An Ontology-based Collaboration Recommender System using Patents

Sandra Geisler[1], Rihan Hai[1] and Christoph Quix[2]
[1]*Databases and Information Systems, RWTH Aachen University, Germany*
[2]*Fraunhofer-Institute for Applied Information Technology FIT, St. Augustin, Germany*

Keywords: Ontology Engineering, Patent Analysis, Ontology Matching.

Abstract: Successful research and development projects start with finding the right partners for the venture. Especially for interdisciplinary projects, this is a difficult task as experts from foreign domains are not known. Furthermore, the transfer of knowledge from research into practice is becoming more important in research projects to enable the quick application of research results. This is in particular relevant for projects in medical engineering. Patents and publications contain technical knowledge which can be exploited to find suitable experts. Patents are usually more product-oriented as the inventors have to describe an application area and products might be protected by patents. On the other hand, scientific publications represent the state-of-the-art in research. The challenge is finding the right mixture of research- or application-oriented experts from different domains. Hence, we propose a recommender system for experts for a certain topic based on patent topic clustering, ontologies, and ontology matching, which maps patents to corresponding innovation fields. The medical engineering domain serves as a first test bed, since projects in this area are highly interdisciplinary.

1 INTRODUCTION

Innovation drives research and industry. It is important in both fields to be up-to-date to what will be promising in the future. Especially, medical engineering (ME) is an "innovative, strongly growing, and promising industry in Germany" [1]. In ME, interdisciplinary projects are very common as experts from medicine, engineering, and other disciplines are required. Furthermore, this domain is highly dependent on its innovative capabilities as product cycles are getting shorter and shorter.

Ventures in research and industry rise and fall with the expertise of the partners in the project team. Hence, it is crucial for the success of innovative projects and their proposals to find suitable partners. Studies demonstrate, that collaboration between research institutions and companies are beneficial for both, product and process innovations (Robin and Schubert, 2013). Especially in interdisciplinary projects, the search for experts in unfamiliar domains is time consuming, cumbersome, and might not be as successful as expected. Hence, to assist the process of finding partners for a venture, a recommendation system is desired which speeds up the search and helps to discover collaboration opportunities.

Patents contain a wealth of technical information used for the development of products, but are at the same time hard to analyze as they are written using special terminology (Aras et al., 2014; Zhang et al., 2015). As patent inventors are not only experts in their field, but also have a product-oriented view on ME research, they constitute interesting projects partners. Therefore, we propose an approach using patent clustering, ontology mappings, and ontology matching to recommend collaboration opportunities.

In the mi-Mappa project[2], we aim at finding suitable experts for ME projects based on patents and innovation fields. According to (Schlötelburg et al., 2008), an innovation field in ME is defined as an area which has significant innovation activity, future potential, and a value chain as complete as possible. The main innovation fields for ME comprise (Schlötelburg et al., 2008; Deutsche Gesellschaft für Biomed. Technik im VDE, 2012): Imaging Techniques, Protheses and Implants, Medical Information Systems and Telemedicine, Interventional Devices, Systems, and Techniques, In-vitro Technology, Special Therapy and Diagnostic Systems, and cross-sectional topics, such as patient safety.

In this paper, we propose an approach that combines two complementary ways: 1. We build a profile

[1]http://www.bvmed.de/branchenbericht

[2]http://www.dbis.rwth-aachen.de/mi-Mappa

Geisler, S., Hai, R. and Quix, C..
An Ontology-based Collaboration Recommender System using Patents.
In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 2: KEOD*, pages 389-394
ISBN: 978-989-758-158-8

of the expert, which includes her publications, websites, and other business related information. This comprises the identification of the inventor with a corresponding author of scientific publications. If we find a corresponding author, we can match her publications to innovation fields to identify her specialization areas. We can use classification terms present in publication databases and semi-automatic ontology matching to create the mappings. 2. If an inventor could not be identified as an author of scientific papers, the patents are clustered by topic and these topics are finally mapped to innovation fields using ontology matching.

In the following section, we will discuss existing works in patent analysis and collaboration recommendations. A full description of our approach is given in Section 3. Finally, Section 4 concludes the paper.

2 RELATED WORK

Collaboration/Expert Recommender Systems. The process of finding collaboration opportunities often involves a manual process. For example, predefined criteria are analyzed, and scores are calculated and weighted based on these criteria (Geum et al., 2013; Awasthi et al., 2015). Partners in supply chains can be found by using supervised and unsupervised learning, statistics, and analyzing criteria (Wu and Barnes, 2011). In the field of finding partners for R&D projects, no related semi-automatic or automatic approach could be found.

Systems to find experts for a certain topic are based on self-disclosure (personal information maintained manually), authored documents, or social network activity (Wang et al., 2013). The systems can also be categorized into expert profiling and expert finding (Balog and De Rijke, 2007). The most recent works are using algorithms from social network analysis, such as the link analysis algorithms PageRank or HITS (Rafiei and Kardan, 2015; Wang et al., 2013) and graph-based algorithms (Rani et al., 2015). We will concentrate on expert finding using authored documents (e.g., patents & publications) as we do not need (yet) a complete profile of a researcher. Many document-based Expert Recommender Systems (ERS) are only using enterprise-level documents and are restricted to employees in the same company. In contrast, we propose a document-based approach which uses information from *any* publications and patents available. The DEMOIR approach (Yimam-Seid and Kobsa, 2003) also uses ontologies and domain models for expert finding, but they use them to model the expertise only.

Patent Analysis using Ontologies. The usefulness of ontologies has also been recognized for the patent domain, especially for patent search (Bonino et al., 2010). The PatExpert system, for example, uses a network of ontologies and knowledge bases to enable patent search, classification, and clustering (Wanner et al., 2008). Trappey et al. propose a system that calculates the conditional probability that, given a specific text chunk is present in the document, the chunk is mapped to a specific concept of a given ontology (Trappey et al., 2009). Patent similarity is then based on the number of common matched concepts. This approach restricts the clustering to the terms of the ontology which might lead to missing important terms not present in the ontology.

Patent Clustering. An overview of patent document contents can be retrieved by clustering. Tseng et al. propose a full-text patent clustering methodology which includes document clustering, term clustering, and multi-stage clustering to avoid skewed distribution among clusters (Tseng et al., 2007). TF or IDF (see section 3.2) filtered terms are clustered according to their co-occurrence. Moreover, each cluster obtains a summary title by statically calculating the most frequent terms in the clusters with correlation coefficient method. A bibliometric approach based on co-citation analysis is introduced in (Mogee and Kolar, 1999). The co-cited documents are linked under the assumption that they share the subject matter. The result of the approach also indicates core competencies in the corresponding industrial field. However, using co-citation to group patents may lead to superficial results due to the lack of internal knowledge of the patents (Yoon and Park, 2004). Another drawback is that patents without references are excluded from this approach. Trappey et al. describe a methodology to cluster patents in three steps. First they extract the key phrases of a patent, i.e., they use an ontology-based, statistical method to extract key phrases which represent an important topic in the document. Afterwards they build Technology Clusters of these key phrases using an non-exhaustive overlapping cluster algorithm proposed by Chen and Hu (Trappey et al., 2010; Chen and Hu, 2006). By calculating the cumulative weights for the key phrases of a document, they can determine the Technology Cluster for a patent. In the last step, they use the same clustering technique to cluster the patent documents.

In summary, all of the approaches may cover a part of our approach, but we present a novel approach which combines the use of patent analysis, clustering, ontology design, and ontology matching to recommend experts for a R&D collaboration.

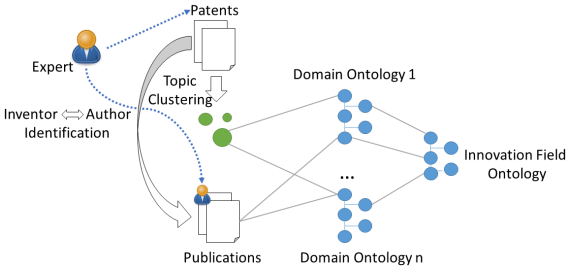

Figure 1: Architecture to map patents to innovation fields.

3 APPROACH

Based on the processed information type, patent analysis can be categorized into quantitative and qualitative approaches [3], which we both utilize in our approach. Qualitative approaches extract patent metadata, e.g., inventors, references, affiliation, while quantitative methods process the full-text, such as claims and abstracts of patent documents. The goal of the overall approach, depicted in Figure 1, is to map inventors of a patent and subsequently the corresponding patents to innovation fields. We propose to do this in two different, but supplementary ways: 1. Match inventors with publication authors. 2. Cluster patents based on the topics they cover. For both approaches, we preselect patents based on a search with keywords extracted from the description of the intended project. The resulting patents can be used to initialize the clustering. In both approaches the results (publications or topics) will be matched to existing medical ontologies. If a mapping between the matched medical terms and concepts of an innovation field exists, a link between patent inventor and innovation fields has been found.

In the following, we detail the two process flows. Subsequently, we describe the design of the utilized ontologies and the matching in the ontology network.

3.1 Inventor-Author-Identification

It is first assumed, that an inventor may be also working as a researcher and is publishing scientific articles in the same domain. We map inventors to authors, because articles published in journals or conferences are often classified by publishers and publication databases. Those classifications usually use terms which are easier to harvest and match to domain ontologies than the abstract International Patent Classification (IPC) or similar classifications. Very promising are bibliographic search engines such as Web of Science, or PubMed, which offer classification of papers according to well known medical taxonomies. We will search the bibliographic databases by author name using a corresponding API. For author identification, we use a multi-step process that uses clustering techniques and statistics to determine the highest probability of an author being the same person as an inventor of a certain patent. We extract the keywords and classification terms from the metadata of the papers and match them with medical ontologies which in turn will be mapped to an innovation field ontology.

3.2 Topic Clustering

A study showed that 70-90% of technological knowledge is only published in patents[4]. Hence, it can be assumed that there exist inventors who have only published patents and an alternative way of mapping patents and inventors to an innovation field has to be found for those persons. We propose to use topic clustering of the patents to do so. This second approach can also be used to verify the results of the first approach.

In patent analysis, classifications such as the IPC are often too broad for specific analytical usage (Tseng et al., 2007) and more detail-oriented categorizations are needed. Clustering methods group objects such that the similarity between objects in the same cluster is greater than the similarity of objects in different clusters. The similarity is usually measured in terms of their relative position in an n-dimensional space using Euclidean or Manhattan distances.

We utilize a set of common preprocessing techniques to facilitate feature extraction, indexing, and clustering. These comprise, amongst others, document parsing, tag removal, tokenization, and lowercasing. Additional steps, such as stemming, pruning, and stopword removal, help to reduce the term set size and increase its quality, improving clustering accuracy (Gonçalves et al., 2010). In each document, only the key terms are selected to present the features of the document, utilizing Inverse Document Frequency (IDF) and Term Frequency (TF) within certain thresholds. These terms are weighed based on TF $\times$ IDF, followed by calculating the similarity using clustering algorithms. Documents are usually merged to clusters successively (hierarchical clustering) or distributed to certain clusters defined in the beginning (partition clustering). We use both kinds of distance-based clustering in our approach, including K-means

[3] http://www.wipo.int/sme/en/documents/patent_information_fulltext.html

[4] http://www.integrityip.com/Patent_Library/Community/Other/GlobalPatentSources.pdf

and K-medoid algorithms with pre-chosen centroids from the query results.

Moreover, it is common in patent analysis that a patent includes multiple features, claims, or inventors. Hence, the non-exhaustive overlapping clustering algorithm (Trappey et al., 2010) is adopted in our approach. Finally, each cluster receives a title generated based on the top k frequent terms (Yang et al., 2000). Furthermore, as performance is an issue in full text analysis of a large document collection, we apply the text analysis only to a part of each patent document (e.g., the first part of the abstract, claims, or introduction). It has been proven that such an approach may achieve better performance than using full texts (Fall et al., 2003).

3.3 Ontologies

Selection of Existing Domain Ontologies. Our approach is heavily relying on the mappings to medical ontologies and subsequently from medical ontologies to the innovation field ontology. A plethora of medical ontologies exist. Hence, we have to analyze which set of ontologies covers as many terms as possible, describing the innovation fields.

We made a first analysis by searching for ontologies in the Bioportal [5] search engine using terms describing innovation fields. The Bioportal search engine is the most comfortable and comprehensive search engine in the life science domain. In addition, it offers several useful tools, e.g., an ontology recommendation tool based on keywords or full-texts. Moreover, we used the Ontology Lookup Service [6] and the Ontobee [7] search engine to have a broad overview. For the search, 174 terms from the six innovation fields extracted from the reports (Schlötelburg et al., 2008; Deutsche Gesellschaft für Biomed. Technik im VDE, 2012) have been used. The most promising four ontologies found were the National Cancer Institute (NCIT) Thesaurus, the Systematized Nomenclature of Medicine - Clinical Terms (SNOMEDCT), MeSH, and the Robert Hoehndorf Version of MeSH (RHMeSH). For these we did a coverage analysis presented in Figure 2. The coverage is the percentage of the innovation field terms present in each of the ontologies.

Note that no ontology really outperforms the others and that the overall coverage is very low. Hence, we decided to analyze the coverage by adding one ontology after another, to see the gain of adding further ontologies. We used the most promising ontologies identified before and started with the NCI Thesaurus. Figure 3 shows the results. It can be noted, that we gain about 10% coverage using all ontologies. The biggest gain is achieved after adding the MeSH ontology. An analysis with the same terms using the Bioportal Recommender tool as depicted in Figure 4 delivers a similar result. The Recommender tool analyzes the annotations which can be found using the given terms. The coverage is calculated taking into account amongst others the mappings to and synonyms from other ontologies, and the size of the ontologies (Jonquet et al., 2010). Hence, it is not directly comparable to the manually created coverage result described above. There are two recommended ontology sets ranked highest: The first comprises the NCIT, MeSH, and the Computer Retrieval of Information on Scientific Projects (CRISP) and resulted in a coverage of 83.6% and an overall result of 68.9%. The second is congruent with our selection: NCIT, MeSH, and SNOMEDCT and resulted in a coverage

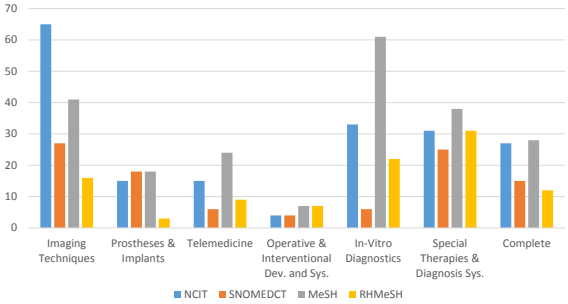

Figure 2: Coverage of search terms in selected ontologies.

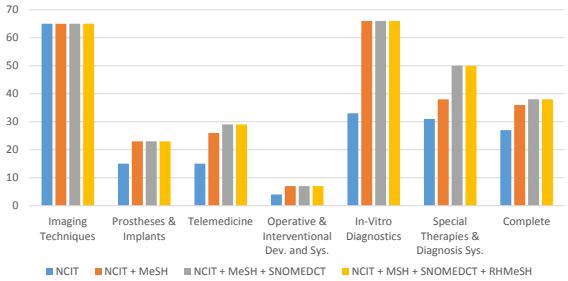

Figure 3: Coverage based on combination of ontologies.

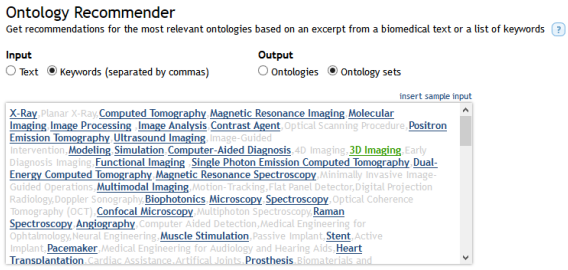

Figure 4: The Bioportal Recommender Tool.

[5] http://bioportal.bioontology.org
[6] http://www.ebi.ac.uk/ontology-lookup
[7] http://www.ontobee.org

of 84.6% and an overall score of 68.9%. Furthermore, all the ontologies have to be assessed according to their quality. A detailed quality analysis will be made using acknowledged quality criteria (Vrandečić, 2009; Gomez-Perez, 2004). We are still in the process of finding a suitable tool, such as the Ontology Pitfall Scanner (OOPS!) [8] (Poveda-Villalón et al., 2012), which will assist us in this regard.

Requirements Analysis & Design of the Innovation Field Ontology. To reach our goal to map publications and clusters to innovation fields, we also need a new ontology which only represents the innovation fields and important terms describing them. To that end, we will make a detailed requirements analysis including interviews with domain experts, analysis of existing ontologies, and an intensive literature research. Where applicable, we will stick to the NeOn methodology (Suárez-Figueroa, 2010) and especially for the requirements analysis, the creation of an Ontology Requirements Specification Document (Suárez-Figueroa et al., 2009) will be useful. In a first step, we have already extracted a preliminary selection of 174 terms (the same terms as for ontology search), which corresponds to the scenario "Reusing and reengineering non-ontological resources" of the NeOn methodology. The terms can be used to make a first draft of a preliminary ontology which is verified during an expert interview. The ontology design will be accompanied by its evaluation using the criteria and tools mentioned above.

3.4 Ontology Matching & Mappings

To identify which collaboration partner is working in which innovation field, we have to integrate all the information we gathered so far. We need to map the cluster terms and the publications, respectively, to the domain ontologies. The mappings between the domain ontologies and the innovation field ontology can be established during design time, because it is not expected that they will change frequently. Ontology matching systems (e.g., our tool GeRoMeSuite (Kensche et al., 2007)) will be used to identify a first set of mappings. Additionally, existing mappings created by the BioPortal can be used to infer further mappings to the innovation field ontology.

Bioportal also provides prepared mappings between ontologies. It creates the mappings either using the LOOM algorithm, the Unified Medical Language System (UMLS) concept unique identifiers (CUI), and the Open Biological and using Biomedical Ontologies (OBO) xref properties[9] (Ghazvinian et al., 2009). These mappings can be retrieved via a REST API offered by the BioPortal website. Afterwards, a domain expert verifies the detected mappings.

If necessary, the creation of mappings between publication classifications and domain ontologies can be prepared during design time. We will also use the semi-automatic matching process described before. More challenging is the creation of mappings between the clusters and the domain ontologies. This has to be done during run time, as the cluster terms are not known in advance. We plan to use also the matching algorithms provided by GeRoMeSuite for this step.

4 CONCLUSION

We have presented an innovative ontology-based approach for recommending experts for research projects in ME. We are making extensive use of ontology engineering in our approach, e.g., analysis, creation, and matching of ontologies, defining requirements for ontologies, and evaluation of ontologies. Also techniques from other areas, such as text mining and patent analysis, are included in our approach. Our current work focuses on the modeling of the ontologies and the selection of the clustering methods. Performance is an issue for text clustering, as we want to have an interactive system.

The work is still in an early stage and we have to see how the integration of text clustering, topic modeling, patent analysis, and ontology matching performs. The various techniques are also challenging tasks if they are considered separately, but the combination of the techniques may show an innovative method for exploring unknown research fields. Our approach is not limited to the field of ME; however, the availability of a huge number of ontologies in the life sciences contributes to our approach.

ACKNOWLEDGEMENTS

This work has been supported by the Klaus Tschira Stiftung gGmbH in the context of the mi-Mappa project (http://www.dbis.rwth-aachen.de/mi-Mappa/, project no. 00.263.2015). We would like to thank our partners in the mi-Mappa project for their fruitful ideas. We would further like to thank Tanja Schmelter for her work on ontology analysis.

[8] http://oops.linkeddata.es

[9] http://www.bioontology.org/wiki

REFERENCES

Aras, H., Hackl-Sommer, R., Schwantner, M., and Sofean, M. (2014). Applications and challenges of text mining with patents. In *Proc. Intl. Workshop on Patent Mining and its Applications*. Stiftung Univ. Hildesheim.

Awasthi, A., Adetiloye, T., and Crainic, T. G. (2015). Collaboration partner selection for city logistics planning under municipal freight regulations. *Applied Mathematical Modelling*.

Balog, K. and De Rijke, M. (2007). Determining expert profiles (with an application to expert finding). In *IJCAI*, volume 7, pages 2657–2662.

Bonino, D., Ciaramella, A., and Corno, F. (2010). Review of the state-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics. *World Patent Information*, 32(1):30–38.

Chen, Y.-L. and Hu, H.-L. (2006). An overlapping cluster algorithm to provide non-exhaustive clustering. *Europ. J. of Operational Research*, 173(3):762–780.

Deutsche Gesellschaft für Biomed. Technik im VDE (2012). Empfehlungen zur Verbesserung der Innovationsrahmenbedingungen für Hochtechnologie-Medizin. Technical report, VDE.

Fall, C. J., Törcsvári, A., Benzineb, K., and Karetka, G. (2003). Automated categorization in the international patent classification. In *ACM SIGIR Forum*, volume 37, pages 10–25. ACM.

Geum, Y., Lee, S., Yoon, B., and Park, Y. (2013). Identifying and evaluating strategic partners for collaborative r&d: Index-based approach using patents and publications. *Technovation*, 33(6):211–224.

Ghazvinian, A., Noy, N., and Musen, M. (2009). Creating mappings for ontologies in biomedicine: simple methods work. In *AMIA Ann. Symp. Proc.*, pages 198–202.

Gomez-Perez, A. (2004). Ontology evaluation. In Staab, S. and Studer, R., editors, *Handbook on Ontologies*, pages 250–273. Springer.

Gonçalves, C. A., Gonçalves, C. T., Camacho, R., and Oliveira, E. C. (2010). The impact of pre-processing on the classification of medline documents. In *PRIS*, pages 53–61.

Jonquet, C., Musen, M. A., and Shah, N. H. (2010). Building a biomedical ontology recommender web service. *J. Biomedical Semantics*, 1(S-1):S1.

Kensche, D., Quix, C., Li, X., and Li, Y. (2007). *GeRoMeSuite*: A system for holistic generic model management. In *Proc. VLDB*, pages 1322–1325.

Mogee, M. E. and Kolar, R. G. (1999). Patent co-citation analysis of eli lilly & co. patents. *Expert Opinion on Therapeutic Patents*, 9(3):291–305.

Poveda-Villalón, M., Suárez-Figueroa, M. C., and Gómez-Pérez, A. (2012). Validating ontologies with oops! In *Knowledge Engineering and Knowledge Management*, pages 267–281. Springer.

Rafiei, M. and Kardan, A. A. (2015). A novel method for expert finding in online communities based on concept map and pagerank. *Human-centric Computing and Information Sciences*, 5(1):1–18.

Rani, S. K., Raju, K., and Kumari, V. V. (2015). Expert finding system using latent effort ranking in academic social networks. *Intl. J. of Information Technology and Computer Science*, 2:21–27.

Robin, S. and Schubert, T. (2013). Cooperation with public research institutions and success in innovation: Evidence from france and germany. *Research Policy*, 42(1):149–166.

Schlötelburg, C., Weiß, C., Hahn, P., Becks, T., and Mühlbacher, A. C. (2008). Identifizierung von Innovationshürden in der Medizintechnik. Technical report, Bundesministeriums für Bildung und Forschung.

Suárez-Figueroa, M. C. (2010). *NeOn Methodology for building ontology networks: specification, scheduling and reuse*. PhD thesis, Universidad Politecnica de Madrid.

Suárez-Figueroa, M. C., Gómez-Pérez, A., and Villazón-Terrazas, B. (2009). How to write and use the ontology requirements specification document. In *Proc. OTM 2009*, pages 966–982. Springer.

Trappey, A. J., Trappey, C. V., Hsu, F.-C., and Hsiao, D. W. (2009). A fuzzy ontological knowledge document clustering methodology. *IEEE Trans. on Systems, Man, and Cybernetics, Part B*, 39(3):806–814.

Trappey, C. V., Trappey, A. J., and Wu, C.-Y. (2010). Clustering patents using non-exhaustive overlaps. *System Science and System Engineering*, 19(2):162–181.

Tseng, Y.-H., Lin, C.-J., and Lin, Y.-I. (2007). Text mining techniques for patent analysis. *Information Processing & Management*, 43(5):1216–1247.

Vrandečić, D. (2009). Ontology evaluation. In Staab, S. and Studer, R., editors, *Handbook on Ontologies*, chapter 13, pages 293–313. Springer.

Wang, G. A., Jiao, J., Abrahams, A. S., Fan, W., and Zhang, Z. (2013). Expertrank: A topic-aware expert finding algorithm for online knowledge communities. *Decision Support Systems*, 54(3):1442–1451.

Wanner, L., Baeza-Yates, R., Brügmann, S., Codina, J., Diallo, B., Escorsa, E., Giereth, M., Kompatsiaris, Y., Papadopoulos, S., Pianta, E., et al. (2008). Towards content-oriented patent document processing. *World Patent Information*, 30(1):21–33.

Wu, C. and Barnes, D. (2011). A literature review of decision-making models and approaches for partner selection in agile supply chains. *Purchasing and Supply Management*, 17(4):256–274.

Yang, Y., Ault, T., Pierce, T., and Lattimer, C. W. (2000). Improving text categorization methods for event tracking. In *Proc. of the 23rd Intl. Annual ACM SIGIR Conf.*, pages 65–72. ACM.

Yimam-Seid, D. and Kobsa, A. (2003). Expert-finding systems for organizations: Problem and domain analysis and the demoir approach. *J. of Organizational Computing and Electronic Commerce*, 13(1):1–24.

Yoon, B. and Park, Y. (2004). A text-mining-based patent network: Analytical tool for high-technology trend. *The Journal of High Technology Management Research*, 15(1):37–50.

Zhang, L., Li, L., and Li, T. (2015). Patent mining: A survey. *ACM SIGKDD Expl. Newsletter*, 16(2):1–19.