CUR: Group Profiling with Community-based Users' Representation

João Emanoel Ambrósio Gomes[1,2], **Ricardo B. C. Prudêncio**[1], **André C. A. Nascimento**[3]

[1]Centro de Informática – Universidade Federal do Pernambuco (UFPE)
Caixa Postal 7851 – 50.732-970 – Recife – PE – Brazil

[2]Campus Serra Talhada
Instituto Federal do Sertão Pernambucano (IFSertão-PE) – Serra Talhada – Brazil

[3]Departamento de Estatística e Informática,
Universidade Federal Rural de Pernambuco (UFRPE) – Recife – Brazil

{jeag,rbcp}@cin.ufpe.br, andrecamara@ufrpe.br

Abstract. *Group profiling methods aim to construct a descriptive profile for communities in social networks. Before the application of a profiling algorithm, it is necessary to collect and preprocess the users' content information, i.e., to build a representation of each user in the network. Usually, existing group profiling strategies define the users' representation by uniformly processing the entire content information in the network, and then, apply traditional feature selection methods over the user features in a group. However, such strategy may ignore specific characteristics of each group. This fact can lead to a limited representation for some communities, disregarding attributes which are relevant to the network perspective and describing more clearly a particular community despite the others. In this context, we propose the community-based user's representation method (CUR). In this proposal, feature selection algorithms are applied over user features for each network community individually, aiming to assign relevant feature sets for each particular community. Such strategy will avoid the bias caused by larger communities on the overall user representation. Experiments were conducted in a co-authorship network to evaluate the CUR representation on different group profiling strategies and were assessed by human evaluators. The results showed that profiles obtained after the application of the CUR module were better than the ones obtained by conventional users' representation on an average of 76.54% of the evaluations.*

1. Introduction

Online social networks are major platforms for production and propagation of new information, which can have its origins inside or outside the network. Also, such data can be used as a primary or complementary source to derive knowledge about the network itself, as well as its members, discussed subjects, communities, among others. Usually, such data include textual descriptions of users' interests and also their interactions inside the network, e.g., their friends, 'likes', comments, quotes, most accessed pages or channels, among others. However, this information almost always occurs as unstructured texts, i.e., in natural language or even in telegraphic or informally coded style. This configures a major challenge for automatic characterization of social communities in such networks [Getoor and Diehl 2005].

Co-authorship networks are among the most studied social networks, given its importance to understanding the structure and evolution of academic societies. A significant amount of openly accessible data about scientific publications have encouraged the use of social network analysis methods to analyze co-authorship structure. Most previous work in this context focused on the prediction of new relationships between collaborators, i.e., focused on link prediction [Lü and Zhou 2011]. The current paper focuses on a distinct task, which is community detection. Communities play a crucial role in co-authorship networks since they reflect the basis of collaboration networks among authors and research groups.

As with other social networks, co-authorship networks may contain many isolated communities, and the natural interconnection of such groups along time is rare. However, the characterization of such groups in thematic areas can enable an external agent to encourage the relations among similar groups, i.e., with the same interests, and eventually lead to new approaches to strengthen and expand scientific collaboration networks. The process of automatic extraction and selection of descriptive features from a network community is referred to as *group profiling* [Tang et al. 2008].

There are two main strategies for group profiling in a given network [Tang et al. 2008]: Aggregation-based Group Profiling (AGP) and Differentiation-based Group Profiling (DGP). In the former strategy, the characterization is done by looking for features that are most likely to occur within the group, without taking into account the rest of the network. On the other hand, the DGP strategy aims to select features which differentiate a group from the others in the network, i.e., all users of the network are considered when a specific community is described. For instance, in [Gomes et al. 2013], a DGP technique is proposed by using the Wilcoxon Rank Sum Test (WRS) to perform group profiling over numeric features (i.e., selecting features which the distribution inside a community is statistically different from the rest of the network). An extension of this method to consider textual features is proposed in [Gomes et al. 2016].

Previous studies demonstrated that AGP is more applicable in relatively noise-free environments [Tang et al. 2011]. For noisy attributes, such as user blog posts or self-reported interests, DGP techniques consistently outperform the AGP approach, nevertheless with a higher computational cost. Alternatively, the Egocentric Differentiation-based Group Profiling approach (EDGP) [Tang et al. 2011] is a variation of DGP which reduces computational cost by selecting only the community neighbors (i.e., the *fringe*[1]) in the differentiation process. EDGP achieves remarkably good results, but still less accurate than global differentiation methods.

More recently, [Gomes et al. 2018] proposed the centrality-based group profiling (CGP) approach, which applies a Centrality Filter module to select the most relevant nodes, according to their relative importance (centrality) in the observed community. The experiments demonstrated that the produced subgraphs could be much smaller, thus drastically reducing the complexity, while at the same time, retaining enough representative nodes to produce a good characterization of the group.

Overall, previous group profiling studies usually define the user representation

[1]The fringe of a community P defined as the set of all vertices, not in P, that have at least one connection to members of P.

uniformly for the entire node set, and then, apply traditional feature selection methods over user features (e.g., a bag of words for textual descriptors). However, such approach ignores the subtleties and latent group characteristics. In fact, some features may look irrelevant when observed under the user representation perspective alone but can be extremely relevant from the network perspective, describing more clearly a particular community. This fact can lead to a limited representation of some communities.

In order to address the above limitations, this work proposes a community-based user's representation module (CUR). In such proposal, feature selection algorithms are applied over user features for each network community individually, followed by an integration of all community representations in a single feature subset. The CUR method aims to assign relevant feature sets for each particular community, avoiding the bias caused by larger communities on the overall user representation.

Experiments were performed in a co-authorship network, collected from the ArXiv repository[2], in which network nodes represent authors. Content information is encoded as textual attributes extracted from the titles and abstracts of the published papers. The CUR module was evaluated under both CGP and DGP-based approaches. The final profiles assigned by the two variations of users' representation, CUR module and the conventional one, were evaluated by human annotators, resulting in a total of 230 collected responses, where each response corresponds to the profile that best represented a given community, among two distinct profiles. Results were consistently better when CUR was adopted.

The rest of this paper is organized as follows: problem statement is formally defined in Section 2. In Section 3, the module proposed in this study are described in more detail, followed by Section 4, in which the experimental methodology is presented. In Section 5, the results and the discussion are presented. Finally, Section 6 presents some conclusions and point to some future works.

2. Problem Statement

In this section, a formal description of the group profiling problem is defined. It is assumed that graph data are modeled together with content data.

Formally, the network of interest is represented as a graph $G = (V, E)$ with vertices $V = \{v_1, v_2, ..., v_n\}$ and edges $E \subseteq V \times V$. For simplicity, we assume an undirected graph without self-loops, i.e. $(v, v') \in E \iff (v', v) \in E$ and $(v, v) \notin E$. Additionally, each vertex is associated to a d-dimensional vector of features, $\mathbf{a} \in \mathbb{D}^d, \mathbf{a} = (a_1, a_2, ..., a_d)$, where $\mathbb{D}$ comprises the attribute domain (e.g., $\{0, 1\}$ or $\mathbb{R}$). For instance, a node can be associated to a set of topics of interest, which can be modeled by binary attributes: $a_j \in \{0, 1\}$, in which $a_j = 1$ indicates the user interest in the j-th topic.

A group/community is represented by a subgraph $P_i = (V_{P_i}, E_{P_i})$, where $V_{P_i} \subseteq V$, $E_{P_i} \subseteq V_{P_i} \times V_{P_i}$, $E_{P_i} \subseteq E$. For simplicity, we assume that the communities are disjoint, i.e., $G = \bigcup_i P_i$. In such setting, the characterization of a given group is defined by a subset of attributes that are relevant for the group, i.e., a vector $\mathbf{c}_{P_i} \in \mathbb{D}^k, k \leq d$. This way there is a total of $\binom{d}{k}$ distinct characterizations of size k for each group. The objective is to select the best k descriptive attributes for each group from the original d candidate

[2]https://arxiv.org/

attributes. For such, one can define a scoring function $f(a_j, P_i)$ in order to assign the importance (i.e., descriptive score) for each attribute in a given partition, and then select the top-k scored attributes.

3. Communities-based Author's Representation

In the group profiling task, the initial feature set used to represent users' content plays an important role since it defines the features that will eventually be used for the group/community characterization. Information about the users can be obtained from heterogeneous side information sources (e.g., structured or unstructured databases) or even from the network itself, and usually undergo a series of preprocessing and feature engineering steps (e.g., NLP techniques, feature selection, among others). Incorrect representation of the users can lead to an erroneous characterization, evidencing the importance of irrelevant characteristics for the groups or discarding a representative feature.

Two main types of errors can occur in the user representation phase: (1) selection of attributes that are not relevant to the domain of the underlying groups, leading to the inclusion of features without semantic values, which can be mistakenly selected to describe the communities. For example, in a community about football, consider the features about religion ('church', 'bible', among others); (2) discarding essential attributes, which leads to even worse consequences, since type (1) errors can eventually be corrected in the group characterization step, i.e., discarding irrelevant features. However, the absence of the important features (type error (2)) is an irreversible error, because the attributes will be disregarded of the process, making it impossible to includes them in the profiles. For example, in a community about football, disregard we have features relevant such as 'world cup', 'FIFA', 'Botafogo', among others.

The traditional group profiling approaches (AGP and DGP) have particular limitations. As seen, the AGP approach may assign a wrong little relevance to a feature since it analyzes it only in a local context (i.e., feature importance only inside the community).On the other hand, the DGP approach has the opposite limitation as it does not consider the internal information of the community. Additionally, DGP has a higher computational cost. To circumvent these limitations, hybrid methods which consider both high local frequency and low global frequency can be more effective.

In this work, we propose a hybrid group profiling process, in which CUR representation relies on an AGP approach, while the final generation of profiles is performed by a DGP method. Group profiling in our work is then composed of three steps:

1. Feature selection: collection, preprocessing and filtering of users' attributes, resulting in an individual representation of each community (i.e., AGP approach is applied to the representation of each community);
2. Feature integration: integration of all communities' representations in a single user representation;
3. Profiling: application of differentiation-based-based group profiling methods to identify the labels characterizing the communities (DGP approach).

CUR representation corresponds to steps 1) and 2) in the above process. Initially, all individual features for the users in the network are collected from the database followed by conventional preprocessing steps (e.g., NLP), disregarding community information.

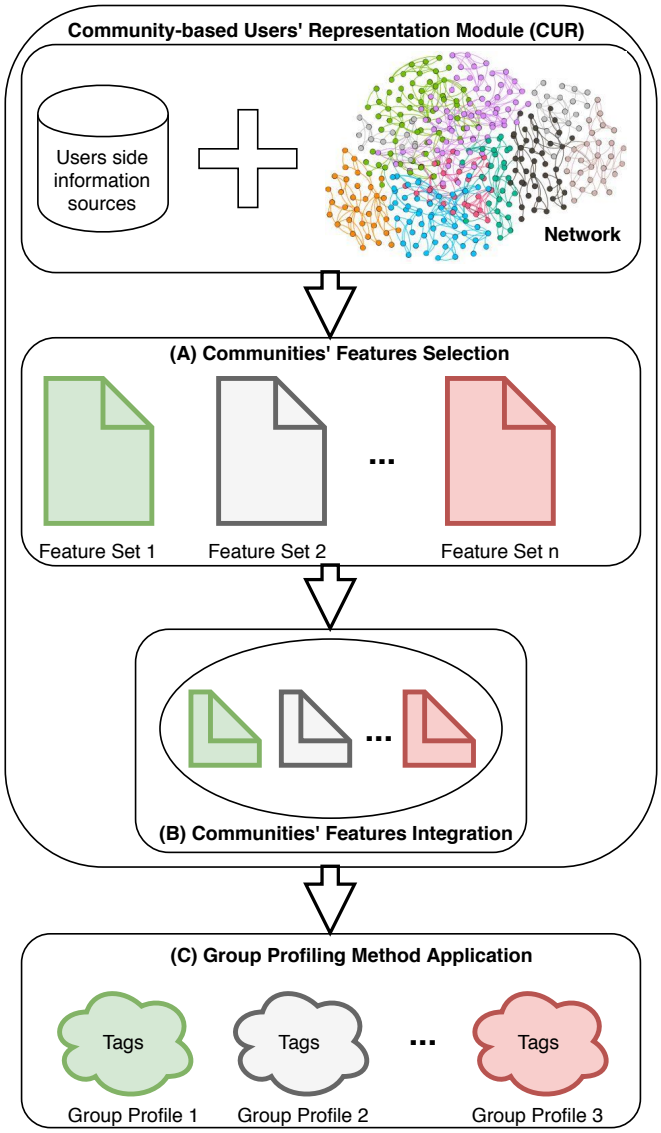

Figura 1. An overview of the Group Profiling with Communities-based Users' Representation.

Next, community labels are incorporated for the filtering/selection of attributes (Figure 1 - A). In this step, feature selection algorithms, such as stopwords, stemming, and removal of non-nouns and non-adjectives [Barrera and Verma 2012], are apply for each community in an isolated manner, selecting the top k most informative features for each group (AGP approach). The objective of this step is to generate an effective and fair feature selection for each group, disregard its relative size to the overall network size. After that, the selected features are then integrated into a single representation (Figure 1 - B), by simply concatenating the selected features (AGP approach). Finally, the differentiation-based group profiling method of choice can be applied over the final user representation (Figure 1 - C)) (DGP approach).

4. Experiment Setup

This section describes the experiment setup by detailing: the co-authorship network dataset adopted, community detection procedure, CUR representation, communities' features integration and group profiling methods application.

4.1. Data Set

As previously mentioned, to conduct a group profiling evaluation, a suite of networked data with individual attributes is necessary. For this, in our case study, we generate a co-authorship network from data collected in arXiv, maintained by Cornell University. This dataset contains millions of bibliographical records and pre-print scientific papers, mostly in mathematics, computer science, biology, finance, and statistics. In our case study, we focused on papers in the field of Artificial Intelligence, published between 2012 and 2014 (single dataset considered). In the constructed network, each node represents an author, and two nodes are connected if they have co-authored at least one paper.

4.2. Community Detection

Since no explicit community has been defined in arXiv co-authored network yet, the application of external algorithms to identify communities groups was mandatory. The Multi-level Aggregation Method (MAM) [Blondel et al. 2008] based on optimizing local modularity, was adopted. According to [Fortunato and Lancichinetti 2009], MAM is among the best-performing methods on non-directed and unweighted networks, such as co-authorship networks.

Detected groups that had fewer than ten users were eliminated since they were considered too small and irrelevant for the study. Also, the communities were filtered according to their density values, resulting in 10 remaining groups to be characterized later (global network).

4.3. Feature Selection

This step consists of the selection of the features to describe a user set. We compared the CUR approach with the conventional user representation. Both started with the same collection of authors' features. As such, published papers of each author were collected and combined into a single document. Then, a series of pre-processing steps were applied for each document: tokenization, removal of stopwords, stemming, removal of non-nouns and non-adjectives [Barrera and Verma 2012] and, finally, extraction of n-grams (with $n = \{1, 2, 3\}$).

In the CUR approach, for each community (in an isolated manner), the top k most frequent features (i.e., highest TF) for each group were selected. In this experiment we considered $k = 1,000$ for each community, resulting in a total of $\leq 10,000$ features (due to repeated terms). For the conventional user representation, $k = 10,000$, calculated over the entire network.

4.4. Feature Integration

In this step, the selected features are integrated into a single representation, by the union of all the representations from the previous step. Even though this is a rather simple approach, it avoids selecting features that are only relevant for large communities. Thus CUR

returns a balanced feature selection for each group, disregarding its relative size to the overall network size. This step is not considered in the conventional users' representation approach. The final number of features for each approach is shown in Table 1.

4.5. Group Profiling Method

In this work, we adopted a two distinct differentiation-based group profiling approaches, Global (DGP) and Centrality-based (CGP). Both were based on Inverse Document Frequency (IDF) [Maqbool and Babri 2005] since it is one of the simplest and traditional documents characterization algorithms. IDF assigns a higher weight when a term occurs in fewer communities and lower weight, otherwise. IDF can be defined by Equation 1:

$$IDF_t = \log \frac{N}{1 + df_t}, \tag{1}$$

where N is the total number of nodes in a network, and df_t is the number of nodes in the network N where the feature t is observed. Features with high IDF values can discriminate one community from the other, and therefore be used as candidates for community descriptors [Maqbool and Babri 2005].

For DGP experiments, each community is compared to the rest of the network (all nodes). In the CGP, only a fraction of the community nodes is considered, based on the node centrality. We consider the same parameters as in [Gomes et al. 2018], i.e., $n = 10$, selecting the ten nodes with the highest PageRank centrality scores in each community, to represent it in the characterization process. Thus, the subnetwork produced from the original co-authorship network is smaller. The general statistics of the resulting networks obtained using global DGP, and CGP (with PageRank) are given in Table 1.

4.6. Evaluation

As there is no guideline or gold standard for the evaluation of the detected profiles, in the performed experiments we rely on the human blind selection of the best labels generated for each group. A total of 23 people with diverse backgrounds (undergraduate, graduate students, university faculty) in the area of Computer Science/Artificial Intelligence evaluated the group profiles. Each evaluator was presented with a form containing:

1. The titles of the ten most cohesive papers[3] in the group (with a link to the abstract in arXiv web page). This selection was necessary since it is impossible for evaluators to consider all papers simultaneously[4];
2. A table with the generated profiles, i.e., the ten most representative terms, detected by each method (one per column);
3. Evaluation question: "Based on these articles, which method produced the best profile for the group?"
4. Finally, a space for the selection of the best method.

In each evaluation form, the two methods were simply denoted as "Method I"and "Method II". Also, the presentation order of group profiles was randomized for each page, to avoid the bias associated with the method names.

[3]One paper is considered cohesive if it presents high content similarity to the content found in the group.

[4]As we notice in one pilot study, subjects tend to assign random ratings if the task takes too long.

5. Experiment Results and Discussion

Experiments were conducted to analyze the effectiveness of the CUR module representation to the overall group profiling task. To achieve this, both CUR and conventional users' representation were used together with two group profiling approaches DGP and CGP.

Tabela 1. Network related measures of the global arXiv co-occurrence network and after centrality filter application

Measure	Global	PageRank
#Authors	372	100
#Links	654	157
#Groups	10	10
Link Density	0.009	0.032
Average Link	3.516	3.14
Diameter	19	16
#Conventional Features	10000	10000
#CUR Features	7.794	7.959

The generated profiles were evaluated by 23 people resulting in 230 evaluations, 115 evaluations for each considered user representation (i.e., CUR and conventional). The results, presented in Figure 2, indicated that in all communities, regardless of the adopted group profiling approach (DGC or CGP), the profiles obtained after the application of the CUR module were better (on an average of 76.54% of the evaluations). These results become more prominent for communities 256, 153, and 6, in which more than 80% of the participants pointed the CUR profiles as the best result.

In order to analyze in deeper detail the results, we present two concrete examples: groups 80 and 256. The Tables 2 and 3, presents the extracted profiles to describe these groups based on the titles and abstracts of the articles published by the authors. The labels are sorted in alphabetical order, in order to facilitate the comparison between the profiles generated by each method.

Tabela 2. Profiles for Group 80 arXiv.

Group 80	
CUR - PageRank	Conventional - PageRank
binary bayesian networks	ancestor
car comparison	bad
discovery dbcl presence	bar
edge-by-edge correction	crisis
half	discrete-valued
knowledge engineering version	half
netview subnetworks leak	illustrative
numerical probabilities	netview
qualitative belief propagation	requirmenets
SPIC	SPIC

The experiments demonstrated the group 80 as a community composed by researchers focused on the study of Bayesian networks. Terms like 'binary bayesian

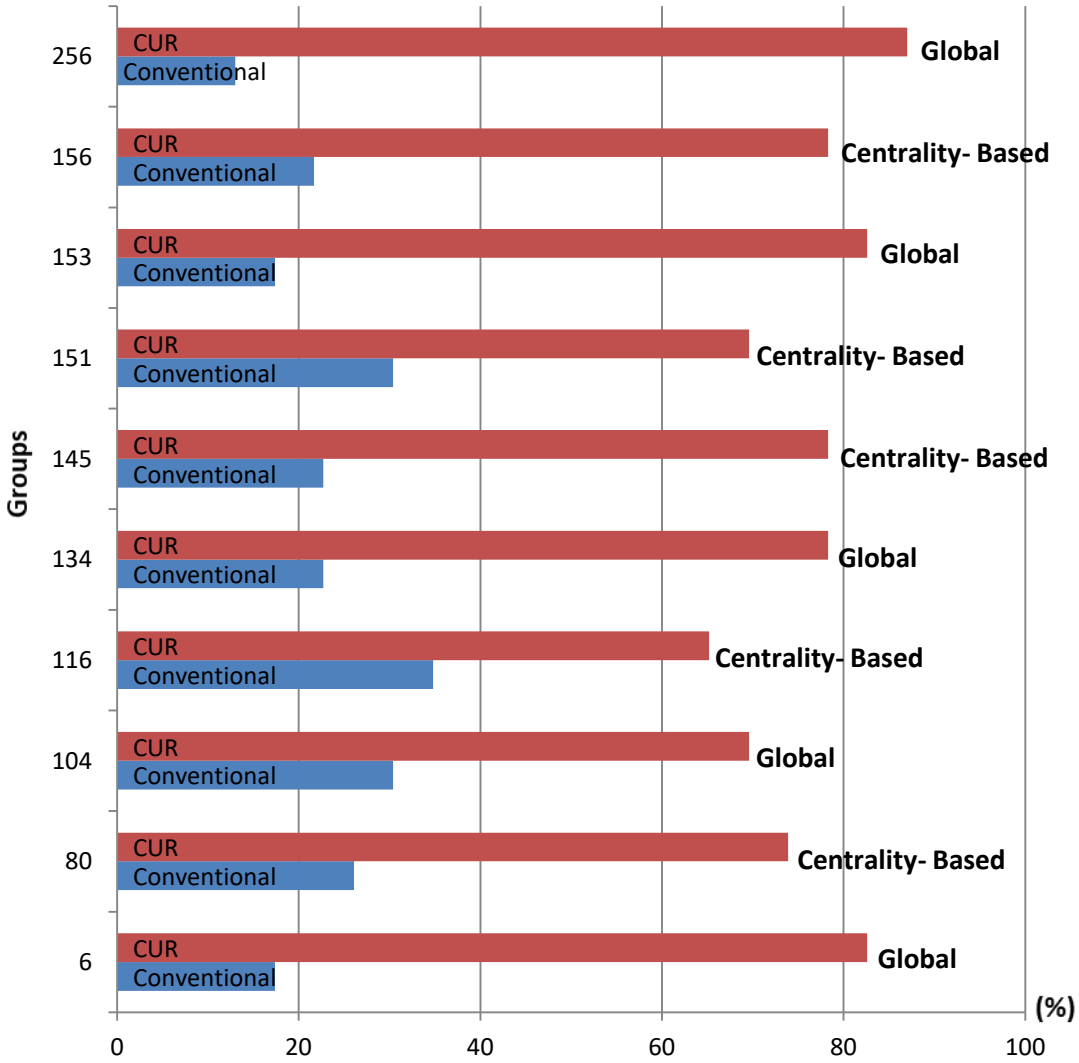

Figura 2. Percentages indicating how many times each evaluated method provided the best profile in the survey. The possible configurations are: DGP + Conventional, DGP + module CUR, CGP + Conventional, and CGP + module CUR.

networks','SPIC'[5], and 'qualitative belief propagation'. This way, the present study adheres to previous analysis [Gomes et al. 2016], i.e., indicating group 80 as a machine-learning group, based on probabilistic models, focused on the study of Bayesian networks and stochastic and dynamic learning models.

For group 256, despite the large number of indications for the CUR generated profiles, we found out that most of the suggested labels were of few descriptive importance, such as: 'real world', 'method single', 'observation construction', 'single outward propagation', 'diagrams approach', and 'tool adaptive systems'. This may have been due to the property of the IDF algorithm, which assigns greater relevance to the terms that differentiate the analyzed community from the others. That is since all authors of the network are

[5]Symbolic Probabilistic Inference with Continuous Variables - an extension of the SPI algorithm to perform the function-handle Bayesian networks with continuous linear gaussian variables.

Tabela 3. Profiles for Group 256 arXiv.

Group 256	
CUR - Global	Conventional - Global
analogous potentials cliques	appendix
diagrams approach	bags
method single	conp/poly
observation construction	distortions
posterior marginal	extent
QPNs	papamichail
qualitative reasoning	plastic
real world	pool
single outward propagation	QPNs
tool adaptive systems	spaghetti

from the area of artificial intelligence, there is the possibility that the community 256 is "eclectic", formed by members acting in different strands. This peculiarity of the community may have led the IDF algorithm to identify less descriptive labels such as the group profiling.

However, how can we justify such adherence of the evaluators to the profile generated by adopting the CUR module? Both profiles suggested labels with lew descriptive power, but one can observe the inclusion of some interesting labels in the CUR module, which would justify a choice for the "less bad" profile. Among the selected features, we highlight: 'posterior marginal' - label that defines the use of marginal posteriors in marginal likelihood estimation via importance-sampling; 'qualitative reasoning' - On reasoning in networks with qualitative uncertainty; and, 'QPNs' (Qualitative probabilistic networks) - label referent the Bayesian networks, this provides a probabilistic semantics for qualitative assertions about likelihood. It can be concluded that, despite the great importance of the users' representation in the quality assurance of the profiles, the group profiling method must be effective in identifying the relevant terms for the description of the communities.

Analyzing the profiles presented in Tables 2 and 3, with the CUR module we obtained appropriate labels not considered in the conventional representation. One can observe that most terms obtained by the conventional representation are composed by 1-grams, while terms from CUR are mostly 2 and 3-grams. This can be explained by the fact that 2 and 3-grams are less frequent when considering the entire dataset, and were probably discarded by the conventional approach, since they weren't among the 10,000 most frequent. This validates the initial hypothesis, confirming that a single look for each community, avoids the disregard of relevant attributes during the users' representing process. As well as, the experiment evidenced the effectiveness of the CUR module, considerably improving the quality of the profiles generated. However, the need for more detailed experiments is emphasized, especially with the focus on applying the CUR module together with more robust group profiling methods (e.g., Frequent and Predictive Words [Popescul and Ungar 2000] and Latent Semantic Indexing [Kuhn et al. 2007]).

6. Conclusions and Future Work

In this work, a new users' representation strategy to the sampling of relevant features in group profiling studies was presented. The proposed community-based user's representation module (CUR), introduced the use of users' feature selection for each network community individually. With the CUR module, we got relevant feature-sets for each particular community, avoiding the bias caused by larger communities on the overall user representation. Experiments with a real-world co-authorship network demonstrated that the application of the proposed representation strategy led to the balance between the conceptions of the communities, while at the same time raising the quality of the profiles generated.

To the best of our knowledge, this was also the first time that communities were considered in the users' representing process. In the context of co-authorship network, the balance between the communities can help to extract latent interests, i.e., the reasons why authors connect with each other in scientific networks. These insights may eventually lead to new approaches to strengthen and expand scientific collaboration networks.

Analyzing the results, with the CUR module we obtained appropriate labels not considered in the conventional representation. One can note that in all communities, regardless of the adopted group profiling approach (DGC or CGP), the profiles obtained after the application of the CUR module were better on an average of 76.54% of the evaluations. This validates the initial hypothesis, confirming that a single look for each community, avoids the disregard of relevant attributes during the users' representing process.

A limitation of the present work is the consideration only of the modest IDF algorithm, making it necessary to carry out a more in-depth experiment, especially with the focus on applying the CUR module together with more robust group profiling methods. However, as noted, relevant new labels were included in the profiles obtained, which proves the qualitative gains with the use of the CUR module. Another aspect subject to future work is to evaluate the use of relational information as weights during the characterization process.

Acknowledgment

The authors would like to thank CNPq (Brazilian Agency) for its financial support.

Referências

Barrera, A. and Verma, R. (2012). *Computational Linguistics and Intelligent Text Processing: 13th International Conference, CICLing 2012, New Delhi, India, March 11-17, 2012, Proceedings, Part II*, chapter Combining Syntax and Semantics for Automatic Extractive Single-Document Summarization, pages 366–377. Springer Berlin Heidelberg, Berlin, Heidelberg.

Blondel, V., Guillaume, J., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 10:8.

Fortunato, S. and Lancichinetti, A. (2009). Community detection algorithms: a comparative analysis: invited presentation, extended abstract. In *Proceedings of the Fourth*

International ICST Conference on Performance Evaluation Methodologies and Tools, VALUETOOLS '09, pages 27:1–27:2.

Getoor, L. and Diehl, C. P. (2005). Link mining: a survey. *SIGKDD Exploration Newsletter*, 7(2):3–12.

Gomes, J. E. A., Prudêncio, R. B. C., and Nascimento, A. C. A. (2018). Centrality-based group profiling: A comparative study in co-authorship networks. *New Generation Computing*, 36(1):59–89.

Gomes, J. E. A., Prudêncio, R. B. C., Meira, L., Azevedo Filho, A., Nascimento, A. C. A., and Oliveira, H. (2013). Profiling for understanding educational social networking. *Software Engineering and Knowledge Engineering (SEKE 2013)*.

Gomes, J. E. A., Prudêncio, R. B. C., and Nascimento, A. C. A. (2016). A comparative study of group profiling techniques in co-authorship networks. In *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 373–378.

Kuhn, A., Ducasse, S., and Gírba, T. (2007). Semantic clustering: Identifying topics in source code. *Inf. Softw. Technol.*, 49(3):230–243.

Lü, L. and Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A*, 390(6):11501170.

Maqbool, O. and Babri, H. (2005). Interpreting clustering results through cluster labeling. In *Emerging Technologies, 2005. Proceedings of the IEEE Symposium on*, pages 429–434.

Popescul, A. and Ungar, L. H. (2000). Automatic labeling of document clusters.

Tang, L., Liu, H., Zhang, J., Agarwal, N., and Salerno, J. J. (2008). Topic taxonomy adaptation for group profiling. *ACM Trans. Knowl. Discov. Data*, 1(4):1:1–1:28.

Tang, L., Wang, X., and Liu, H. (2011). Group profiling for understanding social structures. *ACM Trans. Intell. Syst. Technol.*, 3:15:1–15:25.