Facilitate IT-Providing SMEs in Software Development: a Semantic Helper for Filtering and Searching Knowledge

Riccardo Martoglia
DII – Department of Information Engineering
University of Modena and Reggio Emilia
Modena, Italy
riccardo.martoglia@unimo.it

Abstract— **Software development is still considered a bottleneck in the advance of the Information Society. The recently started FACIT-SME European FP-7 project targets to facilitate the use and sharing of Software Engineering methods and best practices among software developing SMEs. On top of an Open Reference Model (ORM) serving as an underlying knowledge backbone, specific filtering/search mechanisms will support the identification of adequate processes and practices for specific enterprise needs. In this paper, we focus on the proposal of knowledge-based text analysis and retrieval techniques which will form a key component of the advanced filtering mechanisms of the project. The proposed solution is designed to be more powerful and flexible than standard syntactic search techniques, but also to be easily applicable for any SME. The experimental evaluation on the preliminary implementation shows promising results.**

Keywords-software engineering; information retrieval; text analysis; semantic knowledge; semantic similarity.

I. INTRODUCTION AND MOTIVATION

Over the last years, Software Engineering (SE) research has provided more and more advanced and promising techniques for facilitating software development. In particular, the integration between Software and Knowledge Engineering has recently become very important, and several techniques possibly enabling better domain knowledge sharing and assisting developers in specific tasks such as component reuse [1] or software process assessment [2, 3], have been proposed to the research community [4].

Nonetheless, back to the "real" software development world, recent studies have shown that a large number of fundamental challenges still need to be faced. In Europe, software development is becoming a bottleneck in the development of the Information Society [5], while, on a global scale, the quality and productivity of work has not been able to keep up with the society software needs [6]. These issues are especially critical in the case of SMEs in the software development market: indeed, even if innovative SE methodologies are constantly devised and presented, many enterprises are usually not able to take full advantage from them since they generally lack the resources and knowledge needed for the internal deployment of the required methods and tools. Indeed, being software an often underpaid product, SMEs need to allocate mostly all of their available resources on its production rather than, for instance, on new technology training.

This is the challenging scenario of the recently started European FP7 3 years project "Facilitate IT-providing SMEs by Operation-related Models and Methods (FACIT-SME)". The main project goal is to facilitate IT SMEs in using SE methods for design and development, systematizing their application integrated with the business processes. Another fundamental goal is to provide efficient and affordable certification of these processes according to internationally accepted standards, and to securely share best practices, tools and experiences with development partners and customers. In order to achieve these goals, the project will develop a novel Open Reference Model (ORM) [7] for ICT SMEs serving as an underlying knowledge backbone and, on top of that, a customizable Open Source Enactment System (OSES) [8] will provide IT support for the project-specific application of the ORM. More specifically, the ORM will store existing reference knowledge for software-developing SMEs, including different engineering methods, tools, quality model requirements and enterprise model fragments of IT SMEs, in a computer-processable form. On top of the ORM repository, specific search mechanisms, which will be a key part of the OSES, will support the identification of adequate processes and data structures for a specific enterprise. Different application scenarios, identified with the support of the participating SMEs and enumerating the possible use cases of the FACIT-SME solution, will be dealt with. The most notable ones include supporting the organizations in their need to find a new methodology ("From Scratch" scenario) or to modify an existing one in order to better manage its software development projects ("From Methodology" scenario). In all cases, through a filtering phase, which takes as input company and project information and, for the second scenario, existing methodology descriptions, the organization will receive a set of suggestions in the form of the most relevant / useful elements and models in the ORM. Subsequent phases will also include helping the organization to easily check quality constraints and refining the models and results in order to adapt it to its needs. Besides five R&D partners providing the required competences, the project consortium also includes five SMEs operating in the ICT domain which will evaluate the results in daily-life application.

In this paper, we focus on the foundations we are laying for the filtering/searching mechanisms, carefully considering the actual user-targets these techniques will be aimed at. More specifically, we will primarily take advantage of textual

information, a vital knowledge source not only in the ORM defined in the project but also in the documentation already available in each enterprise. In this respect, we propose an innovative approach based on text analysis and semantic retrieval techniques leading to the following achievements:

- it is powerful enough to provide enhanced searching effectiveness over standard syntactic techniques;
- it is general and flexible as a basis of many functionalities offered by the OSES (i.e. for filtering software methodologies for software process assessment and improvement, quality requirements for helping in certification process, best practices for faciliting knowledge sharing, and so on);
- it is devised for IT SMEs, providing them with easy-to-apply methods that do not require big investments or knowledge prerequisites, allowing them to query for the information they need in the way they are used to;
- it exploits the large amounts of textual knowledge (i.e. methodology descriptions, and so on) already available in each enterprise, without requiring complex conversions toward complex structured formats which would be time and cost consuming.

Such approach forms the foundations of a **Semantic Helper** component which will be overviewed in Section II, while the analysis and semantic search techniques themselves will be deepened in Sections III and IV, respectively. Section V shows the promising results of a preliminary experimental evaluation, while Section VI concludes the also by briefly analyzing related works.

II. SEMANTIC HELPER OVERVIEW

The Semantic Helper will support other components of the FACIT-SME solution in filtering/searching/analyzing relevant information available in the ORM, including:

a) **assisted filtering / selection** of ORM elements given specific enterprise objectives (e.g. in "From Scratch" scenario to give pointers to useful information for certification status);
b) **assisted suggestion proposal** for a given enterprise methodology (e.g. in Scenario "From methodology" to help identifying relevant information or gaps between the given methodology and the ORM methodologies);
c) **automatic matching** between ORM documents (such as quality requirements and SE methodologies).

In order to facilitate such processes, a representation of the key parts of the ORM in a semantic and machine-processable way is needed. Given the predominant importance of textual information in the ORM, first of all we provide the Semantic Helper with appropriate **text analysis** techniques (see Section III), which are designed to automatically extract a shared "terminology" from the given set of **documents**. The extracted terminology is enriched with statistical and semantic information (i.e. links to thesauri and domain vocabularies, definitions, synonyms), in order to obtain a computer-processable semantic glossary Note that the analysis can be applied not only to documents coming from the ORM (e.g. about different quality requirements or software methodologies), but also to any "external" document or query submitted by an enterprise. In any case, once documents are reduced to a set of terms with associated information, appropriate **semantic similarity** techniques (detailed in Section IV) are exploited to easily identify relevant documents w.r.t. to a given query document, and to produce a list of suggestions ranked on the similarity (relevance) score.

III. TEXT ANALYSIS TECHNIQUES

A. Text Analysis and Keyword Extraction

The goal for text analysis and keyword extraction in the FACIT project is to design and develop an effective and easy-to-apply technique for automatically extracting terms (and their associated semantic information and statistics) from the submitted text documents. In particular, we wanted to devise a flexible technique to be exploited both for "off-line" analysis, thus working on the textual descriptions already available in the ORM, and for "on-line" querying operations, i.e. applied on the fly to the submitted query documents. Even if many packages are available for keyword extraction purposes, most of them do not allow sufficient configuration and extension options, making their integration with the future FACIT solution very complex. Therefore, we preferred to design a custom-made technique tailored to the FACIT environment. Here is a short summary of the steps performed, for each text document, in the text analysis phase:

1. **Tokenization**: the text is "tokenized" (words are identified, punctuation is removed);
2. **Stemming**: the tokens are "normalized" and "stemmed", i.e. terms are reduced to their base form (managing plurals, inflections, ...) (as we will see this will be very useful to enhance the effectiveness of the similarity computation phase);
3. **POS (Part of Speech) Tagging**: the tokens are "tagged" with Part of Speech tags (i.e. nouns, verbs, ...);
4. **Composite terms identification**: possible composite terms (such as "product action plan" or "product requirement") are identified by means of a simple state machine and of POS tags information;
5. **Filtering and enrichment:** by exploiting external knowledge sources, the most relevant terms are selected and they are associated to additional information (such as definitions, synonyms, …). More specifically we make use of the IEEE Software and Systems Engineering Vocabulary[1], a knowledge source covering specialist terms in the project area, and the WordNet[2] English thesaurus, possibly complementing the specialist source with general knowledge about English concepts;
6. **Term statistics and weights computation**: weights are computed for each of the terms, reflecting their

[1] http://www.computer.org/sevocab
[2] http://wordnet.princeton.edu/

TERM	WN	IEEE	SYNS	DEFS	IDF	DOC_LIST
acquirer	Y	Y	buyer, customer, owner, purchaser	(1) stakeholder that acquires or procures a product or service from a …	7,4961	['QM1372']
acquisition	Y	Y	outsourcing	(1) process of obtaining a system, software product or software service …	5,5491	['QM0392', 'QM0755', …]

Figure 1. An excerpt of the FACIT-SME semantic glossary (global view)

importance and meaningfulness in the text. As we will see, this information is fundamental in computing accurate text similarities (more on this in the following sections).

By applying batch text analysis to the documents currently available in the ORM, we achieved a first significant result in the FACIT-SME project, i.e. the automatic generation of a **semantic glossary**, representing a first step toward the sharing of the most important concepts available in the model and the automatic computation of text similarities. Thanks to the automatic text analysis procedure, this first draft can be easily updated/enriched in case of new content added to the ORM, while more fine-grained user interventions for adding/modifying/eliminating information are also possible. The following section describes the semantic glossary structure more in detail.

B. FACIT-SME Semantic Glossary

The Semantic Glossary consists of a **global view** (all terms in all documents, together with their statistics) and a **per-document view** (terms occurrences in each of the documents with their statistics). The **glossary global view** is an alphabetical sort of all the extracted terms, in a tabular form. Figure 1 shows an excerpt of the glossary global view. The format is:

TERM - the extracted term;
WN – whether the term is present in WordNet thesaurus;
IEEE – whether the term is present in the IEEE vocabulary;
SYNS – possible synonyms for the term (as extracted from the IEEE vocabulary and/or WordNet);
DEFS – possible definitions for the term (as extracted from the IEEE vocabulary and/or WordNet);
IDF – the inverse document frequency of the term in the collection;
DOC_LIST – a list of the documents IDs in which the term occurs.

The **glossary per-document view** is a list of all the term occurrences in the documents, sorted on the document ID, together with their statistics. Figure 2 shows an exercpt of the glossary per-document view. For each term (in each document) the report contains:

DOC – the document ID in which the term occurs;
TERM - the extracted term;
TF - the frequency of this term in the document, normalized by total number of terms in document;
WEIGHT - the TF*IDF weight of the term.

DOC	TERM	TF	WEIGHT(TF*IDF)
QM0001	iso	1	6,8024
QM0002	management	0,3333	0,6931
QM0002	quality	0,3333	1,0303

Figure 2. An excerpt of the FACIT-SME semantic glossary (per-doc view)

The glossary includes both synonyms/definitions and weight information, allowing, as we will see, the similarity functions of the Semantic Helper to draw useful knowledge from both the semantic and the classic text retrieval worlds. As to the semantic information, note that even if all the similarity techniques described in the next section are designed to work (and, as proved in Section V, provide encouraging results) without further intervention, the list of synonyms and definitions retrieved from the external knowledge sources could easily be refined manually by experts or automatically by means of techniques such as sense disambiguation [9, 10]. In addition, as in classic Information Retrieval, the importance (**weight**) of each keyword in each document is estimated. Beside term frequency (**TF**), we compute the inverse document frequency (**IDF**)[3] [11], which provides an estimate of the meaningfulness of each term. The weight is then computed as TF*IDF. In this way, very common terms which are present in a large number of documents have a lower weight and will give a lower contribution to the final similarity, since they are probably less meaningful.

IV. SEMANTIC SIMILARITY TECHNIQUES

As anticipated in the past sections, the need of effectively and efficiently computing similarities between documents is crucial to the project. To this end, we want to define a document similarity formula $DSim(D^x,D^y)$ which, given a source document D^x and a target document D^y, quantifies the similarity of the source document with respect to the target document. Being documents represented by sets of terms, semantic similarity computation becomes a matter of computing similarities between sets of terms. Therefore, document similarity will, in turn, use a combination of the scores provided by a term similarity formula *TSim* between the document terms. The computation of *DSim* between a given D^x and all the possible submitted D^y induces a **ranking** of the available documents with respect to the source one, thus predicting which documents are relevant and which are not with respect to D^x. For instance, by computing the document

[3] IDF is obtained by dividing the total number of documents by the number of documents containing the term and then by computing the logarithm of that ratio

similarities between a given quality requirement description and all the available software methodology descriptions of the ORM, the induced ranking will suggest all the software methodologies that could be relevant/related to the given quality requirement.

We provide different options with respect to the similarity formulas, so to be able to experimentally assess the ones most suited to the project. Equation (1) shows the standard document similarity formula between a given source document D^x and a target document D^y: the similarity is given by the sum of all term similarities between each term in D^x and the term (defined in (2)) in D^y maximizing the term similarity with the term in D^x:

$$DSim(D^x, D^y) = \sum_{t_i^x \in D^x} TSim\left(t_i^x, t_{j(i)}^y\right) \quad (1)$$

$$t_{j(i)}^y = \ argmax_{t_j^y \in D^y} \left(TSim\left(t_i^x, t_j^y\right)\right) \quad (2)$$

Note that (1) is not meant to be symmetric, instead it is conceived so to facilitate the ranking of the documents D^y with respect to document D^x. In case symmetry is needed, the summation in (1) can by extended to the terms of both documents. As to term similarity, the most basic option, only considering equal terms, is shown in (3), where the similarity *TSim* between two terms t_i and t_j is basically a Boolean formula valued 1 if the two terms are equal, 0 otherwise.

$$TSim(t_i, t_j) = \begin{cases} 1 & t_i = t_j \\ 0 & otherwise \end{cases} \quad (3)$$

Equation (4) proposes an extended (and possibly more effective) document similarity option, taking into account the weights as extracted by the text analysis process:

$$DSim(D^x, D^y) = \sum_{t_i^x \in D^x} TSim\left(t_i^x, t_{j(i)}^y\right) \cdot w_i^x \cdot w_{j(i)}^y \quad (4)$$

where $w_i^x = tf_i^x \cdot idf_i$; $w_{j(i)}^y = tf_{j(i)}^y \cdot idf_{j(i)}$. In this case, each term contributes to the final similarity with a different weight *w*, i.e. more frequent and more significant terms contribute more to the similarity between the two documents.

Let us now consider more advanced options for term similarity. Equation (5), besides equal terms, also takes **synonyms** and **semantically related terms** into account:

$$TSim(t_i, t_j) = \begin{cases} 1 & t_i = t_j \ \ or \ \ t_i \ SYN \ t_j \\ r & t_i \ REL \ t_j \\ 0 & otherwise \end{cases} \quad (5)$$

More specifically, the case of maximum similarity (value 1) is extended to the case where the two terms are synonyms (*SYN* relation). Moreover, the formula provides a further case where the two terms are not equal or synonyms, nonetheless they are in some way strongly related from a semantic point of view: such terms will contribute with a similarity of *r*, where $0<r<1$.

While the synonym information straightly comes from the text analysis phase, we consider two different ways of exploiting the extracted information and the external knowledge sources to determine wheter two given terms are semantically related. Equation (7) shows a possible way of computing the similarity by exploiting the glosses (definitions) of the terms:

$$t_i \text{ REL } t_j \iff \text{GSim}\left(\text{gl}(t_i), \text{gl}(t_j)\right) \geq \text{Th} \quad (7)$$

$$\text{GSim}\left(\text{gl}(t_i), \text{gl}(t_j)\right) = \sum \left|\text{ovl}\left(\text{gl}(t_i), \text{gl}(t_j)\right)\right|^2 \quad (8)$$

In this case, two terms are semantically related (*REL* relation) if the gloss similarity *GSim* between their glosses exceeds a given threshold *Th*. The Literature presents many possible ways of computing similarities between glosses. We found the extended gloss overlap measure [12] shown in (8) to be particularly effective in our context, especially with the IEEE vocabulary glosses. It quantifies the similarity between the two glosses by finding overlaps in them (the similarity is the sum of the squares of the overlap lengths). Other gloss similarity measures could also be exploited and investigated in the future, such as the gloss vector one described in [13].

Another possible way of computing semantic relatedness is to exploit the relations between terms coming from the WordNet thesaurus. Indeed, we adopt one of the most widely used methods in knowledge management, relying on the hypernymy relations of the thesaurus:

$$t_i \text{ REL } t_j \iff \text{HSim}\left(t_i, t_j\right) \geq \text{Th} \quad (9)$$

$$\text{HSim}\left(t_i, t_j\right) = \begin{cases} -\ln \frac{\text{len}(t_i, t_j)}{2H}, & \exists \ \text{lca}(t_i, t_j) \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

In this case, two terms are semantically related if their hypernym similarity *HSim* exceeds a given threshold *Th*. In particular, the *HSim* shown in (10) derives from the works [9, 14] and computes a score which is inverserly proportional to the length of the shortest path connecting the (senses of the) two terms. H is a constant, which for WordNet is defined as 16. On the other hand, the similarity is 0 if the two terms are not connected in the WordNet hypernymy structures.

Note that the different document similarity and term similarity formulas presented in this section can be selected in a fully orthogonal way, so to be able to adapt in a flexible way to the specific settings and needs of the project.

V. Experimental Evaluation

In this section we present the results of the preliminary effectiveness evaluation we performed on the proposed techniques in the context of the project. We formed a collection of 1500 documents (i.e. textual descriptions) about quality requirements which will be part of the final ORM and derive from existing quality models such as CMMI [15] and ISO 9000 [16]. Then, we created different queries with reference to this collection; each query is either composed by a short text containing candidate keywords, so to simulate possible querying situations following the "From Scratch" scenario of the project (queries Q1-Q6, selected as the most representative ones), or by a whole new document ideally representing the description of an existing enterprise requirement or methodology, similarly to the "From

Query	Results			No sem syn/rel (baselines)			No kw sel		
	Prec	Rec	F	Prec	Rec	F	Prec	Rec	F
Q1	1,000	1,000	1,000	1,000	1,000	1,000	0,011	0,420	0,022
Q2	1,000	1,000	1,000	1,000	1,000	1,000	0,005	0,330	0,010
Q3	1,000	1,000	1,000	1,000	0,969	0,984	0,946	0,240	0,383
Q4	0,947	1,000	0,973	1,000	0,079	0,146	0,921	0,321	0,476
Q5	0,878	1,000	0,935	1,000	0,077	0,143	0,986	0,235	0,380
Q6	0,923	0,949	0,936	1,000	0,333	0,500	0,967	0,369	0,534

Figure 3. Effectiveness analysis in terms of precision, recall and F-measure (standard results on the left, two baselines on the right)

Methodology" scenario (queries QT1-QT4). Each query will be submitted to the current implementation of the semantic helper (text analysis and similarity computation), so to generate a set of possible "suggestions", i.e. pointers to the relevant documents in the collection. In order to evaluate the effectiveness of our approach, for each query the output of the helper will be compared to a "gold standard", i.e. the relevant answers which were manually selected from the collection by experts in the field.

The first analysis we conducted was to assess the quality of the retrieved results in terms of precision and recall, which are typical evaluation metrics in the information retrieval field[4]. Figure 3 shows the results for Q1-Q6 (left part of figure). The shown results are those obtained with the gloss similarity as the similarity relatedness function (later we will discuss the WordNet hypernym-based one). Besides precision and recall, we also report, as a summarizing figure, their weighted harmonic mean (F-measure). Further, in order to emphasize the contribution of the different applied techniques to the achieved results, in the right part of figure we also present the results concerning two baselines, i.e. a standard retrieval method ignoring semantic synonyms and related terms information and another method not exploiting the text analysis phase (including stemming and composite terms identification). Let us now analyze the results in detail.

As we can see from Figure 3, the precision and recall levels achieved by the described techniques are very satisfying for all queries (equal or higher to 0.84 and 0.94, respectively). The processing of all queries greatly benefits from the text analysis phase: as the results of the second baseline show, without it recall levels significantly drop to 0.2-0.4 (for instance, different inflections of the same word are not correctly identified). Text analysis can also greatly benefit precision as, for instance, in Q1 and Q2: since they contain, among others, such composite expressions as "interface requirement" (Q1) or "configuration management system" (Q2), not correctly identifying them leads to a very large number of irrelevant retrieved documents (in the second baseline precision drops to less than 0.01, compared to 1 for the standard results). Differently from Q1 and Q2, queries Q3-Q6 also require synonyms and related terms management in order to provide satisfying answers: for instance, one of the key terms in Q3 is "supplier", a concept which is expressed as "vendor" in some of the documents (recall goes from 1 to 0.96 of the first baseline), while Q4 contains "purpose" which is

[4] Precision is defined as the fraction of retrieved documents which are known to be relevant, recall is the fraction of known relevant objects which were actually retrieved.

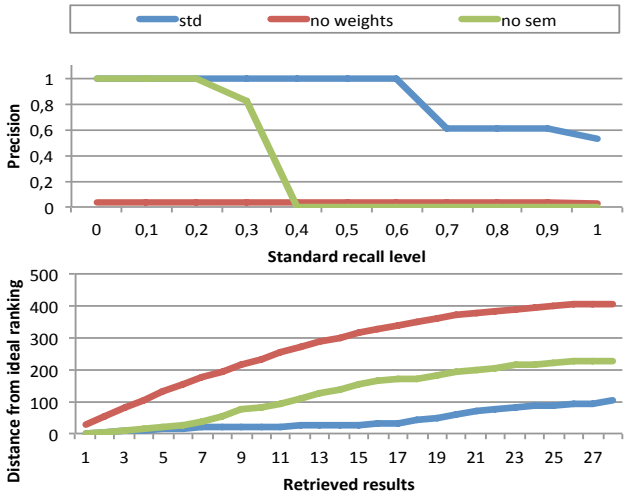

Figure 4. In-depth effectiveness analysis for query QT1: precision at standard recall levels (top) and distance from optimal ranking (bottom)

mostly expressed as "objective" in the collection (recall drops from 1 to less than 0.08). The same holds for the "related terms": by applying the gloss similarity semantic relatedness formulas exploiting the IEEE definitions, we achieve near-perfect recall levels (as opposed to the less than optimal ones of the first baseline) while also maintaining high precision levels. For example, most documents containing "review" are also relevant to Q5, which contains "audit", the ones containing "document" are also relevant to Q6 asking for "documentation", and so on. The WordNet based similarity proved equally useful as the gloss based one (for instance it correctly identifies the relatedness between "attribute" and "property", "procedure" and "process", and others), however we found that in some cases it may lead to several false positives, mainly due to the non-specialized nature of the employed thesaurus. For this reason, we decided to focus on the IEEE gloss similarity in these preliminary tests and to analyze the impact of the WordNet similarity more in detail in future tests.

We will now deepen the effectiveness analysis by considering queries QT1-QT4, in the form of actual text documents for which to find related documents in the collection. In this case, the queries are substantially more complex than Q1-Q6 and can possibly produce a very large number of results: for this reason, it is essential to evaluate not only which answers are returned but also their score and the induced ranking, so to assess whether the best suggestions are returned in the top positions and, thus, whether the proposed weighting scheme is effective. Figure 4 (top) shows, for QT1, the precision values obtained at different recall levels, i.e. when a given percentage of relevant documents have been found. The "standard" technique, which uses all the weights and semantic synonymy/relatedness information, is compared to non-weighted and non-semantic baselines. Notice that the standard technique achieves very high precision levels even at high recall levels: for instance, at recall level 0.6, the precision is still 1, while the baselines' precision levels have already dropped lower than 0.03. This confirms that our techniques are able to identify the most significant terms in the queries, without being misled by non-relevant ones. Figure 4 (bottom)

confirms the goodness of the retrieved results: for each alternative, the curve represents the normalized Spearman footrule distance [17] between the retrieved and the ideal ranking, i.e. the normalized sum of the absolute values of the difference between the ranks. Due to lack of space we do not show this detailed analysis for QT2-QT4, however we found that the good performance of QT1 is fully representative of all the queries. In particular, for QT1-QT4 the most relevant results are always among the first to be retrieved, and the precision at recall level 1 is always higher than 0.5 (orders of magnitude better than our baselines).

VI. Concluding Remarks

Several literature papers have highlighted possible benefits of combined knowledge engineering and software engineering approaches for specific SE tasks [2, 3, 4]. For instance, while standard reuse repositories are limited to plain syntactical search and generally suffer from low precision and recall [4], knowledge-based approaches such as [1] enhance the effectiveness of the component reuse task by proposing the usage of formal descriptions of components (in OWL) to be queried by specific graph query languages such as SPARQL. Other notable proposals have been presented, for instance, for facilitating software process assessment through formal descriptions of specific process improvement approaches such as CMMI [2, 3]. As already noted in [4], however, the discussion on integrating SE and KE approaches has been, in many cases, very academic, focusing on aspects like meta-modeling and neglecting applicability and usability.

The filtering/search approach we presented in this paper is designed to be effective and very easily applicable. In the FACIT-SME scenario, software-developing SMEs will be able to exploit all the advanced functionalities offered by these semantic foundations for a large number of tasks, such as software process and improvement and certification. Moreover, the approach does not have any prerequisite, such as the knowledge of complex formal representation/querying standards or the need of converting/updating the documentation already available in the enterprise. To this end, our proposal leverages on the strengths of both classic information retrieval (incorporating weight information [11]) and of knowledge-based techniques. In particular, semantic similarity techniques which already proved their effectiveness in a number of non-SE scenarios, from information disambiguation [9] to the querying of heterogeneous information in digital libraries and PDMSs [18], are adapted and extended in this new framework.

FACIT-SME has just started and the presented approach is the first step toward the project goals. Future work will include further analysis and refinements of the similarity techniques (especially for the WordNet-based ones), user feedback on the retrieved suggestions, Multilanguage information management and querying support, and the exploitation of other non-textual knowledge which will eventually be available in the ORM repository. Finally, the project evaluation phase, which will start soon, will involve user IT companies in actual scenarios in order to obtain useful opinions and suggestions about the quality of the proposed techniques and their improvement.

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme managed by REA Research Executive Agency (http://ec.europa.eu/research/rea) ([FP7/2007-2013] [FP7/2007 - 2011]) under grant agreement n° 243695.

Our sincere thanks to Sonia Bergamaschi, Domenico Beneventano (UniMoRe), Gorka Benguria (ESI), Frank-Walter Jaekel (Fraunhofer IPK) and to the other project partners for their support to this research.

References

[1] C. Kiefer, A. Bernstein, J. Tappolet. Mining software repositories with ISPARQL and a software evolution ontology. In Int'l Workshop on Mining Software Repositories (MSR), 2007.

[2] K. Soydan, "An OWL Ontology for Representing the CMMI-SW Model". Proc. of 2nd Int'l Workshop on Semantic Web Enabled Software Engineering (SWESE), 2006.

[3] L. Liao, Y. Qu, H. K. N. Leung, "A software process ontology and its application". Proc. of the 4th Int'l Semantic Web Conference, 2005.

[4] H.-J. Happel and S. Seedorf, "Applications of Ontologies in Software Engineering". Proc. of 2nd Int'l Workshop on Semantic Web Enabled Software Engineering (SWESE), 2006.

[5] Aetic (Spain), Agoria (Belgium), AssInform (Italy) et al., "Position paper towards a European software strategy", presented to commissioner Viviane Reding on 24 October 2008.

[6] DG INFSO Internal Reflection Group on Software Technologies, "ITEA", April 2002.

[7] F.-W. Jaekel (Editor), "ORM Architecture and Engineering Models", FP7-SME FACIT-SME (FP7-243695), Deliverable, http://www.facit-sme.eu/FACIT-2-2010-10-18-IPK-deliverable_2_1-37b.pdf, Oct 2010.

[8] G. Benguria (Editor), "OSES Architecture and Component Specification", FP7-SME FACIT-SME (FP7-243695), Delivarable, Dec 2010.

[9] F. Mandreoli, R. Martoglia, "Knowledge-Based Sense Disambiguation (Almost) For All Structures". In Information Systems 36(2), 2011.

[10] L. Po, S. Sorrentino, "Automatic generation of probabilistic relationships for improving schema matching". In Information Systems, 36(2), 2011.

[11] G. Salton, C. Buckley, "Term-weighting approaches in automatic text retrieval". In Information Processing & Management 24(5), 1988.

[12] S. Banerjee and T. Pedersen, "Extended Gloss Overlaps as a Measure of Semantic". Proc. of the Eighteenth Int'l Joint Conference on Artificial Intelligence, pp. 805-810, 2003.

[13] S. Patwardhan, T. Pedersen, "Using WordNet Based Context Vectors to Estimate the Semantic Relatedness of Concepts". Proc. of the EACL 2006 Workshop Making Sense of Sense, pp. 1-8, 2006.

[14] C. Leacock and M. Chodorow, "Combining local context and WordNet similarity for word sense identification". In C. Fellbaum, editor, WordNet: An electronic lexical database. MIT Press, 1998.

[15] Carnegie Mellon University Software Engineering Institute, "CMMI for Development, Version 1.2" (pdf), 2006.

[16] ISO TC176, "DIS 9001:2000 Quality Management Systems - Requirement." (pdf), 1999.

[17] P. Diaconis, R. L. Graham, "Spearman's Footrule as a Measure of Disarray", Journal of the Royal Statistical Society 39(2), 262-268, 1977.

[18] F. Mandreoli, W. Penzo, S. Sassatelli, S. Lodi, R. Martoglia, "Semantic Peer, Here are the Neighbors You Want!". Proc. of the 11th Int'l Conf. on Extending Database Technology (EDBT), pp.26-37, 2008.