result, two relations, `sort hasModifier public` and `sort hasType int`, are detected. The numbers of instances and relations identified

**Table 2** Query comparison: JavaScript interface, nRQL and SPARQL

| JavaScript | nRQL | SPARQL |
|---|---|---|
| `var query = new Query();`<br>`query.declare("M", "V", "C");`<br>`query.restrict("M", "definedIn", "C");`<br>`query.restrict("M", "contains", "V");`<br>`query.retrieve("C", "V");`<br>`var result = ontology.query(query);` | `(RETRIEVE (?C ?V)`<br>`(AND`<br>`(?M ?C definedIn)`<br>`(?M ?V contains)))` | `PREFIX sc: <http://...>`<br><br>`SELECT ?C ?V`<br>`WHERE`<br>`{?M sc:definedIn ?C.`<br>`?M sc:contains ?V}` |



© The Institution of Engineering and Technology 2008

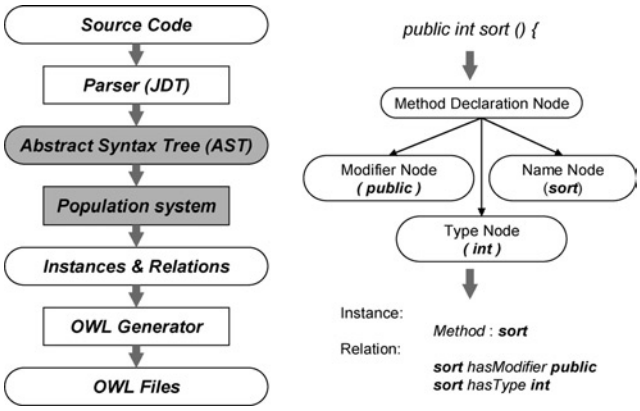



**Figure 10** *Automatic population of the source code ontology*

by our system depend on the complexity of the ontology and the size of the source code to be analysed. At the current stage of our research, we limit the population of the source code ontology to 38 of the higher-level concepts (classes) and 41 types of relations (ObjectProperties). We restrict the ontology population currently to these high-level concepts (e.g. package, class, method) due to the application domain of the ontology. Our objective is currently to support a more general system comprehension rather than focusing on specific, low-level source analysis (e.g. at the expression level).

We have performed several case studies on different open source systems to evaluate the size of the populated source code ontology. The results are described in Section 6.

## 5.2 Documentation ontology population

We developed a custom text mining system to extract knowledge from software documents and populate the corresponding sub-ontology [27]. Note that, in addition to the software documentation ontology, the text mining system can also import the instantiated source code ontology corresponding to the document(s) under analysis.

The system first performs a number of standard preprocessing steps, such as tokenisation, sentence splitting, part-of-speech tagging and noun phrase chunking [35]. Then, named entities modelled in the software ontology are detected in a two-step process: First, an OntoGazetteer is used to annotate tokens with the corresponding class or classes in the software ontology (e.g. the word 'architecture' would be labelled with the architecture class in the ontology). Complex named entities are then detected in the second step using a cascade of finite-state transducers implementing custom grammar rules written in the JAPE language, which is part of GATE. These rules refer back to the annotations generated by the OntoGazetteer, and also evaluate the ontology. For example, in a comparison like `class=="Keyword"`, the ontological hierarchy is taken into account so that a `JavaKeyword` also matches, since a Java keyword is a keyword in the ontology. This significantly reduces the overhead for grammar development and testing.

The next major steps are the normalisation of the detected entities and the resolution of co-references. Normalisation computes a canonical name for each detected entity, which is important for automatic ontology population. In natural language texts, an entity, like a method, is typically referred to with a phrase like 'the myTestMethod provides...'. Here, only the entity myTestMethod should become an instance of the Method class in the ontology. This is automatically achieved through lexical normalisation rules, which are stored in the software ontology as well, together with their respective classes. Moreover, throughout a document a single entity is usually referred to with different textual descriptors, including pronominal references (like 'this method'). In order to find these references and export only a single instance into the ontology that references all these occurrences, we perform an additional co-reference resolution step to detect both nominal and pronominal co-references.

The next step is the detection of relations between the identified entities in order to compute predicate-argument structures, like implements (class, interface). Here, we combine two different and largely complementary approaches: a deep syntactic analysis using the SUPPLE bottom-up parser and a number of pre-defined JAPE grammar rules, which are again stored in the ontology together with the relation concepts.

Finally, the text mining results are exported by populating the software documentation sub-ontology using a custom GATE component, the OwlExporter. The exported, populated ontology also contains document-specific information; for example, for each class instance the sentence it was found in is recorded.

For further details on our software text mining system, we refer the reader to [27].

# 6 Evaluation and application

In this section, we analyse the performance of our approach. The discussion is split into two parts: in the first subsection, we present quantitative numbers for the automatic ontology population. The second subsection then demonstrates the applicability of our approach on a real-world code base.

## 6.1 Ontology population evaluation

In what follows, we first analyse the performance of our system, in particular with respect to the automatic population of the source code and document ontologies.

*6.1.1 Source code analysis performance:* Table 3 summarises the evaluation results of the automatic source code population subsystem. The information provided in the table includes the size of the software system being


© The Institution of Engineering and Technology 2008
doi: 10.1049/iet-sen:20070062

**Table 3** Source code and generated ontology sizes

| | LOC, k | Processing time, s | Instances | Relations |
|---|---|---|---|---|
| java.util | 24 | 13.62 | 10 140 | 47 009 |
| InfoGlue [36] | 40 | 27.61 | 15 942 | 77 417 |
| Debrief [37] | 140 | 67.12 | 52 406 | 244 403 |
| UDig [38] | 177 | 82.26 | 69 627 | 284 692 |

measured in lines of code (LOC). The processing time includes both AST traversal and ontology population time in seconds. The size of the resulting source code ontology is measured in terms of concept instances and relations identified as shown in the last two columns of the table. As it can be observed from Table 3, for the small to medium size systems the processing time, number of instances and relations are directly dependent on the size of the system. It has to be noted that for the analysis of (very) large systems on computers without sufficient memory available to store the complete AST and ontology in memory, the processing time might increase significantly.

*6.1.2 Text mining performance:* So far [27], we evaluated our text mining subsystem on two collections of texts: a set of five documents (7743 words) taken from the Java documentation for the `Collections` framework [39] and a set of seven documents (3656 words) from the documentation of the uDig [40] geographic information system (GIS). The document sets were chosen because of the availability of the corresponding source code.

Both sets were manually annotated for named entities, including their ontology classes and normalised forms, as well as relations between the entities. In what follows, we present results for the named entity recognition and entity normalisation. For performance evaluation of the software named entity recognition, we computed the standard `precision`, `recall` and `F-measure` results. A named entity was only counted as correct if it matched both the textual description and ontology class. Table 4 shows the results for two experiments: first running only the text mining system over the corpora (left side) and secondly performing the same evaluation after running the code analysis, using the populated source code ontology as an additional resource for named entity detection as described in [27].

As it can be observed, the text mining system achieves a very high precision (90%) in the named entity detection task, with a recall of 62%. With the imported source code instances, these numbers become reversed: the system can now correctly detect 87% of all entities, but with a lower precision of 67%. The drop in precision after code analysis is mainly due to two reasons. Since names in the software domain do not have to follow any naming conventions, simple nouns or verbs often used in a text will be mis-tagged after being identified as an entity appearing in a source code. For example, the Java method `sort` from the collections interface will cause all instances of the word 'sort' in a text to be marked as a method name. Another precision hit is due to the current handling of class constructor methods, which are typically identical to the class name. Currently, the system cannot distinguish the class name from the constructor name, assigning both ontology classes (i.e. `Constructor` and `OO_Class` for a text segment, where one will always be counted as a false positive.

Both cases require additional strategies when importing entities from the source code analysis, which are currently under investigation. However, the current results already underline the feasibility of our approach of integrating code analysis and NLP.

We also evaluated the performance of our lexical normalisation rules for entity normalisation, since correctly normalised names are a prerequisite for the correct population of the result ontology. For each entity, we manually annotated the normalised form and computed the accuracy A as the percentage of correctly normalised entities over all correctly identified entities. Table 4 shows the results for both the system running in text mining mode alone and with additional source code analysis. As can be seen from the table (column A), the normalisation component's performance is very satisfying.

## 6.2 Application evaluation

An initial evaluation of the practical applicability of our approach has been performed on a large open source GIS, the 'User-friendly Desktop Internet GIS' (uDig) [41]. The uDig system is implemented as a set of Eclipse plug-ins

**Table 4** Text mining evaluation results: entity recognition and normalisation performance

| Corpus | Text mining only | | | | With source code ontology | | | |
|---|---|---|---|---|---|---|---|---|
| | *P* | *R* | *F* | *A*, % | *P* | *R* | *F* | *A*, % |
| Java collections | 0.89 | 0.67 | 0.69 | 75 | 0.76 | 0.87 | 0.79 | 88 |
| UDig | 0.91 | 0.57 | 0.59 | 82 | 0.58 | 0.87 | 0.60 | 84 |
| total | 0.90 | 0.62 | 0.64 | 77 | 0.67 | 0.87 | 0.70 | 87 |



© The Institution of Engineering and Technology 2008

that provide geographic information management integration. The uDig documents used in the study consist of a set of JavaDoc files and a requirements analysis document [42].

Links between the uDig implementation and its documentation are recovered by first performing source code analysis to populate the source code ontology (Section 5.1). The resulted ontology contains instances of `Class`, `Method`, `Field` etc., and their relations such as inheritance, invocation etc. Our text mining system (Section 5.2) then takes the identified class names, method names and field names as input to populate the documentation ontology. Through this text mining process, a large number of Java language concept instances are discovered in the documents, as well as design level concept instances such as design patterns [23] or architectural styles [43]. The ontology alignment rules are then applied to link both the documentation ontology and the source code ontology. Parts of our results are shown in Fig. 11, with the content of the figure corresponding to the following sentences:

Sentence_2544: 'For example if the class `FeatureStore` is the target class and the object that is clicked on is a `IGeoResource` that can resolve to a `FeatureStore` then a `FeatureStore` instance is passed to the operation, not the `IGeoResource`'.

Sentence_712: 'Use the visitor pattern to traverse the AST'

Fig. 11 shows that in the uDig documents, our text mining system was able to discover that a sentence (sentence_2544) contains both class instance `_4098_FeatureStore` and `_4100_IGeoResource`. Both of these instances can be linked to the instances in source code ontology – `org.geotools.data.FeatureStore` and `net.refractions.udig.catalog.IGeoResource`, respectively. In addition, in another sentence (sentence_712), a class instance (_719_AST) and a design pattern instance (_718_visitor_pattern) are also identified. Instance `_719_AST` can then be linked in a similar manner to the `net.refractions.udig.catalog.util.AST` interface in the source code ontology.

**Script1**

```
var query = new Query();                    // define a new query
query.declare("M1", "M2",  "C");            // declare three query variables
query.restrict("M1", "Method");             // M1 is a method
query.restrict("M2", "Method");             // M2 is also a method
query.restrict("C", "Class");               // C is a class
query.restrict("M1", "definedIn", "C");     // M1 is defined in C
query.restrict("M2", "definedIn",
"org.geotools.data.FeatureStore");          // M2 is defined in FeatureStore
query.restrict("M1", "calls", "M2");        // M1 calls M2
query.retrieve("C");                        // this query only retrieve C
var result = ontology.query(query);         // perform the query
```

**Figure 12** *Query on source code ontology*

After the source code ontology and documentation ontology are linked, queries regarding the source code entities, design level concepts and their occurrences in documents can be performed using the reasoning services provided by our ontology reasoner, Racer. For example, during the comprehension of the class `FeatureStore`, a maintainer may want to study the classes that are related to `FeatureStore`. Within the source code ontology, a query similar to Script 1 (Fig. 12) can be performed to retrieve all classes that contain methods that call class `FeatureStore`.

Unfortunately, the class `IGeoResource`, which has a documented relation with `FeatureStore` (Script 1), will not be returned by such a query, because `IGeoResource` has no explicit invocation relations with `FeatureStore` in the uDig implementation. In addition to these types of source code queries, the reverse engineer can perform queries that cross the boundaries between source code and documentation. Such types of queries are supported due to the already established links between the source code and documentation ontology. For example, Script 2 (shown in Fig. 13) retrieves all classes that occur in the same sentences as class `FeatureStore`. This time, class

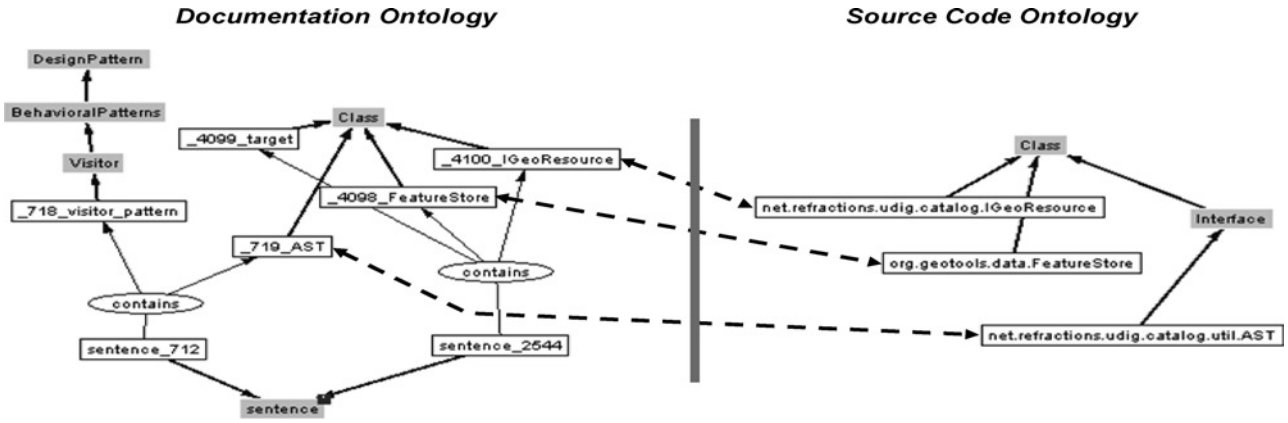



**Figure 11** *Linked source code and documentation ontology*


© The Institution of Engineering and Technology 2008

doi: 10.1049/iet-sen:20070062

**Script 2**

```
var query = new Query();                    // define a new query
query.declare("S", "C");                    // declare two query variables
query.restrict("S", "Sentence");            // S is a Sentence
query.restrict("C", "Class");               // C is a Class
query.restrict("S", "contains",
"org.geotools.data.FeatureStore");          // S contains FeatureStore
query.restrict("S", "contains", "C");       // S also contains C
query.retrieve("C", "S");                   // retrieve C and the sentence S
var result = ontology.query(query);         // perform the query
```

**Figure 13** *Query on documentation ontology*

`IGeoResource` will be returned because both classes occur in sentence_2544. The retrieved classes, as well as the associated sentences therefore provide additional information useful for reverse engineers to understand the class `FeatureStore`.

The linked source code and documentation ontologies also provide the ability to combine semantic information from both software implementation and documentation. For example, our text mining system has detected that class AST is potentially a part of a visitor pattern (Fig. 11). In order to retrieve all documented information related to the detected pattern, the Script 3 (Fig. 14) can be used to retrieve all text paragraphs that describe the sub classes of AST.

This query utilises both, the programming language semantics, such as the inheritance relation between query variable *C* and the class AST, and the structural information of documentation, such as the containing relation between *P* and *C*. The result of this query therefore contains all text paragraphs that describe the sub classes of AST, that is, the visitor pattern. It has to be noted that the role `contains` is a transitive relation to describe the document structure. The ontology reasoner can automatically resolve the transitivity from `Paragraph` to `Sentence` and from `Sentence` to `Class`.

# 7 Related work and discussions

There exists a significant body of work on analysing and modelling source code and its semantic structure [44, 45], however these approaches are mostly limited to modelling and analysing source code artefacts and do not include the recovery of traceability between source code and documents written in natural language.

**Script 3**

```
var query = new Query();                    // define a new query
query.declare("P", "C");                    // declare two query variables
query.restrict("P", "Paragraph");           // P is a paragraph
query.restrict("C", "Class");               // C is a class
query.restrict("C", "hasSuper",
"net.refractions.udig.catalog.util.AST");   // C is a sub-class of AST
query.restrict("P", "contains", "C");       // P contains C
query.retrieve("P");                        // this query only retrieve P
var result = ontology.query(query);         // perform the query
```

**Figure 14** *Query across the source code and documentation ontology*

Traditional approaches dealing with software documents are mainly based on IR techniques [18], which address the indexing, classifying, and retrieving of information in natural language documents. Some existing research recovers traceability links between source code and design documents using IR techniques by indexing software documents and then automatically linking software requirements documents through these indexes to implementation artefacts. Antoniol *et al.* [6] addressed the traceability problems by linking software documents and code using both probabilistic methods and a Bayesian classifier. They apply their IR model on two case studies to trace C++ source code onto manual pages and Java code to functional requirements. Marcus *et al.* [8, 46] used latent semantic indexing (LSI) [47], which is a machine-learning model that induces representations of the meaning of words by analysing the relationships between words and passages in large bodies of text, to recover traceability links and to locate domain concepts in source code. Their semantic indexing approach extracts the meaning (semantics) of the documentation and source code, and then uses this information to identify traceability links based on similarity measures. Hayes *et al.* [48] also demonstrated a tool that uses a vector space model to improve requirements tracing.

In contrast to these IR approaches, our work also utilises structural and semantic information found in both the documentation and the source code by means of text mining and source code parsing. This additional information allows us to recover links that would otherwise not be discovered using traditional IR techniques [17, 18]. Classical IR models only compare the occurrences of individual keywords. However, in many situations, synonyms and concept hierarchies (general concept against specific concept) have to be taken into account. For example, the query 'people' should also match the document that contains 'person' 'student' or 'professor', as 'person' is a synonym of 'people', and 'student' and 'professor' are both more specialised concepts of 'people'. Thesaurus-based retrieval model [49] and ontology-based models [50] address this issue by utilising a collection of information concerning relationship between different terms. However, IR-based approaches typically neglect structural and semantic information in the software documents, therefore limiting their ability in providing results with regards to the 'meaning' of documents.

Very little previous work exists on text mining software documents containing natural language. Most of this research has focused on analysing texts at the specification level, for example, in order to automatically convert use case descriptions into a formal representation [51, 52] or detect



© The Institution of Engineering and Technology 2008

inconsistent requirements [53]. In contrast, we aim to support the automatic recovery of traceability links for legacy systems. To the best of our knowledge, there has been so far no attempt to automatically cross-link entities (e.g. methods, design patterns, architectures) detected by text mining software documents with corresponding entities identified through source code analysis, which is an important contribution of our work. Furthermore, compared with logic programming approaches, using ontologies enables us to extend our resulting model to reflect more closely an 'open' world assumption [11], allowing for the integration of incomplete knowledge (artefacts) and extendibility of our ontological knowledge base to reflect more closely the creation of mental models used during program comprehension [3].

Ontologies have been commonly regarded as a standard technique for representing domain semantics and resolving semantic ambiguities. Existing research on applying DLs or formal ontology in software engineering have been addressed in early works of the LaSSIE [54] and CBMS [3] systems. In addition to above knowledge-based software engineering tools, other researchers have proposed to apply DL in the software engineering domain. In [55], Welty and Ferrucci presented a formal ontology that may capture and re-use architectural level knowledge from software documents. Möller [56] presented an approach that integrates OO programming methodologies and DL knowledge systems. In [57], Yang *et al.* presented an approach that can extract domain ontologies from legacy systems written in COBOL for program comprehension and re-engineering. Ontologies have also been used in software engineering to establish a common terminology or to model specific aspects of software engineering processes [58, 59]. Compared with our approach, these systems are however much more restricted by the expressiveness of their underlying ontology languages. In addition, these systems also lack the support of optimised DL reasoners, such as Racer in our case.

The introduction of an ontological representation for software artefacts allow us to utilise existing techniques, such as text mining and information extraction [17], to 'understand' parts of the semantics conveyed by these informal information resources and thus to integrate information from different sources at finer granularity levels. Note that our approach relies on an ontology that has been explicitly designed for text mining, unlike the existing approaches employing text mining for ontology learning, like the work by Sabou [60] and Sabou and Bontcheva [61]. We believe that the current methods for automated ontology construction do not provide the necessary level of detail as discussed in Section 3.3. However, a future combination of both approaches, that is, automatically enriching a pre-designed ontology with newly detected concepts, might deliver additional benefits, for example, in dealing with domain-specific information.

In our previous work, we have already demonstrated how the ontological model of source code and documentation can support various maintenance tasks, such as program comprehension [59], architectural analysis [24] and security analysis [62]. In another work, we have examined the requirements for software reverse engineering repositories [63], where we focus on dealing with incomplete and inconsistent knowledge on software artefacts obtained from different sources. We aim at incorporating this work into our ontology approach to systematically deal with the possible inconsistencies between source code and document analysis discovered during the automatic ontology population.

# 8 Conclusions and future work

The presented research addresses an important issue in the reverse engineering domain, the recovery and maintenance of traceability links among existing documents and source code artefacts. We present a novel approach that provides formal ontological representations for both source code and document artefacts. The ontologies capture structural and semantic information conveyed in these artefacts, and therefore allow us to recover traceability links between the software implementation and documentation at the semantic level.

In addition, utilising state-of-the-art ontology reasoners such as Racer, our approach also allows inferring implicit relations between discovered concept instances. The linked ontologies provide the capability to perform queries across the boundary between programming language and natural language.

Furthermore, our documentation ontology identifies different design-level concept instances such as design patterns and architectural styles. These identified instances, linked to source code entities, allow users to discover relations between source code and its corresponding design information at different levels of abstraction.

As part of our future work, we will be exploring a hierarchical linking strategy, starting from code, including inline comments (like JavaDoc), over implementation, design, and specification documents to domain-specific knowledge, to allow us to offer a truly complete process life-cycle for the automated support of traceability links.

# 9 Acknowledgments

Qiangqiang Li contributed to the software text mining system.

# 10 References

[1] SEACORD R., PLAKOSH D., LEWIS G.: 'Modernizing legacy systems: software technologies, engineering processes and business practices', Boston: Addison-Wesley, 2003 (SEI Series in SE), AW (2003)


© The Institution of Engineering and Technology 2008

doi: 10.1049/iet-sen:20070062

[2] SOMMERVILLE I.: 'Software engineering' (Addison Wesley, 2000, 6th edn.)

[3] STOREY M.A., SIM S.E., WONG K.: 'A collaborative demonstration of reverse engineering tools', *ACM SIGAPP Appl. Comput. Rev.*, 2002, **10**, (1), pp. 18–25

[4] WELTY C.: 'Augmenting abstract syntax trees for program understanding'. Proc. 1997 Int. Conf. Automated Software Engineering (IEEE Computer Society Press, 1997), November 1997, pp. 126–133

[5] LETHBRIDGE T.C., NICHOLAS A.: 'Architecture of a source code exploration tool: a software engineering case study', Technical Report, TR-97-07, Department of Computer Science, University of Ottawa, 1997

[6] ANTONIOL G., CANFORA G., CASAZZA G., DE LUCIA A.: 'Information retrieval models for recovering traceability links between code and documentation'. Proc. IEEE Int. Conf. Software Maintenance, San Jose, CA, 2000

[7] ARKLEY P., MASON P., RIDDLE S.: 'Position paper: enabling traceability'. Proc. 1st Int. Workshop on Traceability in Emerging Forms of Software Engineering, Edinburgh, Scotland, September 2002, pp. 61–65

[8] MARCUS A., MALETIC J.I.: 'Recovering documentation-to-source-code traceability links using latent semantic indexing'. Proc. 25th Int. Conf. Software Engineering, 2002

[9] ZHANG Y., WITTE R., RILLING J., HAARSLEV V.: 'An ontology-based approach for the recovery of traceability links'. 3rd Int. Workshop on Metamodels, Schemas, Grammars and Ontologies for Reverse Engineering (ATEM 2006), Genoa, Italy, 1 October 2006

[10] JURAFSKY D., MARTIN J.H.: 'Speech and language processing' (Prentice Hall, 2000)

[11] BAADER F., CALVANESE D., MCGUINNESS D.L., NARDI D., PATEL-SCHNEIDER P.F.: 'The description logic handbook' (Cambridge University Press, 2007, 2nd edn.)

[12] HAARSLEV V., MÖLLER R.: 'RACER system description'. Proc. Int. Joint Conf. Automated Reasoning IJCAR'2001, Italy, (Springer-Verlag, 2001), pp. 701–705

[13] GRUBER T.: 'A translation approach to portable ontology specifications', *Knowl. Acquis.*, 1993, **5**, (2), pp. 199–220 Academic Press

[14] OWL Web Ontology Language Reference, W3C Recommendation, 10 February 2004, available at: http://www.w3.org/TR/owl-ref/

[15] HAARSLEV V., MÖLLER R., WESSEL M.: 'Querying the semantic web with Racer + nRQL'. Proc. KI-2004 Int. Workshop on Applications of Description Logics (ADL'04), Ulm, Germany, 24 September 2004

[16] HORROCKS I., KUTZ O., SATTLER U.: 'The even more irresistible SROIQ'. Proc. 10th Int. Conf. Principles of Knowledge Representation and Reasoning (KR 2006), 2006, (AAAI Press), pp. 57–67

[17] FELDMAN R., SANGER J.: 'The text mining handbook: advanced approaches in analyzing unstructured data' (Cambridge University Press, 2006)

[18] BAEZA-YATES R., RIBEIRO-NETO B.: 'Modern information retrieval' (Addison Wesley, 1999)

[19] KIRYAKOV A., POPOV B., TERZIEV I., MANOV D., OGNYANOFFE D.: 'Semantic annotation, indexing, and retrieval', *J. Web Semant.*, 2005, **2**, (1), pp. 49–79

[20] CUNNINGHAM H., MAYNARD D., BONTCHEVA K., TABLAN V.: 'GATE: a framework and graphical development environment for robust NLP tools and applications'. Proc. 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia, July 2002

[21] EHRIG M.: 'Ontology alignment: bridging the semantic gap' (Springer, 2006)

[22] NOY N.F., STUCKENSCHMIDT H.: 'Ontology alignment: an annotated bibliography – semantic interoperability and integration' (Dagstuhl, Germany, 2005) available at http://www.bibsonomy.org/bibtex/24f0e1c4e86993600cc975bfa31f54394/achorley

[23] GAMMA E., HELM R., JOHNSON R., VLISSIDES J.: 'Design patterns – elements of reusable object-oriented software' (Addison-Wesley, 1994)

[24] WITTE R., ZHANG Y., RILLING J.: 'Empowering software maintainers with semantic web technologies'. 4th European Semantic Web Conference (ESWC 2007), Innsbruck, Austria, 3–7 June 2007

[25] SCHNEIDERMAN B.: 'An empirical studies of programmers'. Proc. 2nd Workshop on Empirical Studies of Programmers, Ablex Publichers, NJ, 1986

[26] WITTE R., KAPPLER T., BAKER C.J.O.: 'Ontology design for biomedical text mining', in 'Semantic web: revolutionizing knowledge discovery in the life sciences' (Springer Verlag, 2007)

[27] WITTE R., LI Q., ZHANG Y., RILLING J.: 'Ontological text mining of software documents'. 12th Int. Conf. Applications of Natural Language to Information Systems (NLDB 2007), CNAM, Paris, France, 27–29 June 2007

[28] OWL Web Ontology Language Guide, W3C: Available at: http://www.w3.org/TR/owl-guide/, accessed November 2007




© The Institution of Engineering and Technology 2008

[29] PROTÉGÉ ONTOLOGY EDITOR: Available at: http://protege.stanford.edu/, accessed September 2007

[30] ECLIPSE JAVA DEVELOPMENT TOOLS: Available at: http://www.eclipse.org/jdt/, accessed October 2007

[31] ECLIPSE: Available at: http://www.eclipse.org, accessed November 2007

[32] JENA: Available at: http://jena.sourceforge.net/, accessed July 2007

[33] RHINO: Available at: http://www.mozilla.org/rhino/, accessed March 2007

[34] W3C Recommendation, SPARQL Protocol and RDF Query Language (SPARQL), available at: http://www.w3.org/TR/rdf-sparql-query/, accessed November 2007

[35] GATE DOCUMENTATION: Available at: http://gate.ac.uk/documentation/, accessed July 2007

[36] INFOGLUE: Available at: http://www.infoglue.org, accessed February 2007

[37] DEBRIEF: Available at: http://www.debrief.info, accessed March 2007

[38] UDIG: Available at: http://udig.refractions.net, accessed March 2007

[39] JAVA COLLECTIONS: Available at: http://java.sun.com/j2se/1.5.0/docs/guide/collections/index.html, accessed February 2007

[40] UDIG GIS DOCUMENTATION: Available at: http://udig.refractions.net/, accessed March 2007

[41] UDIG SYSTEM: Available at: http://udig.refractions.net/confluence/display/UDIG/Home, accessed March 2007

[42] UDIG DOCUMENTATION: Available at: http://udig.refractions.net/docs/, accessed April 2007

[43] SHAW M., GARLAN D.: 'Software architecture: perspectives on an emerging discipline' (Prentice Hall Publisher, 1996)

[44] GUEHENEUC Y., MENS K., WUYTS R.: 'A comparative framework for design recovery tools'. Proc. Conf. Software Maintenance and Reengineering, Bari, Italy, 22 March

[45] GUÉHÉNEUC Y.: 'Ptidej: promoting patterns with patterns'. Proc. 1st ECOOP workshop on Building a System using Patterns, July 2005, Springer

[46] MARCUS A., SERGEYEV A., RAJLICH V., MALETIC J.I.: 'An information retrieval approach to concept location in source code'. 11th IEEE Working Conf. Reverse Engineering, Netherlands, 2004

[47] LANDAUER T.K., FOLTZ P.W., LAHAM D.: 'An introduction to latent semantic analysis. discourse processes', *Discourse Processes*, 1998, **25**, (2-3), pp. 259–284

[48] HAYES J.H., DEKHTYAR A., OSBORNE J.: 'Improving requirements tracing via information retrieval'. Proc. 11th IEEE Int. Requirements Engineering Conf., 2003, pp. 138–147

[49] KOWALSKI G.: 'Information retrieval systems: theory and implementation' (Kluwer Academic Publishers, 1997)

[50] KHAN L., MCLEOD D., HOVY E.: 'Retrieval effectiveness of an ontology-based model for information selection', *Int. J. Very Large Data Bases*, 2004, **13**, (1), pp. 71–85

[51] ILIEVA M.G., ORMANDJIEVA O.: 'Automatic transition of natural language software requirements specification into formal presentation'. 10th Int. Conf. Applications of Natural Language to Information Systems (NLDB), Alicante, Spain, 2005

[52] MENCL V.: 'Deriving behavior specifications from textual use cases'. Proc. Workshop on Intelligent Technologies for Software Engineering (WITSE'04), Austria, 2004

[53] KOF L.: 'Natural language processing: mature enough for requirements documents analysis?'. 10th Intl. Conf. on Applications of Natural Language to Information Systems (NLDB), Alicante, Spain, 15–17 June 2005

[54] DEVANBU P., BRACHMAN R.J., SELFRIDGE P.G., BALLARD B.W.: 'LaSSIE: a knowledge-based software information system', *Commun. ACM*, 1991, **34**, (5), pp. 36–49

[55] WELTY C., FERRUCCI D.A.: 'A formal ontology for reuse of software architecture documents'. 1999 Int. Conf. Automated Software Engineering, 1999, IEEE Computer Society Press

[56] MÖLLER R.: 'A functional layer for description logics: knowledge representation meets object-oriented programming'. OOPSLA '96, San Jose, California,

[57] YANG H.J., CUI Z., O'BRIEN P.: 'Extracting ontologies from legacy systems for understanding and re-engineering'. 23rd Int. Computer Software and Applications Conf., Washington, DC, USA, 1999

[58] WONGTHONGTHAM P., CHANG E., DILLON T.S., SOMMERVILLE I.: 'Software engineering ontology – instance knowledge Part I', *Int. J. Comput. Sci. Netw. Secur.*, USA, 2007

[59] MENG W.J., RILLING J., ZHANG Y., WITTE R., CHARLAND P.: 'An ontological software comprehension process model'. 3rd


© The Institution of Engineering and Technology 2008

doi: 10.1049/iet-sen:20070062

Int. Workshop on Metamodels, Schemas, Grammars and Ontologies for Reverse Engineering (ATEM 2006), Genoa, Italy, 1 October 2006

[60] SABOU M.: 'Extracting ontologies from software documentation: a semantic method and its evaluation'. Proc. ECAI-2004 Workshop Ontology Learning and Population, Valencia, Spain, 2004

[61] BONTCHEVA K., SABOU M.: 'Learning ontologies from software artifacts: exploring and combining multiple sources''. Proc. 2nd Int. Workshop on Semantic Web Enabled Software Engineering (SWESE 2006),

[62] ZHANG Y.G., RILLING J., HAARSLEV V.: 'An ontology based approach to software comprehension – reasoning about security concerns in source code'. Proc. 30th Int. Computer Software and Applications Conf., 2006

[63] KÖLSCH U., WITTE R.: 'Fuzzy extensions for reverse engineering repository models'. 10th Working Conf. Reverse Engineering (WCRE), Canada, 2003




© The Institution of Engineering and Technology 2008