Hierarchical Expert Profiling using Heterogeneous Information Networks

Jorge Silva(✉)[1,2][0000−0002−0239−2744], Pedro Ribeiro[1,2][0000−0002−5768−1383], and Fernando Silva[1,2][0000−0001−8411−7094]

[1] CRACS & INESC TEC, Porto Portugal
[2] Departamento de Ciência de Computadores - Faculdade de Ciências, Universidade do Porto, Portugal
jmbs@inesctec.pt;{pribeiro,fds}@dcc.fc.up.pt

Abstract. Linking an expert to his knowledge areas is still a challenging research problem. The task is usually divided into two steps: identifying the knowledge areas/topics in the text corpus and assign them to the experts. Common approaches for the expert profiling task are based on the Latent Dirichlet Allocation (LDA) algorithm. As a result, they require pre-defining the number of topics to be identified which is not ideal in most cases. Furthermore, LDA generates a list of independent topics without any kind of relationship between them. Expert profiles created using this kind of flat topic lists have been reported as highly redundant and many times either too specific or too general.
In this paper we propose a methodology that addresses these limitations by creating hierarchical expert profiles, where the knowledge areas of a researcher are mapped along different granularity levels, from broad areas to more specific ones. For the purpose, we explore the rich structure and semantics of Heterogeneous Information Networks (HINs). Our strategy is divided into two parts. First, we introduce a novel algorithm that can fully use the rich content of an HIN to create a topical hierarchy, by discovering overlapping communities and ranking the nodes inside each community. We then present a strategy to map the knowledge areas of an expert along all the levels of the hierarchy, exploiting the information we have about the expert to obtain an hierarchical profile of topics.
To test our proposed methodology, we used a computer science bibliographical dataset to create a star-schema HIN containing publications as star-nodes and authors, keywords and ISI fields as attribute-nodes. We use heterogeneous pointwise mutual information to demonstrate the quality and coherence of our created hierarchies. Furthermore, we use manually labelled data to serve as ground truth to evaluate our hierarchical expert profiles, showcasing how our strategy is capable of building accurate profiles.

Keywords: Expert Profiling · Topic Modelling · Information Networks.

1 Introduction

With the current exponential growth in web-documents, the problem of linking persons to knowledge areas and vice-versa has gained a lot of attention.

This problem is known as expertise retrieval [1] and it is divided into two sub-problems: expert profiling and expert finding. The former identifies the areas of expertise of a person, while the latter finds experts in a certain topic. In literature, the expert finding task has been receiving considerably more attention than the expert profiling one. In this paper we focus on the expert profiling task. Creating accurate knowledge profiles of a person has several important applications such as [2]: categorizing personal according to their skills, identifying possible collaborations, and tracking individual or group evolution of expertise. Furthermore, the profiles generated could be used as sources of information in the expertise finding task [6, 11].

In most cases, the expert profiling problem does not have a pre-defined set of knowledge areas for the persons. Instead, they are identified in a data-driven fashion using a topic modelling approach. The Latent Dirichlet Allocation (LDA) [3] model is the most widely used strategy to define the knowledge areas/topics in text. Due to its potential, the LDA algorithm was adapted to output the distribution of authors over the discovered topics [15]. This discovery fostered the development of a group of algorithms named Author-Topic models that, not only identify topics in documents, but also profile the author's expertise. Since then, several other Author-Topic models have been proposed [7, 10, 12].

The core of the Author-Topic models is the LDA algorithm and despite it being widely used, there are some known flaws in it [8]: lacks an intrinsic methodology to choose the number of topics, contains several hyper-parameters that can cause overfitting, and it is incompatible with properties of text such as Zipf's law for the frequency of words. In order to avoid these flaws, we propose a different strategy to the topic modelling part. The vast number of Author-Topic models that exist in literature, indicate that adding external sources of information besides text, improves the quality of expert profiles. Therefore, we use documents' meta-data to model their inter-relations in a Heterogeneous Information Network (HIN), and we uncover hidden structures in the linked data that represent topics/knowledge areas which can be used to categorize a person's knowledge. An advantage of this process when compared to LDA is that it does not require defining the number of topics to be discovered.

With respect to the expert profiling task, experts have reported that the profiles assigned to them are redundant, and either too general or too specific [2]. This occurs because the expert profiles are generated from a flat list of topics without any relation between them. A solution to the problem is to create an hierarchy of topics with "sub-topic of" relations. Unfortunately, automatically creating these structures and mapping experts into them is not trivial [16, 21]. In this work, we take advantage of the HIN to organize the topics discovered in an hierarchy and to map the experts into the topics. As a result, we are capable of creating hierarchical profiles that on top represent broad knowledge areas and on the bottom more specific ones. Figure 1 illustrates the differences between a flat and an hierarchical profile.

This paper is structured as follows. In Section 2 we discuss the related word for the topic modelling and expert profiling tasks. In Section 3 we formalize the

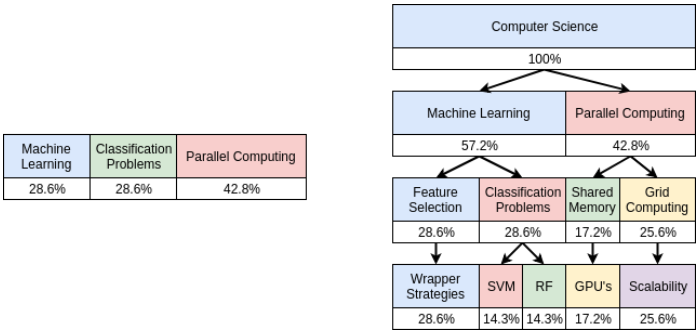

Fig. 1: Example of flat versus hierarchical organization of topics of expert profiling.

task of creating an hierarchical expert profile from an heterogeneous information network. In Section 4 we describe our model and in Section 5 we evaluate the topic modelling and the expert profiles constructed. Finally, in Section 6 we present the conclusions and address future work.

2 Related Work

In the expert profiling domain, the Author-Topic models are widely used for the task. These models are inspired by the Latent Dirichlet Allocation (LDA) algorithm [3] which represents topics as a multinomial mixture over words, and documents as a multinomial distribution over topics. In 2004, Rosen-Zvi et al. [15] added the authors distribution of documents to the LDA model, thus creating the first Author-Topic model and fostering the motivation to several other ones. Tang et al. [20] unveiled the importance of adding the conference distribution to the author-topic models. Later, Wang et al. [22] proposed the Author-Conference-Topic-Connection model which besides adding the conference distribution, also adds the subjects of the conferences. In 2012, Daud [5] added the documents timestamps and proposed the Temporal-Author-Topic which models the topic distribution of an author over time. Later, Jeong et al. [10] proposed the Author-Topic-Flow which allows each author to directly have a temporal pattern of expertise. Duan et al. [7] explored the community information in networks, and proposed the Mutual Enhanced Infinite Community-Topic model which finds communities and the topics they discuss in text-augmented social networks. This work was the pioneer in simultaneously integrating community discovery with topic modelling, while considering communities and topics as different latent variables (i.e. a community may be interested in several topics).

There are some works in literature that rely on information networks to avoid the problems of the LDA model. Gerlach et al. [8] represented a word-document matrix as a bipartite network, and reformulated the problem of topic modelling as the task of finding communities in such network. The authors proposed the hierarchical Stochastic Block Model (hSBM) which is a probabilistic inference approach that is capable of handling the possibility of higher-order

structures. Consequently, the algorithm is capable of generating an hierarchy of topics. Some different approaches that focus on topic modelling using HINs have been proposed. Rankclus [18] was a pioneer algorithm that simultaneous clusters and ranks nodes in a HIN using a generative model that operates on bipartite topologies. Netclus [19] emerged later with the intent to extend the Rankclus to HINs with a star-topology. More recently, CATHYHIN [21] extended the previous algorithms to support the following features: ranked list of attributes for each type along with a ranked list of phrases, any HIN topology, soft-clustering of all the nodes, and developing an hierarchy of topics. With respect to this work, CATHYHIN produces a similar output to our algorithm (i.e. an hierarchy of topics where each topic consists of multiple node types, see Figure 2 for an illustration.). However there are two main differences in our work. To start CATHYHIN uses a generative model to discover the communities while we use modularity optimization. Additionally, CATHYHIN focus on discovering topics in an HIN. We extend this goal and we define strategies to map the experts into the discovered topics to create their hierarchical expert profile.

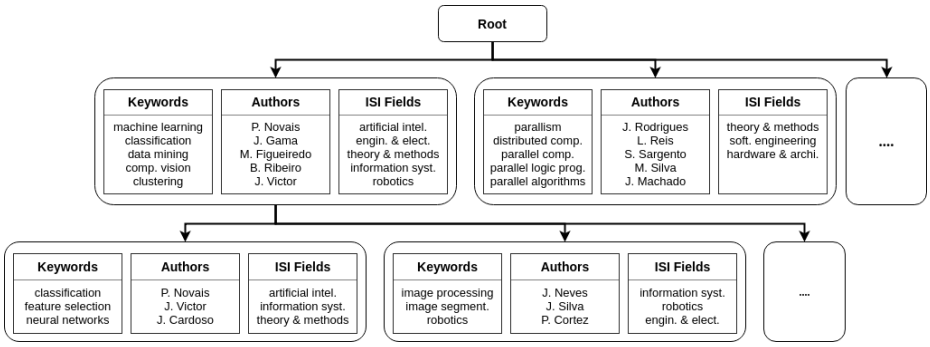

Fig. 2: Sample of the hierarchy of topics obtained from our algorithm.

In literature, there are a few works that create a expertise profile with hierarchical properties. Bin et al. [9] uses explicit feedback from persons and their bookmarks information to extract keywords that reflect their expertise. Afterwards, these keywords are mapped into a pre-defined ontology. Thus constructing an hierarchical profile. Rybak et al. [16] uses publication's meta-data to maps authors into the ACM computation classification system.[3] Since this is organized in hierarchies, the expert profile is also hierarchical. An important aspect of both strategies is the fact that they use a manually created hierarchy which requires a lot of human effort. Moreover, these structures are dynamic. As a result this is not a one time task [21]. Additionally, there is the problem of mapping the expert's knowledge into the hierarchy. In [9] the authors have to restrict the keywords to the ones that are on the ontology. On the other hand, Rybak [16] restricts the author's publication to the ones published in ACM conferences. Both strategies potentially leave out details that may be relevant to characterize the experts' knowledge. In this work we automatically create the topological hierarchy, and

[3] https://www.acm.org/publications/class-2012

since our topics consist of multiple entities, we are capable of mapping experts directly (the author is part of the topics) and indirectly (author is represented by other meta-data attributes) into the hierarchy.

3 Problem Description

We formalize the problem of creating hierarchical profiles for experts as the task of receiving an HIN, generating a topical hierarchy and the mapping expert's knowledge into that structure.

Definition 1. *An information network is defined as a directed graph $G = (N, L)$ with an object type mapping $\psi : N \to A$ and a link type mapping $\varphi : L \to R$. Each node $n \in N$ belongs to an object type $a : \psi(n) \in A$. Furthermore, each link $l \in L$ belongs to a relation type $r : \varphi(l) \in R$. If two links share the same relation type, they both start at a node with type a' and end at a node with type a''.*

An HIN is a type of information network where $|A| > 1$ and/or $|R| > 1$. For a better understanding of the object types and relations, HINs have a meta-level description named network-schema [17].

Definition 2. *We define a topical hierarchy as a tree T where each node is a topic. Each topic t contains $|A'|$ lists of ranked attributes where $A' \subseteq A$ and A is the set of object types in the HIN.*

Definition 3. *An hierarchical expert profile P is a tree such that $P \subset T$. Each $t \in P$ contains a q indicating the percentage of knowledge of the expert on that topic. Additionally, $\forall l \in L, \sum_{t \in P_l} t_q = 1$, where L is the number of levels in the tree and P_l is the set of topics at level l.*

Our proposed model is divided into two parts. The first consists in defining a function θ such that $\theta(G) = T$. Then, we introduce two strategies to create a function λ such that $\lambda(T, e) = P_e$, where e is an expert and P_e his hierarchical expert profile. We address the construction of both functions in the next section.

4 Hierarchical Expert Profile

4.1 Network Construction

The model proposed in this work can be applied to any HIN. However, to ease understanding we present the discussion and evaluation of its components in the context of bibliographic databases. More concretely, we use data from Authenticus[4] which is a bibliographic database for the Portuguese researchers. To construct the HIN we select a set of publications and, for each one, we query the

[4] https://www.authenticus.pt

database for the following meta-data: authors, keywords and ISI fields[5]. Then, the HIN is constructed following a star-schema topology where publications are the star-nodes, and authors, keywords and ISI fields are the attribute-nodes (see Figure 3 for an illustration). There are three different types of relations: publication-author, publication-keyword and publication-ISI field. Each relation has a different W_x that represents the importance of objects of type x in the network. The W_x values are normalized with respect to the number of attributes x connected to the star-nodes (in this case publications). For example, considering that W_a is the publication-author's weight, all the n authors of a certain publication p have a link weight of $\frac{1}{n}W_a$.

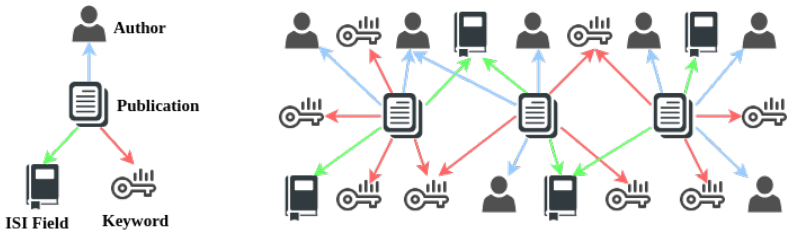

Fig. 3: Network scheme of our proposed bibliographic HIN.

4.2 Topic Modelling

Once we have an HIN we apply a modularity optimization algorithm to unveil communities on the network structure. We assume that the communities represent topics/knowledge areas for the expert profiling task. Given a network community c, modularity [14] estimates the fraction of links within c minus the expected fraction if links were randomly distributed. The value of modularity ranges between -1 and 1. Positive values indicate that the number of links in c, exceeds the number of expected ones at random. A modularity based community detection algorithm aims to maximize the global modularity of the communities in the network. However, due to the time complexity of the task, algorithms must use some heuristics in order to decrease its computational cost. In this work we use Louvain algorithm [4] which is a greedy optimization method with expected runtime $O(n\ \log(n))$, where n is the number of nodes in the network.

With respect to our overall goal of topic modelling in HINs, using Louvain algorithm presents some drawbacks: does not account for nodes and links heterogeneity, ignores network-schema, and produces non-overlapping communities. The first two points lead to a loss of information in the HIN. The latter produces the undesired effect of hard-clustering attribute-nodes (by intuition, some authors/keywords should be part of more than one community). In order to tackle these problems, before applying the Louvain algorithm to detect communities we adapt our HIN to a similarity graph of star-nodes $G' = (N', L')$. In case of our

[5] Research areas created by the Institute for Scientific Information.

bibliographic HIN, all the nodes in G' are publications and the links represent how related two publications are.

The process to construct G' starts with the selection of all the star-nodes from the HIN. Each one represents a different node in G'. The edge weights between every pair of nodes $(p1, p2) \in L'$ are defined by the following formula:

$$l_{p1,p2} \in L' = \sum_{n \in K} l_{p1,n} + \sum_{n \in K} l_{p2,n} \tag{1}$$

where K is the set of nodes that are adjacent to $p1$ and $p2$ in the HIN, and $l_{x1,x2}$ is the edge weight between nodes $x1$ and $x2$.

After the construction of the similarity graph we apply the Louvain algorithm which returns a community partition C that maps nodes into their respective community. Extrapolating C to the HIN, we obtain the community membership of all the star-nodes. On the next step, we expand these communities in the HIN to assign community membership to the attribute-nodes. Due to our star-schema topology, every attribute-node a is connected to at least one star-node p, that belongs to a community $c_j \in C$. Therefore, we estimate the community membership of attribute-nodes as the fraction of their link weights connected to different communities. For example, if a_i is linked to star-nodes $p1, p2$ and $p3$, and $p1$ and $p2$ are members of community c_1 and $p3$ is member of community c_2, then the community membership of a is 67% in c_1 and 33% in c_2.[6]

In the end of the whole process, all the nodes in the HIN are assigned to one or more communities. In the context of the bibliographic data of this work, we aim that our topics consist of three ranked lists of attributes: authors, keywords and ISI fields.[7] Therefore, to rank the attributes within a community, we remove the star-nodes on the network and generate a new HIN with a different network-schema. Figure 4 illustrates the different phases of topic modelling in a HIN.

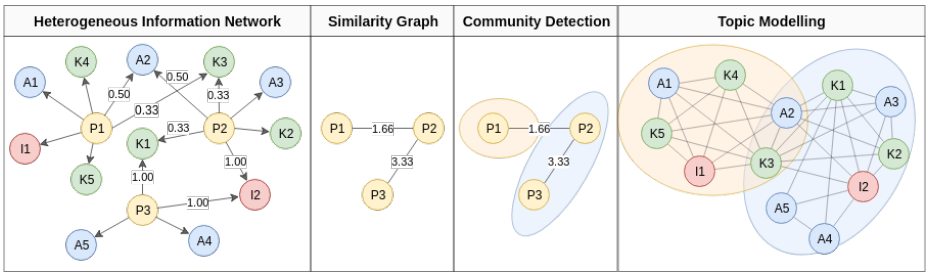

Fig. 4: Topic modelling in HINs using modularity-based community detection.

[6] For simplicity consider that the links have the same weight
[7] As illustrated by figure 2

4.3 Ranking Attributes Within a Topic

With respect to the information network, a topic consists of a sub-network of nodes of three attributes types. In order to better understand the topics discovered, we rank the nodes within each topic according to their importance and type. For the purpose we used several network centrality metrics: node's degree, PageRank, betweenness, closeness and eigenvector. Through experimentation we determined that PageRank seems to be the best metric for our purposes. In this work we use the node's ranking within a topic, to facilitate human interpretation of what a topic represents. However, in the case of extending our expertise profiles to other tasks such as the expert finding one, the rankings could be used to determine who is the best expert in a certain domain.

4.4 Hierarchical Topics

The topic modelling strategy presented in Section 4.2 creates a flat list of topics for a HIN. In this section we summarize the steps necessary to create an hierarchy of topics with a pre-defined number of l levels:

1. Start with HIN $G = (N, L)$
2. Convert the HIN into a similarity graph G' of star-nodes.
3. Apply the *Louvain* community detection algorithm such that $Louvain(G') = C$ where $C = C_1, C_2, ..., C_k$ and each C_i represents a community of star-nodes.
4. Transfer the communities information into the HIN and estimate the community membership of all the attribute nodes.
5. For each $C_i \in C$:
   (a) Create subgraph $G_{Ci} = (N', L')$ where N' is the set of the nodes in community C_i and L' the links between those nodes in G.
   (b) Rank all the attribute nodes according to their importance and object type.
   (c) If the current level is lower than l, set $G = G_{Ci}$ and go back to step 1.

4.5 Mapping Experts into the Hierarchical Topics

One of the problems of using an hierarchy of topics on the expert profiling task is that most of the times, mapping the experts into the hierarchy is either not trivial, or it requires discarding information [9, 16]. In our strategy, we generate topics that consist of multiple attributes. As a result we can use them to map the experts into the topical hierarchy and create expertise profiles. In cases where the expert is represented by a node in the HIN, there is a direct mapping into the hierarchy. Otherwise, the expert can be mapped indirectly using attributes that characterize his expertise and are represented in the HIN.

To create the expert profile of an expert e that is part of the HIN, we transverse the topical hierarchy T and consider all the topics he is part of. For example, let us consider that e at the lowest level of T is 40% in topic *"5-2-2-1"*, 40% in

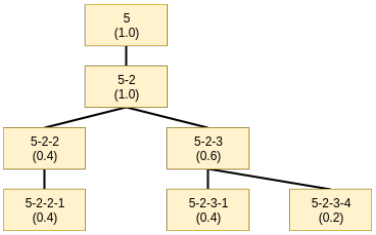

Fig. 5: Example of an hierarchical expert profile.

"5-2-3-1", and 20% in *"5-2-3-4"*.[8] Then, its expert profile p_e considering the complete hierarchy, would be:

- 1st level: 1.0 in topic *"5"*
- 2nd level: 1.0 in topic *"5-2"*
- 3rd level: 0.4 and 0.6 in topics *"5-2-2"* and *"5-2-3"*
- 4th level: 0.4, 0.4 and 0.2 in topics *"5-2-2-1"*, *"5-2-3-1"* and *"5-2-3-4"*

Figure 5 illustrates e's expert profile. In cases where e is not represented in T, we obtain his profile by considering the set of keywords K that he has used in his publications. For each $k_i \in K$ we match it with a keyword node in the HIN by selecting the one with highest Word2Vec similarity [13] to k_i, and obtain its topical profile r_i (similar to the one illustrated in Figure 5.) Then, we sum all the topical profiles into a single one, considering the times the expert used each keyword. For each topic in the merged profile M_p, we estimate its value (V_t) using the following formula:

$$V_t = \sum_{k \in K} \chi(r_i, t) \tag{2}$$

where χ is a function that given a topical profile r_i, extracts the value associated to topic t. On the final step, we normalize the topics' values per hierarchy level in order to make them comparable to profiles extracted directly from T. In this work we are interested in expert profiles, however using the indirect mapping we are capable of creating knowledge profiles for other entities. For example, we can create the profile for a research institution using its authors, or for a conference using the keywords used in it.

5 Experimental Evaluation

In this section we test the efficiency of the discovered topics and the quality of the profiles created using them. For the purpose, we constructed a dataset using all the computer science related publications from the Authenticus database. Our dataset consists of 8587 publications, 2715 authors, 19662 keywords and 120 ISI fields. With this data, we constructed 8 Heterogeneous Information Networks

[8] For clarification, an '-' symbol refers to a different level on the hierarchy

(HIN) changing the weights assigned to each type of relation. For each HIN we applied our model to create a topical hierarchy setting the number of levels to 4.[9] Table 1 shows the relational weights used and the number of topics discovered per hierarchical level.

HIN	relation weights				number of topics per level				
	uniform?	P-K	P-A	P-I	level 0	level 1	level 2	level 3	total
CS_1	Yes	1.0	1.0	1.0	4	9	10	10	33
CS_2	No	1.0	1.0	1.0	4	55	122	200	381
CS_3	No	2.0	1.0	0.5	4	85	352	684	1125
CS_4	No	2.0	0.5	1.0	4	72	253	479	808
CS_5	No	1.0	2.0	0.5	4	51	235	563	853
CS_6	No	0.5	2.0	1.0	4	22	54	94	174
CS_7	No	1.0	0.5	2.0	4	14	30	49	97
CS_8	No	0.5	1.0	2.0	4	9	19	21	53

Table 1: Relational weights and number of topics discovered for the constructed HINs. P-K: publication-keyword. P-A: publication-author and P-I: publication-ISI field.

To evaluate the importance of normalizing the relation weights per publication, we constructed a HIN (CS_1) where the weights are uniform. From the results we observe that the relational weights have a huge impact on the number of topics discovered. Increasing the importance of the publication-keyword relation generates the most topics. On the other hand, decreasing this relation while increasing the publication-ISI field one, generates the least among the HINs with no uniform weights. The uniform HIN generated the fewest number of topics by a high margin.

5.1 Topic Evaluation

In literature, there are several metrics to evaluate the quality of topics modelled. However, they assume that the topics consists only of words, and that they were obtained using statistical inference on text. Our task of constructing an hierarchy of topics, where each topic consists of multiple attributes has only been evaluated by the work of Wang et al. [21]. Therefore, we used the heterogeneous pointwise mutual information (HPMI) metric proposed by the authors to evaluate our topics. HPMI is an extension of the point mutual information metric which is commonly used in topic modelling. For each discovered topic, HPMI calculates the average relatedness of each pair of attributes ranked at top-k:

$$HMPI(v^x, v^y) = \left\{ \begin{array}{cc} \frac{2}{k(k-1)} \sum_{1 \leq i < j \leq k} log(\frac{p(v_i^x, v_j^y)}{p(v_i^x)p(v_j^y)}) & x = y \\ \frac{1}{k^2} \sum_{1 \leq i,j \leq k} log(\frac{p(v_i^x, v_j^y)}{p(v_i^x)p(v_j^y)}) & x \neq y \end{array} \right\} \quad (3)$$

where v^x is a node of type x, ranked among the top-k attributes of type x in a certain topic. The higher the HPMI is, the more coherent the topics are. We calculated the HPMI for the 8 constructed HINs using $k = 20$ and $k = 40$.[10]

[9] Through experimentation we determined that 4 was the number of levels that achieved the most comprehensible topical hierarchy

[10] Following the idea of [21], we setted $k = 5$ for ISI fields since there are only 120 of them in the HIN. In these cases, the part $\frac{1}{k^2}$ of the formula changes to $\frac{1}{5k}$.

HIN	#Topic	K-K	K-A	K-I	A-A	A-I	I-I	Overall
k = 20								
CS_1	33	-1.847	-0.960	-0.726	-1.910	-0.764	-1.056	-1.211
CS_2	381	0.204	1.420	0.222	3.164	0.439	0.057	0.918
CS_3	**1125**	**1.392**	**2.355**	**0.467**	**5.780**	**0.692**	**0.223**	**1.818**
CS_4	808	0.855	1.932	0.347	4.807	0.559	0.144	1.441
CS_5	853	1.025	1.425	0.263	2.735	0.425	0.032	0.984
CS_6	174	0.557	0.479	-0.030	-0.382	0.009	-0.209	0.071
CS_7	97	-1.040	0.492	-0.218	-0.955	-0.135	-0.270	-0.354
CS_8	53	-1.816	-0.946	-0.645	-1.899	-0.671	-0.561	-1.090
k = 40								
CS_1	33	-1.791	-0.966	-0.755	-1.912	-0.757	-1.056	-1.206
CS_2	381	0.289	1.395	0.213	3.171	0.435	0.057	0.927
CS_3	**1125**	**1.443**	**2.349**	**0.467**	**5.777**	**0.691**	**0.223**	**1.825**
CS_4	808	0.902	1.938	0.345	4.808	0.559	0.144	1.449
CS_5	853	1.082	1.423	0.269	2.739	0.422	0.032	0.995
CS_6	174	0.588	0.479	-0.018	-0.394	0.003	-0.209	0.075
CS_7	97	-0.972	0.472	-0.205	-0.969	-0.130	-0.270	-0.346
CS_8	53	-1.730	-0.944	-0.636	-1.922	-0.645	-0.561	-1.073

Table 2: HPMI results for all the HINS. K: keywords. A: authors and I: ISI fields. Highlighted values indicate the highest score for each k.

Table 2 shows the scores obtained. Each column represents the average relatedness of a pair of object types (x, y) for all the topics discovered. The *Overall* column is the average of the values of the 6 possible relations. The results demonstrated that the scores are very similar for k=20 and k=40. Additionally, in 5 out of 8 HINs our strategy was capable of obtaining a positive overall HPMI. Focusing on the best result (CS_3), the topics are highly coherent, specially on the author-author and keyword-author relations. With respect to the HIN construction we observed the importance of using normalization in the relation weights. The only HIN with uniform weights (CS_1) scored the worst. Regarding the non-uniform HINs, CS_7 and CS_8, the only two that assign higher importance to the publication-ISI field relation, are the only ones that achieved an overall negative HPMI. In general we discovered that in order to generate more coherent topics, we must assign an higher importance to the publication-keywords relation. The best results were obtained when doubling the importance of this relation while decreasing the weight of the publication-ISI field one (CS_3).

5.2 Profiles Evaluation

To evaluate the expert profiles created, we selected 12 authors that are computer science professors at the University of Porto. For each one, we crawled their Google Scholar page[11] and collected the research interests that they manually assigned to themselves. In this test, we assume that the research interests of an author reflect his knowledge areas. Table 3 summarizes the name of the authors, the number of publications they have (in the Authenticus database) and their research interests.

For each author we created his hierarchical expert profile using the HIN CS_3 which was the one that yielded the best HPMI results. Then, we compared the profiles against each other to obtain their similarity per hierarchy level. To compute the similarity between authors a_1 and a_2 at a certain level l, we obtain the topical distribution of each authors at l and we sum the topical intersection. The similarity value ranges from 0.0 to 1.0, where 1.0 indicates a perfect match,

[11] https://scholar.google.com/

Name	#Pubs	Google Scholar Interests
Alipio Jorge	133	data mining; machine learning; text mining; recommender systems; artificial intelligence machine learning
Antonio Porto	30	logic programming; coordination; artificial intelligence
Fernando Silva	91	parallel and distributed computing; logic programming; information mining; algorithms; complex networks
Luis Torgo	90	data mining; machine learning
Nelma Moreira	89	automata theory; descriptional complexity; formal verification of software
Pedro Ribeiro	37	complex networks; algorithms and data structures; parallel and distributed computing; computer science education; artificial int
Pedro Brandao	31	communication networks; body area networks; ehealth; distributed systems
Ricardo Rocha	90	logic programming; tabling; parallelism; language implementation
Rita Ribeiro	25	data mining; machine learning
Rogerio Reis	81	formal languages; automata theory; combinatorics
Sergio Crisostomo	16	computer networks; communications; computer science
Veronica Orvalho	40	computer graphics

Table 3: Author's number of publications and google scholar interests.

while 0.0 describes no match between the authors. The total similarity represents the sum of the similarities obtained for all the hierarchy levels.

The aim of this test is to use research interests as evidence to evaluate whether two authors should have high or low similarity profiles. In total we have 132 comparisons. In order to filter some cases on this analysis we divided the results into two groups considering the total similarity between authors. On the first group the total similarity is higher or equal to 2.0, while it is lower or equal to 1.0 on the second one. Tables 4 and 5 show the similarity values and the number of topics shared for both groups.

		Level 0		Level 1		Level 2		Level 3		Total	
Author 1	Author 2	Sim	#To	Sim	#To	Sim	#To	Sim	#To	Sim	#To
Nelma Moreira	Rogerio Reis	1.00	2	1.00	3	1.00	4	1.00	4	4.00	13
Fernando Silva	Pedro Ribeiro	0.73	3	0.73	3	0.60	3	0.60	3	2.66	12
Pedro Brandao	Sergio Crisostomo	0.66	2	0.66	2	0.66	2	0.66	2	2.64	8
Alipio Jorge	Luis Torgo	0.85	3	0.59	4	0.56	4	0.56	4	2.56	15
Fernando Silva	Ricardo Rocha	0.77	3	0.62	4	0.56	4	0.28	3	2.23	14
Pedro Ribeiro	Ricardo Rocha	0.76	3	0.57	3	0.42	3	0.26	2	2.01	11
Luis Torgo	Rita Ribeiro	0.67	2	0.67	2	0.34	2	0.32	2	2.01	8

Table 4: Comparison results for total similarity ≥ 2.0

Only 7 out of 132 comparisons scored a total similarity equal or higher than 2.00. This is expected due to the fact that we have a broad range of interests from the Google scholar, and the lower hierarchical levels refer to very specific topics. Thus, making it more difficult to find similar researchers at those levels. The highest similarity score (Nelma Moreira and Rogerio Reis) represent a perfect profile match at all hierarchical levels. Although their Google scholar interests are very similar, we further looked into this case due to the fact that it represents a wide gap score wise to the other cases. A co-authorship analysis on the network revealed that the two authors are co-authors in 66 publications (81.5% of Rogerio Reis's publications). Therefore, the perfect match is expected. Regarding the other cases, we observe high similarity between pairs of knowledge areas such as: machine learning (Alipio Jorge, Luis Torgo, and Rita Ribeiro), parallel programming (Fernando Silva, Pedro Ribeiro and Ricardo Rocha), and communication networks (Pedro Brandao and Sergio Crisostomo).

An interesting fact is to note that two authors, Veronica Orvalho and Antonio Porto, are not similar enough with any other author. In the case of Veronica Orvalho, this is anticipated due to the fact that her interest on computer graphics is not shared by any other author. However, in the case of Antonio Porto, since his interests refer to areas shared by other authors an higher comparison was expected. A further look into his profile revealed that it is scattered by several topics. As a result, his intersections with other authors are not significant enough.

Author 1	Author 2	Level 0		Level 1		Level 2		Level 3		Total	
		Sim	#To	Sim	#To	Sim	#To	Sim	#To	Sim	#To
Nelma Moreira	Veronica Orvalho	0.25	1	0.25	1	0.25	1	0.25	1	1.00	4
Rogerio Reis	Veronica Orvalho	0.25	1	0.25	1	0.25	1	0.25	1	1.00	4
Antonio Porto	Pedro Ribeiro	0.66	2	0.33	1	0.00	1	0.00	1	0.99	5
Antonio Porto	Pedro Brandao	0.66	2	0.33	1	0.00	1	0.00	1	0.99	5
Fernando Silva	Luis Torgo	0.37	2	0.37	2	0.17	1	0.00	1	0.91	6
Nelma Moreira	Pedro Ribeiro	0.33	1	0.33	1	0.25	1	0.00	1	0.91	4
Nelma Moreira	Pedro Brandao	0.58	2	0.33	1	0.00	1	0.00	1	0.91	5
Nelma Moreira	Sergio Crisostomo	0.58	2	0.33	1	0.00	1	0.00	1	0.91	5
Pedro Ribeiro	Rogerio Reis	0.33	1	0.33	1	0.25	1	0.00	1	0.91	4
Pedro Brandao	Rogerio Reis	0.58	2	0.33	1	0.00	1	0.00	1	0.91	5
Rogerio Reis	Sergio Crisostomo	0.58	2	0.33	1	0.00	1	0.00	1	0.91	5
Alipio Jorge	Pedro Brandao	0.61	3	0.28	2	0.00	1	0.00	1	0.89	7
Alipio Jorge	Antonio Porto	0.64	2	0.14	1	0.00	1	0.00	1	0.78	5
Fernando Silva	Sergio Crisostomo	0.33	1	0.33	1	0.00	1	0.00	1	0.66	4
Pedro Ribeiro	Sergio Crisostomo	0.33	1	0.33	1	0.00	1	0.00	1	0.66	4
Rita Ribeiro	Sergio Crisostomo	0.33	1	0.33	1	0.00	1	0.00	1	0.66	4
Ricardo Rocha	Sergio Crisostomo	0.47	2	0.14	1	0.00	1	0.00	1	0.61	5
Luis Torgo	Sergio Crisostomo	0.34	2	0.17	1	0.00	1	0.00	1	0.51	5
Alipio Jorge	Sergio Crisostomo	0.28	2	0.14	1	0.00	1	0.00	1	0.42	5
Antonio Porto	Sergio Crisostomo	0.33	1	0.00	1	0.00	1	0.00	1	0.33	4
Sergio Crisostomo	Veronica Orvalho	0.25	1	0.00	1	0.00	1	0.00	1	0.25	4

Table 5: Comparison results for total similarity ≤ 1.0

With respect to the least similar results, in general they complement the observations from the top results that some areas (machine learning, parallel programming and communication networks) do not merge into highly similar profiles. In most of the cases we observe that there is a similarity in the level 0 of the hierarchy (i.e. on the broader topics), however as the topics get more specific the intersections between authors fade. An interesting case to highlight is the author Sergio Crisostomo, that matches on the first two levels with almost every other author, but with none (exception to Pedro Brandao, who shares a high similar profile with him) on the last two levels of the hierarchy. This indicates that from the level 2 of the topical hierarchy, there is a clear distinction of the communication network topics (his most specific google scholar interests).

Another case worth to note is the fact that although Veronica's interests are further away in comparison to the others, she still has some comparisons with total similarity higher than 1.0. A further look into her profile revealed that she is scattered through several topics however she is never a highly ranked author of the topic. In our dataset, the computer graphics area does not have as many publications as other areas such as machine learning and parallel programming. As a result, our strategy fails to model the topic correctly and scatters its information among other more predominant topics.

6 Conclusions

In this paper we addressed the problems of topic modelling and expert profiling. We avoided the problems of the LDA-based approaches by using modularity optimization in HINs to discover multi-typed topics. Additionally, we proposed a strategy to use the modelled topics to profile experts whether they are represented in the HIN or not. In order to tackle the current literature problems of constructing profiles that are redundant and either too specific or too broad, we organized the topics into an hierarchy. As a result, we create an hierarchical profile which starts with describing the expert's most broad areas, and it moves

to the most specific ones. We evaluated our model with respect to the topics discovered using a state of the art metric (HPMI). This test revealed that the topics generated are coherent. Furthermore, in order to maximize topic coherency we have to assign the highest importance to the publication-keyword relation in the HIN. In another test, we used Google scholar data to evaluate the quality of the hierarchical profiles constructed. Our test revealed that we are capable of generating high similarity profiles for experts that have common research interests, while generating low similarity profiles for the ones that do not. This test also demonstrated that we need to improve our strategy to model topics that are under represented in the data.

Regarding future work, in the domain of topic discovery, we plan to test other community detection algorithms, specially the ones that not require transforming the HIN into a similarity graph. In the domain of the expert profiling, we aim to take a further look into the rankings of the nodes inside a topic and how they can be used in the profiling step. We also aim at creating an automatic summarization of the topics in such a way that we can construct a visualization of the expert's profile. Furthermore, we will look into considering the timestamps of the expert's meta-data in order to create time-sensible profiles.

7 Acknowledgements

This work is funded by the ERDF through the COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961, and by National Funds through the FCT as part of project UID/EEA/50014/2013. Jorge Silva is also supported by a FCT/MAP-i PhD research grant (PD/BD/128157/2016).

References

1. Balog, K., Fang, Y., de Rijke, M., Serdyukov, P., Si, L., et al.: Expertise retrieval. Foundations and Trends® in Information Retrieval **6**(2–3), 127–256 (2012)
2. Berendsen, R., Rijke, M., Balog, K., Bogers, T., Bosch, A.: On the assessment of expertise profiles. Journal of the Association for Information Science and Technology **64**(10), 2024–2044 (2013)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
4. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment **2008**(10), P10008 (2008)
5. Daud, A.: Using time topic modeling for semantics-based dynamic research interest finding. Knowledge-Based Systems **26**, 154–163 (2012)
6. De Campos, L.M., Fernández-Luna, J.M., Huete, J.F.: Committee-based profiles for politician finding. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **25**(Suppl. 2), 21–36 (2017)
7. Duan, D., Li, Y., Li, R., Lu, Z., Wen, A.: Mei: Mutual enhanced infinite community–topic model for analyzing text-augmented social networks. The Computer Journal **56**(3), 336–354 (2012)

8. Gerlach, M., Peixoto, T.P., Altmann, E.G.: A network approach to topic models. arXiv preprint arXiv:1708.01677 (2017)
9. bin Jamaludin, N.A., Annamalai, M., Jamil, N., Bakar, Z.A.: A model for keyword profile creation using extracted keywords and terminological ontology. In: e-Learning, e-Management and e-Services (IC3e), 2013 IEEE Conference on. pp. 136–141. IEEE (2013)
10. Jeong, Y.S., Lee, S.H., Gweon, G.: Discovery of research interests of authors over time using a topic model. In: Big Data and Smart Computing (BigComp), 2016 International Conference on. pp. 24–31. IEEE (2016)
11. Karimzadehgan, M., White, R.W., Richardson, M.: Enhancing expert finding using organizational hierarchies. In: European Conference on Information Retrieval. pp. 177–188. Springer (2009)
12. Li, C., Cheung, W.K., Ye, Y., Zhang, X., Chu, D., Li, X.: The author-topic-community model for author interest profiling and community discovery. Knowledge and Information Systems **44**(2), 359–383 (2015)
13. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
14. Newman, M.E.: Modularity and community structure in networks. Proceedings of the national academy of sciences **103**(23), 8577–8582 (2006)
15. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: Proceedings of the 20th conference on Uncertainty in artificial intelligence. pp. 487–494. AUAI Press (2004)
16. Rybak, J., Balog, K., Nørvåg, K.: Temporal expertise profiling. In: ECIR. pp. 540–546. Springer (2014)
17. Shi, C., Li, Y., Zhang, J., Sun, Y., Philip, S.Y.: A survey of heterogeneous information network analysis. IEEE Transactions on Knowledge and Data Engineering **29**(1), 17–37 (2017)
18. Sun, Y., Han, J., Zhao, P., Yin, Z., Cheng, H., Wu, T.: Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In: Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology. pp. 565–576. ACM (2009)
19. Sun, Y., Yu, Y., Han, J.: Ranking-based clustering of heterogeneous information networks with star network schema. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 797–806. ACM (2009)
20. Tang, J., Jin, R., Zhang, J.: A topic modeling approach and its integration into the random walk framework for academic search. In: Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on. pp. 1055–1060. IEEE (2008)
21. Wang, C., Liu, J., Desai, N., Danilevsky, M., Han, J.: Constructing topical hierarchies in heterogeneous information networks. Knowledge and Information Systems **44**(3), 529–558 (2015)
22. Wang, J., Hu, X., Tu, X., He, T.: Author-conference topic-connection model for academic network search. In: Proceedings of the 21st ACM international conference on Information and knowledge management. pp. 2179–2183. ACM (2012)