Scientometrics (2022) 127:5005–5026
https://doi.org/10.1007/s11192-022-04449-9

Semantic Academic Profiler (SAP): a framework for researcher assessment based on semantic topic modeling

Felipe Viegas[1] · Antônio Pereira[2] · Pablo Cecílio[2] · Elisa Tuler[2] · Wagner Meira Jr.[1] · Marcos Gonçalves[1] · Leonardo Rocha[2]

Received: 2 February 2022 / Accepted: 20 June 2022 / Published online: 11 July 2022

Abstract

Recent efforts have focused on identifying multidisciplinary teams and detecting co-Authorship Networks based on exploring topic modeling to identify researchers' expertise. Though promising, none of these efforts perform a real-life evaluation of the quality of the built topics. This paper proposes a Semantic Academic Profiler (*SAP*) framework that allows summarizing articles written by researchers to automatically build research profiles and perform online evaluations regarding these built profiles. *SAP* exploits and extends state-of-the-art Topic Modeling strategies based on Cluwords considering n-grams and introduces a new visual interface able to highlight the main topics related to articles, researchers and institutions. To evaluate SAP's capability of summarizing the profile of such entities as well as its usefulness for supporting online assessments of the topics' quality, we perform and contrast two types of evaluation, considering an extensive repository of Brazilian curricula vitae: (1) an offline evaluation, in which we exploit a traditional metric (NPMI) to measure the quality of several data representations strategies including (i) TFIDF, (ii) TFIDF with Bi-grams, (iii) Cluwords, and (iv) CluWords with Bi-grams; and (2) an online evaluation through an A/B test where researchers evaluate their own built profiles. We also perform an online assessment of SAP user interface through a usability test following the SUS methodology. Our experiments indicate that the CluWords with Bi-grams is the best solution and the SAP interface is very useful. We also observed essential differences in the online and offline assessments, indicating that using both together is very important for a comprehensive quality evaluation. Such type of study is scarce in the literature and our findings open space for new lines of investigation in the Topic Modeling area.

Keywords Semantic Academic Profiler · Topic modeling · Word embeddings

Antônio Pereira, Pablo Cecílio, Elisa Tuler, Wagner Meira Jr., Marcos Gonçalves, and Leonardo Rocha have contributed equally to this work.

✉ Felipe Viegas
frviegas@dcc.ufmg.br

Extended author information available on the last page of the article

Introduction

The evolution of science has demanded solving increasingly complex problems, requiring multidisciplinary teams composed of researchers with different expertise. Although researchers usually have defined or preferred research (domains of) interests in specific points in time, these interests may evolve over time due to the rise of new lines of research, change of institutions, or even new research collaborations.

Analyzing the scientific publications of these researchers over time can help capture these evolving research interests over time. Indeed, scientific articles usually contain relevant information (e.g., title, keywords, abstracts) that can be used to infer these research interests, which may temporally shift. In this context, there is a window of opportunity for building automated solutions capable of extracting and analyzing the research domains (or research topics) of interest based on a researcher's history of publications.

An essential step towards building such solutions includes developing and exploring automatic methods for identifying the main topics of investigation explored by researchers in their publications—a set of high-quality extracted topics may shed light on the researcher's expertise. Performing such a task is not easy, though, for several reasons, one being scale. Global scientific production has presented a growth never seen before, concerning the number of published articles and the number of involved researchers. A detailed study Gusenbauer (2019) estimated an amount of 389 million documents in the Google Scholar platform in 2018. All these articles are potential sources of information that can be used as a foundation for an effective and reliable association of researchers with knowledge areas de Siqueira et al. (2020). This fact put even more demand on the proposal of automated solutions for the tasks mentioned above.

Accordingly, in this article, we propose a **S**emantic **A**cademic **P**rofiler framework (*SAP*) that allows filtering, summarizing, and analyzing a large number of articles written by researchers and published on different digital platforms, aimed mainly at extracting the main research topics of these articles. As illustrated in Fig. 1, *SAP* has four main building blocks: (i) Data Representation, (ii) Topic Modeling Decomposition, (iii) Correlation Entities, and (iv) Summarizing Interface. The goal is to build an end-to-end framework capable of extracting relevant information from the collection of scientific articles and providing a visualization interface to graphically present the extracted information, corresponding to our first major contribution.

To guarantee the scale and high quality of the extracted topics, SAP exploits state-of-the-art solutions in its main building blocks. For instance, in step (i—Data Representation), we exploit a recently proposed representation called Cluwords Viegas et al. (2019), which is currently considered state-of-the-art for topic modeling tasks Nunes et al. (2021). As one of our contributions, we extend the original Cluwords proposal by exploiting Bi-grams.

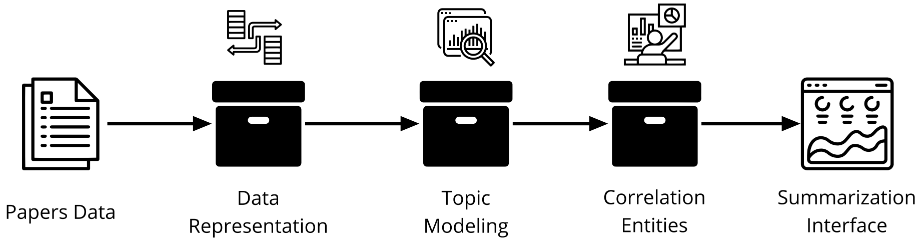


Fig. 1 Building blocks of *SAP*

This original (novel) extension is motivated by characteristics of our target task—extracting researchers' profiles over time—based on the hypothesis that many scientific concepts that capture a researcher's interest are better captured by n-gram based expressions that reflect the semantics of such topic. And as a second major contribution in this article, we also introduce a new visual interface[1] to summarize all the gathered information, highlighting the main topics related to each article, author and institution.

Thus, the two main research questions we aim to answer in this article include:

RQ1 Can *SAP* effectively summarize the profile of researchers, universities and articles?;
RQ2 Is *SAP* a useful tool for supporting **online assessments** of topic modeling strategies, complementing traditional offline assessments?

To answer these questions, we perform an extensive experimental evaluation of four SAP instantiations based on different data representations (TFIDF; TFIDF with Bi-grams; Cluwords, Cluwords with Bi-grams) and one state-of-the-art Topic decomposition strategy—Non-Negative Matrix Factorization (NMF). In our experiments, we consider a dataset composed of articles published by Computer Science Brazilian researchers in international journals written in English. To select the researchers and their articles, we take as starting point the Brazilian curricula vitae official repository (the Brazilian Lattes Platform[2]).

We divide our experiments into two parts, a quantitative analysis for answering RQ1 and a qualitative one focused on answering RQ2. The quantitative analysis is further divided into an offline and an online experiment aimed at evaluating the instantiations of the three first blocks of *SAP*, comparing four topic modeling alternatives - four document representations coupled with NMF. The offline experiment focuses on measuring traditional topic quality metrics in the literature, such as NPMI Nikolenko (2016), the main metric to measure the quality of topics without involving user interactions. For the online analysis of the SAP instantiations, we rely on an A/B testing experiment in which 12 different researchers evaluated each of the SAP instantiations (totaling 48 researchers), scoring the quality of the topics assigned to their articles, profiles, and institutions. Online experiments measure the quality of methods based on user interactions. Usually, in online experimentation, the quality metrics measure how users interact with the methods.

The results of both quantitative experiments (off and online) reinforced recent work in the literature Nunes et al. (2021), Wu et al. (2021) demonstrating the high quality of the topics built by our solutions, especially those built by our new CluWords with Bi-grams proposal. For instance, in terms of NPMI, CluWords with Bi-grams was up to 3% better than the solution based just on CluWords. Furthermore, in the performed A/B online test, the researchers scored the topics built by our proposal (CluWords with Bi-grams) 20% better than those built by the solution solely based on CluWords. These results positively answer **RQ1**. It is worth mentioning that although both experiments indicate that the CluWords with Bi-grams is the best solution, they present essential differences in evaluating the solutions, pointing out that online and offline assessments are complementary and need to be considered together for the type of scenario we are considering in this article. *Online*

[1] Available at https://labpi.ufsj.edu.br/lattestopics/-username:user-test,password:avaliacao.

[2] https://lattes.cnpq.br/.

assessment of topic modeling solutions is a research area that is highly neglected in the current literature, especially in contrast with offline evaluations.

For our qualitative assessment to answer RQ2, we focus on evaluating the summarization interface and employing online experimentation. We evaluate the usability of our visualization interface proposal using the SUS methodology Kocaballi et al. (2018) and consider the same 48 researchers also using an A/B test. In this case, the score achieved for all instantiations was above 68, which is considered an interface with excellent usability. These results, associated with the differences found between offline and online experiments, can be considered as a *third* major contribution of this article since it shows that *SAP*'s visualizations and interface are effective and valuable in the goal of complementing the traditional offline assessments, based on topic quality metrics, with online assessments, also positively answering **RQ2**.

To summarize, the major contributions of this article are:

- A general framework (*SAP*) that allows summarizing the profile of researchers, institutions and articles by analyzing articles written by researchers and published on different digital platforms.
- A visual interface that summarizes the main topics associated with articles, authors and institutes, allowing online assessments to be carried out in addition to offline assessments.
- A novel comparative study contrasting online and offline evaluation strategies for topic modeling solutions. Such type of study, especially supported by a specific visualization interface, is rare, if not absent, in the topic modeling literature.

Background and related work

Data representations for representing research articles

In spite of the type of data (e.g., title, keywords, abstracts) that we will use for representing the researchers' articles from which we will extract the topics and the researcher's interests, the common ground is that all options are textual in nature. Therefore, here we provide a basic description of the textual representations we use in our methods.

Although conceptually simple and cheap (from a computational cost perspective), the traditional TFIDF BoW representation Salton and Buckley (1988) suffers from sparseness problems, as most documents contain only a small portion of the collection's vocabulary. Moreover, this representation usually does not consider the information about the positions occupied by the words in the original document, which can often be correlated with semantic relationships among terms.

In this context, document enrichment strategies, such as n-grams (a.k.a. Bag-of-n-grams) have been adopted (Huang et al., 2018). N-grams are rudimentary models that exploit the ordering of words by capturing the local relation between them. Although N-grams may not capture complex semantic relationships among terms, their use may improve topics' quality since some are better represented with composed terms. There are other alternative representations based on co-occurrences Figueiredo et al. (2011), or in word vector representations, aka word embeddings (Mikolov et al., 2017, 2018; Pennington, et al., 2014), whose similarities correlate with semantic relatedness Bojanowski et al. (2016) or context of use Devlin et al. (2018).

Particularly, in Viegas et al. (2019), the authors proposed CluWords, a data representation that combines statistical evidence with similarity (distances) between word embeddings, being able to capture complex semantic relationships. However, since Cluwords are built employing pre-trained general word embeddings and similarity distances (e.g., cosine), they may contain semantic noise[3] such as terms that should not be grouped in specific applications (e.g., words with opposite polarities in sentiment or discriminate analysis words from different classes in text classification). To mitigate such semantic noise, there is a specific filtering mechanism in the Cluwords generation process to remove pairs of words with cosine similarity lower than a given threshold.

Part of our strategy to answer research question **RQ1** involves determining which of the above data representation alternatives is more adequate for extracting topics from research articles using some Topic Modeling strategy, discussed next.

Topic modeling strategies

Topic modeling (TM) is the main technique we exploit in SAP to infer researchers' interests over time and, as such, we briefly describe some of the main (TM) strategies next.

In a nutshell, TM is an unsupervised technique in NLP used to determine word patterns (topics) in data collection. Such topics can be explored by downstream applications to accomplish a given task. Topic Modeling Strategies can be divided into (i) Probabilistic approaches Lee and Seung (1999) and (ii) Non-probabilistic approaches (Hofmann, 1999; Allahyari & Kochut, 2016; Blei et al., 2003). The main non-probabilistic method is the Latent Dirichlet allocation (LDA) Blei et al. (2003). LDA generalizes estimating the probability distribution over terms w considering documents belonging to the abstract topic z—$P(w \mid z)$. It assumes a Dirichlet distribution—a continuous multivariate distribution that does a small set of words. On the other hand, non-probabilistic methods are based on matrix decomposition. The main method is Non-negative Matrix Factorization (NMF). Under this strategy, the input collection ($A \in R^{n\times m}$) is decomposed into two sub-matrices $H \in R^{n\times k}$ and $W \in R^{k\times m}$, such that $A \approx H \times W$. In this notation, k denotes the number of latent factors (i.e., topics), H encodes the relationship between documents and topics, and W encodes the relationship between terms and topics. Recent studies showed that NMF achieved better results when compared to LDA (Viegas et al., 2019; Nunes et al., 2021). However, Topic Modeling strategies face a major challenge when applied in a certain domain: How to represent the data that will be summarized to topics?

In Viegas et al. (2019), the authors exploited the CluWords combined with the NMF method. This combination has been evaluated in some NLP domains, such as in Crisis Events in Tweeter Nunes et al. (2021), analyzing chronic Pain using Topic Modeling Nunes et al. (2021), Covid-19 Pedro et al. (2021), programming languages evolution de Alencar Almeida et al. (2019) and understanding biases in SocialMedia Wu et al. (2021). *In this work, we extend this approach (CluWords + NMF) by exploiting Bi-grams to enrich the document representation. To the best of our knowledge, this is the first work that exploits CluWords with Bi-grams.*

[3] Semantic noise can be associated with a bias of prediction in the word-vector generation.

Topics decomposition for academic scenario

In this context, some efforts exploit the concept of topic modeling for detecting and strengthening co-Authorship Networks (Hwang et al., 2017; Krasnov et al., 2019; Hu et al., 2020). In Jeong et al. (2016), the authors proposed the Author topic flow that discovers the evolution of the authors' interest over time. In addition, the solution also discovers the temporal topic distribution for each author. In Xuan et al. (2015) designed an approach to detect an infinite number of topics for authors. The method uses a stochastic method to find the optimal number of topics from the collection and a gamma negative binomial method to learn a hierarchical structure of authors, topics and documents. In (Hwang et al., 2017; Krasnov et al. 2019) the authors exploited the LDA method to build the topics in the Co-authorship networks.

In contrast to the solutions in the literature, our solution exploits non-probabilistic methods (i.e., NMF) to detect topics through scientific publications. Non-probabilistic methods are superior for detecting topics in other topic modeling scenarios. In addition, the NMF method allows manipulations from matrices decomposition to indirectly determine correlations of entities, as we will detail in Section "Correlation entities".

Another important issue in the literature regards topic modeling evaluation. In our work, as in most in the field, we adopt offline topic evaluation metrics, such as Normalized Pointwise mutual information (NPMI) (Nikolenko, 2016; Shi et al., 2018; Viegas et al., 2019, 2020), to measure the quality of topics built based on the researcher's articles.

However, we go beyond and, differently from most works, including the ones cited above, we also perform real-life *online* evaluation regarding the quality of the built topics. *Our framework* (*SAP*) *includes a web interface so that authors may check the topics in real-time. Accordingly, we exploit the proposed interface to answer* ***RQ2*** *employing an A/B test experimental evaluation with a group of researchers to evaluate different instantiations of the SAP framework as well as the usability of SAP interface. As mentioned before, online assessments of topic modeling solutions are rarely performed in the literature, especially in contrast with offline evaluations.*

Proposed instatiations

This section presents our proposed framework to automatically extract relevant information from research articles and present them in a visual interface - *SAP*. The goal is to identify research topics covered by the articles, authors, and institutions. We decompose *SAP* into four main building blocks, namely, (i) Data Representation, (ii) Topic Modeling Decomposition, (iii) Correlation Entities, and (iv) Summarization Interface. As we shall see, proper choices for each building block significantly impact the solution's effectiveness. In this work, we go beyond the traditional evaluation of automatically extracting relevant information from research articles, in which an offline metric is used to measure the performance of the solution. The last building block of *SAP* (Summarization Interface) allows the framework's instantiation to be evaluated online through an A/B Test, gathering feedback from the researchers themselves. In Section "Data representation", we detail the data representations adopted in this work. Section "Topic modeling decomposition" briefly details the NMF decomposition matrices used to extract the latent information from research articles. In Section "Correlation entities" we present our solution to extract information from the NMF's decomposed matrices to infer the relationship between topics of authors and institutions, respectively. The use of matrix factorization to extract academic information (i.e., topics for authors) from research articles is a minor contribution of this work. Finally, in Section "Summarization interface", we present the interface that will be used to display the results obtained from the *SAP*'s instantiations.

Data representation

Data representation is one of the main steps of any NPL solution. The effectiveness of an NLP solution has a high correlation to the data representation that will serve it—data representations that capture syntactic, contextual and positional aspects of the text have already been shown to produce significant improvements in many learning tasks. Thus, several efforts in the literature investigate solutions to capture complex representations in texts. In this work, we adopt four data representation solutions—(i) TFIDF, (ii) TFIDF w/ Bi-grams, and (iii) CluWords—which have been shown to be effective in topic decomposition tasks Viegas et al. (2019), and (iv) CluWords w/ Bi-grams—a new data representation that combines the CluWords representation with Bi-grams. To the best of our knowledge, this is the first time CluWords has been exploited with Bi-grams. Since several composite terms define research topics (i.e., Machine Learning, Sentiment Analysis), we believe that Bi-grams can improve the quality of topics. The four representations are described below:

TFIDF

Bag-of-Words representation where each term composes a fixed-length vector that exploits the traditional TFIDF as a feature weighting (Equation 1).

$$TFIDF_{d,t} = TF_{d,t} * log\left(\frac{\mid \mathbb{D} \mid}{1 + d_t}\right) \tag{1}$$

TFIDF with Bi-grams

The previous TFIDF representation adding the two adjacents words from the research articles. To build the Bi-grams representation, we exploit the method Phrases from gemsim [4]. To reduce the length of the vocabulary, we ignore all bigrams with $bigramScore(w_a, w_b) < 0.5$ (Equation 2), where the function $count(\cdot)$ returns the occurrence in the collection, $min\_count = 3$, and $\mid \mathbb{V} \mid$ is the vocabulary size.

$$bigramScore(w_a, w_b) = \frac{(count(w_a, w_b) - min\_count) * \mid \mathbb{V} \mid}{(count(w_a) * count(w_b))} > threshold \tag{2}$$

CluWords

It builds a data representation by exploiting word embedding similarities, and mainly, by filtering out potential noise and properly weighting them. To build a data representation, the CluWords applies three generic steps—(i) Clustering, (ii) Filtering, and (iii) Weighting. The Clustering step exploits the nearest neighborhood strategy to capture semantic relationships between words through embedding models. The Filtering

[4] https://radimrehurek.com/gensim/models/phrases.html#gensim.models.phrases.Phrases

Table 1 Calculating author contributions to topics

	Web Applications	Computer Vision	Information Security
(a) Resulting NMF			
Article 1	30	70	10
Article 2	20	65	40
Article 3	17	80	8
(b) Author’s Pertinence by Topic			
Author 1	67	215	58
Author 2	47	150	18
Author 3	20	65	40
(c) Normalized Author’s Pertinence by Topic			
Author 1	20%	63%	17%
Author 2	22%	70%	8%
Author 3	16%	52%	32%

step filters semantic noise in the neighborhood representation built in the Clustering step. The weighting step combines the semantic information built in the first step (and filtered in the second step) with Term-Frequency (TF) Bag-of-Words representation.

CluWords with Bi-grams It differs from the CluWords in the Clustering and Weighting steps. In the Clustering step, we exploit Bi-grams to build the word embedding exploited to capture the semantic relationship among words. The Weighting step combines the semantic information with the Bi-grams representation, exploiting the Term-Frequency information.

Topic modeling decomposition

The topic modeling step is based on the Non-negative matrix factorization (NMF) strategy. NMF is one of the main topic modeling strategies that have been exploited in a wide range of domains (Pedro et al., 2021; Viegas et al., 2019; Nunes et al., 2021). The NMF Lee and Seung (1999) approach performs a “part-based” decomposition of latent relationships of a non-negative design matrix $A \in \mathbb{R}^{n \times m}$, where n is the number of articles and m the number of terms (i.e., words). The NMF has an input parameter k, which corresponds to k-dimensional approximation of A ($k \ll m$) in terms of non-negative factors $H \in \mathbb{R}^{n \times k}$ and $W \in \mathbb{R}^{k \times m}$. The matrix H encodes the relationship between documents and topics, while the matrix W encodes the relationship between terms and topics. The intuition of the NMF method is to approximate the column vectors of A by linear combinations of non-negative basis vectors (columns of H) and the coefficients given by the columns of W. However, choosing the parameter k for NMF is not a simple task since it may vary depending on the dataset. NMF is a non-convex problem with a unique solution and has no guarantee of finding the global minimum Lin (2007).

Correlation entities

For the scenarios on which our framework will be used, the resulting matrix *H* correlates topics and articles and the matrix *W* correlates the topics with the vocabulary extracted from the scientific articles dataset (words). Let us consider a dataset containing articles related to the Computer Science area with the matrix *H* resulting from NMF application as presented in Table 1(a). In this table, we have articles 1, 2 and 3 associated with the topics "Web Applications", "Computer Vision" and "Information Security". The description of topics are extracted from the matrix *W*, which encodes the relationship between terms and topics. On the other hand, the association between articles and topics are obtained from matrix *H*, which encodes the relationship between documents and topics. The articles have authors associated with them, and, consequently, we can associate the topics with the authors manipulating the matrix *H*, generating a matrix as presented in Table 1(b) and its normalized version Table 1(c). In these tables, we have authors 1, 2 and 3 associated with the topics "Web Applications", "Computer Vision" and "Information Security". In the same way, the authors work for research institutions and, consequently, we can associate the topics with the institutions manipulating the matrix that correlates authors with topics (Table 1(b)). All these manipulations correspond to the third block of our framework. In order to illustrate how our proposal performs this matrix manipulation, we will consider the example presented in Table 1.

First, to identify the main words associated with each topic, we must manipulate the matrix *W*, as widely used in literature. In our example, as previously mentioned, the three main topics are “Web Applications”, “Computer Vision” and “Information Security”. Repaying our example, we can observe the “pertinence” of each topic for each article in Table 1(a), which represents the matrix *H*. For instance, the “pertinence” of “article 1” are 30, 70 and 10 for the topics “Web Applications”, “Computer Vision” and “Information Security”, respectively. Considering that all these articles have the “author 1” as author, we calculate the “pertinence” of each topic for him/her by aggregating the “pertinence” of each article for each topic:

- “Web Applications” topic: $(30 + 20 + 17) = 67$
- “Computer Vision” topic: $(70 + 65 + 80) = 215$
- “Information Security” topic: $(10 + 40 + 8) = 58$

A similar process must be performed for all authors. Considering that “author 2” is author of articles “article 1” and “article 3”, we have:

- “Web Applications” topic: $(30 + 17) = 57$
- “Computer Vision” topic: $(70 + 80) = 150$
- “Information Security” topic: $(10 + 8) = 18$

In the same way, considering that “author 3” is author of article “article 2”, we have:

- “Web Applications” topic: $(20) = 20$
- “Computer Vision” topic: $(65) = 65$
- “Information Security” topic: $(40) = 40$

After all these calculations, we obtain the Table 1(b). Finally, based on this table, it is possible to calculate the distribution of an author among the topics by performing a normalization on the rows representing each author in Table 1(b), achieving as result Table 1(c). In this table, we have the influence of the research of each author on each topic. Observing the "author 1", his/her articles are 20% associated with the topic "Web Applications", 63% with the topic "Computer Vision", and 17% with the topic "Information Security". The articles of "author 2" are 22% associated with the topic "Web Applications", 70% with the topic "Computer Vision", and 8% with the topic "Information Security". For the "author 3", his/her articles are 16% associated with the topic "Web Applications", 52% with the topic "Computer Vision", and 32% with the topic "Information Security". As previously mentioned, as each author works for a research institution, an equivalent analysis can be performed to identify the influence of the research of each institution on each topic.

Summarization Interface

We have designed an interface that intuitively and simply the topics and their associations with researchers and institutions. Thus, all the analyses described in Section “Proposed instatiations” can be achieved quickly and efficiently. We present the data analyses associated with topics and their correlations with articles, researchers, and institutions. All those analyses have a large set of visual conceits in a web application. The web application can present similar results in different forms, including graphs, heat maps, tables, and interactive search methods. This Web App is available at our website https://labpi.ufsj.edu.br/lattestopics/ [5]. This web interface allows researchers to provide feedback on the instantiations performed. Figure 2 shows our application with one of the visual conceits provided, the one detailing the topics related to a researcher. In Section 4.2, we discuss the results regarding this figure. Other visual conceits will be presented in the next section, in which we will present possible analyses that *SAP* can perform.

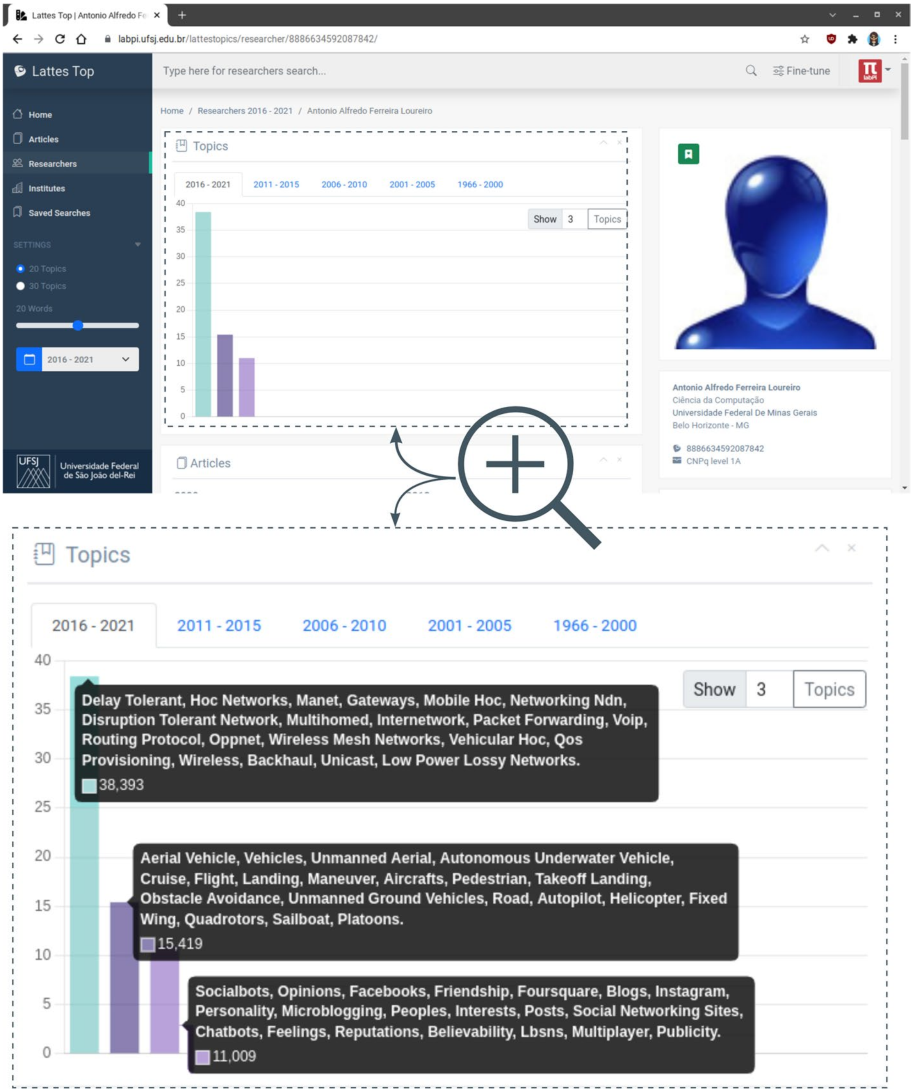
Experimental evaluation

In this section, we evaluate the effectiveness of the proposed solution. First, we describe the dataset and the preprocessing steps applied to it. Next, we report and analyze the obtained experimental results with our proposed framework. We divide our analysis into two parts, a quantitative and a qualitative one. Remind that the main research questions we aim to answer with our experiments are:

RQ1 Can *SAP* effectively summarize the profile of researchers, universities and articles?
RQ2 Is *SAP* a useful tool for supporting online assessments of topic modeling strategies, complementing traditional offline assessments?

We shall see in Section Quantitative analysis, we break our main research questions into seven sub-questions (*q1* to *q7*) that help to answer the main ones (**RQ1** and **RQ2**).

[5] username: user-test, password: avaliacao.



Lattes Top | Antonio Alfredo Fe
labpi.ufsj.edu.br/lattestopics/researcher/8886634592087842/
Lattes Top
Type here for researchers search...
Fine-tune
Home
Articles
Researchers
Institutes
Saved Searches
SETTINGS
20 Topics
30 Topics
20 Words
2016 - 2021
Universidade Federal de São João del-Rei
Home / Researchers 2016 - 2021 / Antonio Alfredo Ferreira Loureiro
Topics
2016 - 2021
2011 - 2015
2006 - 2010
2001 - 2005
1966 - 2000
Show 3 Topics
40
35
30
25
20
15
10
5
0
Articles
Antonio Alfredo Ferreira Loureiro
Ciência da Computação
Universidade Federal De Minas Gerais
Belo Horizonte - MG
8886634592087842
CNPq level 1A

Delay Tolerant, Hoc Networks, Manet, Gateways, Mobile Hoc, Networking Ndn, Disruption Tolerant Network, Multihomed, Internetwork, Packet Forwarding, Voip, Routing Protocol, Oppnet, Wireless Mesh Networks, Vehicular Hoc, Qos Provisioning, Wireless, Backhaul, Unicast, Low Power Lossy Networks.
38,393

Aerial Vehicle, Vehicles, Unmanned Aerial, Autonomous Underwater Vehicle, Cruise, Flight, Landing, Maneuver, Aircrafts, Pedestrian, Takeoff Landing, Obstacle Avoidance, Unmanned Ground Vehicles, Road, Autopilot, Helicopter, Fixed Wing, Quadrotors, Sailboat, Platoons.
15,419

Socialbots, Opinions, Facebooks, Friendship, Foursquare, Blogs, Instagram, Personality, Microblogging, Peoples, Interests, Posts, Social Networking Sites, Chatbots, Feelings, Reputations, Believability, Lbsns, Multiplayer, Publicity.
11,009

Fig. 2 Application interface

Datasets

Our dataset corresponds to one of our minor contributions[6] and is composed of articles published by Computer Science Brazilian researchers in international journals written in English. To select the researchers and their articles, we take as starting point the Brazilian

[6] Available at https://github.com/ufsj-labPi/Semantic-Academic-Profiler.

Table 2 Statistics summary table

	2021-2016	2015-2011	2010-2006	2005-2001	2000-1966	2021-1966
#authors	2,677	2,827	1,894	1,145	586	3,746
#articles (total)	15,522	12,218	6,476	3,313	2,051	39,580
#articles (mean)	5.7982	4.3219	3.4192	2.8934	3.5	10.5659
#articles (std)	7.25	5.0746	4.0497	3.3325	6.0939	15.0116
#articles (min)	1	1	1	1	1	1
#articles (median)	3	3	2	2	2	5
#articles (max)	78	51	43	39	86	199

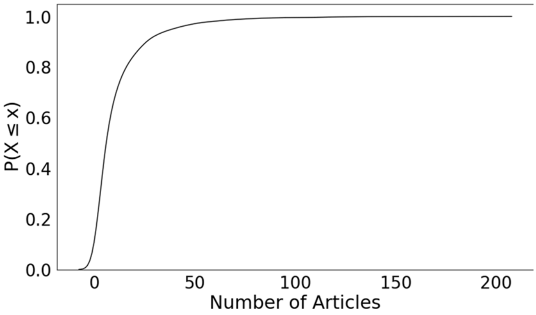


Fig. 3 Cumulative Distribution Function - Total of articles by Researcher (2021–1966)

curricula vitae official repository, named Lattes Platform[7] (Lattes). First, we collected from Lattes the title of the articles related to all Brazilian researchers with a Ph.D. in Computer Science and correlated areas. In addition to the title, we also collected the list of co-authors and consolidated them by performing a disambiguation process. Based on the titles and authors of articles, we collect the keywords and abstract of them in the Semantic Scholar[8]. For each article, we applied four preprocessing strategies: (i) uppercase to lowercase conversion Uysal and Gunal (2014), (ii) stopwords removals Gerlach et al. (2019), (iii) tokenization[9], and (iv) removal of some entities identified in the named entity recognition (NER) processGudivada and Arbabifard (2018). We exploited NER to remove noise referring to names of organizations, dates and people because these terms are not relevant in the modeling process and are not informative for defining a topic.

Table 2 shows the dataset's characteristics after the preprocessing. In our analysis, we split the database into five-time slots—2021-2016, 2015-2011, 2010-2006, 2005-2001, and 2000-1966. The last time range includes a larger number of years due to the low number of articles. We adopted this temporal division because it allows an effective modeling in terms of the number of research articles and allows the monitoring of the evolution of the research lines of the authors present in the dataset. As we can see in Table 2, our dataset is composed of 39,580 articles from 1966 to 2021, related to 3,746 researchers from 487 different institutions. It is possible to observe that researchers' total number of articles and the

[7] https://lattes.cnpq.br/.

[8] https://www.semanticscholar.org/.

[9] https://www.nltk.org/api/nltk.tokenize.html#module-nltk.tokenize.casual.

Table 3 Comparing the results achieved by each topic model solution considering 20 topics and 20 words for NPMI for different periods. The best results are in bold

Period	CluWords w/ Bi-grams	CluWords	TFIDF w/ Bi-grams	TFIDF
2021 - 2016	0.9520	0.9251	0.5151	0.5256
2015 - 2011	0.9550	0.9311	0.5228	0.5524
2010 - 2006	0.9502	0.9373	0.4958	0.5112
2005 - 2001	0.9494	0.9439	0.4827	0.4969
2000 - 1966	0.9477	0.9462	0.5025	0.5250

mean number of articles increases over time. Regarding the total of articles by researchers, in Fig. 3 we present the Cumulative Distribution Function on which we can observe that 60% of the researchers have up to seven articles considering the entire period.

Quantitative analysis

The quantitative analysis focuses on evaluating the instantiation of the three first blocks of *SAP*. As previously presented in Section Proposed instatiations, we instantiate four versions of our framework varying only the document representation strategy (e.g., TFIDF, Cluwords, TFIDF w/ Bi-grams, Cluwords w/ Bi-grams). The first two sub-questions that we want to analyze in this section are: (*q1*) *Which of the combination of data representation and topic modeling strategies is the best to represent the profile of articles, researchers and institutions?* (*q2*) *For all these cases, is there a significant difference between the alternatives?*.

In order to answer these sub-questions, first, we compare the four topic modeling solutions (document representation + NMF) considering a traditional topic quality metric in the literature Nikolenko (2016), the *pairwise point-wise mutual information (PMI)* between the top words in a topic. It captures how much one ”gains” in the information given the occurrence of the other word, taking dependencies between words into consideration. Following a recent work Nikolenko (2016), we compute a *normalized version of PMI* (NPMI), in which, for a given ordered set of top words $W_t = (w_1, ..., w_N)$ in a topic, NPMI is computed as:

$$NPMI_t = \sum_{i<j} \frac{log \frac{p(w_i,w_j)}{p(w_i)p(w_j)}}{-log\, p(w_i, w_j)} \tag{3}$$

This corresponds to the traditional offline evaluation of Topic Modeling approaches and the results are presented in Table 3, regarding different periods. Defining the number of topics and relevant keywords used to name them corresponds to the major drawback of non-probabilistic topic modeling approaches in the literature Viegas et al. (2020). Usually, these parameters are empirically defined and the quality of built topics considers a different number of topics and number of keywords. In our evaluation, we consider 20 topics, each one with 20 words. These parameters can be configured by users in SAP. We assess the statistical significance of our results by exploiting a Two-way ANOVA test with 95% confidence. As we can observe in Table 3, the best results are achieved considering the solution that combines CluWords w/ Bi-grams, followed by CluWords (until then, considered the state-of-the-art), TFIDF w/ Bi-grams and, finally, the traditional TFIDF. These results reinforce results of recent works in the literature (Nunes et al., 2021; Pedro et al.,

Table 4 Comparing the scores achieved by each strategy considering an online evaluation for the period 2021-2016. The best results are in bold

Activity	CluWords w/ Bi-grams	CluWords	TFIDF w/ Bi-grams	TFIDF
Activity 1	**275**	200	205	165
Activity 2	**275**	255	225	215
Activity 3	**210**	180	165	180
Total	**760**	635	595	560

2021; Wu et al., 2021) that demonstrate the high quality of the topics built by solutions that exploit CluWords for data representation. Moreover, these results also emphasize the improvements obtained (marked in bold) by the enhancements proposed in this paper (CluWords w/ Bi-grams) over the original proposal (up to 3%). Notice that the results with the original Cluwords are already very high (close to a perfect 1.0) and therefore are very difficult to improve.

It is very important to emphasize that the literature of topic modeling area focuses primarily (if not only) on offline evaluations, considering traditional metrics such as NPMI. Moreover, most importantly, offline experiments do not involve user interactions. Online experiments are experimental evaluations capable of measuring the quality of methods based on user interaction. Usually, in online experimentation, the quality metrics measure how the user interacts with the methods. Therefore, in this paper, we performed an A/B test (online) experiment, which we consider as an essential contribution of our work. This experiment could only be performed due to the availability of our visualization interface[10], which is another of our main contributions. More specifically, we instantiate four complete versions of *SAP*, including the summarization interface, varying just the topic modeling solutions (data representation + NMF). Twelve different researchers evaluated each of these versions scoring the quality of the topics assigned to their articles, profiles, and institutions, totalizing 48 participants from many institutions for all regions from Brazil. We kept the same proportion of researchers in the four groups according to essential characteristics, such as seniority, total number of publications, institutions, etc. Each researcher evaluated the three following activities:

- **Activity 1**: *Considering the period 2021-2016, are the three main topics assigned to your profile associated with your line of work as a researcher?*
- **Activity 2**: *Considering the period 2021-2016, are the three main topics assigned to the institution's profile associated with the line of work of its researchers?*
- **Activity 3**: *Considering the period 2021-2016, choose an article of your authorship. Are the three main topics highlighted in the article associated with the predominant theme addressed by it?*

As we can see, we focus our online experiment on the more recent research period (2021–2016). The researchers had three different scores for each activity to assign to the three main topics: (i) completely associated - 10pts; (ii) partially associated - 5pts; and (iii) not associated at all - 0pts. Each answer has a score associated with it, which we consider for evaluating the variants of our framework. To illustrate how we use these scores, consider that a researcher X evaluated Activity 1 for the TFIDF instantiation of our framework

[10] https://labpi.ufsj.edu.br/lattestopics/.

as following: Topic 1: completely associated; Topic 2: completely associated; and Topic 3: not associated. We can also suppose that another researcher Y evaluated Activity 1 for the TFIDF w/ Bi-grams instantiation: Topic 1: completely associated; Topic 2: partially associated; and Topic 3: not associated. In those cases, the total score associated by researcher X is 20 (10+10+0) and by the researcher Y is 15 (10+5+0). Based on these two researchers and considering just Activity 1, the best variant of the tool would be the TFIDF. Following this process, we summarize the score associated by all researchers (12 for each variant), for each activity/topic in Table 4.

As we can observe in Table 4, the framework instantiation that achieves the best results is the CluWords w/ Bi-grams, considering all activities (profile evaluation of researchers, institutions and articles). These results corroborate the offline evaluation presented in Table 3, besides answering sub-questions *q1* and *q2* raised at the beginning of this subsection. Considering that CluWords w/ Bi-grams was initially proposed in this paper, it can be considered as a minor contribution of this paper.

As previously mentioned, we were unable to find works in the literature that perform extensive and comparative online evaluations of topic modeling strategies, as presented in this paper, to analyze the quality of the generated topics. This quality is usually measured by exploring only traditional metrics (offline experiments) such as NPMI. Usually, offline experiments are preferred because they are less complex to be performed when compared to online experiments. Online experiments needs a framework, as the proposed SAP, that allows the users to be sorted in one of the evaluated solutions. All the evaluated solutions need user interaction and time for convergence. So, to have a large number of solutions in an online experiment implies in fewer user interacting (because there are more solution to sort the users) and more time for convergence. Thus, the idea is to reconcile both offline and online experiments. In this context, an important refinement of **RQ2** which is not consistently answered in the literature is: (*q3*)—*Is there a correspondence between offline and online assessments? What are the differences?*. Before answering the question, it is important to reinforce the need to define the correspondence between offline and online experiments, since online experiments are expensive to consider, given that participants and a period of time are needed. While in offline experiment, it needs an evaluation methodology that reflects the real world, as demonstrated in Section Quantitative analysis, in the perspective of the NPMI metric. In order to answer this sub-question, we compare the results presented in Tables 3 and 4. Considering just the results related to the period 2021-2016, which corresponds to the period evaluated in the online experiment, we can observe some interesting contrasts in the offline and online results. From the offline perspective (NPMI metric), despite the statistical difference of 3% between the CluWords w/ Bi-grams and CluWords, both can be considered almost equally good. In the online evaluation, researchers scored the topics built by our proposal (Cluwords with bi-grams) 20% better than those built by the solution based on CluWords without bi-grams. This contrast indicates that the NPMI metric captures information shared between the words in the topics built by the CluWords instantiationn, but from the users' perspective, these words (in the topics built by Cluwords without Bi-grams) do not capture well the semantics of the topics for their research interests, or at least not so as the version *with* bi-grams. Moreover, while the offline evaluation points that the TFIDF solution is slightly better them the TFIDF w/ Bi-grams, the users in the online experiment consider the opposite. Despite the difference between the use or not of the bi-grams, it is important to note that the CluWords representation performs better than TFIDF in both experiments. So, we could conclude that there is a tendency of agreement among the online and offline experiments, although they

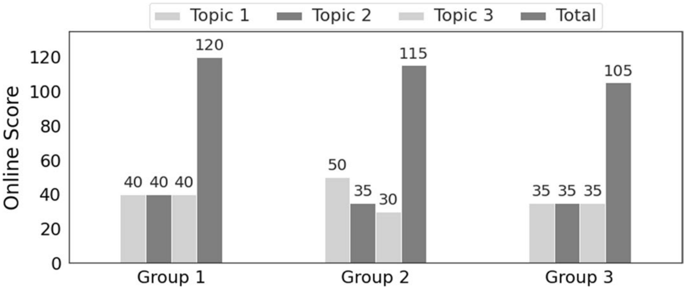


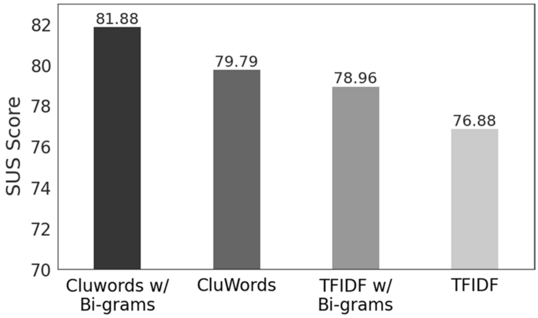
Fig. 4 Comparing the online evaluation considering groups of researchers with different number of articles

are not precise. And this is why it is relevant to evaluate in the real world (online experiments). In sum, what we can take advantage of this correspondence is the fact that offline experiments, in terms of NPMI, can be considered for preliminary evaluations to eliminate weak solutions, given the simplicity in contrast to online evaluations, and in this way, keep the most relevant solutions to be considered in online experiments. Thus, the smaller the number of solutions to consider in the online assessment, the more participants will evaluate the "competitive" solutions.

In addition to the quality issues discussed throughout this section, an orthogonal question that must be analyzed to determine the limitation of these topic modeling solutions to determine profiles from research is related to the quantitative information available. Thus, to conclude our quantitative analysis, we propose to analyze a fourth sub-question: (*q4*)—*How does the amount of information available (total of published articles) influence the identification of the researcher's profile?*. Aiming to answer this sub-question, we analyze the score assigned for researchers for the best instantiation version of our framework (CluWords w/ Bi-grams). We separate the 12 researchers that evaluated this instantiation into three groups: Group 1 composed of the four researchers with the highest number of publications (more than 12 articles in the period); Group 2 composed of four researchers with an intermediate number of publications (between six and 12 articles) and; Group 3 composed of the four researchers with the lowest number of publications (less than six articles). Now, considering these three distinct groups, we perform the same process used to generate the Table 4, also presenting a breakdown of the assessments for each topic, as we can see in Fig. 4.

As we can see, the group of researchers with the highest number of published articles was the one that best evaluated the quality of the profiles defined by SAP. Moreover, this group evaluated almost equally the quality of the three main topics, at high levels. On the other hand, the group of researchers with the lowest number of articles was the one that indicated the worst evaluation for the topics. This observation shows how important is the amount of information available about a researcher to define his/her research profile through automatic topic modeling, answering sub-question *q4*. Observing in detail these results, we realize that the main differences in assessments are in the second and third topics. It indicates that determining researchers' secondary lines of research is even more challenging when there is little information about them. All these analyses need to be taken into account by recent efforts that exploit the concept of topic modeling for detecting and strengthening co-Authorship Networks (Hwang et al.M 2017; Krasnov et al., 2019; Hu et al., 2020). We observe similar results regarding the institution's profiles, in which institutions with the highest number of researchers had their profiles better evaluated.

Fig. 5 SUS Score Average



Qualitative analysis

Our qualitative assessment focuses on evaluating the summarization interface. Through online experimentation, we evaluate the usability of the instantiations of our framework using the System Usability Scale (SUS) methodology (Kocaballi et al., 2018; Bangor et al., 2008) and consider the same 48 researchers as the previous experimentation. Our goal with this experiment is to answer two refinements of *RQ2* (fifth and sixth sub-questions): (*q5*)—*What is the researchers' assessment of the usability of our visualization interface proposal?*; (*q6*)—*Does the quality of the topics, according to the quantitative assessment, influence the users' perception of the proposal's usability?*. Thus, we asked all researchers to score the following ten questions with one of five responses that range from Strongly Agree (5 pts) to Strongly disagree (1 pts):

1. *I think that I would like to use this system frequently.*
2. *I found the system unnecessarily complex.*
3. *I thought the system was easy to use.*
4. *I think that I would need the support of a technical person to be able to use this system.*
5. *I found the various functions in this system were well integrated.*
6. *I thought there was too much inconsistency in this system.*
7. *I would imagine that most people would learn to use this system very quickly.*
8. *I found the system very cumbersome to use.*
9. *I felt very confident using the system.*
10. *I needed to learn a lot of things before I could get going with this system.*

The strategy used to calculate the SUS Score, which represents the degree of system usability, is: the results of the odd questions are taken and 1 is subtracted from each answer; for the pairs, the answer obtained is subtracted from 5. In the end, the values are added and multiplied by 2.5 to convert the original scores of 1-5 to 0-100. In Fig. 5 we present the results related to each framework version. According to SUS methodology, the score achieved for all versions in our evaluation was above 68, which is considered good usability, positively answering the sub-question *q5*. Note that the evaluation above 80.3 is considered excellent, which is the case for the framework instantiation—CluWords w/ Bi-grams.

Moreover, analyzing the differences between the SUS score for each instantiation and comparing with the results presented in Table 4, there is a clear correspondence between them, which means that the quality of the built topics directly influences the perception of usability by researchers, answering sub-question *q6*. The answer to this last sub-question is

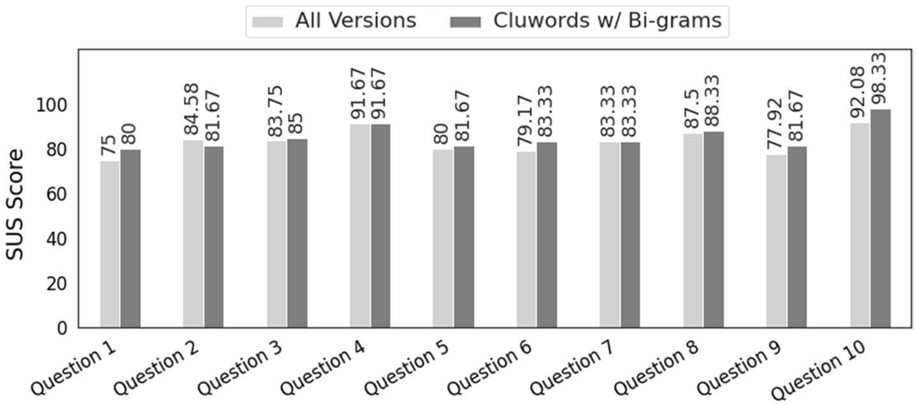


Fig. 6 Evaluating individually each question from SUS

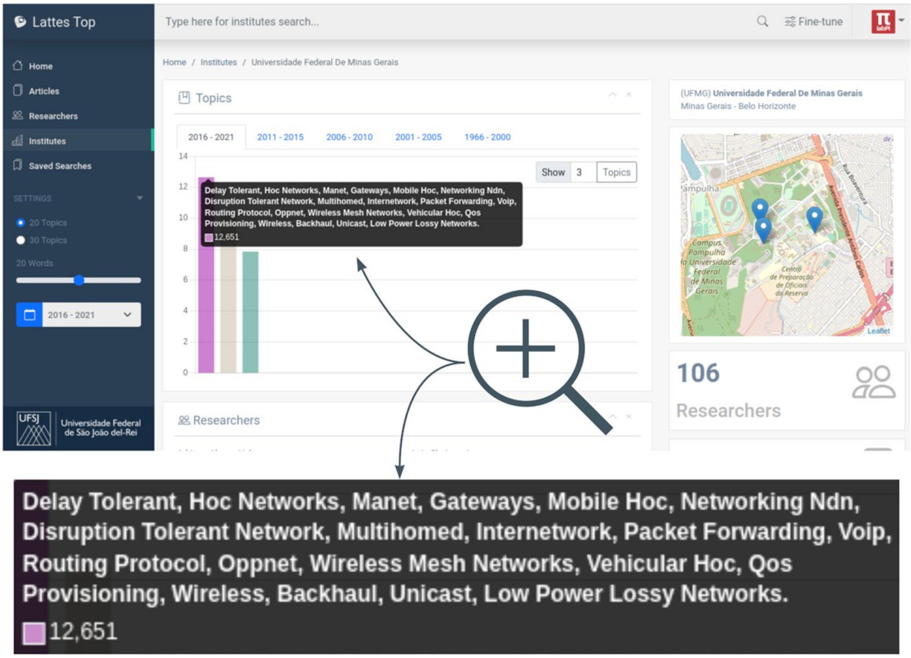
interesting and goes beyond the scope of the current work. In any case, it provides evidence that usability may be closely associated with the effectiveness of a system's purpose, but this requires further investigation in future work.

We also individually analyze the ten questions from the SUS methodology to answer yet another sub-question: (*q7*)—*Which points still need improvement?*. We considered all 48 researchers who participated in this experiment and only those who evaluated the best instantiation of *SAP*, comparing both results. The answers for each question were normalized by the highest obtainable value. Results are shown in Fig. 6. Considering all researchers, we can observe that the worst-performing questions were 1 and 9. Both are associated with the effectiveness of the proposal. On the other hand, we can observe that these questions were not an issue for the 12 researchers who evaluated the best instantiation of the framework. These results reinforce our observations that the quality of the built topics influences the perception of usability by researchers. In any case, the effectiveness is a point that can still be improved as new topic modeling strategies are proposed in the literature.

Finally, considering the data representation proposal that obtained the best results in the quantitative and qualitative assessments, we perform a complementary analysis evaluating the topics assigned to the Brazilian researcher with the highest h-index according to Google Scholar: Professor Antônio Alfredo Loureiro[11] at the Federal University of Minas Gerais (UFMG). In Fig. 2 we detail the three main topics associated with his profile from 2016 to 2021. His research is associated (at 38%) with a topic related to ad-hoc networks and correlated themes (i.e., delay-tolerant, QoS, etc.); 15% on a topic related to aerial vehicle; and 11% on a topic related to social networks. This association, automatically identified by our framework, can be corroborated by the author's description in his institutional pages [12]. We also evaluated the profile associated with his institute, UFMG, presented in Fig. 7. As we can see, the first topic is related to Professor Loureiro's research area (ad-hoc networks and correlated themes), confirming the proper functioning of our framework.

[11] https://scholar.google.com.br/citations?user=GOGlTIMAAAAJ.

[12] https://dcc.ufmg.br/professor/antonio-alfredo-ferreira-loureiro/.



Delay Tolerant, Hoc Networks, Manet, Gateways, Mobile Hoc, Networking Ndn, Disruption Tolerant Network, Multihomed, Internetwork, Packet Forwarding, Voip, Routing Protocol, Oppnet, Wireless Mesh Networks, Vehicular Hoc, Qos Provisioning, Wireless, Backhaul, Unicast, Low Power Lossy Networks.

12,651

Fig. 7 The main topic associated to profile of Federal University of Minas Gerais, Prof. Loureiro'institution

Final Remarks

In this section, we link the six sub-question (listed below) answered in the previous sections with our two main research questions—**RQ1**: *Can SAP effectively summarize the profile of researchers, universities and articles?* and, **RQ2**: *Is SAP a useful tool for supporting online assessments of topic modeling strategies, complementing traditional offline assessments?*

(q1) Which of the combination of data representation and topic modeling strategies is the best to represent the profile of articles, researchers and institutions?
(q2) For all these cases, is there a significant difference between the alternatives?
(q3) Is there a correspondence between offline and online assessments? What are the differences?
(q4) How does the amount of information available (total of published articles) influence the identification of the researcher's profile?
(q5) What is the researchers' assessment of the usability of our visualization interface proposal?
(q6) Does the quality of the topics, according to the quantitative assessment, influence the users' perception of the proposal's usability?

We can conclude that our framework can effectively summarize the profile of researchers, universities, and articles. So, in this case, we can positively answer the **RQ1** since there is evidence shown in the experimental analysis that answers the sub-questions *q1* and *q2*.

Considering offline and online experiments, we observed that the solution that combines CluWords w/ Bi-grams achieved the best results in terms of researches, universities, and articles representations, with significant differences from the other alternatives. In addition, the experimental analysis that answers the sub-questions *q5* and *q6* also showed that real users also evaluated the framework positively, according to the SUS evaluation above 80.3.

Regarding the **RQ2**, we can also conclude that SAP is a useful tool that supports both online and offline evaluation. The discussion that answers the sub-questions *q3*) and *q4* showed it is possible to correlate both offline and online experiments and, more important, to exploit both experiments in a single framework. Our framework can also be instantiated considering different data representation and topic modeling strategies. The proposed graphical interface is useful to present and display the topics generated by each instantiation, which can effectively be evaluated online, complementing the traditional offline assessments widely used in the topic modeling literature.

Conclusions and future work

We proposed a novel framework called Semantic Academic Profiler (SAP) that summarizes and analyzes articles written by researchers and published on different digital platforms. The main motivation for building SAP was to support online (and offline) assessments by the researchers themselves of the quality of the profiles generated for them. Online evaluations are rarely performed in the Topic Modeling literature and their correlation with the most commonly used offline evaluation metrics such as NPMI is very under-exploited.

SAP was built as a flexible end-to-end framework capable of automatically building research profiles for researchers and their institutions and performing online evaluations. SAP has four general main building blocks: (i) Data Representation, (ii) Topic Modeling Decomposition, (iii) Correlation Entities, and (iv) Summarizing Interface. Each of the blocks can be instantiated on its own, exploiting different strategies. To evaluate the SAP framework, we built four instantiations considering the Data Representation Block—(i) TFIDF, (ii) TFIDF w/ Bi-grams, and (iii) CluWords, and (iv) CluWords w/ Bi-grams. As for the offline evaluation, we considered the NPMI metrics. We also evaluated SAP's user interface using an A/B Test for the online evaluation using the SUS methodology. As a result, CluWords w/ Bi-grams instantiation was considered the most effective strategy in offline and online evaluations. However, both evaluations produced different results in terms of topic quality. The offline presented small gains of 3% for the best instantiation, while the online evaluation showed large improvements of about 20%. These results showed the sensitivity of building topics in real-world scenarios, motivating online evaluations to gather feedback. The results of the usability interface also showed some correlation between the quality of the interface and the quality of the topics as perceived by the researchers, further motivating the online evaluations.

As future work, we intend to extend SAP by instantiating other strategies for each block, particularly other data representations and/or topic modeling algorithms, comparing and contrasting them through online and offline experiments. We will also consider larger and more diverse datasets, with researchers worldwide and in all knowledge areas. We also intend to go deeper in the analyses of the differences and correlations between online and offline evaluations, exploring their strengths and complementarities.

Acknowledgements This work was partially supported by CNPq, CAPES, Fapemig and AWS.

References

Allahyari, M., & Kochut, K. (2016). Discovering coherent topics with entity topic models. In *2016 ACM International Conference on Web Intelligence (WI)* (pp. 26–33). https://doi.org/10.1109/WI.2016.0015

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *International Journal of Human-Computer Interaction, 24*(6), 574–594. https://doi.org/10.1080/10447310802205776.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3,* 993–1022.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. CoRR arXiv:1607.04606.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

de Alencar Almeida, R. J., Serapilha Durelli, V. H., Campos Moraes, I., Carvalho Viana, M., Carvalho Fazzion, E., Barbosa Feres Carvalho, D., Colombo Dias, D. R., & Chaves Dutra da Rocha, L. (2019). Combining data mining techniques for evolutionary analysis of programming languages. In *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)* (pp. 1–8). https://doi.org/10.1109/IRI.2019.00015.

de Siqueira, G. O., Canuto, S. D., Gonçalves, M. A., & Laender, A. H. F. (2020). A pragmatic approach to hierarchical categorization of research expertise in the presence of scarce information. *International Journal on Digital Libraries, 21*(1), 61–73. https://doi.org/10.1007/s00799-018-0260-z.

Figueiredo, F., Rocha, L., Couto, T., Salles, T., Gonçalves, M. A., & Meira, W., Jr. (2011). Word co-occurrence features for text classification. *Information Systems, 36*(5), 843–858.

Gerlach, M., Shi, H., & Amaral, L. A. N. (2019). A universal information theoretic approach to the identification of stopwords. *Nature Machine Intelligence, 1*(12), 606–612.

Gudivada, V. N., & Arbabifard, K. (2018). Chapter 3—Open-source libraries, application frameworks, and workflow systems for nlp. In: Gudivada, V. N., Rao, C. R. (Eds.), Computational analysis and understanding of natural languages: Principles, methods and applications. Handbook of statistics (Vol. 38, pp. 31–50). https://doi.org/10.1016/bs.host.2018.07.007. http://www.sciencedirect.com/science/article/pii/S0169716118300221.

Gusenbauer, M. (2019). Google scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics, 118*(1), 177–214.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 50–57).

Hu, X., Li, O. Z., & Pei, S. (2020). Of stars and galaxies—Co-authorship network and research. *China Journal of Accounting Research, 13*(1), 1–30. https://doi.org/10.1016/j.cjar.2019.09.002.

Huang, Q., Chen, Z., Lu, Z., & Ye, Y. (2018). Analysis of bag-of-n-grams representation's properties based on textual reconstruction. CoRR arXiv:1809.06502.

Hwang, S.-Y., Wei, C.-P., Lee, C.-H., & Chen, Y.-S. (2017). Coauthorship network-based literature recommendation with topic model. *Online Information Review, 41,* 318–336.

Jeong, Y.-S., Lee, S.-H., & Gweon, G. (2016) Discovery of research interests of authors over time using a topic model. In *2016 International Conference on Big Data and Smart Computing (BigComp)* (pp. 24–31). https://doi.org/10.1109/BIGCOMP.2016.7425797

Kocaballi, A. B., Laranjo, L., & Coiera, E. (2018). Measuring user experience in conversational interfaces: A comparison of six questionnaires. In *Proceedings of the 32nd International BCS Human Computer Interaction Conference*, p. 21.

Krasnov, F., Dimentov, A., & Shvartsman, M. (2019). Comparative Analysis of Scientific Papers Collections via Topic Modeling and Co-authorship Networks, pp. 77–98. https://doi.org/10.1007/978-3-030-34518-1_6.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401*(6755), 788–791.

Lin, C.-J. (2007). Projected gradient methods for nonnegative matrix factorization. Neural Comput.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2017). Advances in pre-training distributed word representations. CoRR arXiv:1712.09405.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2018). Advances in pre-training distributed word representations. In *LREC'18*.

Nikolenko, S. I. (2016). Topic quality metrics based on distributed word representations. In: SIGIR'16

Nunes, D. A. P., de Matos, D. M., Ferreira-Gomes, J., & Neto, F. (2021). Chronic pain and language: A topic modelling approach to personal pain descriptions. CoRR arXiv:2109.00402.

Nunes, D., Matos, D., Gomes, J., & Neto, F. (2021). Chronic Pain and Language: A Topic Modelling Approach to Personal Pain Descriptions. arXiv:2109.00402 .

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP*.

Pedro, A., Pereira, A., Cecilio, P., Pena, N., Viegas, F., Tuler, E., Dias, D., & Rocha, L. (2021). An article-oriented framework for automatic semantic analysis of covid-19 researches. In *Computational Science and Its Applications—ICCSA 2021* (pp. 172–187). Springer, Cham.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management, 24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0.

Shi, T., Kang, K., Choo, J., & Reddy, C. K. (2018). Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In *WWW '18* (pp. 1105–1114).

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management, 50*(1), 04–112. https://doi.org/10.1016/j.ipm.2013.08.006.

Viegas, F., Cunha, W., Gomes, C., Pereira, A., Rocha, L., & Goncalves, M. (2020). CluHTM - semantic hierarchical topic modeling based on CluWords. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, (pp. 8138–8150). https://doi.org/10.18653/v1/2020.acl-main.724.

Viegas, F., Canuto, S., Gomes, C., Luiz, W., Rosa, T., Ribas, S., Rocha, L., & Gonçalves, M. A. (2019). Cluwords: exploiting semantic word clustering representation for enhanced topic modeling, 753–761.

Wu, F., Du, M., Fan, C., Tang, R., Yang, Y., Mostafavi, A., & Hu, X. (2021). Understanding social biases behind location names in contextual word embedding models. *IEEE Transactions on Computational Social Systems*. https://doi.org/10.1109/TCSS.2021.3106003.

Xuan, J., Lu, J., Zhang, G., Yi Da Xu, R., & Luo, X. (2015). Infinite author topic model based on mixed gamma-negative binomial process. In *2015 IEEE International Conference on Data Mining* (pp. 489–498). https://doi.org/10.1109/ICDM.2015.19.

Authors and Affiliations

Felipe Viegas[1] · Antônio Pereira[2] · Pablo Cecílio[2] · Elisa Tuler[2] · Wagner Meira Jr.[1] · Marcos Gonçalves[1] · Leonardo Rocha[2]

Antônio Pereira
antoniopereira@aluno.ufsj.edu.br

Pablo Cecílio
cecilio@aluno.ufsj.edu.br

Elisa Tuler
etuler@ufsj.edu.br

Wagner Meira Jr.
meira@dcc.ufmg.br

Marcos Gonçalves
mgoncalv@dcc.ufmg.br

Leonardo Rocha
lcrocha@ufsj.edu.br

[1] Department of Computer Science, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

[2] Department of Computer Science, Federal University of São João Del Rey, São João Del Rey, Minas Gerais, Brazil