Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 96 (2016) 627 – 636

Procedia Computer Science

20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems

CoMRing: A framework for Community detection based on Multi-Relational querying exploration

Soumaya Guesmi, Chiraz Trabelsi, Chiraz Latiri

[a]*LIPAH, Université de Tunis El Manar, Faculty of Sciences of Tunis, Tunisia*

Abstract

Community detection in multi-relational bibliographic networks is an important issue. There has been a surge of interest in community detection focusing on analyzing the linkage or topological structure of these networks. However, communities identified by these proposed approaches, commonly reflect the strength of connections between networks nodes and neglect considering the interesting topics or the venues, i.e., conferences or journals, shared by these community members, i.e, authors. To tackle this drawback, we present in this paper a new approach called *CoMRing* for community detection from heterogeneous multi-relational network which incorporate the multiple types of objects and relationships, derived from a bibliographic networks. We firstly propose to construct the Concept Lattice Family (CLF) to model the different objects and relations in the multi-relational bibliographic networks using the Relational Concept Analysis (RCA) methods. Then after we introduce a new method, called $Query_{Exploration}$, that explores such CLF for community detection. Carried out experiments on real-datasets enhance the effectiveness of our proposal and open promising issues.

Peer-review under responsibility of KES International

Keywords: Multi-Relational bibliographic networks; Community detection; Relational Concept Analysis; Multi-Relational querying.

1. Introduction

Social networks are often deemed as heterogeneous networks since they involve multiple typed objects and multiple typed links denoting different relations. Indeed, heterogeneous information networks are not only more aligned with the real world, but also contain a wealth of information, and therefore have the potential to provide us with more accurate and implicit knowledge. An example of an heterogeneous information network is the bibliographic information network. In addition to having multi-relational characteristics, bibliographic information network also contains a wealth of information networks, such as the co-author network, the citation network including titles, authors, affiliations, keywords, venues(conferences), publication data and other entities, which are connected to form a multi-dimensional heterogeneous network. Hence, following different objectives in bibliometrics (relationship study-

E-mail address: soumaya.guesmi@fst.utm.tn

Peer-review under responsibility of KES International
doi:10.1016/j.procs.2016.08.244

ing, ranking, community mining.), dedicated approaches made use of various kinds of techniques such as: Statistics, Data Mining, Formal Concept Analysis, Graph Theory, *OLAP* analysis. Among these different types of analysis, we are more interested by Formal Concept Analysis (FCA) techniques. Despite research attention on FCA, and efficient topological algorithm design, a much more fundamental issue concerning the design of the heterogeneous relational infrastructure has not been addressed. In fact, FCA techniques cannot satisfy relational analysis on bibliographic information data, because they do not consider the relationships between attributes. Central to FCA is the notion of formal context, which is defined as a triple $\mathcal{K} = (O, \mathcal{A}, \mathcal{I})$, where O is a finite set of objects; $\mathcal{A}$ is a finite set of attributes; $\mathcal{I}$ is a binary relation between O and $\mathcal{A}$[8]. In short, the formal context is a binary relation between a set of objects and a set of attributes. However, besides the binary relations between objects and attributes, in the real world each object also has some connections with other objects, which induces the binary relations between any two objects. For example, the author-conference relation in a bibliographic network. Hence, it is also necessary to introduce these real-world relations to the contexts of FCA. Although we can deduce some relations between any two objects author, conference $\in O$ from relation $\mathcal{I}$, for instance, author and conference have the same subset of attributes in context $\mathcal{K}$. But for many real-world scenarios, it is more meaningful to define some binary relations on O in the context, and discuss the direct connections between any two objects of O by virtues of these relations. To expand the scope of application for FCA, recent researchers have introduced the Relational Concept Analysis (RCA) as an extension of FCA techniques[11,3], which allow to design relational contexts to represent attributes relationships. 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. 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 called $Query_{Exploration}$ based on the use of the RCA techniques designed within a multiple academic database 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.

2. Related Work

Previous community detection methods are overwhelmingly focused on the homogeneous multi-relational network which contain different types of edges and only one type of nodes. However, many real-world networks are naturally described as heterogeneous multi-relational hypergraphs which contain different types of nodes and edges. Recently, many researches have addressed community detection in heterogeneous multi-relational networks. Lin et *al.,*[6] proposed a method based on tensor factorization, called MetaFac(MetaGraph Factorization). Authors suppose that a community contains different types of nodes, thus dividing nodes of different types separately means that nodes of different types have the same number of communities. The main limit of the proposed method is that we rarely see this situation in real-world scenario. Furthermore, in Song et *al.,*[13], authors construct a General Heterogeneous Network Clustering algorithm called GenClus to integrate the incomplete attribute information and the network structure information. Zhang et *al.,*[15] proposed a method based on matrix factorization that combine user-generated contents and friendship networks to discover user communities sharing common content interests densely connected. However, a notable drawback of these two approaches is that they require a priori knowledge about the number of communities. This limits their usage in deducing (inferring) the latent organization of a real system. Authors in[7] proposed a new approach for detecting communities in heterogeneous multi-relational network which follows the line of the modularity optimization method. Note that the drawback of this approach is the resolution limit, that may prevent it from detecting communities which are comparatively small in large-scale networks.

Actually, 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[2]. 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[9]. The authors in[10] used the iceberg method as well as the stability method as a Galois lattice reduction methods. Authors in[5] 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, its argued that this approach yields good results for extracting pertinent communities based on concepts. As its described in the survey conducted by Planti and Crampes[9], 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, we introduce in this paper a new approach for academic community discovering based on Relational Concept Analysis(*RCA*) because of its multi-relation nature[11]. In order to find out the multi-relational academic community structure across multiple network dimensions, we have to integrate the information from all dimensions. In particular, we firstly propose to use the *RCA* techniques[11] to model the different relations and entities embedded in the hypergaph bibliographic networks. Then, we introduce a new algorithm, called *Query_Exploration*, based on the exploration of the generated Galois Lattices to extract the multi-relational academic communities.

3. Background on FCA and RCA

A. Formal Concept Analysis (FCA): Formal Concept Analysis is a mathematical approach that derives a set of objects described by attributes into a hierarchy of concepts that is a complete lattice[4]. A 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 $o \in \mathcal{O}$ 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 $\mathcal{O}$. This leads us to the definition of a formal concept.

definition 1. **(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.*

Proposition 1 presents the partial order on formal concepts *w.r.t.* set inclusion[4].

proposition 1. *A partial order on formal concepts 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 $A_1 \subseteq A_2$.*

When partially sorted with set inclusion, formal concepts form a structure called *Galois* (*concept*) *lattice*, defined as follows.

definition 2. **(Galois (concept) lattice)** *Given a context $\mathcal{K}$, the set of formal concepts C is a complete lattice $\mathcal{L}_C = (C, \leq)$, called Galois* (*concept*) *lattice, when C is considered with set inclusion between concepts intents* (*or extents*)*[4].*

B. Relational Concept Analysis (RCA): Relational Concept Analysis is an extension of Formal Concept Analysis (FCA) to the processing of multi-relational datasets *i.e.* datasets in which individuals are described both by their own features and by their relations to other individuals[11].

definition 3. **(Relational Context Family)** *A relational context family RCF is a pair (*K, R*) where* $\mathrm{K} = \{\mathcal{K}_i\}_{i=1,\ldots,n}$ *is a set of (object-attribute) contexts $\mathcal{K}_i = (\mathcal{O}_i, \mathcal{A}_i, \mathcal{I}_i)$ and $\{r_{j,l}\}_{j,l\in\{1,\ldots,n\}}$ 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 $r_{j,l}$ ($dom(r_{j,l})$) and $\mathcal{O}_l$ is called the range of $r_{j,l}$ ($ran(r_{j,l})$)[11].*

A function *rel* is associated with a *RCF* which maps a context $\mathcal{K} = (\mathcal{O}, \mathcal{A}, \mathcal{I}) \in \mathrm{K}$ to the set of all relations $r \in \mathrm{R}$ starting at its object set $\mathcal{K}$: $rel(\mathcal{K}) = \{r \in \mathrm{R}, \text{where } \mathrm{dom}(r) = \mathcal{O}\}$. Hence, given a relation r and a quantifier f chosen within the set $\mathrm{F} = \{\forall, \exists, \forall\exists, \geq, \geq_f, \leq, \leq_f\}$. k maps an object set from $\mathrm{ran}(r)$ to an object set from $\mathrm{dom}(r)$ as $k : \mathrm{F} \times \mathrm{R} \times \cup_{i=1,\ldots,n} \mathcal{P}(\mathcal{O}_i) \rightarrow \cup_{i=1,\ldots,n} \mathcal{P}(\mathcal{O}_i)$[11]. 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 $o_i \in O_i$ and $o_j \in O_j$, 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 $\mathcal{K}_i^0 = (O_i^p, \mathcal{A}_i^p, \mathcal{I}_i^p)$ 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 $r_{ij} \subseteq O_i \times O_j$ be a relational context. The exists scaled relation $r_{ij}^{\exists}$ is defined as $r_{ij}^{\exists} \subseteq O_i \times B(O_j, A, I)$, such that for an object o_i and a concept c:$(o_i, c) \in r_{ij}^{\exists} \Leftrightarrow \exists x, x \in o_i' \cap Extent(c)$.

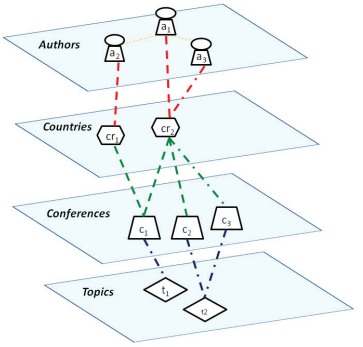

Fig. 1. An example of a multi-relational academic network.

4. A framework for Community detection based on Multi-Relational querying exploration (*CoMRing*)

Our main objective, is to discover a set of multi-relational academic communities, from multi-relational source, formally represented by the Relational Concept Analysis (RCA).
In order to find out the academic community structure across multiple network dimensions, we have to integrate the information from all dimensions. In particular, we are interested of the academic search and mining system, which extracts and integrates the academic data from the distributed Web.
Our *CoMRing* approach is based on two main steps, as represented in Fig.2, which are:

- Modeling the academic network (objects and relations) based on Relational Concept analysis(RCA) techniques.
- Navigate between Concept Lattices based on a new algorithm called $Query_{Exploration}$ to extract the multi-relational academic communities.

4.1. The Multi-relational bibliographic network 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= $\{a_1, a_2, \ldots, a_n\}$, locates in a given Country Cr= $\{cr_1, cr_2, \ldots, cr_p\}$ and works on the same Topics T= $\{t_1, t_2, \ldots, t_k\}$ and a country could hold some Conferences Conf=$\{c_1, c_2, \ldots, c_l\}$); (we use the same notation for the rest of the paper). 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, (conference, country) relation. We

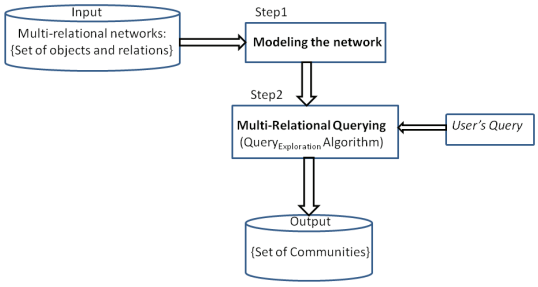

Fig. 2. The CoMRing framework.

use a relational concept family to describe the relations contexts and the objects contexts constructed from a multi-relational bibliographic network. Fig.1 depicts the data schema of the handled multi-relational bibliographic network. The relational concept family is made of 4 objects contexts : $\mathcal{K}_{Authors}$, $\mathcal{K}_{Countries}$, $\mathcal{K}_{Conferences}$,$\mathcal{K}_{Topics}$; and 4 relations contexts : $r_{Locates}$, r_{Holds}, r_{Has}, and $r_{Discusses}$. We report in Fig.3 these 4 objects contexts and in Fig.4 the 4 related relations contexts.

$K_{Authors}$

Authors	a1	a2	a3	a4	a5	a6	a7	a8	a9
O. Willum	×								
D. Wei		×							
Wenhu Wu			×						
Manuel Will				×					
Mariya Das					×				
Roberto Gallo						×			
Henrique Kawakami							×		
Ahmed Seffah								×	
Peter M. Maurer									×

K_{Topics}

Topics	dt	net	is	mm
Datamining	×			
Network		×		
Information Security			×	
Multimedia				×

$K_{Conferences}$

Conferences	c1	c2	c3	c4	c5	c6	c7
SIGIR	×						
DEXA		×					
Wireless Networks			×				
ICNP				×			
TIP					×		
CCS						×	
TIFS							×

$K_{Countries}$

Countries	cr1	cr2	cr3	cr4
China	×			
USA		×		
Canada			×	
France				×

Fig. 3. The objects contexts extracted from the multi-relational bibliographic network.

The overall process of *RCA* follows a multi-FCA method[11] which allows to build a set of lattices called Concept Lattice Family (*CLF*). The *RCA* process is an iterative one 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. According to our example, the generated concept lattice family, consists of four lattices: Authors and Topics in Fig.5, Countries and Conferences lattices in Fig.6.

4.2. Multi-Relational Querying Exploration for Community Detection

In this section, we introduce a new approach to extract a set of academic communities, which gives a general navigation schema that applies to concept lattices built with the existential scaling. We present a $Query_{Exploration}$ method, which allows to select suitable communities that respond to the submitted users' query. Indeed, in order to

$r_{Discusses}$

Discusses	Datamining	Network	Information Security	Multimedia
O. Willum	×			
D. Wei	×			
Wenhu Wu	×			
Manuel Will		×		
Mariya Das		×		
Roberto Gallo			×	
Henrique Kawakami			×	
Ahmed Seffah				×
Peter M. Maurer				×

$r_{Locates}$

Locates	China	USA	Canada	France
O. Willum	×			
D. Wei	×			
Wenhu Wu	×			
Manuel Will		×		
Mariya Das		×		
Roberto Gallo			×	
Henrique Kawakami			×	
Ahmed Seffah				×
Peter M. Maurer				×

r_{Holds}

Holds	SIGIR	DEXA	Wireless Networks	ICNP	TIP	CCS	TIFS
China	×	×					
USA			×	×			
Canada					×		
France						×	×

r_{Has}

Has	Datamining	Network	Information Security	Multimedia
SIGIR	×			
DEXA	×			
Wireless Networks		×		
ICNP		×		
TIP				×
CCS			×	
TIFS			×	

Fig. 4. The relations contexts extracted from the multi-relational bibliographic network.

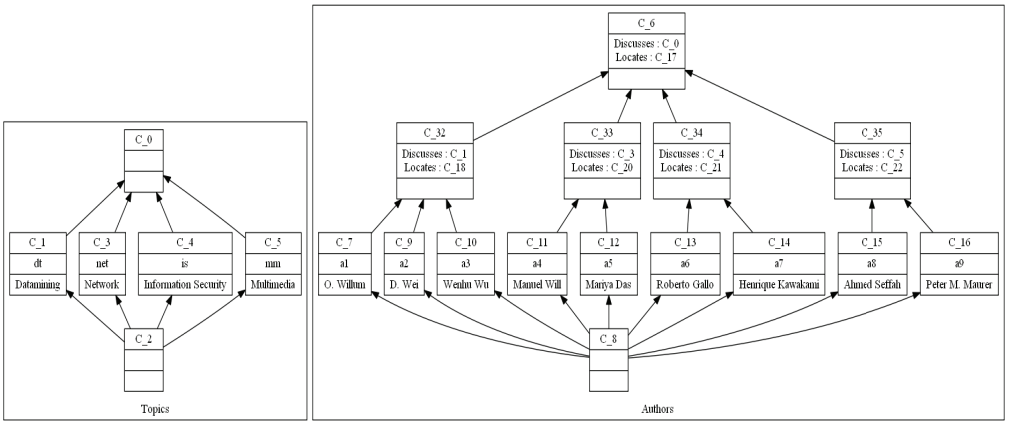

Fig. 5. Authors and Topics Lattices.

navigate on concept lattices to communities discovering, users have to submit a query that respects the *RCF* relations. Firstly, we have to transform such query to a query path *QP* that guides the navigation across a concept lattice family. In fact, a query path *QP* is the inverse order of the relational query *RQ* which is composed of several Simple Queries *SQ* defined as follow:

definition 4. **(Simple Query)** *A Simple Query SQ on a context K=(A,O,I), denoted by SQ=$\{o_q\}$, is a set of objects satisfying the query (or the answer set) with $o_q \subset O$.*

definition 5. **(Relational Query)** *A Relational Query $RQ = \{rq_0, rq_1, \ldots, rq_m\}$ 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 Simple Queries SQ.*
- *r_{st} is the relation between q'_s and q'_t. It leads one-to-one mapping between q'_s and q'_t.*

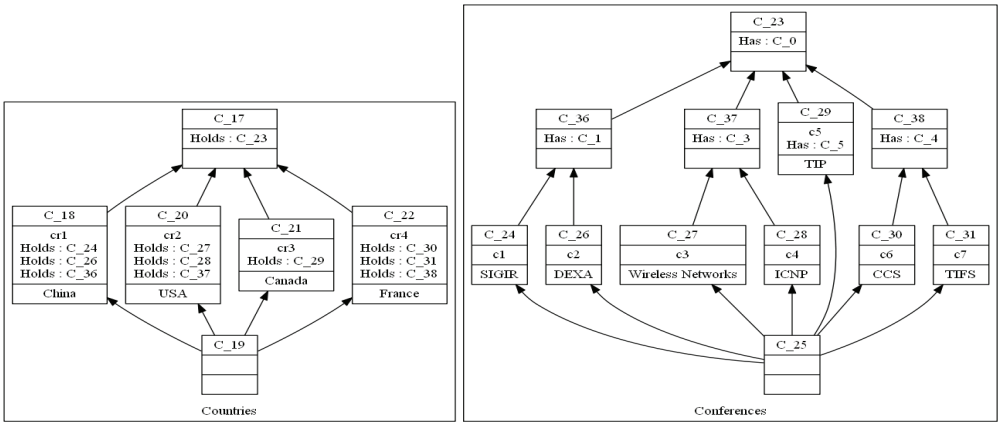

Fig. 6. Countries and Conferences Lattices.

To explore the concept lattice family, we have to construct a query path *QP*. It allows to know the path that we have to follow and specifies the source and the target lattices.

definition 6. (Query Path) *Let QP=$\{qp_0, qp_1, \ldots, qp_n\}$ 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, that we will navigate in it. 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 $q_{t0}=q'_{sm}$*

Algorithm 1: *Query − Exploration*

Input: **- The path QP=$\{qp_k\}$ with qp=$((q_s,L_s),(q_t,L_t))$**
Output: **- Answer to the query (set of communities)**
begin 1
  foreach *qp in QP* **do** 2
    if *k eq 0* **then** 3
      foreach *C in* L_s **do** 4
        $o_s := q_s$; 5
        if $o_s \subseteq Extent(C)$ **then** 6
          $q_t := q_t \bigcup C$; 7
    foreach *Element in* q_t **do** 8
      Let *C* be the $Concept(L_t)$ having $Intent(C) \supseteq$ Element; 9
      $Obj := \text{Obj} \bigcup \text{Extent}(C)$; $q_t := q_t \bigcup C$; 10
return (Obj); 11

In order to detect academic communities, we have to navigate between different lattices aiming to respond to the submitted users' query. Indeed, we propose a new method called $Query_{Exploration}$ 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=$((q_s,L_s),(q_t,L_t))$ and outputs the identified community as an answer to the user query Q.
The $Query_{Exploration}$ algorithm 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 (line3 to line7). It proceeds by identifying the concept extent of the lattice L_s then it extracts the concepts that contains an extent related to the query q_s. The result of the initial phase is a set of concepts C_i that respond to the query q_s. The second phase of $Query_{Exploration}$ (line8 to line10), 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, $Query_{Exploration}$ 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.

Example: Suppose that a user submits the following query: *Q ='I am looking for author community locates in country which holds conference that has Network topic'*. Hence, within this first step, q is transformed into a relational query as follows : rq=((q_A,'Locates',q_{Cr}); (q_{Cr},'Holds',q_{Conf}); (q_{Conf},'Has',q_T)) with q_T='Network'.

The correspond query path, QP=((q_T,L_T), (q_{Conf},L_{Conf}); (q_{Conf},L_{Conf}),(q_{Cr},L_{Cr}); (q_{Cr},L_{Cr}), (q_A,L_A)).

In this case, the $Query_{Exploration}$ method starts by handling all concepts extent C in the source lattice L_T, in order to extract the corresponding concepts which is an answer to the query q_T.

The result of the first phase is the C_3 which contains 'Network' as extension. The second step, consists on handling the corresponding lattice L_{Conf} to extract the set of concepts which contains the C_3 in the concept intent of the lattice L_{Conf}. In this case, C_37 is an answer of this query because it contains C_3 in its intent. $Query_{Exploration}$ proceeds then by following the query path order. The next path in the query path QP is((q_{Conf},L_{Conf}),(q_{Cr},L_{Cr})). Thus, it extracts the set of concepts that contains C_37 in there intents by handling the lattice L_{Cr}. The answer of this query is C_20. In the same way, the $Query_{Exploration}$ algorithm extracts the the set of concepts containing C_20 in their intent with handling the lattice L_A. The result of this query is thus the C_33. Then after, $Query_{Exploration}$ extracts the extent of the final selected concept which is C_33. In this case the extent of C_33 are 'Manual Will' and 'Mariya Das'. These two individuals constitute an academic community returned to the user.

5. Experimental evaluation

We collect data from two bibliographic databases. First we use the well known database DBLP. But in order to complete our conceptual hypergraph model, we access on AMiner database for taking keywords, institutions and research topics. In theses two sources, 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). At the end, we build a data set which contains 914 contributions and 336 authors since 2010. Finally, the navigational query algorithm is developed in JAVA and tested on a Windows 7 with Intel core i5, 2.4GHz and 8GB 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: addresses 4 entities, *i.e.,* Authors, Countries, Conferences and Topics; and 3 relations, *i.e.,* Locates, Holds and Has.

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

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

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

Community quality via ground truth: We consider two different ground truths [14]. The first ground truth $GT1$: each explicit authors' topic in the dataset is a ground truth community $GT1$. $GT1$ contains authors nodes which share the same topic. The second ground truth $GT2$: each explicit author conference is a ground truth community. $GT2$ contains authors nodes which participate in the same conference.

Effectiveness of our approach: The performance is assessed by the measures of *Recall*, *Precision* and Fβ_measure, computed over all vertices [12]. 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* (Fβ_ measure= $\beta \times \frac{Precision \times Recall}{Precision + Recall}$ where $\beta \in \{1,2\}$).

Finally, for performances comparison with the baseline according to the four queries and the two ground truths, an overall average score of the Precision, Recall, $F1\_measure$ and $F2\_measure$ is computed. The results are depicted in Fig.7. Note that aggregated bars on $Q1$ to $Q4$ correspond to the results on the 4 considered measures. Thus, according to the sketched curves in Fig.7, we can point out that *CoMRing* 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 the *CoMRing* 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,

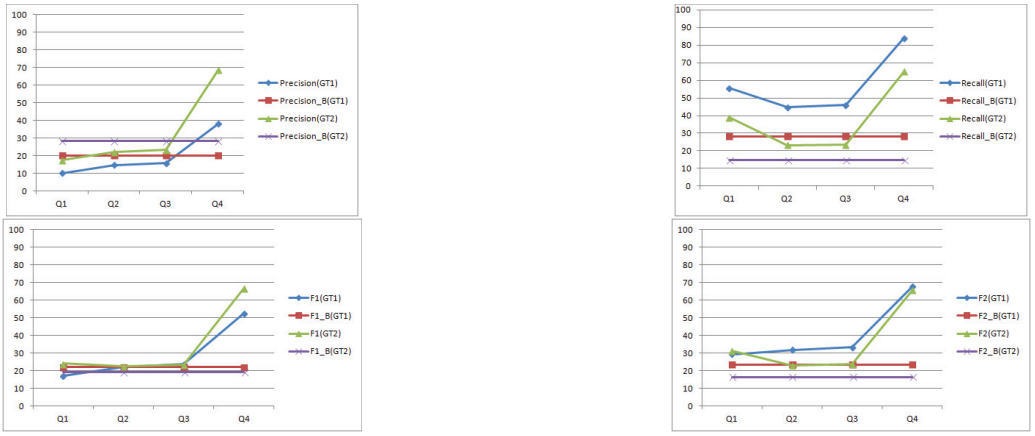

Fig. 7. Average score of the Precision, Recall, F1_measure and F2_measure of *CoMRing* approach *vs.* those of the baseline (B).

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. Carried out experiments show that the use of multi-relations allows to detect hidden relations that respond to two types of queries which are relational and simple queries. We also study the Recall score evolution over the ground truth GT1 with respect to authors number in each extracted community by considering the 4 queries. We can point out from Fig.8 (Left) that detected communities over all queries have different size, communities size ranges between 2 to 300 authors per community. Thus, our approach doesn't suffer of the resolution problem which is a big problem for a lot of approaches such that[7].

Another important observation is that the Recall score increases accompanied to the increasing of the communities size over the queries Q1,Q3 and Q4. It increases from 22.72% (community with 59 authors) to 70.16% (community with 180 authors) over the query Q1 and from 2.02% (community contains 21 authors) to 52.99% for community which contains 132 authors. The obtained results highlights that the proposed approach allows to detect significant communities with different size considering different queries and that our approach is more efficient in communities with big size. We focus now on the most complex relational query

Fig. 8. **Left.**The Recall scores against the number of items in the community with respect to the 4 queries. **Right.**The Recall scores against the number of conferences in each community extracted by the query $Q1$.

$Q1$. Experiments show that our approach rises the challenge of detecting multi-relational communities semantically rich. Indeed, $Q1$ allows to extract 5 communities, each community is labelled by only one topic, an average of 2.2 conferences and an average of 1.7 countries. Fig.8 (Right) represents the Recall evolution against the number of conferences shared in each community extracted by the query $Q1$ over the two ground truths $GT1$ and $GT2$. As we show in the Fig.8 (Right), we have 5 communities: 2 communities sharing 1 conference, 2 communities sharing 2 conferences and 1 community shares 5 conferences.

For the topic communities 'Human Computer' and 'Computer Network' which contain one conference in each community, the average Recall is between 22.72% and 31.33% among the two ground truths (GT1 and GT2). Furthermore, we can note an interesting increasing of the average Recall when the number of communities conferences increases, it reaches 55.24% over the ground truth GT1 and 62.2% over the ground truth GT2 in 'Data Mining' and 'Computer Graphics' communities which has two conferences

in each community. Indeed, we can note also that the community with the highest number of entities (5 conferences) shows a significant improvement by 47.44% and 14.92% over the ground truth GT1 and by 37.13% and 6.26% over the ground truth GT2 comparing with one and two conferences entities communities respectively.
It's clear that more the number of shared entities (conferences) is bigger more the Recall score of communities is higher. An interesting observation is that more the community is rich of shared relations and entities more the Recall increases and thus more the matching between the ground truth communities ($GT1$ and $GT2$) and the extracted communities is important. We can conclude that the increase of relations and entities improves communities structurally and semantically.
In summary, this experiments show that the *CoMRing* approach is better than the state of the art in many areas: first, it detects a set of multi-relational communities semantically rich with a high accuracy; second, it overcomes the resolution limits; third, the *CoMRing* approach detects communities from general networks without requiring a priori knowledge from users; fourth, it extracts different type of communities and responds to different type of queries(simple and relational), thus we don't have to change the model every time we change the type of query provided that it respects the hypergraph model.

6. Conclusion

In this paper, we have introduced a novel approach called *CoMRing* for discovering communities from heterogeneous multi-relational bibliographic networks based on querying exploration. Firstly, we proposed to use the RCA techniques to model the different entities and relations of the bibliographic network. We introduce a new method, called: $Query_{Exploration}$ for academic communities detection, which allows to navigate between the different lattices linked by attributes relations. Our future research will focus on further study other quantifier such as $\forall$ quantifier to address a more diversified set of queries. We also plan to evaluate and test our approach on other real-world multi-relational networks such as Genetic data collection for medical diagnosis.

Acknowledgements

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

References

1. U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner. On modularity clustering. *IEEE Trans. on Knowl. and Data Eng.*, 20:172–188, 2008.
2. M. Crampes and M. Plantié. Détection de communautés chevauchantes dans les graphes bipartis. In *MARAMI 2012 : conférence sur les modèles et l'analyse des réseaux : Approches mathématiques et informatiques*, 2012.
3. X. Dolques, F. Le Ber, and M. Huchard. Aoc-posets: a scalable alternative to concept lattices for relational concept analysis. In *CLA 2013: 10th International Conference on Concept Lattices and Their Applications*, volume 1062 of *CEUR Workshop Proceedings*, pages 129–140, Oct 2013.
4. B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer, Berlin, 1999.
5. N. Jay, F. Kohler, and A. Napoli. Analysis of social communities with iceberg and stability-based concept lattices. In R. Medina and S. Obiedkov, editors, *Formal Concept Analysis*, volume 4933, pages 258–272. Springer, 2008.
6. Y.-R. Lin, J. Sun, P. Castro, R. Konuru, H. Sundaram, and A. Kelliher. Metafac: Community discovery via relational hypergraph factorization. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 527–536, New York, NY, USA, 2009. ACM.
7. X. Liu, W. Liu, T. Murata, and K. Wakita. A framework for community detection in heterogeneous multi-relational networks. *Advances in Complex Systems*, 17(6), 2014.
8. N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Efficient mining of association rules using closed itemset lattices. *INFORMATION SYSTEMS*, 24:25–46, 1999.
9. M. Planti and M. Crampes. Survey on social community detection. In *Book Chapter, Social Media Retrieval, Computer Communications and Networks*, pages 65–85, 2013.
10. C. Roth, S. A. Obiedkov, and D. G. Kourie. On succinct representation of knowledge community taxonomies with formal concept analysis. *Int. J. Found. Comput. Sci.*, 19:383–404, 2008.
11. M. Rouane-Hacene, M. Huchard, A. Napoli, and P. Valtchev. Relational concept analysis: Mining concept lattices from multi-relational data. *Annals of Mathematics and Artificial Intelligence*, 67(1):81–108, Jan. 2013.
12. S. Song, H. Cheng, J. X. Yu, and L. Chen. Repairing vertex labels under neighborhood constraints. *PVLDB*, 7:987–998, 2014.
13. Y. Sun, C. C. Aggarwal, and J. Han. Relation strength-aware clustering of heterogeneous information networks with incomplete attributes. *PVLDB*, 5(5):394–405, 2012.
14. J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. In *ICDM*, pages 745–754. IEEE Computer Society, 2012.
15. Z. Zhang, Q. Li, D. Zeng, and H. Gao. User community discovery from multi-relational networks. *Decis. Support Syst.*, 54(2):870–879, Jan. 2013.