1st Workshop on Patent Text Mining and Semantic Technologies (**PatentSemTech 2019**)

Hidir Aras[1]

FIZ Karlsruhe – Leibniz Institute for Information Infrastructure

Linda Andersson, Florina Piroi[2]

TU Wien, Institute for Information Systems Engineering

Abstract

This volume presents the proceedings of the 1st Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech 2019) co-located with the SEMANTiCS 2019 conference, held in Karlsruhe, Germany. It is a first in series of workshops that aims to establish a long-term collaboration and a two-way communication channel between the IP industry and academia from relevant fields to foster the usage of semantic technologies for answering research questions related to patent text mining and patent analytics as well as adopt them in working applications.

Keywords: Patent Text Mining, Semantic Technologies, Semantics 2019, Intellectual Property, Machine Learning

1. Introduction

The PatentSemTech 2019 workshop[3] is a first in series of workshops that aims to establish a long-term collaboration and a two-way communication channel between the IP industry and academia from relevant fields such as natural-language processing (NLP), text and data mining (TDM) and semantic technologies (ST) in order to explore and transfer new knowledge, methods and technologies for the benefit of industrial applications as well as support research in applied sciences for the IP and neighbouring domains.

We invited scientific contributions as well as proof of concepts that show relevant use cases for patent text mining and analytics. Moreover, we invited researchers to investigate and promote new means for bootstrapping training data generation, e.g. for labelling domain-specific data sets from the Intellectual Property (IP) domain. The articles included in this volume went through a peer-review process where each submission was reviewed by at least three reviewers out of a mixed programme committee of academic researchers and experts from the IP domain. Seven papers passed the review process – 3 long, 2 short and 2 demo papers. Three submissions that passed the reviewing process were proposed for publication to the World Patent Information[4] (WPI) virtual special issue on ”Patent Text Mining and Semantic Technologies”. For these submissions we only included their (extended) abstracts in these proceedings.

In its first year, the workshop was organized as a one day event. We have invited as a keynote speaker Mr. A. Trippe, managing director of Patinformatics LLC., a long year, internationally recognised IP expert, and an adjunct Professor of IP Management and Markets at Illinois Institute of Technology where he teaches courses on patent analysis and landscapes for strategic decision making. His keynote addressed the importance of patent analytics tools based on semantics and machine learning techniques for the strategic decisions that businesses need to take with respect to their long term

[1] https://www.fiz-karlsruhe.de

[2] https://www.ifs.tuwien.ac.at

[3] http://www.ifs.tuwien.ac.at/patentsemtech/

[4] https://www.journals.elsevier.com/world-patent-information

*Published in "Proceedings of The 1st **Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech 2019)**", Eds. L. Andersson, H. Aras, F. Piroi, A. Hanbury, PatentSemTech 2019, 12 September 2019, Karlsruhe, Germany. https://doi.org/10.34726/pst2019.1 | © Authors 2019. CC BY 4.0 License.*

R&D and economic plans.

2. Keynote: Improving Patent Analytics Using Semantic Technologies

The use of patent analytics has increased exponentially over the past ten years. So much so that even the worlds patent offices have devoted resources and staff to create departments responsible for developing insights into technology areas of importance to that country or region using output generated applying patent analytics. At the same time new tools, methods and systems have begun to emerge that seek to make the analysis of patent data easier to accomplish. Included in these new developments are a significant number of approaches that apply machine learning and make use of knowledge modeling and semantic analysis in order to deal with existing challenges for text and data analytics. As these changes continue to occur, it would be useful to review a list of the tasks associated with patent analytics and think about the types of tools, systems, or methods that a patent analyst would like to have at their disposal. Starting with a general overview of patent analytics, and with a focus on patent landscape reports, case studies and perspectives will be provided on why this work is so highly valued. The presentation concluded with a prioritized list of suggestions for how patent analytics and patent landscape creation could be aided by the further development and implementation of semantic technologies.

3. Main Topics and Objectives

In the started workshop series we aim to set the basis for researchers and the IP industry to explore next-generation text and data mining methods and semantic technologies for the enrichment and large-scale analysis of huge amounts (Big Data) of scientific-technical information in general and patent data in particular.

We want to motivate and enable scientists from academia to make use of and exploit the richness of the scientific-technical information that is amassed nowhere else but in the patent data, by, for example, interlinking it to other knowledge sources from domain-specific knowledge graphs (bio-pharma, chemistry or engineering, etc.) or the linked open data cloud.

Starting with publicly available datasets for patent mining and patent retrieval tasks such as classification, passage retrieval, etc. we want to set the focus on developing enhanced methods for analysing patent texts by applying machine learning and making use of implicit and explicit semantic information.

Hence, the workshop series aims to motivate research and development in related areas in order to

- *explore* IP applications with underlying advanced NLP, TDM and artificial intelligence methods, e.g. applying Deep Learning (DL) for generating patent embeddings, etc.
- *apply* enhanced machine or deep learning technologies for the semantic enrichment and analysis of big data of patent texts, e.g. to contribute to use cases such as technology analysis, trend analysis, semantic patent landscaping, competitor analysis, etc.
- *show* proof of concepts for patent and technology analysis use cases such as patent landscaping, portfolio analysis, white and hotspot analysis, technology trends analysis, etc.
- *evaluate* new visual user interface concepts for exploring and analysing large datasets of scientific texts

There have been several text mining initiatives in terms of establishing tools and benchmark collections for widely used data such as news corpora, medical data, etc. However, a set of benchmark collections covering the diversity of the information needs of the IP industry, as for example detailed in [30, 4], is still missing. A long term goal of this series of workshops is therefore to encourage future research collaboration with focus on IP related data – patent documents, non-patent literature, court litigation cases – and combine it with more traditional patent analytic resources, like meta-data, to be used for the above described tasks and use cases.

4. State of the Art and the Impact of Training and Test Data

Patent text mining [48, 24, 22, 36] includes research on the handling and the integration of multiple and diverse information sources, since information related to IP for science and technology are siloed into various repositories consisting of laws, regulations, patents, court litigation, scientific publications etc.

4.1. Patent Retrieval

As a research field, Patent Retrieval belongs to domain-specific information retrieval, hence, represents a sub discipline of information retrieval (IR). The research focus of patent retrieval is to develop techniques and methods that effectively and efficiently retrieve relevant patent documents or paragraphs in response to an information need [36, 50, 51]. IR has received a significant amount of focus from researchers in different computer science disciplines since many decades. In comparison, patent retrieval is poorly treated by the academic scientific communities, with periodic surges of such activities whenever patent data became available to researchers. It is only during the last 20 years that the challenges in patent retrieval have been a target for the research community [30]. On reason for this is that, compared to other types of text, the patent genre presents unique features such as lengthy documents (multi-page), multi-modal documents (e.g. image, text, algorithm and programming codes), multi-language, semi-structured, meta-data rich, stretching over a variety of technologies (heterogeneous). Answers to information needs in IP also vary from a complete multi-page application to a one-page inventor disclosure to just a few keywords [22].

In a scientific context patent retrieval was first introduced in the NIIs NTCIR-1 campaigns (2002 to 2007) [15], followed by several initiatives that included patent retrieval as a research topic, e.g. Dutch Belgian Information Retrieval workshop [41], ASPIRE [17], Patent Information Retrieval (PaIR), TREC-Chem (TExt REtrieval Conference Chemical track) [28], and the Information Retrieval Facility Symposium and Conference (2008-2014) [42, 29]. The largest academic research impact, in Europe, has been made by the CLEF-IP track, which was part of the Cross-Language Evaluation Forum (CLEF). The CLEF-IP track was organized from 2009 to 2013 and included a variety of tasks ranging from image classification, prior art search, and patent text classification.

4.2. Passage Retrieval

In 2012, CLEF-IP introduced the Patent Passage Retrieval task [33]. Given a patent application and selected claims in the document, the aim was to retrieve relevant documents *and* also extract those paragraphs (passages) from them that are found most relevant. Since CLEF-IP used mainly data released by the European Patent Office (EPO), the relevance assessments were semi-automatically extracted from EPO search reports.

The passage retrieval task is very close in spirit with the work of patent examiners done during an invalidity or validity search: examiners need to identify both the prior art documents, as well as each specific paragraph within these documents considered to be Prior Art for specific claims in the patent application [18, 36]. Patent Passage Retrieval could be seen as a cross-over between adhoc document retrieval tasks and question answering (QA) tasks. Concretely, in order to achieve good performance it is required that query formulations include automatic technical term extraction, followed by an advanced ad-hoc IR approach. Furthermore, in order to narrow in on each relevant passage information extraction (IE) approaches needs to be considered as well.

4.3. Enhanced Semantic Analysis and Patent Mining

Segmenting the full text of patent documents (e.g. patent descriptions [39] text, claims [16]) is regarded an important step for the semantic structure analysis of the patent texts. New approaches based on machine learning are increasingly used for a variety of tasks related to patent text mining and large-scale patent analytics [38]. Important examples are trends analysis [47], technology forecasting [43], various clustering algorithms like reinforcement learning [10], support vector clustering [49], and matrix factorization [13].

In the last few years researchers started to apply Deep Learning methods to patent text mining tasks such as keyword extraction [21], synonym extraction [25] or patent classification [12]. Tasks such as calculating patent similarity [37], patent segmentation [11], and patent landscaping [3] can be considered as important sub-tasks to be considered. In addition, various types of embedding [32] such as graph embedding [46], word embedding have been applied to evaluating patent similarity [9] or text classification tasks [44].

4.4. Benchmark Data and Patent Resources for Patent Mining Tasks

After the CLEF-IP and the TREC-Chem evaluation campaigns, and the test collections resulting out of them, further efforts to establish and update benchmark text collections have been made, the latest is the WPI collection [26].

In the World Patent Information (WPI) journal, own data collections are provided in order to support the objectives of the journal, to publish new research and insights covering a broad spectrum of intellectual property information retrieval and patent analytics related practices and methods. The WPI journal editors together with the team at IFI CLAIMS Patent Services have put together a patent research collection, publicly available and for free, to foster scientific good practice: comparability, reproducibility, transparency and repeatability of experiments and results.

The WPI collection is for this reason static. It will not be updated with new data. This decision was made in order to make sure that experimental results are traceable and due to improvements of the proposed mining/retrieval methods are due to algorithmic improvements and not due to changes in the dataset. The WPI collection complements existing test collections, which are vertical (one domain or one authority over many years). Compared to them, the WPI collection is horizontal: it includes all technical domains from the major patenting authorities over the relatively short time span of two years.

Other, industry driven initiatives to establish resources for patent text mining also aim to provide researchers with benchmark data for this area:

- In October 2017, Google launched several patent related data collections and services. Google provides the Google Patents Public Datasets[5] on BigQuery, with a collection of publicly accessible, connected database tables for empirical analysis of the international patent system. The Google Patent Datasets can provide a solution to developing and answering search oriented questions. For instance, it is possible to formulate questions such as "what percentage of the patents have more than one inventor?" or "what funding does the government provide to promote innovation in certain patent areas?"

- Linked Open EP data[6] uses Uniform Resource Identifiers (URIs) to identify patent applications, publications and other resources present in patent data. The URIs make it possible to link the data other datasets. The data set covers the most relevant, but not all available bibliographic data elements for patents and not all data elements from the CPC scheme. It also includes references to the full text publication in PDF, HTML and XML format, which are stored on the European Publication Server. Linked Open EP data creates a public web of interlinked patent data from EPO and other data publishers that can be queried, retrieved and viewed using standardized web technologies like HTTP, URI, RDF and SPARQL.

A common problem in machine learning and particularly in the patent domain is the creation of labelled data (i.e. training and testing data) for a variety of search and analysis tasks. As available labelled corpora are either too small or not accessible to the research community, we want to alleviate this situation with a three-year effort. That is, we plan to target the creation of more datasets open to research and addressing different patent text mining and patent text analysis applications with focus on Information Extraction (IE), classification, clustering, and to establish further datasets that require only semi-supervised training methods.

Currently, the publicly available patent mining datasets involve only a few different types of IR applications (classification, passage retrieval and prior art search[7]). In order to explore and support other patent text mining and text analysis applications that originate from the diversity of IP experts' information needs, we aim for the creation of more IE-oriented datasets. The IE-oriented datasets will be designed for domain-specific terminology extraction, for example extraction of particular token types like mathematical formula, chemical compounds, quantity entity, sequences programming codes.

These datasets will provide to the industry and the research community a variety of benchmark data, a kind of 'PatentPedia', which can support different types of question answering systems ranging from text-based to knowledge-based approaches. Furthermore, a variety of patent retrieval and analysis tasks such as technology analysis, trend analysis, semantic patent landscaping,

[5] https://cloud.google.com/blog/products/gcp/google-patents-public-datasets-connecting-public-paid-and-private-patent-data

[6] https://www.epo.org/searching-for-patents/data/linked-open-data.html

[7] For example see `http://ifs.tuwien.ac.at/patentsemtech/data-sources.html`

competitor analysis, etc. could be explored, developed and evaluated.

The effort to establish these datasets will be undertaken as a community effort with an annotation task. As organisers we intend to provide a small starting set, which is gradually improved and increased as the task is launched and running. A similar procedure has been explored in BioNLP and SemEval [19] successfully.

During the first workshop we used existing data, which, though small in size, allowed us to apply supervised and unsupervised methods to detect technical terms [14]. For the coming years we plan define tasks that build on previous ones, with the aim of creating specific patent text sets of data where completing specific tasks is needed in order to approach the next one.

5. Impact and Expectations

5.1. Target Audience

The workshop is customised for the IP industry experts as well as for academic researchers. To attract the IP industry and especially expert users, we have been in contact with members of the CEPIUG (Confederacy of European Patent Information User Groups) in order to offer Continuing Professional Development (CPD) points.

For securing visibility and increase in research submissions and participation from both industry and academia, the authors of a selected set of accepted papers are invited to submit extended versions of their research to the virtual special issue of WPI, "Text Mining and Semantic Technologies in the Intellectual Property Domain". This year the following papers have been promoted to the virtual special issue:

- *Deep Learning based Pipeline with Multichannel Inputs for Patent Classification*
- *Detecting Multi Word Terms in Patents the same way as Named Entities*
- *Semantic Views - Interactive Hierarchical Exploration for Patent Landscaping*

In the future we aim to span over different scientific disciplines such as Economic[45, 23] and Social Science [20], which also have a long tradition of working with patent analytics, in particular on citation networks. Furthermore, we would like to involve, the Triz community *Invention Innovation creativity and design*, which has run their own conference since 2009. We are working to engage national and international patent offices, such as the EPO, and companies active in the patent industry (e.g. Legit, Patsnap, Landscaping Valuenex, Google Patent, to name a few) and invite them to the workshop events.

Our key speaker in the first workshop in the series we invited a representative of the IP industry side. In our second, upcoming event, we plan to invite an academic key speaker, while in the third event we plan for a panel debate with participants from patent offices, industry and academics.

Mr. Anthony Trippe, the key speaker of the first event, is Managing Director of Patinformatics, LLC. Patinformatics is an advisory firm specializing in patent analytics and landscaping to support decision making for technology based businesses. In addition to operating Patinformatics, Mr. Trippe is also an Adjunct Professor of IP Management and Markets at Illinois Institute of Technology teaching a course on patent analysis, and landscapes for strategic decision making. He has written or contributed to IP related articles that have appeared in the Wall Street Journal, Forbes, The Washington Post, and more than a dozen additional sources. Besides that, Mr Trippe has worked for a variety of organizations, in a number of different capacities, including: P&G where he was responsible for evangelizing the use of patent analytics for business decision, Aurigin Systems where he travelled the world working with companies of all sizes that use patent analytics for competitive advantage.

5.2. How do we want to engage the patent expert to evaluate and assess our tools?

One of the key issues when developing domain-specific mining tools, especially tools requiring labelled data and expert assessment, is to engage a sufficient number of domain-experts to establish a sizable corpora or benchmark collections. Therefore, for the participating patent experts, upon request, we will issue a certificate statement which describes and documents the tasks they have been involved in, certificate signed by the head of the organisation committee. Within the certification body of the International Standard Board for Qualified Patent Information Professionals, the certified professional needs to engage in Continued Professional Development (CPD) on an annual basis. There are four types of group activities:

1. Presenting at a conference and co-author a paper on patent-related topics,
2. Participating in courses related to patent information or patentable subject matter,
3. Reading publication on patent information,
4. Peer reviewing manuscripts or search reports, or attending patent information vendor, webinar.

Within the PatentSemtTech workshop series there is ample opportunity to obtain credit for each of the group activities. Our workshop allows the IP professionals to participate as a reviewer, a presenter, or to learn about new emerging technology as well as design future use cases and contributed to establishing new benchmark collections within the field of patent text mining.

6. Organizing and Programme Committee

6.1. Organizing Committee

The organizing committee consists of persons with experience both in academic research and in close collaboration with experts in the IP domain. Two of the committee members have been key persons in organizing and running the CLEF-IP and TREC-Chem campaigns.

- Dr. Hidir Aras, FIZ Karlsruhe, Germany
- Linda Andersson, TU Wien & Artificial Researcher IT, Austria
- Dr. Lei Zhang, FIZ Karlsruhe, Germany
- Dr. Florina Piroi, TU Wien & Artificial Researcher IT, Austria
- Prof. Dr. Allan Hanbury, TU Wien, Austria
- Dr. Mihai Lupu, Data Science Studio, Research Studios Austria Forschungsgesellschaft, Austria

6.2. Programme Committee

We are grateful to the following people for providing high quality reviews and helping the workshop organizers with the submission selection process:

- Jian Wang, University of Leiden, Netherlands
- Simone Ponzetto, University of Mannheim, Germany
- Hans-Peter Zorn, inovex Gmbh, Karlsruhe, Germany
- Catherine Faron Zucker, University of Nice, France
- Ron Daniel, Elsevier Labs, USA
- Natasa Varytimou, Refinitiv, formerly Thomson Reuters, UK
- Paul Groth, University of Amsterdam, Netherlands
- Natterer Michael, Dennemeyer Octimine GmbH, Munich, Germany
- Pedro Szekeli, USC Viterbi School of Engineering, USA
- Kobkaew Opasjumruskit, German Aerospace Center, Jena, Germany
- Shariq Bashir, University of Islamabad, Pakistan
- Michail Salampasis, International Hellenic University, Greece
- Siegfried Handschuh, University of St. Gallen, Switzerland
- Agata Filipowska, Poznan University of Economics, Poland
- Rene Hackl-Sommer, FIZ Karlsruhe, Germany
- Richard Eckart de Castilho, TU Darmstadt, Germany
- Joni Sayeler, Uppdragshuset Sverige AB, Sweden
- Mustafa Sofean, FIZ Karlsruhe, Germany
- Christoph Hewel, Patent Attorney at Cabinet Beau de Lomnie, France
- Sebastian Pado, University of Stuttgart, Germany
- Parvaz Mahdabi, Swisscom, Switzerland
- Gabriela Ferraro, Australian National University, Australia
- Wlodek Zadrozny, UNC Charlotte, USA
- Bharathi Raja Chakravarthi, Insight Centre for Data Analytics, National University of Ireland, Galway

7. Conclusions

The workshop organized this year addressed researchers from academics as well as industrial experts from relevant domains and aimed to establish a two-way communication channel between both. The general feedback was very positive and participants recommended keeping the good mix of scientific and practical presentations and the demos.

The participating experts expressed that such an event was missing since a while and efforts towards this direction are welcome by both - IP experts as well as academic researchers who supported the workshop actively by, for example, providing data or participating in paper reviewing as part of the programme committee.

The PatentSemTech workshop can be seen as a first initiative to establish a patent data mining community and will be more than a one-day event per year. Our intention is to make it into an active community with webinars on relevant topics, training and assessment activities to promote patent data mining and creating benchmark data to address different patent use cases and tasks.

We plan to run the workshop for three years, and a selected set of peer-reviewed and accepted scientific papers will be invited to be published in a Virtual Special Issue (VSI) of the World Patent Information "Text Mining and Semantic Technologies in the Intellectual Property Domain". Submissions to the VSI is possible also during the years after the workshop has taken place. In addition, it is planned to establish social media channels (Twitter, LinkedIn) in order to publish news related to the workshop and future activities. We recommend all interested researchers to have a look at our workshop website, where we will update datasets and resources, or announce interesting results and events.

References

[1] Assad Abbas, Limin Zhang, Samee U. Khan, A literature review on the state-of-the-art in patent analysis, World Patent Information, Volume 37, 2014.

[2] Aras, Hidir, Ren Hackl-Sommer, Michael Schwantner and Mustafa Sofean. Applications and Challenges of Text Mining with Patents. IPaMin@KONVENS, 2014.

[3] Aaron Abood and Dave Feltenberger. 2018. Automated patent landscaping. Artif. Intell. Law 26, 2 (June 2018), 103-125.

[4] D. Alberts, C. Barcelon Yang, D. Fobare-DePonio, K. Koubek, S. Robins, M. Rodgers, E. Simmons, D. DeMarco. 2017. Introduction to Patent Searching: Practical Experience and Requirements for Searching the Patent Spaces. In [27].

[5] L. Andersson, M. Lupu, J. Palotti, A. Hanbury, and A. Rauber. When is the time ripe for natural language processing for passage patent retrieval monitoring of vocabulary shifts over time. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM 16, 2016.

[6] L. Andersson, M. Lupu, Joo R. M. Palotti, F. Piroi, A. Hanbury, and A. Rauber. Insight to hyponymy lexical relation extraction in the patent genre versus other text genres. In Proceedings of the First International Workshop on Patent Mining and Its Applications (IPaMin 2014) co-located with Konvens 2014, Hildesheim,Germany, October 6-7, 2014., 2014.

[7] Anick, P. G., M. Verhagen, and J. Pustejovsky. "Identification of Technology Terms in Patents." LREC. 2014.

[8] J. Alex, H. Schtze, and S. Brgmann. "Unsupervised training set generation for automatic acquisition of technical terminology in patents." Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical Papers. 2014.

[9] Aras, H.; Trker, R.; Geiss, D.; Milbradt, M.; Sack, H. Get Your Hands Dirty: Evaluating Word2Vec Models for Patent Data, In Proc. of the 14th Int. Conf. on Semantic Systems (SEMANTICS 2018), P&D Track, CEUR workshop proceedings vol. 2198, 2018.

[10] H. Beltz, A. Fueloep, R. R. Wadhwa, P. Erdi, From ranking and clustering of evolving networks to patent citation analysis, in: Neural Networks 350 (IJCNN), 2017 International Joint Conference on, IEEE.

[11] Carvalho, Danilo & Nguyen, Minh-Le. (2017). Efficient Neural-based patent document segmentation with Term Order Probabilities. ESANN 2017 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 26-28 April 2017.

[12] Don, S & Min, Dugki. (2016). Feature Selection for Automatic Categorization of Patent Documents. Indian Journal of Science and Technology.

[13] R. Du, B. Drake, H. Park, Hybrid Clustering based on Content and Connection Structure using Joint Nonnegative Matrix Factorization, 2017, arXiv:1703.09646.

[14] T. Fink. Improving Multi Word Term Detection in the Patent Domain with Deep Learning. 2018 Master thesis. TU Wien

[15] Fujii, M. Iwayama, and N. Kando. Introduction to the special issue on patent processing. Information Processing & Management, 43(5):1149–1153, 2007. Patent Processing.

[16] Hackl-Sommer, Rene; Schwantner, Michael. Patent Claim Structure Recognition. Archives of Data Science, Series A, 2017, v. 2(1), 15

[17] Hanbury, V. Zenz, and H. Berger. 1st international workshop on advances in patent information retrieval (aspire10). SIGIR Forum, 44(1):1922, August 2010.

[18] D. Hunt, L. Nguyen, and M. Rodgers. Patent Searching: Tools & Techniques. Wiley, 2007.

[19] N. Ide and J. Pustejovsky, eds. Handbook of Linguistic Annotation. Springer, 2017.

[20] A. B. Jaffe, S. R. Peterson, P. R. Portney and R. N. Stavins. 1995. Environmental Regulation and the Competitiveness of U.S. Manufacturing: What Does the Evidence Tell Us? In Journal of Economic Literature. Vol. (33). No (1). pages 132-163. American Economic Association

[21] Hu, Jie, Shaobo Li, Yong Yao, Liya Yu, Guanci Yang and Jianjun Hu. Patent Keyword Extraction Algorithm Based on Distributed Representation for Patent Classification. Entropy 20 (2018): 104. Sunghae Jun, Sang-Sung Park, and Dong-Sik Jang. 2014. Document clustering method using dimension reduction and support vector clustering to overcome sparseness. Expert Syst. Appl. 41, 7 (June 2014), 3204-3212.

[22] J. Jrgens, C. Womser-Hacker, and T. Mandl. Modeling the interactive patent retrieval process: an adaptation of marchioninis information seeking model. In Fifth Information Interaction in Context Symposium, IIiX 14, Regensburg, Germany, August 26-29, 2014, pages 247250, New York, NY, USA, 2014.

[23] D. Franz, Kogler, G. Heimeriks and L. Leydesdorff. 2018 Patent portfolio analysis of cities: statistics and

maps of technological inventiveness, In European Planning Studies, Routledge, Vol. (26), No (11), pages 2256-2278, Routledge

[24] R. Krestel and P. Smyth. Recommending patents based on latent topics. In Proceedings of the 7th ACM Conference on Recommender Systems, RecSys 13, pages 395398, New York, NY, USA, 2013. ACM.

[25] A. Leeuwenberg, M. Vela, J. Dehdari, J. van Genabith. A Minimally Supervised Approach for Synonym Extraction with Word Embeddings. The Prague Bulletin of Mathematical Linguistics volume 105, Pages 111-142, De Gruyter, Berlin, Germany, 4/2016.

[26] M. Lupu, L. Papariello, R. Alentorn, M. Baycroft, J. List, The WPI patent test collection, World Patent Information, Volume 56, 2019, Pages 78-85, ISSN 0172-2190.

[27] Mihai Lupu, Katja Mayer, Noriko Kando, and Anthony J. Trippe. Current Challenges in Patent Information Retrieval (2nd ed.). Springer Publishing Company, Incorporated, 2017.

[28] M. Lupu, K. Mayer, J. Tait, and A. J. Trippe. Current Challenges in Patent Information Retrieval. Springer Publishing Company, Incorporated, 1st edition, 2011.

[29] M. Lupu, A. Hanbury, and A. Rauber. 4th international workshop on patent information retrieval (pair11). In Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011, pages 26232624.

[30] M. Lupu and A. Hanbury (2013), ”Patent Retrieval”, Foundations and Trends in Information Retrieval: Vol. 7: No. 1, pp 1-97.

[31] M. Lupu, J. Huang, and J. Zhu. Evaluation of chemical information retrieval tools. In Current Challenges in Patent Information Retrieval. Springer, 2011.

[32] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of the 27th Annual Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, NV, USA, 2013.

[33] F. Piroi, M. Lupu, and A. Hanbury. Passage retrieval starting from patent claims. A clef-ip 2013 task overview. In Working Notes for CLEF 2013 Conference, Valencia, Spain, September 23-26, 2013.

[34] O. Nekhayenko. Eigennamenerkennung fr Technologien. Implementierung und Evaluierung eines Prototyps fr Patente, 2016 Master Thesis, Stiftung Universitt Hildesheim.

[35] H. Menkge. Computer Aided Patent Processing: Natural Language Processing, Machine Learning, and Information Retrieval. PhD Thesis Drexel University library 2018.

[36] W. Shalaby and W. Zadrozny. Patent retrieval: A literature review. CoRR, abs/1701.00324, 2017.

[37] P. Sharma, R. Tripathi and R. C. Tripathi, ”Finding similar patents through semantic expansion,” 2016 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, 2016, pp. 1-5.

[38] Sofean M., Aras H., Alrifai A. (2018) A Workflow-Based Large-Scale Patent Mining and Analytics Framework. Information and Software Technologies, In: Damaeviius R., Vasiljevien G. (eds) Information and Software Technologies. ICIST 2018. Communications in Computer and Information Science, vol 920. Springer, Cham.

[39] Sofean, M. Automatic Segmentation of Big Data of Patent Texts in: Proceedings of the International Conference on Big Data Analytics and Knowledge Discovery. DaWaK 2017. Springer pp 343-351.

[40] S. Taduri, Gloria T. Lau, Kincho H. Law, and Jay P. Kesan. A patent system ontology for facilitating retrieval of patent related information. In Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance, ICEGOV, 2012.

[41] J. Tait, M. Lupu, H. Berger, G. Roda, M. Dittenbach, A. Pesenhofer, E. Graf, and van Rijsbergen K. Patent search: An important new test bed for ir. In The Dutch-Belgian Information Retrieval Workshop, 2009.

[42] J. Tait. Information retrieval facility symposium in Vienna. SIGIR Forum, 42(1):67, 2008.

[43] C. V. Trappey, H.Y. Wu, F. Taghaboni-Dutta, and A. J. C. Trappey, Using patent data for technology forecasting: China RFID patent analysis, Advanced Engineering Informatics, 2011.

[44] Trker, R.; Zhang, L.; Koutraki, M.; Sack H. The Less Is More for Text Classification”, In Proc. of the 14th Int. Conf. on Semantic Systems (SEMANTICS 2018) P&D Track, CEUR workshop proceedings vol. 2198.

[45] B. Van Looy, J. Callaert and K. Debackere. 2006. Publication and patent behavior of academic researchers: Conflicting, reinforcing or merely co-existing?. Research Policy. Vol. (35), No (4),pages 596 - 608,

[46] Wu, T.; Zhang, D.; Zhang, L.; Qi, G. Cross-Lingual Taxonomy Alignment with Bilingual Knowledge Graph Embeddings, 7th Joint International Semantic Technology Conference (JIST 2017), Cold Coast, QLD, Australia, November 10-12, 2017.

[47] J. Yoon and K. Kim, TrendPerceptor. A property function based technology intelligence system for identifying technology trends from patents, Expert Systems with Applications, 2012.

[48] H. Yu, S. Taduri, J. Kesan, G. Lau, and H. Law Kincho. Retrieving information across multiple, related domains based on user query and feedback: Application to patent laws and regulations. In Proceedings of the 4th International Conference on Theory and Practice of Electronic Governance, ICEGOV 10, pages 143151, New York, NY, USA, 2010.

[49] S. Jun, S.-S. Park, D.-S. Jang, Document clustering method using dimension reduction and support vector clustering to overcome sparseness, Expert Systems with Applications, 2014.

[50] L. Zhang, L. Li, and T. Li. Patent mining: A survey. SIGKDD Explor. Newsl. 16(2):119, May 2015.

[51] L. Zhang. An Integrated Framework for Patent Analysis and Mining. PhD thesis, Florida International University, USA, 2016.