Semantic Exploitation of Implicit Patent Information

Klaus Ulmschneider
Institute of Artificial Intelligence
Ulm University, Germany
Email: klaus.ulmschneider@uni-ulm.de

Birte Glimm
Institute of Artificial Intelligence
Ulm University, Germany
Email: birte.glimm@uni-ulm.de

***Abstract*— In recent years patents have become increasingly important for businesses to protect their intellectual capital and as a valuable source of information. Patent information is, however, not employed to its full potential and the interpretation of structured and unstructured patent information in large volumes remains a challenge. We address this by proposing an integrated interdisciplinary approach that uses natural language processing and machine learning techniques to formalize multilingual patent information in an ontology. The ontology further contains patent and domain specific knowledge, which allows for aligning patents with technological fields of interest and other business-related artifacts. Our empirical evaluation shows that for categorizing patents according to well-known technological fields of interest, the approach achieves high accuracy with selected feature sets compared to related work focussing on monolingual patents. We further show that combining OWL RL reasoning with SPARQL querying over the patent knowledge base allows for answering complex business queries and illustrate this with real-world use cases from the automotive domain.**

I. Introduction

Knowledge workers and decision makers are confronted with a massive load of data from numerous heterogeneous sources, making it difficult for them to identify the relevant information for performing their tasks [1]. One particular challenge is the identification, analysis and monitoring of patents with regard to a firm's strategic orientation, i.e., matching and utilizing them with respect to the firm's internal technological fields of interest or existing product portfolios. Specifically challenging is the abstruse language style with complex syntactic structure and legal terminology, the richness and novelty of technical terms, the lack of meaningful keywords (see e.g., [2], [3], [4], [5]) as well as patents being written in foreign languages. Moreover, considering the sheer amount of patent information, the scope of analysis as well as the richness of information uncovered and, therefore, the business value, can be limited [6]. The typically required manual patent analysis can be considered as time-consuming and requires a degree of technical and legal knowledge. Hence, the scientific and technological knowledge, which can be gathered from patents, is often not used to its full potential [3], [7], [8]. In contrast, patent information can provide essential technological information to define business strategies [6] and support decision making, e.g., in the context of a firm's innovative processes [9], [10]. More precisely, patents can be used in state-of-the-art and infringement analyses, prior art search, technology planning, R&D portfolio management, human resource management (e.g., to identify internal and external experts in a specific technological field), external knowledge generation, competitor and technology assessments, exploitation of emerging markets as well as forecasting future trends and business opportunities (see e.g., [9], [11], [12], [6]). Hence, the outcome of patent-related activities can certainly affect a firms value and performance [13]. Therefore, searching, analyzing, and monitoring patent information to gain commercial intelligence has become crucial from legal, R&D, and managerial viewpoints. The retrieval, analysis and evaluation of patents must be institutionalized to ensure the continuous and systematic use of patent information in a firm's R&D and decision-making processes [12]. An important aspect of institutionalizing the patent process is the alignment of patent information to the prevalent, usually domain- or firm-specific paradigms of its consumers to achieve enhanced accessibility and comprehensibility for non-experts. Consider knowledge workers and decision makers, who usually have limited legal expertise. They are mostly not or only partially familiar with the patent-specific terminology, phraseology, or existing patent classification systems. Consequently, (1) the alignment and integration of patent information to well-known cognitive patterns (e.g., domain- or firm-specific conceptual systems) and (2) computational intelligence, i.e., utilizing patent information seamlessly in daily business, is desirable.

Conventional patent analysis, while helpful for gaining technological information and identifying the present condition of technology assets, does not attempt to integrate technological and commercial perspectives by identifying promising new business opportunities [6] and does not release knowledge workers from laborious tasks such as searching, evaluating or monitoring patents. However, machines can assist in automating analysis processes and utilizing patents by extracting valuable information. Particularly, implicit information has to be extracted from higher level associations among patent documents, their textual content as well as between other business-related artifacts. Picking up this challenge, we present an integrated framework, which allows for deep analysis of patents, i.e., being capable of analyzing structured and unstructured patent information and exploiting it in various dimensions. Specifically, we combine techniques from several fields of computer science, namely natural language processing (NLP), information extraction (IE), machine learning (ML), information retrieval (IR) and enhance traditional text processing components with Semantic Web (SW) technologies, which allow for identifying interdependencies between entities and inferring knowledge from extracted and normalized information. With the alignment of patent information to well-known paradigms, such as technological fields of interest, which usually already have references to ongoing projects, persons or products (among others), new business opportunities can be generated (e.g., for proactive patent management or decision

978-1-5090-4240-1/16/$31.00 ©2016 IEEE

making), and, therefore, economic value. Further, the increased accessibility, understandability and utilizability achieved by the alignment of patents with well-known conceptual systems enables non-experts, such as research engineers, business analysts, HR managers, or executives, to utilize patent information in their daily business. Hence, the traditional, mostly isolated patent analysis is shifted towards an integrated, business-focused viewpoint.

The paper is organized as follows. Section II presents an integrated analysis framework to exploit patent information and shows how to map it to a domain of interest, i.e., existing business artifacts. Section III demonstrates the applicability of the approach and presents our findings along with examples from the automotive domain. Finally, Section IV discusses related work and Section V concludes with a summary and an outlook.

II. Analysis Framework

The analysis and monitoring of patent-related information requires the integration of complementary data as well as the discovery of hidden facts and relationships by means of natural language processing and semantic exploitation [11]. Therefore, in this section, a framework based on three major components is presented: A preprocessing component, which is responsible for integrating and unifying available business artifacts, a semantic pipeline (SP) for analyzing patent content, and a knowledge base (KB) which is represented as an RDF dataset.

A. Integration and Processing of Patent Information

In order to integrate and thoroughly exploit patent information, patent analysis is formalized with an SP (cf. Figure 1), which allows for merging various analysis *components* in a structured and extensible way, i.e., the SP is capable of conducting several processing steps to structure, normalize, and link patent information.

Definition 1: Let C be a set of *processing components* such that each $c \in C$ provides a *processing function* f that can be affected by a set of *input parameters* P. A *semantic pipeline* SP w.r.t. C consists of an input I, an output O, and a sequence $c_1, \ldots, c_n$, $n \geq 2$, of processing *components* from C such that each $c_i \in C$, $1 \leq i \leq n$, is parameterized by a set of *input parameters* P_i. Each *component* c_i, $1 \leq i \leq n$, processes an input I_i and produces an output O_i that may serve as the input for the following *components*, i.e., $I_i = \{I, O_1, ..., O_{i-1}\}$, $O_i = I_{i+1}$, $1 \leq i < n$, or acts as an *input parameter* P_{i+1}, $1 \leq i < n$.

An SP therefore constitutes the symbiosis of *components*, which usually perform syntactic, linguistic, semantic or statistical analyses on business artifacts, such as patents or project fact sheets, and their existing relations. Available *components* are compliant with a specified interface agreement and therefore can be arranged consecutively and adjusted by the plug-in – plug-out principle. In particular, *components* which implement NLP-algorithms or machine learning techniques can take care of extracting hidden facts and aligning patents with technological fields of interest. Furthermore, SW technologies are capable of bridging the gap between patents and other business-related artifacts and allow for inferring new knowledge (e.g., unknown facts) from existing relations to well-known structures as well as from processed and normalized content, to uncover implicit information (e.g., latent interdependencies). As example, consider extracted facts, such as spatial information, which can be reused as *input* by a reasoning *component* or a detected language, which can be the *input parameter* for a lemmatizing or a categorization *component*. Consequently, the *output* of an SP can be new artifacts, relations between them as well as new, updated, or enriched object or relation properties, which we call *annotations*.

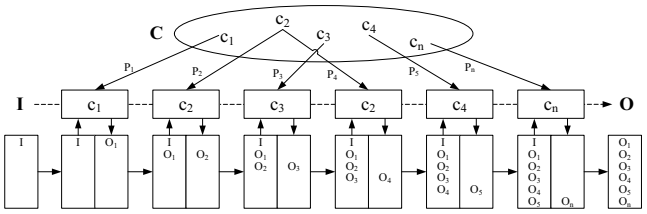

Fig. 1. Semantic pipeline

In the following, we illustrate the composition of the SP used in our scenario. Note that the adjustment and the correct arrangement of *components* is an important step when setting up an SP. In a first step, if divergent, possibly heterogeneous data representations (possibly from different sources) of business artifacts (e.g., patents with meta information like the owning company) are gathered in their original format, transformed to homogeneous objects [1], and provided to the SP. The SP, orchestrated as shown in Figure 2, takes over each artifact and starts with several preprocessing steps. The first *component*, the Language Detector, identifies the language of an object and serves as *input parameter* in further processing steps. The second *component*, the Tokenizer, splits raw text into smallest processible units, i.e., tokens. The tokens are then normalized with the help of a Lemmatizer, such that higher accuracy is ensured in further processing steps. With these general *components*, the scope of analysis and the richness of information uncovered can be limited, making it difficult to locate significant patents which might be of relevance for a firm's business [6]. Therefore, the SP is enriched with further annotators. A natural language processing (NLP) *component* allows for deep linguistic analysis, such as part of speech tagging (POS), noun phrase chunking (NPC), and named entity recognition (NER). As example, noun phrases are distilled by their frequencies, since they are often used to describe an invention, regularly represent novel terms, and can provide indications to identify the actual claim or topical information [2], [5]. Additionally, the resultant noun phrases are filtered by general language-specific and patent-specific stop words.

Custom annotators, which usually serve domain- or firm-specific purposes, complete the processing of patent content. Overall, linguistic, syntactic and semantic *components* are used for these tasks. For example, competitors are extracted from full text in order to be contextualized, i.e. linked with the corresponding entities in succeeding processing steps. Other patent-specific *components* parse and process metadata fields (also known as "bibliographic metadata") which occur in structured (e.g., application date, prioritization date) and semi-structured (e.g., inventors, assignees, citations, or

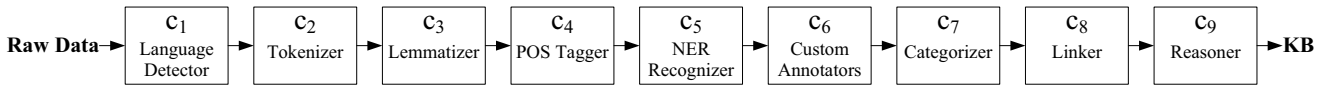

Fig. 2. Basic SP components and their orchestration

classifications) formats. Consequently, explicit, i.e., directly extracted information (e.g., person and organization names, cited patents) and corresponding relations (e.g., associations to patent classification systems, patent families, locations, persons, or organizations) can be created. One classification system, for example, is the International Patent Classification System (IPC) provided by the World Intellectual Property Organization (WIPO). However, IP classes, which are extracted from patent metadata fields, follow a specific syntax and usually cannot directly be matched with the WIPO IP class symbols. Thus, in order to allow assignments of patents to their corresponding IP classes, a respective *component* normalizes the extracted IPC symbols to match the WIPO standard.

Next, after processing the textual content and metadata of an artifact, the Categorizer annotates each artifact with topical information. In a final processing step, the above mentioned linking *component* integrates explicit (e.g., IP classes, locations) and implicit (e.g., competitors, similar patents) assertions between business artifacts, which are derived from preceding processing steps, in the KB.

The KB itself is a directed, labeled and weighted multi-edge graph represented as an RDF dataset, which allows for persisting and querying all underlying data, i.e., conceptual systems (e.g., the WIPO IP Classification System or a location taxonomy), which are represented as hierarchically organized entities, as well as business artifacts (e.g., patents, persons) and all (semantic) relationships among them.

Definition 2: RDF is based on the set I of all *Internationalized Resource Identifiers* (IRIs), the set L of all RDF *literals*, and the set B of all *blank nodes*. The set T of RDF *terms* is $I \cup L \cup B$. We generally abbreviate IRIs using prefixes rdf, rdfs, owl, and xsd to refer to the RDF, RDFS, OWL, and XML Schema Datatypes namespaces, respectively. We use the empty prefix for our domain-specific IRIs. An RDF *graph* G is a set of RDF *triples* of the form $(\text{subject}, \text{predicate}, \text{object}) \in (I \cup B) \times I \times T$. An RDF *dataset* is a collection of RDF graphs $\{G_d, \langle i_1, G_1 \rangle, ..., \langle i_n, G_n \rangle\}$, and comprises exactly one RDF graph G_d, called the *default graph*, which is unnamed and may be empty, as well as zero or more *named graphs*, consisting of a pairwise disjoint set of one $i \in I$ or one $b \in B$, which represents the unique *graph name*, and one RDF graph G_n.

RDF graphs can be interpreted in a number of ways based on various W3C recommendations. The simple semantics of RDF specifies the graph structure, whereas more elaborated semantics, such as RDFS or OWL, provide a special meaning to certain terms and allow additional RDF statements to be inferred from explicitly given assertions. Hence, the abstract specifications allow for describing business artifacts (entities) and their interdependencies (relations) in an ontology. However, in order to answer business-related questions, the general RDF definitions must be enriched and formalized with additional semantics.

Entities are identified by an IRI and comprise a set of properties such that all types of entities, e.g., patents, profiles or technological fields of interest, share a common set of properties (e.g., "title", "abstract", "creationDate"). Thus, each entity can be described by assertions of the form (resource, predicate, literal), where the resource represents an entity instance, the predicate a property and the literal denotes the actual value. Analogously, distinct types of semantic relations (e.g., "isAuthorOf", "isSimilarTo") are defined as properties, such that each entity can be linked to another one by triples of the form (resource, predicate, resource), where the resource represents an entity instance and the predicate denotes a specific type of relation.

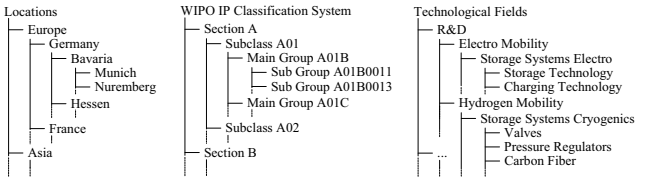

Fig. 3. Exemplary conceptual systems utilized for evaluation

In order to gain additional value from the KB, the general ontology must be refined with domain-specific details. Therefore, patent-related properties, such as "applicationDate", "description" or "claim" are defined. Moreover, appropriate types of relations, such as "hasInventor", "sameFamilyAs", "cites", "hasTopic", or "assignedToIPC" link patents with each other or other types of entities, such as profiles, technological fields of interest, locations or IP classes. We further categorize the above described relations into hierarchical, associative, and equivalent ones. For this purpose three properties `:hierarchical`, `:associative`, and `:equivalent` are introduced and appropriate subproperty statements are added. Moreover, we declare properties as symmetric, transitive, or reflexive, respectively, e.g., subproperties of `:hierarchical` are declared as transitive. The following exemplary statements illustrate such declarations:

```
:hasInventor rdfs:subPropertyOf :hasRole .
:hasRole rdfs:subPropertyOf :associative .
:isPartOf rdfs:subPropertyOf :hierarchical,
 owl:TransitiveProperty .
:hierarchical rdfs:subPropertyOf
 owl:TransitiveProperty .
:sameFamilyAs rdfs:subPropertyOf :equivalent .
:equivalent rdfs:subPropertyOf
 owl:TransitiveProperty, owl:SymmetricProperty,
 owl:ReflexiveProperty .
```

Transitivity and the corresponding transitive closure of the relations allows for traversing conceptual systems like rooted tree graphs, such as technological fields of interest (cf. Figure 3) within the KB and resolving all related business

entities (e.g., patents, profiles, or products) including their properties. Consequently, once patents are aligned with such conceptual systems, the reasoner is capable of contextualizing them with other business artifacts in various ways. In general, network analyses (e.g., network structure and evolution analyses like cluster identification or resolving paths based on specific semantic relations) are enabled, i.e., the identification of unknown correlations and dependencies. Explicit semantic relations between business entities and conceptual systems bridge the gap between patents and other business artifacts. Consider a patent and a technology message from a scout which are both associated with the same technological field of interest or one of its parents or children of the corresponding conceptual system. They can be implicitly associated with each other, e.g., for the purpose of an infringement analysis or to be brought in context for a specific engineering problem. Corresponding (SPARQL) queries can be narrowed down from network structure (entities and their relations) to property level.

Moreover, the illustrated abstraction level for relations (i.e., associative, hierarchical, and equivalent) allows for deducting implicit information on a higher level for each instance. For example, all persons who have a role for all patents in the field of electro mobility, where each role can be associated with its subordinate role definitions like inventor or assignee and electro mobility covers associations with its child entities such as storage systems. In consequence, further implicit information and interdependencies can be deducted on conceptual level based on the output of the SP (e.g., processed IPCs, extracted competitors, associated profiles, or topical and similarity analyses). Figure 4 demonstrates a simplified KB with such exemplary inferences.

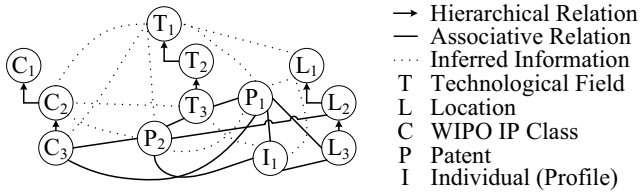

Fig. 4. Exemplary business network with possible inferences

An important aspect for real-life applications is, however, not yet covered. Consider a patent which was identified by the SP as being similar to another one. Basically, this assertion can be expressed in a single statement: `:patentA :isSimilarTo :patentB`. However, the assertion does not provide any information to which degree `:patentA` is similar to `:patentB`. Therefore, querying for similar patents may result in an unordered (and possibly large) list of patents, which makes it difficult for patent analysts to identify the relevant patents when performing their tasks. Weights, which define the strength of a semantic relationship between two entities, address this aspect. A weight is expressed as a numeric range between 0 and 1, with 1 indicating the strongest possible semantic relation. The assignment of weights to RDF statements is, however, not directly possible, but the concept of reification can be used to further describe each relation with meta statements (i.e., statements about RDF statements), such as “hasWeight”, “creationDate” or “expirationDate”. The following example illustrates the concept of reification for the purpose of assigning weights:

```
[] rdf:subject :patentA ;
   rdf:predicate :isSimilarTo ;
   rdf:object :patentB ;
   :hasWeight "0.75"^^xsd:decimal .
```

By adding additional semantics to relations a patent analyst is not only capable of identifying related business artifacts, but the most relevant ones (e.g., identifying similar patents is important for infringement analyses).

Summing up, the SP constitutes the symbiosis of processing *components*, which usually perform syntactic, linguistic, semantic or statistical analyses on business artifacts, such as patents or project fact sheets, and their existing relations. Available *components* can be arranged consecutively and orchestrated by the plug-in – plug-out principle, such that the SP, which is constituted by the composition of several *components*, serves an intended purpose (e.g., metadata enrichment, detection of implicit relations between entities). Each *component* can be affected by a set of *input parameters* and generates an *output* which can be reused by succeeding *components*. The SP is capable of analyzing structured and unstructured (patent) content, creates *annotations*, and establishes relationships between business artifacts. Moreover, new (implicit) knowledge can be inferred by a reasoning *component*. In particular, the KB is used by the SP to (1) align processed business artifacts with existing ones and to (2) store the business artifacts along with the resulting *annotations* and relationships persistently. The KB is modeled according to the above-mentioned principles and is compliant with the latest W3C standards. Data from the KB is accessible and processible by all implemented *components*, new information can be added at runtime, and the underlying model can be adapted to any domain of interest. Further note that the underlying model allows for integrating conceptual systems like rooted tree graphs, which are capable of representing general, domain- or firm-specific structures (e.g., locations, product-related structures, or organizational structures) and interconnect business entities with high semantic expressiveness. In addition, the uniform representation of business artifacts allows for answering complex business questions on multiple levels by providing a SPARQL endpoint on top of the KB, i.e., the reasoner can be used to deduct implicit relations. Examples include deriving additional IPC assignments by inferring the children of aligned IP classes, detecting spatial patterns, topical analyses, or identifying competences by analyzing roles of profiles based on underlying semantic associations.

B. Aligning Patents with Domain-Specific Systems

An important step while analyzing patents is to extract thematic information, which allows for aligning patent information with other business-related artifacts. Although each patent has pre-assigned technological fields, such as IPC symbols, many of these assignments are either too general or too specific to fit the intended knowledge structures for topic interpretation [4], [5] and are difficult to understand to its full extend by non-experts. Therefore, restricting analyses on such given classifications does not meet the requirement of making patent

information more accessible to non-experts. However, domain- or firm-specific paradigms, such as organizational, spatial, product-related, or technological structures (e.g., departments, sales regions, product range, technological fields), which can usually be represented as conceptual systems, can be excellent connectors to detect and create relations between business-related artifacts (e.g., patents and project fact sheets associated with the same technological field). In particular, most of such structures can be formalized (and mostly already are) and represented as hierarchical conceptual systems. Consequently, aligning patents with such existing conceptual systems allows for the integration to existing business environments and, therefore, making them comfortably accessible for knowledge workers and decision makers. Following this, we select technological fields of interest as connectors for the experimental validation (cf. Section III), since they are usually defined with a firm's strategy.

III. Validation

In this section we present experimental results conducted with a dataset from the automotive domain.

A. Research Design

The presented analysis framework builds upon several components, namely an SP which incorporates an integration component, which is responsible for providing and preprocessing (business) entities in their raw formats, and a KB. The backbone of the system is constituted by a general top-level ontology, which defines (business) entities and their possible relationships, extended by a lightweight patent ontology (a domain-specific ontology). The resulting enhanced ontology belongs to the OWL 2 RL profile and comprises 10 classes, 28 object, and 54 data properties. Among the properties there are 5 transitive and 3 symmetric properties and among the 78 logical axioms, there are 9 object and 13 data property domain axioms, 23 object property range axioms, 24 subproperty axioms, and 1 inverse object property axiom. The ontology is stored in the KB along with several conceptual systems and the processing results from the SP. Consequently, the KB is the basis for answering complex business queries. The software itself is written in Java and all elementary components are publicly available: for data integration, several interfaces to access various data formats are provided, namely Apache Tika[1] and Apache POI[2] (e.g., to access PDF documents or spreadsheets) as well as several database connectors (e.g., MySQL JDBC). Connectors to additional sources can be integrated with reasonable effort. The SP is orchestrated as illustrated in Figure 2 and realized with Apache UIMA,[3] a de facto standard used in industry, which provides a robust infrastructure to integrate *components* in a modular manner. In order to extract facts, create noun phrase vectors or further enrich patent documents, UIMA is extended with a Language Detector, several NLP *components* (e.g., Tokenizer, Lemmatizer, POS tagger) as well as several custom *annotators* (e.g., a competitor annotator or a named entity annotator to extract person, organization or location names). The alignment of patents with predefined technological fields is done with the OpenNLP Document Categorizer,[4] which is based on the Maximum Entropy framework. The KB, which is storing the processed data persistently, compromises a triple store (TDB), which is built for large datasets, and Apache Jena[5] for manipulating and querying data.

B. Dataset

Before analyzing patents, several conceptual systems, which act as connectors in the overall business network, have to be added to the knowledge base. In particular, two exemplary rooted tree graphs, i.e., hierarchically organized entities, which are publicly available, are used to contextualize patents: the WIPO IP Classification System[6] (≈72.000 entities) and locations[7] (≈43.000 entities). Further, domain-specific technological fields of interest (≈200 entities), organized in the same manner, are added (cf. Figure 3). In order to analyze patents and align them to technological fields of interest, we used a multi-lingual patent corpus which origins from several exports of a commercial patent database.[8] The patents, almost 10,000 in total, were selected by a patent expert with regard to two emerging technological fields of interest in the automotive domain: *electro mobility* and *hydrogen mobility* (cf. Figure 3). Each given category from the sample data, i.e., a technological field of interest, which was assigned by a patent expert to each patent for training purposes, denotes the most specific category within the hierarchy. Consequently, the respective technological fields represent the leaves of the corresponding rooted tree graph to which patents can be aligned.

C. Experimental Setup

In order to validate the alignment of patents to technological fields of interest, we conducted a controlled experiment. We analyzed patents in English, German, Chinese and Japanese, which are among the most important languages in the patent domain. The distribution of patents across the studied languages for the training set was 58.28% English, 27.40% Japanese, 8.04% German, and 6.28% Chinese. The patent sections title (t), abstract (a), description (d), and claim (c) were selected as features and the training and categorization task was performed using four different combinations of these features for each language: 1) t, a 2) t, a, d 3) t, a, c and 4) t, a, d, c. The resulting training subsets may contain only a small amount of sample patents for each technological field when selecting them by language and availability of content within the mentioned patent sections (e.g., some old patents are physically available, but the full text is not available in a digital format). The nature of patents allows, however, for minimizing this effect to some extend with the help of patent families: Since family members belong to the same invention, almost identical content can be found in patents belonging to the same family. Therefore, a family member, if available and preferably containing sample content in English, was selected for categorization. The motivation behind the

[1] http://tika.apache.org/
[2] http://poi.apache.org/
[3] http://uima.apache.org/
[4] http://opennlp.apache.org/
[5] http://jena.apache.org/
[6] http://www.wipo.int/classifications/ipc/en/ITsupport/
[7] http://www.geonames.org/
[8] http://www.minesoft.com/patbase.php

method is to increase the recall for tasks where high recall is preferred (e.g., searching patents for a technological field of interest or analyzing patent strategies of competitors in different countries). Note that this method can also be employed as an alternative for machine translation tasks. Since given categories for training are provided as the leaves of a conceptual system, i.e., a taxonomy of technological fields, the choice for sample content is further reduced. Consequently, the repeated random sub-sampling method is selected for validation. The idea behind this method is to randomly split the set of sample patents into unique training and validation sets. For each split, the categorizer is trained with the training set and validated with the validation set (which contains only unseen patents). Following this, the experiment was conducted over five runs (cross-validation) and the means and standard deviations of the results were calculated respectively. Specifically, the categorizer was trained and validated by changing exactly one parameter for each run, i.e., content language, a unique combination of patent sections (features), and a threshold. The threshold is based on the assumption that the higher the probability of the best class, the more accurate a patent is categorized. If the best class confidence exceeds the threshold, the categorization task was successful. In order to meet this requirement, we added further indicators to the categorization task. Every assigned technological field, equipped with a probability value, is reweighted by (1) IPC match and (2) noun phrase vector similarity. Expert interviews revealed that some IP classes can directly be assigned to technological fields and that noun phrases describe a patent more accurately than single keywords. Therefore, the IPC match indicator has influence on the probability if one of the (normalized) IP classes found in patent metadata matches with a set of predefined IP classes for a (leaf) sample. Further, we created a noun phrase vector, cleared from stop words, from each patent document and calculate the similarity with a given noun phrase vector for each (leaf) sample. Note that further indicators can be added to the reweighting function on demand (e.g., CP classes, citations, patent families, similarity to other patents in the same category). Accuracy, defined as the percentage of correctly categorized patents, was used as performance measure for our results. Consequently, the average accuracy denotes the mean of all runs. After completion of the categorization task, each patent was added to the KB by populating the integrated ontology. Further, semantic relations to existing conceptual systems (e.g., technological fields of interest) and other business artifacts (including other patents) were created by the SP, i.e., each patent was contextualized with its corresponding entities in the KB. Overall, the RDF dataset resulted in more than 10.6 million statements for the roughly 10,000 patents. Most patent searches, e.g., regarding patentability or validity, are highly business-sensitive and missing relevant documents would be unacceptable [11]. Therefore, the reasoner can further be used for deriving the actual parents of the categorized technological fields. Remember, that the sample categories are provided as the leaves of the hierarchy, and, consequently, are very specific. Therefore, overlapping of similar samples can occur. Consider, for example, valves and pressure regulators, which serve more ore less the same purpose. Since their descriptions might be quite similar, the prediction probabilities for the best classes converge. However, a patent expert searching for patents is primarily interested in high recall and is usually not evaluating very specific technological fields. Therefore, it is sufficient to satisfy higher level queries (e.g., patents for storage systems cryogenics). The reasoner can resolve such queries and categorizes all children as correct results of their parent category. The advantage of the approach is that not all members of a hierarchy have to be trained, but all are resolved on demand with help of the reasoner. Further, false positives derived from the categorization *component* can turn correct when queried over their parent (e.g., both, valves and pressure regulators, in their context, belong to storage systems cryogenics). This bottom-up approach is based on the assumption that the higher the level of a queried hierarchically organized entity, the higher the precision and recall of the results. Following this, we achieved an extra (averaged) 3–8% accuracy in terms of correctly categorized documents with respect to all mentioned parameters when querying for the first parent level. Next, to support categorization tasks, the reasoner enables the detection of implicit knowledge and supports answering business-related questions. In particular, queries over hierarchically organized conceptual systems and their associations with business-related objects are empowered. For example, consider firms, which protected an invention in the European market for electric storage systems since 2010. The reasoner resolves all patent applications in European countries (including European patent applications, i.e., directly filed at the European Patent Office EPO), selects the patents which are assigned to the technological field of interest (including its children) and returns a list of organizations (e.g., derived from assignees or extracted firms in patent content) according to the given time range.

D. Findings

The most important aspect of aligning patent information with business-related artifacts constitutes the correct categorization according to well-known technological fields of interest, since they act as connectors to other business entities in the KB, such as project or technology fact sheets. Table I illustrates the average categorization accuracy over five runs with different combinations of patent sections and the standard deviation enclosed in parentheses. In general, the categorization task was completed with notably high accuracy over all languages. Therefore, we can confirm that language independent alignment to a given set of technological fields is feasible for patent information. However, we observed that some parameters are influencing the results more than others. As example, the categorization of Japanese patents did not return adequate results when training only specific features (i.e., patent sections) and a threshold enforced a high categorization probability (cf. Table I, indicated in bold). This effect can be explained, as a patent expert approved, with divergent requirements of content within the actual sections in patent applications from the Japanese Patent Office (JPO) as well as the common attitude to protect inventions on micro-level. In turn, all other languages performed very well and mostly achieved an accuracy beyond 80% with minimal features (t, a) as well as for the combination of all sections. Surprisingly, the categorization tasks which incorporated the

TABLE I
AVERAGE CATEGORIZATION ACCURACY FROM FIVE REPLICATIONS

	ϕ	10				20				30			
β	*lang*	t/a	t/a/d	t/a/c	all	t/a	t/a/d	t/a/c	all	t/a	t/a/d	t/a/c	all
0.3	**en**	0.94 (0.089)	0.72 (0.130)	0.56 (0.134)	0.74 (0.195)	0.86 (0.096)	0.68 (0.148)	0.59 (0.108)	0.78 (0.045)	0.85 (0.056)	0.77 (0.078)	0.64 (0.080)	0.81 (0.064)
0.3	**de**	0.80 (0.000)	0.80 (0.200)	0.76 (0.167)	0.88 (0.084)	0.82 (0.091)	0.77 (0.130)	0.74 (0.074)	0.87 (0.076)	0.78 (0.077)	0.85 (0.073)	0.67 (0.062)	0.86 (0.083)
0.3	**cn**	0.90 (0.100)	0.96 (0.055)	0.68 (0.130)	0.96 (0.089)	0.92 (0.076)	0.95 (0.050)	0.72 (0.115)	0.97 (0.045)	0.89 (0.061)	0.96 (0.028)	0.63 (0.131)	0.95 (0.038)
0.3	**ja**	0.88 (0.110)	0.92 (0.084)	0.76 (0.134)	0.94 (0.089)	0.87 (0.084)	0.95 (0.050)	0.59 (0.114)	0.92 (0.045)	0.93 (0.043)	0.92 (0.030)	0.58 (0.141)	0.93 (0.078)
0.5	**en**	0.90 (0.100)	0.64 (0.167)	0.54 (0.182)	0.70 (0.100)	0.79 (0.139)	0.63 (0.097)	0.51 (0.108)	0.59 (0.082)	0.79 (0.072)	0.64 (0.055)	0.50 (0.122)	0.65 (0.090)
0.5	**de**	0.74 (0.114)	0.60 (0.071)	0.68 (0.084)	0.84 (0.055)	0.76 (0.055)	0.76 (0.147)	0.66 (0.065)	0.79 (0.119)	0.73 (0.078)	0.71 (0.109)	0.65 (0.077)	0.79 (0.045)
0.5	**cn**	0.90 (0.122)	0.98 (0.045)	0.56 (0.152)	0.96 (0.055)	0.98 (0.027)	0.95 (0.035)	0.71 (0.082)	0.95 (0.050)	0.90 (0.053)	0.95 (0.018)	0.63 (0.082)	0.95 (0.038)
0.5	**ja**	***0.36*** (0.207)	0.96 (0.089)	***0.16*** (0.114)	0.98 (0.045)	***0.19*** (0.065)	0.96 (0.042)	***0.18*** (0.110)	0.95 (0.035)	***0.25*** (0.102)	0.92 (0.030)	***0.24*** (0.134)	0.95 (0.056)
0.7	**en**	0.76 (0.152)	0.58 (0.130)	***0.18*** (0.084)	0.64 (0.114)	0.71 (0.096)	0.56 (0.147)	***0.32*** (0.076)	0.55 (0.100)	0.70 (0.085)	0.51 (0.092)	***0.24*** (0.109)	0.48 (0.104)
0.7	**de**	0.72 (0.110)	0.70 (0.141)	***0.44*** (0.195)	0.76 (0.134)	0.71 (0.096)	0.67 (0.160)	***0.49*** (0.114)	0.67 (0.115)	0.73 (0.097)	0.72 (0.061)	***0.35*** (0.107)	0.67 (0.092)
0.7	**cn**	0.88 (0.130)	1,00 (0.000)	***0.62*** (0.130)	0.94 (0.089)	0.89 (0.022)	0.97 (0.045)	***0.58*** (0.045)	0.96 (0.042)	0.85 (0.077)	0.95 (0.038)	***0.65*** (0.087)	0.95 (0.018)
0.7	**ja**	***0.10*** (0.071)	0.92 (0.084)	***0.14*** (0.089)	0.92 (0.084)	***0.11*** (0.082)	0.89 (0.055)	***0.10*** (0.100)	0.95 (0.035)	***0.12*** (0.056)	0.94 (0.043)	***0.09*** (0.037)	0.93 (0.043)

β = threshold, *lang* = trained language, ϕ = # trained patents, features: t = title, a = abstract, d = description, c = claim, all = all sections

claim section generally resulted in poor accuracy. In contrast, Japanese patents performed well for (t, a, c). Thus, we can deduct, that paying attention to actual patent sections during the training task can have a strong effect on the categorization accuracy. Nevertheless, if patents with unstudied languages are processed with the framework, some evaluation on that point is recommended, even if training all sections always resulted in adequate results. Further, not surprisingly, the more patents in the training set and the more thoroughly the training set is prepared, the more accurate one can expect the results to be. Despite the fact that many authors proposed improvements for categorizing patents, the achieved accuracy seems to be limited at a certain level. To the best of our knowledge no related work can reduce incorrect categorizations less than 10-15% (e.g., see [14], [15] for overviews). Following this, we define an accuracy of 85% as baseline for our experiments. Overall, with the identified optimal settings, our approach can compete and even outperform similar related work with an accuracy ranging between 75% and 95%. Note that our results are based on multilingual patents, a comparably small training set and the training task was performed on the lowest hierarchical level. Therefore, it is more difficult to achieve high accuracy with our general approach in contrast to fine-tuned solutions for a specific (monolingual) problem. Moreover, the query accuracy (e.g., for search and infringement tasks) generally increases for higher level queries using the reasoner. In addition, the proposed methodology is complementary with conventional approaches, adds new analytical capabilities, and the overall framework with its graph-based knowledge representation is designed in a generic way, i.e., it is not tailored to any specific technological fields of interest or use cases.

In consequence, the presented framework enables a wide range of business applications. We will briefly illustrate some exemplary real-world use cases which were realized. For example, we conducted a competitor analysis based on organizations, which were derived from patent metadata or extracted from patent content, for the technological field of storage systems electro. We found that suppliers hold more patents in this field than the manufacturers themselves and Japanese companies, in general, hold comparably more patents than others. Surprisingly, Tesla is not strong in this area with regard to their patent applications. A patent expert explained this outcome with many patent applications directly assigned to the company's founder compared to applications which are associated with the company itself. Moreover, we analyzed competences for a specific company, i.e., persons which were mentioned in patent applications since 2010. The results were used, in combination with other business-related information (e.g., persons working on R&D projects in a specific technological field), to identify gaps between technological fields of interest and experts with a focus on future products, i.e., whether the company has enough expertise to realize its planned projects. Furthermore, the proposed framework allows for creating several visualizations to support decision makers. Examples include citation networks, technology (road)maps or other patent-product-related questions (e.g., which patents exist for a specific part).

IV. RELATED WORK

Patent analysis has challenged researchers for more than a decade. General approaches and the state-of-the-art of patent analysis have already been summarized in several studies (see e.g., [9], [16], [11], [14], [5]). Evolutions are ranging from the adoption of natural language processing (NLP) and semantics for automatic processing of information to the design of innovative and efficient user interfaces, from the integration of information coming from less traditional sources (such as the World Wide Web) to the exploitation of hidden information [11]. In general, we distinguish between quantitative (i.e., statistical) and qualitative (i.e., declarative) approaches.

Numerous studies have attempted to create patent networks with a focus on technology planning or technology roadmaps based on quantitative analyses (see e.g., [4], [8]). The majority of authors use text mining or similar techniques (e.g., to extract keywords or process patent metadata such as citations) as a basis for their analyses (e.g., for creating networks) and use statistical methods for evaluation. Some authors provide visualizations for their results which can be beneficial for decision makers. Another aspect of patent analysis, with respect to knowledge workers and decision makers, is to determine the quality of patents [12], [13], [17]. Some authors explicitly relate their work to business- and decision-making processes in this context (e.g., see [12], [18], [6], [15]). Declarative approaches are, in general, semantic solutions using formal representations such as taxonomies or ontologies and primarily focus on integrating and representing patent information from heterogeneous sources to support patent retrieval (see e.g., [10], [3], [7], [19], [20]). Other related

approaches have attempted to categorize patents according to existing classification systems, such as IPC, CPC, EPO, or USPTO, or cluster them to derive topical information (see e.g., [18], [21], [15]). However, these methodologies are mainly keyword-based and therefore language-dependent to some extent. None have, however, fully exploited multilingual patent information in its structured, semi-structured or unstructured formats, aligned it with well-known business artifacts, which can be represented as conceptual systems, as well as reusing extracted facts to derive implicit information from higher level associations to address concrete business-related requirements. Moreover, the presented approach is not limited to a certain use case, and thus, allows for performing further statistical, semantic and graph-based analyses, e.g., to determine a patent's value, discover similarities between (other) business artifacts, support patent retrieval, or allow dependency analyses or technology forecasting (e.g., citation-based, temporal) as illustrated in related work. In conclusion, the proposed framework is capable of supporting most aspects of a firm's innovation- and patent-related processes.

V. Summary and Outlook

This paper presents an integrated framework for analyzing patents and aligning them with business-related artifacts. A controlled experiment with multi-lingual patents originating from a real-world case in the automotive domain demonstrates how to align multilingual patents with technological fields of interest and other business-related artifacts with high accuracy. In contrast to related work, the illustrated solution meets the general requirements of a holistic tool, i.e., integrating complementary data, the ability to process large sets of multilingual patents, discovering hidden facts and relationships, eliciting topical information as well as empowering statistical and semantic exploitation [16], [11] to create additional business value with computational intelligence. The combination of methods from different fields of computer science, an adaptable and standard-compliant graph-based uniform representation as well as the capability to process structured, unstructured and interconnected information allows various types of analyses (linguistic, syntactical, statistical, semantic, or network-based exploitation), and ensures the interoperability for various types of business artifacts and their relationships. Moreover, implicit information from higher level associations can be derived with a reasoning *component*. With the achieved analysis opportunities and the ability to integrate patent intelligence into supplementary business processes, the pro-active management of patents can be institutionalized and new business cases are enabled. Furthermore, users with different roles, such as technical engineers or decision makers, can explore and utilize patent information in their daily tasks with less manual effort and expertise. New functionality or domain-specific requirements can easily be incorporated due to the plug-in – plug-out principle. Examples include adding, removing or reconfiguring *components* of the SP, the integration of further background knowledge in the ontology or further conceptual systems (e.g., organizational units and their roles, products and their components). Based on the proposed framework, further linguistic, semantic, graph-based and time-based analyses (e.g., technology roadmapping and forecasting, pattern analysis over citations, deep analysis of patent claims, competitor and activity analyses) will be integrated in future work. Moreover, by shifting our attention to the evaluation of patent determinants with regard to their quality and impact, i.e., examining their additional value for a business, will enhance the benefits of the proposed framework and enable further business cases. In addition, the general applicability of the proposed framework, i.e., for other domains and datasets, will be evaluated.

References

[1] K. Ulmschneider, B. Michelberger, B. Glimm, B. Mutschler, and M. Reichert, "On maintaining semantic networks: Challenges, algorithms, use cases," *Int. J. of Web Inform. Systems*, vol. 11, no. 3, pp. 291–326, 2015.

[2] H. Aras, R. Hackl-Sommer, M. Schwantner, and M. Sofean, "Applications and challenges of text mining with patents," in *Proc. of the 1st Int. Works. on Patent Mining and Its Applications (IPaMin'14)*, vol. 1292, 2014.

[3] D. Eisinger, J. Mönnich, and M. Schroeder, "Developing semantic search for the patent domain," in *Proc. of the 1st Int. Works. on Patent Mining and Its Applications*, vol. 1292, 2014.

[4] Y.-H. Tseng, C.-J. Lin, and Y.-I. Lin, "Text mining techniques for patent analysis," *Information Processing and Management*, vol. 43, no. 5, pp. 1216–1247, 2007.

[5] S. Verberne, E. D'hondt, C. H. Koster, and N. Oostdijk, "Quantifying the challenges in parsing patent claims," in *Proc. of the 1st Int. Works. on Advances in Patent Information Retrieval (AsPIRe'10)*, 2010, pp. 14–21.

[6] S. Lee, B. Yoon, C. Lee, and P. Jinwoo, "Business planning based on technological capabilities: Patent analysis for technology-driven roadmapping," *Technological Forecasting and Social Change*, vol. 76, no. 6, pp. 769–786, 2009.

[7] M. Giereth, A. Stäbler, S. Brügmann, M. Rotard, and T. Ertl, "Application of semantic technologies for representing patent metadata," in *Informatik 2006*, vol. 1. Köllen, 2006, pp. 297–304.

[8] B. Yoon and Y. Park, "A text-mining-based patent network: Analytical tool for high-technology trend," *J. of High Technology Management Research*, vol. 15, no. 1, pp. 37–50, 2004.

[9] A. Abbas, L. Zhang, and S. Khan, "A literature review on the state-of-the-art in patent analysis," *World Patent Information*, vol. 37, pp. 3–13, 2014.

[10] M. Bermudez-Edo, M. V. Hurtado, M. Noguera, and N. Hurtado-Torres, "Managing technological knowledge of patents: HCOntology, a semantic approach," *Computers in Industry*, vol. 72, pp. 1–13, 2015.

[11] D. Bonino, A. Ciaramella, and F. Corno, "Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics," *World Patent Information*, vol. 32, no. 1, pp. 30–38, 2010.

[12] H. Ernst, "Patent information for strategic technology management," *World Patent Information*, vol. 25, no. 3, pp. 233–242, 2003.

[13] B. Fabry and H. Ernst, "How to make investors understand the value of IP assets," *les Nouvelles*, vol. 40, no. 4, pp. 201–208, 2005.

[14] J. C. Gomez and M.-F. Moens, "A survey of automated hierarchical classification of patents," in *Professional Search in the Modern World*. Springer, 2014, vol. 8830, pp. 215–249.

[15] S. Venugopalan and V. Rai, "Topic based classification and pattern identification in patents," *Technological Forecasting and Social Change*, vol. 94, pp. 236–250, 2015.

[16] F. Baudour and A. v. d. Kuilen, "Evolution of the patent information world: Challenges of yesterday, today and tomorrow," *World Patent Information*, vol. 40, pp. 4–9, 2015.

[17] A. J. C. Trappey, C. V. Trappey, C.-Y. Wu, and C.-W. Lin, "A patent quality analysis for innovative technology and product development," *Advanced Engineering Informatics*, vol. 26, no. 1, pp. 26–34, 2012.

[18] G. J. Kim, S. S. Park, and D. S. Jang, "Technology forecasting using topic-based patent analysis," *J. of Scientific and Industrial Research*, vol. 74, no. 5, pp. 265–270, 2015.

[19] S.-H. Liu, H.-L. Liao, S.-M. Pi, and J.-W. Hu, "Development of a patent retrieval and analysis platform - a hybrid approach," *Expert Systems with Applications*, vol. 38, no. 6, pp. 7864–7868, 2011.

[20] S. Taduri, G. T. Lau, K. H. Law, and J. P. Kesan, "A patent system ontology for facilitating retrieval of patent related information," in *Proceedings of the 6th Int. Conf. on Theory and Practice of Electronic Governance: (ICEGOV'12)*. ACM, 2012, pp. 146–157.

[21] J. Ma and A. L. Porter, "Analyzing patent topical information to identify technology pathways and potential opportunities," *Scientometrics*, vol. 102, no. 1, pp. 811–827, 2015.