Günter Reiner – Department of Law, Helmut Schmidt University, Germany
Philipp Adämmer – Department of Mathematics and Statistics, Helmut Schmidt University, Germany

Similarities Between Human Structured Subject Indexing and Probabilistic Topic Models

Abstract:
This paper adds statistical findings from natural language processing to an ongoing interdisciplinary research project between lawyers and information scientists. The project proposes an indexing scheme that follows a content grid of six predefined categories, also called facets in a broad sense. These facets try to capture the essential structure of legal information and are based on Ranganathan's fundamental facets as well as on the famous Roman lawyer's Gaius tripartite division of persons, things and actions. The prototype database, built as part of the project, consists of nearly 2,500 cases which have been manually indexed. The present study uses the prototype database to investigate whether similarities exist between human subject indexing and automatically clustered terms. We first examine the similarities between the indexing terms and terms generated by a term frequency–inverse document frequency approach. Then we investigate the similarities between the indexing terms and word clusters generated by unsupervised probabilistic topic models, namely latent Dirichlet allocation (LDA) and the correlated topic model (CTM). On average, the similarities of the manually indexed terms with the topic terms are not high but statistically significant, and for some cases we find strong similarities in which the clustered terms of the topic models match well the humanly indexed terms. Correlations are slightly higher when using the CTM instead of LDA. Our results indicate that topic modelling could be beneficial at least for semi-automatic indexing. Disentangling further which facets contain those terms that predominantly cause similarities with automatically clustered terms could further enhance the support for human structured subject indexing.

1.0 Introduction

Indexing is one of the oldest tools to foster the retrieval of written information. Even in the age of electronic full-text searches, it has proven its worth (Gross et al. 2015). The present project ties in with an ongoing project between lawyers and information scientists, which is funded by the Social Sciences and Humanities Research Council in Canada and in which one of the authors is involved (Cumyn et al. 2019). The Canadian project proposes an indexing scheme that follows a content grid of six predefined categories, also called facets in the broadest sense. These facets (Person, Action, Thing, Context, Legal category and Sanction) attempt to capture the essential structure, so to speak the "grammar" (Reiner et al. 2019, 352-353), of legal information in two ways: first in terms of the fundamental division between the factual elements of a case (the first four facets) and the legal consequences (the last two facets). Second, in terms of the way factual information is analysed, which is based on Ranganathan's universal facets (Personality, Matter, Energy, Space, Time) and the famous Roman jurist Gaius' famous tripartite distinction of persons, things and actions (Cumyn et al. 2018; 2019; Reiner et al. 2019).

A test database (called "Gaius"), which is an excerpt from the offering of the commercial Quebec database operator SOQUIJ (Société québécoise d'information juridique), has been created as part of the Canadian project. It contains 2,500 court decisions from Quebec (mostly in French) which are divided into five sub-databases according to those fields of law that SOQUIJ had assigned to them (administrative law, labour law,

contract law etc.). These decisions were manually re-indexed using a faceted scheme on the basis of a controlled vocabulary (thesaurus) being developed gradually and kept as lean as possible.

Human indexing is expensive and time consuming, which is why we attempt to find methods for semi- or even fully automatic indexing. It is not linked to the Canadian project in terms of its research goal, yet it is based on the test database described above, that is an interesting source of information, since it allows to compare human (manual) indexing with automated indexing. Legal databases that are systematically indexed according to content criteria are rare. In addition, the Gaius database has the advantage that the indexing is structured, which enables specific statistical analyses revealing indications of human indexing patterns. Probabilistic methods of computerized text analysis are more similar to the human understanding of (legal) texts than one might think; it is well-known that the natural acquisition and processing of language is based not on the application of rigid rules but on experience, which can be simulated using an inductive process based on the estimation of probabilities (Chater and Manning 2006, 340). Our project aims at investigating whether probabilistic topic models, such as latent Dirichlet allocation of Blei et al. (2003), create and assign word clusters to legal documents (court decisions) in a manner that is similar to human facet indexing. According to the literature, there has been experience with topic modeling in the field of legal information for several years (George et al. 2014; Livermore et al. 2017), but so far they remain isolated and do not serve the purpose of indexing.

We are approaching the purpose with the following sub-questions:

1. Are the similarities between human indexing and automatically generated keywords generally higher when using the words of the topic models instead of the words generated using a frequency–inverse document frequency approach?

2. Is there a statistically significant relationship between human indexing and keywords generated by topic modeling? If so, topic models can be useful for (semi-)automatic indexing.

3. Do facet indexing and topic keywords created by topic models correlate more strongly than unstructured indexing (*e.g.*, SOQUIJ indexing) and topic modeling keywords? If so, this could be an indication that faceted indexing - beyond its supposed advantages for queries - offers advantages in (semi) automation.

4. Are there some of the six facets which, by the terms assigned hereto, predominantly cause similarities with automatically clustered terms? If so, this insight could help to enhance human structured subject indexing (conceptually and in individual cases) up to semi-automatic indexing.

2.0 Empirical analysis

To test the hypotheses, we have automatically generated word lists using the term frequency–inverse document frequency approach (2.1) and the topic model approach (2.2). We compared the results with the manual indexing of the Gaius database using cosine similarities (2.3).

2.1 Term frequency–inverse document frequency

The term frequency inverse document frequency (tf-idf) aims to measure the relevance of a word within a certain document. The term frequency (tf) simply counts how often a word (hereafter *w*) occurs in a document (hereafter *d*). Yet the informational content of terms that occur frequently in many documents (*e.g., the, and, for*, . . . , etc.) is mostly low. It is rather those terms that appear frequently in a small number of documents but rarely in the other ones that tend to be informative (Huang, 2008, p. 51). The tf-idf accounts for this aspect whose formula can be written as follows:

$$tf\text{-}idf(w,d) = tf(w,d) \times log\left(\frac{N}{df(w)}\right), \tag{1}$$

where *t f* (*w*, *d*) is the term frequency of *w* in *d* and *df* (*w*) denotes the number of documents in which *w* occurs. *N* equals the number of documents in a corpus and *log*() is the logarithm with base 10.[1] Terms that occur frequently in some documents but only rarely in the overall corpus yield high tf-idf values. We computed tf-idf values for each word in each legal document of the sub-databases *Admin* (administrative law), *Contrats* (contract law) and *Travail* (labor law). The legal documents are grouped according to their sub-database. We used separately the 10, 20, . . . , 50 words of each document with the highest tf-idf values for comparison with the corresponding index terms from the Gaius database.

2.2 Probabilistic topic models

Probabilistic topic models (TM) are algorithms for the analysis of large document collections. In contrast to the tf-idf approach, TM can assign terms to documents that are not included in the document itself (cross-referencing). In addition, TM assume that documents are written by a stochastic process in which all documents share *K* common *topics*. A topic is a discrete probability distribution over words. All topics contain the same words, namely the totality of all words in the database, but the probabilities given to each word differ. For example, a topic about *damages* would give high probabilities to words such as *negligence* and *causation*, while a topic about *labor contracts* would put high probabilities on words such as *employee* and *dismissal*. Each document is then assumed to be a mixture of those corpus wide topics. The topic mixture for each document is given by the so-called *topic proportion*.

The most popular and most cited TM is latent Dirichlet allocation (LDA) by Blei et al. (2003). The model owes its name to the fact that the topics and the topic proportions are assumed to be drawn from a Dirichlet distribution. Each element of a randomly drawn Dirichlet vector is between zero and one. In addition, the elements of a Dirichlet vector sum up to one, thereby complying with the requirements of probabilities. However, one drawback is that the Dirichlet distribution cannot account for correlations. For example, if a legal document writes about contracts it is more likely that it also deals with *frustration* than if it deals with administrative law. Therefore, we also used the

[1] The base can certainly be changed, but we decided to stay with the default settings given and justified in the R-package *quanteda* by Benoit et al. (2018).

correlated topic model (CTM) by Blei and Lafferty (2007). As is common in natural language processing, we removed a whole bunch of stop words (*e.g. aux, notre, nous, que*), hyphens, apostrophes, numbers, etc., before estimating (computing) the models.[2] We also removed words that have been classified as irrelevant by ourselves (*e.g., demandeur*). To avoid the particular difficulties of a multilingual database, we have also excluded automatically the few English language decisions. Finally, we replaced certain words in the documents according to the synonym list of the Gaius thesaurus. Although being unsupervised learning algorithms, LDA and the CTM require the selection of the number of topics *K*. Both TM were estimated with $K = 10, 20, \ldots, 50$.[3] For each *K*, we assigned successively the 1, 2, 3 most probable topic(s) to each document (indicated by the topic proportions). We then chose successively the 5, 10, 15 most probable words of each topic. With a fixed number of topics *K*, we thus had 9 (3x3) combinations of terms to compare with the human indexing for each legal document.

2.3 Measuring similarities with human indexing

To measure how similar the words from the automated approach are to the humanly indexed terms of the Gaius database, we used the concept of cosine similarities (see, *e.g.*, Huang 2008). In the first step, we converted the total Gaius indexing and the automated word lists into a single so-called *document-term matrix* (dtm), where each row (*i*) corresponds to a classification of one legal document (Gaius or automated) and each column (*j*) denotes a unique word. Each cell entry thus indicates how often the word *j* occurred in the classification document *i*. A classification document can thus be represented as a vector in high dimensional space.

The idea of the cosine similarity is to measure the angle between two vectors. In our case, it measures how close the classification vectors from the Gaius database and each of the corresponding automated approaches are. On the one hand, if two vectors are identical, the angle between them is zero. The cosine of zero equals one; this value therefore represents the highest similarity. On the other hand, if the two vectors are orthogonal to each other (no words intersect), they have an angle of 90º, whose cosine value equals zero. We thus have a measure bounded between zero and one that indicates how close the Gaius indexation and our automated word lists are.

The cosine similarity between two vectors can be computed as:

$$\cos(\theta) = \frac{\mathbf{V}_G \cdot \mathbf{V}_M}{||\mathbf{V}_G||\ ||\mathbf{V}_M||} = \frac{\sum_{w=1}^{Z} V_{G,w} \cdot V_{M,w}}{\sqrt{\sum_{w=1}^{Z} V_{G,w}^2} \cdot \sqrt{\sum_{w=1}^{Z} V_{M,w}^2}}, \quad (2)$$

[2] It does not make sense to try to eliminate all irrelevant words in advance. This is not only time-consuming but above all, there are many terms whose (missing) relevance to the document content depends on the context.

[3] Choosing the optimal number of topics depends on the purpose of the analysis. For example, predicting unseen documents is a different task than trying to find the optimal number of semantically meaningful topics for a fixed set of documents. Several metrics have been proposed in the literature to find the optimal number of topics *K* (see, *e.g.*, Roberts et al. 2019). We used the R-package *topicmodels* by Grün and Hornik (2011) to estimate LDA and the *stm* package by Roberts et al. (2019) to estimate the CTM.

where *G* and *M* denote the vector of word counts for the Gaius database (*G*) and the automated (tf-idf or TM) word lists (*M*). The number of unique words is given by *Z*. The nominator computes the dot product of two vectors and the denominator is the product of the vectors' Euclidean norms. Since the Gaius database uses a lot of compounded words for indexation (*i.e.*, *permis d'alcool*), we have split them into separate strings to make the terms comparable. We also broke down the words to their stem, such that, for instance, words in plural are given in singular form.

3.0 Empirical Results

3.1 Quantitative Results

Table 1 shows an example of the five most probable words of the seven most prevalent topics in the sub-database *Contrats*. The topics were estimated with the CTM and the total number of topics was 20. As the topic model randomly assigns topic numbers, we renumbered the topics for illustration purposes from 1 to 7.

Table 1: Word distributions for seven selected topics[4]

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
somme	malfaçon	travaux	véhicule	somme	vente	preuve
preuve	acheteur	preuve	vente	contrat	immeuble	contrat
services	vente	contrat	prix	prix	contrat	somme
contrat	eau	dommages	garantie	travaux	promesse	être
payer	être	somme	bien	argent	représentant	droit

The terms of the TM were used to compute the cosine similarities in the manner as described above. Figures 1 and 2 show estimated kernel densities for the cosine similarities between indexed terms of the Gaius database and words created by both text mining approaches (tf-idf and TM). The overall area of each estimated density sums up to one. The further the density area is shifted to the right, the more often occur higher cosine similarities. We have chosen the number of words per topic and the number of topics assigned to each document so that (i) the standard deviation of all cosine similarities is the lowest (shown in Figure 1) and (ii) the average of the cosine similarities is the highest (Figure 2). With regard to the optimal number of topics, this resulted in a uniform value of three per document.

The figures show that the cosine similarities between words of the TM models and the Gaius indexing are higher than those between the words given by the tf-idf and the Gaius indexing. This visual impression is statistically confirmed by t-tests regarding the mean values. Additional t-tests on the differences of cosines similarities between LDA and the CTM indicate that the CTM approach yields higher cosine similarities or, put differently, CTM words are, on average, more similar with the human indexed database.

4 The table shows the five most probable words in descending order for the seven most prevalent topics of the sub-database *Contrats*, estimated by the CTM. The topic numbers have been changed for illustration purposes.

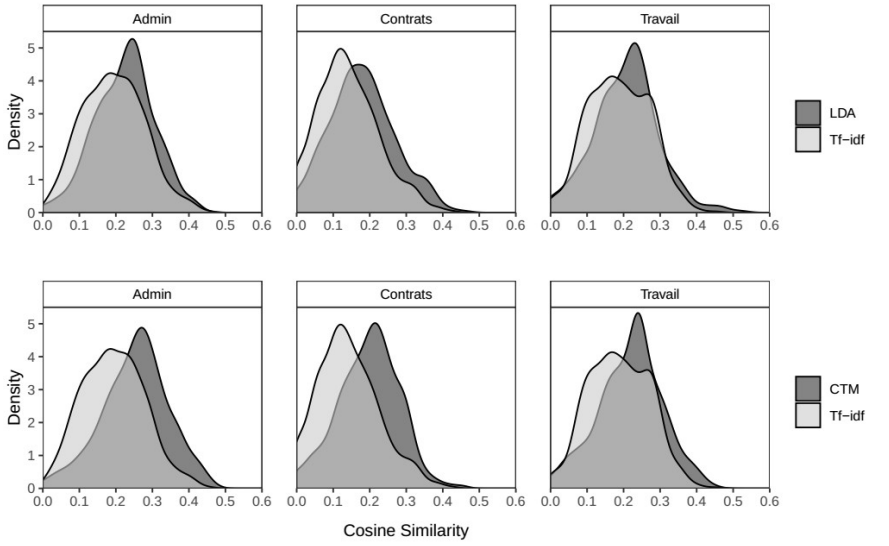

Figure 1: The figure shows estimated kernel densities for cosine similarities between words created by the tf-idf/topic models (LDA and CTM) and words by the Gaius database. The number of words for the tf-idf and the number of words/number of topics of the topic models have been chosen in such a way that the standard deviations of the cosine similarities have been minimal.

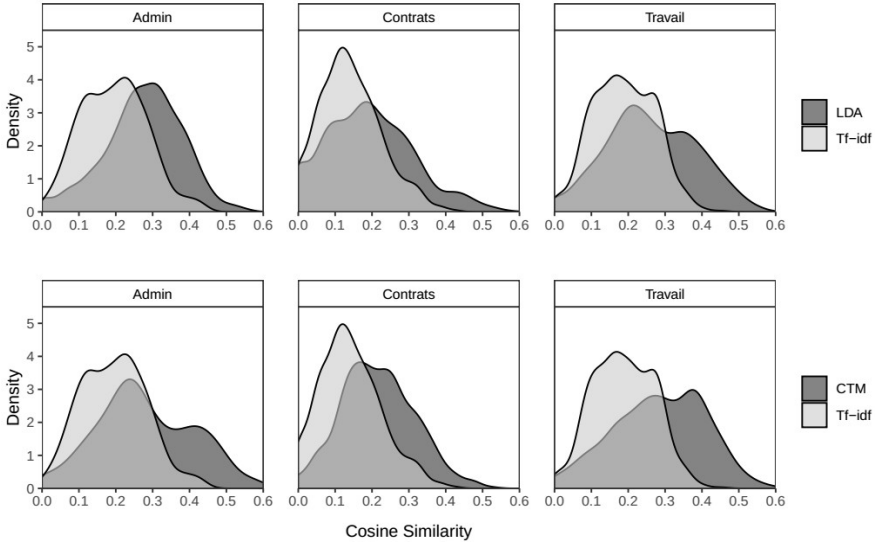

Figure 2: The figure shows estimated kernel densities for cosine similarities between words created by the tf-idf/topic models (LDA and CTM) and words by the Gaius database. The number of words for the tf-idf and the number of words/number of topics of the topic models have been chosen in such a way that the average of the cosine similarities is highest.

Apart from computing cosine similarities between TM (CTM and LDA) keywording and the Gaius indexing, we have also computed similarities between the TM terms and the intuitive, non-structured SOQUIJ indexing. For the sub-database *Admin*, the results indicate that the TM terms coincide, on average, more with Gaius than with SOQUIJ. For the sub-databases *Contrats* and *Travail*, however, the results are reverse: the terms from the TM have higher similarities with SOQUIJ than with Gaius. Yet, the differences are minor.

3.2 Which facets and which words drive similarities

Having shown that words generated by TM correlate with human subject indexing, we are also interested in which facets mostly correlate with the automated terms. To do so, we have computed, again, the cosine similarities on the basis of five of the six facets (for the rest as described above), but leaving out successively a different one of the six facets. Deleting the terms from facet 5 (Legal category) predominantly caused the largest drop in similarities, indicating that the terms from this facet are the most important drivers for similarities. We have verified that this finding is not caused by the fact that the facets contain different relative and overall numbers of terms. Investigating in further detail which terms from which facets are the most important drivers for the similarities remains a subject for future research, especially when considering using TM for semi-supervised indexing.

3.3 Qualitative results

Above we have used quantitative methods to investigate similarities between TM keywords and human (faceted) indexing. Another, qualitative question is whether TM is capable of providing the legally interesting core of the indexed decisions and how it performs in comparison with the Gaius and SOQUIJ indexing. The first impression, which is based on our own legal expertise and a selected sample of 25 decisions, is that the quality of the automatic TM keywording as an indicator for the decision content is – with fl from decision to decision – overall encouraging. This result is certainly subjective and needs empirical validation by, for example, conducting expert and user tests.

In all three sub-databases, the vast majority of decisions (*Contrats*: 0.87; *Admin*: 0.96; *Travail*: 0.89) were assigned to topics, at least one of which was relevant for the decision in question with a probability of > 0.5, and a considerable proportion of the decisions (*Contrats*: 0.36; *Admin*: 0.61; *Travail*: 0.43) even to topics with a relevance of > 0.9, while the probability for the other two topics was much lower (mostly between 0.0 - 0.2). For those decisions where even the probability of the most relevant topic was < 0.5, there was usually another topic of considerable relevance (> 0.3).

The legal information content of the (20 or 50) topics of the three tested sub-databases was mixed. There were topics whose most probable five or ten terms associated certain facts with certain types of legal issues (certain types of cases) and others which had a rather low distinctiveness. An example of a more meaningful keywording from the *Contrats* sub-database is topic 2, shown in extracts in table 1 above. The word *être* has passed through our filter, because it was not included in our applied stop word dictionary. Topic 7 from the same sub- database, also shown in extracts in table 1, is an example of a rather meaningless topic with a large portion of irrelevant words (*être*, *droit*, *peut* and *comme*). For our qualitative analysis, we have limited ourselves to those decisions

where at least one topic, with usable distinctiveness in the aforementioned sense, has been assigned to with a minimum probability of 0.3. In the future, it should be examined whether it is useful to manually filter the list of topics in advance. However, as our analysis has revealed, it is possible that relevant terms may originate from topics which are, taken as a whole, of little significance.

In any case, it can be determined that the keywords from those topics that were assigned to the respective decision with a high probability (> 0.8) were predominantly relevant to describe the broad outline of the decision (*i.e.,* the type of contract, the concern). The terms, however, did not always hit exactly those points that were of particular legal interest (*i.e.* controversial) in the decision in question. Admittedly, that would be a high standard, which neither the Gaius indexing nor the SOQUIJ indexing have met consistently. In sum, comparing the three approaches of keywording, the Gaius indexing predominantly met best the content of the decision, which could be related to the facet scheme that pushes the indexers' view in the direction of the legally decisive dimensions. However, Gaius does not always describe the legal core of the decision and occasionally even points in the wrong direction. This may be the consequence of the standardisation brought about by the use of a controlled vocabulary. Occasionally, the SOQUIJ indexing has been more accurate, possibly because it is not bounded by a thesaurus. TM keywording also classifies well at times as illustrated by the following example on the decision *Robert c. Bergeron* (2013 QCCQ 5859) from the *Contrats* subdatabase. In that case, the buyer of a 24-year-old wooden house (plaintiff) is suing the seller asking for a purchase price reduction due to the defect of water seeping through the basement fl or. Prior to the purchase, the plaintiff had inspected the property for about only two hours without consulting an expert. When he discovered water entering next spring, he removed one of the insulation boards to find that the cement was irregularly shaped. The court dismissed the claim holding that it was not a *vice caché*. The buyer could have identified the defect himself by careful inspection in accordance with his duty under Art. 1726 of the Civil Code of Québec (CCQ), considering the age of the house and especially since the promise of sale (*promesse d'achat*) had even expressly referred to the possibility of water ingress during spring. What is legally interesting about this decision is the scope of the buyer's duty of inspection under Art. 1726 CCQ.

Our algorithm has assigned to this decision - in this order - the topics 2, 4 and 6 with the probabilities 0.998, 0.000 and 0.000 respectively, leading to a cosine similarity to the Gaius index of 0.479. The first five terms of the topics are as shown in table 1 above. Since these terms are unigrams, but legal concepts, especially in the French language, often consist of several terms, the interpretation of the topic terms requires a certain legal expertise to recognize related terms (*e.g.* from *salariale*, *équité* to *équité salariale*). If one interprets the above terms of topic 2 in this way, and cleansing them of legally irrelevant words and redundancies as well as sorting them in a meaningful way, they read as follows: *vente, immeuble, eau, malfa¸con, [vice] caché*.

The rough context of the case is thus already drawn. The aspect of the seller's responsibility is included in the term *(vice) caché* and reinforced by topic 4 (*garantie*; (diminution du) *prix*). However, topic 4 also misleads with the terms *véhicule* and *moteur.* Yet, the algorithm puts a probability close to zero for topic 4 (compared to 0.998 for topic 2). The probabilities must therefore be taken into account when interpreting the TM indexing. The special aspect of the buyer's obligation to inspect, which is part of

the legal regime of *vice caché*, is not directly expressed in the TM keywording; the Gaius indexing is more explicit on this point (the term *inspection* in the Action facet), but without being unambiguous. This is also true, albeit in a diff t way, regarding the SOQUIJ indexing (*vice apparent* as opposed to *vice caché*). Nevertheless, neither of the two indexings, Gaius and SOQUIJ, grasps fully the legal focus of the case.

The sub-database *Contrats* contains four further decisions to which the TM algorithm has assigned the same topic combination 2, 4 and 6 in the same sequence ordered by the topic probabilities. The four documents deal with the seller's liability for defects, and interestingly, the legal focus in three of these decisions is also on the buyer's duty to inspect the goods, identical to the example outlined above. Only in one decision, to which our algorithm assigns topic 2 with a significantly lower probability (0.317) than with the others (0.793, 0.993, 0.646), this aspect is not the court's main focus. This example strengthens the impression that topic interpretation must take probabilities into account; it also suggests that TM keywording might be suitable for recognizing legally similar decisions. In a corresponding manner, TM enable to use quantitative metrics such as the Hellinger distance and the Kullback-Leibler divergence to find similar documents.

4.0 Conclusion and Outlook

We have shown that similarities exist between human structured subject indexing and automatically generated terms based on methods from natural language processing. The cosine similarities between human indexing and automatically generated keywords are generally higher when using the words of topic models (TM) instead of the words generated by a frequency–inverse document frequency approach. Our quantitative and qualitative results indicate that TM can be (at least) a useful tool to support human indexing in a semi-automated approach. For example, TM can provide clustered terms that can be used, in addition to a standardized vocabulary, for indexing completion. In addition, TM can useful to identify similar documents.

We propose that future research on semi-automated indexing could try to combine TM with the multifaceted approach in two directions: first one can try to optimize TM keywording by assigning the topic terms to exogenously defined facets. This approach could also be beneficial to identify similar documents. When creating a faceted vocabulary, however, it must be ensured that the majority of the faceted terms are contained in the full text and that each word of the vocabulary is only represented within one facet. Second, it would be interesting to check whether the quality of the topics can be improved when using a vocabulary of predefined facets for the TM estimation.

References

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. "Quanteda: An R Package for the Quantitative Analysis of Textual Data." *Journal of Open Source Software* 3, no.30: 774.

Blei, David M. and John D. Lafferty. 2007. "A Correlated Topic Model Of Science." *The Annals of Applied Statistics* 1, no.1: 17–35.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993-1022.

Chater, Nick and Christopher D. Manning. 2006. "Probabilistic Models of Language Processing And Acquisition." *Trends in Cognitive Sciences* 10, no.7: 335–344.

Cumyn, Michelle, Michèle Hudon, Sabine Mas, and Günter Reiner. 2018. “Towards a New Approach to Legal Indexing Using Facets.” In *Recent Trends and Future Technology in Applied Intelligence. IEA/AIE 2018*, edited by Malek Mouhoub, Samira Sadaoui, Otmane Ait Mohamed, and Moonis Ali. Lecture Notes in Computer Science 10868. Cham: Springer, 881–888

Cumyn, Michelle, Günter Reiner, Sabine Mas, and David Lesieur. 2019. “Legal Knowledge Representation Using a Faceted Scheme.” In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*. New York: Association for Computing Machinery, 258–259.

George, Clint P., Sahil Puri, Daisy Zhe Wang, Joseph N. Wilson, and William F. Hamilton. 2014. “SMART Electronic Legal Discovery Via Topic Modeling.” In *Proceedings of the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference*, edited by Edited by William Eberle and Chutima Boonthum-Deneck. Palo Alto, California: The AAAI Press, 327–332.

Gross, Tina, Arlene G. Taylor, and Daniel N. Joudrey. 2015. “Still a Lot to Lose: The Role of Controlled Vocabulary in Keyword Searching.” *Cataloging & Classification Quarterly* 53, no.1: 1–39.

Grün, Bettina and Kurt Hornik. 2011. “topicmodels: An R Package for Fitting Topic Models.” *Journal of Statistical Software* 40, no.13: 1–30.

Huang, Anna. 2008. “Similarity Measures for Text Document Clustering.” In *Proceedings Of The Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand* 4, pages 9–56.

Livermore, Michael, Allen Riddell, and Daniel Rockmore. 2017. “The Supreme Court And The Judicial Genre.” *Arizona Law Review* 59: 837-901.

Reiner, Günter, Michelle Cumyn, Michèle Hudon, and Sabine Mas. 2019. “Designing a Database to Assist Legal Thinking: A New Approach to Indexing Using Facets.” *In Internet of Things : proceedings of the 22nd International Legal Informatics Symposium: IRIS 2019*, edited by Erich Schweighofer, Franz Kummer, and Ahti Saarenpää. http://hdl.handle.net/20.500.11794/34753

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. 2019. “stm: An R Package for Structural Topic Models.” *Journal of Statistical Software* 91, no.2: 1–40.

ADVANCES IN KNOWLEDGE ORGANIZATION VOL. 17

INTERNATIONAL SOCIETY FOR KNOWLEDGE ORGANIZATION [ISKO]

Knowledge Organization at the Interface

Proceedings of the
Sixteenth International ISKO Conference, 2020
Aalborg, Denmark

Organized by
International Society for Knowledge Organization (ISKO),
Marianne Lykke, Tanja Svarre,
Mette Skov and Daniel Martínez-Ávila

Edited by
Marianne Lykke
Tanja Svarre
Mette Skov
Daniel Martínez-Ávila

Ergon

Knowledge Organization at the Interface

Advances in Knowledge Organization, Vol. 17 (2020)

Knowledge Organization at the Interface

Proceedings of the Sixteenth International ISKO Conference, 2020 Aalborg, Denmark

Organized by
International Society for Knowledge Organization (ISKO),
Marianne Lykke, Tanja Svarre,
Mette Skov and Daniel Martínez-Ávila

Edited by

Marianne Lykke
Tanja Svarre
Mette Skov
Daniel Martínez-Ávila

ERGON VERLAG

The Deutsche Nationalbibliothek lists this publication in the
Deutsche Nationalbibliografie; detailed bibliographic data
are available on the Internet at http://dnb.d-nb.de.

www.ergon-verlag.de

ISBN 978-3-95650-775-5 (Print)
ISBN 978-3-95650-776-2 (ePDF)
ISSN 0938-5495

Table of Contents