Recommender System for Researchers and Research Proposals within UM[1]

S. Chatzopoulou, A. Grigoriu, L. Hufkens, O. Manhar, and M. Stokvis

June 29, 2018

Abstract

Multidisciplinary research is becoming increasingly popular due to its distinctive characteristics compared to research focused on one area. Combining research from separate fields can lead to high impact research, as a high level of creativity can positively impact the project. Finding the appropriate researchers for a multidisciplinary project can be a difficult task in comparison with finding the researchers in the same field. The researcher may lack the necessary information for selecting persons outside of his/hers field. A common solution to this problem is selecting collaborators from the researcher's network, but this doesn't necessarily lead to efficient collaborations. The present paper presents a method which uses LDA topic modeling to extract topics from the title of the collaboration project. Based on these, it can recommend appropriate collaborators by six proposed algorithms. Especially algorithm 1 shows promising results, which prefers authors with more diversity in their topics over authors that have high probabilities in just a few topics. By increasing the size of the training set the results are expected to improve and become more robust.

1 Introduction

Nowadays, research is becoming more and more advanced due to rapid developments in all areas of expertise. Research projects are increasingly often multidisciplinary research, because it leads to creativity and high impact research. It takes into account the viewpoints of multiple fields, which makes it relevant to a broader area and therefore has a higher chance to receive funding[2].

One of the main challenges remains finding those researchers that would like to collaborate and that lead to effective research[2, 3]. Due to the multidisciplinary approach, researchers are bound to leave their area of research in order to find collaborators. As they may enter less known areas or even unknown ones, the search for collaborators can become increasingly difficult. Most often, collaborators are found by the network of supervisors and co-authors, causing the network not to be extended to its full potential and authors sticking to their known hub. This can cause a decrease research quality, as most probably the researchers won't collaborate with the needed experts in the fields.

1.1 Finding collaborators automatically

By automating the process of finding collaborations, this problem can be solved efficiently. New collaborations can be proposed by a recommender system, and ideally, can be applied to Maastricht University. With such a system, it saves the researchers the time that they otherwise would have spent on finding collaborators by hand or waiting for collaborators to come to them, while now they can focus on doing research instead. This leads to the following research questions:

1. Can a recommender system be built by publicly available data?
2. Can a recommender system be automatically updated when new data is available?
3. Can a recommender system suggest useful research collaborations for UM, despite the limited amount of data available from the UM?

1.2 Outline

The remainder of the paper is structured as follows. In the next section the data set is discussed. In section 3 the configurations of the methods are given, followed by the experiments and results in section 4. Section 5 gives a conclusion of the paper, with the future work in section 6 and the social impact in section 7.

2 Dataset

An important part of building a recommender system is the dataset used for training and testing. Maastricht

[1] This paper was prepared in partial fulfillment of the requirements for the Degree of Master of Science in Data Science for Decision Making and the Degree of Master of Science in Artificial Intelligence, Maastricht University, supervisor: Michel Dumontier.

University provides system PURE, which contains the public profiles of the university's researchers. Because this data might not be enough for both a proper training and test set, an additional existing database, Scopus, has been used for training purposes.

PURE is a current research information system with which research management and reporting is being done[4]. In the PURE system, researchers of Maastricht University may show their research work. Permission of the library services to use the system has been received, after the data has been obtained as an XML file from the library contact person. Currently, there are 62715 research outputs in the PURE system (24 April 2018). This information can then be used to test our system.

For both datasets, only English, published, non-confidential scientific papers and contributions to journals or conferences have been used to make the dataset consistent. The Scopus data has been reduced to only contain abstracts. The abstracts will be used for topic modelling, which is further explained in section 3.1. The PURE dataset consists of two parts: 1) publications, and 2) authors. The publications dataset has been reduced to only contain the research titles, and the associated authors of the research. Several other fields have been stored for future work, such as number of citations in Scopus, number of pages, and publication date. The authors simply contain names. As research is linked to the authors, authors are recognised regardless of different versions of spelling, including middle names and abbreviations. The PURE dataset contains 30633 publications, with a total number of authors of 5813.

3 Methods

3.1 Building the model

The recommendation model will first extract topics from the abstracts from the Scopus data using the Latent Dirichlet Allocation (LDA), after which it is calculated how relevant those topics are to the authors from the PURE dataset. The relevance of topics for researchers is based on how much those topics are relevant to the documents that those researchers have contributed to. A researcher searches for future collaborators by entering the proposed research title, from which the relevant topics are extracted. Then, a list of researchers is retrieved that is ranked according to a sorting algorithm based on the relevance of the topics that the researcher and research title have in common. Thus, there are two main parts to the recommender system, that will both be further explained in this chapter: topic modeling, and ranking the possible collaborators.

Topic modeling

Topic modeling has been performed by Latent Dirichlet Allocation (LDA), a statistical model that is often used for topic modeling in text mining. It views a document as a collection of topics, and assumes that each word in the document is an attribute of those topics. Additionally, it uses a Dirichlet topic distribution, hence the name of the model. The Dirichlet distribution is a probability distribution just as the normal distribution, but instead of sampling from real numbers it samples from probability simplex, which is X numbers that add up to 1, where X is the number of categories, or topics[1]. Instead of a word, a topic can be seen as a collection of words that contain the information of a document. The pseudocode of the model can be found in Algorithm 1.

```
input  : list of documents $d$,
         number of topics $t$
output: list of probabilities that topic $t'$
         generate word $w$ for all topics $T$
for $i \leftarrow 1$ to $d$ do
    for $j \leftarrow 1$ to $\#words \in i$ do
        randomly assign $j$ to one of $t$ topics
    end
end
while steady state has not been reached do
    for $i \leftarrow 1$ to $d$ do
        for $j \leftarrow 1$ to $\#words \in i$ do
            compute p(topic t | i)
            compute p(j | topic t)
        end
        reassign $j$ to topic $t'$ with probability
          p(topic $t'$ | i) * p(j | topic $t'$)
    end
end
```

Algorithm 1: Latent Dirichlet Allocation

The documents in the algorithm are the abstract from the Scopus dataset, which first need to be preprocessed before they can be used for proper text mining. First, punctuations and other non-alphabetic characters are removed from the abstracts. Then, the texts are tokenized, after which stop words are deleted, as stop words are not significant or useful for topic detection. The remaining tokenized words are then stemmed and used to create a dictionary which is fed to the LDA model. LDA assumes a sparse Dirichlet topic distribution, which is based on normal text. In general, research titles are very specific and most words are important and refer to topics, while running texts like abstracts contain more details about research itself, and therefore more possibly useful topics, and have a different distribution of important words. Therefore, the LDA is only trained on the abstracts of the Scopus data. The model considers 1300 topics. A larger number of topics (1500, 2000, 3500) resulted in the inclusion of unimportant words and therefore worse results, and cause memory issues, as the memory requirements increase exponentially with the number of topics. A smaller number of topics (800, 1000) resulted in similar results to having 1300 topics, however, the results with 1300 topics are more consistent and therefore are preferred.

Retrieving researcher's skillsets

Once the topic model has been trained, the data from PURE can be used to retrieve the topics from past research at the University of Maastricht. The recommendation model examines the topic vectors for each document and adds the probabilities of topics and the probabilities of the corresponding words to the author's topic profile. After adding all the probabilities, the probabilities are normalized to be able to use them for ranking.

Ranking suggested collaborators

Once skill sets have been retrieved, a ranked list has to be created after a query will be entered. The research fields of researchers have not been included, as researchers should objectively be recommended as collaborator based on their experience instead of their environment (department). For ranking, the research field of researchers have not been included as they may skew the results and may leave out experts that may work in different departments, regardless of their objectively defined skill levels. When using the LDA model, topics are not defined within research fields which automatically allows researchers to be present in multiple distinct topic clusters. Depending on the algorithm used, the model will reward researchers with multiple skill sets relevant to the proposed research project.

In order to explore various recommendation options, different strategies of ranking the collaborators have been implemented. Depending on the project or on the user's preferences, some might be more fitted then others for the current situation. The user can choose between the types, in order to receive a more tailored recommendation. The types of implemented algorithms are described in the following paragraphs.

1. Algorithm 1 prefers authors with more diversity in their topics over the specialized authors that have high probabilities in just a few topics. This might be helpful in a project where a smaller, limited amount of researchers is needed but their topics need all the necessary topics for the future paper.

   Algorithm 1

   (a) For each topic in the query, find all related authors
   (b) Remove duplicate authors
   (c) Create a table-like list of all found authors and their relevant topic probabilities
   (d) Sort the list using every author's highest probability (descending)
   (e) Sort by smallest distance between order of topics in query and skill set of researcher (ascending)
   (f) Sort by the number of topics per author (descending)
   (g) Take the top X entries for each topic count, except 1
   (h) For topic count of 1, first sort by authors' highest probability (descending) before taking the top X entries
   (i) Return the merged list of all top X entries

2. In order to offer the alternative option, a second algorithm was created. Algorithm 2 performs similar steps to algorithm 1, except for dealing with topic count sorting. It favors specialized authors that have high probabilities in just a few topics over authors with more diversity and thus lower probabilities.

   Algorithm 2

   (a) For each topic in the query, find all related authors
   (b) Remove duplicate authors
   (c) Create a table-like list of all found authors and their relevant topic probabilities
   (d) Sort the list using every author's highest probability (descending)
   (e) Sort by smallest distance between order of topics in query and skill set of researcher (ascending)
   (f) Sort by the number of topics per author (ascending)
   (g) Take the top X entries for each topic count, except 1
   (h) For topic count of 1, first sort by authors' highest probability (descending) before taking the top X entries
   (i) Return the merged list of all top X entries

3. The following algorithm makes use of a voting system to determine where the ranks of sublist entries go in the merged list. It heavily benefits authors with high probabilities, which usually means that the authors are very specialized. It does not tend to recommend more broadly oriented authors as a result of this.

   Algorithm 3

   (a) For each topic in the query, find all related authors
   (b) Remove duplicate authors

(c) Create a table-like list of all found authors and their relevant topic probabilities

(d) For each topic, create a sublist of the relevant authors

(e) Sort each sublist by their topic's probability (descending) and take the top X entries.

(f) Record the ranks of all authors on each sublist

(g) The mean for each author's ranks becomes the new rank

(h) Sort the merged list by the new ranks

(i) If two or more authors have the same rank, order them by their highest probability (descending)

4. This algorithm is a variation of algorithm 3. It has the same characteristics (heavily favors authors with high probabilities), but calculates the new ranks differently, which is more visible when a higher number topics is involved.

Algorithm 4

(a) For each topic in the query, find all related authors

(b) Remove duplicate authors

(c) Create a table-like list of all found authors and their relevant topic probabilities

(d) For each topic, create a sublist of the relevant authors

(e) Sort each sublist by their topic's probability (descending) and take the top X entries.

(f) Record the ranks of all authors on each sublist

(g) The median for each author's ranks becomes the new rank

(h) Sort the merged list by the new ranks

(i) If two or more authors have the same rank, order them by their highest probability (descending)

5. In the following method, the suggested collaborators are recommended based on their overall topic weight. It tends to offer a higher rank to experienced collaborators with multiple topics covered from the query.

Algorithm 5

(a) For each topic in the query, get a list of related authors

(b) Create an array containing all the authors received from step 1

(c) Sort the array by their topic weights

(d) For each author, add the number of topics covered to their topic weights

(e) For above step, in order to make sure the weight is not highly skewed by topic counts, taking a log of the count makes sure the additions are steady

$$Weight_{overall} = Weight_{accumulated} + log(N_{topics}) \tag{1}$$

(f) Sort the array based on the overall estimated weight

(g) Return top required amount of authors as potential collaborators

6. In Algorithm 6, the suggested collaborators are recommended based on their affinity to a topic which is important from the query. This method prefers authors with higher probabilities in topics they share with the query.

Algorithm 6

(a) For each topic in the query, get a list of related authors in the decreasing order of topic's probabilistic value

(b) Create an array of authors for each topic and combine them

$$Authors = [[Authors_{topic1}], [Authors_{topic2}]....] \tag{2}$$

(c) Sort the array(from step 2) of arrays based on topic weights

(d) For each topic, select authors while preferring the authors with higher weights

(e) Select higher number of authors for topics with higher weight (can be done by taking a ratio based on the weight)

$$N_{topic1}^{SelectedAuthors} = Weight_{topic1} * N_{topic1}^{Authors} \tag{3}$$

(f) Repeat step 5 until authors have been selected for each topic

(g) Return all the collaborators selected from above step as a list

Figure 4 shows the flow-chart of the system.

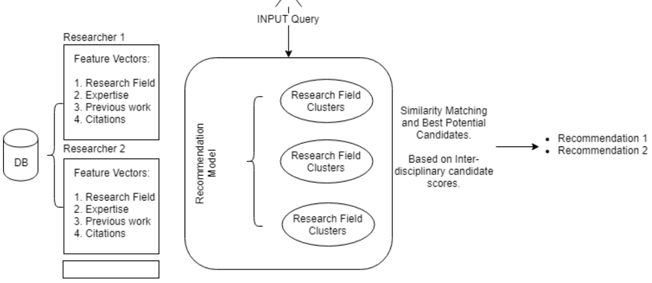


Figure 1: Flow-chart of the proposed recommender system.

3.2 Tools and Technologies

The latest edition of Python(3) provides a very well defined structure and platform to easily implement data models and apply learning algorithms on them. The primary advantages of having this as a development language are its simplicity and efficiency, comparatively smaller compile and run-time and the number of active open source libraries (e.g., sci-kit learn, NumPy) [5] which are specifically designed for solving Scientific Computing and Machine Learning related problems.

3.3 Testing the performance of the models and determining the best approach

The recommender system should be able to produce a single ranked list of the most suited researchers, based on the project proposal title used as input. The results of this recommender system are dependent on the following components:

- identifying the field(s) a research proposal belongs to,
- identifying the field(s) researchers are part of, and
- ranking prospective partners by suitability for participation in the project.

4 Experiments & results

4.1 Topic Model

The LDA topic model is a model that has no right or wrong answers, and therefore testing it is not trivial. Test cases are hand-picked, and results are reviewed manually. Experiments for the topic model are run to find topics from research titles, as well as researchers to create user profiles. Part of the research titles that have been tested can be found in Appendix A. For each research title in this graph, a list of topics in descending order of relevance is returned. A topic consists out of a set of words, in descending order of relevance. From the first query it is very interesting to see that the query contains the word Amsterdam, and the first returned topic speaks of European and city. Art and creativity are too represented in the topics. Although the returned topic including carbon, dioxide and concentration may be relevant to the urban renewal in general, it may not be relevant to this research. Additionally, the meaning of the word slow in this research is dependent on art, as slow art is a concept on its own, while the topic model connects slow to a topic regarding memory.

Another interesting feature of the LDA becomes visible in the fourth query, where the word ”data” occurs in both topics. This shows that words can occur more often due to having multiple meanings or simply because they are relevant to more than one topic. With a larger training set, the word ”slow” from the first query might have been recognized to be related to ”art”. As seen in the second query, it also recognizes words that are similar, such as ”organization” and ”organizational”. The differences in British English and American English can be discarded when using preprocessing techniques when training the data, when running the queries, or both. The fourth query shows that the topics regarding machine learning and big data have been identified, but it misses the scope of the word ”beyond” from the query.

Semantics are, however, a research area within the field of text mining[], and using semantic analysis in this project might be irrelevant. A more simple way of retrieving more relevant topics for researchers using the system would be to rephrase the proposed query to a more specific title: what is beyond machine learning?

A sample of the results of creating a research profile by the topic model can be viewed in Appendix B. For each researcher, the five most relevant topics are presented in decreasing order, with their corresponding relevant words in decreasing order. When looking at the first query, Cavill R., it becomes clear this researcher is focused on cells, metabolisms and molecular mechanisms. The fourth topic implies that this researcher uses time series, a common approach that shows change over time, which is common within biology. This is close to her actual research interests, as she works in the field of bioinformatics. Besides time series, however, it does not provide insights into what specific methods she uses in her research. The second researcher has to do with international relations, politics and law. It is therefore clear that this person’s expertise lies in the field of international law. According to the PURE data, this person’s expertise also includes human rights and jurisprudence.

Again, not all expertise and details become clear from the first five topics from the user profiles created by LDA. Nonetheless, it shows an adequately first portrait of the user’s profile, which might be useful as a first impression for users looking for collaborators. The last query shows Dumontier, M., but it is not immediately clear what the particular area of interest is of this researcher. According to the topics, he has focused on ontologies or ontology engineering, semantics, drugs, and some performed procedures possibly relating animals. The sparsity of this topic model might be due to the large amount of research publications (nearly a hundred), causing many topics to have some relevance to the researcher. Having more papers could be an advantage when researchers are open for collaborations.

4.2 Ranking algorithms

Despite the fact that LDA does not include all details of researchers in the first five topics, less relevant or more specific topics will also be taken into account by most of the proposed algorithms. Hence, it is to be expected that the results will also be sufficient for more specific queries. The results of the sample queries of collaborator suggestions can be found in Appendix C. When looking at most queries, algorithm 3 and 4 do not provide good suggestions, as most of the collaborators are not even considered in the list of potentially relevant researchers. These two algorithms are therefore not taken into account from now. When comparing algorithm 1 and 2, algorithm 2 does not provide the researchers from the publication in a better rank in any case. Thus, algorithm 2 will also be discarded. In general, algorithm 1 seems to be the most promising candidate, as it returns most of the researchers, and when it returns researchers their rank is quite high. A high rank means that collaborators would be suggested with the best score - according to the algorithm they would be more fit to join the research team.

It is clear that algorithm 6 does not consider nearly as many relevant researchers to be a potential collaborator, and when it does, the returned ranks are quite low, meaning that when researchers receive a list of suggested collaborators, they will most likely not consider this person. However, for example in query 12, the algorithm performs best with a returned rank of 12, where algorithm 5 returns 116 and algorithm 1 fails to return this person at all. This might be because of the difference in algorithms: it may suggest that preferring researchers that have top knowledge about some topic over researchers that have a little knowledge about everything is usually not good, but in some cases it is wanted and useful. However, because algorithm 6 is so unstable the potential does not seem to be high. When looking closer at the differences between algorithm 1 and 5, whenever algorithm 1 does not find a researcher relevant, algorithm 5 returns a large rank number. This may be due to the implementation of the algorithm, as algorithm 1 takes the entire list of researchers and then starts sorting. The researchers will still be relevant to the system, but will be placed late in the list, resulting in a low rank.

4.3 Interdisciplinary queries on the recommender system

When combining queries from different research areas, visible in query 21 - 24, the recommendation system may not always return the desired mix of authors. This, however, was expected due to several reasons. The research titles that are combined may have nothing in common, causing the topic model to come up with a poor topic model of the combined query. Additionally, the errors of the individual queries will most likely add up, and thus, the retrieved ranks are not to be expected. In actual multidisciplinary research the topics will have more than nothing in common, and therefore better results are expected in actual queries. A larger training set for the LDA topic model will also contribute to improved results for research, including multidisciplinary research.

5 Conclusion

The recommender system shows a lot of potential to facilitate the lives of many researchers searching for collaborators. It aims to encourage researchers to collaborate beyond their social network connections, suggesting collaborations based on the requirements of the project. Researchers are encouraged to provide correct and complete information about their previous work for the best results. This information is also of help for those who wish to form collaborations without assistance from the system, as they too make use of it to find prospective collaborators.

The topics that are retrieved from the proposed research titles, or queries, are highly dependent on the training data. If the topic model would have been trained on the PURE data, it would likely overfit the model so that future research would not be recognized. Moreover, the research titles alone are not enough to create a topic model which gives meaningful results. The model is dependent on the training data, which also means that it does not work well with new research topics which may contain new words that have not been covered by the training data (yet). Nonetheless, the topic model gives very meaningful results.

One must also keep in mind that in real use cases, the search queries are usually generic and simpler rather than complicated. As the results show the recommender system suggests authors with a background in those topics or a mix of authors who would cover all the topics as provided by the topic model. The various discussed algorithms to rank the authors give a good insight into what can be expected from a recommender system of this type. The selection of the eventual algorithm would completely depend on the end users, e.g. whether the user wants authors with multiple topics coverage or a collection of authors with higher weights in individual topics. Among all six implemented algorithms, algorithm 1 seems to be the most promising.

Results have also shown that the more information about the author's publication the system has, the more accurately it groups them to specific topics. It therefore encourages researchers to provide correct and complete information about their previous work, which is also of help for those who wish to form collaborations without assistance from the system. It is also expected that the results will significantly improve when the Latent Dirichlet Allocation topic model will be trained on a larger dataset, which was not possible for this research due to the limited amount of publications that could be downloaded with the used API.

6 Future work

In order to improve the recommendation, a number of features available from the papers could be added. The high number of Scopus citations could help in retrieving the most relevant authors for a certain topic. In addition, each paper has a list of authors in which each individual has a rank. The current system considers the authors as being equally important for the paper. However this might not represent the reality of their importance, as the rank of an author in the paper is relevant. The first author could have contributed more to the paper compared to the one ranked as last, therefore has more knowledge about the topic. The paper's number of pages could also improve the recommendation. Researchers that have papers with the same (or similar) number of pages as the one that is querying the system might have more in common with the user. This could mean that their style of writing is similar, therefore they could work together.

As the topic model extracts every topic from the title used as a query, the user might be interested in seeing these topics and selecting some of them. Using a selection step might help the recommendation process as the user could select only the topics believed to be most relevant for the title from the extracted ones.

Using the departments researchers work in according to PURE to promote multidisciplinary could improve the performance of the system. However, this could have the downside that collaborator suggestions are not objective anymore and a black-box approach might be better. Authors should be recommended based on their skill levels and not based on in which department they work. If the latter is applied, the quality of research and research collaborations could decrease, or research will not include multiple disciplines.

One of the key components to be implemented in the recommendation model is to take the recent trends in researcher's latest work into consideration. It will also play a factor in determining the measure of how fit they are as a collaborator for a certain project proposal. However, that would depend on how far back in time the researcher's profile data goes.

References

[1] (2017). Dirichlet distribution. `https://www.quora.com/What-is-an-intuitive-explanation-of-the-Dirichlet-distribution`.

[2] (2012). Science europe position statement, horizon 2020: Excellence counts. *International Journal of Educational Research and Information Science*.

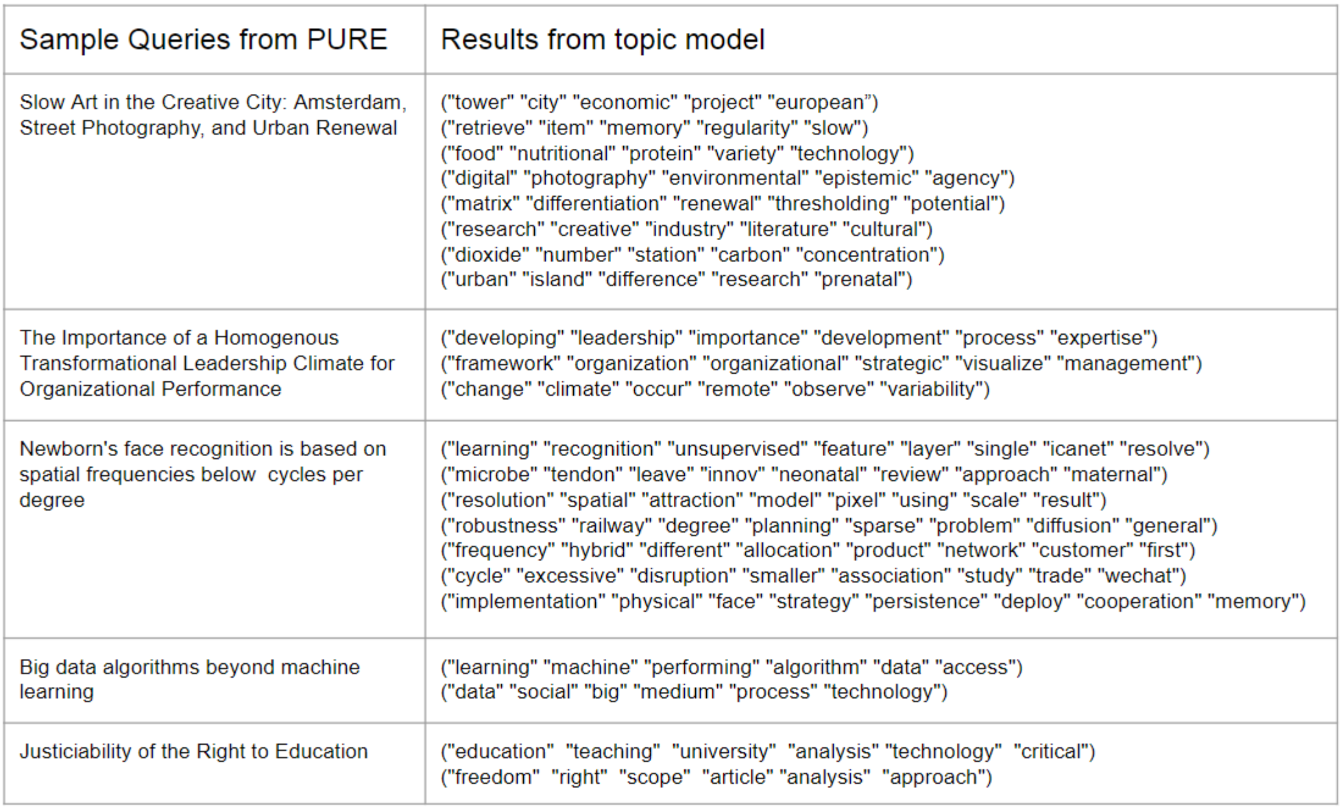
[3] (2015). Creativity and innovation in research: Scope for multidisciplinary research. *International Journal of Educational Research and Information Science*, pp. 54–60.

[4] `https://library.maastrichtuniversity.nl/cris-support/about/`.

[5] `http://scikit-learn.org/stable/index.html`.

Appendices

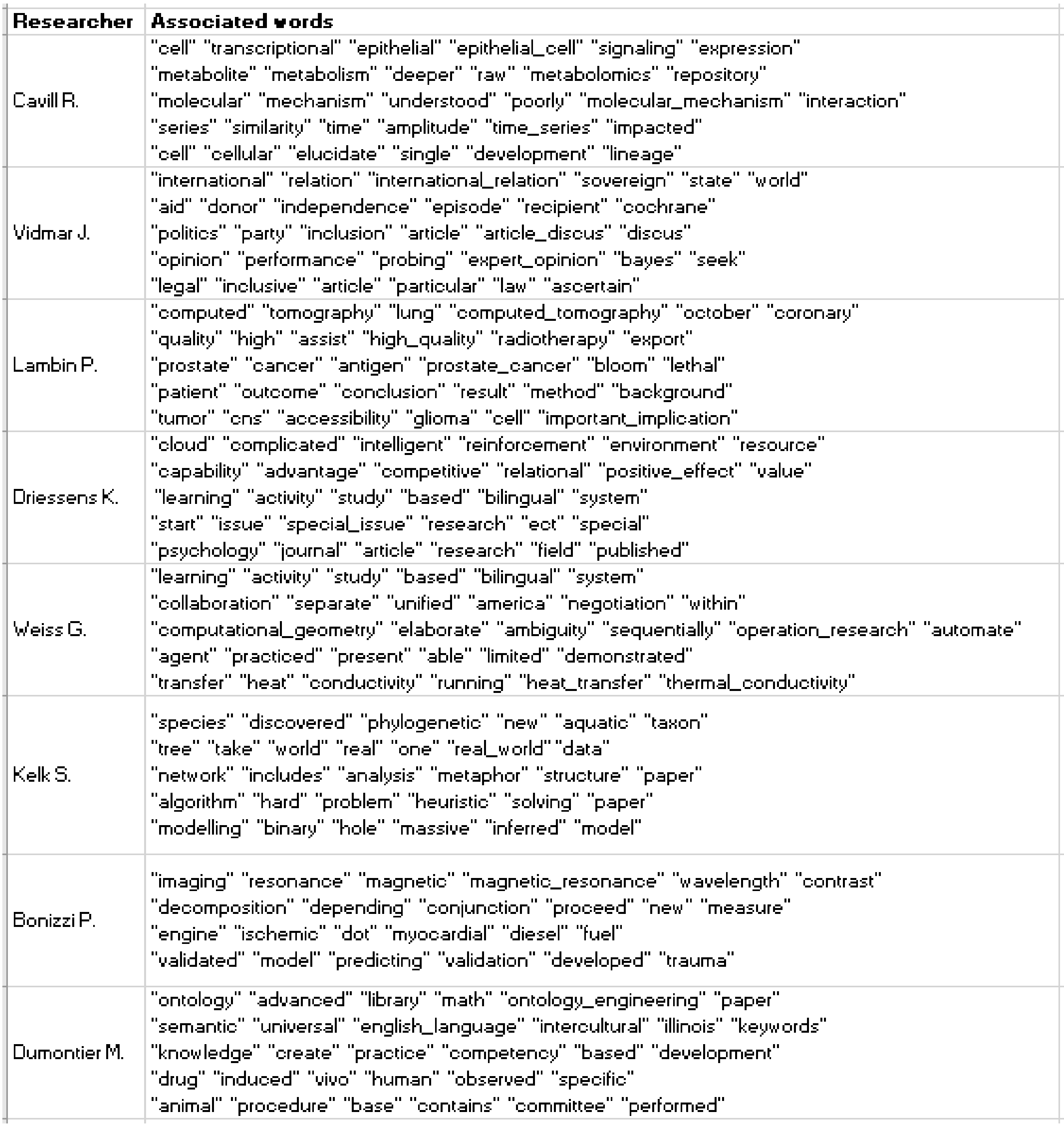
A Sample research titles from PURE and their results from the topic model



Sample Queries from PURE	Results from topic model
Slow Art in the Creative City: Amsterdam, Street Photography, and Urban Renewal	("tower" "city" "economic" "project" "european”) ("retrieve" "item" "memory" "regularity" "slow") ("food" "nutritional" "protein" "variety" "technology") ("digital" "photography" "environmental" "epistemic" "agency") ("matrix" "differentiation" "renewal" "thresholding" "potential") ("research" "creative" "industry" "literature" "cultural") ("dioxide" "number" "station" "carbon" "concentration") ("urban" "island" "difference" "research" "prenatal")
The Importance of a Homogenous Transformational Leadership Climate for Organizational Performance	("developing" "leadership" "importance" "development" "process" "expertise") ("framework" "organization" "organizational" "strategic" "visualize" "management") ("change" "climate" "occur" "remote" "observe" "variability")
Newborn's face recognition is based on spatial frequencies below cycles per degree	("learning" "recognition" "unsupervised" "feature" "layer" "single" "icanet" "resolve") ("microbe" "tendon" "leave" "innov" "neonatal" "review" "approach" "maternal") ("resolution" "spatial" "attraction" "model" "pixel" "using" "scale" "result") ("robustness" "railway" "degree" "planning" "sparse" "problem" "diffusion" "general") ("frequency" "hybrid" "different" "allocation" "product" "network" "customer" "first") ("cycle" "excessive" "disruption" "smaller" "association" "study" "trade" "wechat") ("implementation" "physical" "face" "strategy" "persistence" "deploy" "cooperation" "memory")
Big data algorithms beyond machine learning	("learning" "machine" "performing" "algorithm" "data" "access") ("data" "social" "big" "medium" "process" "technology")
Justiciability of the Right to Education	("education" "teaching" "university" "analysis" "technology" "critical") ("freedom" "right" "scope" "article" "analysis" "approach")

Figure 2: Sample queries from the PURE dataset, with the corresponding topics according to the LDA model. Each row in the right column contains a set of words, together forming a topic.

B Sample researchers from PURE and their results from the topic model

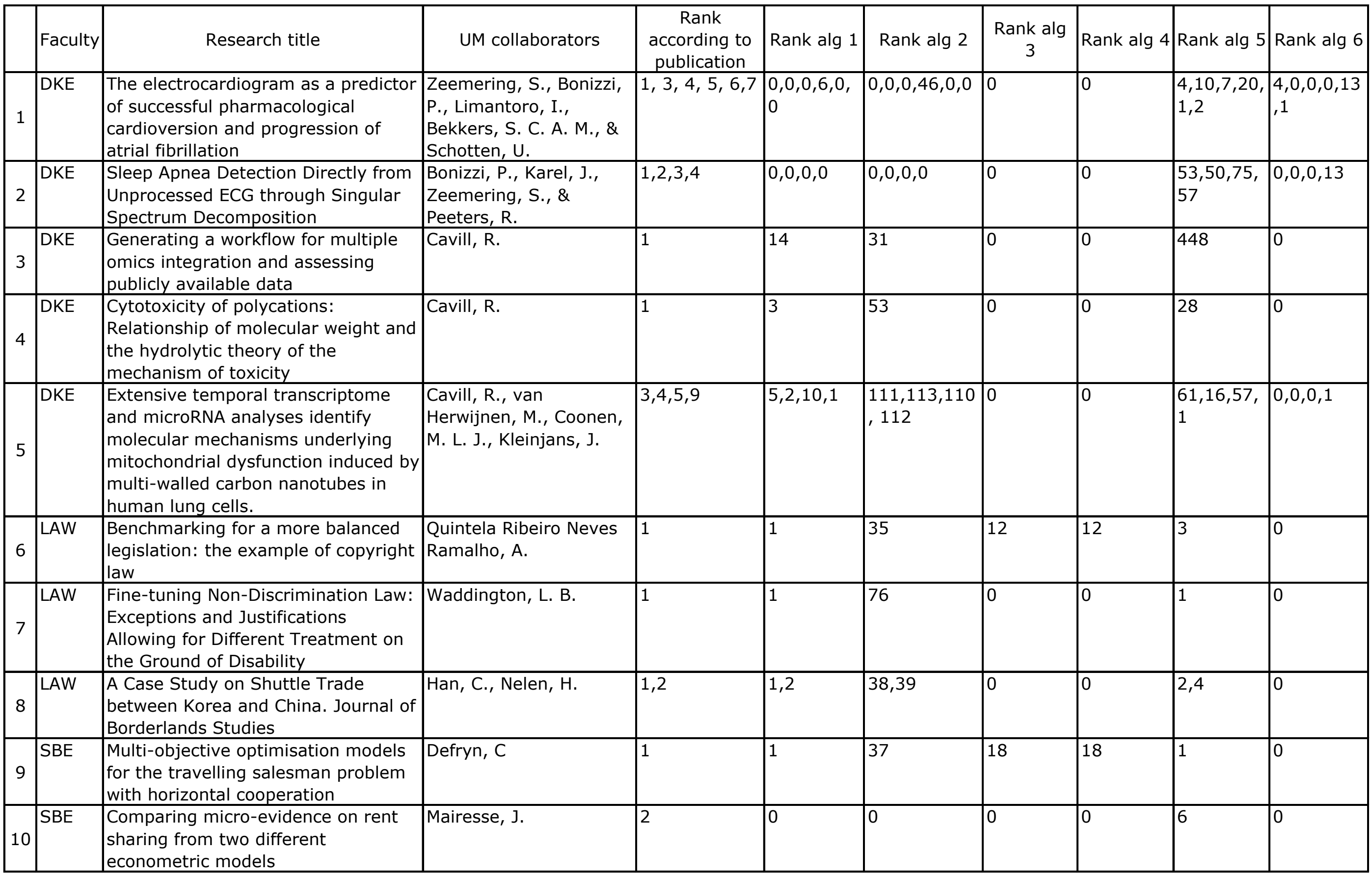


Researcher	Associated words
Cavill R.	"cell" "transcriptional" "epithelial" "epithelial_cell" "signaling" "expression" "metabolite" "metabolism" "deeper" "raw" "metabolomics" "repository" "molecular" "mechanism" "understood" "poorly" "molecular_mechanism" "interaction" "series" "similarity" "time" "amplitude" "time_series" "impacted" "cell" "cellular" "elucidate" "single" "development" "lineage"
Vidmar J.	"international" "relation" "international_relation" "sovereign" "state" "world" "aid" "donor" "independence" "episode" "recipient" "cochrane" "politics" "party" "inclusion" "article" "article_discus" "discus" "opinion" "performance" "probing" "expert_opinion" "bayes" "seek" "legal" "inclusive" "article" "particular" "law" "ascertain"
Lambin P.	"computed" "tomography" "lung" "computed_tomography" "october" "coronary" "quality" "high" "assist" "high_quality" "radiotherapy" "export" "prostate" "cancer" "antigen" "prostate_cancer" "bloom" "lethal" "patient" "outcome" "conclusion" "result" "method" "background" "tumor" "cns" "accessibility" "glioma" "cell" "important_implication"
Driessens K.	"cloud" "complicated" "intelligent" "reinforcement" "environment" "resource" "capability" "advantage" "competitive" "relational" "positive_effect" "value" "learning" "activity" "study" "based" "bilingual" "system" "start" "issue" "special_issue" "research" "ect" "special" "psychology" "journal" "article" "research" "field" "published"
Weiss G.	"learning" "activity" "study" "based" "bilingual" "system" "collaboration" "separate" "unified" "america" "negotiation" "within" "computational_geometry" "elaborate" "ambiguity" "sequentially" "operation_research" "automate" "agent" "practiced" "present" "able" "limited" "demonstrated" "transfer" "heat" "conductivity" "running" "heat_transfer" "thermal_conductivity"
Kelk S.	"species" "discovered" "phylogenetic" "new" "aquatic" "taxon" "tree" "take" "world" "real" "one" "real_world" "data" "network" "includes" "analysis" "metaphor" "structure" "paper" "algorithm" "hard" "problem" "heuristic" "solving" "paper" "modelling" "binary" "hole" "massive" "inferred" "model"
Bonizzi P.	"imaging" "resonance" "magnetic" "magnetic_resonance" "wavelength" "contrast" "decomposition" "depending" "conjunction" "proceed" "new" "measure" "engine" "ischemic" "dot" "myocardial" "diesel" "fuel" "validated" "model" "predicting" "validation" "developed" "trauma"
Dumontier M.	"ontology" "advanced" "library" "math" "ontology_engineering" "paper" "semantic" "universal" "english_language" "intercultural" "illinois" "keywords" "knowledge" "create" "practice" "competency" "based" "development" "drug" "induced" "vivo" "human" "observed" "specific" "animal" "procedure" "base" "contains" "committee" "performed"

Figure 3: Sample queries from the PURE dataset, with the corresponding topics according to the LDA model. Each row in the right column contains a set of words, together forming a topic.

C Sample queries from the PURE dataset, with the corresponding suggestions for collaborations

Figure 4: Part 1/3 of the sample queries from research titles from the PURE dataset. For each research title, the UM researchers have been noted with the corresponding rank according to the actual research. The remaining columns state the rank of those same researchers in that order, according to algorithm 1 to 6. A rank of 0 means the researcher has not been considered in the full list of potential relevant researchers.

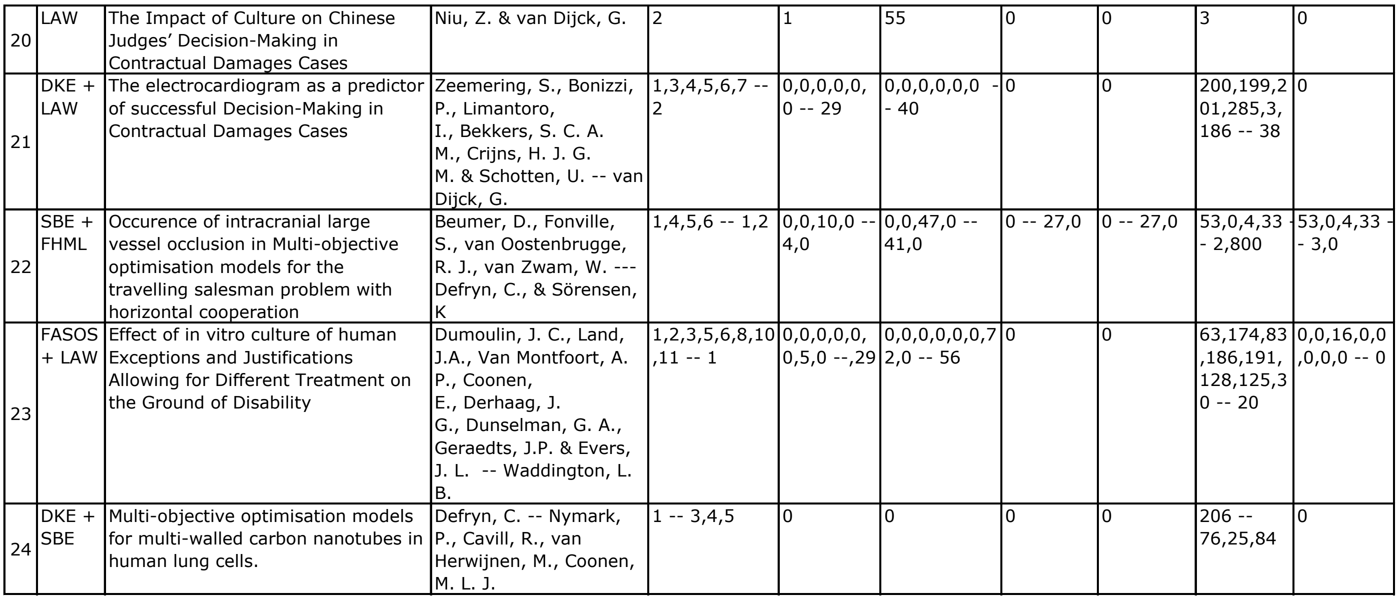


	Faculty	Research title	UM collaborators	Rank according to publication	Rank alg 1	Rank alg 2	Rank alg 3	Rank alg 4	Rank alg 5	Rank alg 6
1	DKE	The electrocardiogram as a predictor of successful pharmacological cardioversion and progression of atrial fibrillation	Zeemering, S., Bonizzi, P., Limantoro, I., Bekkers, S. C. A. M., & Schotten, U.	1, 3, 4, 5, 6,7	0,0,0,6,0,0	0,0,0,46,0,0	0	0	4,10,7,20,1,2	4,0,0,0,13,1
2	DKE	Sleep Apnea Detection Directly from Unprocessed ECG through Singular Spectrum Decomposition	Bonizzi, P., Karel, J., Zeemering, S., & Peeters, R.	1,2,3,4	0,0,0,0	0,0,0,0	0	0	53,50,75,57	0,0,0,13
3	DKE	Generating a workflow for multiple omics integration and assessing publicly available data	Cavill, R.	1	14	31	0	0	448	0
4	DKE	Cytotoxicity of polycations: Relationship of molecular weight and the hydrolytic theory of the mechanism of toxicity	Cavill, R.	1	3	53	0	0	28	0
5	DKE	Extensive temporal transcriptome and microRNA analyses identify molecular mechanisms underlying mitochondrial dysfunction induced by multi-walled carbon nanotubes in human lung cells.	Cavill, R., van Herwijnen, M., Coonen, M. L. J., Kleinjans, J.	3,4,5,9	5,2,10,1	111,113,110, 112	0	0	61,16,57,1	0,0,0,1
6	LAW	Benchmarking for a more balanced legislation: the example of copyright law	Quintela Ribeiro Neves Ramalho, A.	1	1	35	12	12	3	0
7	LAW	Fine-tuning Non-Discrimination Law: Exceptions and Justifications Allowing for Different Treatment on the Ground of Disability	Waddington, L. B.	1	1	76	0	0	1	0
8	LAW	A Case Study on Shuttle Trade between Korea and China. Journal of Borderlands Studies	Han, C., Nelen, H.	1,2	1,2	38,39	0	0	2,4	0
9	SBE	Multi-objective optimisation models for the travelling salesman problem with horizontal cooperation	Defryn, C	1	1	37	18	18	1	0
10	SBE	Comparing micro-evidence on rent sharing from two different econometric models	Mairesse, J.	2	0	0	0	0	6	0

Figure 5: Part 2/3 of the sample queries from research titles from the PURE dataset. For each research title, the UM researchers have been noted with the corresponding rank according to the actual research. The remaining columns state the rank of those same researchers in that order, according to algorithm 1 to 6. A rank of 0 means the researcher has not been considered in the full list of potential relevant researchers.

11	FASOS	Do Transnational Child-Raising Arrangements Affect Job Outcomes of Migrant Parents? Comparing Angolan Parents in Transnational and NonTransnational Families in the Netherlands	Haagsman, K.	1	1	87	57	62	8	0
12	FASOS	Beyond Helpless Victims and Survivor Trauma: New Historiography on Jews in the Age of	Laczo, F.	1	0	0	31	31	116	12
13	DKE	The electrocardiogram as a predictor of successful pharmacological cardioversion and progression of atrial fibrillation	Zeemering, S., Bonizzi, P., Limantoro, I., Bekkers, S. C. A. M., Crijns, H. J. G. M. & Schotten, U.	1,3,4,5,6,7	0,0,0,6,0,0	0,0,0,46,0,0	0	0	4,10,7,20,1,2	1,0,0,0,5,4
14	DKE	Graph kernels and Gaussian processes for relational reinforcement learning	Driessens, K.	1	1	33	12	12	1	5
15	FHML	Toll-like receptor 9 gene expression in the post-thrombotic syndrome, residual thrombosis and recurrent deep venous thrombosis: A case-control study	Bouman, A. C., Castoldi, E., Wielders, S. J., Spronk, H. M. H., ten Cate, H., Hoek - ten Cate, A.	2,3,4,5,6,7	0,2,1,3,0,5	0,96,95,97,0,99	0	0	380,70,100,30,6,58	0
16	FHML	Occurence of intracranial large vessel occlusion in consecutive, non-referred patients with acute ischemic stroke	Beumer, D., Fonville, S., van Oostenbrugge, R. J., van Zwam, W.	1,4,5,6	4,0,1,2	72,0,77,78	57,0,0,51	57,0,0,51	6,0,1,2	0,0,4,2
17	LAW	Family Rights and Immigration Law: a European Perspective	Forder, C. J.	1	1	41	0	0	4	0
18	LAW	When moral intuitions are immune to the law: a case study of euthanasia and the Act-Omission Distinction in The Netherlands	Hauser, M. D., Tonnaer, F. M. C. L. & Cima - Knijff, M. J.	2,3	1,2	65,66	0	0	6,8	0
19	FASOS	Effect of in vitro culture of human embryos on birthweight of newborns	Dumoulin, J. C., Land, J.A., Van Montfoort, A. P., Coonen, E., Derhaag, J. G., Dunselman, G. A., Geraedts, J.P. & Evers, J. L.	1,2,3,5,6,8,10,11	7,6,9,8,0,4,3,5	37,36,39,38,0,34,33,35	0,0,0,0,28,0,0,0	0,0,0,0,28,0,0,0	1,18,6,10,27,680,8,4	1,5,0,0,0,15,6,0

Figure 6: Part 3/3 of the sample queries from research titles from the PURE dataset. For each research title, the UM researchers have been noted with the corresponding rank according to the actual research. The remaining columns state the rank of those same researchers in that order, according to algorithm 1 to 6. A rank of 0 means the researcher has not been considered in the full list of potential relevant researchers.



20	LAW	The Impact of Culture on Chinese Judges’ Decision-Making in Contractual Damages Cases	Niu, Z. & van Dijck, G.	2	1	55	0	0	3	0
21	DKE + LAW	The electrocardiogram as a predictor of successful Decision-Making in Contractual Damages Cases	Zeemering, S., Bonizzi, P., Limantoro, I., Bekkers, S. C. A. M., Crijns, H. J. G. M. & Schotten, U. -- van Dijck, G.	1,3,4,5,6,7 -- 2	0,0,0,0,0,0 -- 29	0,0,0,0,0,0 -- 40	0	0	200,199,201,285,3,186 -- 38	0
22	SBE + FHML	Occurence of intracranial large vessel occlusion in Multi-objective optimisation models for the travelling salesman problem with horizontal cooperation	Beumer, D., Fonville, S., van Oostenbrugge, R. J., van Zwam, W. --- Defryn, C., & Sörensen, K	1,4,5,6 -- 1,2	0,0,10,0 -- 4,0	0,0,47,0 -- 41,0	0 -- 27,0	0 -- 27,0	53,0,4,33 -- 2,800	53,0,4,33 -- 3,0
23	FASOS + LAW	Effect of in vitro culture of human Exceptions and Justifications Allowing for Different Treatment on the Ground of Disability	Dumoulin, J. C., Land, J.A., Van Montfoort, A. P., Coonen, E., Derhaag, J. G., Dunselman, G. A., Geraedts, J.P. & Evers, J. L. -- Waddington, L. B.	1,2,3,5,6,8,10,11 -- 1	0,0,0,0,0,0,5,0 --,29	0,0,0,0,0,0,72,0 -- 56	0	0	63,174,83,186,191,128,125,30 -- 20	0,0,16,0,0,0,0,0 -- 0
24	DKE + SBE	Multi-objective optimisation models for multi-walled carbon nanotubes in human lung cells.	Defryn, C. -- Nymark, P., Cavill, R., van Herwijnen, M., Coonen, M. L. J.	1 -- 3,4,5	0	0	0	0	206 -- 76,25,84	0