An algorithm for topic modeling of researchers taking into account their interests in Google Scholar profiles

Serhiy Shtovba[a] and Mykola Petrychko[b]

[a] *Vasyl' Stus Donetsk National University, 600-richchia str., 21, Vinnytsia, 21021, Ukraine*
[b] *Vinnytsia National Technical University, Khmelnytske Shose, 95, Vinnytsia, 21021, Ukraine*

Abstract

An algorithm for topic modeling of researchers based on their interests from Google Scholar's profiles is proposed. As topics for modeling, we took research groups from research classification system ANZSRC – Australian and New Zealand Standard Research Classification. Researchers' distribution to research groups is found based on their interests' usage statistics in categorized publications from Dimensions. Topic modeling is conducted accordingly to principles of statistical support, multi-labeling, noise filtering, ignoring stop-words, solidarities, focusing, compactness and research groups' interactions. We compare topic modeling based on data with low level of information from researchers' profiles in Google Scholar with topic modeling based on a few dozen authored publications categorized by Dimensions. Comparison is made by modified Czekanowski metric that takes into account the interaction between research groups. By comparing the results of topic modeling based on different sources of initial information a good match was found. It allows to use the proposed algorithm as the intellectual core of information technology in regards to scientific staff, in particular, for the selection of candidates as opponents of a dissertation, as reviewers for research projects, for forming a team to collaborate on mutual research projects etc.

Keywords

topic modeling, Google Scholar, Dimensions, ANZSRC, researcher's profile, research interests, research group, Czekanowski metric, Jaccard index.

1. Introduction

Google Scholar aggregates the most volumetric collection of researchers' profiles. The most used information from Google Scholar profiles is citations. It, for example, is used as primary information for university ranking in Webometrics. Several studies, in particular [1, 2], are concerned with comparing concordance of Google Scholar citations with different scientometrics systems such as Scopus, Web of Science, Dimensions and others, that use only meta-information from publishers. A researcher's profile in Google Scholar contains not only publications and their citations, but also other information. In particular, a researcher provides his or her interests. A researcher chooses the interests in a loose manner without any limitations. Google Scholar provides a web interface to search researchers by an interest. However, the results are formed by literal coincidence. That is why the results for *fuzzy set* and *fuzzy sets* are different; the same applies for synonymous interests such as *fuzzy evidence* and *fuzzy inference*. Moreover, Google Scholar does not take into account an interconnection of the interests, that is the search by an interest is done independently and isolated. Given that, the search and analytical services that provide information about many researchers in Google Scholar are relatively straightforward.

The goal of this paper is topic modeling of researchers based on their interests from Google Scholar. Methods that process a researcher's interests from Google Scholar profile are not studied

CMIS-2021: The Fourth International Workshop on Computer Modeling and Intelligent Systems, April 27, 2021, Zaporizhzhia, Ukraine
EMAIL: shtovba@donnu.edu.ua (S. Shtovba); mpetrychko@vntu.edu.ua (M. Petrychko)
ORCID: 0000-0003-1302-4899 (S. Shtovba); 0000-0001-6836-7843 (M. Petrychko)

well. We identified only two relevant publications. One of them is [3], it describes a recommendation system that recommends supervisors based on some information and interests of candidates from Google Scholar profiles as well. Another paper [4] presents an information technology that synthesizes a research profile of institute or research laboratory. It also uses interests of researchers from their profiles at Google Scholar. Articles [3, 4] are based on pairwise comparison using cosine similarity metric between researcher and a set of keywords from a given topic. Such a topic in [3] is an article at Wikipedia. Unlike these methods, we strive to categorize researchers by a given research classification system, that is to assign a research group to each of them.

Automatic researchers' categorization is usually done as a result of generalizing the topics of their publications. One of the methods for this is presented at [5]. The authors present a statistical model "Author-Topic" that is based on topic modeling model Latent Dirichlet Allocation [6]. This model represents a researcher as a distribution over some abstract topics. The topics are clusters of similar words. One of the drawbacks of this model is low interpretation of the topics because they are formed by words frequency in a document. To improve the interpretation another model "Author-Subject-Topic" is proposed in [7]. This model additionally uses a research specialty that is defined by journal in which an analyzed publication is published. In [8] another improvement of "Author-Topic" model is presented – "Author-Persona-Topic" model. In this model rather than representing all researcher's documents as single topic distribution, authors group all documents into different clusters, each with its own topic distribution. These clusters represent "personas" under which an author writes.

Apart from topic modeling methods there are also methods based on word embedding. They generally perform better than topic models because they can incorporate semantic relationships. One of the most popular models of embedding is word2vec [13]. It is used in [9] as a part of similarity metric between researchers using their publications. They assess the similarity between words that comprise publications of different researchers using representation of words defined by word2vec. The authors of [10] use publications' titles as source information for solving the problem of collaboration recommendation. The words from the titles are represented as vectors using word2vec. These vectors are then clustered using k-means to partition researchers into different academic domains. The representation of a researcher is further improved by using his co-authorship and the random walk method to find his influence in different domains. In [11] authors represent a researcher as a set of documents he/she has written. The words of the documents are defined as vectors trained by word2vec model. These representations are then used to solve the problem of expert finding by utilizing a restricted convolutional neural network. In [12] a researcher is represented as a concatenation of all his/hers abstracts. Each word in the concatenation is then represented as a vector from word2vec model to solve the problem of reviewer recommendation.

Analyzed methods assume to have enough number of publications for a given researcher with selected keywords. At the same time, they do not account for the fact that co-author contribution is sometimes relative to a small subset of the paper keywords. A researcher, especially a young one, may have only a few publications that may not be enough for a valid categorization. On the other hand, the researcher can manually specify at the profile a set of keywords that describe his (or her) activities. As the time goes on a researcher may change his research direction, for example, move to another laboratory or another project. Given that there is no change when a researcher is categorized based on his publications, categorization based on his keywords may find these changes. By taking that in mind, we study the topic modeling based on actual interests that a researcher specified by himself (or herself) at the current moment.

2. Problem statement

We use the following notations:

$W = (w_1, w_2, ..., w_n)$ is a set of keywords that are equal to researcher interests in Google Scholar profile;

$T = (t_1, t_2, ..., t_m)$ is a set of research topics from a research classification system;

$D_1, D_2, ..., D_m$ is a set of topic-marked collections of texts; each collection contains only publications from topics $t_1, t_2, ..., t_m$, respectively;

$B = D_1 \cup D_2 \cup ... \cup D_m$ is the general collection of topic-marked texts; each element of B belongs to one or more topics from the set T;

$R(D,T) \subset D \times T$ is a relation that describes membership of a publication to topic-marked collections.

The problem is to find out topics from T that correspond to the set of interests W. The results of mapping $W \to T$ is a fuzzy set $\widetilde{W}$ defined on the universal set of topics T as follows:

$$\widetilde{W} = \left(\frac{\mu_W(t_1)}{t_1}, \frac{\mu_W(t_2)}{t_2}, ..., \frac{\mu_W(t_m)}{t_m} \right),$$

where $\mu_W(t_p) \in [0,1]$ denotes membership degree of the set of interests W to topics t_p, $p = \overline{1,m}$.

We set the following restrictions on $\widetilde{W}$:

1) the cardinality of the fuzzy set support must be small $1 \le \left| \sup\left(\widetilde{W}\right) \right| \le T_{\max}$, for example, with $T_{\max} \in \{2,3,4\}$ a researcher will be assigned only to a few topics;
2) $\sum_{p=\overline{1,m}} \mu_W\left(t_p\right) = 1$, which is equivalent to the topic modeling regularization condition.

3. Data acquisition and preprocessing

We use a researcher's profile from Google Scholar to get the keywords. For example, in Figure 1 we have a researcher's profile with three keywords that are marked with blue color. For this researcher: $w_1 =$"*Computational Intelligence*"; $w_2 =$"*Fuzzy Logic*"; $w_3 =$"*Artificial Intelligence*". The order of keywords in the set W is not important, this corresponds to the bag of words model. Interests often complement each other thus making their research topics more focused. To take that into account we synthesize additional keywords defined as pairs of initial interests. Interests in a pair are combined by a logical operation *AND*. For researcher from Figure 1 additional keywords are defined as follows:

$w_4 =$"*Computational Intelligence*" *AND* "*Fuzzy Logic*";

$w_5 =$"*Computational Intelligence*" *AND* "*Artificial Intelligence*"

$w_6 =$"*Fuzzy Logic*" *AND* "*Artificial Intelligence*"

If a researcher's profile has 3 interests, additional 3 keywords are synthesized, if it has 4 interests then 6 additional are synthesized etc.

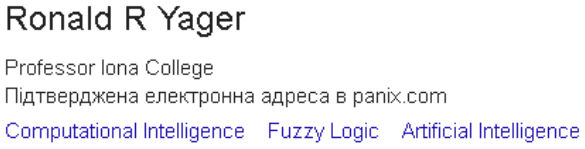

Figure 1: An example of a researcher's profile with 3 interests

For researchers' topic modeling, we need to choose a research classification system. There are a lot of them, but when choosing we take into account not only their semantic advantages and disadvantages, but also that there is an information system with this research classification system that has available search services. In addition, we require that the information system must indexes a large number of categorized publications over all research. The information system that satisfies these requirements is Dimensions.

Dimensions indexes more than 110M of publications. All publications are categorized by the two-level variant of Australian and New Zealand Standard Research Classification (ANZSRC) with 22 research divisions and 154 research groups (Table 1). In this work we use the research groups to model a researcher's interests.

Table 1.
Research classification system ANZSRC, that is used in Dimensions

Research Division	Research Group
Mathematical Sciences	A1 - Pure Mathematics; A2 - Applied Mathematics; A3 - Numerical and Computational Mathematics; A4 – Statistics; A5 - Mathematical Physics
Physical Sciences	B1 - Astronomical and Space Sciences; B2 - Atomic, Molecular, Nuclear, Particle and Plasma Physics; B3 - Classical Physics; B4 - Condensed Matter Physics; B5 - Optical Physics; B6 - Quantum Physics; B7 - Other Physical Sciences
Chemical Sciences	C1 - Analytical Chemistry; C2 - Inorganic Chemistry; C3 - Macromolecular and Materials Chemistry; C4 - Medicinal and Biomolecular Chemistry; C5 - Organic Chemistry; C6 - Physical Chemistry (incl. Structural); C7 - Theoretical and Computational Chemistry; C8 - Other Chemical Sciences
Earth Sciences	D1 - Atmospheric Sciences; D2 - Geochemistry; D3 - Geology; D4 - Geophysics; D5 - Oceanography; D6 - Physical Geography and Environmental Geoscience; D7 - Other Earth Sciences;
Environmental Sciences	E1 - Ecological Applications; E2 - Environmental Science and Management; E3 - Soil Sciences; E4 - Other Environmental Sciences;
Biological Sciences	F1 - Biochemistry and Cell Biology; F2 - Ecology; F3 - Evolutionary Biology; F4 - Genetics; F5 - Microbiology; F6 – Physiology; F7 - Plant Biology; F8 - Zoology; F9 - Other Biological Sciences
Agricultural and Veterinary Sciences	G1 - Agriculture, Land and Farm Management; G2 - Animal Production; G3 - Crop and Pasture Production; G4 - Fisheries Sciences; G5 - Forestry Sciences; G6 - Horticultural Production; G7 - Veterinary Sciences; G8 - Other Agricultural and Veterinary Sciences
Information and Computing Sciences	H1 - Artificial Intelligence and Image Processing; H2 - Computation Theory and Mathematics; H3 - Computer Software; H4 - Data Format; H5 - Distributed Computing; H6 - Information Systems; H7 - Library and Information Studies; H8 - Other Information and Computing Sciences
Engineering	I1 - Aerospace Engineering; I2 - Automotive Engineering; I3 - Biomedical Engineering; I4 - Chemical Engineering; I5 - Civil Engineering; I6 - Electrical and Electronic Engineering; I7 - Environmental Engineering; I8 - Food Sciences; I9 - Geomatic Engineering; I10 - Manufacturing Engineering; I11 - Maritime Engineering; I12 - Materials Engineering; I13 - Mechanical Engineering; I14 - Resources Engineering and Extractive Metallurgy; I15 - Interdisciplinary Engineering; I16 - Other Engineering
Technology	J1 - Agricultural Biotechnology; J2 - Environmental Biotechnology; J3 - Industrial Biotechnology; J4 - Medical Biotechnology; J5 - Communications Technologies; J6 - Computer Hardware; J7 – Nanotechnology; J8 - Other Technology
Medical and Health Sciences	K1 - Medical Biochemistry and Metabolomics; K2 - Cardiorespiratory Medicine and Haematology; K3 - Clinical Sciences; K4 - Complementary and Alternative Medicine; K5 - Dentistry; K6 - Human Movement and Sports Science; K7 – Immunology; K8 - Medical Microbiology; K9 – Neurosciences; K10 - Nursing; K11 - Nutrition and Dietetics; K12 - Oncology and Carcinogenesis; K13 - Ophthalmology and Optometry; K14 - Paediatrics and Reproductive Medicine; K15 - Pharmacology and Pharmaceutical Sciences; K16 - Medical Physiology; K17 - Public Health and Health Services; K18 - Other Medical and Health Sciences;

Research Division	Research Group
Built Environment and Design	L1 - Architecture; L2 – Building; L3 - Design Practice and Management; L4 - Engineering Design; L5 - Urban and Regional Planning; L6 - Other Built Environment and Design;
Education	M1 - Education Systems; M2 - Curriculum and Pedagogy; M3 - Specialist Studies In Education; M4 - Other Education
Economics	N1 – Economic Theory; N2 – Applied Economics; N3 – Econometrics; N4 – Other Economics
Commerce, Management, Tourism and Services	O1 - Accounting, Auditing and Accountability; O2 - Banking, Finance and Investment; O3 - Business and Management; O4 - Commercial Services; O5 – Marketing; O6 – Tourism; O7 - Transportation and Freight Services;
Studies in Human Society	P1 - Anthropology; P2 - Criminology; P3 - Demography; P4 - Human Geography; P5 - Policy and Administration; P6 - Political Science; P7 - Social Work; P8 - Sociology; P9 - Other Studies In Human Society
Psychology and Cognitive Sciences	Q1 - Psychology; Q2 - Cognitive Sciences; Q3 - Other Psychology and Cognitive Sciences;
Law and Legal Studies	R1 – Law; R2 - Other Law and Legal Studies
Studies in Creative Arts and Writing	S1 - Art Theory and Criticism; S2 - Film, Television and Digital Media; S3 - Journalism and Professional Writing; S4 - Performing Arts and Creative Writing; S5 - Visual Arts and Crafts; S6 - Other Studies In Creative Arts and Writing
Language, Communication and Culture	T1 - Communication and Media Studies; T2 - Cultural Studies; T3 - Language Studies; T4 – Linguistics; T5 - Literary Studies; T6 - Other Language, Communication and Culture
History and Archaeology	U1 - Archaeology; U2 - Curatorial and Related Studies; U3 - Historical Studies; U4 - Other History and Archaeology
Philosophy and Religious Studies	V1 - Applied Ethics; V2 - History and Philosophy of Specific Fields; V3 - Philosophy; V4 - Religion and Religious Studies; V5 - Other Philosophy and Religious Studies

A query to Dimensions is formed separately by each element of the set *W*. If an element is a phrase, then it is surrounded by quotes. As a search scope, we use *Title and Abstract* and we search only the last 5 years – 2016 – 2020. An example of a search result for the query "*fuzzy logic*" is presented in Figure 2. For each research division and research group there is a number of publications that has the query mentioned in either the title or abstract. The results are sorted by the number of publications descending. We can also find the overall number of publications for each research division and group, that is without any query.

4. Topic modeling algorithm

We perform the topic modeling of researchers based on the following principles:

- *the principle of statistical support* – the more publications from a specific research group a given keyword contains, the more membership degree of this keyword to this research group is;
- *the principle of multi-labeling* – a keyword can belong to a few research groups;
- *the principle of noise filtering* – we ignore research groups with low membership degree to a given keyword;
- *the principle of ignoring stop-words* – we ignore keywords that appear in a very large number of publications;

- *the principle of solidarities* – the more keywords belong to the same research group the larger the chance that the researcher belongs to this research group;
- *the principle of focusing* – if a topic-marked collection of publications contains a few keywords of a researcher at once then the chances to assign this researcher to the respected topic increase;
- *the principle of compactness* – a researcher can only be assigned to a few research groups;
- *the principle of research groups interaction* – when cutting the tail of topic distribution, the contribution of minor research groups is redistributed on leaders by taking into account their similarity.

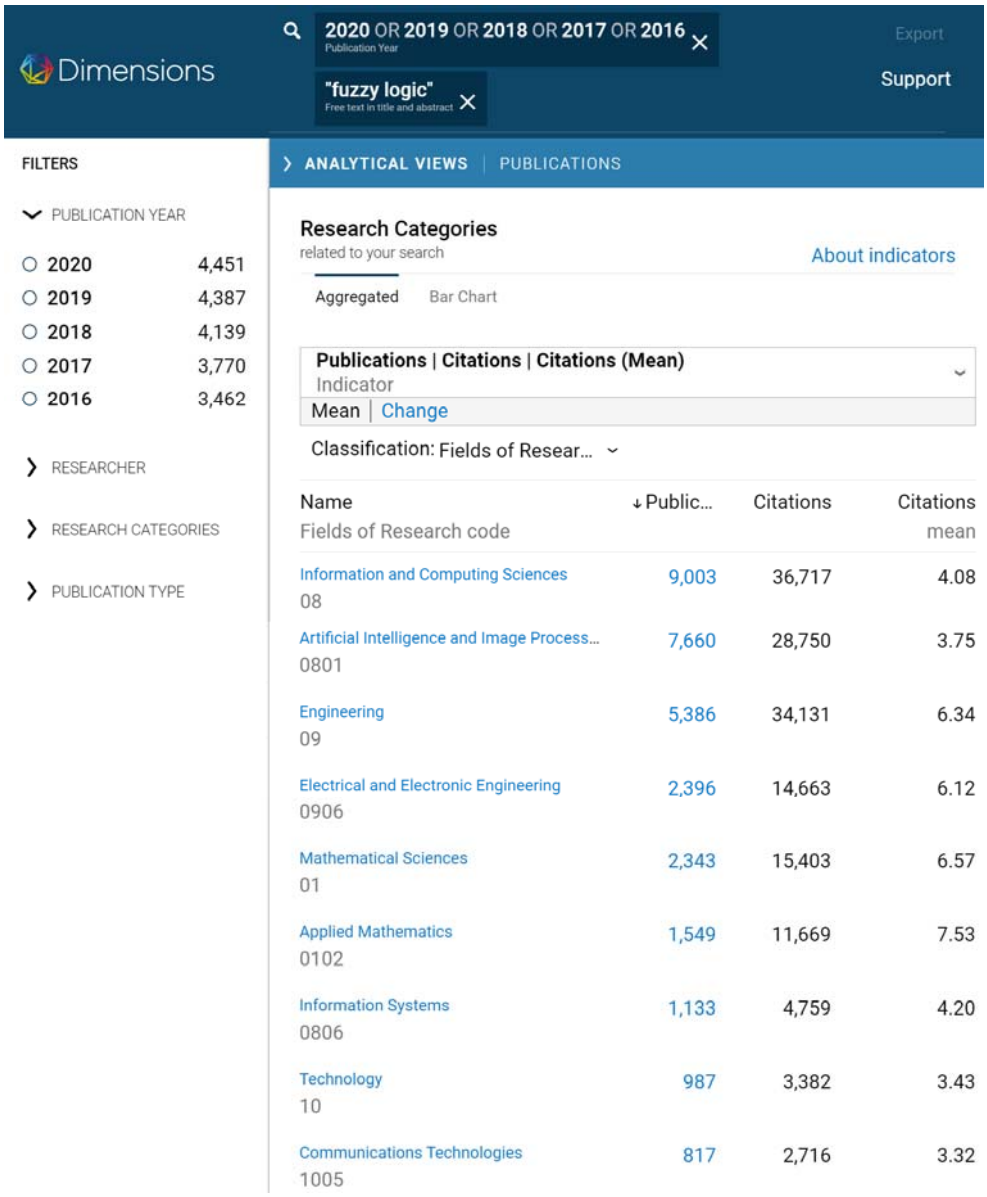

Figure 2: The results from Dimensions by search query “fuzzy logic” for the period 2016-2020

We propose an algorithm to implement the proposed principles that consists of 3 stages. On the first stage the set of queries based on keywords and their combination is formed. We use only pairs of keywords because the results using triples of keywords are often empty and increase the processing time. The second stage performs topic modeling by each query separately. Research groups are chosen by the frequency of mentions at a topic-marked collection. Stop-words and noise are filtered by the frequency of mentions in research groups at all topic-marked collections. The minor research groups are left out using cumulative principle, by cutting the tail of the distribution. On the third stage

all membership degrees of queries are averaged, the resulted distribution is cut and research groups with the low membership degree are dropped. To ensure compactness we only allow 1 to 4 research groups.

```
%Topic modeling algorithm
%  #1 – creating the set E of search queries from the keywords
E=W
for i=1:length(W)
      for j=i:length(W)
            E={E; [‘“’ W(i) ‘AND’ W(j) ‘”’] }
      end
end
%  #2 – compute membership degrees to research groups by each query
< Find the number of publications at each topic-marked collection
  N=[N(1), N(2), …, N(m)] >
Counter=0  % the counter of successful query responses
for i=1:length(E)
      < Find Q – the number of publications D, that contain E{i} >
      If Q>Threshold_SW continue  % stop-words
      end
      If Q<Threshold_noise continue;  % noise
      end
      < Find t(1), t(2),…, t(m) – the number of publications in the topic-
        marked collections in each research group for query E{i} >
      % Ignore topics with a low number of publications:
      index=find(t<Threshold_topic)
      t(index)=0
      if max(t)==0 continue
      end
      % Compute the frequency of E{i} at topic-marked collections:
      Gamma=t./N
      < Choose the most popular research groups that have cumulative
        contribution in Gamma not lower than Tail_1. Research groups that have
        cumulative contribution lower than Tail_1 are put in vector Rejected >
      %  Ignore research groups with contribution lower than Tail_1:
      Gamma(Rejected)=0
      Gamma=Gamma./sum(Gamma)  % norm to be in [0, 1]
      Counter=Counter+1
      Mu(Counter)=Gamma
end
If    Counter==0
      return (‘Unsuccessful’)
end
%  #3 – compute membership degrees using all queries
Mu_mean=mean(Mu)  % averaging all successful queries
< Form leaders of research groups that have cumulative contribution Mu_mean not
  lower than Tail_2. We restrict the number of leaders to be at most 6 with the
  largest cumulative contribution. If we have more than 6 leaders their numbers
  will be in the vector Rejected >
% Ignore research groups with contribution lower than Tail_2:
Mu_mean(Rejected)=0
Mu_mean= Mu_mean./sum(Mu_mean)  % norm to be in [0, 1]
%  Current number of research groups:
Current_N_fields=sum(Mu_mean>0)
```

```
%  Set the max number of research groups for a researcher:
T_max=min(4, Counter+1)
< Find Mu_worst – the smallest membership degree among the leaders >
while (Current_N_fields>T_max OR Mu_worst<Tail_3)
      < Drop the minor group and redistribute its contribution to others based
        on their similarity >
      Current_N_fields=Current_N_fields-1;
      Mu_mean= Mu_mean./sum(Mu_mean)  % norm to be in [0, 1]
      < Find Mu_worst – the smallest membership degree among chosen
        research groups >
end
```

On the last stage of the algorithm when dropping a minor research group its contribution is redistributed to other research groups based on the similarity defined at [14, 15]. For example, let us say that on an intermediate stage a researcher is assigned to research groups in the following way: $\widetilde{W}=\left(\frac{0.5}{H6},\frac{0.2}{O5},\frac{0.2}{O6},\frac{0.1}{O4}\right)$. Let us drop the minor group *O4*. For this, first using method from [14, 15] we compute Jaccard indexes between *O4* and other research groups. For the data from 2016 – 2020 they are:

$J(O4,H6)=0$;

$J(O4,O5)=0.13$;

$J(O4,O6)=0.22$.

By taking into account the similarity, the contribution of the minor specialty *O4* is redistributed in the following way:

$$\widetilde{W}=\left(\frac{0.5+0\cdot 0.1}{H6},\frac{0.2+0.13\cdot 0.1}{O5},\frac{0.2+0.22\cdot 0.1}{O6}\right).$$

As a result, we get:

$$\widetilde{W}=\left(\frac{0.5}{H6},\frac{0.213}{O5},\frac{0.222}{O6}\right).$$

After norming to be in [0,1] we have:

$$\widetilde{W}=\left(\frac{0.535}{H6},\frac{0.228}{O5},\frac{0.237}{O6}\right).$$

5. Checking example

Let us illustrate how the algorithm works using as an example topic modeling of the researcher from Figure 1. Using three interests, we form six queries. Figure 3 shows frequency of queries at topic-marked collections. Figure 4 shows the results after cutting the first tail of the distribution. Next, we average by all queries (Figure 5) and cut the tail of the distribution (Figure 6). The resulting distribution is overfilled due to the broad usage of interests for the given researcher. To make the results more focused the final stage of the algorithm reduces the number of research groups to 2 (Figure 7). As a result, we get that a researcher with interests at artificial intelligence and neural networks has the largest membership degree in research groups *H1 – Artificial Intelligence and Image Processing* with membership degree 0.441 and *H2 – Computation Theory and Mathematics* with membership degree 0.559. Such a categorization of the researcher does not contradict with the authors' viewpoint. The example shows that even with two initial keywords the proposed algorithm finds a good enough membership relation between the researcher and research groups.

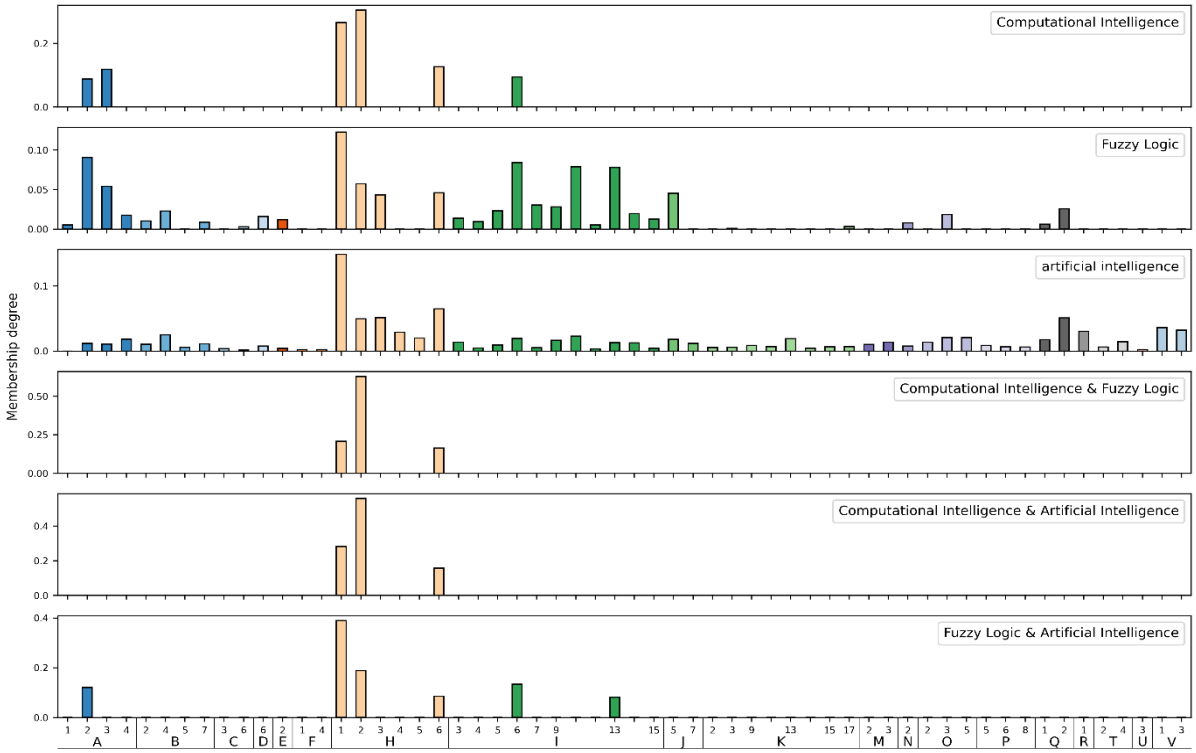

Figure 3: Initial membership distribution of each interest to research groups

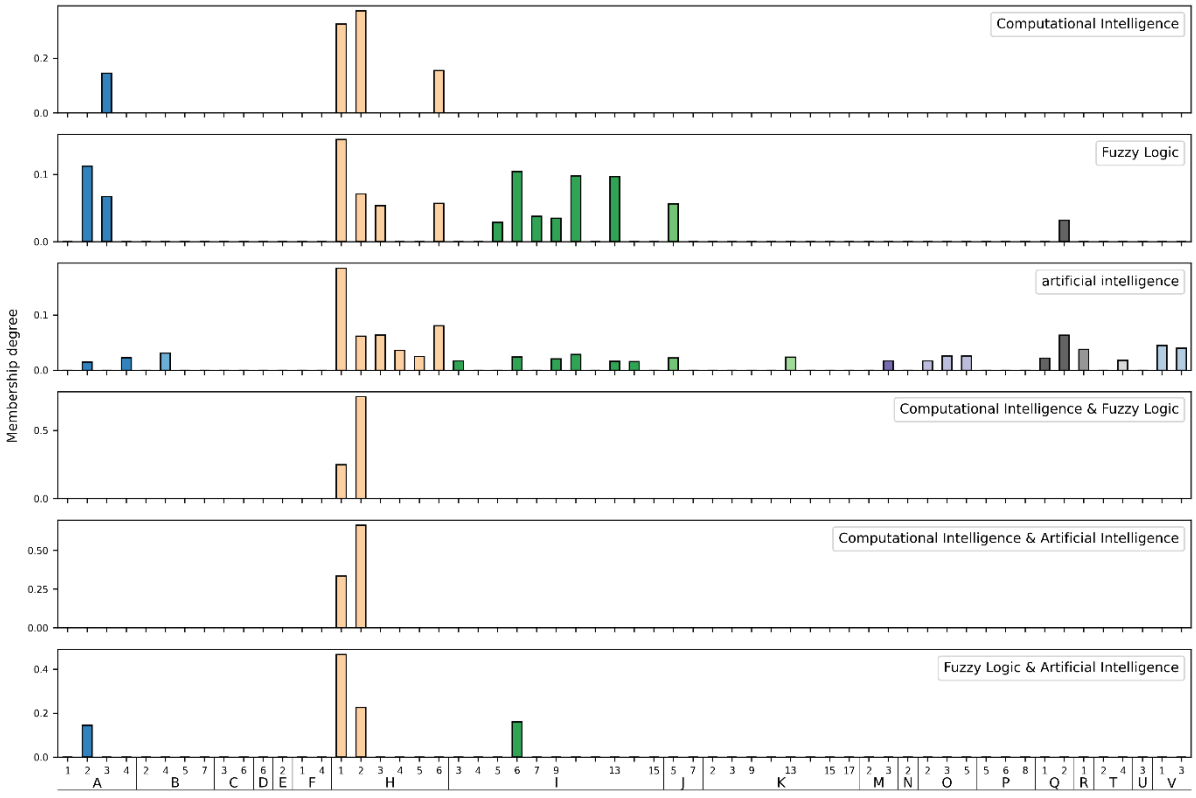

Figure 4: Distributions after the first noise filtering

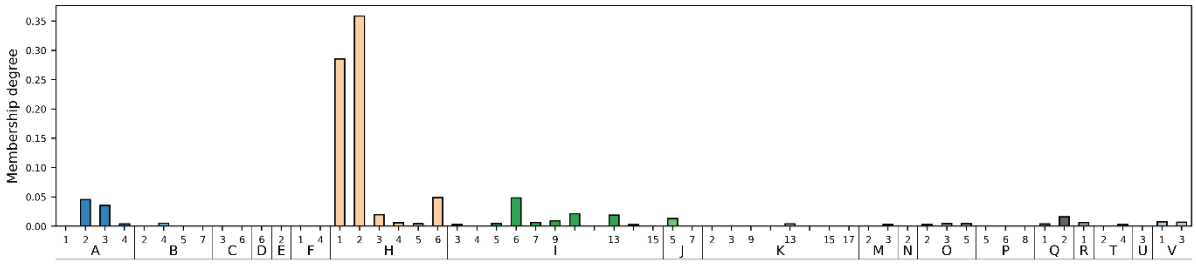

Figure 5: Averaging distributions of all keywords

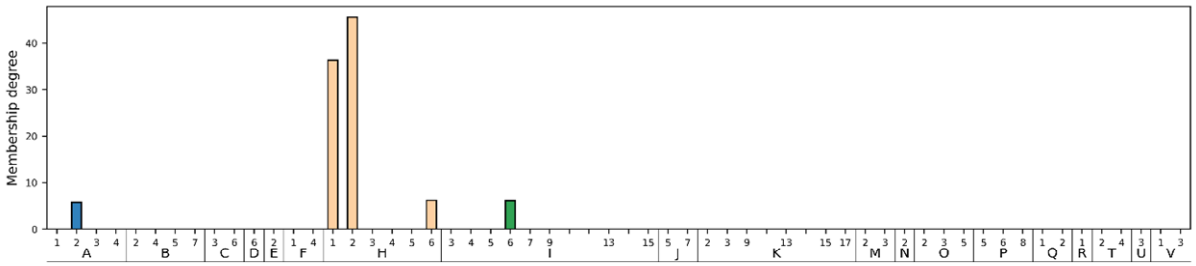

Figure 6: Distributions after the second noise filtering

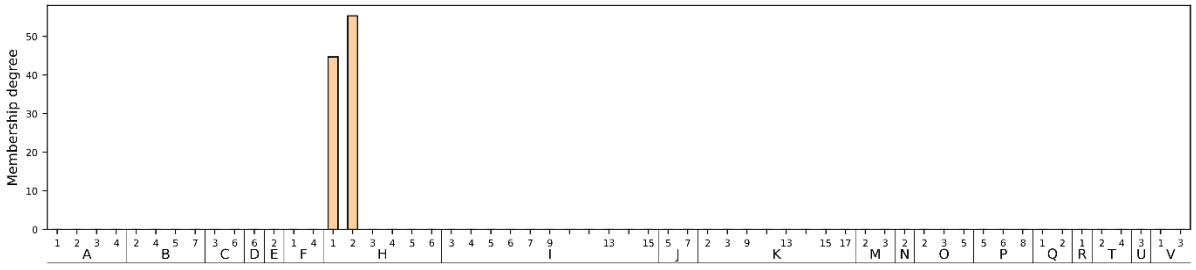

Figure 7: The result of topic modeling for the researcher from Figure 1

6. Comparing with categorized papers

Let us compare the topic modeling results based on keywords from researchers' profiles in Google Scholar and based on categorized publications of the researchers in Dimensions. For this we take three researchers:

- Ronald Yager with the interests *Computational Intelligence*, *Fuzzy Logic*, and *Artificial Intelligence*;
- Nataliia Kussul with the interests *Machine Learning*, *Remote Sensing*, *Data Science*, *Disaster Management*, and *Agricultural Monitoring*;
- Yevgeniy Bodyanskiy with the interests *Computational Intelligence*, *Data Mining*, *Data Stream Mining*, and *Big Data*.

The results of topic modeling of researchers with the above-mentioned interests are presented in Table 2.

These researchers have a considerable number of publications in Dimensions for the last 5 years that allows getting statistically significant results.

During the analyzed period, Ronald Yager published 141 papers that are categorized by 20 research groups. The most publications – 63 are assigned to the research group *H1*. Yevgeniy Bodyanskiy published 88 papers. They are categorized by 12 research groups. The most publications – 59 are assigned to the research group *H1*. Nataliia Kussul published 47 papers that are categorized by 14 research groups. The most publications – 21 are assigned to research group *I9*. By using the third stage of the proposed algorithm on papers distributions, we get membership degrees to research groups (Table 2).

Table 2
The results of researchers' topic modeling

Research group	Ronald Yager		Nataliia Kussul		Yevgeniy Bodyanskiy	
	Dimensions	Google Scholar	Dimensions	Google Scholar	Dimensions	Google Scholar
D6				0.283		
I9			0.675	0.447		
H1	0.63	0.441	0.172	0.346	0.797	0.295
H2		0.559				0.199
H6	0.37		0.153		0.203	0.506

Comparing the results, we see that topic modeling based on interests from Google Scholar – laconic subjective information, with the proposed algorithm categorizes researchers good enough. For quantitative assessment of the results, we used Czekanowski metric. For the case when membership degrees are in [0, 1], Czekanowski metric between two researchers W_1 and W_2 is computed in the following way:

$$Fit(W_1, W_2) = \sum_{p=\overline{1,m}} \min(\mu_{t_p}(W_1), \mu_{t_p}(W_2)). \quad (1)$$

The metric (1) can be interpreted as a sum of membership degrees of the intersection of fuzzy sets $\widetilde{W}_1$ and $\widetilde{W}_2$, that represent the topic modeling results based on two source of information – interests from Google Scholar and categorized publications in Dimensions.

Based on the data from Table 2 we get the following assessments using metric (1):

$Fit(Yager) = 0.441$;

$Fit(Kussul) = 0.619$;

$Fit(Bodyanskiy) = 0.498$.

Using metric (1) the match is computed with isolate assumption – only in the scope of each individual research group. To take into account the contribution of similar research groups we propose to the value of metric (1) to add the following interactive addend:

$$\Delta Fit(W_1, W_2) = \sum_{v=\overline{1,M}} \sum_{p=\overline{1,M}} J(t_v, t_p) \cdot \min(\varepsilon_{t_v}(W_1), \varepsilon_{t_p}(W_2)) \quad (2)$$

where $J(t_v, t_p)$ denotes Jaccard index between research groups t_v and t_p;

$\varepsilon_{t_v}(W_1) = \min(0, \mu_{t_v}(W_1) - \mu_{t_v}(W_2))$ denotes residual of membership degree to research group t_v in $\widetilde{W}_1$ after taking into account the matching between $\mu_{t_v}(W_1)$ and $\mu_{t_v}(W_2)$ in (1);

$\varepsilon_{t_p}(W_2) = \min(0, \mu_{t_p}(W_2) - \mu_{t_p}(W_1))$ denotes residual of membership degree to research group t_p in $\widetilde{W}_2$ after taking into account the matching between $\mu_{t_p}(W_1)$ and $\mu_{t_p}(W_2)$ in (1).

To filter information noise, we use the formula (2) only for pairs of research groups with a high level of similarity – with Jaccard index greater than 0.02. For the research groups from the Table 2 we have 2 such pairs. Jaccard indexes for them are the following:

$J(D6, I9) = 0.083$;

$J(H1, H6) = 0.071$.

Substituting the indexes in (2), we get:

$\Delta Fit(Yager) = 0$;

$\Delta Fit(Kussul) = \min(0.675 - 0.447, 0.283) \cdot 0.083 + \min(0.346 - 0.172, 0.153) \cdot 0.071 = 0.03$;

$\Delta Fit(Bodyanskiy) = \min(0.797 - 0.295, 0.506 - 0.203) \cdot 0.071 = 0.022$.

By taking into account the similarity of research groups the level of matching the topic modeling results takes the following values:

$Fit_{sim}(Yager) = 0.441 + 0 = 0.441$;

$Fit_{sim}(Kussul) = 0.619 + 0.03 = 0.649$;
$Fit_{sim}(Bodyanskiy) = 0.498 + 0.022 = 0.52$.

7. Conclusions

We proposed researchers' topic modeling based on their interests in Google Scholar profiles. Interests in profiles researchers specify based on their discretion without using any vocabulary of keywords. In the paper, we propose an approach to categorization of such researchers using the research classification system ANZSRC. A mapping "researcher – research groups" is done using information system Dimensions that contains more than 110 millions of publications categorized according to ANZSRC.

The algorithm of researchers' topic modeling has 3 stages. The first stage forms a set of queries based on keywords and their combination. On the second stage we perform topic modeling using each query separately with filtering stop-words and underused words. On the third stage we average membership degrees of all queries and cut the distribution to a few research groups. When dropping minor research groups their contribution is redistributed to the leaders. As a result, we get membership degrees for a researcher to a few research groups that correspond to the set of his interests the most. Such mapping of interests can be viewed as an analog to the word2vec procedure.

We compared topic modeling based on small amount of information from researchers' profiles at Google Scholar with topic modeling based on a few dozens of authored publications categorized by Dimensions. As a result of comparison, we get a good matching of topic modeling results based on different sources of initial information. It allows using the proposed algorithm as the intellectual core of information technology in regards to scientific staff, in particular, for the selection of candidates as opponents of a dissertation, as reviewers for research projects, for forming a team to collaborate on shared research projects etc.

8. References

[1]. A. Martín-Martín, M. Thelwall, E. Orduna-Malea, E.D. López-Cózar, Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations, Scientometrics 126 (2021) 871–906. doi: 10.1007/s11192-020-03690-4.

[2]. A. W. Harzing, S. Alakangas, Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison, Scientometrics 106(2) (2016) 787–804. doi: 10.1007/s11192-015-1798-9.

[3]. B. Rahdari, P. Brusilovsky, D. Babichenko, E. B. Littleton, R. Patel, J. Fawcett, Z. Blum, Grapevine: A profile-based exploratory search and recommendation system for finding research advisors, Proceedings of the Association for Information Science and Technology 57(1) (2020). doi: 10.1002/pra2.271.

[4]. J. Saad-Falcon, O. Shaikh, Z. J. Wang, A. P. Wright, S. Richardson, D. H. Chau, PeopleMap: Visualization Tool for Mapping Out Researchers using Natural Language Processing, arXiv preprint (2020), arXiv:2006.06105.

[5]. M. Rosen-Zvi, T. Griffiths, M. Steyvers, P. Smith, The author-topic model for authors and documents, In Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, AUAI Press (2004) 487-494.

[6]. D. Blei, A. Ng, M. Jordan, Latent Dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993-1022.

[7]. J. Jian, G. Qian, M. Haikun, C. Chong, Author–Subject–Topic model for Reviewer Recommendation, JIS-Journal of Information Science 4 (2019). doi: 10.1177/0165551518806116.

[8]. D. Mimno, A. McCallum, Expertise modeling for matching papers with reviewers, in: Proceedings of the 13th ACMSIGKDD international conference on knowledge discovery and data mining, KDD' 07, ACM, San Jose, CA, 2007. doi: 10.1145/1281192.1281247.

[9]. C. Sun, T. J. King, P. Henville, R. Marchant, Hierarchical Word Mover Distance for Collaboration Recommender System, Springer 996 (2018) 289-302. doi: 10.1007/978-981-13-6661-1_23.
[10]. K. Xiangjie, J. Huizhen, Y. Zhuo, A. Tolba, X. Zhenzhen, X. Feng, Exploiting Publication Contents and Collaboration Networks for Collaborator Recommendation, PlosOne 11(2): e0148492 (2016). doi: 10.1371/journal.pone.0148492
[11]. Y. Zhao, J. Tang, Z. Du, EFCNN: A Restricted Convolutional Neural Network for Expert Finding, volume 11440 of Lecture Notes in Computer Science, Springer, Cham, 2019. doi: 10.1007/978-3-030-16145-3_8.
[12]. A. Omer, G. Hongyu, B. Suma, H. Wen-Mei, X. JinJun, PaRe: A Paper Reviewer Matching Approach Using a Common Topic Space, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLPIJCNLP), Association for Computational Linguistics, Hong Kong, 2019. doi: 10.18653/v1/D19-1049.
[13]. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Neural Information Processing Systems 2 (2013) 3111–3119.
[14]. S. Shtovba, M. Petrychko, Jaccard Index-Based Assessing the Similarity of Research Fields in Dimensions, CEUR Workshop Proceedings 2533 (2019) 117-128.
[15]. S. Shtovba, M. Petrychko, An Informetric Assessment of Various Research Fields Interactions on Base of Categorized Papers in Dimensions, CEUR Workshop Proceedings 2845 (2021) 159-169.