Scientometrics (2012) 91:645–670
DOI 10.1007/s11192-012-0635-7

A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis

Jan M. Gerken · Martin G. Moehrle

Received: 6 June 2011 / Published online: 31 January 2012

Abstract Given that in terms of technology novel inventions are crucial factors for companies; this article contributes to the identification of inventions of high novelty in patent data. As companies are confronted with an information overflow, and having patents reviewed by experts is a time-consuming task, we introduce a new approach to the identification of inventions of high novelty: a specific form of semantic patent analysis. Subsequent to the introduction of the concept of novelty in patents, the classical method of semantic patent analysis will be adapted to support novelty measurement. By means of a case study from the automotive industry, we corroborate that semantic patent analysis is able to outperform available methods for the identification of inventions of high novelty. Accordingly, semantic patent information possesses the potential to enhance technology monitoring while reducing both costs and uncertainty in the identification of inventions of high novelty.

Keywords Novelty measurement · Semantic patent analysis · Inventive progress · Technology monitoring · Citation analysis · Classification analysis

Mathematical Subject Classification (2000) 68U15

JEL Classification O32 · O34

Introduction

Technology proves to be a central driving agent of economic success, as technological change is able to influence the competitive environment to a considerable extent as well as in different directions. On the one hand, technological change is the most powerful driver of growth (Sood and Tellis 2005), offering companies the opportunity to gain advantages in competition. On the other hand, technological change may also result in the decline of long-established institutions (Lichtenthaler 2004).

J. M. Gerken (✉) · M. G. Moehrle
Institute of Project Management and Innovation (IPMI), University of Bremen,
Wilhelm-Herbst-Str.12, 28359 Bremen, Germany
e-mail: jgerken@uni-bremen.de

Even though the definitions of technological change, e.g. as introduced by Achilladelis et al. (1990), Christensen and Overdorf (2000), Chandy and Tellis (1998) and Anderson and Tushman (1990), differ in some respects (e.g. as concerns customer benefits), they share the notion that technological change is enforced by inventions which are characterized by a high degree of novelty. For this reason, companies are confronted with the challenge not only to create inventions of high novelty, but also to identify competitive inventions of high novelty at an early stage, as they may provide the basis for a prospective technological change in the competitive environment.

At present the greatest challenge does not lie in the availability of information about inventions of high novelty, but rather in the assessment thereof, as companies are confronted with the problem of information overflow (Bergmann et al. 2008). Information overflow causes the monitoring of inventions by manual efforts to be extremely expensive. In order to meet this challenge, several articles address the identification of inventions of high novelty by way of a computer-based patent analysis (see Table 1 with their characteristics). Major shortcomings are (i) the level of analysis, which is not often valid on the basis of a single patent, and (ii) the use of linguistic variables for novelty measurement, which are mostly based on citation links and keywords and neglect functional relationships in texts.

To overcome these shortcomings we will develop a specific form of semantic patent analysis. Our approach focuses on functional structures in patents, which may be extracted by using semantic knowledge. In our approach we are going to focus on three research queries:

- How can semantic patent analysis be adapted to measure novelty?
- How well does semantic patent analysis perform compared with existing methods in the identification of inventions of high novelty?
- What implications are to be expected in theory and management practice?

In Theoretical background, we shall discuss the theoretical background of novelty measurement. We will adapt semantic patent analysis for novelty measurement in Semantic patent analysis for identifying inventions of high novelty. In particular, we will discuss issues of variable measurement and novelty calculation. Using a case study from the automotive industry in Data and research design, we will examine in detail, whether semantic patent analysis is able to outperform available methods in identifying inventions of high novelty. The results, which will be discussed in Results, may serve to confirm our assumptions: semantic patent analysis offers an opportunity to identify inventions of high novelty. However, there is still room for improvement. Limitations of this approach and the case study will be discussed in Discussions and conclusions, which also refers to theoretical and managerial implications.

Theoretical background

Term and dimensions of novelty

Novelty is in fact a highly complex concept. In analogy to innovations, we differentiate between recombinant and pioneering novelty. Recombinant novelty dates back to Schumpeter, (Schumpeter 1934) who interpreted novelty as new combinations of more or less known resources (see also Hargadon and Sutton 1997; Fleming and Sorenson 2001; Fleming 2001; Fleming et al. 2007; Schoenmakers and Duysters 2010). In contrast, pioneering novelty can be found in innovations made up of elements, which are not drawn from the 'made' world (Fleming 2001). We do not consider recombinant and pioneering novelty as a dichotomy but as poles of a continuum, as the combinations of resources used

Table 1 Prior methods for the analysis of novelty based on patents

Approach(es)	Novelty indicator	Novelty measurement	Focus of measure	Shortcomings
Achilladelis et al. (1987)	Patents	Counting of patents in technological field	S-curves	Novelty measurement only on a macro level but not on the level of a single patent
Achilladelis et al. (1990)	Patents	Counting of patents in technological field	S-curves	Inadequate for the identification of single patents with a high degree of novelty
Andersen (1999)	Patents	Counting of patents in technological field	S-curves	
Frietsch (2007)	Patents	Counting of patents in technological field	S-curves	
Haupt et al. (2007)	Patents	Counting of patents in technological field	S-curves	
Ahuja and Lampert (2001)	Patents and US patent classification	Counting of new US patent classes entered by a firm in the previous three years	Novel technologies	
Trajtenberg et al. (1997)	Backward citation	Counting of backward citations to scientific literature	Basicness	70% of all patents are cited less often than three times, citation analysis focuses exclusively on bibliographic information and in many cases ignores the description section of patents (Lee et al. 2009a), on average, the time lags between citing and cited patents are more than ten years (Hall et al. 2001), the references stated in a patent are subject to strategic decisions, as patent owners can also file citations for their own patents (USPTO 2007)
Trajtenberg et al. (1997)	Backward citation	Weighting of backward citation by incorporating a three-digit classification scheme	Distance in technology space	
Ahuja and Lampert (2001)	Backward citations	Number of patents without backward citations	Pioneering technologies	
Dahlin and Behrens (2005)	Backward citation	Similarity structures between backward citations	Novelty of radical inventions	
von Wartburg et al. (2005)	Backward citations (multi-stage)	Multi-stage backward citations weighted by the inverse number of backward citations in a patent	Technological value added	
Schoenmakers and Duysters (2010)	Backward citations	Number of backward citations	Novelty of radical inventions	

Table 1 continued

Approach(es)	Novelty indicator	Novelty measurement	Focus of measure	Shortcomings
Kim et al. (2008)	Keywords (experts' choice)	Clustering based on keyword existence matrix, patent mapping	Emerging technologies	Dependent on experts' judgement of keywords, limited to single words without contextual information
Lee et al. (2009b)	Keywords (computer-based approach with expert screening)	Distinguishing between and counting of emerging, declining, core and established keywords in consumption with patent mapping	Evolutionary directions of technologies	
Lee et al. (2009a)	Keywords (computer-based approach with expert screening)	Distinguishing between and counting of adjacent patents with emerging and declining keywords in consumption with patent mapping	Discovering new technology opportunities	
Lee et al. (2011)	Keywords (computer-based approach with expert screening)	Similarity calculation between keyword vectors and analysis of presence and absence of keywords	Monitor trends of technological change	
Choi et al. (2011)	SAO networks	Density and cohesion of SAO sub-networks	Novel technology functions	Evaluation of novel technology functions as elements of patents in a technological domain but no evaluation on the level of patents
Yoon and Kim (2011b)	SAO structures	Similarity measurement based on SAO structures, construction of patent networks based on similarity, calculation of density and cohesion of patent sub-networks	Novel patent clusters	Novel inventions are identified on the level of sub-domains but not on the level of single patents

for recombinant novelty may come from a narrow field, but also from wider fields and then approach pioneering novelty.

In this article, we are focussing on the measurement of novelty of patents for the assessment of technologies. As we will analyse novelty especially in patents and their respective inventions, it is necessary to highlight three particular aspects of novelty in this context. First of all, Witt (2009) defines novelty "as something that was unknown before a particular point in time that, hence, was discovered or created at that time". If novelty is defined as something that had hitherto been unknown, this implies that novelty is also directly related to what was known before (see also Dahlin and Behrens 2005). In the context of patents "what was known before" is usually referred to as "state of the art" or "prior art". Comprehending what the novel impulse of an invention really is, relies on a diligent selection of prior art. Secondly, in patents novelty is combined with so-called non-obviousness (USPTO 2007). A patent should not be granted "if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains" (§ 103). In our article we will consider novelty and non-obviousness as a joint phenomenon. In a more advanced approach it could be useful to model non-obviousness as a moderating effect of the relationship between a patent's novelty and its impact on other variables like patent value. Thirdly, in the context of technological change the novel item or method should be useful, as we are not interested in detecting changes in fashion or similar phenomena. As concerns patents we shall take this aspect for granted, because the patent law recommends that an invention be new, inventive and useful in order to get filed as a patent (USPTO 2007).

Available methods for the identification of novel technologies in patents

Patents are generally considered to have various advantages over other information sources, such as their coverage and extent (Fendt 1988; Trajtenberg et al. 1997; Debackere et al. 2002; Knight 2004), availability (Debackere et al. 2002) and the precondition of industrial applicability or usefulness (Granstrand 2000), to mention only a few. Due to the obvious merits of patent information, several methods for determining novelty have already been established: On the linguistic level, they are based on patent existence, citations and keywords.

The relationship between patent existence (measured in counts) and the technological life cycle has been analysed in various studies, and in many cases an S-shaped curve or a double S-shaped curve was observed (e.g. Achilladelis et al. 1987; Achilladelis et al. 1990; Andersen 1999; Frietsch 2007; Haupt et al. 2007). Even though S-shaped curves facilitate the anticipation of technological change on a macro level, methods based on patent counts are not appropriate for the identification of patents with a high degree of novelty, as the counting of patents relies on the assumption that patents are equally significant This assumption is contradictory to the existence of 'key patents' and 'basic patents' (Debackere et al. 2002). 'Basic patents' are patents that protect a fundamental breakthrough in a technical field (Grupp 1997). Consequently, basic patents are characterized by a high degree of novelty. Key patents' constitute the basis for further patents. They are of outstanding value (Debackere et al. 2002). Hence, patent counts are inadequate for the identification of single patents with a high degree of novelty.

Patent citation analysis has often been considered able to overcome these drawbacks and provide an appropriate index for estimating the value and the importance of patents (Trajtenberg 1990). Patent citation analysis is frequently used for the analysis of patents and gaining insight into a technological field (Debackere et al. 2002; Yoon and Park 2004). The application of citation analysis in this context has its origin in the bibliometric studies

of scientific publications (Trajtenberg 1990) and has been applied to analyse the knowledge flow between technological sectors (e.g. Han and Park 2006; Park et al. 2005) and geographical regions (e.g. Jaffe et al. 1993; Jaffe and Trajtenberg 1999), to discover patterns in patent literature (Lee et al. 2009a) and to assess the value (Trajtenberg 1990) as well as the basicness of patents (Trajtenberg et al. 1997). Backward citations are generally applied as an indicator of novelty (Reitzig 2003b).

The available methods that are based on patent citation can be divided into three categories: Methods, (i) which simply count backward citations, (ii) which qualify and count backward citations and (iii) which compare the structures of backward citations.

(i) Ahuja and Lampert (2001) as well as Schoenmakers and Duysters (2010) refer to the quantity of backward citations for a measurement of novelty. As patents are bound to state their technological origins and limitations by citing all preceding patents on which they are based (Ahuja and Lampert 2001) and which may have an impact on the patentability of the patent (USPTO 2007), a quantity of zero backward citations indicates that there are no closely related preceding patents. This also means that a low number of backward citations suggests a high degree of novelty. In contrast, a high number of backward citations indicates a close relation to preceding patents and to the state of the art (Rost 2010), which may be an indication of a low degree of novelty. In contrast to this, Wartburg et al. (2005) found no evidence to support a relationship between counts of backward citations and the value added by a patent family.

(ii) An enhancement to the count of backward citation is to be found in methods which qualify references. This can be achieved by distinguishing between certain types of references, e.g. patents and scientific literature, through multi-stage citation analysis (von Wartburg et al. 2005) and also by incorporating patent classifications (Trajtenberg et al. 1997).
Methods that distinguish between certain types of reference rely on the fact that patent references do not only include backward citations to other patents but also to other types of publications, such as journals, books and proceedings. Trajtenberg et al. (1997) suggest that basic research is closely linked to scientific literature. Hence, scientific linkages of patents may indicate a high degree of novelty. Wartburg et al. (2005) present a multi-stage citation approach to the measurement of inventive progress. In this approach, the authors introduce a weighting of references by the inverse number of references listed in a patent. Based on this weighting, several measures were calculated, relating to the technological value added. A qualification of references based on classification schemes can be found in an approach by Trajtenberg et al. (1997). In this approach, 'technological distance' (TECHB) is computed by incorporating a three-digit classification scheme, as described by Hall et al. (2001). Hall et al. (2001) summarized the US patent classes in a three-digit classification scheme consisting of categories and sub-categories, which supports a comparison of technological domains on several levels. Based on this classification scheme, the average distance between a certain patent and the cited patents is calculated. The distance between the citing patent and the cited patent is (a) set to zero if both patents are in the same 3-digit class, (b) set to 0.33 if cited and citing patent class are in the same 2-digit class, (c) set to 0.66 if cited and citing patent class are in the same 1-digit class and (d) set to 1 if citing and cited classes are not in the same class. TechB is calculated as the average distance to all cited patents.

(iii) Dahlin and Behrens (2005) applied an approach which is closely related to bibliographic coupling as introduced by Kessler (1963). Dahlin and Behrens (2005) assume that inventions of high novelty differ from older patents in terms of their citation structure. Dahlin and Behrens (2005) do not only count the overlapping

references of two patents, instead they calculate the similarities in the citation structures of two patents by dividing the number identical references by the total number of references in both patents. Thus, the authors ensure that the results are not influenced by the total number of references.

On the whole, citation analysis represents a valuable enhancement to the field of patent analysis. However, it also has several drawbacks: (i) the scope of information is limited, as citation analysis focuses exclusively on bibliographic information and ignores the description section of patents (Lee et al. 2009a), (ii) patent citation merely indicates individual links between patents and is thus not suitable for an analysis of overall relationship (Yoon and Park 2004) and, most importantly, (iii) the references stated in a patent are subject to strategic decisions, as patent owners can also file citations for their own patents (USPTO 2007). This may influence the number and choice of references.

With respect to the limitations of patent citation analysis, some keyword-based methods for patent analysis have already been introduced by various authors (e.g. Kim et al. 2008; Lee et al. 2009a, b). These methods involve a comparison of the occurrence of keywords in patents, for example by means of measuring term frequency (Li et al. 2009) or extracting and classifying keywords (Yoon and Park 2005). Most commercially available patent analysis tools have long been relying on the measurement and comparison of term frequency (Trippe 2003). But especially in patent analysis, the extraction and comparison of single words is often too unspecific for a detailed analysis. Therefore, the extraction of multi-words is preferable. Multi-words are more precise in representing the contents of a patent (Tseng et al. 2007).

SAO-based patent analysis

To overcome the drawbacks of the preceding methods, approaches based on SAO-structures (Subject-Action-Object-structures) deserve special attention. SAO-structures are also referred to as problem–solution-structures, as the combination of object and action incorporates the problem or required function, while the subject provides the solution (see Moehrle and Geritz 2007; Bergmann et al. 2008), and thus represent the functional relationship of an invention (Cascini et al. 2004; Choi et al. 2011; Yoon and Kim 2011b). Accordingly, SAO-structures give insights into the relationship between components of a technical system (Yoon and Kim 2011b). For example, the SAO structure "central differential (S)—comprise (A)—sun gear (O)" clearly shows, that the central differential involves a sun gear and consequently represent structural properties of the invention. Thus, SAO-structures are expected to be a high-quality representation of patents' technological contents. SAO-based patent analysis, generally labeled as semantic patent analysis, has already been used for several tasks, e.g. inventor profiling (Moehrle et al. 2005), M&A decision support (Moehrle and Geritz 2007), patent infringement analysis (Bergmann et al. 2008; Park et al. 2011), technology monitoring (Gerken et al. 2010b) and the identification of technological trends (Choi et al. 2011; Yoon and Kim 2011b). Yoon and Kim (2011a) apply SAO-analysis to detect signals of new technological opportunities. For this purpose, outlying patents are identified by calculating the distance to their *k*-nearest neighbours.

Semantic patent analysis for identifying inventions of high novelty

In order to overcome the drawbacks of the available methods, we are going to introduce the use of semantic patent analysis for identifying inventions of high novelty. In this approach, we are relying on semantic patent analyses that have been developed for other tasks. We

will extend and apply available methods of semantic patent analysis to calculate degrees of novelty on the level of single patents.

The process of semantic patent analysis in general

Our method consists of four successive steps, which mainly focus on the extraction of information and the analysis of the extracted information: First of all, semantic structures have to be identified in and extracted from patent texts. Secondly, a specific domain- or situation-related linguistic analysis has to be performed in order to ascertain the relevance of the extracted semantic structures for the given field of interest. Thirdly, the similarity measurement takes place, based on the consideration of the significance of specific structures within the extracted structures. Fourthly, the resulting similarity matrices are used for the calculation of the novelty of inventions in patents. These four steps will be described in greater detail in the subsequent sections.

Natural language processing: the extraction of semantic structures

The first step of semantic patent analysis for identifying inventions of high novelty in patents involves the extraction of semantic structures. The extraction of semantic structures aims at the analysis of a vast quantity of textual data, which includes the preparatory processing of this textual data, such as natural language text, for further analysis. Hence, the basis of semantic patent analysis lies in a syntactic analysis of the patents' full texts, which can be achieved by use of part-of-speech tagging. This means that syntactic information is added to the text.

The gap between syntactical analysis and semantic analysis is bridged by focussing primarily on the technical aspects of inventions, as represented by functional aspects of the invention. In order to do this, we extract SAO-structure from patents. The extraction of SAO-structures[1] can be achieved with the Knowledgist™ software by Invention Machine.[2] As characteristic SAO-structures might occur more frequently than less characteristic SAO-structures, we extract SAO-structures and also count the frequency of these SAO-structures.

Specific domain- or situation-related linguistic analysis

In step two, the identification of semantic structures is followed by a linguistic analysis of domain- and situation-related elements. The necessity of employing filter modules is caused by the problem of (i) synonymy and (ii) stop words.

(i) The problem of synonymy arises from the fact that different words may represent the same general meaning in a specific domain or situation. For example 'motor vehicle' and 'automotive vehicle' can be seen as general synonyms, as they both describe "a self-propelled wheeled vehicle that does not run on rails" (Princeton University 2006). As a matter of fact, general synonyms are very rare, as most synonyms arise from the context. In addition to this, words are often related to each other in hierarchical terms. Hence, the

[1] Although we will focus on SAO-structures in this article, it is noteworthy to show an alternative: Word *n*-grams. Word *n*-grams can be extracted with or without regard to syntactical classes and functions. Extracting *n*-grams regardless to syntactical functions cause loss of syntactical information. Nevertheless, *n*-grams still have semantic information. *n*-grams take the co-occurrence of words into account and hence, highlight a relationship between these words on the content level, as they show that *n* words co-occur close together in a patent.

[2] For further information on Invention Machine and the Knowledgist™ see invention-machine.com.

necessity to apply a domain- and situation related linguistic analysis is based on the fact that words may be seen as synonyms in a specific context, though not in general.

(ii) Furthermore, some extremely common words are of little value for the selection and the comparison of patents; they are thus referred to as stop words (Manning et al. 2008). Comparable to the problem of synonymy, some words may be identified as stop words in a specific context but not in another. Hence, we distinguish between domain-specific and general stop words. Domain-specific stop words are extremely common in the respective technological field and thus of little value for the discrimination of patents. In contrast, we will call words that occur in practically every text or sentence, such as 'the' and 'a', general stop words.

Generating similarity matrices

Prior to the measurement of novelty, a similarity measurement is carried out to highlight those aspects of the invention that have already been described in older patents. The statistical comparison of the semantic structures of patents within a given patent set results in a matrix of similarity values. According to Moehrle (2010) a two-step approach to establishing textual similarities between patents can be applied: Firstly, a variable measurement is concerned with a pairwise count of overlapping semantic structures between patents and the overall count of semantic structures within each patent. Secondly, applying similarity coefficients to the variables leads to similarity measures.

First of all, a variable measurement is required. For this purpose, complete linkage is applied, as structures are compared without any modifications of their quantity, e.g. removing of semantic structures which occur twice (see Moehrle 2010). The identical structures of two or more patents are linked by means of directed connections. The count of links between two patents represents the overlap of these patents, which is typical of single-sided measurement. As we are focussing on novelty, we add a chronological dimension to the variable measurement. Due to this, the variable measurement now involves forward and backward looking variables—comparable to patent citation analysis. Forward looking and backward looking overlapping semantic structures and the semantic structures of each patent are measured according to their availability (see Fig. 1).

(i) As long as only patent i exists, the semantic structures in patent i (c_i) can be counted.

(ii) Once patent j occurs, the semantic structures of patent j (c_j) and the semantic structures of patent i, which are also part of patent j ($c_{i(j)}$), and the semantic structures of patent j, which are also part of patent i ($c_{j(i)}$), can be counted.

(iii) Once patent k occurs, the semantic structures of patent k (c_k), the semantic structures of patent i, which are also part of patent k ($c_{i(k)}$), the semantic structures of patent j, which are also part of patent k ($c_{j(k)}$), the semantic structures of patent k, which are also part of patent i ($c_{i(i)}$), and the semantic structures of patent k, which are also part of patent j ($c_{k(i)}$), can be counted.

(iv) Analogous to steps i–iii, for each occurring patent, the semantic structures and semantic structures overlapping with older patents can be counted.

$c_{i(j)}$, $c_{i(k)}$ and $c_{j(k)}$ are forward looking variables. Forward looking variables indicate the share of semantic structures that have been retained throughout successive patents. For example, $c_{i(j)}$ equal 3 indicates that three semantic structures of patent i also occurred in patent j. In contrast, $c_{j(i)}$, $c_{k(i)}$ and $c_{k(j)}$ are backward looking variables. For example, $c_{j(i)}$ equal 4 indicates, that four semantic structures of patent j were already mentioned in patent i.

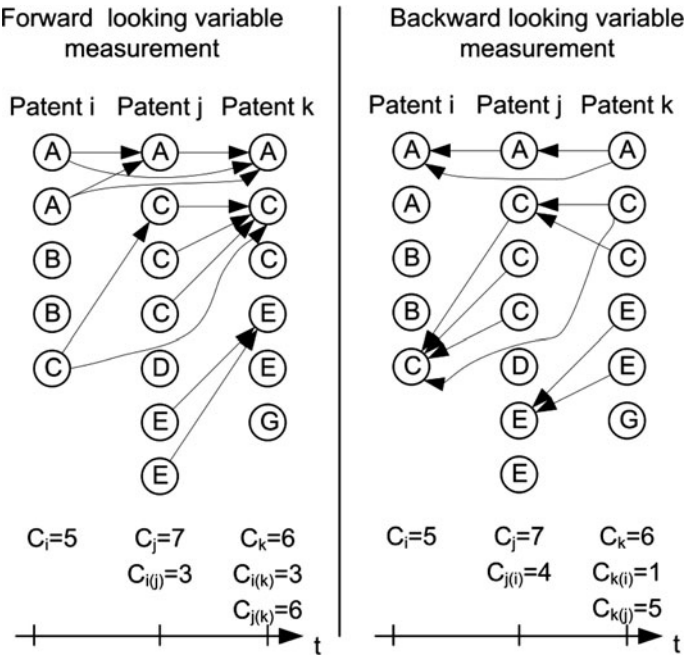

Fig. 1 Example of single-sided variable measurement based on Moehrle (2010), with the addition of a chronological dimension. Ⓐ to Ⓖ represent the semantic structures of patents *i*, *j* and *k*. *Arrows* link identical semantic structures in patents *i*, *j* and *k*

On the basis of the introduced variables, single-sided-inclusion is a useful similarity coefficient for measuring the share of a patent that is consistent with preceding patents (Eq. 1):

$$\text{Backward looking single sided inclusion} = c_{j(i)}/c_j \tag{1}$$

where $c_{j(i)}$ is the number of semantic structures of patent *j*, which are also part of patent *i*, and c_j is the number of semantic structures of patent *j* (Moehrle 2010).

Novelty calculation

In this article, novelty measures will be calculated on the basis of a similarity matrix. As we have already outlined that novelty can only be determined in relation to established technologies, a novelty calculation of the inventions described in the patents of the two patent sets by means of semantic patent analysis is based on the similarity between a patent and all preceding patents. The maximum similarity of a patent to the preceding patents can be regarded as its "oldness": In contrast to this, we considered novelty as the particular share of a patent that does not resemble the preceding patent in terms of maximum similarity. Accordingly, we calculate the novelty of a patent in proportion to preceding patents by subtracting the maximum similarity of a patent to its preceding patents from one (Eq. 2):

$$N_i = 1 - \max(s_{i(n)}) \text{ for all } n < i \tag{2}$$

where N_i is the novelty of patent *i* and $s_{i(n)}$ is the similarity of patent *i* to each patent *n* filed prior to patent *i*.

Data and research design

In order to test the method of semantic patent analysis for the identification of inventions of high novelty, we compiled a data set from the automotive industry. For companies in this industry, competitiveness depends upon the ability to innovate as much as on the ability to defend their particular market niche. In this context, the four-wheel drive technology is a noteworthy technology seen from a market perspective but also from a technological point of view. Especially the leading manufacturer of four-wheel drive vehicles SUBARU[3] seems to be an interesting example, as the market for four-wheel drives has expanded steadily in recent years and is expected to grow still further (Kurmaniak 2008; Pope 2009; Stockmar 2004). In the subsequent sections we will first describe the compilation of patent data, followed by the analysis and preparatory processing of data.

Technological foundation: SUBARU's four-wheel drive patents

Patents corresponding to SUBARU's four-wheel drive technologies were identified by means of two different search threads, and two patent sets were established: a limited and a comprehensive set. The limited patent set no. 1 is narrowly focussed on four-wheel drive technology, especially regarding differentials. The comprehensive patent set no. 2 focuses on four-wheel vehicles, in which case the four-wheel drive represents one technology amongst others.

For both patent sets, patents were collected from the USPTO database. The study was restricted to patents which had been granted between 1976 and 2006. The outset of the timeframe was defined by the USPTO database itself, as it provides no full-text patents dating back to any year before 1976. The end of the timeframe was not defined precisely, but when the patent search was conducted in 2009, the latest relevant patent had been issued in 2006, so this date served as a further boundary.

Concerning patent set 1, we focussed on US patent class 475, which is entitled "planetary systems and components". This class is highly significant for four-wheel drive trains. For instance, sub-class 475/220 includes inventions with a differential gear as part of the power train, and sub-class 475/230 refers to inventions with a differential based on bevel gears. The patents that were to be analysed had to bear a close relation to four-wheel drive technology. In addition to the classification and the company name, it was decided that they should contain the keyword "four wheel". The patent search according to the search thread produced a result of 57 patents in total (Table 2).

The second patent search was to produce a wider range of patents in patent set 2. The basic idea behind this patent set was that there may be inventions related to four-wheel drive trains that have been allocated to other classes, as complimentary technologies, technological predecessors or for the purpose of concealment. Therefore, the search was extended to all USPC classes. We collected additional patents from SUBARU with the aid of the keyword "four wheel". As a result, we found 225 patents in total (Table 2).

[3] For detailed information about SUBARU: http://www.subaru-global.com/.

Table 2 Search strings and results for patent sets 1 and 2

	Search string	Results
Patent set 1	(AN/"fuji jukogyo") AND ("four-wheel" OR "four wheel") AND CCL/475/$ AND ISD/19760101->20061231	57 Patents
Patent set 2	(AN/"fuji jukogyo") AND ("four-wheel" OR "four wheel") AND ISD/19760101->20061231	225 Patents

Patent search was carried out in the USPTO database (uspto.gov)

Data analysis and preparatory processing

The data was analysed and processed in three steps: (i) Patents were evaluated in close collaboration with experts from the FVA[4], (ii) novelty indices were computed by means of different methods, and (iii) computer-based results were compared with each other, using the experts' evaluation as a quality indicator.

First of all, the patents of both patent sets were evaluated manually by use of a qualitative scale. Starting with patent set 1, those patents where identified, in which certain aspects of SUBARU's four-wheel drive technology had been described for the first time. In accordance with a four-wheel drive technology scheme by Naunheimer et al. (2007) the type of transmission, e.g. manual and automatic transmission, the torque transfer between the axes, the lock mechanism, and the transfer case, were distinguished. On the basis of SUBARU's four-wheel drive history, the factor of market relevance was also used as an additional indicator of an inventions' novelty (Table 3). All in all, eleven inventions of high novelty were selected form patent set 1; the novelty of the inventions of the remaining 46 patents was classified as low.

According to this, patent set 1 can be considered a controlled patent set. Patent set 1 forms the core set of patent set 2. Based on the controlled patent set 1, we investigated, whether any of the inventions had been published before in patents of patent set 2. Furthermore, we identified product technologies, which had not been described in any patents of patent set 1 (Fig. 2). All in all, thirteen inventions of high novelty were identified in patent set 2 (Table 3).

Secondly, subsequent to the evaluation of the patents, novelty indices were computed. For this purpose, we selected methods which analyse novelty on the basis of (i) citation counts, (ii) qualified citations, (iii) set-related as well as (iv) non-set related citation structures and (v) semantic patent analysis. In total, ten investigations were carried out for both patent sets.

(i) The analysis of citation counts is based on the method applied by Ahuja and Lampert (2001). No or few backward citations point to a patent with a high degree of novelty; a high number of backward citations indicates a relatively low degree of novelty. Backward citations were thus counted for both patent sets.

(ii) For the analysis of qualified citations the method by Trajtenberg et al. (1997) was chosen. In addition to analysing patent citations in terms of categories, sub-categories and patent classes, US sub-classes were equally taken into account in this article, to enable a more detailed analysis of novelty, or respectively the class-related distance between patents.

(iii) For the analysis of the citation structure, we adapted the approach of Dahlin and Behrens (2005). We set the novelty of a patent to one minus the maximum of

[4] For detailed information about the FVA: http://www.fva-net.de/. The FVA can be seen as the leading innovation network in the field of drive train technology in Germany. The FVA enhance the collaboration between industry and science in the field of drive train technology.

Table 3 History of the market launch of SUBARU's four-wheel drive technology, new aspects of these technologies and their corresponding patents (according to SUBARU Deutschland GmbH (Ed.) 2005 with additional information and slight modifications, own representation, for a detailed description of innovation history see Gerken et al. 2010a, c)

Year of market launch	New aspects of the 4 × 4 system	US-patent (Set 1)	US-patent (Set 2)	Vehicle
1980	Mechanically insertable four-wheel drive and "dual range" for manual transmission	4,241,621	4,170,273	SUBARU 1800/Leon II
1981	Automatic transmission with multi-platetransfer-4WD	4,480,505	4,480,505	SUBARU 1800/Leon II
1983	Electro-pneumatically insertable 4WD for manual transmission	–	4,545,457	Libero and Justy
1988	Permanent 4WD with central differential for manual transmission and automatic transmission (ACT-4)	4,787,269	4,805,721	Coupé XT
1989	Permanent 4WD with self-locking viscous coupling and central differential	4,819,506	4,819,506	Legacy
1989	ECVT and insertable 4WD	4,846,765	4,715,467	Justy
1991	Permanent 4WD with VTD: planetary gear-type central differential with a multi-disc-clutch	5,066,268	5,066,268	Gran Turismo SVX
1998	Integration of VDC with optimized sensor system	–	5,734,595	Premium models of the legacy type series
2004	Permanent 4WD with optimized VTD, VDC, μ-estimator, automatic transmission and communication between different control units	6,595,086	6,595,086	Legacy 3.0 since 2004
2005	STi-AWD with			Impreza WRX STi
	(a) six-speed manual transmission	6,830,528	6,830,528	
	(b) variable dial	6,878,085	6,878,085	
	(c) DCCD, planet set—centre differential, electromagnetic clutch	7,127,343	7,127,342	
	(d) helical geared front differential	7,029,415	7,029,415	

similarity to preceding patents (Eq. 2). We calculated the similarity of citation structures for a set-related citation structure.

(iv) On the basis of the same method, we also calculated the novelty in terms of a non-set-related citation structure.

(v) Semantic patent analysis was conducted according to the process described in Semantic patent analysis for identifying inventions of high novelty. In order to overcome the problems arising from synonymy and the occurrence of stop words (described in Specific domain- or situation-related linguistic analysis), certain filter modules that have been devised by Forschungsvereinigung Antriebstechnik (FVA) and are set to specific technical domains and situations (e.g. the purpose of the semantic patent map), find application. We apply a synonym filter that serves to replace more than 6,500 synonyms (e.g. 'epicyclic gear' and 'satellite gear' were replaced by 'planetary gear'). General and domain-specific stop word filter were applied, which all in all eliminated almost 400 stop words (e.g. 'figure' and 'example' were deleted). For reasons of robustness we conducted three tests: (i) We applied two different similarity coefficients, an Inclusion and a Jaccard coefficient (Moehrle 2010). (ii) We analysed the effect of speech filters in total. As can be seen in Table 4, speech filters ensure that several SAO structures containing no content-oriented information were deleted and the quality of the SAO analysis was enhanced. (iii) We

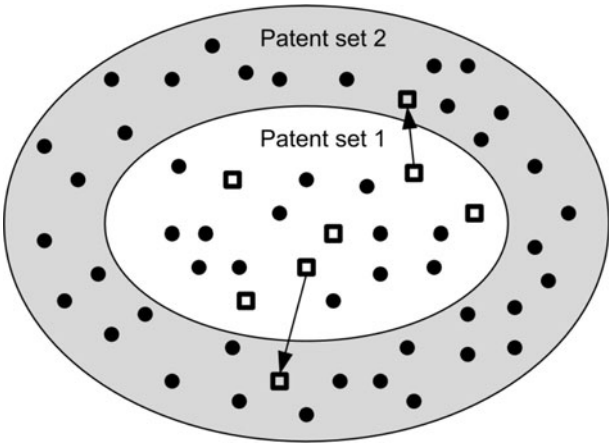

Fig. 2 Schematic illustration of the identification of patents related to inventions of high novelty. *White squares* symbolize patents related to inventions of high novelty; *black circles* symbolize the remaining patents, *arrows* indicate that a patent of patent set 2 includes earlier information than the corresponding patent in patent set 1

conducted a semantic patent analysis without using a synonym filter but with a stop word filter. The effect was that the mean value of similarities between all patents in a set decreased by about 13% from 0.0310 to 0.0269 in patent set 1 and by about 14% from 0.0137 to 0.0117 in patent set 2. Remarkably, the influence on the novelty measure was considerably lower. The effect was that the mean value of novelty increased by about 3% in patent set 1 and by about 5% in patent set 2. Obviously, the synonym filter mainly increased the similarity between patents that are not closely related to one another, but not of patents that are per se similar.

Thirdly, the results of our investigations were compared with each other, using the experts' evaluation as a quality indicator. For this purpose, we conducted a *U* test, we calculated precision and recall and Spearman's rank correlation.

In order to test, whether inventions of high novelty identified by experts differ in their novelty indices from patents not related to inventions with high novelty, Mann–Whitney *U* tests were conducted (see Mann and Whitney 1947; Buehl 2010).

Next, precision and recall were calculated. Precision and recall are quality measures in the field of information retrieval, where precision refers to the fraction of relevant documents within the corpus of retrieved documents while recall refers to the total number of relevant documents that have been retrieved (Manning et al. 2008). In other words, recall can be seen as the completeness of the retrieved documents (Eq. 3), precision signifies the accuracy of the retrieved documents (Eq. 4) (Stock 2007):

$$\text{Recall} = |\text{A} \cap \text{B}|/|\text{A}| \tag{3}$$

$$\text{Precision} = |\text{A} \cap \text{B}|/|\text{B}| \tag{4}$$

where A symbolizes relevant documents while B stands for retrieved documents (van Rijsbergen 1981).

In research tasks, it is necessary to maintain a balance between precision and recall, as these two influence each other (Manning and Schütze 2005): Reading all patents leads to a recall equal one. Due to the fact, that not all patents are relevant, precision decreases. In

Table 4 Results of SAO extraction with and without SAO filtering

Set	Filtering	Number of SAO structure	Median	Mean value	Max.	Min.
1	Yes	7,874	125	138.14	586	36
1	No	16,615	202	291.49	1,966	54
2	Yes	32,968	135	146.52	586	31
2	No	56,612	215	251.61	1,966	54

contrast, a selection of high-ranking patents leads to an increase in precision while the recall is reduced, because some relevant patents do not belong to the high-ranking patent set (Manning et al. 2008). In our concept, eleven recall levels form a basis for the computation of corresponding precision values. Precision is calculated for those patents which are the first to exceed recall levels between 0 and 100% in steps of 10% per level (see Manning and Schütze 2005).

For the purpose of calculating precision we developed a specific algorithm (see appendix A), in which we listed the patents according to their respective level of novelty. For each recall level we then identified the one patent that exceeded a predetermined limit and, if this appeared feasible, established the respective value of precision[5].

Finally, rank correlations between the investigations were established. For this purpose, Spearman's correlation (Buehl 2010) was applied. Rank correlations highlight the similarity between measures of novelty and constitute a basis for the characterization of novelty measures.

Results

The Mann–Whitney U test serves to determine whether there is a significant difference in the ranking of highly novel and less novel patents. In our case, the only significant difference related to novelty calculated by means of semantic patent analysis (Table 5). For patent set 1, the novelty indices calculated with TechB also pointed to a significant difference. This suggests that semantic patent analysis may be more suitable for the identification of inventions of high novelty than other methods.

Table 6 shows the results of the calculation of precision and recall for six novelty indices, including one novelty index based on semantic patent analysis for reasons of robustness, with respect to each patent of both patent sets. Two major findings deserve to be mentioned first of all: Semantic patent analysis outperforms other available methods of identifying inventions of high novelty. And novelty calculated by means of semantic patent analysis is more differentiated than novelty calculated by other methods. In greater detail, our findings are:

(i) In general, the smallest number of patents has to be read in order to reach a specific recall level, if semantic patent analysis is applied. Consequently, precision reaches its highest level, if semantic patent analysis is used for novelty assessment. Only in very few cases, semantic patent analysis is outperformed by other methods. This can be

[5] In the algorithm we took into account, that several patents may have the same level of novelty. In such cases we assume that analysts read patents stepwise. Every novelty value has to be considered as one step. Hence, if an analyst reads one patent with a novelty equal 0.2, he also reads all other patents with a novelty of 0.2 independent of the relevance of the first patent he has read with a novelty equal 0.2. In some of these cases it makes no sense to report a precision value on a specific level of recall.

observed with respect to both patent sets. Assessing novelty by means of citation counts performs well in both patent sets on a very low recall level. On higher recall levels, citation counts are outperformed by TechB as well as by non-set related citation similarity and set related citation similarity. On most recall levels, TechB has a higher precision than non-set related citation similarity and set related citation similarity. On high recall levels (0.9 and 1.0), set related and non-set related citation similarity perform slightly better than TechB.

(ii) Some methods proved inapplicable for a calculation of precision on all recall levels. In the first patent set precision could only be calculated on all recall levels for novelty measured by means of semantic patent analysis. In the second patent set precision could be ascertained on all recall levels for novelty established by means of semantic patent analysis as well as by TechB. Accordingly, semantic patent analysis possesses the greatest discriminatory power in novelty assessment.

(iii) The results of semantic patent analysis without synonym filter are comparable to the results of semantic patent analysis with synonym filter. This may suggest that semantic patent analysis is also a useful method for novelty calculation without domain-specific modifications.

The results regarding rank correlation highlight three major findings (Table 7): novelty calculated by means of semantic patent analysis has the strongest correlation with the experts' choices. The performance of novelty indices is influenced by the extent of patent data. And novelty indices do not only differ in terms of performance but also in terms of results.

The results of Spearman's rank correlation predominantly confirm our findings concerning recall and precision. Semantic patent analysis performs best in the identification of inventions of high novelty, as it shows the strongest correlation with the experts' choices regarding both patent sets. TechB is also marked by a strong and significant correlation with experts' choices in regard to patent set 1, but less so regarding patent set 2.

Furthermore, we already observed that all methods perform better in the restricted patent set 1. Here, the results of Spearman's rank correlation differ in so far, as the set related citation similarity correlates slightly more with the experts' choice in reference to patent set 2 than in reference to patent set 1. But the correlation between set related citation similarity as well as non-set related citation similarity and the experts' choice is of minor significance for both patent sets.

Considering the correlations between different methods it becomes obvious that some methods are similar in terms of results. But in some cases, the results also bear a minor

Table 5 Results of Mann–Whitney *U* Test

	Semantic (Jaccard)	Semantic (Inclusion)	Non-set related citation similarity	Set related citation similarity	Citation counts	TechB
Patent set 1						
Mann–Whitney *U*	103.000	95.000	226.000	244.000	252.500	117.000
Z	−3.033	−3.195	−0.555	−0.241	−0.010	−2.756
Asymp. Sig. (2-tailed)	0.002	0.001	0.579	0.810	0.992	0.006
Patent set 2						
Mann–Whitney *U*	761.00	796.00	1200.50	1287.50	1238.00	1192.50
Z	−2.71	−2.55	−0.79	−0.48	−0.62	−0.82
Asymp. Sig. (2-tailed)	0.01	0.01	0.43	0.63	0.54	0.41

Table 6 Recall and precision for patent sets 1 and 2

Recall level (%)	Patent set 1						Patent set 2					
	Semantic (Jaccard)	Semantic (Inclusion)	Non-set related citation similarity	Set related citation similarity	Citation counts	TechB	Semantic (Jaccard)	Semantic (Inclusion)	Non-set related citation similarity	Set related citation similarity	Citation counts	TechB
Recall reached by top *n* patents												
0.00	1	1			2		1	1	73		12	39
0.10	3	3				5	12	13	78		32	45
0.20	4	4	18		8	6	34	16	84		62	55
0.30	8	8	21		11	7	38	49	108		85	68
0.40	10	12				10	45	61	138			84
0.50	13	15	24		32	13	64	67	139			86
0.60	17	16	32		38	14	69	83	154		150	93
0.70	18	17	36	43	43	23	94	89		166	170	114
0.80	25	22	41	45	50	24	98	103	158		185	139
0.90	29	23	44	53	56	33	109	105	168	178	203	198
1.00	41	40	47	56	57	52	157	160	169	182	217	209
Precision												
0.00	1.00	1.00			0.50		1.00	1.00	0.01		0.08	0.03
0.10	0.67	0.67				0.40	0.17	0.15	0.03		0.06	0.04
0.20	0.75	0.75	0.17		0.38	0.50	0.09	0.19	0.04		0.05	0.05
0.30	0.50	0.50	0.19		0.36	0.57	0.11	0.08	0.04		0.05	0.06
0.40	0.50	0.42				0.50	0.13	0.10	0.04			0.07
0.50	0.46	0.40	0.25		0.19	0.46	0.11	0.10	0.05			0.08
0.60	0.41	0.44	0.22		0.18	0.50	0.12	0.10	0.05		0.05	0.09
0.70	0.44	0.47	0.22	0.19	0.19	0.35	0.11	0.11		0.06	0.06	0.09

Table 6 continued

Recall level (%)	Patent set 1						Patent set 2					
	Semantic (Jaccard)	Semantic (Inclusion)	Non-set related citation similarity	Set related citation similarity	Citation counts	TechB	Semantic (Jaccard)	Semantic (Inclusion)	Non-set related citation similarity	Set related citation similarity	Citation counts	TechB
0.80	0.36	0.36	0.22	0.20	0.18	0.38	0.11	0.11	0.07		0.06	0.08
0.90	0.34	0.43	0.23	0.19	0.18	0.30	0.11	0.11	0.07	0.07	0.06	0.06
1.00	0.27	0.28	0.23	0.20	0.19	0.21	0.08	0.08	0.08	0.07	0.06	0.06

Grey-coloured fields indicate that precision could not be calculated on a specific level of recall. Semantic patent analysis by use of the Jaccard coefficient was conducted for robustness reasons

resemblance to those obtained by means of other methods. A strong correlation for both patent sets can be found between set related similarity and non-set related similarity. Both set related and non-set similarity also bear a resemblance to novelty measured by semantic patent analysis. Furthermore, there is a weak correlation between citation counts and TechB. This may indicate that citation counts marginally influence TechB, as TechB is also based on citations. For both patent sets, the correlation between TechB on the one hand, and set related citation similarity, non-set related citation similarity and semantic similarity on the other, is insignificant.

A detailed analysis of the differing results obtained by TechB and semantic patent analysis shows that in both patent sets TechB outperforms semantic patent analysis with regard to US patent 4,819,506 and US patent 6,878,085. US patent 4,819,506 is primarily assigned to US patent class 74, which is entitled 'machine element or mechanism'. Only one out of seven patents cited by US patent 4,819,506 is also assigned to US patent class 74. Comparable to this, US patent 6,878,085 is primarily assigned to US patent class 475, but only one out of nine of its cited patents is also assigned to US patent class 475. This indicates that TechB performs well, if there is a swift from one technological domain, as represented by certain patent classes, to another technological domain. In the case of US patent 4,819,506 we observed a swift from US patent class 180, which contains the majority of cited patents, to US class 74, which is the primary patent class of US patent 4,819,506. Similar to that, we observed that a swift from US patent class 180 to US patent class 475 was the most common.

In contrast, TechB works less well for US patents 6,830,528, 4,480,505 and 7,029,415 in both patent sets and for US patent 4,805,721 in the second patent set. In these patents,

Table 7 Spearman's rank correlation coefficients between different types of novelty calculation for patent set 1 and patent set 2

		Semantic (Jaccard)	Semantic (Inclusion)	Non-set related citation similarity	Set related citation similarity	Citation counts	TechB	Expert
Semantic (Jaccard)	Set 1	1.000	0.963**	0.446**	0.156	0.069	0.044	0.405**
	Set 2	1.000	0.981	0.557	0.295	−0.133	0.072	0.181
Semantic (Inclusion)	Set 1	0.963**	1.000	0.437**	0.190	0.087	−0.006	0.427**
	Set 2	0.981	1.000	0.549	0.293	−0.129	0.073	0.171
Non-set related citation similarity	Set 1	0.446**	0.437**	1.000	0.527**	−0.094	−0.056	0.074
	Set 2	0.557	0.549	1.000	0.537	−0.139	0.047	−0.053
Set related citation similarity	Set 1	0.156	0.190	0.527**	1.000	−0.075	0.063	−0.032
	Set 2	0.295	0.293	0.537	1.000	−0.036	0.093	0.032
Citation counts	Set 1	0.069	0.087	−0.094	−0.075	1.000	0.092	−0.001
	Set 2	−0.133	−0.129	−0.139	−0.036	1.000	0.137	0.041
TechB	Set 1	0.044	−0.006	−0.056	0.063	0.092	1.000	0.373**
	Set 2	0.072	0.073	0.047	0.093	0.137	1.000	0.056
Expert	Set 1	0.405**	0.427**	0.074	−0.032	−0.001	0.373**	1.000
	Set 2	0.181	0.171	−0.053	0.032	0.041	0.056	1.000

** The correlation is significant at the 0.01 level (two-sided)

Grey-coloured fields indicate that semantic analysis by use of the Jaccard coefficient was included for robustness reasons

only minor patents from other patent classes are cited. But they strongly differ from preceding patents in semantic terms, as these patents involve new components, such as a permanent four-wheel drive and a helical geared front differential in a four-wheel drive train. Hence, a semantic patent analysis obviously outperforms TechB in comparing patents from identical technological domains. On the other hand, TechB seems to be a reliable indicator of changing technological domains between a patent and its cited patents. In this context, TechB focuses on trajectories, as represented by citation links. Semantic patent analysis is fully independent of direct linkage between patents.

Discussions and conclusions

Semantic patent analysis, especially its application for measuring the novelty of inventions, still is a fairly new field of research. In addition to a summary of our results the following conclusions will refer to theoretical perspectives and to managerial and political implications as well as to certain limitations. Furthermore, we shall put a special emphasis on scientific perspectives in the further advancement of semantic patent analysis.

Conclusions

In this article, we introduced a new method for the identification of patents related to inventions of high novelty. For this purpose, we applied semantic patent analysis and established each examined patent's distance from prior art as represented by preceding patents. In the course of a case study, we compared our method with three established methods, based on patent citation data, with reference to two different patent sets and analysed which method is the most adequate for the identification of patents related to inventions of high novelty. We evaluated the results by use of measures borrowed from information science: precision and recall. In both patent sets semantic patent analysis mostly manages to outperform other methods; only under certain conditions TechB performs better.

Our findings contribute to (i) the field of novelty measurement, to (ii) the theory of patent and technology management but also to (iii) the theory of technological change. Our findings lead to an advancement of the theory of novelty measurement, as we have introduced a semantic method, using functions within patent texts to generate measures of novelty. Our findings help to advance the theory of technology and patent management, because measuring the novelty of an invention protected by a patent may also be an enhancement to patent valuation, as the value of a patent also depends upon its novelty (Reitzig 2003a). These results contribute to the measurement of inventions' novelty, which is a basic element for the understanding of technological change. This may provide a valuable support to scientists and practitioners alike.

Seen from a managerial perspective, derived from the phenomenon of information overflow (Bergmann et al. 2008), the use of sematic patent analysis for the identification of patents related to inventions of high novelty may have two implications, one for patent management and one for technology monitoring. For patent management it considerably expedites the reduction of efforts and uncertainty in monitoring patents, which has so far been a particularly time-consuming task. The reliable identification of highly important information opens up the opportunity to focus resources on this particular information while less important information is only given a fast screening. The application of novelty measures enables a ranking of new (and old) patents, directing the patent assessor's

attention to high-ranking patents which may be more significant than the low-ranking ones. Novelty measures can also find application in the valuation of patents.

As concerns technology monitoring, our method is not a complete monitoring instrument as it is not able to suggest a way of devising new technologies. But it can shed light on technology fields where intensive progress seems to be emerging (measured in the development of the related patents' novelty) and thus help identify promising technology areas.

Regarded from a political point of view, a semantic patent analysis may provide valuable support to patent offices. As novelty is a precondition for the granting of a patent, patent offices are required to evaluate the novelty of each patent that has been filed. Consequently, the use of semantic patent analysis for the identification of patents related to inventions of high novelty and the appropriate assessment of patents' novelty is also capable of changing the process of patent assessment in patent offices. Even in cases where the search has to be carried out across a range of patent classes, a semantic patent analysis may be helpful: Separating the analysis over the classes (which represent different technological fields) can provide class specific measures for novelty and help the patent officer evaluate the invention.

Our case study is subject to certain limitations concerning (i) the scope of the case study as such, (ii) the scope of the patent searches, (iii) the fact, that the relationship between semantic structures and functional relationships have not yet been explored in detail and (iv) the necessity to develop different kinds of filters.

Firstly, the exploration of our method has so far been limited to the field of mechanical engineering. This area appears suitable for semantic patent analysis, as the wording tends to be simpler as well as more standardized than in other technological fields. For example, patents from the pharmaceutical area seem to be more difficult, as they involve chemical symbols and often highly complex chemical compounds.

Secondly, by comparing the results from both patent sets, we found that all approaches showed a better performance with regard to the first patent set. Hence, the assessment of novelty is not only dependent on the chosen method, but also on the data. In patent set 1 we were unable to find a corresponding patent for each technology involved in SUBARU's innovations. Hence, this data set would be too small and incomplete for an effective technology monitoring in the field of four wheel drive trains. In contrast, patent set 2 rather seems to be too extensive, as several patents of high novelty that are not directly related to the four wheel drive train can actually be found in this patent set. Therefore, we have come to the conclusion that there is still a gap concerning the unequivocal definition of technological fields in patent data, even though patent classification serves to facilitate orientation.

Thirdly, we used semantic structures, i.e. subject-action-object structures, to analyse the novelty of inventions. In this context, we presumed that these semantic structures represent functional aspects of the inventions. Up to now, the relationship between SAO-structures extracted from patents' full texts and the functional aspects of inventions has not been explored in detail.

Fourthly, we developed different kinds of speech filter in this case study. For robustness reasons we tested the influence of domain-specific filters. Our results actually indicate that semantic patent analysis for novelty measurement may be applied without such filters. But one has to be aware of the fact that we analysed patents from a clearly defined patent set. Elaborate speech filters may be more important in broader technological fields. The development of filters may seem to be time-consuming. However, these filters merely have to be generated once for a technological domain and can then be reused. Furthermore, we see a perspective in approaches that deal with this problem by using available ontologies (e.g. WordNet).

Scientific perspectives

Although semantic patent analysis worked very well in our particular patent sets, there still are some drawbacks and a need for enhancement, especially (i) in testing further full-text based methods for patent analysis, (ii) in profiling available patent-based methods, (iii) in SAO weighting, (iv) in similarity calculation, (v) in answering the question, what is new and, maybe most importantly, (vi) in differentiating between types of novelty.

(i) In this article, SAO-based patent analysis was tested for novelty measurement, which was mainly motivated by drawbacks of patent citation analysis and keyword analysis. Comparable to SAO-based patent analysis, the analysis of patents by use of *n*-grams (word chains with *n* words) (for details see e.g. Egghe 2000; Manning and Schütze 2005) might also be a promising approach and remains to be tested in detail.

(ii) As mentioned before, TechB performs better with regard to patents citing patents form different patent classes. This phenomenon calls for meticulous investigation, and we are challenged to find ways of combining both approaches, as they both have their respective advantages and disadvantages.

(iii) In our approach, we used SAO structures all with the same weights. Prospectively, different weights may be assigned to the SAO structures. For instance, the weights may be calculated using an SAO structure's frequency within one patent in proportion to the frequency in a textual corpus.

(iv) Similarity calculation has been discussed in respect to various areas (Sternitzke and Bergmann 2009; Jeong et al. 2008; Kangasabai and Pan 2008) and there are several methods of linking similar objects and calculating the similarity between these objects (Moehrle 2010). Hence, similarity calculation has to be approached from a more theoretical direction, as in our case study we applied a very common similarity measure. We defined novelty as the distance between a patent and the most similar preceding patent. But there possibly are more complex relationships between patents, as novelty may be limited to a certain aspect in one patent, and to completely different aspects in another. Hence, limitations of novelty are not restricted to the most similar patent, but can, in different forms, also be traced back to the second, third and fourth most similar patents.

(v) Our approach could be usefully complemented by co-word-analysis (see for example Callon et al. 1991; Engelsman and van Raan 1994; An and Wu 2011). This would not only facilitate a quantitative measure of novelty, but would also allow for gaining insights into textual structures that might characterize a technological field.

(vi) In this article, we focussed on a very general understanding of novel inventions, even though there are various types of novelty (see Theoretical background). Indeed, it is possible to distinguish between different types of novelty, especially recombinant and pioneering novelty, by way of patent information.

Acknowledgments The cited case study was produced in the course of a joint project with the Forschungsvereinigung Antriebstechnik (FVA). We wish to thank the FVA and all industrial members for their contributions and their support. Furthermore, we would like to thank Dipl.-Ing. (FH) Jens Potthast for extensive programming efforts on the PatVisor®, Dr. Lothar Walter for commenting an earlier version of this paper and two anonymous reviewers for their constructive and helpful comments.

Appendix A: algorithm for the calculation of precision

See Fig. 3.

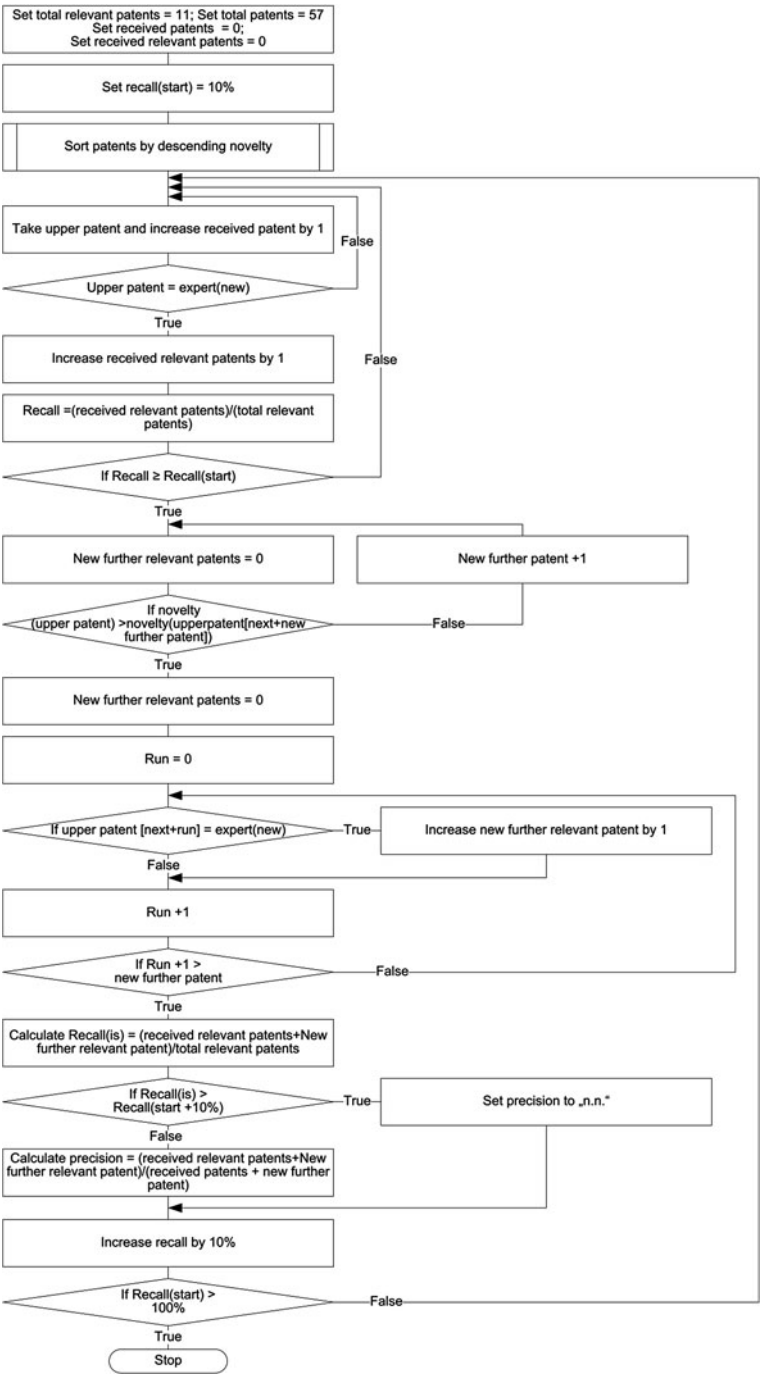

Fig. 3 Flow chart of the algorithm for the calculation of precision

References

Achilladelis, B., Schwarzkopf, A., & Cines, M. (1987). A study of innovation in the pesticide industry: Analysis of the innovation record of an industrial sector. *Research Policy, 16*(2–4), 175–212.

Achilladelis, B., Schwarzkopf, A., & Cines, M. (1990). The dynamics of technological innovation: The case of the chemical industry. *Research Policy, 19*(1), 1–34.

Ahuja, G., & Lampert, C. M. (2001). Entrepreneurship in the large corporation: A longitudinal study of how established firms create breakthrough inventions. *Strategic Management Journal, 22*(6–7), 521–544.

An, X. Y., & Wu, Q. Q. (2011). Co-word analysis of the trends in stem cells field based on subject heading weighting. *Scientometrics, 88*(1), 133–144.

Andersen, B. (1999). The hunt for S-shaped growth paths in technological innovation: A patent study. *Journal of Evolutionary Economics, 9*(4), 487–526.

Anderson, P., & Tushman, M. L. (1990). Technological discontinuities and dominant designs: A cyclical model of technological change. *Administrative Science Quarterly, 35*(4), 604–633.

Bergmann, I., Butzke, D., Walter, L., Fuerste, J. P., Moehrle, M. G., & Erdmann, V. A. (2008). Evaluating the risk of patent infringement by means of semantic patent analysis: The case of DNA-chips. *R&D Management, 38*(5), 550–562.

Buehl, A. (2010). *PASW 18: Einführung in die moderne Datenanalyse* (12th ed.). München: Pearson Studium.

Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics, 22*(1), 155–205.

Cascini, G., Fantechi, A., & Spinicci, E. (2004). Natural language processing of patents and technical documentation. In S. Marinai & A. Dengel (Eds.), *Document analysis systems VI* (pp. 89–92). Berlin: Springer.

Chandy, R. K., & Tellis, G. J. (1998). Organizing for radical product innovation: The overlooked role of willingness to cannibalize. *Journal of Marketing Research, 35*(4), 474–487.

Choi, S., Yoon, J., Kim, K., Lee, J. Y., & Kim, C. (2011). SAO network analysis of patents for technology trends identification: A case study of polymer electrolyte membrane technology in proton exchange membrane fuel cells. *Scientometrics, 88*(3), 863–883.

Christensen, C. M., & Overdorf, M. (2000). Meeting the challenge of disruptive change. *Harvard Business Review, 78*(2), 66–77.

Dahlin, K. B., & Behrens, D. M. (2005). When is an invention really radical? Defining and measuring technological radicalness. *Research Policy, 34*(5), 717–737.

Debackere, K., Verbeek, A., Luwel, M., & Zimmermann, E. (2002). Measuring the progress and evolution in science and technology—II: The multiple uses of technometric indicators. *International Journal of Management Reviews, 4*(3), 213–231.

Egghe, L. (2000). The distribution of N-grams. *Scientometrics, 47*(2), 237–252.

Engelsman, E. C., & van Raan, A. F. J. (1994). A patent-based cartography of technology. *Research Policy, 23*(1), 1–26.

Fendt, H. (1988). Technische Trends rechtzeitig erkennen—Patentschriften gewähren Blicke hinter die Kulissen von F&E. *Havard Manager, 10*(4), 72–80.

Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science, 1*, 117–132.

Fleming, L., Mingo, S., & Chen, D. (2007). Collaborative brokerage, generative creativity, and creative success. *Administrative Science Quarterly, 52*(3), 443.

Fleming, L., & Sorenson, O. (2001). Technology as a complex adaptive system: Evidence from patent data. *Research Policy, 30*(7), 1019–1039.

Frietsch, R. (2007). *Patente in Europa und der Triade: Strukturen und deren Veränderung*. Karlsruhe: Fraunhofer Institut für System- und Innovationsforschung.

Gerken, J. M., Moehrle, M. G., & Walter, L. (2010a). Patents as an information source for product forecasting: Insights from a longitudinal study in the automotive industry. *R&D Management Conference 2010 Proceedings.* Manchester.

Gerken, J. M., Moehrle, M. G., & Walter, L. (2010b). Semantische Patentlandkarten zur Analyse technologischen Wandels: Eine Längsschnittstudie aus der Allradtechnik. In J. Gausemeier (Ed.). *6. Symposium für Vorausschau und Technologieplanung* (pp. 325–349). Paderborn: Heinz Nixdorf Institut

Gerken, J. M., Walter, L., & Moehrle, M. G. (2010c). Semantische Patentlandkarten. Einsatz semantischer Patentlandkarten im Anwendungsfeld der Antriebstechnik—Eine explorative Analyse am Beispiel der Planentengetriebe. *Heft Nr. 924 der Forschungsvereinigung Antriebstechnik.* Frankfurt/Main: VDMA.

Granstrand, O. (2000). *The Economics and management of intellectual property: Towards intellectual capitalism*. Cheltenham: Edward Elgar.
Grupp, H. (1997). *Messung und Erklärung des technischen Wandels: Grundzüge einer empirischen Innovationsökonomik*. Berlin: Springer.
Hall, B. H., Jaffe, A. B., & Trajtenberg, M. (2001). The NBER patent citations data file: lessons, insights and methodological tools. *NBER Working Paper 8498*.
Han, Y., & Park, Y. (2006). Patent network analysis of inter-industrial knowledge flows: The case of Korea between traditional and emerging industries. *World Patent Information, 28*(3), 235–247.
Hargadon, A., & Sutton, R. I. (1997). Technology brokering and innovation in a product development firm. *Administrative Science Quarterly, 42*(4), 716–749.
Haupt, R., Kloyer, M., & Lange, M. (2007). Patent indicators for the technology life cycle development. *Research Policy, 36*(3), 387–398.
Jaffe, A. B., & Trajtenberg, M. (1999). International knowledge flows: Evidence from patent citations. *Economics of Innovation and New Technology, 8*(1/2), 105–136.
Jaffe, A. B., Trajtenberg, M., & Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *The Quarterly Journal of Economics, 108*(3), 577–598.
Jeong, B., Lee, D., Cho, H., & Lee, J. (2008). A novel method for measuring semantic similarity for XML schema matching. *Expert Systems with Applications, 34*(3), 1651–1658.
Kangasabai, R., & Pan, H. (2008). Method of text similarity measurement. *US-Patent 7,346,491 B2*.
Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation, 14*(1), 10–25.
Kim, Y. G., Suh, J. H., & Park, S. C. (2008). Visualization of patent analysis for emerging technology. *Expert Systems with Applications, 34*(3), 1804–1812.
Knight, H. J. (2004). *Patent strategy for researchers and research managers*, (2nd ed.). Chichester: Wiley.
Kurmaniak, C. (2008). Electromagnetics comes through in the clutch. *ANSYS Advantage, 2*(3), 30–31.
Lee, C., Jeon, J., & Park, Y. (2011). Monitoring trends of technological changes based on the dynamic patent lattice: A modified formal concept analysis approach. *Technological Forecasting and Social Change, 78*(4), 690–702.
Lee, S., Yoon, B., Lee, C., & Park, J. (2009a). Business planning based on technological capabilities: Patent analysis for technology-driven roadmapping. *Technological Forecasting and Social Change, 76*(6), 769–786.
Lee, S., Yoon, B., & Park, Y. (2009b). An approach to discovering new technology opportunities: Keyword-based patent map approach. *Technovation, 29*(6–7), 481–497.
Li, Y., Wang, L., & Hong, C. (2009). Extracting the significant-rare keywords for patent analysis. *Expert Systems with Applications, 36*(3), 5200–5204.
Lichtenthaler, E. (2004). Technological change and the technology intelligence process: A case study. *Journal of Engineering and Technology Management, 21*(4), 331–348.
Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 18*(1), 50–60.
Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press.
Manning, C. D., & Schütze, H. (2005). *Foundations of statistical natural language processing*. Cambridge: MIT Press.
Moehrle, M. G. (2010). Measures for textual patent similarities: A guided way to select appropriate approaches. *Scientometrics, 85*(1), 95–109.
Moehrle, M. G., & Geritz, A. (2007). Developing acquisition strategies based on patent maps. In T. Khalil & Y. Hosni (Eds.), *Management of technology: New directions in technology management* (pp. 19–29). Oxford: Elsevier.
Moehrle, M. G., Walter, L., Geritz, A., & Müller, S. (2005). Patent-based inventor profiles as a basis for human resource decisions in research and development. *R&D Management, 35*(5), 513–524.
Naunheimer, H., Novak, W., & Ryborz, J. (2007). *Fahrzeuggetriebe: Grundlagen, Auswahl, Auslegung und Konstruktion*. Berlin: Springer.
Park, H., Yoon, J., & Kim, K. (2011). Identifying patent infringement using SAO based semantic technological similarities. *Scientometrics, 90*(2), 515–529.
Park, Y., Yoon, B., & Lee, S. (2005). The idiosyncrasy and dynamism of technological innovation across industries: Patent citation analysis. *Technology in Society, 27*(4), 471–485.
Pope, B. (2009). All-wheel-drive suppliers get grip on changing market. Resource document. Accessed March 8, 2010 http://wardsauto.com/ar/suppliers_grip_market_090427/.
Princeton University. (2006). WordNet 3.0. Resource document. Accessed March 23, 2011, from http://wordnetweb.princeton.edu/perl/webwn.

Reitzig, M. (2003a). What do patent indicators really measure? A structural test of 'novelty' and 'inventive step' as determinants of patent profitability. *LEFIC Working paper 2003–1*. Copenhagen, DK.
Reitzig, M. (2003b). What determines patent value? Insights from the semiconductor industry. *Research Policy, 32*(1), 13–26.
Rost, K. (2010). The strength of strong ties in the creation of innovation. *Research Policy*. doi: 10.1016/j.respol.2010.12.001.
Schoenmakers, W., & Duysters, G. (2010). The technological origins of radical inventions. *Research Policy, 39*(8), 1051–1059.
Schumpeter, J. A. (1934). *The theory of economic development: an inquiry into profits, capital, credit, interest, and the business cycle*. Cambridge: Harvard University Press.
Sood, A., & Tellis, G. J. (2005). Technological evolution and radical innovation. *Journal of Marketing, 69*(3), 152–168.
Sternitzke, C., & Bergmann, I. (2009). Similarity measures for document mapping: A comparative study on the level of an individual scientist. *Scientometrics, 78*(1), 113–130.
Stock, W. G. (2007). *Information retrieval: Informationen suchen und finden*. München: Oldenbourg.
Stockmar, J. (2004). *Das große Buch der Allradtechnik*. Stuttgart: Motorbuch-Verl.
SUBARU Deutschland GmbH (Ed.). (2005). 33 Jahre SUBARU-Allradantrieb. Resource document. Accessed May 15, 2009, from http://www.subaru-presse.de/fileadmin/templates/downloads/awd/PressemappeSubaruAllrad-Technologie04-2005SUB.doc.
Trajtenberg, M. (1990). A penny for your quotes: Patent citations and the value of innovations. *The Rand Journal of Economics, 21*(1), 172–187.
Trajtenberg, M., Henderson, R., & Jaffe, A. (1997). University versus corporate patents: A window on the basicness of invention. *Economics of Innovation and New Technology, 5*(1), 19–50.
Trippe, A. J. (2003). Patinformatics: Tasks to tools. *World Patent Information, 25*(3), 211–221.
Tseng, Y., Lin, C., & Lin, Y. (2007). Text mining techniques for patent analysis. *Information Processing & Management, 43*(5), 1216–1247.
USPTO (Ed.). (2007). *Manual of patent examining procedure* (8th ed.). Alexandria.
van Rijsbergen, C. J. (1981). *Information retrieval* (2th ed.). London: Butterworth.
von Wartburg, I., Teichert, T., & Rost, K. (2005). Inventive progress measured by multi-stage patent citation analysis. *Research Policy, 34*(10), 1591–1607.
Witt, U. (2009). Propositions about novelty. *Journal of Economic Behavior & Organization, 70*(1–2), 311–320.
Yoon, J., & Kim, K. (2011a). Detecting signals of new technological opportunities using semantic patent analysis and outlier detection. *Scientometrics*. doi:10.1007/s11192-011-0543-2.
Yoon, J., & Kim, K. (2011b). Identifying rapidly evolving technological trends for R&D planning using SAO-based semantic patent networks. *Scientometrics, 88*(1), 213–228.
Yoon, B., & Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. *The Journal of High Technology Management Research, 15*(1), 37–50.
Yoon, B., & Park, Y. (2005). A systematic approach for identifying technology opportunities: Keyword-based morphology analysis. *Technological Forecasting and Social Change, 72*(2), 145–160.