A Combination of Semantic Annotation and Graph Mining for Expert Finding in Scholarly Data

Stella Zevio[1], Guillaume Santini[1], Henry Soldano[1,2,3], Haïfa Zargayouna[1], and Thierry Charnois[1]

[1] LIPN CNRS UMR 7030, Université Sorbonne Paris Nord, Villetaneuse, France
[2] Muséum d'Histoire Naturelle, ISYEB, Paris, France
[3] NukkAI, Paris, France
{zevio,santini,zargayouna,charnois}@lipn.univ-paris13.fr
soldano@mnhn.fr

Abstract. Scholarly data is a relevant source of knowledge for expert finding in academia. Although peer validation is widely spread in academia, state-of-the-art methods do not take peer validation into consideration for expert finding but exploit similarity between experts instead. We propose a new definition of expertise, taking peer validation into consideration. That leads us to suggest a new method for expert finding from scholarly data, combining semantic annotation with graph mining. First, we extract expertise from a corpus of scientific publications thanks to semantic annotation through an ontology. We then represent extracted knowledge in the form of attributed graphs. The originality of our approach lies in the consideration of the scientific collaboration links that individuals maintain in the scientific community, in order to identify the experts on a given topic. Indeed, scientific collaboration links are carried by the scientific publications themselves. They can easily be extracted from scholarly data and enable to take into account an implicit peer validation within a community when identifying experts. To take into account this peer validation, we use an emerging graph mining method called core closed pattern mining and use a recent pattern set selection method that enables to reduce the number of patterns to consider. As we consider two-mode networks relating authors to the articles they write, we exploit bi-pattern mining and we introduce a new way to reduce enumeration by constraining the component patterns to have common items. We apply our method to a sample of the ACL Anthology corpus and demonstrate that we can identify relevant sets of experts and their shared expertise.

Keywords: Graph mining · Semantic annotation · Scholarly data

1 Introduction

In academia, expert finding is a recurring problem. Indeed, it is essential to assign appropriate experts when setting up program committees for scientific events, for

example. The identification of experts and their associated expertise is a critical task for expert finding. However, determining that an individual is an expert is not only related to the identification of competencies they master [1] but also on the receipt of their work by other members of the scientific community. For example, an individual who is frequently cited on a recurring topic is probably a relevant expert on the same topic. That leads us to suggest a new definition of expertise, taking into account peer validation which is widely spread in academia. State of the art methods do not take peer validation into consideration but exploit similarity between experts through text mining or graph mining methods [2] instead. Most promising results have been recently obtained by combining text mining with graph mining approaches [3].

We propose a method for discovering experts and their associated expertise from scholarly data using such a combination. A key idea of our work is to exploit pattern mining in attributed graphs. More precisely, our approach combines classical semantic annotation methods through an ontology with an emerging method of attributed graph mining called core closed pattern mining. From scholarly data, the expertise and collaboration links between experts are extracted. We obtain an annotated corpus that we represent in the form of attributed graphs. In a second step, we apply core closed pattern mining to the graphs in order to extract dense subgraphs under constraints. This method allows to enumerate maximum sets of common expertise shared by experts and identify the associated sets of experts while taking into consideration peer validation embodied by the constraints. In this paper, we introduce restricted bi-pattern mining through some experiments about bipartite graphs. We also use a recent pattern selection method, the $g\beta$ pattern selection, to reduce the number of patterns to explore.

The rest of the paper is organized as follows. After giving an overview of related work on expert finding in Section 2, background on core closed pattern mining and core bi-pattern mining which are useful for our method is given in Section 3 as well as the pattern set selection process. The proposed method is presented in Section 4 and the results in Section 5. Finally, Section 6 shows that our approach allows to identify relevant sets of experts as well as their shared expertise and discusses the leads to further investigate.

2 Related work

Initially, expert finding systems were based on the self-assignment of expertise from a selection of predefined keywords [4]. Since self-assignment of expertise is a time-consuming task requiring continuous maintenance [5], it is essential to automate the profiling of experts from other sources of knowledge [2]. Two sources were explored : documents (mainly text), constituting the predominant source, as well as social networks [2]. Limits of the analysis of social networks are due to the weak investment of scientists in them. Indeed, all researchers are not registered or active on social networks. Therefore, despite the relevance of altmetrics used to rank individuals in social networks [6], they do not alone define individuals as experts in the scientific community [2]. However, abundant knowledge about

individuals' expertise is buried in the research papers, which are considered as universal working documents in academia [7]. Peer validation is widely used in academia and scholarly data carries itself the scientific collaboration links that are vital to evaluate the insertion in the scientific community. Still, the automation approaches of expert finding from texts are based on the capture of expertise buried in the text and their assignment to the individuals with whom they are associated [8,2]. In that case, interdisciplinarity and peer validation are not considered, as a researcher can be considered an expert on any topic of publication tackled in their research papers. Taking advantage of the approaches inspired by the analysis of social networks is therefore essential to capture expertise in scientific publications as well as to represent scientific collaboration links as an implicit peer validation. Indeed, the state of the art suggests that problems related to the identification and classification of experts can be avoided by combining the identification of expertise with the identification of relationships between experts [4,2].

According to the literature, the main methods of expert finding are based on graphs or machine learning [2]. One of the leading systems analysing scholarly data is Rexplore [9]. It is based on a Temporal Topic-Based Clustering algorithm [10] which is an unsupervised machine learning algorithm based on clustering of the researchers' trajectories. This approach seems more appropriated to trend analysis rather than expert finding as it does not distinguish a hierarchy between researchers through any kind of peer validation or classification. Furthermore, graph-based methods provide better performance than machine learning based methods [2]. Indeed, representation of knowledge extracted from data through graphs is quite common in expert recommendation systems. Apart from machine learning methods, graph-based methods have been used for expert finding, principally in the analysis of social networks and expertise graphs [11]. Expertise graphs are social graphs in which nodes are experts and non experts, edges a relation between them [11,2]. Graph-based methods fall into two main categories: graph properties (correctness, centrality) and computing algorithms (HITS, PageRank or other algorithm variations) [2]. From RDF graphs, experts can also be discovered from rules [12]. Bayesian approaches such as clustering thanks to a Dirichlet algorithm can also be used to detect clusters in a star network [13]. The limitations of graph-based methods lie in their failure in taking into account the content of documents [2], and therefore in their lack of consideration of the expertise buried in the working documents. Recent work indicated that the combination of text mining and graph mining methods leads to better performance than using one of the two methods alone, although their combination remained anecdotal [14].

Thus, hybrid methods combining machine learning and graph-based methods have been developed recently. The most effective expert finding system as far as we know is Wiser [3]. It is an unsupervised system combining document-centric approaches with entity linking from Wikipedia Knowledge Graph. Every academic author indexed in Wiser is associated to a graph issued from Wikipedia representing the Wikipedia entities mentioned in the author's publication. Entities,

namely topics of publication, are linked by their semantic relatedness in the graph. At present, Wiser is considered the state of the art expert finding system. Their method, though, does not take into account any kind of peer validation but is based on the analysis of semantic likeness between expert profiles and expert finding queries.

On the other hand, identifying experts and their associated expertise in a semantic network can relate to knowledge discovery in attributed graphs. Recent work focuses on combining the consideration of attribute patterns with connectivity constraints enabling to enumerate closed patterns occurring in cores, that is dense areas of an attributed graph respecting a core property [15]. In attributed graphs describing experts and non experts connected by collaboration links and described by features, recent work demonstrated that the core closed pattern mining method is relevant for enumerating maximum sets of common features shared within cores along their support sets [16]. Adapted to our problem of expert finding, this method could be relevant for investigating expertise graphs and enumerating maximum sets of shared expertise along with identifying their associated experts while taking into account peer validation embodied by the core properties. The method has already been used in citation networks, with hub-authority core (also named h-a-HAcore), which is a HITS inspired core property [16]. In the Section 3, we recall the state of the art of core closed pattern mining and introduce restricted bi-pattern mining that will be useful in the modelization of our experiments presented in Section 5.

3 Restricted bi-pattern mining unifies single and bi-pattern mining

We rely on core closed single pattern mining and bi-pattern mining of attributed network [17]. In closed pattern mining, a pattern q has an *extension* also called a *support set* $e = \text{ext}(q)$ representing its set occurrences in a set of objects V. This support set defines the equivalence class of all patterns with support set e. The most specific pattern c with support set e is unique as far as the pattern language is a lattice and will be considered as the representative of this class. We may then enumerate closed patterns that represents a *condensed representation* of all patterns in the object dataset. The closed pattern is obtained by using an intersection operator int that applies the lowest upper bound operator $\wedge$ to the set of object descriptions $d[e]$[4]. In the attribute pattern setting, objects are described as itemsets i.e. subsets of a set of items I. In this case the intersection operator simply is the set theory intersection operator $\cap$.

When applying an operator p to $\text{ext}(q)$ that reduces the support set into a so-called *abstract* support set, the most specific pattern c of the class of pattern with same abstract support set $e = p(\text{ext}(q))$ as pattern q is defined and obtained as:

$$c = f(q) = \text{int} \circ p \circ \text{ext}(q) \tag{1}$$

[4] $d[e]$ is the image of e by d, i.e. $d[e] = \{d(v) | v \in e\}$

This requires p to be an interior operator, which is the case when the object set V is the set of vertices of a graph $G = (V, E)$ and that p is an operator that, given a vertex subset W, returns the *core* of the subgraph induced by W[15] i.e. the largest subset S of W whose vertices all satisfy some *core property* P within the subgraph induced by S. The resulting abstract closed patterns are called *core closed patterns*.

We display Figure 1 an attributed network together with its 2-core pattern subgraph, whose vertices all have degree at least 2, and its 3-nearstar core pattern subgraph, whose vertices are stars, i.e. have degree at least 3 or are satellite connected to some star.

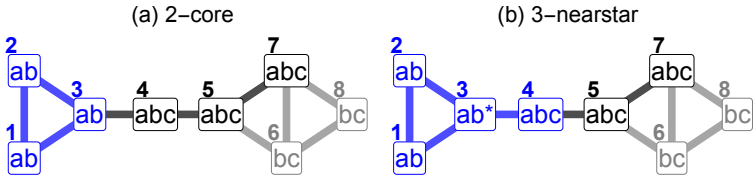

Fig. 1: An attributed network. Pattern a occurs in vertices 123457 so inducing the pattern a subgraph represented as bold vertices and edges. (a) On the left within the pattern a 2-core subgraph depicted in blue, the vertices have in common the core closed pattern ab. (b) On the right is pictured the pattern a 3-nearstar core subgraph, made of a star (vertex 3) and three satellites (vertices 1,2, 4). The associated core closed pattern is again ab.

Bi-pattern mining [17] allows to mine two-mode networks $G(V_1, V_2, E)$ whose edges in E relate vertices from V_1 to vertices from V_2. A vertex subset pair (W_1, W_2) induces then a subgraph $G_{(W_1,V_2)}$ whose edges relate W_1 to W_2. Bi-pattern mining follows from the remark that one may extend single pattern mining by:

- considering pairs of patterns $q = (q_1, q_2)$ called bi-patterns. q_1 is a subset of some set of items I_1. while q_2 is a subset of another set of items I_2. I_1 and I_2 may have common items.
- extending accordingly support sets to support set pairs $\text{ext}(q) = (\text{ext}_1(q_1), \text{ext}_2(q_2))$ where $\text{ext}_i(q_i)$ is the support set of q_i within the object set V_i. We also writes, when convenient, $\text{ext}_i(q_i)$ as $\text{ext}(V_i, q_i)$.
- defining the intersection operator *int* that applies to a pair of object subsets $e = (W_1, W_2) \in V_1 \times V_2$ as $\text{int}(e) = (\text{int}_1(W_1), \text{int}_2(W_2))$ where $\text{int}_i(W_i)$ is the intersection of the objects descriptions in W_i.
- considering an interior operator which reduces the pair of object subsets $e = (W_1, W_2)$ to a smaller pair $e' = (W_1', W_2')$ i.e. a pair such that $W_1' \subseteq W_1$ and $W_2' \subseteq W_2$.

Accordingly to Equation 1 the core closed bi-pattern c is defined as the most specific bi-pattern (considering both components) whose support set pair is the same as bi-pattern q. Cores have then to be pairs of vertex subsets, and are called *bi-cores*. Bi-core definitions rely on pairs of bi-core properties (P_1, P_2). As an example, in the *h-a* BHA bi-core subgraph $G_{(C_1,C_2)}$ all nodes from C_1 have degree at least h and all nodes from C_2 have degree at least a.

3.1 Restricted bi-pattern mining

In this section we introduce a new way to restrict bi-pattern mining: whenever the two sets of items I_1 and I_2 used to describe the two kind of nodes intersect, we may consider only bi-patterns q_1, q_2 such that q_1 and q_2 have the same items on a part F of this intersection. We have then two extreme cases:

- $F = \emptyset$ is the unrestricted bi-pattern mining case discussed above.
- $F = I_1 = I_2 = I$ is the single pattern mining case. It concerns networks in which the vertex sets represents objects of same type and results in bi-patterns of the form (q, q).

The natural scenario where restricted bi-pattern mining appears is whenever investigating a two-mode network in which there is some common attribute subset $I_1 \cap I_2$ shared by nodes from the two modes. It occurs in our experiments because authors and articles they write both are concerned with some scientific domain which is associated to an item. Though the meaning of such an item may slightly differ whether the node represents an author or an article, it still makes sense to investigate bi-patterns in which the scientific domain has to be shared by all nodes in the core subnetwork.

We define F-restricted bi-pattern mining by considering a subset F of $I_1 \cap I_2$, called the *constrained common part*, and requiring that in any F-restricted bi-pattern (q_1, q_2), and for any item i from F, either i belongs both to q_1 and q_2 or i belongs to neither of them. Fortunately, in such a restricted bi-pattern language, given some restricted bi-pattern (q_1, q_2) there still is a unique most specific restricted bi-pattern whose support set pair is the same as (q_1, q_2). As a consequence, we may define a closure operator and obtain core closed bi-patterns.

3.2 $g\beta$ Pattern set selection in attributed graphs

Now we deal with a recurrent question in pattern mining: how to select a few number of relevant and non redundant patterns ? $g\beta$ is a simple and general post-processing pattern subset selection scheme[18] that selects within a pattern set P a pattern subset S such that i) in S pairwise distances between patterns all exceed some threshold β and ii) S maximizes the sum of the individual interestingness g of its patterns. This supposes that we have some distance definition on patterns together with some positive interestingness measure g. The $g\beta$ algorithm is a greedy algorithm that guarantees the distance constraint and efficiently returns an approximate optimal solution S. The greedy $g\beta$ algorithm has worst case

complexity $\mathcal{O}(|P||S|)$ both in number of comparisons and number of distances to compute. As a consequence it is very efficient when a strong distance constraint β is applied. It consists in *i)* an initialisation step in which an empty list S is defined and in which patterns from P are sorted in decreasing order of g values ii) a search step in which each pattern in P is in turn either rejected, when its distance to some pattern from the current S list is smaller or equal to β, or added to S.

Bi-Pattern Set Selection To apply $g\beta$ to bi-pattern set selection we need to define the distance d between bi- pattern as well as the interestingness measure g.

Distances In the single pattern mining case, the core closed pattern q is associated to a vertex subset, its core support set $p \circ \text{ext}(q)$. As a distance $d(q, q')$ between patterns q and q' we will use the Jaccard distance between their core support sets. Recall that the Jaccard distance between two subsets X and X' of some set has range [0,1] and is defined as $d_J(X, X') = 1 - \frac{|X \cap X'|}{|X \cup X'|}$. We then have: $d(q, q') = d_J(W, W')$ where W is the core support set of q and W' is the core support set of q'.

Regarding bi-patterns we compute the distances between their pattern components and take the maximum value. This is a conservative choice: when bi-pattern q is selected, to remove bi-pattern q' both components of q' have to be at distance less than β from q. We have then:

$$d(q, q') = \max(d_J(H, H'), d_J(A, A')) \tag{2}$$

where (H, A) is the core support set pair of bi-pattern q and (H', A') is the core support set pair of bi-pattern q'.

Selecting and ordering patterns In the $g\beta$ algorithm, g only role is in the (bi)-pattern ordering by decreasing g values. This pattern ordering may have a high impact on the pattern set we obtain. We further consider local modularity as in [19], i.e. a measure that have high positive values whenever there are more inside links in a subgraph, with respect to all links whose extremities are within the subgraph, than expected.

4 Methodology and datasets

In the light of the state of the art, we propose an approach for expert finding whose originality is based on the combination of the capture of underlying expertise within working documents with the consideration of an implicit peer validation between experts thanks to collaboration links carried by working documents. The preliminary step consists in extracting the underlying expertise and scientific collaboration links within scholarly data. Then, the knowledge previously extracted is represented through attributed graphs. The final step

consists in using core closed pattern mining to identify sets of experts connected in the graphs along with the expertise they share. We also set up an evaluation framework.

4.1 ACL Anthology corpus

We consider a sample of the ACL Anthology corpus [20], composed of 13322 research papers, written by 10724 authors and published between 1985 and 2008 in the areas of computational linguistics and natural language processing [5]. The original ACL Anthology corpus [21] is comprised of 48104 research papers. The sample thus covers approximately 28% of the original corpus. For each paper, descriptors are available including an identifier, authors list, title, year of publication, abstract, cited authors, titles of the cited publications and years of publication of cited publications.

4.2 Labeling publications with ontology concepts

From a corpus of scientific publications, underlying expertise and scientific collaboration links can be extracted. Abstracts carry semantic concepts corresponding to topics of publication taken as expertise. Furthermore, coauthor and citation links convey information about endorsement of authors' work by the scientific community. Indeed, if a researcher can be considered as a prominent member of the scientific community when maintaining coauthorship with numerous others researchers. In addition, if a researcher is frequently cited by others, it is likely that this researcher is eminent and can be considered as an expert. If machine learning methods are commonly used in order to extract expertise buried in texts [8,2], semantic annotation through an ontology can be employed to identify semantic concepts within texts. Using ontologies to extract semantic concepts from text enables to have interpretable and interoperable results and to exploit richer and more specialized languages than simple term recognition. The Computer Science Ontology has recently been released [22] and used to classify research papers [23] in computer science [22] through a syntactic and a semantic recognition modules [23]. The syntactic recognition module identifies semantic concepts from the ontology whose labels are explicitly recognized in the abstract [23]. The semantic recognition module infers the presence of semantic concepts from the ontology using a morphosyntactic labeling of terms and lexical embedding [23]. The classifier enables to automatically classify research papers according to the rich taxonomy of fine-grained concepts issued from the Computer Science Ontology [22,23]. For example, the semantic web is described by more than 40 sub-topics in the Computer Science Ontology, such as Linked Data or SPARQL. We performed a semantic annotation of the ACL Anthology corpus. From abstracts, 2714 semantic concepts derived from the Computer Science Ontology [22] have been extracted on our behalf. We consider that semantic concepts from the Computer Science Ontology found in abstracts are underlying expertise within scientific publications. Indeed, they correspond to topics of publication.

[5] ACL Anthology corpus sample : url to dataset anonymized

4.3 Evaluation and gold standard

We use a review article [24] from 2010 on the domain of *information extraction* as a source to set up a gold standard. Information extraction belongs to the areas covered by the ACL Anthology corpus. Furthermore, the review article was published few years after the last ACL Anthology corpus' articles. Among the 90 references belonging to the review article, 47 also belong to the ACL Anthology corpus. The semantic concept from the Computer Science Ontology the most used to describe the abstracts of the 47 publications is *information extraction* (IE for short). This motivates our choice for using the review article as an evaluation framework. Thus, we select closed abstract patterns describing the *information extraction* domain for each of our graph. We use classic precision and recall measures to evaluate the relevance and coverage of the publications and authors we obtain as key documents and experts on the closed pattern selected containing the domain of *information extraction*. Among the authors who wrote the 47 publications, 97 also belong to the authors of the ACL Anthology corpus sample. 133 semantic concepts were used to describe the 47 publications.

5 Experiments

In this section we will present the experiments that we conducted when applying our method of expert finding combining semantic annotation with core closed (bi-)pattern mining on the ACL Anthology corpus and evaluate results obtained according to our evaluation framework.

5.1 Encoding the information in a labeled graph

Our goal is to identify experts and their associated expertise with the assumption that expertise in the academic community is not self-assigned by authors but derives from peer approval. We formulate the hypothesis that this peer approval can be effectively encoded in form of links in a network of scientific collaboration and that the identification of experts and their associated expertise benefits from taking into account this relational dimension. Many labeled graphs can be generated from the ACL data corpus. Authors of publications are considered as potential experts: they form the vertices of the graph. Semantic concepts identified in publications are considered as potential expertise: they are used to label vertices. Links in the graph represent the network of scientific collaboration.

Given a scientific network graph whose vertices are potential experts labelled with potential expertise, the application of abstract pattern mining to the graph should allow for the identification of subsets of related experts, linked in the scientific collaboration network and sharing the same set of peer-reviewed expertise. Obviously, among the set of graphs that can be constructed from the ACL dataset, not all of them translate with the same quality the dimensions of peer endorsement. In this study we will present, for the sake of clarity, results for two of them only: G_{co} and $G_{A \to P}$ graphs.

- Co-authors graph G_{co} offers the least confidence but is also the simplest way to represent a scientific endorsement. Indeed, in this graph each co-author is a potential expert of all the concepts found in the publications he co-authors.
- The bipartite citation graph $G_{A \to P}$ offers more confidence. The scientific collaboration network connects authors with the bibliography they use in their articles. Authors and publications are labeled with the same semantic concepts. A group of experts will be identified because they all use the same semantic concepts and cite the same publications that use the same semantic concepts.

Depending on the nature of the graph we will apply different abstractions to it. The choice of the abstraction determines a constraint of connectivity between the authors that we want to see verified in the sub-graph of the scientific collaboration network induced by the sets of authors identified as experts because retained by the enumeration.

For the co-authoring simple graph G_{co} we explore the results of the k-core and k-nearstar-core abstractions. The authors and pattern language is L_A which is described hereunder as part of the language pair (L_A, L_P) used in the bipartite graph $G_{A \to P}$.

Regarding the citation graph $G_{A \to P}$, we consider the F-restricted-h-a-BHA bi-core abstraction as introduced in the section 3.1. In the graph $G_{A \to P}$, the pattern language on authors is $L_A = T_A \cup C$ where

- T_A is a set of constraints on the publication period of authors in A. A publication period is an interval $\Delta_A =]l, r]$ where l and r are both thresholds belonging to the set of years $T = \{1992, 1999, 2004, 2007\} \cup \{-\infty, \infty\}$. For instance the publication period of an author who has published between 1997 and 2006 is $\Delta_A =]1992, 2007]$. Consider then some author, T_A allows to encode two kind of constraints on Δ_A:
  - $\Delta_A \cap]l, r]$ occurs whenever Δ_A has a non void intersection with the interval $]l, r]$.
  - $\Delta_A \supseteq]l, r]$ occurs whenever Δ includes the interval $]l, r]$.

  For instance, consider two authors with respective publication periods $]1992, 2007]$ and $]-\infty, 2004]$ they are both occurrences of pattern $\Delta_A \supseteq]1992, 2004]$ and also both occurrences of pattern $\Delta_A \cap]1999, 2004]$. Clearly $\Delta_A \supseteq]x, y]$ is more specific (is a stronger constraint) than $\Delta_A \cap]x, y]$.
- C constrains the semantic concepts appearing in publications. Let X be set of semantic concept appearing in the publications authored by some author. A Pattern in C is then of the form "$X \supseteq S$" where S is a semantic concept subset. As an example, the pattern "$X \supseteq \{information\ extraction, languages\}$" occurs if the author has published articles concerning *information extraction*, *languages* and possibly other semantic concepts. We further simply write such patterns as sequences, e.g. *information extraction, languages*, and when convenient in a concise way as *infor._extr., languages*.

In the same way, the description language of publications in P is $L_P = T_P \cup C$ where T_P is a set of time constraints on the publication year Y_P. The part

constraining Y_P in a pattern is then of the of the form $t_i < Y_P \leq t_j]$, where t_i and t_j belongs to the same threshold set T as in T_A. We also use semantic concept patterns from C regarding publications, in which case X is the semantic concept subset whose concepts appear in the publication.

Regarding the citation graph $G_{A \to P}$, we restrict our search to restricted bi-patterns in which the concepts are the same in both of its components, i.e. we consider bi-pattern q such that $q = (q_A, q_P)$ satisfies $q_A = q_{T_A} \cup q_C$ and $q_P = q_{T_P} \cup q_C$.

Example 1.
$((inf.\_extr., \Delta_A \supseteq [1992, 2004[), (inf.\_extr., Y_P \in ([1999, 2007[))$
is a restricted bi-pattern while
$((inf.\_extr., \Delta_A \supseteq [1992, 2004[[), (inf.\_extr., languages, Y_P \in [1999, 2007[))$ is not.

The occurrences of the former is represented as a pair (S_A, S_P) where

- S_A contains authors whose publication period contains $[1992, 2004[$ and whose publications contain the semantic concept *information extraction*.
- S_P contains cited publications whose publication year is within $[1999, 2007[$ and that contains the semantic concept *information extraction*.

We further denotes restricted bi-patterns in a simpler way, namely in this example:
$inf.\_extr., \Delta_A \supseteq [1992, 2004[, Y_P \in ([1999, 2007[$.
Furthermore, as we consider 3-3 BHA bi-cores, authors from S_A should have cited at least 3 publications in S_P while publications in S_P should have been cited by at least three authors from S_A.

In order to be able to evaluate the quality of our results throughout this study, we will rely on the gold standard we have set up with identified experts in the field of information extraction. We will therefore initially focus on experts and publications labeled with the semantic concept *i*nformation extraction (IE for short).

5.2 High degree authors are not always experts

First we explore the naive idea that authors who are experts in a field would be more connected in the scientific network induced by that field of expertise. If we follow this idea, authors who are experts in information extraction (IE) should be among the authors with the highest degree in the graph induced by the authors labelled with the IE concept.

As shown in figure 2, selecting the topK authors of the highest degree in these graphs does not allow efficient selection of experts in the field of information extraction. In order to identify experts in a field among a graph of authors, it is therefore necessary to go beyond the identification of single authors highly connected. One way to do this is to consider that expert of a field should be close in the network. Though it would be more efficient to search for connected sub-networks of authors sharing the same expertise. A solution consists in enumerating the sub-networks corresponding to core closed patterns of a labeled graph.

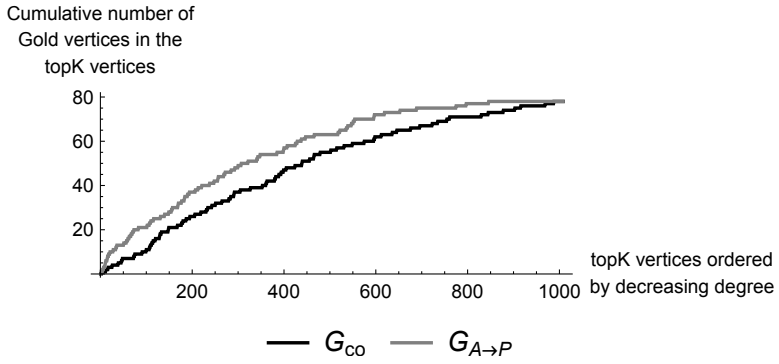

Fig. 2: G_{co} and $G_{A\to P}$ vertices are ordered by decreasing number of degree. In black (resp. grey) the cumulative number of gold vertices (among 97) in the topK vertices.

5.3 Identifying the abstraction best preserving the information

As shown in figure 3, the lower the core constraint is - i.e. low value of k-, the better is the coverage of the graph vertices and the better is the recall the gold standard IE experts in the union of the found core closed patterns.

First, the core closed patterns satisfying the k-core property in G_{co} have been enumerated. The overall low recalls indicates that the abstraction of k-core is probably not the best way to identify experts (here in the field IE).

The core constraint property has therefore been relaxed to test the efficiency of the k-nearstar core. It appears that the number of core closed patterns explode very fast when decreasing the k-nearstar constraint. For a similar number of patterns, the k-nearstar property gives a better recall at the cost of increasing the size of the support set of the patterns. The much larger size of the k-nearstar core patterns appear to be not very suitable to identify the experts who may be few in number in a particular domain of expertise.

Graph	$\|P\|$	$\|V_P^A\|$	$\|V_P^{IE}\|$	Precision	Recall
G_{co}(6-core)	17875	2258	441	0.091	0.412
G_{co}(20-nearstar)	14530	4619	667	0.109	0.753
$G_{A\to P}$(3-3-BHA)	56934	5878	801	0.096	0.794

Table 1: See notations in figure 3

Finally, the F-constrained-h-a-BHA core property has been explored on the bipartite graph $G_{A\to P}$. In order to determine whether the parameters h and a should be chosen balanced ($h = a$) or unbalanced ($h \neq a$), the mean out-degree from the 'author'/'hubs' vertices $\bar{H}$ and the mean in-degree entering from the

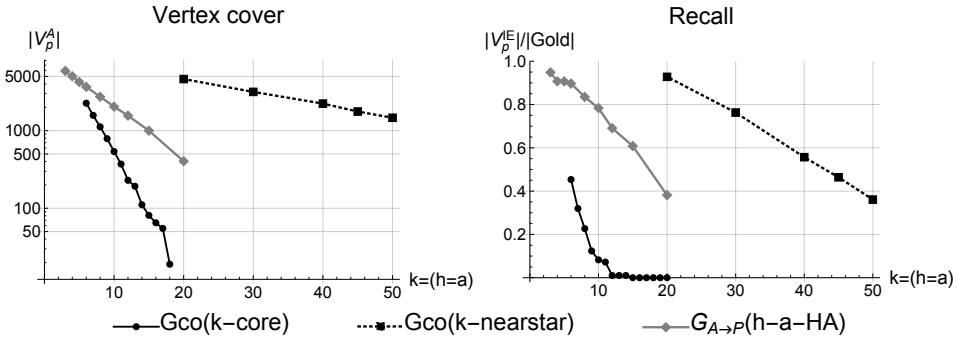

Fig. 3: P: the number of patterns, V_P^A the 'author' vertices covered by the extensions of P. V_P^{IE} the 'author' vertices labelled with the concept IE covered by the extension of P. Left: The cover of the authors of G_{co} (resp. $G_{A\to P}$) by the patterns enumerated with the k-core and the k-nearstar properties (resp. balanced h-a-HA core property). Right: The recall over the 97 gold standard IE authors covered by P.

'publication'/'authorities' vertices $\bar{A}$ of this graph were calculated. The values of $\bar{H}$ and $\bar{A}$ found are similar ($\bar{H} = 9.7$ and $\bar{A} = 7.8$) indicating that the graph is globally balanced. Therefore it is chosen to explore in the following of this study balanced values of (h, a) pairs of parameters.

As shown in figure 3 and table 1 the F-constrained-h-a-BHA core property offers a better recall and a similar precision than the two other properties. Moreover it allows to use rather low constraint parameters allowing to identify small groups of experts. As expected, the best recall is obtained for the balanced pair $(h, a) = (3, 3)$. For these reasons, the rest of the study will focus on the study of the F-constrained-3-3-BHA core property patterns.

5.4 Identifying the experts with F-constrained-3-3-BHA core property

The enumeration of the F-constrained-3-3-BHA core closed patterns on the bipartite graph $G_{A\to P}$ leads to 56934 patterns. As it is impossible to extract any knowledge from such a large number of patterns it is necessary to reduce the number of patterns. First, in order to make the expertise associated with the expert sets specific enough, the size of the support set of the patterns of interest is limited to 100 authors at maximum. The 14509 patterns that have more than 100 authors are removed. The remaining 42425 patterns are still too numerous so we seek to reduce their number by selecting a interesting subset of them.

The patterns are ordered in decreasing local modularity[19] and are $g\beta$-selected for different β-values. As expected the number of selected patterns S decreases sharply when the distance criterion β increases. However the number of 'author' vertices V_S^A present in the union of the extensions of the patterns S

remains rather stable indicating that the selection preserves the coverage of the graph vertices. Similarly, the number of selected patterns containing the semantic concept IE (S_{IE}) is also greatly reduced to 12 patterns for $\beta = 0.8$. Interestingly, the number of authors $V^A_{S_{IE}}$ covered by the patterns S_{IE} decreases rather slightly from 215 authors for $\beta = 0.0$ to 163 for $\beta = 0.8$ preserving the precision and the recall with respect to the gold standard.

The best precision (0.28) is obtained for $\beta = 0.8$ which selects 12 patterns containing the semantic concept IE with a recall of 0.46. These patterns are listed in table 3 and described in table 2.

Pattern	1	2	3	4	5	6	7	8	9	10	11	12
$\lvert V_S \rvert$	167	82	101	35	26	34	25	14	10	7	6	6
$\lvert V^A_S \rvert$	100	51	48	19	16	18	14	8	6	3	3	3
Precision	0.31	0.45	0.25	0.42	0.38	0.61	0.64	0.50	0.67	0.00	1.00	-
Recall	0.32	0.24	0.12	0.08	0.06	0.11	0.09	0.04	0.04	0.00	0.03	0.00

Table 2: Metrics on the 12 S_{IE} patterns $g\beta$-selected with $\beta = 0.8$ from the F-constrained-3-3-BHA core patterns.

P	Description
1	inf._extr., $\Delta_A \cap]2007, \infty]$, $Y_P \geq 1999$
2	inf._extr., nat._lang._processing, $\Delta_A \cap]1992, 2007]$, $Y_P \leq 2007$
3	inf._extr. $\Delta_A \supseteq]-\infty, 1999]$, $Y_P \leq 2004$
4	inf._extr., languages.., $\Delta_A \supseteq]1999, \infty]$, $Y_P \leq 1999$
5	inf._extr., user_information, $\Delta_A \cap]1992, 1999]$, $1992 < Y_P \leq 2007$
6	inf._extr., learning, $\Delta_A \supseteq]1999, 2007]$, $1992 < Y_P \leq 2007$
7	inf._extr., $\Delta_A \supseteq]1992, 2007]$, $2004 < Y_P \leq 2007$
8	inf._extr., cond._random_field, $\Delta_A \cap]2004, 2007]$, $1999 < Y_P \leq 2007$
9	inf._extr., languages, named_entity_recog., $\Delta_A \supseteq]1999, \infty]$, $1992 < Y_P \leq 2007$
10	inf._extr., correl._analysis, $\Delta_A \supseteq]1992, \infty]$, $Y_P \leq 1999$
11	inf._extr., languages, nat._lang._processing, $\Delta_A \supseteq]1992, 2004]$, $1992 < Y_P \leq 1999$
12	inf._extr., named_entity_recognition, $\Delta_A \cap]1992, 2004]$, $1999 < Y_P \leq 2004$

Table 3: The 12 S_{IE} patterns $g\beta$-selected with $\beta = 0.8$ from the F-constrained-3-3-BHA core patterns.

The 9th pattern has a precision of 0.67. It is composed of 6 authors and 4 cited publications. 4 of the authors belong to the gold standard. A manual evaluation of the two remaining authors showed that they are also experts even if they are not listed in the gold standard.

In order to test our method on another domain, we followed the same procedure (F-constrained-3-3-BHA core closed pattern enumeration followed by a $g\beta$-selection for $\beta = 0.8$) and extracted the patterns related to the semantic concept 'sentiment_analysis' (SA). 2 patterns are identified : P_1 = *sentiment_ analysis*, *semantic_ orientation*, $\Delta_A \supseteq]2004, \infty]$, $1999 < Y_P \leq 2007\}$ and $P_2 = \{$semantic_analysis, $\Delta_A \cap]2004, \infty]$, $Y_P > 1999\}$. Manual validation of the experts

obtained on these patterns with the chapter named "Sentiment Analysis" (SA) belonging to the book from which IE is issued. 3 authors support P_1 while 49 authors support P_2 and 3 and 38 authors are also found in the references of SA respectively.

6 Discussion

In this paper, we have showed that our approach combining semantic annotation and graph mining while taking into account peer validation for expert finding enables to identify relevant sets of experts along their shared expertise from textual documents. Thus, we suggest to extend the definition of expertise. We recommend that determining than an individual is an expert on a specific competency should not only rest on the assignment of their competencies but also require a validation process through a professional network. We have also introduced gβBi-Pattern set selection and restricted bi-pattern mining for the sake of our experiments on core closed pattern mining on bipartite graphs. Future work should focus on generalizing patterns which are very specific. Our idea is to extend a pattern by suggesting frequency analysis of patterns at distance $< \beta$ in vertices belonging to its extension.

References

1. Fotis Draganidis and Gregoris Mentzas. Competency Based Management: a Review of Systems and Approaches. *Information Management & Computer Security*, 14(1):51–64, 2006.
2. Mohammed Zuhair Al-Taie, Seifedine Kadry, and Adekunle Isiaka Obasa. Understanding Expert Finding Systems: Domains and Techniques. *Social Network Analysis and Mining*, 8(1):57, 2018.
3. Paolo Cifariello, Paolo Ferragina, and Marco Ponza. Wiser: A Semantic Approach for Expert Finding in Academia Based on Entity Linking. *Information Systems*, 82:1–16, 2019.
4. Milena Angelova, Veselka Boeva, and Elena Tsiporkova. Advanced Data-driven Techniques for Mining Expertise. In *30th Annual Workshop of the Swedish Artificial Intelligence Society (SAIS)*, number 137, pages 45–52, 2017.
5. Dawit Yimam-Seid and Alfred Kobsa. Expert-Finding Systems for Organizations: Problem and Domain Analysis and the DEMOIR Approach. *Journal of Organizational Computing and Electronic Commerce*, 13(1):1–24, 2003.
6. Min-Chun Yu, Yen-Chun Jim Wu, Wadee Alhalabi, Hao-Yun Kao, and Wen-Hsiung Wu. Researchgate: An Effective Altmetric Indicator for Active Researchers? *Computers in Human Behavior*, 55:1001–1006, 2016.
7. F. Xia, W. Wang, T. M. Bekele, and H. Liu. Big Scholarly Data: A Survey. *IEEE Transactions on Big Data*, 3(1):18–35, 2017.
8. Georgeta Bordea. *Concept Extraction Applied to the Task of Expert Finding*, pages 451–456. 2010.
9. Francesco Osborne, Enrico Motta, and Paul Mulholland. Exploring Scholarly Data with Rexplore. In *International Semantic Web Conference*, pages 460–477, 2013.

10. Francesco Osborne, Giuseppe Scavo, and Enrico Motta. Identifying Diachronic Topic-based Research Communities by Clustering Shared Research Trajectories. In *European Semantic Web Conference*, pages 114–129, 2014.
11. Aditya Pal and Joseph A Konstan. Expert Identification in Community Question Answering: Exploring Question Selection Bias. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 1505–1508, 2010.
12. Jie Li, Harold Boley, Virendrakumar C Bhavsar, and Jing Mei. Expert Finding for eCollaboration Using FOAF with RuleML Rules.
13. Yizhou Sun, Jie Tang, Jiawei Han, Cheng Chen, and Manish Gupta. Co-evolution of Multi-typed Objects in Dynamic Star Networks. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):2942–2955, 2013.
14. Soumyajit Ganguly and Vikram Pudi. Paper2vec: Combining Graph and Text Information for Scientific Paper Representation. In *European Conference on Information Retrieval*, pages 383–395, 2017.
15. Henry Soldano and Guillaume Santini. Graph Abstraction for Closed Pattern Mining in Attributed Networks. In *European Conference in Artificial Intelligence (ECAI)*, volume 263, pages 849–854, 2014.
16. Henry Soldano, Guillaume Santini, Dominique Bouthinon, and Emmanuel Lazega. Hub-Authority Cores and Attributed Directed Network Mining. In *International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1120–1127, 2017.
17. Henry Soldano, Guillaume Santini, Dominique Bouthinon, Sophie Bary, and Emmanuel Lazega. Bi-pattern Mining of Attributed Networks. *Applied Network Science*, 4(1):37, 2019.
18. Henry Soldano, Guillaume Santini, and Dominique Bouthinon. Attributed graph pattern set selection under a distance constraint. In *COMPLEX NETWORKS (2)*, volume 882 of *Studies in Computational Intelligence*, pages 228–241. Springer, 2019.
19. Martin Atzmueller, Henry Soldano, Guillaume Santini, and Dominique Bouthinon. Minerlsd: efficient mining of local patterns on attributed networks. *Applied Network Science*, 4(1):43, 6 2019.
20. Kata Gábor, Haïfa Zargayouna, Davide Buscaldi, Isabelle Tellier, and Thierry Charnois. Semantic Annotation of the ACL Anthology Corpus for the Automatic Analysis of Scientific Literature. Proceedings of the LREC 2016 Conference, 2016.
21. Steven Bird, Robert Dale, Bonnie J Dorr, Bryan Gibson, Mark Thomas Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R Radev, and Yee Fan Tan. The ACL Anthology Reference Corpus: a Reference Dataset for Bibliographic Research in Computational Linguistics. 2008.
22. Angelo Salatino, Thiviyan Thanapalasingam, Andrea Mannocci, Francesco Osborne, and Enrico Motta. The Computer Science Ontology: A Large-Scale Taxonomy of Research Areas. 2018.
23. Angelo Salatino, Thiviyan Thanapalasingam, Andrea Mannocci, Francesco Osborne, and Enrico Motta. Classifying Research Papers with the Computer Science Ontology. 2018.
24. Jerry R Hobbs and Ellen Riloff. Information Extraction. *Handbook of Natural Language Processing*, 2, 2010.