Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 158 (2019) 443–449

Procedia Computer Science

www.elsevier.com/locate/procedia

3rd World Conference on Technology, Innovation and Entrepreneurship (WOCTINE)

Forecasting of innovation in the light of semantic networks

Cigdem Baskici [a]*, Suat Atan[b], Yavuz Ercil[c]

[a] *Assistant Professor, Başkent University, Ankara, 06790, Turkey*
[b] *PhD, Agriculture and Rural Development Support Institution, Ankara, 06790, Turkey*
[c] *Professor, Başkent University, Ankara, 06790, Turkey*

Abstract

The subject of this study is a strategic and competitive global innovation management. In this frame, a study was carried to reveal the dynamics of innovations from idea to product in three steps. In the first step, semantic network analysis was used in 400 papers with the highest citations published in 20 journals with the highest h5-index in the health field from 2013 to 2015. Web scraping and text mining tools based on R language were used to create semantic network structures. Semantic network analysis revealed that the focus of the articles was on cancer. The analysis carried out using the Sankey diagram revealed that scientists who work in cancer are most frequently involved in the lung, however, the scientists who related with the lung are not focused on treatment, and heart. The experts' comments are due to challenges in the treatment of the lung cancer scientist may be focused on areas like diagnosis and phases of cancer. In the second step, 260,000 rows of Clinical Trials cases were analyzed. In the third step, an analysis was made on a total of 1000 patent documents selected from the lens.org site, which contains the databases in global patent offices. Methods used in the analysis of the articles in the first step were carried out in these documents. According to this, diagnosis, treatment and therapy words the most common pass with lung cancer through the documents. In addition, the most cited authors in the field of lung cancer were searched in patent documents. These authors are considered as academic backgrounds that feed new technologies. According to this, it has been found that there are no major matches among the most cited authors of 400 papers and most cited authors in the patent documents in the field of lung cancer. It has been found that patents feed on more specific journals rather than major medical journals. In addition, it was also found that the most cited authors of 400 papers were different from the patent owners.

Peer-review under responsibility of the scientific committee of the 3rd World Conference on Technology, Innovation and Entrepreneurship

Keywords: strategic technology forecasting; socio-technological systems; text mining; artificial intelligence

* Corresponding author. Tel.: +90312 246 66 66; fax: +90312 246 66 66.
E-mail address: cbaskici@baskent.edu.tr

1877-0509
Peer-review under responsibility of the scientific committee of the 3rd World Conference on Technology, Innovation and Entrepreneurship
10.1016/j.procs.2019.09.074

1. Technology forecasting models

The aim of this study is to propose a model for strategic and competitive global innovation management. The developed model has been tested in the health sector. In this context, the study aimed to answer the question if innovation could be forecasted by a nonlinear, dynamic socio-technological model based on semantic networks.

Nomenclature

TF	Technology Forecasting
LDA	Latent Dirichlet Allocation
TM	Text Mining
BGR	Bigram

In literature, most of the innovation forecasting models focus on inclination, substitution, and transfer slopes in the innovation process. However, the limited empirical results of these models have been discussed in the literature [1]. Watts and Porter [1] propose the technology foresight model on three basic dimensions to fill this gap;

- Technology Life Cycle Status,
- Innovation Context Receptivity,
- Product Value Chain and Market Prospects,

As Watts and Porter [1] proposed, indicators within the technological life cycle status are defined as the number of basic research, number of applied research, number of patents, number of news and the number of saturation in the social impact and the number of companies established. Innovation context receptivity indicators are identity technologies, technology decomposition status, corporate players, patent concentration profiles as specified in articles related to the target technology, government support, and special support, restrictions, regulations, functional equivalence definition, determining the interests of competitors, are indications such as the tabulation of problems. At last product value chain and market prospects indicators are self-profile through component technologies, scope and definition of trained personnel resources, range of possible applications recorded, sectoral activity concentration and location of geographical distribution.

The variables included in the technology life cycle status, which is the first dimension mentioned here, have been the subject of myriad studies, in particular, after Fenn began to be discussed his most effective hype cycle model [2]. The impact of the technology cycle model on strategy innovation management literature has continued nowadays. Today much research [3,4] claims that Fenn's Hype circle method is valid not only in the IT sector but also in all other traditional sectors.

In this study, the technology cycles are questioned in detail. The first query is the use of indicators in the technology cycle. In many studies, the indicators are calculated on a linear formula. However, Acemoglu and et al.'s [5] remarkable study suggests that there are relations between patents in a field of technology (e.g. pharmaceuticals) and another technology (e.g. chemistry). These relations are seen in correlations according to patent classes [6]. These relationships can help patents in technology to be the predictor of patents in other technology. For example, the fact that patents in the field of chemistry nourish the field of pharmacy is also seen through citation networks. In other words, there are non-cyclic effects of indicators on each other [5].

In another query, as shown in the technology forecasting model [7], which is defined by a series of indicators, it is likewise advised to use network structures over quantitative indicators, and the number of citations is heavily used in these network analyses. In this case, it is not possible to follow the premise of the cited study [8]. For this reason, a database has been used predominantly in the studies on technology forecasting. Interactions within this database have been studied in most of the literature. In general, this database is containing patent data [9]. However, ideas that create innovation are published as articles and books before patents. In addition, those who create citations are not scientific publications, but their authors, and usually an article has more than one author. In this case, it is difficult to follow the relations between the authors.

Another query is the spread of innovation. In the literature, the spread of innovation has been tried to be followed through a linear trace from the beginning to the end of the technology cycle. This development is also influenced by a user-driven attraction. Expectations and readiness of the users are also an important factor in the emergence of innovation which neglected in most of the literature. The technology prediction models that draw attention to this subject [10,11] argue that the unit of analysis should be socio-technical systems rather than sectoral innovative systems. In this way, competence will be gained in understanding the institutions and actors that create innovation.

In the light of all these inquiries, it is necessary to create a non-linear, dynamic and integrated model (including actors, institutions and rules) where qualitative and quantitative indicators can be included in the analysis. In this study, the research question was tried to be answered with the creation and testing of such a model. In the following sections, the construction of the model and the results obtained were shared.

2. Research

This study based on sophisticated web scraping and text mining tools. In the study, innovation has been described as a function of past literature and today's patents [5] (Formula-1). Past literature has been defined by data mining. At this stage, non-structural data were first structured.

$$\Delta(t)=f(N(t), R(t)) \quad (1)$$

The research is designed as 3 basic steps (Fig.1). The first step is to define the ideas of innovation which means identification of ideas that leads to innovations. In order to achieve this, 400 papers published in 20 journals in the h-5 index of the health sector are examined. The basic concepts mentioned in the papers are extracted. This is done by specially coded software. The relationships between the basic concepts obtained by coding are defined as bigrams. These concepts, which are related to these basic concepts, are defined as 2nd step concepts. The 3rd step concepts related to these 2nd step concepts were also defined as bigrams. The bigrams are word pairs like “treating cancer” or “cancer treating”. Bigrams use to help to explore the concepts in their contexts in text mining. Bigrams offer a database suitable for network topology. With the help of the network, the concepts that are in the focus can be examined as well as the relations between the other concepts.

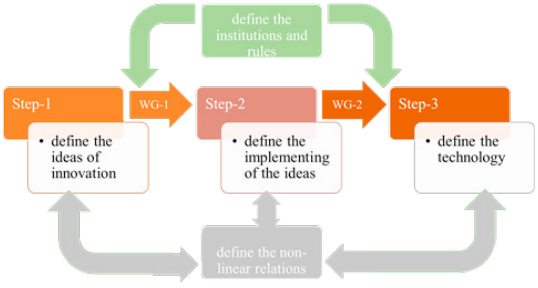


Fig.1. Design of the research

Thus, the concepts in the most cited articles were extracted by considering their relations [12]. The authors of these articles were taken out one by one and the source of each idea was defined. After this stage, a Sankey (Alluvial) diagram was constructed and focus groups of the related studies were revealed.

The second step of the research is to define the implementing ideas. In the second step, the trace of the concepts in the first step was followed in the practical area. Thus, the process from idea to the practical area has been followed. For this purpose, using the database of clinical trials, the ideas in the articles were tried to be found in this database. 260,000 rows of Clinical Trials cases were analyzed. It was possible to follow the interactions between ideas and practical concepts by using dynamic network analysis method in time.

In the third step, the transition from practical to technology has been tried to be followed. At this stage, 1000 data

were examined in patent databases. For this purpose, the researchers who develop the most patents related to the concepts have been followed. The authors of the most cited patents in the field of cancer (who feed the patents) are identified. These authors were cited within the patents as an academic background feeding new technologies. The LDA algorithm was used at this stage. LDA algorithm is an unsupervised machine learning algorithm which does not require external intervention according to the clusters formed by the concepts in text mining [13].

Among the stages of the study, expert meetings were held and the findings obtained at each stage were evaluated with expert intuition. Two workshops were organized for this purpose. 8 scientists from the field participated in the workshops. Thus, the findings obtained by machine learning were evaluated with expert intuition and experience. Through these evaluations, the causality of relations has been revealed.

3. Findings

During the definition of the idea, 400 articles of 20 journals in the h5-index at the health sector were studied. In this first step, the most popular scientific subject was found as cancer (Fig.2).

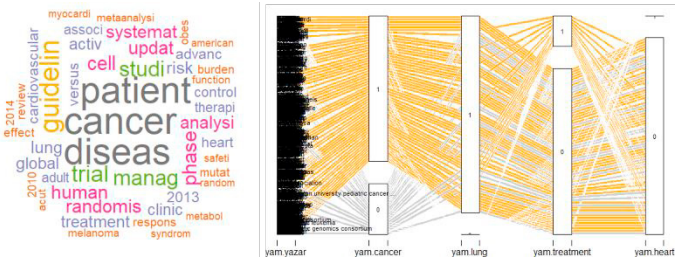


Fig.2. Basic concepts and fields of study in the h-5 index

The concept of cancer has been observed in a Sankey diagram. The analysis carried out using the Sankey diagram revealed that scientists who work in cancer are most frequently involved in the lung, however, the scientists who related with the lung are not focused on treatment, and heart. The experts' comments are due to challenges in the treatment of the lung cancer scientist may be focused on areas like diagnosis and phases of cancer. Thus, the study discipline areas associated with the concept of cancer could be defined.

The 2nd step concepts related to cancer were found as lung, chest, non-small cell, phase, and advance (Fig.3). These concepts are important in terms of expressing the areas in which cancer studies are focused on.

```
findAssocs(tdm_1, terms = "cancer", corlimit = 0.3)

$cancer
      lung    breast nonsmallcel     phase    advanc
      0.41      0.35      0.34      0.33      0.33
```

Fig.3. 2nd step concepts

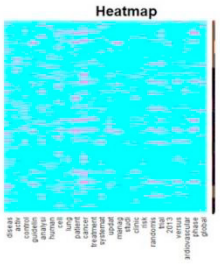
The concepts of non-small cell, fusion, egfr, afatinib, transfusion, alk, rejection, ros1, erlotinib, squamous cell, alkaloid, crizotinib, cisplatin, and analysis have been extracted as the related concepts of lung.

```
findAssocs(tdm_1, terms = "lung", corlimit = 0.3)

$lung
 nonsmallcel       cancer       fusion         egfr    afatinib transfusionrel          alk
        0.58         0.41         0.38         0.38         0.38         0.31         0.31
         ret         ros1    erlotinib   squamouscel     alkposit    crizotinib      luxlung
        0.31         0.31         0.31         0.31         0.31         0.31         0.31
   cisplatin       analys
        0.31         0.31
```

Fig.4. 3rd step concepts

The scientists who dealt with these concepts in their works were also extracted. Thus, the subject - scientist match could be made (Fig.5). The heatmap matrix obtained based on this extraction has formed a network. Analyzing this network makes it possible to reach the degrees of centrality and network roles of each author. Thus, according to the subjects, the roles of scientists in these concepts can be defined.



	<fctr>	diseas <dbl>	activ <dbl>	control <dbl>
1	i benilova	1	0	0
2	rj zatorre	0	0	0
3	t klengel	0	0	0
4	al giraud	0	0	0
5	bg dias	0	0	0
6	ae renton	0	0	0
7	g buzsáki	0	0	0
8	o butovsky	0	0	0
9	l madisen	0	1	0
10	ve miron	0	0	0

Fig.5. Scientist-subject relations

At the workshop after the first step, experts with at least 10 years of international academic experience in the field of cancer diagnosis and treatment were invited. The experts were asked about the discrepancies in the results and the causal relationships one by one. In this workshop, the causal structures of the relations between the concepts are explained. With the results of the workshop, the second step of the research has been redesigned and the efforts have been defined in the implementation phase.

In the 2nd step of the study, 260,000 studies in ClinicalTrials.gov were examined by TM and explored what kind of treatments were performed and what cases occurred in the frame of the 1st step. The maps obtained by examining the cases here were asked to the experts who had at least 15 years of international clinical experience in the field of expert cancer diagnosis and treatment. The experts who participated in the second workshop were able to establish meaningful relationships between the results obtained in the first step of the study and clinical cases. It has been found that the clinical application of each concept leads to consistent causal relationships as in the first step. In this workshop, the effect of clinical studies on academic studies has been demonstrated. Thus, nonlinear relationships between the two networks were revealed. It was also possible to reveal the relationships between the scientists identified in the 2nd step of the research and the scientists in the 1st step.

In step 3 of the research, it has been tried to understand how the findings of the previous 2 steps have evolved into patents. At this stage, patent documents were examined and an analysis was made on their owners. In the first two steps, firstly, patent studies on lung cancer were examined. According to the findings obtained here, it was observed that the patents were primarily oriented towards the method and the issues of diagnosis were given more intensity than the treatment subjects. According to the findings obtained from patent studies, developing technologies are mostly related to pre-cancer prediction technologies and post-cancer survival period. The determination of variables will be the areas of technologies that will be introduced in the first phase. According to the bigram networks (Fig.6) obtained in this step of the research, the most strong and weak ties have been revealed, including the central concept of cancer. Strong ties reveal relationships that should be expected to yield efficient results, while weak ties reveal creative relations [14]. In this case, strong ties are established between the concepts related to cancer and its diagnosis, and the concepts related to lung and its treatment, while the weak ties have been formed between cancer and the concepts such as gene and microRNA.

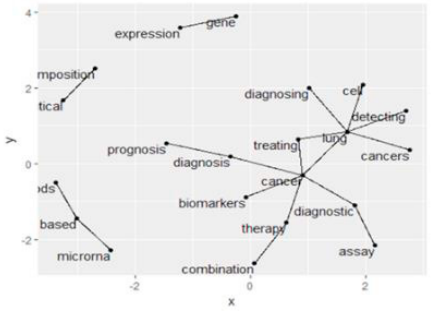


Fig.6. Bigram networks derived from patent data

At this stage of the research, an analysis was made on the relationships between the concepts in the patents and clinical concepts. The LDA algorithm was used as the analysis tool to find out conceptual patterns in this stage. In this phase, the existence of concepts diagnosis, treatment, and biomarker in category 3 showed that these concepts were a special group within the patents (Fig.7).

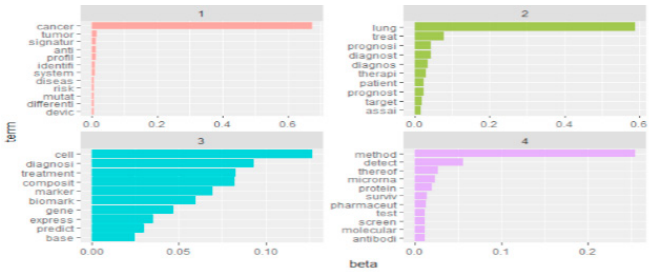


Fig-7: Findings obtained by LDA algorithm

Methods used in the analysis of the articles in the first step were carried out in these documents. According to this, diagnosis, treatment and therapy words the most common pass with lung cancer through the documents. In addition, the most cited authors in the field of lung cancer were searched in patent documents. These authors are considered as academic backgrounds that feed new technologies. According to this, it has been found that there are no major matches among the most cited authors of 400 papers and most cited authors in the patent documents in the field of lung cancer. It has been found that patents feed on more specific journals. In addition, it was also found that the most cited authors of 400 papers were typically different from the patent owners.

4. Discussion

As a result, a model is proposed for strategic and competitive global innovation management. This model is based on TM, machine learning algorithms and analytics on unstuctured data and metadata. The distinguishing feature of this model is that it is a technology cycle model that uses non-linear relationships, qualitative and quantitative indicators. It has been defined that the technologies both focus on diagnosing the disease in pre-cancer phase and survival period in post-cancer phase are being developed at the end of the technological life cycle in the field of health.

This study presents two fundamental results. The first result is theoretically developing the nonlinear dynamic structure of the technology cycle. Thus, it was possible to create nonlinear dynamic models based on semantic networks. The second result is the analysis of the weak and strong ties between the actors, institutions/rules of the

process from idea to product by practically making dynamic inquiries within the framework of the model's prediction. Thus, it was possible to uncover efficient and creative relationships in the innovation structure of the sector.

The analysis of the difference between the predictions of this model and other models in the literature may be inspiring for further studies. Thus, efficient scientific discussions can be realized between designs based on different insights and design based on the concept of dynamic nonlinear semantic network model.

Acknowledgements

This work was supported by the Başkent University, Turkey and STM ThinkTech Future Technologies Institute, Turkey funds.

References

[1] Watts, Robert. J., and Alan. L. Porter. (1997) "Innovation forecasting." *Technological Forecasting and Social Change*, **56** (**1**): 25-47.

[2] Fenn, Jackie. (1995) "When to leap on the Hype Cycle", *Advanced Technologies & Applications*, Gartner.

[3] Lee, Changyong, Ohjin Kwon, Myeongjung Kim, and Daeil Kwon. (2018) "Early identification of emerging technologies: A machine learning approach using multiple patent indicators." *Technological Forecasting and Social Change*, **127**: 291-303.

[4] Jun, Seung-Pyo. (2012) "A comparative study of hype cycles among actors within the socio-technical system: With a focus on the case study of hybrid cars." *Technological Forecasting and Social Change*, **79** (**8**): 1413-1430.

[5] Acemoglu, Daron, Ufuk Akcigit, and William R. Kerr. (2016). "Innovation network," *Proceedings of the National Academy of Sciences*, **113** (**41**): 11483-11488.

[6] Järvenpää, Heini M., Saku J. Mäkinen, and Marko Seppänen. (2011) "Patent and publishing activity sequence over a technology's life cycle." *Technological Forecasting and Social Change,* **78** (**2**): 283-293.

[7] Robinson, D. K., Lu Huang, Ying Guo, and Alan L. Porter. (2013) "Forecasting Innovation Pathways (FIP) for new and emerging science and technologies". *Technological Forecasting and Social Change*, **80 (2)**: 267-285.

[8] Gutiérrez, R., A. Nafidi, and R. Gutiérrez Sánchez. (2005) "Forecasting total natural-gas consumption in Spain by using the stochastic Gompertz innovation diffusion model." *Applied Energy*, **80 (2)**: 115-124.

[9] Meade, N., and Towhidul Islam. (2006) "Modelling and forecasting the diffusion of innovation–A 25-year review". *International Journal of Forecasting*, **22** (**3**): 519-545.

[10] Mann, Darrell L. (2003). "Better technology forecasting using systematic innovation methods." *Technological Forecasting and Social Change*, **70** (**8**): 779-795.

[11] Kamakura, Wagner. A., and Sıva K. Balasubramanian. (1987) "Long-term forecasting with innovation diffusion models: The impact of replacement purchases." *Journal of Forecasting*, **6** (**1**): 1-19.

[12] Bolasco, Sergio, Alessio Canzonetti, Federico M. Capo, Francesca della Ratta-Rinaldi, and Bhupesh K. Singh. (2005) "Understanding text mining: A pragmatic approach", in Sirmakessis S. (eds) *Knowledge Mining. Studies in Fuzziness and Soft Computing*, Springer, Berlin, Heidelberg.

[13] Ni, Xiaochuan, Jian-Tao Sun, Jian Hu, and Zheng Chen. (2011) "Cross lingual text classification by mining multilingual topics from wikipedia." *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 375-384). ACM.

[14] Granovetter, Mark S. (1973) "The strength of weak ties." *American Journal of Sociology*, **78** (**6**): 1360-1380.