Simulating Patent Knowledge Contexts

Jussi Kantola and Aviv Segev
Department of Knowledge Service Engineering
KAIST - Korea Advanced Institute of Science and Technology
291 Daehak-ro, Yuseong-gu, Daejeon, Korea
{jussi, aviv}@kaist.edu

Abstract — **Patent users such as government, inventors, and manufacturing organizations strive to identify the directions in which the new technology is advancing. The organization of patent knowledge in maps aims at outlining the boundaries of existing knowledge. This article demonstrates the methodology for simulating alternative knowledge contexts beyond the border of existing knowledge. The process starts with extracting knowledge from patents and applying self-organizing maps for presenting knowledge. The knowledge extraction model was tested earlier on patents from the United States Patent and Trademark Office. A demonstrator tool is then used to perform "what-if" type of analysis/simulation on the clusters in the dataset to see alternative knowledge contexts for the new knowledge "entity". This may open up new directions and help to plan for the future. The demonstrator tool has been tested earlier on other datasets. The proposed knowledge context simulation shows promise for the future development and applications.**

Keywords: Knowledge context; patent; self-organizing map; simulation; SIMU_SOM

I. INTRODUCTION

Government services attempt to forecast main research areas that would be beneficial to fund. Similarly, researchers try to map knowledge and identify possible gaps that would be relevant to the advancement of science. The extraction of relevant information from patents allows the analysis of main research areas and the mapping of the current topics of interest. The creation of such a service, which allows analysis of patents over time, will provide decision makers with a top level overview of the direction of new inventions. In addition, the service could support knowledge seekers in identifying worthwhile research tasks. A knowledge map service can enable a researcher to identify the need for specific research directions considered "hot". In addition, research and government institutions providing funding will be able to preplan with a longer horizon and divert research funds to necessary fields. Knowledge maps of patents can assist in the classification of directions of research in the past and in the attempt to predict future discovery directions.

The patent service is unique compared to other knowledge based services because of the requirement to identify whether similar knowledge exists as opposed to the need to locate knowledge. Contemporary knowledge based services are based on using existing information, while the patent support service is required to assist in identifying similar domains and patterns that would result in the rejection of a patent request. Furthermore, patents in different countries are not classified under one classification system.

The premise of the patent system lies in its mutual benefit to both the inventor and the public. In return for full public disclosure, a patent offers certain rights to an inventor for a limited period of time, during which the inventor may exclude all others from making, using, importing, or selling his or her invention. The patent is published and disseminated to the public so that others may study the invention and improve upon it. The constant evolution of science and technology, spurred by the monetary incentive the patent system offers to inventors, strengthens the economy. New inventions lead to new technologies, create new jobs, and improve our quality of life.

The work analyzes patents to create an outline of knowledge. The research aims at building a simulation model that predicts the identification of new critical research areas that can exponentially speed up the overall research in specific fields. The patent project analyzes patents from the United States Patent and Trademark Office. The patent analysis process is the following: Existing patents -> Patent knowledge extraction -> Knowledge representation using Self-Organizing Maps -> Knowledge representation analysis / simulation. The first step includes parsing existing patents text. In the analyzed cases the entire patent description was used. Alternative methods include parsing the patents according to dates or according to topics. The model includes three major steps: patent knowledge extraction and knowledge representation using self-organizing maps. The patent knowledge extraction extracts key features from each patent. The knowledge representation creates an evolving map using the self-organizing map technique to represent the patents research topics. The last step involves the analysis /simulation of the knowledge representation map evolution.

One benefit of using simulation is the insight gained into the importance of variables and their interaction [2] [8]. Knowledge elements in the existing patents resemble those variables. Another benefit is the possibility to experiment with new policies before their implementation [2] [8], thus saving both money and time. Knowledge contexts can be tested before implementation. Simulation allows answering "what-if" questions that are important when new systems are being developed [2] [8]. These benefits seem attractive for knowledge context simulation as well.

The next section describes the Self-Organizing Maps. Section 3 describes the SIMU_SOM demonstrator tool. Section 4 describes knowledge extraction process and section 5

describes patent knowledge context simulation approach Section 6 presents a discussion and some concluding remarks.

II. Self-Organizing Maps

The Self-Organizing Map (SOM) is a two-layer unsupervised neural network that maps multidimensional data onto a two dimensional topological grid [6]. The data are grouped according to similarities and patterns found in the dataset, using some form of distance measure, usually the Euclidean distance. The results are displayed as a series of nodes on the map, which can be divided into a number of clusters based upon the distances between the clusters. Since the SOM is unsupervised, no target outcomes are provided, and the SOM is allowed to freely organize itself, based on the patterns identified, making the SOM an ideal tool for exploratory data analysis [3].

According to Kaski and Kohonen [5], exploratory data analysis methods, such as SOM, are like general-purpose instruments that illustrate the essential features of a data set, such as its clustering structure and the relations between its data items. The SOM perform visual clustering of data [3]. More information about the methodology of applying self-organizing maps is provided by Back et al. [1]. The most commonly used method for visualizing the final self-organizing map is the unified distance matrix method, or U-matrix [11]. The U-matrix method can be used to discover otherwise invisible relationships in a high-dimensional data space. It also makes it possible to classify data sets into clusters of similar values. Feature planes, representing the values in a single vector column, are used to identify the characteristics of these clusters [3]. This helps in explaining the meaning of the SOM.

III. SIMU_SOM Demonstrator

To display the interactive approach on SOMs, the SIMU_SOM demonstrator tool was constructed at Tampere University of Technology, Finland by Vesanen, Toivonen and Visa. SIMU_SOM is described in [4]. The demonstrator allows the user to interactively view and comprehend the structure of the constructed SOM and to perform sensitivity analysis on the map. The prototype was coded in the Linux operating system with the Perl Toolkit (Perl). The SIMU_SOM application takes a constructed SOM vector file as its input and draws the map for it. The vector file contains the numerical results of the whole group. Figure 1 shows a sample SIMU_SOM screenshot to illustrate the elements of the demonstrator.

For each variable in the dataset a slider is created and presented on the right side of the window. In this case, patent context elements are the variables. The user can analyze the effect of each variable on the map position using sliders to change parameter values. A pointer, a white ball, changes its position to the closest matching node of the map as the user changes the values of the variables. The pointer shows the simulated map position of the patent application. The closest match is determined using the smallest Euclidean distance between the map node vector and the current value vector of the sliders. The labels that belong to the current position of the pointer are shown below the map. In Figure 1, the arrows can be interpreted as follows: the starting point of an arrow is the current map position and the end of each arrow is the simulated optional target map position. This means that a person can roughly see where incremental increases in the level of context match put the new knowledge entity.

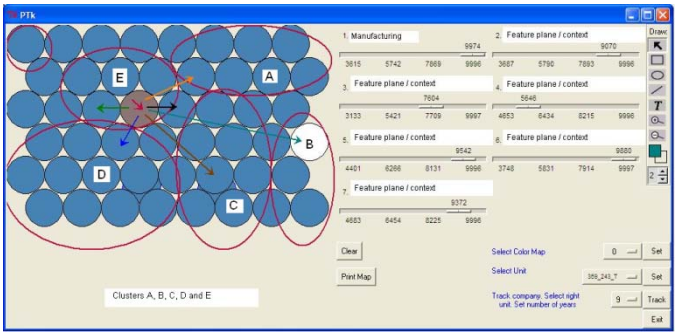

Figure 1. The SIMU_SOM demonstrator has the SOM and feature plane sliders [4].

IV. Patent Knowledge Extraction

Each claim is analyzed separately through the Domain Representation process. To analyze the claims, a context extraction algorithm can be used. To handle the different vocabularies used by different information sources, a comparison based on context is used in addition to simple string matching. For each document the context is extracted by the Patent Knowledge Extraction and then compared with the ontology concept by the Patent Domain Representation.

We define a context descriptor c_i from domain DOM as an index term used to identify a record of information [7], which in our case is a patent. It can consist of a word, phrase, or alphanumerical term. A weight $w_i \in R$ identifies the importance of descriptor c_i in relation to the patent. For example, we can have a descriptor c_1 = Length and $w_1 = 2$. A descriptor set $\{<c_i,w_i>\}$ is defined by a set of pairs, descriptors and weights. Each descriptor can define a different point of view of the concept. The descriptor set eventually defines all the different perspectives and their relevant weights, which identify the importance of each perspective.

By collecting all the different viewpoints delineated by the different descriptors, we obtain the context. A context $C = \{\{<c_{ij},w_{ij}>\}_i\}_j$ is a set of finite sets of descriptors, where i represents each context descriptor and j represents the index of each set. For example, a context C may be a set of words (hence DOM is a set of all possible character combinations) defining a patent and the weights can represent the relevance of a descriptor to the patent. In classic Information Retrieval, $<c_{ij},w_{ij}>$ may represent the fact that the word c_{ij} is repeated w_{ij} times in the patent.

The Patent Knowledge Extraction process uses the World Wide Web as a knowledge base to extract multiple contexts in multiple languages for the textual information. The algorithm input is defined as a set of textual propositions representing the claim information description. The result of the algorithm is a set of contexts - terms that are related to the propositions in multiple languages. The context recognition algorithm was adapted from [10] and consists of the following three steps:

1) Context retrieval: Submit each parsed claim to a Web-based search engine. The contexts are extracted and clustered from the results.

2) Context ranking: Rank the results according to the number of references to the keyword, the number of Web sites that refer to the keyword, and the ranking of the Web sites.

3) Context selection: Assemble the set of contexts for the textual proposition, defined as the outer context.

The algorithm then calculates the sum of the number of Web pages that identify the same descriptor and the sum of number of references to the descriptor in the patent. A high ranking in only one of the weights does not necessarily indicate the importance of the context descriptor. For example, high ranking in only Web references may mean that the descriptor is important since the descriptor widely appears on the Web, but it might not be relevant to the topic of the patent.

The external weight of each context is determined according to the number of retrieved Web references related to the concept and the number of references to the concepts in the patents. In addition, the Term Frequency/Inverse Document Frequency (TF/IDF) method analyzes the patent from an internal point of view, i.e., what concept in the text best describes the patent. The patent knowledge extraction is described in more detail in [9].

The experiments included a set of 81 patents randomly selected from the United States Patent and Trademark Office. Each patent included a vector with 43 top ranking context extracted values. The tool used to create SOMs was eSom2. The tool suggested six different clusters, Figure 2. Each cluster is represented by a different color.

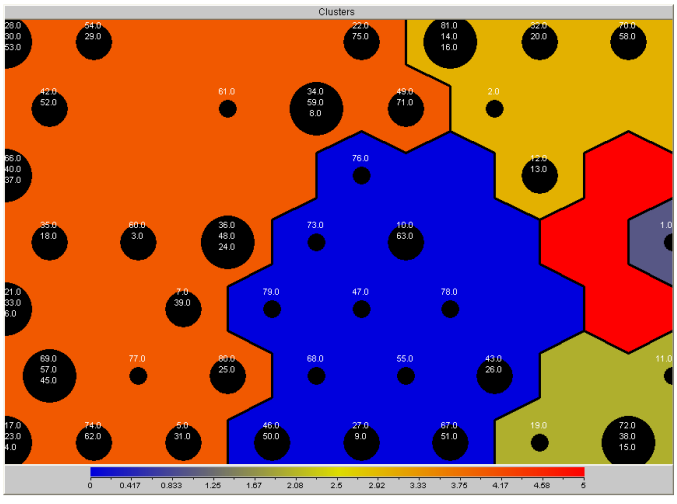

Figure 2. The dataset of 81 patents formed six context clusters.

One example of context classification is displayed in Figure 3, which presents the self-organizing map feature plane *Manufacturing*. We can see that *Manufacturing* is relevant to only one patent and slightly relevant to 20% of the patents according to context matching index.

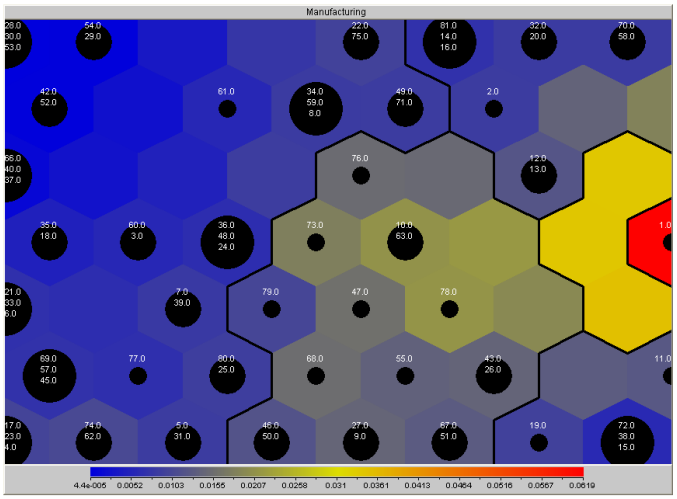

Figure 3. Manufacturing feature plane of the SOM

V. Patent Knowledge Context Simulation

To visualize the values of a single variable of the SOM, it is possible to change the coloring of the map. The user can, for example, select a color map where red represents big values of a variable and blue represents low values. These colors further help the user to understand what the different positions on the map represent on a single variable level. The user can also keep track of the change in position on the map. Each change is marked to represent the previous change in position. The user can select how many changes are tracked. In the SIMU_SOM demonstrator it is also possible to select with the mouse a desired position on the map and see the context match values that constitute that selected position. The context match values for the selected position on the SOM can be seen on the sliders when selecting a node on the map with the mouse.

By knowledge context simulation we refer to a "what-if" type of analysis of alternative knowledge contexts with the SIMU_SOM demonstrator. The basic idea is to show the approximate effect of optional knowledge contexts. This kind of simulation provides an idea of where incremental changes in the new knowledge content would place the new knowledge entity. In simulation, the individual moves the sliders (context elements) she is willing to change or develop. The amount of movement in sliders resembles the amount of change in knowledge; a slight increase signifies a slight change and a large change signifies a major change required in the new knowledge entity. In Figure 1, this can be interpreted as follows: the starting point of an arrow is the current knowledge context position of a new patent application and the end of each arrow is the simulated alternative knowledge context position of a new patent application This simulation gives an idea of how much change is required in the new knowledge entity to achieve a desired knowledge context cluster of the future. The amount of desired change in context that the simulation indicates refers to the effort required in the real world. In practice, this means changing technology plans, investment plans, design plans, etc. The course of the simulation helps the individual to recognize and understand the meaning of context elements and especially to recognize such elements that are the most meaningful in the path towards desired knowledge

context in the future. Context simulation can be an eye-opener for the participants. The aim is to involve participants in the course of taking action that will lead to the personal experience required to internalize something new. This is more than just providing optional vision for the future. Participatory methods usually result in good commitment and motivation.

Knowledge context simulation may save the enterprise time, money, and resources as the impact of patent plans can be roughly estimated on the computer screen first, instead of experimenting in the real world directly. Knowledge context simulation can take the quality and the meaning of planning to a new, targeted level by showing what kinds of development paths could be taken. Do the current patent content and plans meet the vision of the company? Context simulation can help in selecting those actions to which the organization and individuals can commit.

VI. Discussion and Future Work

The patent service model described in the paper allows a self-organizing map to be created on the boundaries of existing knowledge. The model shows promise in extending the field of patent service. This paper describes a work-in-process, and we are currently working on validating and expanding the data set of the proposed joined approach.

We are in the process of verifying the results predicted by SOM by tagging the patents according to their year and evaluating which contexts have become realities for those patents that are a few years old. We may be able to say something about the predicting power of context SOMs and about the "lead time" from patent context to reality.

In this work, SOM produces results based on existing patents. Therefore a question arises on how we can forecast something that lies ahead in the future, based on the existing dataset embedded in the SOM. When we use context time-series in SOM we may see tendencies that show us a way that leads to "out-of-scope" knowledge or to knowledge that is beyond existing knowledge. If we can get some additional hint or clue on what will be important in the future, we individuals and our organizations can be proactive in many ways. We believe that the usefulness of the proposed patent context simulation is in the increased change to see what will be important in the future.

We are suggesting that knowledge context simulation provides increased understanding of the patent contents and patenting plans and a way to improve the existing patent applications and the overall patenting plan.

Future work involves evaluating the patent search model against patents over a timeline to evaluate change in knowledge. We will finish the on-going verification, as discussed above, and explore in what other ways context simulation can be implemented on existing knowledge bases. Another direction is to extend the model to multiple languages. In addition, the demonstrator described in this paper needs further development to suit well to patent context knowledge simulation.

VII. Acknowledgment

This research was partially supported by the Korean Government IT R&D Program of MKE/KEIT (10035166, Development of Intelligent Tutoring System for Nursing Creative HR).

References

1. Back B., Sere K., Vanharanta H., Managing complexity in large data bases using self-organizing maps. Accounting Management and Information Technologies, 8, 4, pp. 191–210 (1998)
2. Banks, J, J.S. Carson II, L.B. Nelson, and D.M. Nicol. 2005. Discrete-event system simulation. International edition. Prentice Hall International Series in Industrial and Systems Engineering. Upper Saddle River, NJ 07458: Pearson Prentice Hall
3. Eklund T., Back B., Vanharanta H., Visa A., Using the self-organizing map as a visualization tool in financial benchmarking. Information Visualization, 2, 3, pp. 171–181 (2003)
4. Kantola, J.,Piirto, A., Toivonen, J. and Vanharanta H., Simulation with Occupational Work Role Competences, Proceeding of Conference on Grand Challenges in Modeling and Simulation (GCMS'09), Istanbul, Turkey, July 13-16 (2009)
5. Kaski S., Kohonen T., Exploratory data analysis by the self-organizing map: structures of welfare and poverty in the world. [in:] Apostolos PN, Refenes YA-M, Moody J., Weigend A. (eds.) Neural Networks in Financial Engineering. Singapore, World Scientific, 498-507 (1996)
6. Kohonen T., Self-Organizing Maps. Springer-Verlag, Leipzig, Germany (2001)
7. Mooers C., Encyclopedia of Library and Information Science. Marcel Dekker, vol. 7, ch. Descriptors, pp. 31-45 (1972)
8. Pegden, C.D., R.E. Shannon, and R.P. Sadowski. 1995. Introduction to Simulation Using SIMAN. 2nd ed. New York: McGraw-Hill.
9. Segev, A. and Kantola, J., Knowledge Discovery System for Patents, 7th International Conference on Knowledge Management (ICKM2010), Pittsburgh, Pennsylvania USA, 22-23 October 2010
10. Siegel, M., Madnick, S.E.: A metadata approach to resolving semantic conflicts. In: Proceedings of the 17th International Conference on Very Large Data Bases, pp. 133–145 (1991)
11. Ultsch A., Self organized feature planes for monitoring and knowledge acquisition of a chemical process. The International Conference on Artificial Neural Networks. Springer-Verlag, London, pp. 864–867 (1993)