

Community Detection in Multi-relational Bibliographic Networks

Soumaya Guesmi^(✉), Chiraz Trabelsi, and Chiraz Latiri

LIPAH, Faculty of Sciences of Tunis, Université de Tunis El Manar, Tunis, Tunisia
soumaya.guesmi@fst.rnu.tn

Abstract. In this paper, we introduce a community detection approach from heterogeneous multi-relational network which incorporate the multiple types of objects and relationships, derived from a bibliographic networks. The proposed approach performs firstly by constructing the relation context family (RCF) to represent the different objects and relations in the multi-relational bibliographic networks using the Relational Concept Analysis (RCA) methods; and secondly by exploring such RCF for community detection. Experiments performed on a dataset of academic publications from the Computer Science domain enhance the effectiveness of our proposal and open promising issues.

Keywords: Multi-relational bibliographic networks · Community detection · RCA

1 Context and Motivation

The primary focus of this work is to extract emergent academic community structure from the bibliographic through the analysis of the different relationships among the multi-relational bibliographic data. Although research attention on heterogeneous networks representation and efficient topological algorithm design, a much more fundamental issue concerning the exploration of the heterogeneous organization infrastructure and communities detection have not been skilfully addressed. Indeed, A wide range of approaches have been proposed in the literature for communities detection in heterogeneous networks. However, they have deeply focused on topological properties of these networks, ignoring the embedded semantic information. To overcome this limitation, in recent years, Formal Concept Analysis (FCA) techniques are used for a conceptual clustering. Using FCA aims to extract communities preserving knowledge shared in each community. In such FCA based approaches, the inputs are bipartite graphs and the output is a Galois hierarchy that reveals communities semantically defined with their shared knowledge or common attributes. Vertices are designed as lattice extents and edges are labeled by lattice intents (*i.e.*, shared knowledge). However, a Galois hierarchy is not a satisfactory scheme since an exponential number of communities may be obtained. Therefore, reduction methods should be introduced. In fact, only very few researches have actually focused on this difficulty [4]. The authors in [5] used the

iceberg method as well as the stability method as a Galois lattice reduction methods. Authors in [3] identify concepts with frequent intents above a set threshold. The main limit of this purpose, that some important concepts may be overlooked. Brandes et al. [1] combine both the iceberg and stability methods, it's argued that this approach yields good results for extracting pertinent communities based on concepts. As it's described in the survey conducted by Planti and Crampes [4], discovering communities based on FCA techniques is the most accurate, because it extracts communities using their precise semantics. Nonetheless, they fall short of giving simple and practical results. Therefore, a new research challenge consists on detecting communities from heterogeneous multi-relational networks. In order to discover communities with a well defined set of properties, we first need to extract the corresponding relations among multiple existing relations. In this paper, we introduce a query navigation approach based on the use of the RCA techniques [6] designed within a multiple academic databases for hidden relationships (or links) detection. This will have significant impact, it can help foster new collaborative teams, help with expertise discovery and in the long term, guide research teams reorganization consistency with collaboration patterns.

The paper is organized as follows. In the next section, we describe our community detection approach. Section 3 presents our experimental results, while Sect. 4 summarizes our contributions.

2 Proposed Community Detection Approach

In this section, we present our community detection approach which aims to model and to extract academic community structure from multi-relational bibliographic data. In order to achieve these goals, the proposed approach relies on two main stages: the multi-relational bibliographic hypergraph modelling stage; and the query navigation for communities discovering stage. We firstly proceed by describing the preliminary concepts of our proposal.

A. Preliminary Concepts

• **Formal context:** is a triplet $\mathcal{K} = (\mathcal{O}, \mathcal{A}, \mathcal{I})$, where \mathcal{O} represents a finite set of objects, \mathcal{A} is a finite set of items (or attributes) and \mathcal{I} is a binary (incidence) relation (*i.e.*, $\mathcal{I} \subseteq \mathcal{O} \times \mathcal{A}$). Each couple $(o, a) \in \mathcal{I}$ expresses that the object

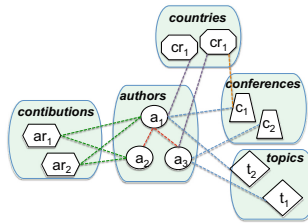


Fig. 1. An example of a multi-relational bibliographic hypergraph.

$o \in \mathcal{O}$ contains the item $a \in \mathcal{A}$. \mathcal{O} is called one-valued context. A worth of interest link between the power-sets $\mathcal{P}(\mathcal{A})$ and $\mathcal{P}(\mathcal{O})$ associated respectively to the set of items \mathcal{A} and the set of objects \mathcal{O} [2].

- **Formal concept:** A pair $c = (O, A) \in \mathcal{O} \times \mathcal{A}$, of mutually corresponding subsets, i.e., $O = \psi(A)$ and $A = \phi(O)$, is called a formal concept, where O is called *extent* of c and A is called its *intent*.

- **A partial order:** on formal concepts, w.r.t. set inclusion [2], is defined as: $\forall c_1 = (O_1, A_1)$ and $c_2 = (O_2, A_2)$ two formal concepts, $c_1 \leq c_2$ if $O_2 \subseteq O_1$, or equivalently $A_1 \subseteq A_2$.

- **Galois concept lattice:** Given a context \mathcal{K} , the set of formal concepts \mathcal{C} is a complete lattice $\mathcal{L}_{\mathcal{C}} = (\mathcal{C}, \leq)$, called *Galois (concept) lattice*, when \mathcal{C} is considered with set inclusion between concepts intents (or extents) [2].

- **Relational Context Family (RCF):** is a pair (K, R) where $K = \{\mathcal{K}_i\}_{i=1, \dots, n}$ is a set of (object-attribute) contexts $\mathcal{K}_i = (\mathcal{O}_i, \mathcal{A}_i, \mathcal{I}_i)$ and $\{r_{j,l}\}_{j,l \in \{1, \dots, n\}}$ is a set of relational (object-object) contexts $r_{j,l} \subseteq \mathcal{O}_j \times \mathcal{O}_l$, where \mathcal{O}_j (called the domain of $r_{j,l}$) and \mathcal{O}_l (called the range of $r_{j,l}$) are the object sets of the contexts \mathcal{K}_j and \mathcal{K}_l , respectively. \mathcal{O}_j is called the domain of $r_{j,l}$ ($\text{dom}(r_{j,l})$) and \mathcal{O}_l is called the range of $r_{j,l}$ ($\text{ran}(r_{j,l})$) [6].

A function rel is associated with a RCF which maps a context $\mathcal{K} = (\mathcal{O}, \mathcal{A}, \mathcal{I}) \in K$ to the set of all relations $r \in R$ starting at its object set $\mathcal{K} : rel(\mathcal{K}) = \{r \in R, \text{ where } \text{dom}(r) = \mathcal{O}\}$. Hence, given a relation r and a quantifier f chosen within the set $F = \{\forall, \exists, \forall\exists, \geq, \geq_f, \leq, \leq_f\}$. k maps an object set from $\text{ran}(r)$ to an object set from $\text{dom}(r)$ as $k : F \times R \times \cup_{i=1, \dots, n} \mathcal{P}(\mathcal{O}_i) \rightarrow \cup_{i=1, \dots, n} \mathcal{P}(\mathcal{O}_i)$ [6]. Scaling a context along a relation consists in integrating the relation to the context in the form of one-valued attributes using a scaling operator. A context is scaled upon all the relevant relations originating from the context by augmenting \mathcal{K} with all the resulting relational attributes. Thus, an object owns an attribute depending on the relationship between its link set and the extent of the concept, i.e., the instances of a relation r , say $r_k(o_i, o_j)$, where $o_i \in \mathcal{O}_i$ and $o_j \in \mathcal{O}_j$, are called links. The evolution of each context $\mathcal{K}_i \in K$ from the input RCF yields a sequence \mathcal{K}_i^p whose zero member $\mathcal{K}_i^0 = (\mathcal{O}_i^p, \mathcal{A}_i^p, \mathcal{I}_i^p)$ is the input context \mathcal{K}_i itself. From there on, each subsequent member is the complete relational expansion of the previous one upon the relations r from $rel(\mathcal{K}_i)$. This yields a global sequence of context sets \mathcal{K}^p and the corresponding sequence of lattice sets, called the Concept Lattice Family (CLF). Thus, the concept lattice family is a set of lattices that correspond to the formal contexts, after enriching them with relational attributes.

In this work, we consider the exists scaling. Hence, let $r_{ij} \subseteq \mathcal{O}_i \times \mathcal{O}_j$ be a relational context. The exists scaled relation r_{ij}^{\exists} is defined as $r_{ij}^{\exists} \subseteq \mathcal{O}_i \times B(\mathcal{O}_j, \mathcal{A}, \mathcal{I})$, such that for an object o_i and a concept $c:(o_i, c) \in r_{ij}^{\exists} \Leftrightarrow \exists x, x \in o_i' \cap \text{Extent}(c)$.

B. Multi-relational Bibliographic Hypergraph Model

Three concepts are involved in our model: object context, relation context, and concept lattice family. As illustrated in Fig. 1, a set of authors $\{a_1, a_2, \dots, a_n\}$, locates in a given country $\{cr_1, cr_2, \dots, cr_p\}$, work closely with each other, under

K_{Authors}											
Author		ai	aj	ak	al	am	an	ao	ap	aq	ar
O. Willum		x									
D. Wei			x								
Wenhui Wu				x							
Manuel Will					x						
Mariya Das						x					
Roberto Gallo							x				
Henrique Kawakami								x			
Ahmed Seffah									x		
Peter M. Maurer										x	

K_{Contributions}											
Contribution		ar1	ar2	ar3	ar4	ar5	ar6	ar7	ar8	ar9	ar10
AR1		x									
AR2			x								
AR3				x							
AR4					x						
AR5						x					

K_{Countries}											
Country		c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
China		x									
USA			x								
Canada				x							
France					x						

K_{Conferences}											
Conference		c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
SIGIR		x									
DEXA			x								
Wireless Networks				x							
ICNP					x						
TIP						x					
CCS							x				
TIFS								x			

K_{Topics}											
Topic		dt	nt	is	mm						
Datamining			x								
Network				x							
Information Security					x						
Multimedia						x					

Fig. 2. Top. The objects contexts. Bottom. The relations contexts.

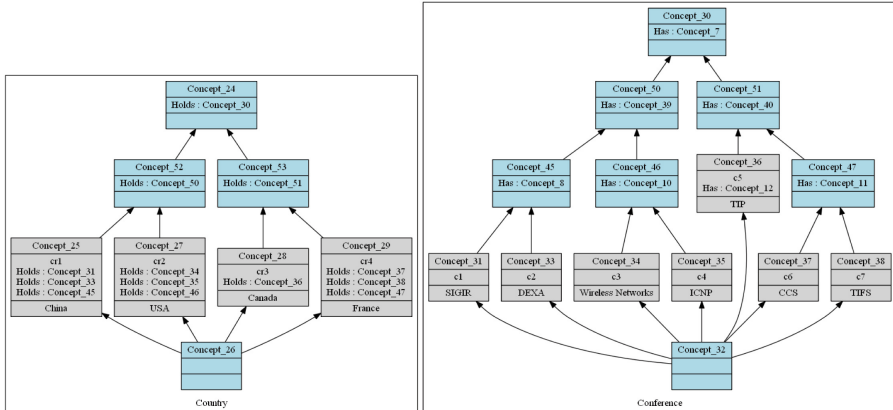


Fig. 3. Country and Conference lattices.

different topics $\{t_1, t_2, \dots, t_k\}$; some of them share scientific contributions (contributions $\{ar_1, ar_2, \dots, ar_m\}$ within a set of conferences $\{c_1, c_2, \dots, c_l\}$). To generally describe such collaboration data, we define an object context as a set of objects or entities of the same type, *e.g.*, an author context is a set of authors and define a relation context as the interactions among objects contexts, *e.g.*, (author, topic) relation, (country, conference) relation., etc. We use a relational concept family to describe the relations contexts and the objects

contexts constructed from a multi-relational bibliographic hypergraph. Figure 1 depicts the data schema of the handled multi-relational bibliographic hypergraph. The relational concept family is made of 5 objects contexts: $\mathcal{K}_{Authors}$, $\mathcal{K}_{Countries}$, $\mathcal{K}_{Conferences}$, \mathcal{K}_{Topics} , $\mathcal{K}_{Contributions}$; and 5 relations contexts: $r_{Locates}$, r_{Holds} , r_{Has} , $r_{Discusses}$ and $r_{Addressed-By}$. We report in Fig. 2 (Top) these 5 objects contexts and in Fig. 2 (Bottom) the 5 related relations contexts.

The overall process of RCA follows a multi-FCA method [6] which allows to build a set of lattices called Concept Lattice Family (*CLF*). It's an iterative process which generates at each step a set of concept lattices. First, the process constructs concept lattices using the objects contexts only. Then, in the following steps, it concatenates objects contexts with the relations contexts based on the existential scaling operator that produce scaled relations. Hence, the exists scaled relation translates the links between objects into conventional FCA attributes and extracts a collection of lattices whose concepts are linked by relations. Figure 3 depicts an example of Country and Conference lattices of the generated *CLF*.

C. Query Navigation for Communities Discovering

The second stage of the proposed approach aims to extract a set of academic communities by performing the following three steps:

- **Step 1: users' relational query submission:** the aim of this step is to transform the submitted user query to a Relational Query RQ which is composed of several Simple Queries (SQ). Hence, for a context $K = (A, O, I)$, a simple query denoted by $SQ = \{o_q\}$, is a set of objects satisfying the query (or the answer set) with $o_q \subset O$.

Definition 1 (Relational Query). A Relational Query $RQ = \{rq_0, rq_1, \dots, rq_m\}$ on a relational context family (K, R) is a triplet $RQ = (q'_s, r_{st}, q'_t)$ with:

- q'_s and q'_t , source query and target query respectively, are a set of SQ .
- r_{st} is the relation between q'_s and q'_t . It leads one-to-one mapping between q'_s and q'_t .

- **Step 2: concept Lattice Family Exploration:** to explore the concept lattice family, we have to construct a query path QP which allows to know the path that we have to follow and specify the source and the target lattices.

Definition 2 (Query Path). Let $QP = \{qp_0, qp_1, \dots, qp_n\}$ and qp_i is a pair $((q_s, L_s), (q_t, L_t))$ where L_s and $L_t \subset CLF$, the source and target lattices respectively. The Query Path QP is the inverse order of the relational query. It means $qp_0 = rq_m$ and $qp_n = rq_0$; with $q_{s0} = q'_{tm}$ and $q_{t0} = q'_{sm}$

- **Step 3: community detection:** in order to detect academic communities, we propose a new method called *Querying_Navigation* that leads to navigate between Galois Lattices based on the extracted query path QP . It takes as input the query path $QP = qp_k$ with $qp = ((qp_s, L_s), (qp_t, L_t))$ and outputs the identified community as an answer to the user query Q . *Query_Navigation*

starts by handling all concepts C of the source lattice L_s , in order to extract the corresponding concepts (C_i) of the initial query path qp_0 . *Query_Navigation* proceeds by identifying the concept extent of the lattice L_s and then extracts the concepts that contains an extent related to the query qp_s . The result of the initial phase is a set of concepts C_i that respond to the query qp_s . The second phase of *Query_Navigation*, consists in generating iteratively a set of concepts containing the set of concepts (C_i) extracted in the initial phase. It consists on handling the corresponding concept intent of the lattice L_t , for extracting the set of concepts (C_{i+1}) containing the C_i . If there is no more query path to be explored, *Query_Navigation* extracts the extent of the last selected concept (C_k). This set of C_k extent represents the set of individuals that constitutes the academic community returned to the user.

3 Experimental Evaluation

We collect data from two bibliographic databases. We use the well known database DBLP and we access on AMiner database for taking keywords, institutions and research topics in order to complete our conceptual hypergraph model. We keep only four research topics (Data Mining, Computer Network, Artificial Intelligence, Human Computer, Computer Graphics) and we pick only a few representative conferences for the five areas (11 conferences). The built dataset contains 914 contributions and 336 authors since 2010. The *Query_Navigation* algorithm is developed in JAVA and tested on a Windows 7 with Intel core i5 2.4 GHz and 8 GB of Ram.

Baseline Model: for enhancing the effectiveness of our approach, we have selected the most popular baseline communities structure which suggests communities as a set of authors belonging to the same affiliation. To carry out our experiments, we consider two simple queries (Q3 and Q4) and two relational queries (Q1 and Q2). We study whether our approach is able to capture the hidden relations between authors and if it can responds to different type of queries:

Q1: 4 entities, *i.e.*, Authors, Countries, Conferences and Topics; and 3 relations, *i.e.*, Locates, Holds and Has.

Q2: 3 entities, *i.e.*, Authors, Countries and Conferences; and 2 relations, *i.e.*, Locates and Holds.

Q3: 2 entities, *i.e.*, Authors and Countries; and 1 relation, *i.e.*, Locates.

Q4: 2 entities, *i.e.*, Authors and Topics; and 1 relation, *i.e.*, Discusses.

Furthermore, we consider two different ground truths [8]. The first ground truth *GT1*: each explicit authors' topic in the dataset is a ground truth community, it contains authors nodes which share the same topic. The second ground truth *GT2*: each explicit author conference is a ground truth community, it contains authors nodes which participate in the same conference.

Effectiveness of our approach: the performance is assessed by the measures of Recall, Precision and $F\beta$ -measure, computed over all vertices [7]. These measures attempt to estimate whether the prediction of this vertices in the same community was correct. Given a set of algorithmic communities C and the ground truth communities S , precision indicates how many vertices are actually in the same ground truth community ($Precision = \frac{|T \cap S|}{|T|}$). The Recall indicates how many vertices are predicted to be in the same community in a retrieved community ($Recall = \frac{|T \cap S|}{|S|}$), and $F\beta$ -measure is the harmonic mean of Precision and Recall ($F\beta_measure = \beta \times \frac{Precision \times Recall}{Precision + Recall}$ where $\beta \in \{1, 2\}$).

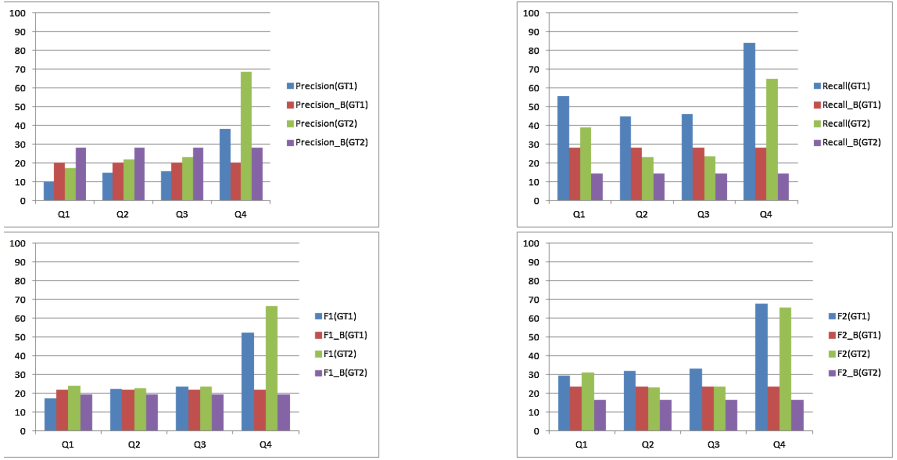


Fig. 4. Average score of the Precision, Recall, F1-measure and F2-measure of our approach *vs.* those of the baseline (B).

Thus, according to the sketched histograms in Fig. 4, we can point out that our approach outperforms the baseline. In fact, as expected, the Recall values of the baseline are much lower than those achieved by our approach among the two ground truths ($GT1$ and $GT2$). As we show, the average Recall achieves 83.87% and 65.02% comparing with the baseline which has 28.31% and 14.58% in term of Recall *vs.* an exceeding about 55.5% and 50.4% over the query $Q4$ among the two ground truths respectively. Indeed, in term of F2-measure our approach (67.61%, 65.7%) outperforms considerably the baseline (23.44%, 16.5%) over the query $Q4$ among $GT1$ and $GT2$ respectively, in this case we can say that the baseline have only a small number of communities detected fairly well and not many detected communities reflect to the ground truth communities.

However, the percentage of Precision for the baseline outperforms slightly our approach according to $Q1$, $Q2$ and $Q3$. Whereas, for $Q4$, our approach has an average of 68.57% showing a drop of 28.31% *vs.* an exceeding about 40.2% against the baseline. A significant observation shows that the relational query

$Q1$ have better Recall (55,68 %) than that of the simple query $Q3$ and that of the relational query $Q2$ (44,85 %). We can conclude that the relational query improves the community structure and leads to extract relevant communities. Hence, considering four different queries, our approach outperforms the baseline in terms of Recall, F1_measure and F2_measure often by a large margin in the Recall score.

4 Conclusion

In this paper, we have presented a novel approach for academic communities discovering from heterogeneous multi-relational bibliographic networks. Our approach takes into account the different entities and relationships expressed in a bibliographic hypergraph. Indeed, we made use of the RCA techniques to model and explore heterogeneous multi-relational bibliographic network via a new introduced method, called: *Query_Navigation*, for academic communities detection. As part of our future work, we plan to address a more diversified set of queries by the integration of other quantifier such as \forall quantifier.

Acknowledgements. This work is partially supported by the French-Tunisian project PHC-Utique RIMS-FD 14G 1404.

References

1. Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., Wagner, D.: On modularity clustering. *IEEE Trans. Knowl. Data Eng.* **20**, 172–188 (2008)
2. Ganter, B., Wille, R.: *Formal Concept Analysis: Mathematical Foundations*. Springer, Berlin (1999)
3. Jay, N., Kohler, F., Napoli, A.: Analysis of social communities with iceberg and stability-based concept lattices. In: Medina, R., Obiedkov, S. (eds.) *ICFCA 2008*. LNCS (LNAI), vol. 4933, pp. 258–272. Springer, Heidelberg (2008)
4. Planti, M., Crampes, M.: Survey on social community detection. In: *Social Media Retrieval, Computer Communications and Networks*, pp. 65–85 (2013)
5. Roth, C., Obiedkov, S.A., Kourie, D.G.: On succinct representation of knowledge community taxonomies with formal concept analysis. *Int. J. Found. Comput. Sci.* **19**, 383–404 (2008)
6. Rouane-Hacene, M., Huchard, M., Napoli, A., Valtchev, P.: Relational concept analysis: mining concept lattices from multi-relational data. *Ann. Math. Artif. Intell.* **67**(1), 81–108 (2013)
7. Song, S., Cheng, H., Yu, J.X., Chen, L.: Repairing vertex labels under neighborhood constraints. *PVLDB* **7**, 987–998 (2014)
8. Yang, J., Leskovec, J.: Defining and evaluating network communities based on ground-truth. In: *ICDM*, pp. 745–754. IEEE Computer Society (2012)

Database and Expert Systems Applications
27th International Conference, DEXA 2016, Porto,
Portugal, September 5-8, 2016, Proceedings, Part II
Hartmann, S.; Ma, H. (Eds.)
2016, XXVII, 465 p. 155 illus., Softcover
ISBN: 978-3-319-44405-5