Technological Forecasting & Social Change xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

Firms' knowledge profiles: Mapping patent data with unsupervised learning

Arho Suominen [a,b,*], Hannes Toivanen [a,b,c], Marko Seppänen [d]

[a] VTT Technical Research Centre of Finland, PL 1000, Espoo, Finland
[b] Teqmine Analytics Ltd, Pasilanraitio 5, 00240 Helsinki, Finland
[c] Lappeenranta University of Technology, School of Business, Lappeenranta, Finland
[d] Tampere University of Technology, Department of Pori/Industrial Management, PL 300, 28101 Pori, Finland

ARTICLE INFO

Article history:
Received 17 February 2016
Received in revised form 28 September 2016
Accepted 28 September 2016
Available online xxxx

Keywords:
Technology management
Patent analysis
Unsupervised learning
Topic modelling
Telecommunication industry

ABSTRACT

Patent data has been an obvious choice for analysis leading to strategic technology intelligence, yet, the recent proliferation of machine learning text analysis methods is changing the status of traditional patent data analysis methods and approaches. This article discusses the benefits and constraints of machine learning approaches in industry level patent analysis, and to this end offers a demonstration of unsupervised learning based analysis of the leading telecommunication firms between 2001 and 2014 based on about 160,000 USPTO full-text patents. Data were classified using full-text descriptions with Latent Dirichlet Allocation, and latent patterns emerging through the unsupervised learning process were modelled by company and year to create an overall view of patenting within the industry, and to forecast future trends. Our results demonstrate company-specific differences in their knowledge profiles, as well as show the evolution of the knowledge profiles of industry leaders from hardware to software focussed technology strategies. The results cast also light on the dynamics of emerging and declining knowledge areas in the telecommunication industry. Our results prompt a consideration of the current status of established approaches to patent landscaping, such as key-word or technology classifications and other approaches relying on semantic labelling, in the context of novel machine learning approaches. Finally, we discuss implications for policy makers, and, in particular, for strategic management in firms.

1. Introduction

Operationalising a company's knowledge base in terms of its depth and breadth and creating trajectories to the future is challenging (Zhang and Baden-Fuller, 2010). The intensified complexity of emerging technologies (Breschi et al., 2003; Garcia-Vega, 2006) requires improved understanding of the nature and effect of cross-disciplinary activities in innovation processes (Wang and von Tunzelmann, 2000). Increasingly, companies must rely on broad knowledge bases covering diverse technology areas, while simultaneously having significant depth in their core competence. This creates a new type of tension for the management of technology and innovation. This is particularly problematic in highly dynamic industries. We examine the effects and potential of big data approaches in managing this increased complexity of company knowledge bases with a study on the telecommunication industry, and develop perspectives to exploit big data foresight approaches in support of strategic planning.

Previous studies on the depth and breadth of knowledge and technological trajectories have used patent information. As Moorthy and Polley (2010) point out, patents are the most feasible approach for analysing the breadth and depth of knowledge within a company as the data provides an insight to its competences. The simplest approach to quantifying the knowledge base is to use the patent classification scheme provided in the patent archive as a basis for evaluation – breadth correlating with the diversity in patent classifications and depth with the concentration of patent classifications in a company patent portfolio. This approach was used for example by Zhang and Baden-Fuller (2010) to analyse technology collaboration. Moorthy and Polley (2010) and SubbaNarasimha et al. (2003) use the approach to analyse the impact of breadth and depth of knowledge to company performance. Wu and Shanley (2009) operationalise the role of exploration in company knowledge stock by means of patent metrics.

Analysing classification metadata, in addition to citations, can be regarded as the de facto standard of utilizing patent metrics (e.g. Huang et al. (2015) as a case in point). This approach in analysing breadth and depth is not without limitations. Connecting patent classifications directly to industry sectors poses a challenge (Schmoch, 2008). Different patent classification systems have struggled to establish a tool to clearly distinguish industries into specific classes, limiting the applicability of classifications for sectoral analysis. Classifications are also of

* Corresponding author at: VTT Technical Research Centre of Finland, PL 1000, Espoo, Finland.
E-mail addresses: arho.suominen@vtt.fi (A. Suominen), hannes.toivanen@teqmine.com (H. Toivanen), marko.seppanen@tut.fi (M. Seppänen).

http://dx.doi.org/10.1016/j.techfore.2016.09.028
0040-1625/

limited value in directing inventive effort (Loh et al., 2006), which is understandable due to the information retrieval nature of patent classifications. Patent classifications are a tool for the patent process, and the human process related to assigning classes is valuable in the intellectual process, even to the extent that automated classifications fall short of providing similar results (Richter and MacFarlane, 2005).

The subjectiveness of the classification process of patents remains a major limitation for the usefulness patent data (Venugopalan and Rai, 2015), making it an inadequate measure to satisfy the needs of corporate planning (Archibugi and Planta, 1996; Lai and Wu, 2005). Nakamura et al. (2015) review the managerial challenges of analysing patent data, pointing to the need for frequent updates (Herrero et al., 2010) and cost of data collection (Nakamura et al., 2015, Kajikawa et al., 2006) with a limited success rate in producing practical results. They conclude, through expert interviews, that even though patent data is a relevant decision making tool for practitioners its usefulness is hindered by the significant limitations embedded in the patent classification based metric. From this perspective, machine learning provides a valuable approach for strategic foresight and technology management (see e.g. Ventura et al., 2015). Machine learning opens the possibility for cost effective analysis of full text patent data, which can mitigate the limitations of de facto standard metadata based approaches.

By employing big data approaches to manage technology intelligence, companies can foster new forms of adaptive learning in innovation and strategy. Such approaches require the augmentation of human judgement in the categorisation and analysis of knowledge with machine learning methods, prompting serious challenges to the existing corporate foresight traditions. Leveraging these efforts within companies requires their systematic integration to existing strategic foresight processes. Using an automated continuous monitoring based on topic modelling with Latent Dirichlet Allocation and network analysis, we will show how a semantic analysis leads to the identification of opportunities for learning and innovation in complex environments. From a total dataset of 157,718 full-text telecommunication patents from USPTO, we have monitored and detected changes in the knowledge patterns of companies, e.g. how semantic analysis shows the change from a hardware-focused knowledge domain in telecommunication towards software-dominated knowledge foci. In this paper, we explore the latent knowledge dimensions of patents in global telecommunication companies, focusing on two questions: 1) Can we identify topical knowledge foci of different companies with unsupervised learning, and if so, 2) What are the dynamics of knowledge domains among the companies?

2. Background

Informetric analysis focuses on operationalising developments in the science and technology system. Informetrics can focus on a science, technologies or companies creating insight on the historical developments and forecast future trajectories. At a company level, Porter and Newman (2011) write about competitive technical information (CTI), the information companies need to survive in the dynamic marketplace. Suominen (2013) reviews the established metrics used to create quantitative insights, highlighting that metrics used to profile developments need to be objective and reproducible, while responding to Ayres (1989) call for accurate decision-making tools.

Much of the current informetrics analyses have focused on the metadata level (cf. Suominen, 2013) creating measures of activity, linkage or impact (Moed et al., 1995). Text analytics have most commonly been limited to keyword or abstract analysis. New open datasets and increases in computational efficiency have made full-text analysis possible (for example Glenisson et al., 2005). Tseng et al. (2007) have reviewed text mining techniques for patent analysis highlighting different methodological options and steps, such as text segmentation, summary extraction, feature selection, term association, cluster generation, topic identification, and information mapping. Tseng et al. (2007) describe the approaches to filter irrelevant content and retrieving the core features of the patent. Kang et al. (2007) have reviewed different clustering approached to summarizing patents, although much of this work is based on utilizing the international patent classification systems. Kim and Choi (2007) on the other hand analyse patents using the semantic structure of the patent as a starting point. Patent text mining studies often either rely on filtering text based on established knowledge on the structure of patent text or show trends of classification or aggregated technology areas.

2.1. Depth and breadth of knowledge

The increased complexity of technologies has changed the dynamics of innovation in that there is an increased need for cross-disciplinary activities (Wang and von Tunzelmann, 2000; Subramaniam and Youndt, 2005). Studies have shown that technologically diverse knowledge systems are a dominant feature in companies, as multiple fields of knowledge are integrated in the innovation process (Mendonça, 2006). To analyse change in knowledge resources we are forced to understand the multi-dimensional knowledge base of an industry (Kauffman et al., 2000).

Knowledge depth can be defined as the level of expertise within a confined technological area (George et al., 2008; Zhang et al., 2007) described by Wang and von Tunzelmann (2000) as "analytical sophistication". In contrast, breadth of knowledge refers to the number of adjacent technologies in the relevant multi-dimensional knowledge space of a company (Wang and von Tunzelmann, 2000; Zhang et al., 2007). Moorthy and Polley (2010) showed that rather than the stock of knowledge, the breadth and depth of knowledge in fact represent more important variables to explain a firm's performance. Companies are required to have a minimum depth of knowledge in a specific area and breadth enables them to cope with rapid technological change. An evaluation of these two variables at a company level allows us to follow strategic trajectories.

The breadth and depth of knowledge in companies is in parts visible outside the company in codified information such as patents, where patent classifications provide a tool for analyses. Codification has enabled easy access to analysing the knowledge structure through a posteriori labels given to new information. With patents, this metadata is in fields such as application data, patent classification, and assignee, which codify the actual information to make it more accessible.

Patent classifications have remained as the most practical approach in understanding the structure of the information. There are, however, significant caveats to this approach. Patent classifications are subjective in nature, prone to classifications errors and overall noisiness (Dahlin and Behrens, 2005; Nemet, 2009). The classifications are by definition an information retrieval system, which scholars and practitioners use as a proxy metric for example analysing the breadth and depth of knowledge. At the same time we are acutely aware of several significant limitations in the proxy we are using. We know that the implicit notions and underlying taxonomy of patents are often misunderstood (McNamee, 2013). There are clear challenges to link patent classifications to either industry (Schmoch, 2008) or market sectors (Jaffe, 1986). Classifications are also of limited value in directing inventive effort (Loh et al., 2006). Using a priori determined classification, new topics pose a challenge. Classification based on historical knowledge lacks the ability to adapt to new knowledge (for discussion on approaches, see van Merkerk and van Lente, 2005; Kuusi and Meyer, 2007).

There is a clear need for a more adaptive approach to analysing patent data, suggesting that automated classification drawn from the actual text could be a better approach for showing the actual breadth and depth of the knowledge base.

2.2. Unsupervised learning and topic modelling in patent data

Unsupervised learning produces an outcome based on an input while not receiving any feedback from the environment. As an automated

Please cite this article as: Suominen, A., et al., Firms' knowledge profiles: Mapping patent data with unsupervised learning, Technol. Forecast. Soc. Change (2016), http://dx.doi.org/10.1016/j.techfore.2016.09.028

classification method, unsupervised learning differs from supervised or reinforced learning by its reliance on a formal framework that enables the algorithm to find patterns. The majority of unsupervised methods rely on a probabilistic model of the input data. An unsupervised learning method estimates the model that represents the probability distribution for an input, either based on previous inputs or independently.

Topic models are unsupervised learning methods and the Latent Dirichlet Allocation (LDA) is one topic model that draws out latent patterns from text. In 2007, Blei and Lafferty (2007) showed the usability of topic models in modelling the structure of semantic text. In presenting the methodology, Blei and Lafferty (2007) noted that topic models " ...can extract surprisingly interpretable and useful structure without any explicit "understanding" of the language by computer". The basic idea behind the model is that each document in a corpus is a random mixture over latent topics, and each latent topic is characterized by a distribution over words. In the LDA model, each document is a mixture of a number of topics based on the words attributable to each of the topics. LDA allows us to uncover these latent probability distributions based on the semantic text used in the document, thus classifying the documents based on the latent patterns within them. For a detailed explanation on algorithms, refer to for example Blei and Lafferty (2009) and for an evaluation in analysing scientific publications, please see Yau et al. (2013).

Recent works have described several tailored or updated versions of LDA. These methods include correlated topic models (CTM), Hierarchical Latent Dirichlet Allocation (Hierarchical LDA) and the Hierarchical Dirichlet Process (HDP), Dynamic Topic Model (DTM) and the Document Influence Model (DIM). All of the methods offer potential within the specific problem setting and data. LDA could be seen as a general model, for which more focused algorithms limit the impact of a constraint or consider for example temporal changes. In this study, we apply the LDA to validate the approach in its generic form.

Topic modelling has been recently applied to patent data in a number of studies. Venugopalan and Rai (2015) used a topic based approach to analysing the structure of patent data. Their analysis used the contextual frame of knowledge spillovers, resulting in the use of patent abstracts and claims as the basis of analysis. A patent abstract is known to carry a low information value (Cascini et al., 2007). This is due to the fact that the author of a patent has no incentive to write a clear and concise abstract that helps in discovering its content – rather the opposite. An uninformative abstract can reduce a patent's discovery by competitors, which may result in strategic benefits. Lack of an informative abstract does not affect the legal protection of the patent. However, the low information value of patent abstract has consequences for tech mining, as abstracts in scientific publications and patents are considered being similar. This inbuilt difference between the texts must be taken into account because in science the authors strive to make their work discovered and for patents, the opposite might hold true. Similarly claims as a semi-structured short descriptors leave much of the dimensions of the patent content hidden, as the claims only focus on the inventive claim. We propose that by using the patent description as source data, we are better capturing the complexity of the inventive effort and the knowledge of a company.

Hu et al. (2014) used LDA, integrated K-means and Principal Component Analysis, to create a knowledge organization system. The study focused in producing a more agile information retrieval system based on refined corpus that is classified with LDA and then aggregated with dimension reduction methods. Lee et al. (2015) analysed patterns of technological convergence and employ LDA as a part of their methodology, which in its majority still relies on IPC classifications. The authors used patent classifications co-citations and supported their findings with an exploratory topic modelling analysis. Zhang et al. (2014) used Latent Semantic Indexing in a case study of dye-sensitized solar cells to create technological intelligence focusing on the usefulness of pre-processing (e.g. term clumping) in producing practical results. Their analysis focused on identifying the impact of pre-processing in topic modelling, a relevant issue in utilizing unsupervised learning methods.

3. Data, pre-processing and analysis

The study uses the mobile telecommunication industry as a case study for the analysis. We took a sample of globally operating telecommunication companies, namely Alcatel-Lucent, Apple, Google, Huawei, Microsoft, Nokia and Samsung Electronics, and analysed their knowledge base with unsupervised learning, using patent data as a proxy. These companies were selected based on their expected difference between patent portfolios. Companies were expected to for groups where some are software oriented (e.g. Microsoft), some hardware (e.g. Nokia) and some have a diverse portfolio far extending telecommunication (e.g. Samsung). The analysis was expected to draw out these differences between the sample companies. The sample could have been extended significantly to other major companies in the sector. This selection, however, kept the analysis in a practical size.

For the analysis, we used a repository of full-text patent descriptions filed in the United States Patent and Trademark Office (USPTO), containing approximately six million patents. The repository, hosted by Teqmine Analytics Ltd., was searched based on one of the sample companies being mentioned as a patent applicant. Companies were searched using aggressive search terms allowing for mistakes in the names of different endings in names. False positives where controlled by manually checking a random sample of the results. The analysis was limited to the period from 2001 to 2014. For identified applicants, we extracted the full-text description of the patent, company name, year of filing and the IPC (International Patent Classification) of each patent. This data was extracted to an analysis file, containing a total of 157,718 records. We analysed the data with the process outlined below and in Fig. 1.

Prior to analysis, the full-text data was pre-processed with a Python script. The script first checked the data validity by transforming the character set to utf-8. The data was also manipulated in the Python script by removing stop words, punctuation and tokens containing numbers or consisting solely of numbers. Terms occurring only once in the whole data were also removed at this stage. After all of the before-mentioned terms had been removed, the text was tokenized and each token was transformed to a corresponding number, to further reduce the complexity of the data.

The tokenized data was analysed with an LDA algorithm implemented in Python. The algorithm is based on an online variational Bayes algorithm for LDA (Hoffman et al., 2010) that divides the corpus into chunks. The algorithm takes a chunk of the corpus as a starting point and updates the model with following chunks until the whole corpus has been analysed. The LDA algorithm is dependent on the user to input the number of topics the patent documents are classified in. Some efforts have been made to use metrics to estimate the number of latent topics in the text, such as the perplexity value (Blei et al., 2003). However, model fit methods are of limited value when humans interpret the results. In fact, "[t]raditional metrics are, indeed, negatively correlated with the measures of topic quality" (Chang et al., 2009). The number of topics should be fitted with an evaluation of real-world performance. To estimate the real-world performance of different K values, we used a trial-and-error approach, where we tested several K values and evaluated each of the result based on the coherence of terminology in a specific topic, thus resulting in 75 topics.

As a result, the algorithm created two matrices, document probabilities and word probabilities, which were used for further analysis. LDA is a soft-partitioning method, thus it creates a probability distribution for each patent to belong to one of the topics (a document probability matrix). A patent can have, for example, a large probability of belonging to two different topics. The algorithm also creates a probability distribution for each word in the corpus to belong to a certain topic (a word probability matrix).

We extracted the topic probability distribution of each document, omitting small probabilities, and created a directed network

Please cite this article as: Suominen, A., et al., Firms' knowledge profiles: Mapping patent data with unsupervised learning, Technol. Forecast. Soc. Change (2016), http://dx.doi.org/10.1016/j.techfore.2016.09.028

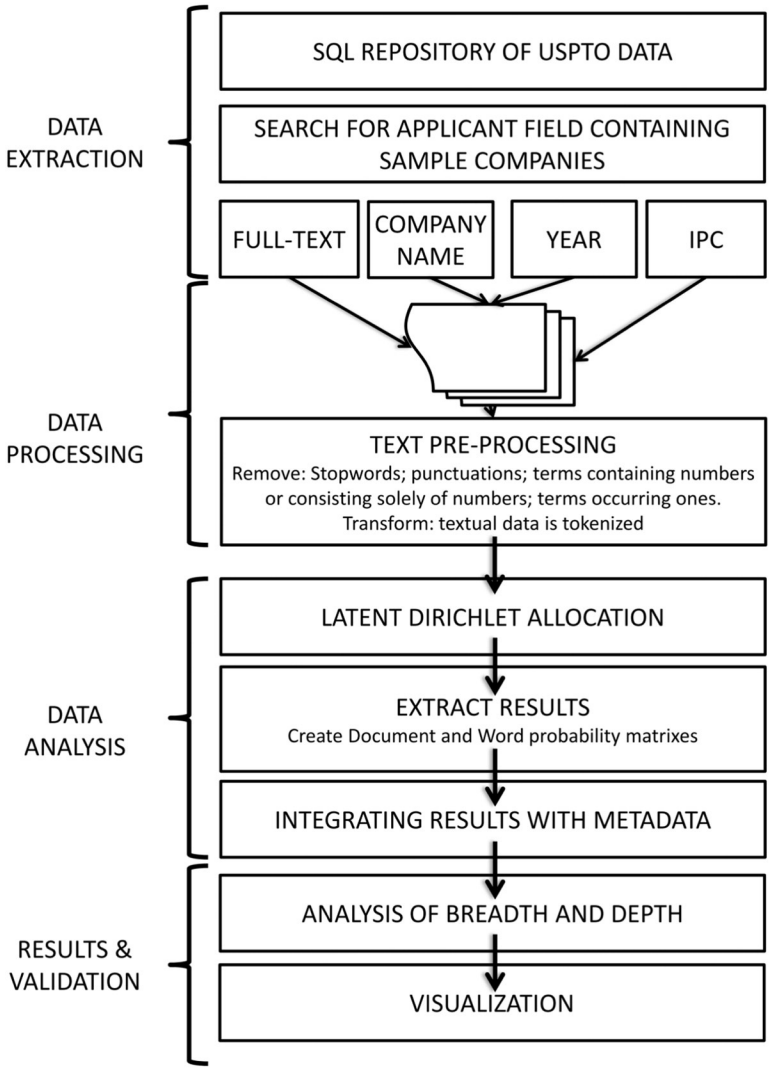


Fig. 1. Graphical representation of the research design.

dataset allowing for visual inspection of the knowledge base of the companies. In practice, we created a directed network where the nodes were latent topics created by the algorithm and individual documents in the dataset. The edges between the nodes were directed from document to topic and the weight of an edge defined by the probability of the document belonging to a certain topic. The network data was imported to Gephi network visualization software for visual inspection and calculations of basic network measures. During this visual inspection process, false positives in the sample where controlled by inspecting the network graph, topic assignment and assignee names. This process resulted in 1169 patent documents being excluded from further analysis.

We used the word probability distributions to create word clouds representing the top words for each topic. This was used to create an overview of the content of each topic. To evaluate the content and quality of each topic, we used the IPC classifications to create a co-occurrence matrix of LDA-based topics and IPC classifications. To create a more simplified structure, the IPC classes were consolidated using the concordance table into aggregated classes. Thereafter, the co-occurrence of a topic and concordance class were calculated using the topic probabilities of a patent as the weighting for topics and a whole counting scheme for concordance class weights. The sum of the product of topic probability and concordance class weights over the whole sample was used as the co-occurrence value of the topic and concordance class.

The document topic probabilities were thereafter used to calculate the technological diversification. The Herfindahl index is commonly used as a measure of diversification (Gambardella and Torrisi, 1998; Garcia-Vega, 2006; Quintana-García and Benavides-Velasco, 2008). A transformation of the Herfindahl index has previously been applied to patent portfolios, demonstrating the technological diversification of a company (Leten et al., 2007; Chiu et al., 2008). This Technological Diversity (TD) index, defined as

$$TD = 1 - \sum_{i=1}^{T} \left(\frac{N_{ij}}{N_i}\right)^2$$

is used in this study. In the definition, N_i is the sum of probabilities of company i patents over latent topics N_{ij} is the sum of probabilities of company i patents in latent topic j and T is the number of topics. A higher TD value suggests that a company is striving for a broad portfolio and a small value suggests a relatively narrow technological focus. This gives us a clue of the depth and breadth of knowledge in a company. In analysing the results, we should also note that the concentration of the dirichlet distribution in LDA is controlled by the algorithm parameters. This limits analysis to a qualitative comparison within the sample and the quantitative results cannot be used outside the confines of the study.

The sum of probabilities of a company i patents in latent topic j is also used to cluster companies and topics. The sum of topic probabilities is rescaled by company to values between 0 and 1. The rescaled values

Please cite this article as: Suominen, A., et al., Firms' knowledge profiles: Mapping patent data with unsupervised learning, Technol. Forecast. Soc. Change (2016), http://dx.doi.org/10.1016/j.techfore.2016.09.028

are transformed into percentiles to form five sample categories. The resulting matrix is clustered using hierarchical clustering based on rows and columns. The resulting matrix and clustering are visualized as a heat map, using R statistical software.

To analyse the dynamics of the knowledge base, the sum of probabilities is linked, via patent metadata, to years. This allows the analysis of allocation by year of patent documents to latent topics. We created a matrix showing the company aggregated sum of document probabilities per topic per year and used this to evaluate growing knowledge areas in the telecommunication system.

Finally, the developments of topics were forecasted. This effort exemplifies how the soft classification based aggregated topic time series can easily be extended to the future creating a forward looking aspect central to technology management. In this study we employ a grouped time series model proposed by Hyndman and Athanasopoulos (2014). The grouped forecasting method allows leveraging of the structure in the data. As the data is grouped, "the forecasts of each group must be equal to the forecasts of the individual series making up the group" (Hyndman et al., 2015). This approach allows us to ensure that the results of the analysis are consistent across the levels of aggregation (i.e. individual topics, clusters of topics or the whole sample of patents). In our data, grouping is based on the hierarchical clustering, topics, and sample companies (bottom level). This allows us to forecast knowledge trajectories of individual companies and compare against other companies in the sample, creating a managerial view on developments in different thematic areas.

4. Results

The topic probability matrix created by the algorithm was first evaluated against the patent IPC classification system, or more precisely the concordance classes created from the IPC classifications. As was expected, the highest co-occurrence sum came from the concordance classification "computer technology". This is inherent to the IPC classification system, where patents from sample companies such as the ones selected here are given a general patent class and then more specific contextual classes. As seen in the Appendix A data, the concordance classes are much more concentrated whereas the LDA produced topics are more equally distributed, even though Topic 54 can be seen as a general theme for all companies. Evaluating the LDA results in concordance classes, we used the Herfindahl index to describe whether a topic is either highly concentrated in one or two concordance classes, suggesting a similar thematic area, or if there is no strong association between the topics and concordance classes. Only five of the 75 topics have a Herfindahl index of over 0.5 (average 0.27 with st. dev. 0.129), whereas most of the topics have a significantly low association with any of the concordance classifications. This suggests that the LDA approach creates a significantly different classification for the patent sample than that obtained using IPC classes.

The graphical representations of the unsupervised learning results are shown, filtered by company, in Fig. 2. The network visualizations are created from the 75 latent topics and eight companies forming the nodes and approximately 1.7 million edges between the nodes. The network data is also available in gexf data format.[1] The network layout is based on the Force Atlas algorithm. The figure highlights the dense network created by soft clustering. Overlaying the company maps, two dense clusters of patent nodes emerge, one at the top of the figure and one in the lower left corner, which suggests that patent knowledge has two major knowledge centres. The clear layering of colours also suggest that companies have positioned themselves differently in the knowledge space.

Individual companies can be identified by filtering the graph. The eight sub-networks, seen in Fig. 2, highlight the differences in company knowledge bases. Based on a visual inspection, Alcatel-Lucent and Huawei are concentrated on the far right of Fig. 2. Google and Microsoft are concentrated in the top portion of the graph in Fig. 2 with a seemingly focused portfolio of knowledge. Apple dominates in the same portion while showing more diversity in the knowledge portfolio. Based on a visual estimation, Motorola, Nokia and Samsung operate in all the latent topics of the knowledge base seen in Fig. 2. Microsoft (Fig. 2 (e)) and Samsung (Fig. 2 (h)) clearly have the strongest portfolios with visible concentrations of patent nodes.

While not showing individual knowledge topics, the TD values operationalise the network visuals seen in Fig. 2. In Fig. 3, the TD values are plotted with the overall volume of the companies' patent portfolios. Divided into four quadrants, the majority of the companies (Alcatel-Lucent, Apple, Huawei, Motorola and Nokia) have a high-diversity portfolio but by volume, they lag significantly behind the volume of the largest IPR holder (Samsung). Microsoft and Google have a more technologically focused portfolio – lower TD values. Microsoft has a high volume of patents, specifically compared to Google and the five companies with a broader technological portfolio. Overall, the graph highlights several features embedded into the sample. Microsoft and Google are most focused on software only, clearly seen in the lower diversity. For Samsung, the breadth is to a significant extent due to Samsung business that far extends the telecommunication industry. This makes Samsung to seem as having more breadth and depth than its competitors. The companies at the lower right corner of Fig. 3 are companies with both hardware and software patents. These companies are or have been device manufacturers. Excluding Apple, all have activities in networks. All of them also have patents in software. Fig. 3 highlights the fact that the sample companies are positioned differently in the space between volume and diversity, which suggests different focus points in the knowledge space.

In Fig. 4, the company TD profile is merged with the latent patterns found in the unsupervised learning process. Hierarchical clustering is used to cluster latent patterns and companies based on the probability of patent assignments. In Fig. 4, the latent topics are aggregated on the basis of a qualitative assessment of words occurring with a high probability in the topics. The seven aggregate topics seen in the figure are labelled, starting from the top, as:

1. Cluster 1: multipurpose technologies: flash memory, displays, testing, switches, battery materials, semiconductor, LEDs
2. Cluster 2: mechanics, image rendering, mixed topics
3. Cluster 3: device functionalities (e.g. video, camera, location, maps, human interaction) and relevant data processing and indexing
4. Cluster 4: radio hardware technologies (base station, terminal, antenna, battery, transmission)
5. Cluster 5: electronics (analog) and hardware assembly
6. Cluster 6: network data, devices, and authentication
7. Cluster 7: networks

In Fig. 4, the heat map is normalized volume-wise showing the knowledge focus of an individual company. Nearly all of the companies, excluding Samsung, have a strong focus on the core network (Cluster 7) and network traffic (Cluster 6). Differences between the companies are apparent in either having the focus on device functionalities (Cluster 3) or radio hardware technologies (Cluster 4), and we see a separation of Google, Apple and Microsoft from Huawei, Nokia, Alcatel-Lucent, and Motorola. In the clustering, Samsung remains as an isolate with notable different focus.

In Cluster 5, electronics and assembly, and Cluster 1, multipurpose technologies, Samsung asserts a strong presence. For Samsung this is explained by the significant breadth of the Samsung Electronics business portfolio, which extends far beyond telecommunication. Cluster 2, mechanics, image rendering, mixed topics, is a selection of latent patterns not clustered in any other classes, where most companies have a limited presence.

[1] http://arhosuominen.fi/ICT/Telecommunication_data.gexf

Please cite this article as: Suominen, A., et al., Firms' knowledge profiles: Mapping patent data with unsupervised learning, Technol. Forecast. Soc. Change (2016), http://dx.doi.org/10.1016/j.techfore.2016.09.028

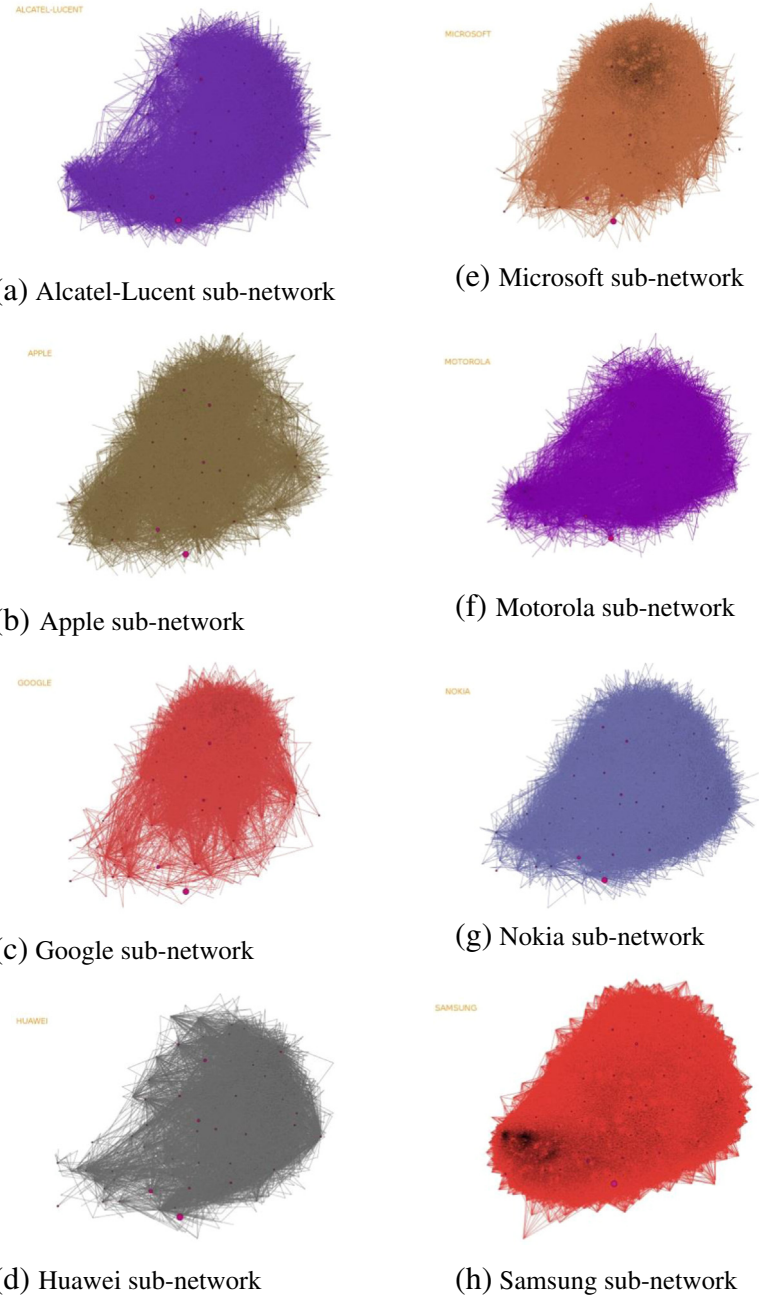


Fig. 2. Filtered sub-networks of individual companies.

Overall, the latent topics show that there is a clear qualitative difference in the knowledge foci of the companies and the volume and diversity of the portfolios further differentiate the companies. Looking at the column-wise dendrogram, Google and Microsoft are clustered together. The four device/network manufacturers are also merged together.

The temporal dynamics of the clusters is analysed further by aggregating the document topic probabilities on a yearly basis. To simplify the visualization, Fig. 5 shows the relevance of each cluster in Fig. 4. Cluster 6, network data, devices and authentication, grows in relative importance, with a significant increase in 2005. At the same time, Cluster 4, radio hardware technologies (base station, terminal, antenna, battery, transmission), and Cluster 5, electronics (analog) and hardware assembly, have diminished in relative importance as a knowledge domain. This clear temporal change in knowledge domain can be attributed to the clear transition from a hardware-driven industry to the increased importance of software. We could argue that our results show how hardware and assembly related development has move to subcontractors, such as Foxconn, while the companies selected to this study have moved their focus towards the more value adding areas.

Looking behind the aggregate clusters, Fig. 6 shows the word clouds of the two highest growing latent topics (Fig. 6 (a) and Fig. 6 (b)) and the two topics with the largest decrease in patenting (Fig. 6 (c) and Fig. 6 (d)). The growth value was calculated based on average of growth between 2011 and 2012 and 2012–2013 to capture recent and stable growth pattern. The highest growth topic, with on average a 24% growth, focuses on navigation and map technologies. The second highest area of growth, with a growth percentage of 16, is sensor technologies. Both of the decreasing topics relate to auxiliary devices, e.g. printing, which decreased by over 16% yearly. At an aggregate level, the temporal data shows a strong increase in software-based technologies and a decrease in hardware patenting, although niche development

Please cite this article as: Suominen, A., et al., Firms' knowledge profiles: Mapping patent data with unsupervised learning, Technol. Forecast. Soc. Change (2016), http://dx.doi.org/10.1016/j.techfore.2016.09.028

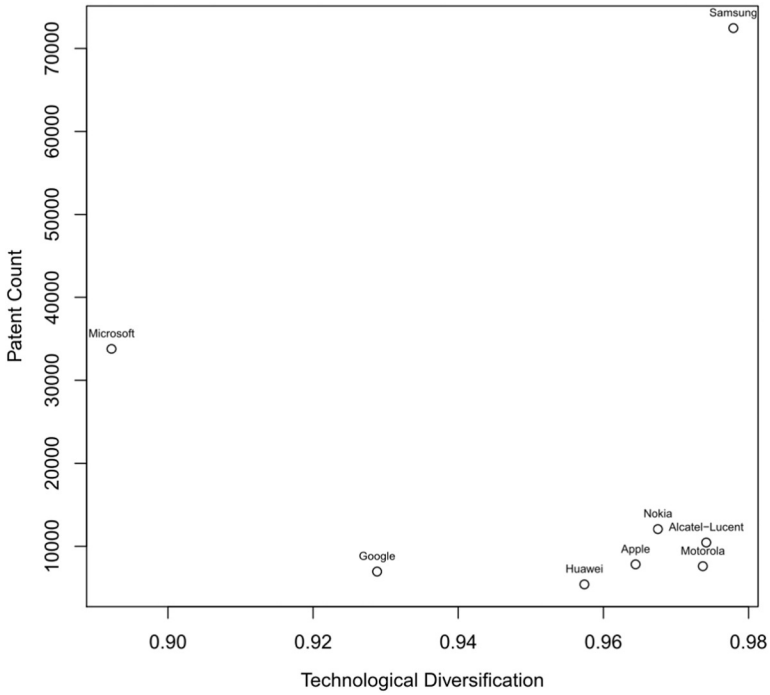


Fig. 3. Scatter plot of company patent portfolio size and technological diversity.

areas in fields such as sensors are visible. Patenting for auxiliary devices and hardware components in general decreased.

By merging the graphical illustration in Figs. 4 and 6, we are able to quantify the dynamics of knowledge domains among companies. The majority of the companies are well positioned in the growth of Cluster 6, with significant portfolios in the area. Strong portfolios in Clusters 2,4 and 5 seem to have lost relevance. Cluster 3 has exhibited a pattern of growth, but not strong growth. The relative patent position of different companies in high growth areas varies. As seen in Fig. 5, companies' knowledge portfolios have different emphasis. A visual inspection highlights a block of knowledge shared by neighbouring companies.

To fully capture the firm level profiles, we need to add a forecasting dimension that can describe if a company is currently building an increasing portfolio in a topical area or decreasing presence in some. At a firm level we are able to produce insights that can be used for managerial decision making on future knowledge investments – comparing current and forecasted knowledge profiles against competitors. A grouped time series model was used to estimate the trend of patenting overall, in clusters and by topic. Fig. 7 shows the behaviour of the time series at the top of the hierarchy (a) and on a cluster level (b). Individual topics are not shown to keep the figure at a practical size. The ARIMA model grouped time series was used to forecast ten years forward, and it clearly shows an increasing patenting in Cluster 6, while the other clusters are at near zero growth or decrease. Clusters 1 and 5, focusing on hardware, seem to decrease in their relative importance the most.

A more in depth view can be taken at a topic level. Taking high growing topics from Fig. 7, we can describe the dynamics of knowledge profiles at a firm/topic level. Fig. 8 (a) shows the second highest growing topic. It shows a pattern where Samsung and Microsoft have had a position, but since reduced investments in the area. Google and Apple show sharp increase since 2012 (Google) and 2013 (Apple). Apple has reduced its investments, while Google invests heavily and, as seen from the estimates, by the end of the time series dominates the topic. Fig. 8 (b) highlights a topic where Samsung has a strong presence, but where forecast suggest the position to be challenged by the investments by Apple and Google.

5. Discussion

Connecting patent information to industry is a challenge (Schmoch, 2008) and elaborating on the underlying changes in knowledge profiles an even harder one. However, focusing on strategic foresight and the dynamic capabilities of a firm, we need to be able to quantify the knowledge resources embedded in an organization. The existing information retrieval based classification system has a limited value in this effort (Loh et al., 2006) and methods that can go beyond using patent classes can open a more dynamic view of the knowledge landscape.

By using an unsupervised learning based approach to quantifying the knowledge profiles of the sample companies, a holistic view of the knowledge profiles in the sample companies can be produced. As we have shown, differences emerge between software-oriented companies (such as Google and Microsoft) and technology-driven firms (such as Nokia or Huawei), underlining that they have a different focus in their knowledge base. By connecting the temporal dimension to the analysis, we were able to show the systemic transition of the telecommunication industry towards a software-driven knowledge frame, but were also able to detect those hardware-related areas that are growing against the overall trend.

Most importantly, topic modelling created classification departs from the IPC classification of patents, producing a very different view of the knowledge of companies. We argue that, by employing unsupervised learning, it is possible to create a more accurate description of the knowledge companies actually possess. In our experience, this requires running rather large full-text datasets that describe the information embedded in companies thoroughly. This is of specific importance with patents, where the abstracts are seldom a good representation of what is actually described in the patent.

The managerial implications of the study suggest that we should be aware of the fact that the objectives of analysis determine the suitability of different approaches. The machine learning approaches have inherent advantages, of which most important ones are the versatility and agility of analysis: Any large data set can easily be analysed from a range of perspectives, whereas the labour intensiveness of traditional methods is a highly restricting factor. Unsupervised learning methods are also able to detect more reliably latent patterns, such as emerging

Please cite this article as: Suominen, A., et al., Firms' knowledge profiles: Mapping patent data with unsupervised learning, Technol. Forecast. Soc. Change (2016), http://dx.doi.org/10.1016/j.techfore.2016.09.028

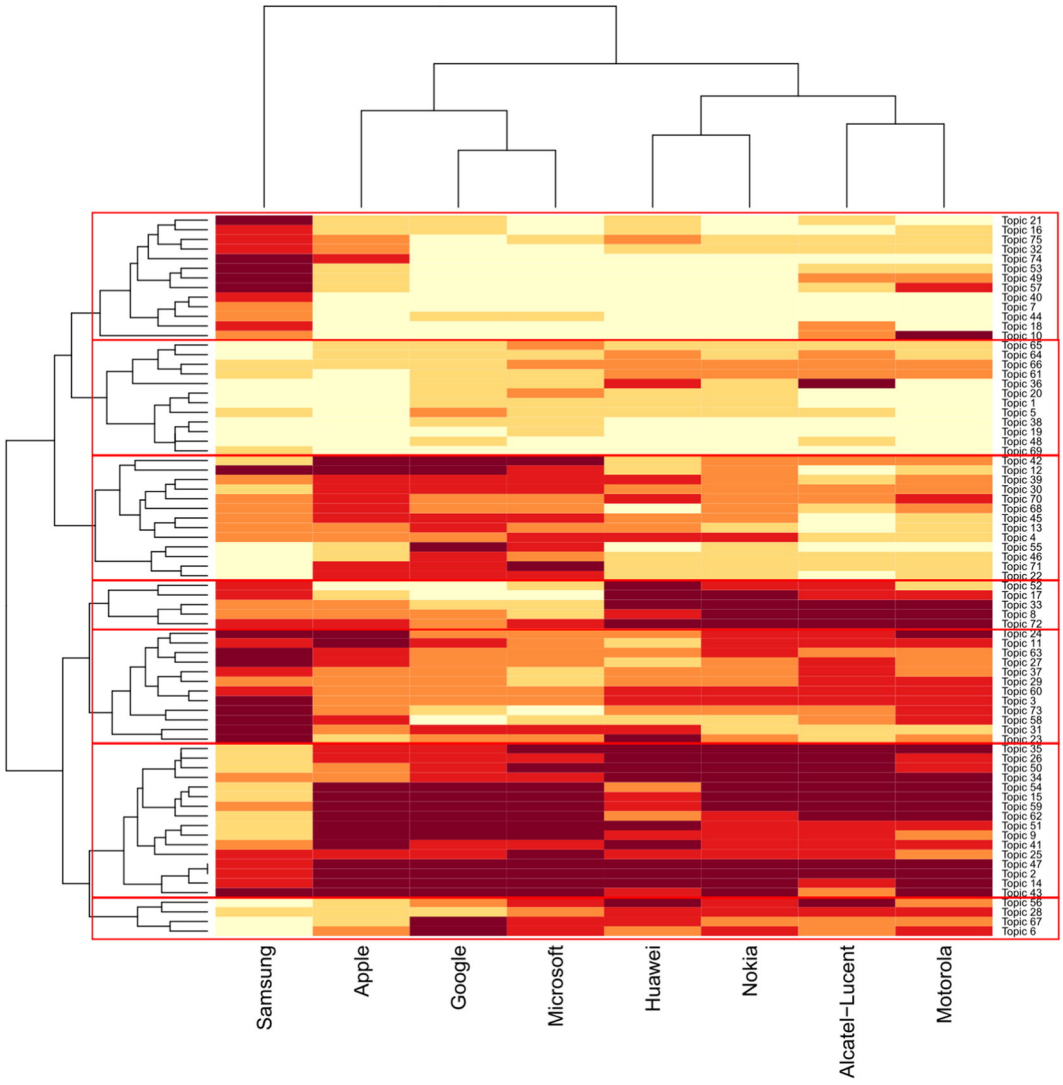


Fig. 4. Heat map of company focus on latent topics emerging from the unsupervised learning process. Dendrograms are based on hierarchical clustering. Red frames highlight the clustered topics referred to in the text.

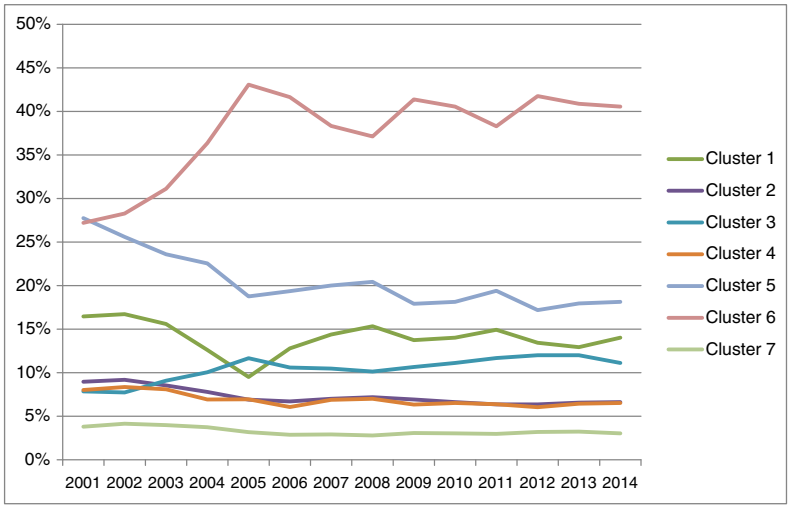
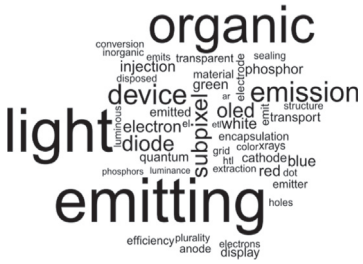


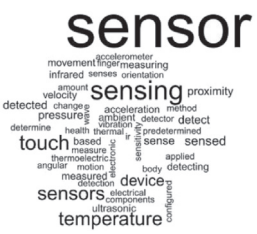
Fig. 5. Relative yearly importance of each of the seven aggregate clusters.

Please cite this article as: Suominen, A., et al., Firms' knowledge profiles: Mapping patent data with unsupervised learning, Technol. Forecast. Soc. Change (2016), http://dx.doi.org/10.1016/j.techfore.2016.09.028

(a) Highest average growth in the last three years (Topic 46)



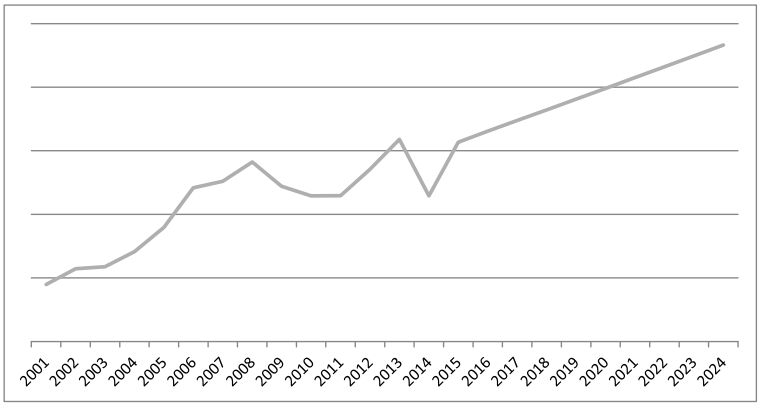
(c) Highest average decrease in the last three years (Topic 7)



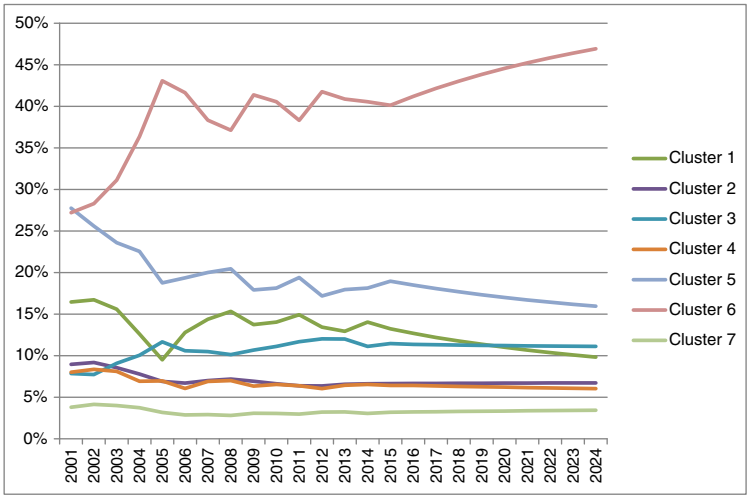
(b) Second highest average growth in the last three years (Topic 68)

(d) Second highest average decrease in the last three years (Topic 69)

Fig. 6. Word clouds of latent topics with highest growth (a and b) and decrease (c and d) in patenting within the three last years.



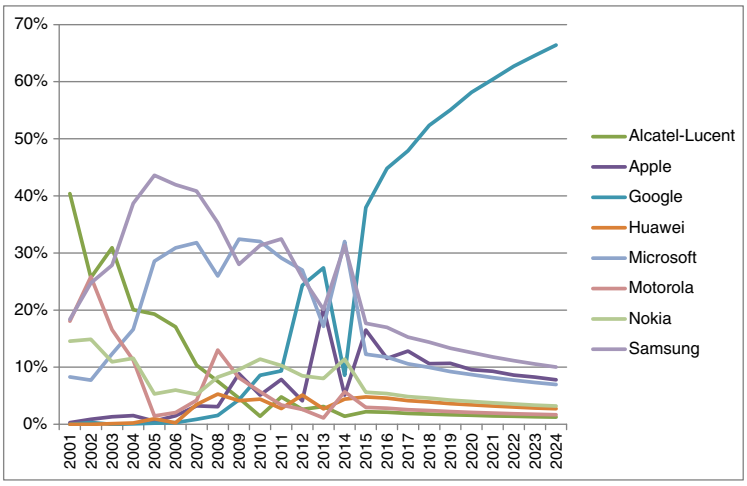
(a) Trend of patenting for the sample. Data until 2014 and from 2015 forecasted with grouped time series model.



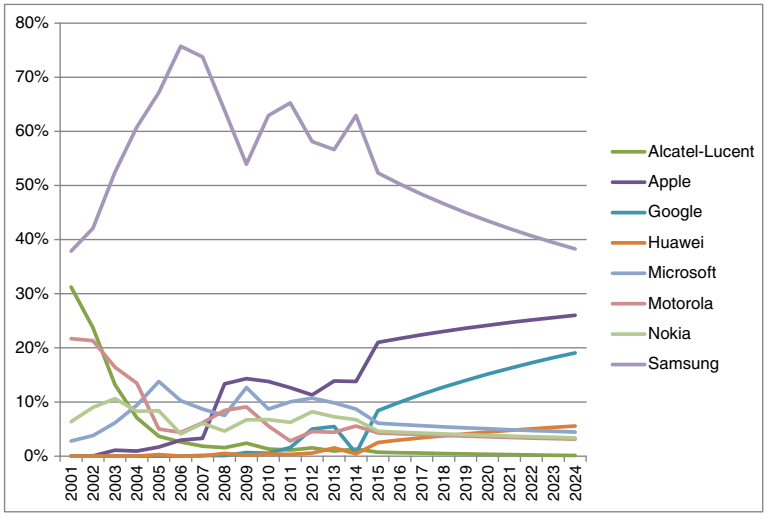
(b) Relative yearly importance of each of the seven aggregate clusters. Data until 2014 and from 2015 forecasted with grouped time series model.

Fig. 7. Grouped time series model of the LDA-generated classifications at the top of hierarchy and cluster level.

Please cite this article as: Suominen, A., et al., Firms' knowledge profiles: Mapping patent data with unsupervised learning, Technol. Forecast. Soc. Change (2016), http://dx.doi.org/10.1016/j.techfore.2016.09.028



(a) Relative shares of company patents in highest average growthtopic (Topic 46). Data until 2014 and from 2015 forecasted with grouped time series model.



(b) Relative shares of company patents in second highest average growth topic (Topic 68). Data until 2014 and from 2015 forecasted with grouped time series model.

Fig. 8. Grouped time series model of the LDA-generated classifications at a topic level.

technologies, whereas human labelling dependent approaches often apply historical classification models to new-to-the-world phenomenon (Suominen and Toivanen, 2015). This has implications for corporate patent portfolio management and for example a mergers and acquisitions situation, where a view on the complementarity of patent portfolios can be assessed. A recent example from our sample firms is the acquisition of Alcatel-Lucent by Nokia (confirmed in early 2016) – the knowledge repositories of these firms complement each other well as can be seen from the dendrograms in Fig. 4.

However, this approach has its limitations. The analysis requires access to patent data that has been pre-processed for automated processing. During the patenting process, specifically in the case of large companies, the semantic text in the patent documents is to a large extent written by patent professionals, such as patent attorneys. Even though the core explanation of the invention is done by the inventors, on most cases the patent professional will rewrite significant portions of the text so that it is in line with legislation and patent practises. This process impacts the natural language processing, as aggressive pre-processing steps are needed to limit the classification to the text describing the invention not the process. Even with extensive pre-processing some of the topics might be more the results of the patenting process than actual descriptions of a company's knowledge profile. Another limitation to the approach used in this paper is that it requires a fairly large dataset to produce practical results. As Leydesdorff and Nerghes (2015) point out, topic modelling seems to be an unpractical approach to smaller datasets, where traditional methods yield better results. We argue that for a large dataset unsupervised learning methods are able to reduce the complexity of the data and more easily produce meaningful results for human interpretation. The results in the paper are also limited by the sample selected for the analysis. At this stage, we confined the analysis to the selected companies, omitting a large portion of the telecommunication industry. For example, Nokia Corporation's small focus on map technologies is easily explained by the fact that this IPR was or is held at a subsidiary, Navionics. In addition, our analysis has omitted a number of large and small companies known to have relevant IPR.

Altogether, we should consider the role played by input data. This paper has argued that using full-text patent descriptions, ranging from few to very numerous pages, has its advantages for the accuracy of text analysis over short abstract or claims data. Patent abstracts are easily purposefully obfuscated, or easily lack much of the relevant information, and therefore offer highly truncated window to the complexity embodied in patent landscapes. An abstract and keyword driven approach also loses much of the narrative of natural language. This can

Please cite this article as: Suominen, A., et al., Firms' knowledge profiles: Mapping patent data with unsupervised learning, Technol. Forecast. Soc. Change (2016), http://dx.doi.org/10.1016/j.techfore.2016.09.028

mean losing information on the application area of an invention or its origin. In addition, the analysis shown here is at its optimal level, run using all the patents in the repository and extracting latent patterns relevant to telecommunication. As noted before, larger more complex datasets are more conveniently structured using e.g. topic modelling, while small datasets are more conveniently structured using other methods. The results of such a broad analysis are challenging to present within the confines of an article. However, the research design suggested in this paper demonstrates the advantage of an automated process for integrating big data into foresight processes.

Further studies should focus on applying other more tailored algorithms in patent data. Other approaches such as Dynamic Topic Model and the Document Influence Model could add value in modelling patent data as these approaches can take into consideration for example changes in time. A rigorous analysis of which algorithms produce the best qualitative result for human analysis with a specific data would be valuable in analysing the capabilities of different algorithmic options.

Acknowledgements

Tekes—the Finnish Funding Agency for Technology and Innovation-supported this work with the research grant (2004/31/2011) awarded for the project “Co-evolution of knowledge creation systems and innovation pipelines”, and research grant (3431/31/2014) awarded to the “Radical and incremental innovation in industrial renewal”, as well as the Academy of Finland research grant (288609) “Modeling Science and Technology Systems Through Massive Data Collections”. Patent data was obtained from the Teqmine Analytics Ltd. patent databases, and it provided also computing resources for the analysis. The funders or data and computing resource providers had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. A previous version of the paper has been presented at the 2015 PICMET conference.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.techfore.2016.09.028.

References

Archibugi, D., Planta, M., 1996. Measuring technological change through patents and innovation surveys. Technovation 16 (9), 451–519.
Ayres, R.U., 1989. The future of technological forecasting. Technol. Forecast. Soc. Chang. 36 (1), 46–60.
Blei, D.M., Lafferty, J.D., 2007. A correlated topic model of science. Ann. Appl. Stat. 1 (1), 17–35.
Blei, D.M., Lafferty, J.D., 2009. Text Mining: Classification, Clustering, and Applications. 10 ed. Taylor and Francis chapter Topic Models, pp. 71–94.
Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.
Breschi, S., Lissoni, F., Malerba, F., 2003. Knowledge-relatedness in firm technological diversification. Res. Policy 32 (1), 69–87.
Cascini, G., Russo, D., Zini, M., 2007. Computer-aided patent analysis: Finding invention peculiarities. Trends in computer aided innovation. Springer, US, pp. 167–178.
Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., Blei, D.M., 2009. Reading tea leaves: How humans interpret topic models. Advances in neural information processing systems, pp. 288–296.
Chiu, Y.-C., Lai, H.-C., Lee, T.-Y., Liaw, Y.-C., 2008. Technological diversification, complementary assets, and performance. Technol. Forecast. Soc. Chang. 75 (6), 875–892.
Dahlin, K.B., Behrens, D.M., 2005. When is an invention really radical?: Defining and measuring technological radicalness. Res. Policy 34 (5), 717–737.
Gambardella, A., Torrisi, S., 1998. Does technological convergence imply convergence in markets? Evidence from the electronics industry. Res. Policy 27 (5), 445–463.
Garcia-Vega, M., 2006. Does technological diversification promote innovation?: An empirical analysis for European firms. Res. Policy 35 (2), 230–246.
George, G., Kotha, R., Zheng, Y., 2008. Entry into insular domains: a longitudinal study of knowledge structuration and innovation in biotechnology firms. J. Manag. Stud. 45 (8), 1448–1474.
Glenisson, P., Glänzel, W., Persson, O., 2005. Combining full-text analysis and bibliometric indicators. A pilot study. Scientometrics 63 (1), 163–180.
Herrero, Á., Corchado, E., Saiz, L., Abraham, A., 2010. DIPKIP: a connectionist knowledge management system to identify knowledge deficits in practical cases. Comput. Intell. 26 (1), 26–56.
Hoffman, M., Bach, F.R., Blei, D.M., 2010. Online Learning for Latent Dirichlet Allocation. In Advances in Neural Information Processing Systems, pp. 856–864.
Hu, Z., Fang, S., Liang, T., 2014. Empirical study of constructing a knowledge organization system of patent documents using topic modeling. Scientometrics 100 (3), 787–799.
Huang, Y., Zhang, Y., Ma, J., Porter, A., Wang, X., 2015. Tracing Technology Evolution Pathways by Combining Tech Mining and Patent Citation Analysis. 2015 Portland International Conference on Management of Engineering & Technology.
Hyndman, R.J., Athanasopoulos, G., 2014. Optimally reconciling forecasts in a hierarchy. Foresight Int. J. Appl. Forecast. 35, 42–48.
Hyndman, R.J., Athanasopoulos, G., Shang, H.L., 2015. hts: An R Package for Forecasting Hierarchical or Grouped Time Series. R Package (Available at http://cran.unej.ac.id/web/packages/hts/vignettes/hts.pdf).
Jaffe, A.B., 1986. Technological Opportunity and Spillovers of R&D: Evidence from firms' Patents, Profits and Market Value.
Kajikawa, Y., Abe, K., Noda, S., 2006. Filling the gap between researchers studying different materials and different methods: a proposal for structured keywords. J. Inf. Sci. 32 (6), 511–524.
Kang, I.S., Na, S.H., Kim, J., Lee, J.H., 2007. Cluster-based patent retrieval. Inf. Process. Manag. 43 (5), 1173–1182.
Kauffman, S., Lobo, J., Macready, W.G., 2000. Optimal search on a technology landscape. J. Econ. Behav. Organ. 43 (2), 141–166.
Kim, J.H., Choi, K.S., 2007. Patent document categorization based on semantic structural information. Inf. Process. Manag. 43 (5), 1200–1215.
Kuusi, O., Meyer, M., 2007. Anticipating technological breakthroughs: using bibliographic coupling to explore the nanotubes paradigm. Scientometrics 70 (3), 759–777.
Lai, K.-K., Wu, S.-J., 2005. Using the patent co-citation approach to establish a new patent classification system. Inf. Process. Manag. 41 (2), 313–330.
Lee, W.S., Han, E.J., Sohn, S.Y., 2015. Predicting the pattern of technology convergence using big-data technology on large-scale triadic patents. Technol. Forecast. Soc. Chang. 100, 317–329.
Leten, B., Belderbos, R., Van Looy, B., 2007. Technological diversification, coherence, and performance of firms. J. Prod. Innov. Manag. 24 (6), 567–579.
Leydesdorff, L., Nerghes, A., 2015. Co-Word Maps and Topic Modeling: A Comparison from a User's Perspective (arXiv preprint arXiv:1511.03020).
Loh, H.T., He, C., Shen, L., 2006. Automatic classification of patent documents for TRIZ users. World Patent Inf. 28 (1), 6–13.
McNamee, R.C., 2013. Can't see the forest for the leaves: similarity and distance measures for hierarchical taxonomies with a patent classification example. Res. Policy 42 (4), 855–873.
Mendonça, S., 2006. The revolution within: ICT and the shifting knowledge base of the world's largest companies. Econ. Innov. New Technol. 15 (8), 777–799.
Moed, H., De Bruin, R., Van Leeuwen, T.H., 1995. New bibliometric tools for the assessment of national research performance: database description, overview of indicators and first applications. Scientometrics 33 (3), 381–422.
Moorthy, S., Polley, D.E., 2010. Technological knowledge breadth and depth: performance impacts. J. Knowl. Manag. 14 (3), 359–377.
Nakamura, H., Suzuki, S., Sakata, I., Kajikawa, Y., 2015. Knowledge combination modeling: the measurement of knowledge similarity between different technological domains. Technol. Forecast. Soc. Chang. 94, 87–201.
Nemet, G.F., 2009. Demand-pull, technology-push, and government-led incentives for non-incremental technical change. Res. Policy 38 (5), 700–709.
Porter, A.L., Newman, N.C., 2011. Mining external R&D. Technovation 31 (4), 171–176.
Quintana-García, C., Benavides-Velasco, C.A., 2008. Innovative competence, exploration and exploitation: the influence of technological diversification. Res. Policy 37 (3), 492–507.
Richter, G., MacFarlane, A., 2005. The impact of metadata on the accuracy of automated patent classification. World Patent Inf. 27 (1), 13–26.
Schmoch, U., 2008. Concept of a Technology Classification for Country Comparisons. Final Report to the World Intellectual Property Organisation. WIPO.
SubbaNarasimha, P.N., Ahmad, S., Mallya, S.N., 2003. Technological knowledge and firm performance of pharmaceutical firms. J. Intellect. Cap. 4 (1), 20–33.
Subramaniam, M., Youndt, M.A., 2005. The influence of intellectual capital on the types of innovative capabilities. Acad. Manag. J. 48 (3), 450–463.
Suominen, A., 2013. Analysis of technological progression by quantitative measures: a comparison of two technologies. Tech. Anal. Strat. Manag. 25 (6), 687–706.
Suominen, A., Toivanen, H., 2015. Map of science with topic modeling: comparison of unsupervised learning and human-assigned subject classification. J. Assoc. Inf. Sci. Technol. (Early view).
Tseng, Y.H., Lin, C.J., Lin, Y.I., 2007. Text mining techniques for patent analysis. Inf. Process. Manag. 43 (5), 1216–1247.
Van Merkerk, R., Van Lente, O.H., 2005. Tracing emerging irreversibilities in emerging technologies: the case of nanotubes. Technol. Forecast. Soc. Chang. 72 (9), 1094–1111.
Ventura, S.L., Nugent, R., Fuchs, E.R., 2015. Seeing the non-stars: (some) sources of bias in past disambiguation approaches and a new public tool leveraging labeled records. Res. Policy 0.44 (9), 1672–1701.
Venugopalan, S., Rai, V., 2015. Topic based classification and pattern identification in patents. Technol. Forecast. Soc. Chang. 94, 236–250.
Wang, Q., Von Tunzelmann, N., 2000. Complexity and the functions of the firm: breadth and depth. Res. Policy 29 (7), 805–818.

Please cite this article as: Suominen, A., et al., Firms' knowledge profiles: Mapping patent data with unsupervised learning, Technol. Forecast. Soc. Change (2016), http://dx.doi.org/10.1016/j.techfore.2016.09.028

Wu, J., Shanley, M.T., 2009. Knowledge stock, exploration, and innovation: research on the United States electromedical device industry. J. Bus. Res. 62 (4), 474–483.
Yau, C.-K., Porter, A.L., Newman, N., Suominen, A., 2013. Clustering scientific documents with topic modeling. Scientometrics 00 (3), 767–786.
Zhang, J., Baden-Fuller, C., 2010. The influence of technological knowledge base and organizational structure on technology collaboration. J. Manag. Stud. 47 (4), 679–704.
Zhang, J., Baden-Fuller, C., Mangematin, V., 2007. Technological knowledge base, R&D organization structure and alliance formation: evidence from the biopharmaceutical industry. Res. Policy 36 (4), 515–528.
Zhang, Y., Porter, A.L., Hu, Z., Guo, Y., Newman, N.C., 2014. "Term clumping" for technical intelligence: a case study on dye-sensitized solar cells. Technol. Forecast. Soc. Chang. 85, 26–39.

Dr. Arho Suominen, PhD, is a Senior Scientist at the Innovations, Economy, and Policy unit at the VTT Technical Research Centre of Finland, also lecturing at the Department of Information Technology at the University of Turku. Dr. Suominen's research focuses on qualitative and quantitative assessment of emerging technologies and innovation management specifically ecosystems. His research has been funded by the Academy of Finland, the Finnish Funding Agency for Technology, Turku University Foundation and the Fulbright Center Finland. Dr. Suominen has published work in several journals such as the Journal of the Association for Information Science and Technology, Science and Public Policy, Scientometrics, Journal of Systems and Software, Futures, and Foresight. Dr. Suominen has a Doctor of Science (Tech.) degree from the University of Turku and holds an Officers basic degree from the National Defence University of Finland.

Dr. Hannes Toivanen is the CEO and co-founder of Teqmine Analytics Ltd., a patent and technology intelligence company. He specializes in technology and patent analysis, technology forecasting, and business development. Toivanen is also affiliated with the Lappeenranta University of Technology. He has previously worked as principal scientist at VTT - Technical Research Centre of Finland, where he led research groups that investigated globalization of innovation, innovation in the emerging economies & developing countries, and quantitative analysis of science and technology. He was a Fulbright Fellow 2000–2004 and received his PhD from the Georgia Institute of Technology in 2004.

Prof. Marko Seppänen, PhD, is a Full Professor in the field of industrial management in Tampere University of Technology, Finland. Prof. Seppänen is an expert in managing value creation in business ecosystems, business concept development and innovation management. In his latest research he has examined e.g. platform-based competition in business ecosystems, and innovation management in business networks. His research has appeared in high quality peer-reviewed journals such as the Journal of Product Innovation Management, Technological Forecasting and Social Change, Journal of Systems and Software, and International Journal of Physical Distribution & Logistics Management.

Please cite this article as: Suominen, A., et al., Firms' knowledge profiles: Mapping patent data with unsupervised learning, Technol. Forecast. Soc. Change (2016), http://dx.doi.org/10.1016/j.techfore.2016.09.028