Comparing Statistical Measures for Discovering Emerging Terms in Scopus Publications in the Area of Decision Support in Smart City

Nikolay Shilov
SPIIRAS
St.Petersburg, Russia
nick@iias.spb.su

Nikolay Teslia
ITMO University
St.Petersburg, Russia
teslya@iias.spb.su

***Abstract*—Discovery of emerging research topics is an important task for scientists, conference organizers, policy-makers, and scientific foundations. The paper aims at comparative analysis of statistical models that can be used for discovering emerging terms in a corpus of documents. Three models are evaluated based on calculation of the TF*IDF, TF*PDF and Energy measures. As a case study, a corpus of abstracts of scientific publications related to decision support in smart city is used that was downloaded from Scopus for 2015-2020. The models are compared and directions of future research to improve the results, namely usage of combinations of models, analysis of synonyms, and usage of additional rules for filtering out non-emerging terms, are identified.**

I. Introduction

Literature analysis is one of the first steps of any research. When deciding on the research topic scientists study publications in a certain area and identify which research questions are currently paid the most attention and which are just starting to be touched.

Obviously, today no techniques can perform a good state of the art analysis, however, in many cases finding emerging topics in an area could be enough. For example, scientific foundations, conference or workshop organizers, policy makers would be interested to identify potentially promising research topics.

The authors of [1] note that there is no common definition of the emergence. The existing definitions are ambiguous and inconsistent. Together with the authors of [2], [3] they come to the following definition of the emerging research topic: the emerging research topic is characterized by the following attributes: (1) radical novelty, (2) relatively fast growth, (3) coherence, (4) prominent impact, and (5) uncertainty and ambiguity. Let us discuss these attributes one by one.

Radical novelty relates the notion of emergence to a certain period of time. This means that the topic can be emerging in the given period (we will refer to it as "emergence period") and not emerging in another period. Thus, the radical novelty assumes that the research topic had not been mentioned (or almost had not been mentioned) until the emergence period started.

Relatively fast growth is also intuitively clear attribute assuming that the attention paid to the topic during the emergence period growth significantly. Since "relatively" and "significantly" are ambiguous words for definitions, for testing the models we will use expert evaluations based on their experience and understanding of what the emerging term is.

Coherence refers to the ability of the research topic to separate into a cluster from research topics where it appeared. This is basically the process of forming a new research topic. Since in our research we aim at one relatively narrow domain, it is very unlikely that we can identify subdomains within it, so we cannot use this criterion. However, in this regard, we cannot say that we are discovering emerging research topics but emerging terms that do not necessarily have to form own research fields.

Prominent impact means how the research topic was accepted by the research community and it is usually measured through various bibliographic indices based on reference counts [2], [4]. This criterion is also mostly applicable to significant research topics that form own research fields during few years, since the topic has to be "invented" and proposed first, then noted by other researchers, cited in their publications that in turn have to be indexed. In order for the topic to become noticeable, this cycle has to be repeated several times that would take some years. In case of emerging term discovery this period is too long, so we will have to rely only on the number of publications considering it.

Uncertainty and ambiguity can be associated with immaturity of the emerging topic. When a new method or model emerge, it is still unclear how they can be efficiently implemented in technology, what benefits and limitations they provide. This attribute is rather a feature of the emerging topic but not a criterion for its identification.

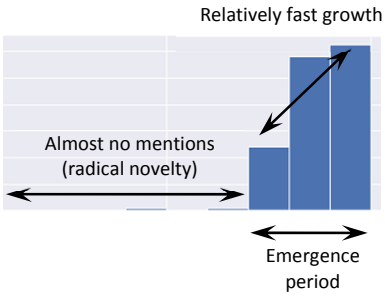


Fig. 1. Number of documents mentioning "covid19" during July 01, 2019 – April 30, 2020 as an example of emerging term

As a result, during the discovery of an emerging topic we have to rely on the first two criteria: radical novelty and relatively fast growth. Integration of these two criteria results in the dynamics of the number of mentions of the emerging term in research papers depicted in Fig. 1, where almost no mentions had been done before the February 2020, but after that the term has become highly mentioned.

The scientific question solved in the paper is what statistical measures can be used for discovering emerging topics in a corpus of documents. The study is done on an example of collection of 1725 abstracts extracted from Scopus in the area of decision support in smart city. The paper is structured as follows. The next section presents the state-of-the art review in the considered area. It is followed by the description of the procedure of forming the corpus of documents (sec. III) and the research framework description (sec. IV). Then various statistical techniques for discovering emerging topics are analyzed and compared (sec. V). Finally, the results are discussed (sec. VI) and conclusions are made (sec. VII).

II. Related Work

Researchers have been interested in discovering hot and emerging topics for some time. In 2001 Bun and Ishizuka proposed a system for emerging topic tracking in several domains [5]. They considered "the most discussed" issue as an emerging topic, which is rather a hot topic than emerging topic. They proposed a Proportional Document Frequency measure (TF*PDF) that has been used by many researchers. For example, the TF*PDF-based approach was successfully used in [6] for discovering hot terms in community question answering services.

Hot topic extraction was done in [7] on the basis of analysis of topic mentions taking into account the factor of time. The authors applied the "Energy" measure to evaluate the "importance" of a topic over different periods of time on different information channels. Unlike previous research efforts, the proposed algorithm mostly found nouns that better match the notion of "topic". Different combinations of the Energy and TF*PDF were later used for different domains, such as news streams [8], or patent libraries [9].

The problem of hot topic detection from microblogs was solved in [10] on an example of Twitter. The authors used own model, which took into consideration not only mentions of the topics but also numbers of references made to analyzed posts. Later, the analysis of cross-references for identification of hot and emerging topics was used in a number of application areas, such as scientific publications [1] or social networks [11].

Emerging topics discovery in microblogs on the example of Twitter was considered in [12] through identification of novel documents and their clustering. The analysis of clusters produced new topics that formed the "core" of the clusters. Similar approach was used in [13] and [14] but for bigger documents (not microblogs) and taking into account cross-citations.

Whereas all the above techniques either apply certain thresholds in order to separate emerging or new topics or use the "Top N" approach, a group of works can be identified, where the pre-computed characteristics are analyzed via machine learning techniques [2], [15].

III. Forming a Corpus of Documents for Studying

As a case study for the research, the topic of decision support in smart city was selected. This is a topic currently being studied under an ongoing project and its development is of interest.

Scopus is a "source-neutral abstract and citation database curated by independent subject matter experts" [16]. It indexes a significant amount of peer-reviewed publications (mostly written in English) and provides access to the bibliographic information including title, authors, abstract, publisher, as well as references in the publications and references to them. Though extensive study of literature through Scopus requires a subscription, some basic information can be extracted free of charge.

In order to get access to the publications we used PyScopus [17], which is a wrapper for Scopus API for Python language. For the purpose of this research, articles for 2015 – 2020 were extracted with the following query string (for each year):

```
ALL("smart city" AND "decision support") AND
SUBJAREA(COMP)
AND PUBYEAR IS 2020
AND (SRCTYPE(p) OR SRCTYPE(j))
```

which means that "smart city" and "decision support" were used as keywords, computer science was chosen as the subject area (we did not consider, for example, sociology or healthcare), and regular papers and journal articles as paper types (book chapters were excluded since the publication cycle for books is usually long and could create a delayed appearance of emerging terms).

For the identified period, 1727 document records meeting the above criteria were extracted together with their abstracts. Bibliographic information for two of the documents could not be accessed, so the formed corpus consists of 1725 documents. The year-based distribution of the documents in the corpus is shown in Fig. 2. One can see that the interest of the researchers in the topic identified is growing (the amount of documents in 2020 is smaller since only four months of 2020 have passed), but uneven distribution of documents requires us to use relative measures instead of absolute ones.

IV. Research Framework

The designed research framework is shown in Fig. 3. The analyzed text is based on the union of the publication title and abstract. Then, tokenization procedure is applied, which selects "tokens" ("words") in the text based on the predefined rules. In this research we eliminate all non-letter symbols (they are

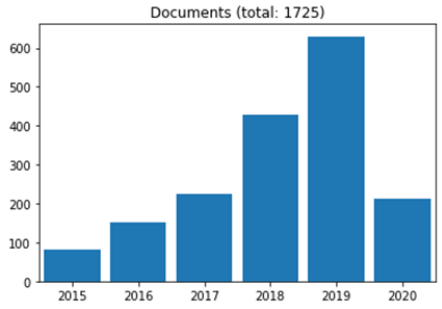


Fig. 2. Year-based distribution of the documents in the corpus

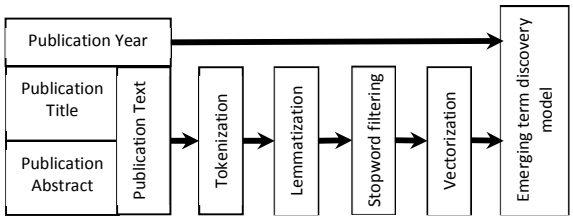


Fig. 3. Research framework for emerging term discovery

replaced with spaces)

The resulting tokens are lemmatized based on the WordNet Lemmatizer module of Python Natural Language Toolkit [18]. The lemmatization procedure determines the "lemma" (initial form) of a word. For example, "cities" is changed into "city", "men" is changed into "man". The reason of using lemmatization instead of stemming (reducing words to their base form) is that, for example, "personal" and "personalization" would both be stemmed to "person", however, from the researcher's viewpoint these are different terms as "personalization" assumes a number of technologies, whereas "personal" or "person" do not.

Stopword filtering is aimed at removal of so called "stopwords" (words that are not significant for the analysis). Usually, stopwords include articles ("a", "an", "the"), prepositions ("in", "on", etc.), conjunctions ("and", "or", etc.), pronouns ("we", "it", etc.) and other similar words.

Having the above procedures carried out for each document we get the sequence of meaningful words for them. After this, each word (or sequence of words) are considered as terms and a matrix is built, with rows being documents, columns being terms, and values being term occurrence numbers in the documents.

In this research we used single words, bi-grams and tri-grams as terms what resulted in 335 796 different terms. Then, based on the information from this matrix, different models for emerging term discovery can be applied and compared.

V. Usage of Different Statistical Models for Emerging Term Discovery

In this section we consider several techniques aimed at discovery of emerging terms based on different statistical measures.

A. TF*IDF Measure

TF*IDF (sometimes "TF-IDF" or "TFIDF") is one of the fundamental statistical measures [15], [19]. It is based on the bag-of-words scheme when the document is represented by a collection of words used in it. The main idea of TF*IDF is that the more important the term for a document is, (1) the more often it appears in it, and (2) the less often it appears in other documents.

The first condition is evaluated via TF (term frequency) measure:

$$TF_{i,j} = \frac{n_{i,j}}{N_i}, \text{ where}$$

$n_{i,j}$ – is the number of occurrences of term j in document i;

N_i – is the number of words in document i.

The second condition is evaluated via IDF (inverse document frequency) measure:

$$IDF_j = \log\left(\frac{D}{d_j+1}\right), \text{ where}$$

D – is the number of documents in the corpus;

d_j – is the number of documents where term j appears;

1 – is a constant added to avoid division by zero (different numbers are used in the literature).

Usage of the logarithm is also subject to variations in different research efforts.

One can see that variations in the calculation of IDF do not allow to perform cross-project comparisons, however, they do not significantly affect the relative results within one research.

The final measure is calculated as:

$$TFIDF_{i,j} = TF_{i,j} \cdot IDF_j$$

Since, in order to identify an emerging term, we have to analyze the radical novelty and relatively fast growth, the first model is based on comparison of the IDF measures before the given period (old) and in it (new) via dividing IDF_{old} by IDF_{new} and sorting the results descending. Then, the certain amount of top results can be sorted via, for example, average TF measure to identify the most mentioned ones. The top 20 results for 2018 and 2019-2020 are shown in Fig. 4.

It can be seen that though some "garbage" terms such as "argues", "measured", "algorithm find" are present, a number of substantial terms can be observed as well. For the further analysis we will identify terms that constitute research fields, technologies, approaches, etc. and do not contain generally

	2018	2019	2020
0	edge computing	keywords	equation modelling
1	convolutional	based machine	algorithm find
2	learning model	based machine learning	modeling sem
3	classification accuracy	mining machine	equation modeling sem
4	convolutional neural network	accepted	abc
5	measured	connects	argues
6	convolutional neural	paper new	generated content ugc
7	feature extraction	one significant	open question
8	using sensor	data mining machine	performance however
9	hierarchy process	many field	term precision recall
10	language processing	available resource	simulation change
11	quality service qos	various research	based community
12	mitigate	tested using	content ugc
13	analysis result	mining machine learning	structural equation modelling
14	natural language processing	mainly focus	internalized
15	recall	application study	ugc
16	expectation	multi aspect	approach one
17	character	different stakeholder	data sparsity
18	ai	smart contract	deep learning method
19	hash	law enforcement	comparison result

Fig. 4. Emerging terms discovered via IDF analysis

used words:

2018: “edge computing”, “convolutional neural network”, “natural language processing”, “quality of service”;

2019: “smart contract”, “law enforcement”;

2020: “structural equation modelling”, “UGC” (standing for “User-Generated Content”), “data sparsity”.

The term “ABC” had to be removed since in different documents it had different meanings, namely “Artificial Bees Colony” [20], [21] and “Agent-Based Computing” [22]. The terms “data mining machine” and “mining machine learning” appeared due to the increased usage of the phrase “data mining and machine learning”, which could not be identified because the longest word sequence analyzed was 3 words. This might be a sign of increased popularity of the joint usage of data mining and machine learning technologies, but we leave it out of the scope of this research.

To evaluate the IDF_{old} / IDF_{new} ratio avoiding division by 0, the following formula was used:

$$IDF_j^{ratio} = \frac{D_{old} \cdot d_{j,new}}{(d_{j,old}+\alpha)\cdot(D_{new}+\alpha)}, \text{ where}$$

“old” means that the value is calculated for the documents dated before the beginning of the considered period;

“new” means that the value is calculated for the documents dated within the considered period.

The variable α is used to avoid division by zero. It has to be noted that the value of this variable affects the final results. Thus, small values (e.g., 0.01) make it possible to identify terms, which had not been mentioned before and were mentioned at least once. The bigger value (e.g., 1) reduces sensitivity, so the terms mentioned in bigger amount of documents are rated higher. In this research an average value of 0.1 was used.

*B. TF*PDF Measure*

TF*PDF was proposed in [5] for defining terms that describe hot topics and it is widely used by different researchers when evaluating hot and emerging topics [6]–[9]. It is calculated by the following formula:

$$TFPDF_j = |F_j| \cdot e^{\left(\frac{d_j}{D}\right)}, \text{ where}$$

$|F_j|$ – is the normalized frequency of term j calculated as:

$$|F_j| = \frac{F_j}{\sqrt{\sum_{k=1}^{K} F_k^2}}, \text{ where}$$

F_j – is the term frequency for term j in the corpus of documents;

K – total number of terms.

As an illustration for the considered corpus of documents, the top 10 terms with the highest TF*PDF (or the most “important” terms) for 2018, 2019 and 2020 are presented in Table I. One can see that “data” and “system” are the most important terms for the field, and, for example, term “model” has been gaining importance.

Thus, our second model is aimed to check for significant growth of the TF*PDF value for a given period of time, what means that the considered term is an emerging term. In fact, this approach is not new. For example, in works [5], [6], [9], various ways of analyzing emergence of terms are considered.

We have tried the same approach as in the previous model through calculation of the ratio between the new TF*PDF (related to the considered period) and the old one (prior to the considered period):

$$TFPDF_j^{ratio} = \frac{TFPDF_{j,new} + \alpha}{TFPDF_{j,old} + \alpha}$$

The issue of the influence of the α on the result in case of TF*PDF is more important since absolute values of TFPDF are generally lower than those of IDF, so we tried 0.01, 0.1, and 1.

The top 20 results for different values of α are shown in Fig. 5. The terms identified for the further research are:

2018: “healthcare”, “disease”, “machine learning”, “neural network”, “edge computing”, “fog computing”, “diagnosis”, “sensor”, “IoT”, “service”, “traffic flow”, “traffic”, “prediction”;

2019: “machine learning”, “blockchain”, “AI” (standing for Artificial Intelligence), “ML” (standing for Machine Learning), “fog computing”, “artificial intelligence”, “disaster”, “neural network”, “game”, “security”, “traffic”;

2020: “deep learning”, “charging station”, “taxi”, “AI”, “UGC”, “spatial utilization”, “neutrosophic”.

TABLE I. THE MOST IMPORTANT TERMS FOR THE CONSIDERED DOCUMENT CORPUS

	2018		2019		2020	
	Term	TFPDF	Term	TFPDF	Term	TFPDF
1	data	0.700	data	0.606	data	0.726
2	system	0.457	system	0.422	system	0.493
3	based	0.330	based	0.339	based	0.351
4	smart	0.269	smart	0.289	model	0.157
5	paper	0.228	model	0.274	research	0.251
6	city	0.224	paper	0.248	information	0.215
7	model	0.219	research	0.224	paper	0.238
8	service	0.209	study	0.202	method	0.170
9	decision	0.201	network	0.187	study	0.136
10	information	0.194	application	0.187	proposed	0.125

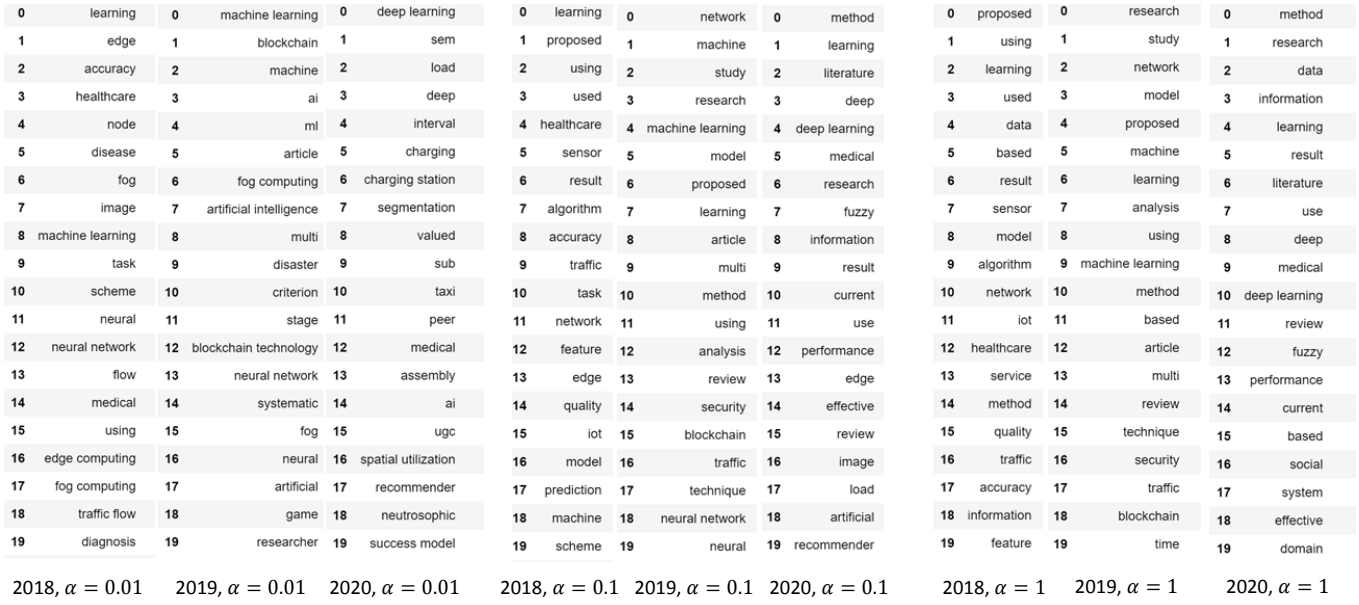


Fig. 5. Emerging terms discovered via TFPDF analysis

The term "segmentation" was not included because it was used in different contexts.

The first observation compared to the previous method is that usage of the TFPDF measure produces more substantial results and less "garbage" results. Besides, one can note that the used methodology does not pay attention to synonyms (such as "machine learning" and "ML", "artificial intelligence" and "AI") what can be attributed to one of the limitations of the current research and a subject for future improvements.

C. Energy

The third statistical measure that can be used for discovering emerging terms is Energy. This measure was originally proposed in [23] for measuring popularity of an event at different stages of its lifecycle (birth, growth, decay, and death). Then it was used in [7] for calculation of the popularity of a term. The authors of [9] and [8] also use this measure to identify whether a topic is hot or not. Unlike the two previous measures, the Energy takes into account the time period when it is measured. Thus, our third model is based on the usage of the Energy as the indicator of the term emergence.

The Energy is calculated by the following formula:

$$E_{j,\tau} = \frac{\left(A_{j,\tau} + B_{j,\tau} + C_{j,\tau} + D_{j,\tau}\right) \cdot \left(A_{j,\tau} \cdot D_{j,\tau} + B_{j,\tau} \cdot C_{j,\tau}\right)^2}{\left(A_{j,\tau} + B_{j,\tau}\right) \cdot \left(C_{j,\tau} + D_{j,\tau}\right) \cdot \left(A_{j,\tau} + C_{j,\tau}\right) \cdot \left(B_{j,\tau} + D_{j,\tau}\right)}$$

where

τ – is the time interval analyzed

$A_{j,\tau}, B_{j,\tau}, C_{j,\tau}, D_{j,\tau}$ are explained via Table II.

TABLE II. EXPLANATION OF PARAMETERS FOR ENERGY CALCULATION

	Documents within τ	Documents before τ
Documents containing term j	$A_{j,\tau}$	$B_{j,\tau}$
Documents not containing term j	$C_{j,\tau}$	$D_{j,\tau}$

The emerging terms for 2018-2020 discovered in top 20 results using the Energy measure are shown in Fig. 6. The identified emerging terms are:

2018: "neural network", "machine learning", "fog computing", "healthcare", "smart grid";

2019: "machine learning", "blockchain", "ML", "artificial intelligence";

2020: "equation modelling" (from structured or structural equation modelling), "UGC", "data sparsity".

VI. COMPARISON OF RESULTS AND DISCUSSION

Table III summarizes the results from the previous sections. It uses the following notations for models: (1)TFIDF,

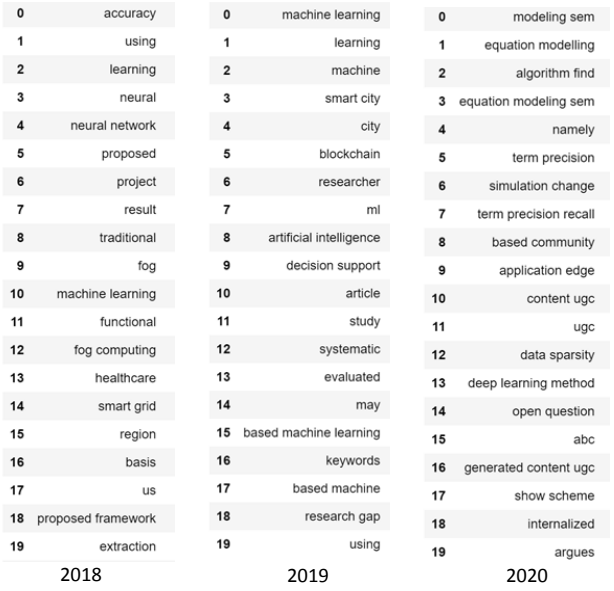


Fig. 6. Emerging terms discovered via the Energy measure

TABLE III. COMPARISON OF THE RESULTS OBTAINED VIA DIFFERENT MODELS

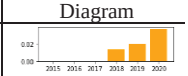
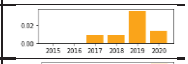
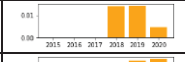
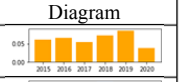
Year	2018						2019						2020						Document Frequency Diagram
Term \ Model	1	2.1	2.2	2.3	3	4	1	2.1	2.2	2.3	3	4	1	2.1	2.2	2.3	3	4	
AI						+		+				+		+				+	
artificial intelligence								+			+	+						+	
blockchain								+	+	+	+	+							
charging station														+					
convolutional neural network	+					+													
data sparsity													+				+	+	
deep learning						+								+	+	+		+	
diagnosis		+				+													
disaster								+				+							
disease		+				+	+												
edge computing	+	+				+													
equation modelling																	+	+	
fog computing		+			+	+		+											
game								+				+							
healthcare		+	+	+	+														
IoT			+	+		+													
law enforcement							+					+							
machine learning		+			+	+		+	+	+	+	+							
ML								+			+	+							
natural language processing	+					+						+							
neural network		+			+	+		+	+			+							
neutrosophic												+		+				+	
prediction			+			+													
quality of service	+					+													
security									+	+									
sensor			+	+															
service				+															
smart contract							+					+							
smart grid					+														
spatial utilization														+					
structural equation modelling													+					+	

TABLE III. COMPARISON OF THE RESULTS OBTAINED VIA DIFFERENT MODELS (CONT.)

Year	2018						2019						2020						Document Distribution Diagram
Term \ Model	1	2.1	2.2	2.3	3	4	1	2.1	2.2	2.3	3	4	1	2.1	2.2	2.3	3	4	
traffic			+	+					+	+									
traffic flow		+				+													
taxi														+				+	
UGC													+	+			+	+	

(2.1) TFPDF ($\alpha = 0.01$), (2.2) TFPDF ($\alpha = 0.1$), (2.3) TFPDF ($\alpha = 1$), (3) Energy, (4) expert evaluation. For illustrative purposes, the relative year distribution of documents $\left(\frac{d_j}{D}\right)$ is given in the last column for the year range 2015 – 2020. The expert evaluation is subjective and is based on the experts' experience. The experts had access to both relative and absolute numbers of documents per year.

Calculated first and second order errors (Table IV) show that the best result is achieved by TFPDF ($\alpha = 0.01$), however it is still relatively low. Besides, we considered the selection of emerging terms from the sets produced by models as a part of the models.

TABLE IV. COMPARISON OF THE MODELS PERFORMANCE

Model	Precision	Recall	F-Measure	Accuracy
TFIDF	0.900	0.257	0.400	0.270
TFPDF ($\alpha = 0.01$)	0.833	0.571	0.678	0.558
TFPDF ($\alpha = 0.1$),	0.545	0.171	0.261	0.244
TFPDF ($\alpha = 1$)	0.400	0.114	0.178	0.213
Energy	0.833	0.286	0.426	0.308

Possible improvements of the considered techniques could include the following:

1. Usage of combinations of techniques. Thus, for example integration of TFPDF ($\alpha = 0.01$) and TFIDF (the best combination of the studied models) produces recall 0.771, F-measure 0.806, and accuracy 0.711. However, at the same time the user will get a bigger amount of "garbage" terms, which have to be sorted out manually. The detailed analysis of possible combinations of the models is the subject of future work.

2. As it was mentioned before, the analysis of synonyms can also improve the results.

3. Additional rules can be used in order to filter results. For example, experts do not consider terms that were published in 1-2 papers as emerging.

4. Currently, only the top 20 terms returned by the models are considered. Having a dynamic value for the number of considered terms can improve the results.

These issues are also subject of future work.

VII. CONCLUSION

The contribution of the paper is the comparative analysis of techniques for discovering emerging terms in a corpus of documents based on statistical models and outlining the ways of future research to improve the results. It was found that the best technique is based on the evaluation of the ratio between the new TF*PDF measures for the considered period and before it. The F-measure for the results produced by this model is 0.68, and the accuracy is 0.59. These results do not seem to be very high, but they can be helpful in identifying emerging terms. Besides, it was also stated that the definition of term's "emergence" is subjective and its formalization has not been done in research.

Different ways of improvement have been identified: usage of combinations of models, analysis of synonyms, application of additional rules for filtering out non-emerging terms, and development of a dynamic procedure for selecting emerging terms within the results returned by models.

ACKNOWLEDGMENTS

State of the art analysis (sec. II) and experimentation results (sec. V and VI) are due to the grant from RFBR (project number 18-07-01272). Identification of documents related to decision support in smart city (sec. III) is due to State Research, project number 0073-2019-0005. The motivation (sec. I) and research framework (sec. IV) are due to the grant of the Government of Russian Federation (grant 08-08).

REFERENCES

[1] Q. Wang, "A bibliometric model for identifying emerging research topics," *J. Assoc. Inf. Sci. Technol.*, vol. 69, no. 2, pp. 290–304, Feb. 2018, doi: 10.1002/asi.23930.

[2] S. Xu, L. Hao, X. An, G. Yang, and F. Wang, "Emerging research topics detection with multiple machine learning models," *J. Informetr.*, vol. 13, no. 4, p. 100983, Nov. 2019, doi: 10.1016/j.joi.2019.100983.

[3] D. Rotolo, D. Hicks, and B. R. Martin, "What is an emerging technology?," *Res. Policy*, vol. 44, no. 10, pp. 1827–1843, Dec. 2015, doi: 10.1016/j.respol.2015.06.006.

[4] I. Jarić, J. Knežević-Jarić, and M. Lenhardt, "Relative age of references as a tool to identify emerging research fields with an application to the field of ecology and environmental sciences," *Scientometrics*, vol. 100, no. 2, pp. 519–529, Aug. 2014, doi: 10.1007/s11192-014-1268-9.

[5] K. K. Bun and M. Ishizuka, "Emerging Topic Tracking System," *Proc. - 3rd Int. Work. Adv. Issues E-Commerce Web-Based Inf. Syst. WECWIS 2001*, pp. 2–11, 2001, doi: 10.1109/WECWIS.2001.933900.

[6] Z. Zhang and Q. Li, "QuestionHolic: Hot topic discovery and trend analysis in community question answering systems," *Expert Syst. Appl.*, vol. 38, no. 6, pp. 6848–6855, Jun. 2011, doi: 10.1016/j.eswa.2010.12.052.

[7] K.-Y. Chen, L. Luesukprasert, and S. T. Chou, "Hot Topic Extraction Based on Timeline Analysis and Multidimensional Sentence Modeling,"

IEEE Trans. Knowl. Data Eng., vol. 19, no. 8, pp. 1016–1025, Aug. 2007, doi: 10.1109/TKDE.2007.1040.

[8] H.-F. Ma, "Hot topic extraction using time window," in *2011 International Conference on Machine Learning and Cybernetics*, Jul. 2011, pp. 56–60, doi: 10.1109/ICMLC.2011.6016664.

[9] K.-L. Nguyen, Byung-Joo Shin, and Seong Joon Yoo, "Hot topic detection and technology trend tracking for patents utilizing term frequency and proportional document frequency and semantic information," in *2016 International Conference on Big Data and Smart Computing (BigComp)*, Jan. 2016, pp. 223–230, doi: 10.1109/BIGCOMP.2016.7425917.

[10] Y. Chen, H. Amiri, Z. Li, and T.-S. Chua, "Emerging topic detection for organizations from microblogs," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '13*, 2013, p. 43, doi: 10.1145/2484028.2484057.

[11] T. Takahashi, R. Tomioka, and K. Yamanishi, "Discovering Emerging Topics in Social Streams via Link-Anomaly Detection," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 120–130, Jan. 2014, doi: 10.1109/TKDE.2012.239.

[12] S. P. Kasiviswanathan, P. Melville, A. Banerjee, and V. Sindhwani, "Emerging topic detection using dictionary learning," in *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*, 2011, p. 745, doi: 10.1145/2063576.2063686.

[13] W. Glänzel and B. Thijs, "Using 'core documents' for detecting and labelling new emerging topics," *Scientometrics*, vol. 91, no. 2, pp. 399–416, May 2012, doi: 10.1007/s11192-011-0591-7.

[14] H. Small, K. W. Boyack, and R. Klavans, "Identifying emerging topics in science and technology," *Res. Policy*, vol. 43, no. 8, pp. 1450–1467, Oct. 2014, doi: 10.1016/j.respol.2014.02.005.

[15] D. Kim, D. Seo, S. Cho, and P. Kang, "Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec," *Inf. Sci. (Ny).*, vol. 477, pp. 15–29, Mar. 2019, doi: 10.1016/j.ins.2018.10.006.

[16] Elsevier, "Scopus," 2020. http://scopus.com (accessed May 26, 2020).

[17] Z. Zuo, "PyScopus," 2020. http://zhiyzuo.github.io/python-scopus/ (accessed May 26, 2020).

[18] NLTK Project, "Natural Language Toolkit (NLTK)," 2020. https://www.nltk.org/index.html (accessed May 26, 2020).

[19] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF," *J. Doc.*, vol. 60, no. 5, pp. 503–520, Oct. 2004, doi: 10.1108/00220410410560582.

[20] M. Venkatesh and S. Sathyalakshmi, "Smart learning using personalised recommendations in web-based learning systems using artificial bee colony algorithm to improve learning performance," *Electron. Gov. an Int. J.*, vol. 16, no. 1/2, p. 101, 2020, doi: 10.1504/EG.2020.105253.

[21] H.-S. Chiang, A. K. Sangaiah, M.-Y. Chen, and J.-Y. Liu, "A Novel Artificial Bee Colony Optimization Algorithm with SVM for Bio-inspired Software-Defined Networking," *Int. J. Parallel Program.*, vol. 48, no. 2, pp. 310–328, Apr. 2020, doi: 10.1007/s10766-018-0594-6.

[22] C. Savaglio, M. Ganzha, M. Paprzycki, C. Bădică, M. Ivanović, and G. Fortino, "Agent-based Internet of Things: State-of-the-art and research challenges," *Futur. Gener. Comput. Syst.*, vol. 102, pp. 1038–1053, Jan. 2020, doi: 10.1016/j.future.2019.09.016.

[23] C. C. Chen, Y.-T. Chen, Y. Sun, and M. C. Chen, "Life Cycle Modeling of News Events Using Aging Theory," in *Machine Learning: ECML 2003. Lecture Notes in Computer Science*, vol. 2837, Springer, 2003, pp. 47–59.