ELSEVIER

World Patent Information xxx (2007) xxx–xxx

WORLD PATENT INFORMATION

www.elsevier.com/locate/worpatin

Towards content-oriented patent document processing

Leo Wanner [a,c,*], Ricardo Baeza-Yates [a,c], Sören Brügmann [b], Joan Codina [c], Barrou Diallo [d], Enric Escorsa [e], Mark Giereth [f], Yiannis Kompatsiaris [g], Symeon Papadopoulos [g], Emanuele Pianta [h], Gemma Piella [c], Ingo Puhlmann [i], Gautam Rao [e], Martin Rotard [f], Pia Schoester [i], Luciano Serafini [h], Vasiliki Zervaki [g]

[a] *Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain*
[b] *Brügmann Software, Bokeler Straße 18, 26871 Papenburg, Germany*
[c] *Pompeu Fabra University, Passeig de Circumval.lació, 8, 08003 Barcelona, Spain*
[d] *European Patent Office, Postbus 5818, 2280 HV Rijswijk, The Netherlands*
[e] *IALE Inc., Balmes, 48, 2° 1ª, 08007 Barcelona, Spain*
[f] *University of Stuttgart, Universitätsstr. 38, 70569 Stuttgart, Germany*
[g] *Informatics and Telematics Institute, 1st Km Thermi-Panorama Road, Thermi-Thessaloniki 57001, Greece*
[h] *Istituto di Cultura de Trentino, Via Sommarive, 18, 38050 Povo-Trento, Italy*
[i] *Fraunhofer Gesellschaft, Leonrodstraße 68, 80636 Munich, Germany*

Abstract

In this article, we present ongoing work on an advanced patent processing service PATExpert. The central assumption underlying PATExpert is that in order to meet the needs of the users of patent processing services, recourse must be made to the content of patent material. We introduce a content representation schema for patent documentation and sketch the design of techniques that facilitate the integration of this schema into the patent processing cycle. Two types of techniques are discussed. Techniques of the first type facilitate the access to the content of patent documentation provided in a textual format – be it by the human reader or by the machine – in that they rephrase and summarize the documentation and map it onto a formal semantic representation. Techniques of the second type operate on the content representation. At this stage, PATExpert is explored in two technology areas – optical recording devices and machine tools. The work is being carried out in the framework of an R&D-project partially funded by the European Commission.

Keywords: Patent content representation; Patent retrieval; Content extraction; Paraphrasing; Summarization; Visualization; Navigation; Valuing; PATExpert; Classification; Translation; Documentation ontologies; Knowledge base

1. Introduction

Currently, patent material is maintained in a textual format (be it in electronic or paper form). In order to retrieve, classify, interpret or assess it, the user must hypothesize how surface textual clues reflect the content. This is costly and the positive outcome is less than guaranteed. An alternative would be to specify the content representation of patent material explicitly in terms of a formal and unambiguous semantic representation. The advantages of this alternative are obvious. On the one hand, such a representation would make the examination and invalidation (by both machine and humans) much more straightforward and, on the other hand, it would facilitate retrieval, classification and interpretation of patent material. As a consequence, the patent processing techniques would be

* Corresponding author. Address: Institució Catalana de Recerca i Estudis Avançats (ICREA), Pompeu Fabra University, Passeig de Circumval.lació, 8, 08003 Barcelona, Spain. Tel.: +34 935422963; fax: +34 93542517.
E-mail address: leo.wanner@upf.edu (L. Wanner).

0172-2190/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.wpi.2007.03.008

semantics-driven, which would imply a change of the paradigm in patent processing from textual (viewing patents as text blocks enriched by ''canned'' picture material, or sequences of morpho-syntactic tokens) to semantic (viewing patents as multimedia knowledge objects). The recent advances in semantic web technologies [1] and the determination of the steering institutions to formalize the input and processing formats of patent documentation[1] speak for the implementation of the paradigm change. We are convinced that in the long run, this change will culminate in the compilation of *patent knowledge bases* (instead of or along with patent data bases).

Two strategies can be pursued to obtain a patent knowledge base: (1) extraction of the content from patent material rendered in text format and its subsequent mapping onto the content representation; (2) explicit representation of patent material in terms of a content representation (such that patent applications are already submitted as formal semantic descriptions).[2] The second strategy is more straightforward and more reliable. However, given the vast amount of patent material available in text format and taking into account that the text format continues to be the unique format of patent documentation, the first strategy is for the time being more practical. Unfortunately, a closer look at the state of the art techniques reveals that they do not fully account for the implementation of this strategy. Even recent initiatives that stress the importance of semantics and seek to develop techniques that extract the content of patent documentation for further use fall short of obtaining a true semantic representation since they rely exclusively upon surface-oriented criteria such as term frequency, term co-occurrence, and morpho-syntactic categories of the terms (i.e., noun, verb, adjective, etc.). In other words, the use of semantic web oriented notations for the resulting representation does not automatically imply that this representation is indeed a semantic (=content) representation. In order to obtain the representation of the content of a given document, ''deep'' analysis is required, and, in order to be able to make proper use of the content representation, knowledge-oriented techniques that operate on content rather than on the text surface are required.

PATExpert[3] addresses the problem of meaning representation and processing of patent documentation. The goal of PATExpert is twofold: (i) to push forward the adoption of the semantic paradigm for patent processing; (ii) to provide the user techniques for better access to the content of textual patent documentation. To achieve this goal, PATExpert focuses on the following four topics:

- content representation that is suitable for the description of inventions in several technology areas,
- semantics-based techniques that operate on the content representation of patent documentation,
- techniques that facilitate the mapping of the existing textual patent documentation to its content representation,
- techniques that facilitate a better access to the content of textual patent documentation.

In this article, we present PATExpert's general approach to these four topics.[4] The feasibility of this approach will be demonstrated within the life time of the still ongoing project for two technology areas: optical recording devices and machine tools.

The remainder of the article is structured as follows. We assume that the representation required for encoding the content of patent documentation must depend on the techniques that make use of it, and the techniques, in their turn, must reflect the needs of the users. Therefore, we start with the analysis of the needs of the users and an assessment of the consequences of these needs for the definition of the semantic representation (Section 2). Section 3 provides a sketch of the content representation framework in PATExpert. In Section 4, first the architecture of the PATExpert-service is presented and then the individual modules that realize the whole range of techniques offered by the service are discussed. Section 5, finally, contains a short summary and an outline of the future work plan within PATExpert.

2. Patent content representation from the user's point of view

The available commercial and experimental patent processing services can be assumed to reflect the central needs of the users[5] – although, obviously, only to the extent to which the state of the art allows for the implementation of a technique that meets a specific need of the user. Thus an attempt to meet some of the user needs requires a workaround. In this case, a deeper analysis of the service is required to identify the real need of the user behind the implemented technique.

In this section, we first examine the central services offered so far and draw then conclusions for the definition of an adequate content representation framework.

[1] ST36, which defines the XML-based format of patent(s) (applications), provides evidence for this determination.

[2] Obviously, an editor supporting the authoring of such semantic descriptions would be needed.

[3] PATExpert [2] is partially funded by the European Commission in its Sixth Framework Programme (FP6 028116).

[4] Note, however, that the development and implementation of the individual techniques within the showcase that will demonstrate the viability of PATExpert's approach has not yet been terminated.

[5] When we speak of users, we primarily mean professional examiners of patent applications, inventors and patent offices of research and industrial institutions and patent lawyer's agencies. These user profiles are represented either by members of the PATExpert-Consortium or by clients of members of the PATExpert-Consortium, which have been interviewed to obtain information on their needs. All user requirements are summarized in an internal working document of the project.

Please cite this article in press as: Wanner L et al., Towards content-oriented patent document processing, World Patent Informat (2007), doi:10.1016/j.wpi.2007.03.008

2.1. User's needs and available patent processing services

The available products cover the following primary patent processing services:[6] (1) abstracting, (2) classification and clustering, (3) absolute and relative (i.e., similarity) patent search, (4) extraction of content and meta-information, (5) translation, (6) linguistic (word or sentence oriented) processing, and (7) meta-information analysis. To facilitate an easier comprehension of the output, most of the techniques also offer its visualization in terms of graphics, maps, etc. To support active examination of patent spaces (such as data bases) by the user, in general, intelligent browsing (navigation) within these spaces is appropriate. Such navigation is supported by some of the services – although by far not by all.

A further service of increasing relevance – although still not fully mastered due to the complexity of the topic – is patent valuing. Furthermore, many services offer OCR correction software which improves the electronic access to printed material [3]. Also worth mention is multilingual access to patent DBs, as developed, e.g., by *Lingway* and used by WIPO under the name TACSY [4].

Let us now briefly review the services mentioned above and their adequacy to meet the corresponding user requirements.

2.1.1. Patent abstracting

As a rule, author written patent summaries do not contain information on all aspects of the patented invention that are of relevance to the reader (be the reader examiner, inventor looking for prior art or any other interested party). Also, they follow the same complex linguistic style in which patent documentation is written – which makes them difficult to read and to interpret. To get around this obstacle, several companies offer either a manual (as, e.g., *Derwent* [5]) or an automatic (as, e.g., *Questel Orbit*'s *PatFam Plus* [6]) abstracting service which provides a concise description of the invention, its novelty and use. They may also provide a description of the contained drawings, or, along with the description of the invention and the sketch of the object of the patent, its advantage and the disadvantages of the previous patents in the same area.

In general terms, what the abstracting services attempt to provide and what the user is interested in, are a concise content-oriented summary of an invention and its delimitation from other inventions.

2.1.2. Classification and clustering

In order to be usable and maintainable, patent document collections must be reasonably structured with respect to a given classification schema. The most common schema is, obviously, the IPC, but other quasi-standard classification schemata – such as the Derwent classification schema – are also used. As a consequence, existing automatic classification techniques serve to classify an unstructured patent library in terms of such reference classification schemata; cf., for instance, [7–9,3,38].

However, many users desire to define and maintain their own highly individual classification schemata. These schemata may drastically deviate from any established schema; they may be also highly heterogeneous with respect to classification criteria. Thus, one classification criterion may be the year of application, and another criterion the function principle of the inventions. The resulting schema is therefore multidimensional (in contrast to the mono-dimensional standard schemata). In order to classify an unstructured patent library in terms of this schema or assign a new patent to a specific class of this schema, new classification techniques are thus required.

The clustering task is related to the classification task. It consists in grouping patent documents with respect to specific criteria. As a rule, the criteria are of semantic nature such that patents describing similar inventions are clustered. A number of available services address the user need for clustering. For instance, *Thomson Inc.* provides two products, *Delphion* [10] and *MicroPatent* [11], which target clustering. *MicroPatent* additionally offers complex visualizations of the cluster space in terms of "concept maps". Further commercial software to be mentioned in this context is *PatentCafe.com Inc.* [12], which uses Latent Semantic Indexing (LSI) [13] for patent document clustering.

However, it remains to be seen how well the known techniques perform when the clustering criteria are fairly heterogeneous – as required by some users.

2.1.3. Patent search

The search for relevant patent documentation in DBs is one of the basic procedures any user of patent processing techniques usually performs. The majority of the search engines available for this purpose are keyword-based. Some of them incorporate a query preprocessing procedure and allow for the use of wild cards, weighting of query terms, query phrases, query expansion by using thesaurus relations, proximity search, etc.[7] However, from the perspective of the user, keywords often substitute deeper, semantic criteria. In general, we can assume that when the user carries out a content-related keyword search (in contrast to meta-information keyword search) she/he would be better served by a semantic search engine. In this context, especially the *Patent-Café* search engine [12] and *IPCentury*'s DECOPA search engine [15], which allow for semantic queries (the so-called *feature-impact pairs*), are to be mentioned.

In order to cover all types of user needs related to patent search, the following searches should be supported: (a) key-

[6] For a review of the state of the art in patent document processing, see [2].

[7] For a detailed review of the available search engines, see [2]. A contrastive assessment of the retrieval engines is also given in [14].

Please cite this article in press as: Wanner L et al., Towards content-oriented patent document processing, World Patent Informat (2007), doi:10.1016/j.wpi.2007.03.008

word-based fuzzy search,[8] (b) similarity search, (c) semantic-criteria search, and (d) image-related search.

Apart from the search engines proper, a patent search service should facilitate query and retrieved patent collection management. Patent search usually comprises a list of inquiries until the search is accomplished such that search queries often evolve step by step, becoming more and more complex. Query management must thus allow for at least the formulation of new queries based on previous queries or on patent lists retrieved before and the storage and annotation of queries. Furthermore, storage and interactive visualization of the search history must be facilitated, as well as recurrent queries that can be processed by the system with a predefined periodicity. The retrieved patent collection management should facilitate a relevance ranking not only with respect to the query, but also with respect to auxiliary information (e.g., the purpose of the query and the technical field). It should also allow for a manually structured storage of retrieved patents (the so-called "ordering in stacks").

Available patent retrieval engines offer only limited query and retrieved patent collection management – although, e.g., the Derwent search engine allows for linking previous queries with new search elements. Advanced techniques are needed that ensure a handling of the evolution and dependencies of queries and this, in its turn, interactive visualization. In fact, browsing documents can be guided by the document structure, the results of a query or a particular search task.

2.1.4. Content and meta-information extraction from patent documentation

The ultimate goal of any reader of patent documentation is to extract the content description of the invention from a given patent (application) or to obtain certain meta-information (which can be explicit or implicit) related to given patent material.

For content and meta-information extraction, *text mining* (and partially also *data mining*) strategies are needed that distil content elements and relations between content elements – either from predefined areas of a patent (application), as, e.g., the claims, or from a patent (application) as a whole. Text mining is acknowledged to be a highly promising technology for the information market in general (see, e.g., [16]), and for patent processing in particular. Some proposals (such as [17]) underline the importance of a combination of text mining and data mining techniques for patent processing. However, in practice, this task is still nearly completely left to the reader. Two of the few exceptions are the PAT-Analyser program [18,19], which can be considered as retrieving binary relations between content elements, and [39], which extracts "conceptual models", i.e., fine-grained document content structures (also known as "concept maps") from patent and other specialized document material.

2.1.5. Translation of patent material

Users ask for translation of patent material. In the past, this translation has been carried out manually by human translators. It is now more and more commonly performed by machine (aided) translation [20]. This is especially true for Japanese into English translation, but also, for instance, for translation from English to Danish [21] and from German to English [5].

However, the goal of the user is not always indeed fully translated patent material. As pointed out by Cavalier in [20], the user might want to examine whether a given patent contains information of interest or to obtain the essence of a given patent. To meet this need, instead of translation, multilingual gist generation would be more appropriate.

2.1.6. Reading aids

It is generally acknowledged that patent documentation is difficult to read and comprehend due to its very complex (linguistic) style and writing conventions. Aids that support the reader in this task are thus appreciated.

Available reading aids as offered, e.g., by *Lingway* [22] provide access to multilingual dictionaries that contain the translation, definition and synonyms of the terms. Advanced options relate terms to topics to which these terms refer (the "topics" can be mere hyperonyms of the terms in question or more abstract), highlight the occurrences of the terms in the document, indicate the frequency of the occurrence of the terms, etc. All these aids are in a sense of a passive nature; they do not spare the reader the necessity to read (fragments of) the original material. Much more appropriate would be a paraphrasing of the original material – linguistically simplified, but semantically equivalent.

2.1.7. Metadata analysis

Metadata contained in patent documentation (in particular bibliographic and legal data) constitute a valuable source of information for analysts of patent documentation. They allow for the identification of relations between patent(s) (applications) in a relatively straightforward way – e.g., in terms of inventor, holder, year of issue (respectively, filing), etc. – providing, thus, the analysts with basic material they need for their analyses. Available aids (as, e.g., Thomson's *Delphion* and *PatentLab II*) facilitate, first of all, the search and visualization of explicit relations between patents within large collections of patent material by a wide variety of customizable charts, graphs, tables, etc. However, the needs of the users go deeper. For instance, they demand the handling of complex inquiries such as "S*how me all patents from 2006 filed by an applicant whose patents prior to 2006 belong exclusively to classes X or Y*". If at all possible, existing solutions require multiple steps in order to solve these kinds of inquiries.

[8] "Fuzzy" is meant here in the morphological and spelling sense, i.e., including all word forms of a term, writing conventions, etc.

Please cite this article in press as: Wanner L et al., Towards content-oriented patent document processing, World Patent Informat (2007), doi:10.1016/j.wpi.2007.03.008

Users also demand mechanisms that enable an acceleration of metadata analysis. One such mechanism is the precompilation of often recurring types of queries.

Furthermore, users need access to *implicit* metadata information such as the number of (main/sub-) claims, inventor profile (private individual or company), whether the patent is still alive, etc. The processing of metadata is particularly important with respect to legal information. This is partially due to the lack of a world-wide standard for legal data. As a consequence, such services as *Inpadoc*,[9] which compiles legal data for patent documentation from a significant number of national patent offices, require special queries for the data of each patent office. Users demand a methodology that is able to cope with this issue.

2.1.8. Patent valuing

Automatic support of patent and patent application valuing is a service increasingly in demand. So far, valuing has been carried out mainly manually by highly trained specialists, who evaluate a whole range of general economic conditions, criteria related to the invention area, and, most importantly, criteria related to the given patent material content and object of invention. Often, complex economic models such as DCF (*Discounted Cashflow*) or OPT (*Option Price Theory*) are applied; cf. [23] for a somewhat outdated overview of patent valuing strategies. Automatic valuing techniques still focus, first of all, on superficial text criteria (such as length, occurrence of specific key words, etc.).

Due to its complexity, reliable monetary-oriented patent valuing is very difficult. The goal must be to offer as alternative index-based, i.e., relative, valuing. First commercial products and services that provide this kind of automatic valuing on statistical grounds are already available; cf., e.g., the service of the *Danish Patent Office* [24] and the product of the *PatentCafe Inc.* [12]. The service of the Danish patent office calculates the value of an application, patent or technology taking into account a series of parameters to be provided by the user. *PatentCafe* offers a fully automatic software service that calculates the score of a patent (application) using 25 parameters – among them ''claim scope breadth'', ''in-license opportunity'', ''technical sophistication'', etc. However, users have no possibility to influence the calculation.

2.2. What do the user's needs tell us with respect to content representation?

The analysis of the user needs and of the services offered so far for processing of patent information shows that the ''primary'' user needs are better served and some of them can only be met if the corresponding techniques have access to the content of patent material. For instance, high quality abstracting can only be achieved if the representation of the content of the corresponding patent material is available. Content extraction (i.e., text mining) is a task of high prominence because no accessible content representation is available. It can also be assumed that the need for translation and linguistic processing of patent material will largely disappear when content representation of patent documentation is available: starting from a content representation, concise multilingual patent descriptions can be generated in a language that the reader in question can understand. The criteria for patent classification and clustering are often meta-information oriented. However, if criteria are semantic access to content is needed. Finally, from the users' point of view, it would be desirable to also incorporate semantic information into the process of patent valuing in order to make the valuing more reliable.

In short, the advantages of a content representation for patent material are obvious. In order to account for the above needs, such a representation should have the following four central features:

1. It should abstract from concrete terms and surface-oriented linguistic structures of the sentences of a patent.
2. It must contain ontologies such that semantic links between concepts can be determined. Only then it is possible to perform, e.g., similarity or semantic search, clustering of patent documentation with respect to content criteria, and abstracting.
3. It must allow for the description of the composition, functioning, etc. of the objects of invention. In other words, it must support the general description of knowledge and the instantiation of this knowledge when a concrete invention is described.
4. It should support the link between semantic and lexical resources such that a correspondence between lexemes (=words in one of their senses) and concepts can be established – in order to facilitate, e.g., automatic content analysis.

In the next section, we present the first version of a content representation of this kind.

3. Towards a content representation for patent documentation

In modern knowledge representation, the definition of knowledge elements and relations between them is separated from their instantiation, i.e., their occurrence in a concrete knowledge space (such as patent material). The definition is dealt with in terms of ontologies; the instantiation is handled in the *knowledge base*. In the patent processing scenario, the ontologies must contain the definition of the concepts of the technological areas of interest. The knowledge base is most conveniently partitioned such that each partition contains the concept and relation instances of a single patent document.

[9] *Inpadoc* is operated by the European Patent Office and is the most comprehensive source for legal information. PATExpert uses *Inpadoc* as a source as well.

Please cite this article in press as: Wanner L et al., Towards content-oriented patent document processing, World Patent Informat (2007), doi:10.1016/j.wpi.2007.03.008

In PATExpert, the ontologies and the knowledge base are encoded in the extension of the *Ontology Web Language* (namely, OWL-DL), which is a semantic web representation language based on the *Resource Description Framework* (RDF). We use a semantic web technology in order to ensure that PATExpert's framework is compatible with other efforts in the field and to be able to use the off-the-shelf support programs offered in the meantime for semantic web technologies.

Apart from the two principal knowledge repositories, the ontologies and the knowledge base, in the patent scenario, an additional "meta" knowledge repository is of relevance: the classification schema in accordance with which patent documents are grouped into larger, homogeneous sets.

In what follows, we discuss the three knowledge repositories that form the PATExpert content representation framework.

3.1. Ontologies in patent documentation

A closer look at the knowledge in representative patent material reveals that it contains, on the one hand, common sense knowledge (as, e.g., that a *patent* is a *document* and that an *object* may consist of several *components*), and, on the other hand, knowledge specific to the patent material considered. Patent-specific knowledge can be further divided into:

(i) knowledge concerning explicit (filing date, owner, etc.) and implicit (relations to other patents, inventor and holder background information, etc.) meta-information,
(ii) knowledge on patent structure (introduction of markers for titles, headings, abstract, claims, etc.),
(iii) knowledge concerning multimedia objects within patent material,
(iv) knowledge of the domain (i.e., of the technological area and the invention),
(v) linguistic knowledge used to describe inventions in the given technological areas.

As the common sense knowledge ontology, PATExpert uses the *Suggested Upper Merged Ontology*, SUMO [25], which is an open source ontology developed by the IEEE Standard Upper Ontology Working Group. Despite its small size (the current version of SUMO contains about 1000 concepts), SUMO has been chosen for use in PATExpert because it is linked to the lexical ontology WordNet [26]. Furthermore, it is well maintained and can be expected to be periodically extended. Each type of patent-specific knowledge (cf. (i)–(v) above) is captured by an own ontology. To bridge the gap between the highly abstract common sense concept ontology and the concrete patent ontologies, we introduce the *Patent Upper Level Ontology* (PULO). This is in accordance with the recent trend in knowledge representation to minimize the abstraction discrepancy between adjacent ontology levels; cf. [27] for a proposal of a *Mid-Level Ontology*, MILO. Fig. 1 shows how the different ontologies in PATExpert are interrelated.

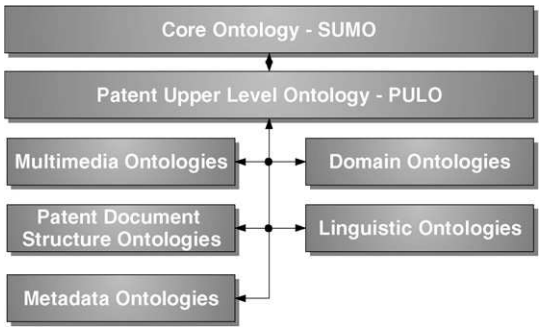

Fig. 1. PATExpert's ontologies and their mutual dependencies.

As already mentioned in the Introduction, PATExpert focuses on two technology areas: optical recording devices and machine tools. The implementation of the ontologies reflects this restriction.

3.2. The knowledge base

As already mentioned, the knowledge base contains a separate partition for each patent document. Each partition consists of the instances of the ontology elements encountered in the document and of the relations between them. Each instance is annotated by the reference to the starting and ending position of the token it is denotated by in the textual material.[10] That is, if the token consists of one word, the starting and the end position will be the same; if the token consists of several words, the starting position will be given by the position of its first word and the end position by the position of its last word.

For illustration, assume that the linguistic analysis retrieved in the patent document US6021104 has the following relation: "recorder contains optical pickup". Assume further that the position of the token "recorder" in the patent is 2_26 (i.e., second line, token number 26) and the position of the token "optical pickup" is 2_29–2_30. Then, the RDF representation of this information in the PATExpert knowledge base will be as follows:

US6021104_A:2_26-2_26 – rdf:type – ordo:recorder
US6021104_A:2_29-2_30 – rdf:type – ordo:optical_pickup
US6021104_A:2_26-2_26 – sumo:hasPart – US6021104_A:2_29-2_30

The prefix (namespace) "ordo" is used to refer to concepts in the optical recording device ontology; the prefix "sumo" to refer to concepts in the core ontology.

[10] By "token", we mean the denotation of a semantic (or, conceptual) unit. A token can be a single word (such as APPARATUS) or a complex word (such as AUDIO INFORMATION).

To preserve the access to the textual form of patent material by surface-oriented techniques, along with the knowledge base, a relational patent data base is maintained. The textual material and the knowledge base can be cross-referenced via the ontologies. Thus, each token of the textual material can be annotated with a pointer to the corresponding concept (relation) in one of the ontologies. Such a cross-referencing is of advantage when key word-based and semantic patent retrieval are used as complementary search techniques.

3.3. Patent classification schema

As mentioned in Section 2, the user often desires to classify a patent collection with respect to her/his own criteria, i.e., criteria that deviate from criteria underlying "standard" classification schemata such as the IPC. PATExpert aims to support the user in defining and modifying personalized classification schemata. More precisely, PATExpert will provide support for:

- heterogeneous classification of patent material,
- multi-dimensional classification,
- integration of the user classification schema with an official schema (such as IPC or ECLA),
- definition of multiple overlapping classification views on the same elements of the schema.

The user will be able either define a completely new schema or draw upon basic classification dimensions offered by the service. Such basic dimensions include, among others, the following dimensions:

(a) *Time (interval) of filing*: Technology does not develop continuously; it mostly evolves in "waves" or "epochs". In other words, inventions addressing a specific problem area will accumulate in specific times. Therefore, many users want to create special classification hierarchies on the basis of epochs.
(b) *Main characteristics of the invention*: The main technical characteristic of an invention is crucial to every classification of technologies. The IPC makes predominantly use of this information.
(c) *Component(s) of the invention*: Engineers in certain companies have often a constructive view on patents. Accordingly, they prefer a hierarchic classification that is oriented to the component structure of the invention.
(d) *Principle of functioning of the invention*: The technical or scientific principle on which an invention is based is occasionally desired as a criterion for patent classification.
(e) *Inventor/inventor category*: "Key inventors" are meaningful for classification since they allow for the ordering of a technological field in terms of one or a small group of inventors.
(f) *Applicant/applicant category*: Classification of patents by applicants provides a good means for assessing the IP-portfolios of all parties active in a special technology field.
(g) *IPC*: The IPC provides a solid basis for any other classification. Although the IPC is usually not fine-grained enough for a company, it is suitable for a rough classification on the top or in the middle level of a classification schema.
(h) *Legal status*: The legal status of a patent allows for assessing whether a patent is filed, examined, granted or abandoned. Furthermore the legal information allows, to a limited extent, for classification of patents into those available for licensing and those that are not.
(i) *Type of patent*: In general, patents describe a method, an artefact or a combination of both. Due to the general nature of this information, it is relevant for classification

As pointed out above, PATExpert's goal is to provide a framework that allows the user to set up her/his own classification schema – using, among others, the default categories listed above. Its goal is not to provide comprehensive classification schemata for any area. To evaluate the performance of this framework, PATExpert will develop operational classification schemata for the two technological areas it focuses on (optical recording devices and machine tools). However, given the general nature of the default categories and the possibility to easily add new categories to the classification schema, the framework will be largely technology area independent.

4. The PATExpert-service

The PATExpert-service consists of seven main modules (with each module meeting one of the specific user needs), an auxiliary module "Linguistic Workbench", which consists of such (mainly off-the-shelf) linguistic processing tools as *tokenizers*, *lemmatizers*, *taggers* and *syntactic parsers* used by various modules, and data and knowledge resources: the patent data base (DB), the patent knowledge base (KB), different types of ontologies, and a user profile DB. The latter contains such user-specific information as preference settings, intermediate search states, etc. The main modules of the service are:

1. the content and metadata extraction module,
2. the patent retrieval module,
3. the patent classification and clustering module,
4. the paraphrasing/readability improvement module,
5. the gist (summary) generation module,
6. the patent space visualization and navigation module,
7. the patent valuing and technology area watch module.

Certain modules consist of several submodules. Thus, the patent retrieval module consists of five different search engines: (1) the fuzzy keyword-based retrieval engine; (2)

Please cite this article in press as: Wanner L et al., Towards content-oriented patent document processing, World Patent Informat (2007), doi:10.1016/j.wpi.2007.03.008

the semantics-based retrieval engine; (3) the patent image retrieval engine; (4) similarity-based retrieval engine, and (5) metadata retrieval engine. An auxiliary user dialogue module facilitates the query dialogue between the user and the individual search engines.[11] Three main indices are being provided for the search engines: a word-based index, a concept-based index and an image index. The indexing is done by separate auxiliary indexing modules.

The patent classification and clustering module is accessible in two modi. The first mode allows the user to classify a selected patent document with respect to a given patent classification schema. In the second mode, the module is used by the patent retrieval module to cluster the retrieved patent collection with respect to predefined similarity criteria. It remains to be examined whether it is desirable to make the clustering mechanism also accessible via the user interface to facilitate the clustering of user selected patent document collections.

All modules can be accessed via a modular user interface which allows for a differentiated use of each technique offered within the service as well as for a combination of several techniques.

In what follows, we briefly introduce the individual modules.

4.1. Content and metadata extraction module

The goal of the content and metadata extraction module is twofold: (a) to meet the user requirements with respect to this task as identified in Section 2, and (b) to populate the knowledge base with the content of the analyzed patent material.

Content extraction is a multistage procedure which relies heavily on the linguistic analysis of the material. The first stage consists in the morpho-syntactic annotation of the material, the second stage in dependency-oriented parsing (i.e., dependency-oriented linguistic analysis) of the material and extraction of the triples of the kind <subject> <verb> <object>, <attribute> <object-denotation>, <attribute> <event-denotation>, and the third stage in the abstraction of the relations (as denoted, e.g., by verbs) that hold between content elements. The abstraction is done by using hyperonym terms of the terms in question from WordNet and generic terms from a manually compiled relation typology to which the individual verbal relations are associated. Consider, for illustration, a few generalized triples as obtained by the content extraction technique:

ordo:optical_disc sumo:hasPart ordo:lead-in
ordo:optical_disc sumo:hasPart ordo:memory
ordo:optical_disc sumo:hasPart ordo:objective_lens

The use of linguistic dependency for the identification of relations between content elements is not new. It underlies the relational paradigm of lexical resources (cf. [28] for an overview, and in particular *WordNet* [26]). It has also already been used for the analysis of patent material [29,18], but without the abstracting stage.[12] However, without abstraction, synonym and quasi-synonym names multiply the number of available relation labels. This makes it more difficult to understand the content and to see the similarity between certain relations.

The extraction of meta information in PATExpert encompasses both explicit and implicit bibliographic, legal, image, and text information.

Bibliographic data extraction triggers certain background analysis procedures – for instance, examination of the names of the applicants and inventors with the goal to determine whether the assignee is a private person, a company, group or public institution, calculation of such time spans as the time between application and request for examination or granting,[13] etc.

Legal data extraction involves the acquisition of information on legal events related to a patent (application) – including opposition filing, granting, examination request filing, etc. Given that the amount of available information varies significantly across patent collections, legal data extracted for a specific patent or patent application may be limited.

Advanced content extraction from images also targeted by PATExpert provides some meta data concerning figures contained in patent documents. In particular, the number, type (table, chart, drawing, etc.) and position of figures is determined.

Extraction of meta information from text targets a variety of different data – including the number of words in the individual sections of a patent, citations that do not appear in bibliographical data, references to companies, products and trademarks, pointers of highly frequent words to PATExpert-ontologies, etc. The claim section receives special attention. Thus, the claim structure (i.e., the dependency between the individual claims) is determined and each claim is identified as an independent or a dependent claim.

Valuable meta information is also derived with respect to the relation between patents. For this purpose, on the one hand, explicit connections given by citations and priority references are evaluated. On the other hand, implicit connections suggested, e.g., by coinciding applicant or inventor names, related inventions, etc. are acquired.

[11] As outlined in Section 4.2, query processing is conceived in PATExpert as a dialogue between the user and the machine. In the course of this dialogue, the original query of the user is iteratively refined and adjusted, based on the feedback provided by the user. Such a dialogue is considered more adequate than a series of isolated queries.

[12] Cascini's PAT-Analyzer also often suffers from an erroneous output of the linguistic tools used in the system. For instance, adjectives in nominal constructions such as "lower jaw" and "opposing flat surface" are occasionally interpreted as nouns (which leads to the introduction of a component "lower", respectively, "opposing flat" into the description of the invention).

[13] Such time spans may give valuable hints with respect to the patent strategy the patent owner follows.

Please cite this article in press as: Wanner L et al., Towards content-oriented patent document processing, World Patent Informat (2007), doi:10.1016/j.wpi.2007.03.008

4.2. Patent retrieval module

Patent retrieval in PATExpert is characterized by two prominent features: it strongly supports interactive (or, feedback-oriented) search, and it offers a number of search engines with different search criteria, which can be used either independently of each other or combined in the same search.

Interviews with users show that the user must be integrated into the search procedure because, on the one hand, it is difficult to capture all relevant search criteria in terms of a query, and, on the other hand, it is difficult to retrieve all material that is supposed to match a query. Therefore, in the PATExpert retrieval module, the user can intervene in the search at different levels of the process. For instance, she/he can

- classify the retrieved document collection in accordance with specific criteria and restrict the search to documents with specific characteristics;
- mark documents in the retrieved document collection as being relevant or irrelevant to her/his search, such that the search query is adapted to the relevant subset;
- annotate text or content passages of selected retrieved documents with search related information (this can be a simple relevance/irrelevance tag or a Boolean expression) in order to facilitate a more targeted search.[14]

As mentioned above, patent search is carried out in PATExpert by five different retrieval engines, which can be accessed via a unique interface. Given that metadata-based retrieval and keyword-based text retrieval follow the same principles, namely the principles of a full-text retrieval engine, we can speak of four engines.

As full-text retrieval engine, PATExpert uses LUCENE [30], which provides a Boolean query language and support for fuzzy syntactic search. LUCENE is also used as the basic platform of the semantic retrieval engine. Once in the operational state, the semantic retrieval engine will allow the user to search for patent documents according to semantic criteria. Such criteria may refer to the material of which an object is made, the availability of a component with a specific functionality, purpose of a component of the invention, etc.

The similarity search engine will allow the user to search for documents that are semantically similar to the document (or to the text passage) she/he provides as sample. Similarity search is especially relevant, e.g., for invalidation and for prior art description.

The image retrieval engine has a twofold purpose. It aims to associate patent figures with both textual and visual cues. The association of patent figures with text takes place by performing optical character recognition on each figure and associating the figure with its description in the patent text, thus allowing for text-based image search and retrieval. At the same time, image analysis techniques are applied to patent figures for search and retrieval based on visual similarity.

The semantic, image and similarity based retrieval engines will operate (at least partially) on the knowledge base. The SESAME RDF framework is used to support storage, querying and inference on the extracted RDF data; cf. [31] for an introduction to SESAME.

4.3. Patent classification and clustering

PATExpert aims to facilitate classification according to user-defined multidimensional classification schemata and clustering based on preselected categories. Both techniques will be available as post-processing on the retrieved patent document collection and as a stand-alone service.

The key concept of PATExpert's classification is the support of the user with respect to creation of her/his individual classification schema. User defined classification schemata facilitate the construction of elaborated and continuously revised ordering structures for large document collections; they may encompass hundreds of classes nested in a number of levels. The user is offered a set of categories upon which she/he can draw in order to compose the overall structure of the schema and to define the individual classes within the schema. The extensive number of categories within the set (including semantic categories; see Section 3.3) allows for a precise organization of any patent collection within the two PATExpert areas in terms of multiple dimensions. For classes which cannot be easily described manually by means of the available categories, a machine learning algorithm that operates on a defined set of criteria is being developed. With a user-provided training set for each class at hand, the algorithm derives first class definitions, and then applies these definitions during the automatic assignment of further patent documents to the most appropriate class(es).

The classification module is operated in two modi: the exclusive mode and the multiple mode. In the exclusive mode, a patent can be assigned only to one single class, whereas the multiple mode allows for assigning a patent to more than one class (obviously, if the characteristics of a patent comply with the specifications of a number of classes).

In addition to classification, PATExpert offers a patent document clustering mechanism. In contrast to classification, clustering does not require an explicit fine-grained classification schema. It suffices to define at least two clusters, to assign some patent documents to each of them, and to select a list of criteria in order to setup the clustering process. Then, definitions for each cluster are computed based on the previously assigned patents and the criteria chosen.

Clustering is considered as particularly useful during a retrieval session when the user needs to structure the

[14] The user can also annotate documents in the retrieved collection with maintenance related information.

retrieved documents for a better overview. Typically, a small set of clusters is sufficient.

4.4. Readability improvement of patent material

To account for the fact that the linguistic style of patent documentation is very complex and thus hard to read – even for native speakers, let alone for users with a less than perfect mastering of the language in which a patent is written – PATExpert offers a paraphrasing module. Paraphrasing can be applied to a patent document as a whole or to a text passage of the document (by selecting the passage in question in the interactive mode of the interface).

In the context of patent documentation, paraphrasing is, first of all, simplification of the linguistic style of the material. Such a simplification can be viewed as consisting of two global stages: (i) decomposition of the linguistic structures into smaller and simpler substructures taking the text and discourse structure of the given material into account, and (ii) fusion and partial transformation of the substructures by a text generator following predefined well-formedness criteria. As text generator, we use MATE [32,33]. MATE is a flexible multilingual dependency-based generator already applied to a number of other applications.

For illustration of paraphrasing, consider an original patent claim (from US6788341B2):

> A recording media storage and player unit, comprising: playback means for playing back data retrieved from a recording medium maintained at said recording media storage and player unit; communication means for obtaining from an external database continually updated expanded information associated with said data retrieved from said recording medium, but generated independently of said retrieved data, recording media or recording media storage and player unit; memory means for storing said expanded information within said recording media storage and player unit; operation means for directing the operation of said recording media storage and player unit based upon at least a portion of said expanded information to perform at least one of obtaining additional expanded information and selecting playback of said recording medium; and display means for displaying at least a portion of said expanded information when said data retrieved from said recording medium is played back.

and one of the possible results of stage (ii):[15]

> 1. A recording media storage and player unit which comprises a playback device, a communication device, a memory device, an operation device, and a display device. 2. The playback device plays back the data retrieved from the recording medium maintained at the unit. 3. The communication device obtains from an external database continually updated expanded information, which is stored within the unit by the memory device. 4. This information is associated with the data retrieved from the recording medium, but is generated independently of the retrieved data, recording media or the unit. 5. The operation device directs the operation of the unit. 6. The direction is based upon at least a portion of the expanded information to obtain additional expanded information and/or select playback of the recording medium. 7. The display device displays at least a portion of the expanded information when the data retrieved from the recording medium is played back.

Note that both stage (i) and stage (ii) involve several substages not discussed here.

4.5. Patent material multilingual gist generation

The generation of concise multilingual summaries (i.e., gists) of given patent material takes into account that a user may be interested in a short overview of patent documentation in order to assess its relevance, or may not master the language in which the documentation is written; see also Section 2. PATExpert will offer gist generation in the three official European patent languages: English, French, and German.

Technically, gist generation can be viewed from two different angles: the surface angle and the deep angle. When viewed from the surface angle, gist generation shares the first stage, i.e., the decomposition of linguistic structures, with the paraphrasing procedure. Once the original linguistic structures are split and simplified, a number of different criteria are applied to each obtained sentence structure to judge its relevance to the summary. Both content- (and domain-) oriented and linguistic (discourse, syntax, and lexis) criteria are used. In the last stage, the remaining structures are ''aggregated'', i.e., fused and regenerated – again using MATE.

In contrast to the surface-oriented gist generation, deep gist generation starts from the content representation in the knowledge base. First, relevance criteria are applied to the fragment of the KB which contains the content of the text chunk to be summarized. Concept triples that are considered relevant to the summary are passed as input to the generator, which generates from them a coherent concise summary in the language requested. For the introduction to generation, we refer the reader to [34].

4.6. Visualization and navigation

Interactive visualization is a central feature of advanced user interfaces that support efficient work with large amounts of data or detailed information.

[15] Due to the lack of space, we skip the illustration of the result of the intermediate stage (i). We also dispense with the presentation of the details of the paraphrasing procedure.

Please cite this article in press as: Wanner L et al., Towards content-oriented patent document processing, World Patent Informat (2007), doi:10.1016/j.wpi.2007.03.008

A user requirement study revealed that visualization is in particular required with respect to three types of patent related information: 1. patent metadata, e.g., bibliographic data, patent classification, citations, patent families, or legal events; 2. patent content structures, e.g., claim dependencies or implicit links between figures and text; 3. semantic relations between the patent and ontology concepts such as "is-a", "part-of", "causes", or "similar-to" relations.

In order to provide rich interaction capabilities, PATExpert's visualization and navigation module is currently realized as a browser plug-in. It uses a declarative mapping approach for mapping different information sources to a flexible visual representation model [35]. Data sources are the PATExpert knowledge base, the PATExpert database and external services such as the *Open Patent Services* [36]. For rendering and layout, we use the *Prefuse* visualization toolkit [37].

In a later stage of the development, the current browser plug-in of the visualization module will be supplemented by a web standards based approach that will use *Scalable Vector Graphics* (SVG) and *Compound Document Formats* (CDF).

Fig. 2 shows an example for browsing patents contained in a given database using the International Patent Classification (IPC). The tree map shows the IPC-groups in terms of different colours. IPC-groups that are represented by patents in the database are displayed in light grey. The user can interactively search the IPC for keywords. Matching IPC-groups are shown in dark grey. In order to minimize the necessary display space, search results are presented in a hyperbolic list. The selected item is shown in a large font whereas the other items are shown in a smaller font. Further details are provided in a separate area at the bottom. Groups that match the query and have patent documents in the database are displayed in black. The patent graph in the middle shows the available patents of a selected IPC group and a customizable set of metadata of each patent.

4.7. Patent valuing

Nearly any user of patent processing services is interested in the assessment of the profitability of patents and patent applications. In order to make this service accessible to all users and speed up the procedure, costly manual techniques currently used for the assessment as well as for technology watch should be replaced by automatic knowledge-based techniques.

However, patent valuing is a complex task, cf., for instance [40]. The value of a patent heavily depends on

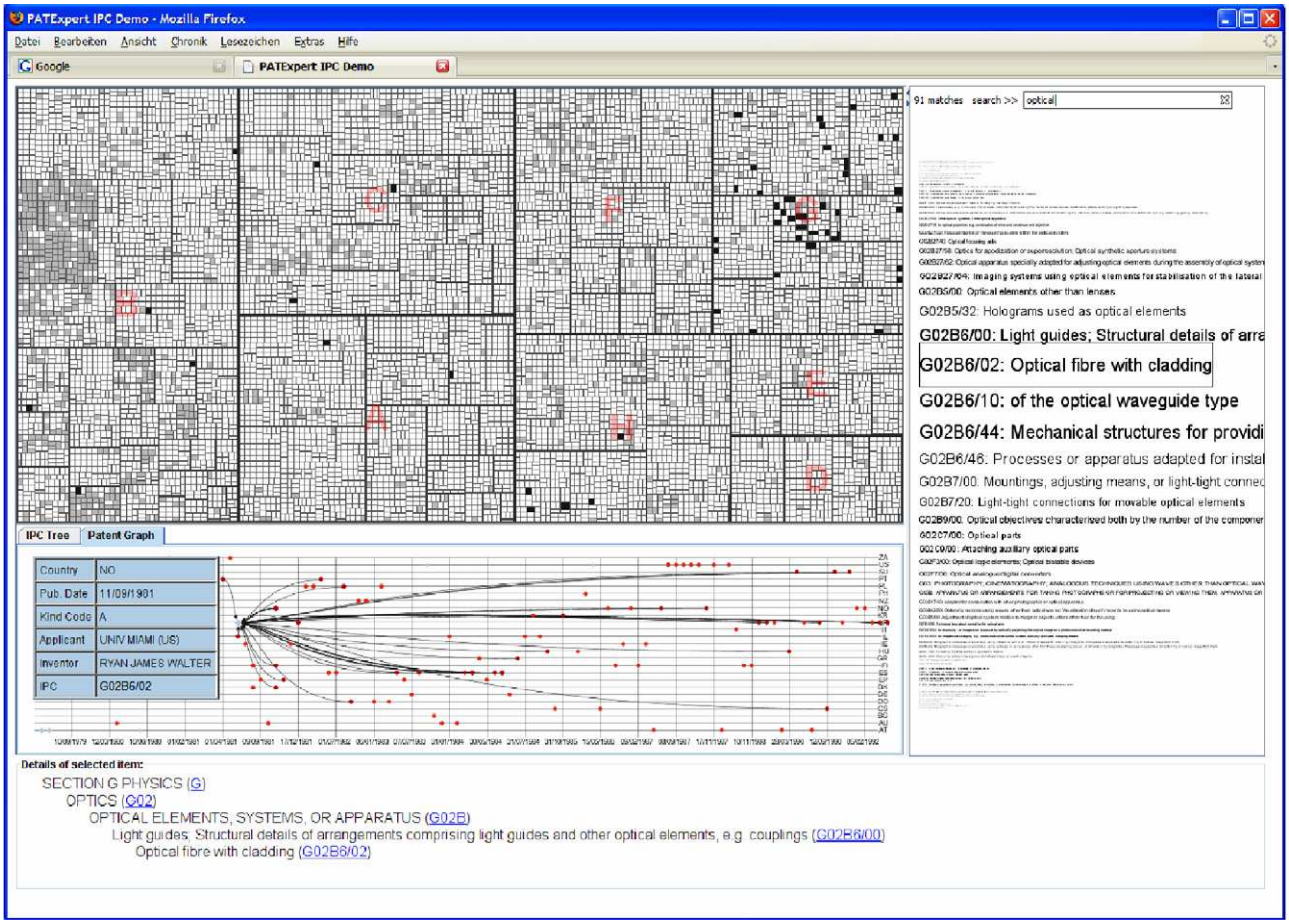

Fig. 2. Example of visualization in PATExpert.

the perspective of the user (group) and on the time at which the valuing is carried out. Therefore, PATExpert does not attempt to define a fixed monetary assessment schema.[16] Rather, it will allow the user to create her/his own assessment model. For the definition of assessment models, PATExpert provides in its *Patent Data-Warehouse* a wide range of patent-related information. The information concerns, in particular, the market and the invention. The market related information includes the market size, turnovers or market coverage, etc. Since nowadays this kind of information is not available in standardized databases, it has to be provided mainly by the user.

Invention related information includes, on the one hand, the fees (i.e., all fees accrued for filing and prolongation of a patent) and numerical data such as the number of inventors, applicants, classes, claims, number of words in the description, oppositions, citations, investment needed, stage of development or the cycle of a product, etc., and content information such as the kind of the invention, the scope of the invention, etc. The *Patent Data-Warehouse* is currently compiled partially by applying data-mining technologies to patent databases. However, as already mentioned above, some information (such as market-related information, investment data and stage of development) must be provided by the user manually.

From the information in the *Patent Data-Warehouse*, parameters that are to be taken into account to determine the value of a patent (application) are derived and, subsequently, their influence on patent valuing is determined. Several methods to analyse the parameter influence have been used – among them, machine learning techniques.

PATExpert's Patent Assessment aims to provide both absolute and relative valuing (i.e., ''patent scoring''). An example for absolute patent valuing is cost-based valuing, which is done by defining an assessment model that includes comprehensive information about costs which usually accrue with regard to patents like costs for translation, fees for application and renewal, and agent fees. In contrast, relative valuing relies upon the comparison of a set of features of a given patent chosen by the user with the features of (a set of) other patents.

A further objective within PATExpert is the development of knowledge-oriented assessment techniques to detect technology areas with high ''patent generating potential''. The aim is to provide technical trends and assessment of technologies as well as their development. This task is expected to be handled by the same techniques as patent valuing.

5. Summary

The PATExpert-service presented in this article reveals three features that we consider essential for any next generation patent processing service. First, it is semantically oriented in that it heavily draws upon the content of patent material rather than only on its textual surface. Its semantic techniques (be it the content distillery, the semantic patent search, or any other technique) are based on a state-of-the-art semantic web content representation technology. Second, it offers a unique range of techniques integrated into a single service that covers all major user requirements. Third, the strategies applied by the individual techniques are transparent and nearly all techniques allow for interaction and intervention by the user. The user can thus comprehend how the results are obtained and tune the service accordingly.

The work in PATExpert is still going on. A prototypical operational version of the service is expected to be available by mid 2007. The service will reach its full functionality by mid 2008.

[16] It can thus dispense with the use of financial information (such as stock market data) used by most models; cf. [40] – even more so because this information is usually available only for users with large patent portfolios.

References

[1] Berners-Lee T, Hall W, Hendler JA, O'Hara K, Shadbolt N, Weitzner DJ. A framework for web science. Found Trend Web Sci 2006;1 (1).
[2] http://www.patexpert.org (includes a review of the state of the art in patent processing, at 'publications', then 'D8.1').
[3] Hull D, Ait-Mokhtar S, Chuat M, Eisele A, Gaussier É, Grefenstette G, et al. Language technologies and patent search and classification. World Patent Inform 2001;23 (3):265–8.
[4] http://claims.wipo.int/tacsy.
[5] http://www.derwent.com.
[6] http://www.questel.orbit.com/en/prodsandservices/pluspat.htm#fam-patplus.
[7] Larkey LS. Some issues in the automatic classification of US patents. In: AAAI-98 Working Notes; 1998.
[8] Fall CJ, Benzineb K. Literature survey: issues to be considered in the automatic classification of patents, WIPO; 2003.
[9] Tikk D, Biró G, Yang JD. Experiments with a hierarchical text categorizer. In: Proceedings of the IEEE international conference fuzzy systems; 2004. p. 1191–6.
[10] http://www.delphion.com.
[11] http://www.micropat.com.
[12] http://www.patentcafe.com/.
[13] Berry MW, Fierro RD. Low-rank orthogonal decompositions for information retrieval applications. Numer Linear Algebra Appl 1996;1 (1):1–27.
[14] http://www.infonortics.com/chemical/ch04/slides/lambert-new.pdf.
[15] http://www.ipcentury.de.
[16] Grimes S. The developing text mining market. A white paper prepared for the text mining summit 2005, Boston, June 7–8, 2005. www.TextMiningNews.com.
[17] Trippe AJ. The importance of being Ernest. Why gathering and cleaning all the relevant data matters for patent analysis. In: Talks presented at the 2005 PIUG north-east meeting, Iselin, NJ; 2005.
[18] Cascini G, Rissone P. PAT-analyzer: a tool to speed-up patent analyses with a TRIZ perspective. In: Proceedings of the ETRIA world conference – TRIZ future; 2003.
[19] Cascini G, Neri F. Natural language processing for patents analysis and classification. In: Lecture notes in computer science. Heidelberg: Springer; 2004.
[20] Cavalier T. Perspectives on machine translation of patent information. World Patent Inform 2001;23 (3):367–71.
[21] http://www.zacco.com.
[22] http://www.lingway.com.
[23] Pitkethly R. The valuation of patents: a review of patent valuation methods with consideration of option based methods and the

potential for further research. The Judge Institute, Working Paper Series. Cambridge WP 21/97; 1997.
[24] Nielsen P-E. Evaluating patent portfolios – a Danish initiative. World Patent Inform 2004;26 (2):143–8.
[25] Niles I, Pease A. Towards a standard upper ontology. In: Proceedings of the 2nd international conference on formal ontology in information systems (FOIS-2001), Ogunquit, Maine, 17–19; 2001.
[26] Fellbaum C. WordNet: an electronic lexical database. Cambridge (MA): MIT Press; 1998.
[27] Niles I, Terry A. The MILO: a general-purpose, mid-level ontology. In: Proceedings of the 2004 international conference on information and knowledge engineering, Las Vegas, NE; 2004.
[28] Evens M. Relational models of the lexicon. Representing knowledge in semantic networks. Cambridge: Cambridge University Press; 1988.
[29] Sheremetyeva S, Nirenburg N. Interactive knowledge elicitation in a patent expert's workstation. IEEE Comput 1996;29:57–63.
[30] http://lucene.apache.org/java/docs.
[31] Broekstra J, Kampman A, Harmelen Fv. Sesame: a generic architecture for storing and querying RDF and RDF schema. In: Proceedings of the first international semantic web conference (ISWC2002). Sardinia, Italy: Springer; 2002.
[32] Bohnet B. Textgenerierung durch Transduktion linguistischer Strukturen, Ph.D. Thesis, University of Stuttgart, Germany; 2005.
[33] Bohnet B, Langjahr, A, Wanner L. A development environment for an MTT-based sentence generator. In: Proceedings of the first international conference on natural language generation, Mitzpe Ramon, Israel; 2000.
[34] Bohnet B, Wanner L. On using a parallel graph rewriting grammar formalism in generation. In: Proceedings of the 8th European natural language generation workshop at the annual meeting of the association for computational linguistics, Toulouse, France; 2001.
[35] Heer J, Agrawala M. Software design patterns for information visualization. IEEE transactions on visualization and computer graphics (TVCG) 2006;12 (5).
[36] http://ops.espacenet.com/.
[37] Heer J, Card SK, Landay JA. Prefuse: a toolkit for interactive information visualization, CHI 2005, Human Factors in Computing Systems; 2005.
[38] Krier M, Zaccà F. Automatic categorisation applications at the European patent office. World Patent Inform 2002;24 (3):187–96.
[39] Hui B, Yu E. Extracting conceptual relationships from specialised documents. Data Knowledge Eng 2005;54 (1):29–55.
[40] John Sykes J, King K. Valuation & exploitation of intellectual property & intangible assets. Emis London: Professional Publishing; 2003.

Leo Wanner holds a Diploma in Computer Science from the University of Karlsruhe, Germany and a Doctorate in Linguistics from the University of the Saarland, Germany. After occupying positions at the German National Centre for Computer Science, University of Stuttgart and University of Waterloo, he is currently ICREA research professor at the Technology Department of the Pompeu Fabra University, Barcelona.

Please cite this article in press as: Wanner L et al., Towards content-oriented patent document processing, World Patent Informat (2007), doi:10.1016/j.wpi.2007.03.008