"A Semantic Information Retrieval Framework within the scope of IPS²-PLM"  

"Michael Abramovici, Philip Gebus*, Jens Christian Göbel, Hoang Bao Dang"

*Corresponding author. Tel.: +49 (0)234/32-28492; fax: +49 (0)234 32-14443. E-mail address: philip.gebus@itm.rub.de

Abstract

The Product Lifecycle Management (PLM) approach faces new challenges if transferred for Industrial Product Service Systems (IPS²). The vast amount of heterogeneous data generated throughout an IPS²’s lifecycle complicates the retrieval of required information for IPS² actors. However, these actors’ risky decisions determine an IPS²’s success during use phase. Thus, an approach is needed that supports actors in finding targeted information. This paper presents a framework that utilizes semantic and text mining techniques in order to improve the information retrieval process in IPS²-PLM and to allow IPS² actors to focus on their value-adding tasks rather than spending a lot of time for finding information.

Keywords: IPS², Semantic Information Retrieval, IPS²-PLM, Semantic Search

1. Introduction

The lifecycles of Industrial Product Service Systems (IPS²) are shaped by a close interaction between providers and customers as well as interconnected product and service structures with strong dynamics of change during the use phase. New business models obligate providers to guarantee their IPS²'s contracted utility through a network consisting of the provider, different customers and service-partners, such as third-party service providers or suppliers [1]. Due to the vast amount and heterogeneity of the information available, it is difficult to realize a transparent, efficient and consistent information flow throughout the IPS² lifecycle in order to meet contractual obligations.

Because of the previously mentioned strong dynamic of change, especially during the use phase, a number of risky decisions made by different actors determine an IPS²’s success. Such decisions imply far-reaching consequences and therefore require an effective and fast information retrieval. As an example, an IPS² provider schedules major maintenance intervals every few weeks. If a maintenance worker locally identifies irregularities, he requires further inquiry in documents or comments, which might imply information necessary in order to decide whether unscheduled (in addition to the regular intervals), proactive maintenance is required. This valuable information is mainly hidden in various unstructured text documents, such as protocols, (maintenance) reports or comments [2]. Therefore, these documents need to be available for all involved parties in the IPS² partner network.

Product Lifecycle Management (PLM) approaches support the integration and management of product data and documents as well as engineering processes and applications [3]. Unfortunately currently available PLM solutions mainly support the single management of product, service or software-related information but not the combination of the above-mentioned [1]. Therefore, in our previous work we introduced an ontology-based PLM concept, which is suitable for IPS². Firstly because it enables a consistent information flow through the IPS²-specific heterogeneous partner networks and secondly supports the integrated management of product, service and software-related data along the whole lifecycle [4].

Commercial PLM solutions only implement traditional search engines that allow a keyword-based Boolean search in a document’s metadata or at most full text search. Due to the diversity of IPS² partners, documents in IPS² provider networks differ in terms of structure and terminology used, which render a targeted information retrieval as very time-consuming or even impossible with current state of the art solutions. Thus,
PLM solutions deliver too many or irrelevant results on queries [2]. The vast amount of documents that are created along the whole IPS²’s lifecycle and the limited search possibilities (keyword-based) compound the aforementioned problem.

Therefore, a semantic approach is necessary that allows an intuitive, targeted and fast information retrieval within a vast amount of heterogeneous data in order to enable efficient knowledge management in IPS² partner networks.

2. Basic Requirements

Considering IPS²-specific lifecycle and data aspects in order to support actors in finding necessary information, the following groups of requirements have been identified:

A dynamic and semantically rich data model: The framework has to provide a data model that links document contents and metadata with their IPS²-PLM-specific contexts on meta data level as well as on instance level, such as concrete IPS²s, goods, services and business models. It must support heterogeneous documents due to the diverse nature of IPS² partner networks. Furthermore, it has to model the semantic relationships between diverse unstructured documents and has to be dynamically expandable when new documents and instances are added.

Exploitation of semantic data: The framework has to allow actors to exploit the data model intuitively, meaning in an easy to learn and quick manner. It must enable information retrieval within the semantic (IPS² context-specific) structure as well as the content of the data provided for actors with different knowledge bases and languages.

Facilitation of targeted and outright retrieval: The framework must provide outright and targeted information retrieval results. These results have to be weighed according to a user’s information needs, which he formulates within a query. Only a suitable ratio between search result precision and recall reduces the time necessary for searching information.

3. State of the Art

Traditional information retrieval mainly bases on the frequency of words in documents (text mining), whereas semantic search can be defined as a search process that exploits domain knowledge and uses formal semantics in any of the three stages of a search process: query, search process or information presentation [5,6]. Utilizing domain knowledge and semantics helps in providing targeted results to a user’s specific query.

The authors in [7], for example, integrate a semantic relationship manager into the PLM approach as a framework between different coexisting software systems. They focus on semi-structured data in the beginning of life phase of products and therefore, do not consider IPS²-specific data in the use-phase or unstructured data. In [8], the authors developed the infrastructure for a semantic search engine but did not yet implement the engine itself and do not consider IPS²-specific relationships. The authors in [9] present an ontology-based context model that exploits users’ context information, such as e.g. his current task, role, location or status in order to recommend and reason possibly required information. They rather present a recommendation approach than a search engine based on an ontology that was implemented in protégé and did not regard the aforementioned specifics of integrated products and services. Due to the lack of IPS² consideration in the related work, a new information retrieval approach for IPS²-PLM had to be designed.

State of the art information retrieval approaches differ on the one hand in the documents considered and on the other hand in the methods of applying semantic technologies in order to exploit domain knowledge. Current semantic technologies in the search process can be divided into the following four basic techniques: graph traversal, query expansion, spread activation and RDFS/OWL reasoning [6].

In the following, methods and scientific approaches belonging to research areas (domain knowledge representation, text mining and semantic technologies during the retrieval phase) that are necessary to develop a semantic information retrieval for IPS²-PLM are presented. Section 4 describes the precise field of application and the relations between the presented techniques.

3.1 Representing Domain Knowledge

Several methods exist to represent domain or enterprise-specific information, which differ in the way of structuring information. Glossaries, for example, are a simple collection of terms. Taxonomies classify terms by clustering them in categories. Thesaurus are also used in order to cluster terms but further express the relationships between words in form of synonyms, hypernyms or hyponyms [10]. The framework can use a thesaurus in order to overcome the difficulty of extracting information from heterogeneous documents with different terminologies.

Map representations focus on linking topics to each other and provide an overview of similar and related topics [11].

Semantic networks describe relations between concepts and terms as a graph that consists of vertices (concepts and terms) and edges (relationships) that link vertices to one another [12].

Ontologies are based on the semantic network concept. Gruber defines ontologies as “a specification of a conceptualization” [13]. Compared to semantic networks, ontologies are more strict and formal due to their axiomatic structure [13]. As the framework requires a dynamically changing data model, which does not follow strict axiomatic rules, semantic networks are preferred over ontologies for representing the relations of documents.

3.2 Text Mining

As stated before, valuable information is mainly hidden in various unstructured text documents. Documents generated during the use phase are mainly of this type. Text mining describes the process of analyzing unstructured text documents. Documents are transformed into their vector representation, where each term is stemmed and represented by its’ normalized term frequency multiplied by its’ inverse document frequency (measure for a term’s importance) [14]. The vector space representation can be exploited to determine a document’s relevancy to a given query and the similarity between two documents by calculating the cosine similarity
between two vector representations [14]. The described text mining methods are necessarily used within the framework to analyze document contents and to extract information from documents.

3.3 Semantic technologies in the retrieval phase

Graph traversal describes a process that uses search algorithms to traverse graphs, which represent domain knowledge. Among others, these algorithms exploit graph edge weights for determining a document’s relevance to a specific domain or query. Shortest path algorithms for example help to quantify relationships between two nodes in the graph [15]. Most relevant research approaches are “MinG” [16] or “SRelation” [17]. Graph traversal is a useful tool set to extract correlations between entities in the semantic network.

Query expansion is defined as a technique to manipulate and extend user queries [15]. By Exploiting a thesaurus, user queries can be extended by hypernyms, synonyms or hyponyms and therefore increase the recall or precision dependent on the implementation method. Other techniques are stemming and spelling correction. A sample research approach is [18], where query keywords are extended by exploiting Linked Open Data. In the context of the semantic framework, query expansion can help to satisfy the requirements for mining heterogeneous documents by fixing spelling mistakes in user queries and by adding synonyms based on a (IPS² partner network-specific) thesaurus as well as translations of the keywords.

Spread activation extends the graph traversal approach. Algorithms do not only consider weights on the edges but also the amount of incoming relationships [15] in order to find correlations between documents and therefore, increase the amount of query results. In [19] for example, the authors propose an ontology-based vector space model that exploits spread activation in order to extend user queries. Within the semantic framework, spread activation helped to find similarities between documents and hypernyms of terms used within a query.

RDFS/OWL Reasoning contains a set of algorithms that helps to conclude relations in the ontology languages RDFS and OWL [15]. On the one hand, logical correctness of queries on RDFS or OWL documents can be verified and on the other hand, implicit statements can be included to the result set by using reasoning. Due to their formal strictness, ontologies are the foundation for using reasoning. Nevertheless, reasoning is not a prioritized requirement as we assume that the user does not use contradictory semantics in the search query. Therefore, it is not used within the framework.

4. Semantic Information Retrieval Framework for IPS²-PLM

The semantic information retrieval framework for IPS²-PLM is set to support IPS² actors to find required information along the whole lifecycle. It focuses on unstructured text documents because of the vast amount of text documents generated during the use phase and their importance for the IPS² actors’ decisions as stated above.

The framework consists of three major components. First, the IPS² specific data model has been defined and extended. Afterwards text mining methods have been used to extract information from documents in order to integrate them into the data model. Finally, a methodical approach has been developed to process user search queries within the scope of IPS²-PLM semantically and to exploit the semantic information from the IPS² domain, which is described in the data model. Fig. 1 shows an overview of the framework’s relevant components and their coherence.

Fig. 1: Semantic IPS² Information Retrieval Framework Components

In the following, the components and their interfaces are presented in detail.

4.1 Extended Data Model

In our previous work a (top-domain) ontology-based PLM concept has been presented, which is suitable for IPS². The top domain ontology enables the management of IPS²-specific information throughout the whole lifecycle across different companies. It is based on our meta data reference model [20] and serves as an integrated information platform that unites different company specific data models. Nevertheless, this ontology does not consider information from any kind of documents, which makes it difficult to retrieve targeted information. Hence, the model had to be extended in order to provide efficient information retrieval in an IPS² environment with large document sets.

In a first step, a semantic network on entity level has been designed. Entities represent documents in the graph and are connected to several instances of concepts from the IPS² top domain ontology, such as related products and services. Due to the nature of PLM, being a hierarchically structured, user-centered and process-driven approach, this background information (meta data) can be exploited to create documents’ semantic relations. Upon checking in a document into a PLM platform, the aforementioned information can be extracted and used to create the semantic network.
The framework creates a document node in the semantic network and stores important attributes to it, e.g. name, date, release status, file type and file path. It connects the user entity from the ontology with the document through a "::AUTHOR_OF" relationship. At the same time, it links the new document to the document container, which can be a specific version of a specification for example. A document container groups several documents of the same kind or topic. This document container is linked to a service or good instance, while the service or good itself is connected via a "::PART_OF" relationship to the IPS² and it is bound to its’ current lifecycle phase. Of course, a good or service can be part of another good or be a subservice. Dependent on the linked service or good entity’s current lifecycle phase, the document gets connected to this phase by a "::BELONGS_TO" relationship accordingly.

Thus, the framework creates a rich semantic network for each document. Fig. 2 displays a snippet of the semantic network on IPS² entity level.

![Fig. 2: Snippet of the Semantic Network on IPS² Entity Level](image)

Being able to find documents that are connected to specific ontology entities did not yet meet the aforementioned requirements (see Section 2). The amount of time needed to check all returned documents’ contents for specific information was still too large after being able to reduce the result set by queries like for instance “All documents belonging to the spindle of Omichron IPSS”. Therefore, in a second step, the data model has been extended in order to also process document content, which makes it possible to search for e.g. “Documents of Omichron IPSS about overheatting”.

Using text mining techniques, documents are analyzed (see Section 4.2) and linked to their terms in the data model via "::CONTAINS" relationships as well as to similar documents via "::SIMILAR_TO" relationships. A term’s normalized frequency is stored as an attribute in the relationship between the term and the document, while the cosine similarity value between two documents is stored as an attribute in the "::SIMILAR_TO" relationship. This enables fast term weighing computations in the retrieval phase as well as additional query possibilities. Fig. 3 presents the data model extension on document level.

Analyzing unstructured text documents with text mining techniques is the prerequisite for the extension of the semantic network on document level. This process is described in the following section.

![Fig. 3: Snippet of the Semantic Network on IPS² Document Level](image)

4.2 Information Extraction (Text mining)

The framework uses text mining techniques in order to prepare documents to be stored in the semantic network. Upon check-in of a document into the IPS²-PLM system, algorithms analyze the document’s content.

In a first step the document is transformed into its’ vector representation where each term that occurs in a document has a non-zero value in the vector. The frameworks uses a stemming algorithm in order to group similar words in their base form (here, referred to as terms) and calculates every term’s tf-idf (term frequency – inverse document frequency) value. A term’s normalized frequency represents the value how often a term occurs in a document divided by the document’s length. This value is constant and therefore can be stored as an attribute of the "::CONTAINS" relationship as mentioned above. The idf value is a measure of a term’s information importance. Both values multiplied represent a term in the vector. This method has several advantages. It enables fast tf-idf calculations in the retrieval phase and weights terms considering their importance to a document in a corpus (collection of documents).

Adding text mining to the semantic information retrieval framework for IPS² enables users to query the semantic network for a document’s content, such as “All documents belonging to the spindle of Omichron IPSS about overheating”.

4.3 User Queries Processing

In order to satisfy the requirements, an easy to learn and user-friendly method to query the semantically rich data model (semantic network) was the only way to increase a user’s efficiency. Complex query languages, such as SPARQL (SPARQL Protocol And RDF Query Language) for example require a time consuming induction and a complex query formulation. Thus, they do not meet the aforementioned requirements.

In order to overcome these difficulties, a natural language query interface for the IPS²-PLM context has been designed due to its’ natural and therefore, easy to use characteristics. Fig. 4 shows the framework’s three-phased query processing concept.

In phase 1 of the query processing, the framework analyzes and prepares the query for parsing. Given the example natural language query “I’m looking for all documents belonging to...”
Omichron IPSS", the engine maps the query terms to IPS²-PLM-specific entities defined in the data model. In this particular example, it maps the term “documents” to the entity “Document”, which can be any kind of document and the terms “Omichron IPSS” to the entity “IPSS-id:123”, which defines the terms as being the object with id:123 of the IPS² entity in the data model.

At the same time, the framework analyzes the query semantically. It uses an algorithm based on a 2- and 3-gram analysis to find phrases that can be mapped to relationships in the data model. In the aforementioned example the phrase “belonging to Omichron IPSS” is interpreted by the engine as “PART_OF IPSS-id:123” (which represents the equivalent relation in the semantic network) and mapped accordingly.

Unnecessary terms and phrases, such as “I’m looking for” and “all” are identified by a stop-words list and removed from the query. After this process, a transformed version of the query is forwarded to the next phase: “(type:documents) of (IPSS-id:123)”. In a next step, the prepared query is semantically parsed in order to check whether the query is part of the language interpreted by the framework. Therefore, a context-free grammar has been developed. The semantic parser uses this grammar to perform a leftmost derivation in order to substitute each query term. While doing so, it constructs the semantic dependencies of the query terms. If the query is parseable and thus belongs to the interpretable language, the framework uses all interpreted information to query the data model. Regarding the example from above, the result set would contain all documents linked to Omichron IPSS. If an IPS²-PLM user enters a query that is not part of the language, e.g. “overheating errors of Omichron IPSS”, the parser computes the leftmost derivation and returns that the query is not parseable. In this case, the framework crops all semantic and unnecessary information and executes a semantically enriched full-text query with the terms remaining. More precisely: the example presented before would generate a query for all text documents, which are linked to Omichron IPSS that contain the keywords “overheating” or “error”. Additionally, an (IPS² partner network-specific) thesaurus identifies and extends the user query with synonyms and equivalent terms in different languages in order to overcome heterogeneous document formulation difficulties. The result documents are finally sorted by their cosine similarity between the query terms “overheating, error” (as well as their synonyms defined in a thesaurus) and the returned documents.

5. Implementation

The presented conceptional framework has been validated and implemented as an extension of our ontology-based IPS²-PLM prototype. The semantic data model is represented in a Neo4J® graph database (see Fig. 5) due to the fact that Neo4J implements fast traversal algorithms and supports the semantic features necessary to implement the data model.

The semantic information retrieval framework runs in a JAVA environment and communicates with our IPS²-PLM platform via web services that exchange JSON values. The framework’s underlying architecture uses a Model View Controller (MVC) pattern combined with an extended Data Access Object (DAO) pattern. The combination of the web service architecture with the aforementioned patterns enables a flexible exchange of the underlying data model implementation, e.g. the usage of a different database, and thus supports Enterprise Application Integration (EAI).

Different IPS² actors can utilize the prototypically implemented framework in order to find necessary information within their provider network. In order to continue the use case from section 1: If a maintenance worker locally identifies irregularities concerning a milling machine’s spindle, e.g. the sensor displays an unusually high temperature although the machine seems to function as usual, he can search for “Documents belonging to Omichron IPS² about high temperature of the spindle” within the system. At first the search engine semantically analyzes and parses the query and returns all documents and comments linked to the spindle of Omichron IPS². This enables the technician to narrow down the considered documents. Finally, it utilizes a predefined thesaurus and displays the necessary information, which in this

---

* Neo4J Graph Database: www.neo4j.com
particular example is hidden in a comment regarding “error 505” (which is the code for overheated components) and hints that the spindle needs to be replaced or it will break after approximately 50 hours of further operation.

State of the art full-text search engines would not display the required information and therefore render the information retrieval very difficult due to the vast amount of documents in a PLM system about “high temperature” and the different terminology used.

6. Conclusion and Future Work

A number of risky decisions made by different actors determine an IPS²’s success, especially during the use phase. In order to make these decisions, actors need relevant information. Finding this information is a time-consuming process. The paper on hand presents a semantic information retrieval framework in order to support IPS² actors in finding required information and enabling actors to focus on their main value-adding tasks by providing intuitive query possibilities that are semantically analyzed in order to traverse the semantic network and thus present the required information.

The framework consists of three main components: Firstly, creating semantic networks on entity level enables linking documents to different IPS²-PLM related ontology objects. Secondly, text mining techniques allow the analysis and processing of document contents, which are saved in the semantic network for performant information retrieval. Thirdly, the natural language user query processing concept facilitates users to exploit domain knowledge and retrieve targeted information. An example consistently used throughout this paper emphasizes the practical implementation of the concept.

The concept described in this paper will be enhanced in future research. A use-case will be developed in order to test different scenarios during use and development phase. Different query result weighing algorithms based on graph traversal will be added to the framework.

References


