Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 159 (2019) 291–300

Procedia Computer Science

www.elsevier.com/locate/procedia

23rd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Community detection in multi-relational social networks based on relational concept analysis

Soumaya Guesmi[a*], Chiraz Trabelsi[a], Chiraz Latiri[a]

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

Abstract

Multi-relational community discovering in heterogeneous social networks is an important issue. Many approaches has been proposed for community discovering in heterogeneous networks. However, they have focused only on topological properties of these networks, ignoring the embedded semantic information. As the solution to this information glut limit, we propose, in this paper, a new multi-relational community discovering approach which incorporates the multiple types of objects and relationships, derived from heterogeneous networks. Firstly, we propose to construct the Concept Lattice Family *CLF* to represent the different objects and relations of the heterogeneous social networks based on the relational concept analysis techniques. Then after we introduce a new algorithm that explores such *CLF* and extract the multi-relational communities. Carried out experiments on real-datasets enhance the effectiveness of our proposal and open promising issues.

Keywords: Heterogeneous social networks, Multi-relational community mining, Relational concept analysis

1. Introduction

In the recent years, the rise of social networks has allowed human interaction with unprecedented entities. The huge growth of scale networks in social media has led to the increasing attention in social network analysis in many areas such as economics, marketing, computer sciences and human behavior. Therefore, a new task has emerged, which is the exploration of a set of members who interact more frequently with each other and who form a cohesive community. The community detection area can be applied in several disciplines including stable clusters, recommender systems, relational learning, visualization and emotional analysis. In real life, people interact with each other in various ways, for example: by phone, by social networks or face-to-face. Thus, by combining the different activities, we can obtain a heterogeneous network representing the variety of members in interaction. Despite the research atten-

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

10.1016/j.procs.2019.09.184

tion on heterogeneous network representation and efficient topological algorithm design, a much more fundamental issue concerning the exploration of the heterogeneous organization infrastructure and community detection has not been skillfully addressed. Therefore, the new research challenge consists in 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 ones. Most multi-relational community discovering approaches have specially focused on the topological properties of social networks, ignoring the semantic information; (*i.e.,* shared properties between vertices). However, in many applications, the social networks can be presented with communities that involve a set of users' attributes that are used to detect more effective communities. As a consequence, the new community detection challenge consists in combining the structural information with users' attributes; (*i.e.,* the semantic information).

To overcome this limitation, we put forward, in this paper, a new approach called $CoMRCA$ based on Relational Concept Analysis (RCA) techniques in order to find out the multi-dimensional social community semantically rich across multiple network dimensions. We firstly suggest to model the different relations and entities embedded in the structure of social networks across multiple network dimensions using the RCA techniques [8, 2], that model such dataset to a set of Galois lattices. Since, the RCA could only visualize and model the dataset into a set of lattices, we have suggested a new algorithm called $SearchCommunity$ aiming to explore such Galois lattices and extract the appropriate multi-relational social communities. The remainder of the paper is outlined as follows. In section 2, related works are reviewed. Section 3 reminds the main principles of Formal Concept analysis (FCA) and RCA. Section 4 explains the new approach of community detection in multi-relational networks $CoMRCA$. In order to evaluate the effectiveness and the efficiency of our approach, extensive experiments are conducted on real-life datasets. The descriptions of these data sets, the experimental results and their analyses are given in section 5. Finally, the conclusions are given in section 6.

2. Related works

Recently, several researches have addressed community detection in heterogeneous multi-relational networks focusing on optimizing a quality function to assess the relevance of a given partition. Sun et *al.* [11] designed ranking-based community detection of heterogeneous information networks with a star network. Comar et *al.* [1] developed a framework to cluster multiple networks based on a joint to factorise their adjacency matrices. Unfortunately, we cannot apply these two methods to general networks because they are restricted to a particular subclass of networks. Lin et *al.,* [6] proposed a method based on tensor factorization, called MetaFac (MetaGraph Factorization). The authors supposed that a community would contain different types of nodes, thus dividing nodes of different types separately; *i.e.,* nodes of numerous types would have the same number of communities. The main limit of the suggested method is that we rarely see this situation in a real-world scenario. Furthermore, in Sun et *al.,* [10], the authors construct a general heterogeneous network clustering algorithm called GenClus to integrate incomplete attribute information and network structure information. In another important work, Zhang et *al.,* [13] put forward a method based on matrix factorization which combine user-generated contents and friendship networks to discover user communities sharing common content interests densely connected. Yet, 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 the latent organization of a real system.

A lot of approaches have been interested in optimizing the composite modularity [4] which is widely used for detecting communities in a homogeneous single-relational network. The authors in [7] proposed a new approach, called Louvain-C, for detecting communities in a heterogeneous multi-relational network which followed the line of the modularity optimization method. However, this approach tended to detect communities in each subnetwork separately, and afterwards they combined the obtained partitions. As a result, this separation induced the loss of knowledge while it could be used to filter their sources, improve their relevance and accuracy and customize their results. Another drawback of this approach was the resolution limit, which might prevent it from detecting communities which are comparatively small in large-scale networks.

Moreover, communities have not been labeled and they have deeply focused on the topological properties of these networks, ignoring embedded semantic information. In fact, no research has addressed the problem of semantic community detection from a multi-relational networks.

To overcome the limitations mentioned above, we introduce in this paper a new approach for semantic community discovering from a multi-relational network structure, based on *RCA* techniques [8] called *CoMRCA*.
The main objective of our approach is to find out the multi-dimensional community structure across multiple network dimensions, so as to model the different relations and entities embedded in the structure of social networks across multiple network dimensions, we have to integrate the information from all dimensions. In particular, we firstly propose to use the *RCA* techniques to model the various relations and entities embedded in the social network. Secondly, we use a new algorithm called *SearchCommunity* in order to explore the generated Galois lattices and extract the multi-dimensional social communities.

3. Background on FCA and RCA

Formal Concept Analysis (*FCA*): the *FCA* is a mathematical approach that derives a set of objects described by attributes into a hierarchy of concepts which is a complete lattice [3]. 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 $a \in \mathcal{A}$. Here $\mathcal{O}$ is called one-valued context. It is worth linking 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.*

Definition 2. (Partial order formal concepts) *A partial order on formal concepts is defined as:* $\forall$ $c_1 = (O_1, A_1)$ *and* $c_2 = (O_2, A_2)$ *are two formal concepts,* $c_1 \leq c_2$ *if* $O_2 \subseteq O_1$*, or equivalently* $A_1 \subseteq A_2$ *[3].*

The concepts of a *FCA* lattice are arranged in a hierarchical order, often referred to as a partial order, based on a $\leq$ relation between concepts. A concept (A_1, B_1) is a sub-concept of a concept (A_2, B_2) if $A_1 \subseteq A_2$ or $B_2 \subseteq B_1$. Correspondingly, (A_2, B_2) is a super-concept of $(A_1, B_1$, hence $(A_1, B_1) \leq (A_2, B_2)$.
Relational Concept Analysis (*RCA*): the *RCA* is an extension of the *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 [8].

Definition 3. (Relational Context Family) *A Relational Context Family RCF is a pair* (K, R) *where* K = $\mathcal{K}_i$*, with i={1,...,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(*$r_{j,l}$*)).*

A function *rel* is associated with an *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 $\mathcal{K}$: $rel(\mathcal{K}) = \{r \in \mathrm{R}$, where $\mathrm{dom}(r) = \mathcal{O}\}$. Hence, given a relation r and a quantifier f chosen within the set $\mathrm{F} = \{\forall, \exists, \forall\exists, \geq, \geq_f, \leq, \leq_f\}$, then k maps an object set from $\mathrm{ran}(r)$ to an object set from $\mathrm{dom}(r)$ as $k : \mathrm{F} \times \mathrm{R} \times \cup_{i=1\ldots n} \mathcal{P}(\mathcal{O}_i) \rightarrow \cup_{i=1\ldots n} \mathcal{P}(\mathcal{O}_i)$ [8]. 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.

Definition 4. (The existential scaling operator) *Given* $K = (O, A, I)$ *and* $r \in rel(K)$*, let* i_r *be such that* $ran(r) = O_i^r$ *where* $K_i^r = (O_i^r, A_i^r, I_i^r)$*. Let also* L_i^r *be the lattice of* K_i^r*. The existential scaling operator* $S_{(r,\exists)}, L_i^r$ *maps K into the derived context* $K^+ = (O^+, A^+, I^+)$ *where:*

- $O^+ = O$
- $A^+ = \{\exists r : c | c \in L_i^r\}$*, where each* $\exists r : c$ *is a relational attribute*
- $I^+ = \{(o, \exists r : c) | o \in O, c \in L_i^r, r(o) \bigcap Ext(c) \neq \emptyset\}$

4. Multi-relational community mining in heterogeneous social networks

Our main target, is to discover a set of multi-relational communities, from multi-relational sources, formally represented by the *RCA*. In order to find out the community structure across multiple network dimensions, we have to

integrate the information from all dimensions. In particular, we are interested in the social network search and mining system, which extracts and integrates the data from the distributed web.
Our approach is based on two main steps, which are: modeling the social network (objects and relations) based on *RCA* techniques and extracting the set of multi-relational communities based on the *SearchCommunity* algorithm, which enables navigating between the *CLF* and extracting the appropriate communities.

4.1. Social network modeling

Three concepts are involved in our model: the object context, the relation context, and the *CLF*. Suppose that we have a set of Users U= $\{u_1, u_2, \ldots, u_n\}$, who live in a given Country C= $\{c_1, c_2, \ldots, c_p\}$, watch the same set of Movies M= $\{m_1, m_2, \ldots, m_k\}$ and have friendship relations between each other. A movie can have a Genre G=$\{g_1, g_2, \ldots, g_l\}$) and be projected in the same Country (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.,* a user context is a set of users and define a relation context as the interactions among objects contexts, *e.g.,* (User, Movie) relation and (Movie, Genre) relation.
We use a the *RCF* to describe the relation contexts and the object contexts constructed from a multi-relational network. The *RCF* is made of four object contexts : $\mathcal{K}_{User}$, $\mathcal{K}_{Country}$, $\mathcal{K}_{Genre}$,$\mathcal{K}_{Movie}$, and three relation contexts: r_{Lives_In}, r_{Has_Genre} and r_{Watch}. We report in Fig. 1 these two object contexts and in Fig. 2 three related relation contexts. The overall process of the *RCA* follows a multi-FCA method [8] which allows building a set of lattices called *CLF*. The construction of the set of lattices associated to the *RCF* is an iterative process that alternates the pure lattice construction and expansion of the contexts through relational scaling. First, the process constructs concept lattices using the object contexts k_i only starting with the lattice L_i built with the original attribute set A_i. The resulting lattices provide the basis for relational scaling which is universally applied on every relation of the *RCF* at the next iteration. Then, in the following steps, whenever the *RCF* is expanded, their corresponding lattices extend as well. In this step the process concatenates object contexts with relation contexts based on the relational scaling operator that produces scaled relations. Hence, the relational scaling translates the links between objects into conventional *FCA* attributes and extracts a collection of lattices whose concepts are linked by relations; and this goes on until the lattices stop evolving between iterations (two consecutive steps produce lattices that are isomorphic). This means that a fixed point for the *RCF* scaling operator has been met and the computation ends up.
According to our example, the generated *CLF* consists of four lattices, we present only two lattices which are: User lattice in Fig. 1 and Movie lattice Fig. 2.

Since, the *RCA* could only visualize and model the dataset into a set of lattices, we have suggested a new algorithm called *SearchCommunity* aiming to explore such Galois lattices and extract the appropriate multi-relational social communities.

Table 1. Object contexts extracted from multi-relational network.

$\mathcal{K}_{User}$	u1	u2	u3	u4	u5	u6	u7	u8	u9
Jack	×								
Alexender		×							
Jakline			×						
John				×					
Mariya					×				
Robert						×			
Henrique							×		
Sofia								×	
Peter M. Maurer									×

$\mathcal{K}_{Movie}$	m1	m2	m3	m4	m5	m6	m7
Movie1	×						
Movie2		×					
Movie3			×				
Movie4				×			
Movie5					×		
Movie6						×	
Movie7							×

4.2. Multi-relational community exploring algorithm

In this section, we introduce a new algorithm called *SearchCommunity*, which leads to extract a set of multi-relational communities and their shared relational properties (labels). To achieve this goal, our algorithm follows two main steps. The first step consists on extracting a set of communities from the Main Lattice (*ML*); (*i.e.,* the Users' lattice which contains the community members (line4 to 11 of the *SearchCommunity* algorithm)). In the second step, we navigate between lattices in order to extract

Table 2. Relation contexts extracted from multi-relational network.

r_{Lives_In}	Chine	America	France
Jack	×		
Alexender	×		
Jakline	×		
John		×	
Mariya		×	
Robert		×	
Henrique			×
Sofia			×
Peter M. Maurer			×

r_{Watch}	Movie1	Movie2	Movie3	Movie4	Movie5	Movie6	Movie7
Jack	×	×					
Alexender	×	×					
Jakline	×	×					
John			×	×			
Mariya			×	×			
Robert			×	×			
Henrique					×	×	×
Sofia					×	×	×
Peter M. Maurer					×	×	×

$r_{Has-Genre}$	Action	Romantic	Science_Fiction
Movie1	×		
Movie2	×		
Movie3		×	
Movie4		×	
Movie5			×
Movie6			×
Movie7			×

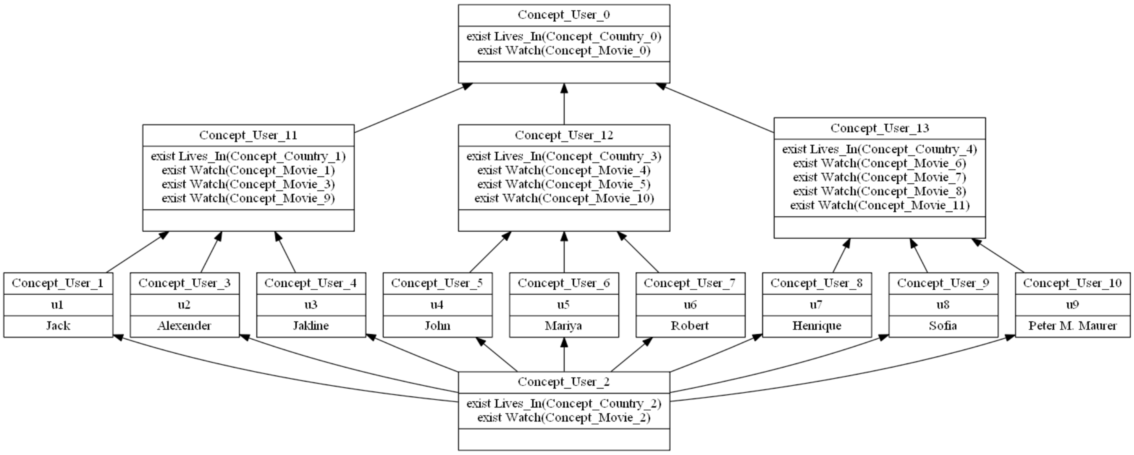
Concept_User_0
exist Lives_In(Concept_Country_0)
exist Watch(Concept_Movie_0)
Concept_User_11
exist Lives_In(Concept_Country_1)
exist Watch(Concept_Movie_1)
exist Watch(Concept_Movie_3)
exist Watch(Concept_Movie_9)
Concept_User_12
exist Lives_In(Concept_Country_3)
exist Watch(Concept_Movie_4)
exist Watch(Concept_Movie_5)
exist Watch(Concept_Movie_10)
Concept_User_13
exist Lives_In(Concept_Country_4)
exist Watch(Concept_Movie_6)
exist Watch(Concept_Movie_7)
exist Watch(Concept_Movie_8)
exist Watch(Concept_Movie_11)
Concept_User_1
u1
Jack
Concept_User_3
u2
Alexender
Concept_User_4
u3
Jakline
Concept_User_5
u4
John
Concept_User_6
u5
Mariya
Concept_User_7
u6
Robert
Concept_User_8
u7
Henrique
Concept_User_9
u8
Sofia
Concept_User_10
u9
Peter M. Maurer
Concept_User_2
exist Lives_In(Concept_Country_2)
exist Watch(Concept_Movie_2)

Fig. 1. User Lattice.

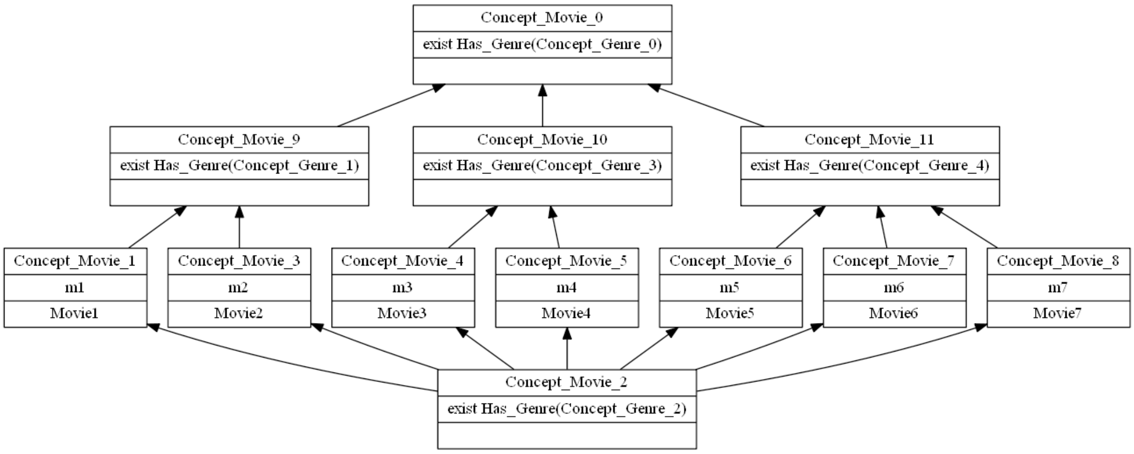

Fig. 2. Movie Lattice.

the set of communities' labels based on the $FindLabel$ function. The pseudo code of the $SearchCommunity$ algorithm is sketched by algorithm 1. It takes as an input a CLF as well as a set of relations R and it outputs a set of communities $Comm$ and a set of corresponding shared Labels Lab. Firstly the algorithm selects all concepts Cpt and their intents ($intent$) of the first level of the main lattice. Then it extracts the concepts' extents of the lowest child of the selected concepts Cpt. The selected extents transform the set of community members $Comm$. The second step consists on surfing between different lattices aiming to extract the communities' shared properties (Labels) based on the function $FindLabel$.

```
Algorithm 1: $SearchCommunity$
Input: -$CLF$, the set of relations $R$
Output: -Set of communities $Comm$ and set of their shared properties (labels) $Lab$
begin                                                              1
    root ← getRootLattice(ML) ;                                    2
    NS ← getChild(root);                                           3
    foreach $Cpt = NS$ do                                          4
        if $getExtent(Cpt)! = \emptyset$ then                      5
            Res.Comm = Res.Comm $\bigcup$ getExtent(Cpt);          6
        $IntCpt_i$← getIntent(Cpt);                                7
        foreach $eltInt_i^j$ in $IntCpt_i$ do                      8
            if $eltInt_i^j \in R$ then                             9
                SLC← SeekLowestChild(Cpt) ;                        10
                Res.Comm= Res.Comm ∪ Extent(SLC) ;                 11
                FindLabel($eltInt_i^j$, $L_i$) ;                   12
        Res = Comm ∪ Lab                                           13
```

```
Algorithm 2: $FindLabel(eltInt, L)$
begin                                                              1
    if $eltInt\ ! = NULL$ then                                     2
        NCp ← SeekConceptNode(L,eltInt) ;                          3
        if $getIntent(NCP) \in R$ then                             4
            LowLab ← SeekLowestChild(NCp) ;                        5
            Lab = Lab ∪ Extent(LowLab) ;                           6
            FindLabel(getIntent(NCp),$T_i$) ;                      7
        else                                                       8
            Res.Lab = Res.Lab ∪ getExtent(NCp);                    9
```

For the first step, the $SearchCommunity$ method starts by handling all concepts Cpt of the first level of the ML and extracting their corresponding intents $IntCpt$. After that, it selects the lowest child (SLC) extents ($Extent(SLC)$), of the extracted concepts Cpt which constitute the set of discovered communities $Comm$ (lines 4 to 7).

The second step, is an iterative process (sketched in the Function $FindLabel$). It generates at each step a set of ConceptNodes (NCP) containing the set of elements ($eltInt$) extracted in the previous step. If the intent (NCp) is a relation (line 4 to 7 of Algorithm2), we extract the LowestChild extent ($LowLab$) which constitutes the set of corresponding community labels Lab. Subsequently, it navigates into the corresponding lattice L_i in order to extract the other labels. If the Intent (NCp) is not a relation, we extract the NCp extent which constructs the set of community labels (lines 8-9). We repeat this step until no more relation is found.

Example: Suppose that the input of our $SearchCommunity$ algorithm is as follows: let User, Movie, Country and Genre lattices constitute the CLF (we have sketched only User and Movie lattices in the Fig. 1 and Fig. 2) where R={Lives_In, Watch and Has_Genre} is the set of relations. In this example, we have the User lattice is the ML. Indeed, we extract the set of concepts (Cpt) of the level 1, which are Concept_User_11, Concept_User_12 and Concept_User_13. The first step consists in extracting the extent of the lowest child SLC of the Cpt concepts that constitute the set of community members $Comm$. As a result of this first step, we have: Comm1: {Jack, Alexander and Jackline}; Comm2: {John, Mariya and Robert}; Comm3: {Henrique, Sofia and Peter.M.Maurer}.

The second step consists in extracting the community labels (Lab) by navigating between the different concept lattices. For example, we have the intent of the 'Concept_User_1' which contains four relation attributes which are: 'exist Lives_In(Concept_Country_1)', 'exist Watch(Concept_Movie_1)', 'exist Watch(Concept_Movie_3)' and 'exist Watch(Concept_Movie_9)'. We take the two relations 'exist Lives_In(Concept_Country_1)' and 'exist Watch(Concept_Movie_9)'. For the first relation 'exist Lives_In(Concept_Country_1)', we have to surf on the country lattice and extract the 'Concept_Country_1' extent (line8-9 of the $FindLabel$ function), which is China. For the relation 'exist Watch_Concept_1', we have to navigate on the Movie lattice and extract the 'Concept_Movie_9' extent which is Movie1 and Movie2, then we repeat the same process for the 'exist Has_Genre(Concept_Genre_1)'.

The final result of this example is as follows: $Comm1$: ({Jack, Alexander, Jackline}; {Movie1,Movie2}; {Chine}; {Action}); $Comm2$: ({John, Mariya, Robert}; {Movie3, Movie4); {Americ}; {Romantic}}; $Comm3$: {{Henrique, Sofia, Peter.M.Maurer}; {Movie5, Movie6, Movie7}; {France} {Science_Fiction}).

5. Experimental evaluation

5.1. Datasets Analysis

We use three data sets: Bibliographic database, Epinion and MovieLens , which are described as follows:

1. Bibliographic database: we collect data from two bibliographic databases. We use the well known database DBLP [1]. In order to complete our conceptual hypergraph model, we access to the AMiner [2] database to take keywords, institutions and research topics. In these two sources, we keep only five 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 that has contained 914 contributions and 336 authors since 2010.
2. Epinion [3]: it is a website where people can review various types of products (software, music, television shows, hardware, office appliances, ...). Products have names and belong to one unique category. In a given category, items may show a common description structure. 720 users are extracted, located in 87 Locations, who review 410 Products belonging to 58 Categories.
3. MovieLens [4]: is a movie recommender system project. Two MovieLens corpora are utilized, first, we use the $MovieLens1$ database to extract the movie information ('User', 'Movie', 'Genre', 'Director', 'Country', 'Tag'), then to complete our hypergraphic model, we make the matching with the $MovieLens2$ database to extract the user profile ('User', 'Movie', 'Age', 'Sex', 'Occupation', 'Zip-code'). We do the evaluation on a sample of the corpus that contains: 500 users, 7 intervals of users' ages, 21 occupations, 1070 published movies from 1992 to 1995, 27 tags, 14 country and 18 genres.

After the collection process, we have proceeded to the data preprocessing step aiming to eliminate the datasets' noise.

Finally, the multi-relational community detection approach is developed in JAVA and tested on a windows 7 with intel core i5, 2.4GHz and 8GB of Ram.

Our $CoMRCA$ approach tends to extract a set of labeled multi-relational communities that are semantically rich. For this reason, we present in Fig. 3 the number of labels (properties) in each detected community from all datasets. $CoMRCA$ approach permits extracting 6 communities from the Bibliographic dataset. The size of the communities is between 9 and 181 authors. Each community is labelled by an average of 1.33 topics, 4.33 conferences, 3.33 countries and 150.3 contributions. For Epinion dataset, our approach extracts 12 communities which contains between 17 and 220 users, each community is labelled by an average of 3.41 locations, 2.08 reviews, 33.25 products and 5.41 categories. The $CoMRCA$ approach detect 31 communities, from the MovieLens dataset, which vary between 5 and 180 users per community. Each community is labelled by 20.61 movies, 2.06 countries, 2.16 occupations, 5.03 tags, 2.09 ages and 1.93 genres. Thus, according to those evaluations, we can say that our approach raises the challenge of detecting semantically rich multi-relational communities.

[1] http://dblp.uni-trier.de/

[2] https://aminer.org/

[3] http://www.epinions.com/

[4] https://movielens.org/

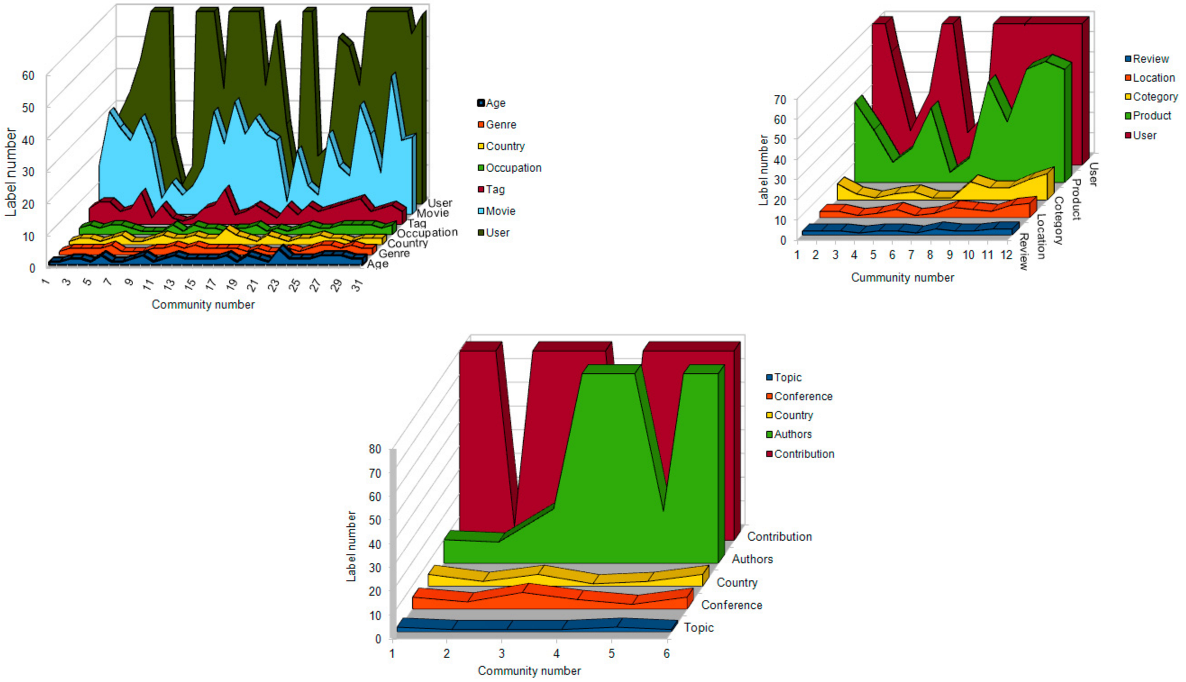

Fig. 3. Number of properties in each community over MovieLens (left), Epinion (right) and bibliographic (down) datasets

5.2. *Effectiveness of CoMRCA approach*

We use the set of ground truths [12] which are defined as follows: for bibliographic database, we consider two different ground truths. For the first ground truth GT_1^A, each explicit authors' topic in the dataset is a ground truth community GT_1^A. GT_1^A contains authors nodes which share the same topic. The second ground truth GT_2^A, each explicit author conference is a ground truth community. GT_2^A contains authors' nodes which participate in the same conference. For Epinion dataset, we consider the set of users who trust each other and belong to the same ground truth community GT^E. Consequently, in the Epinion dataset, when a user consistently gives you good advice, you are likely to trust that person's recommendations in the future and you can add this person to your "trusts" list. For MovieLens, the explicit tag of each user is a community belonging to the ground truth GT^M; *i.e.,* GT^M contains all users that tag the same tag.

In order to enhance the effectiveness of our approach, for the Bibliographic and MovieLens datasets, we calculate the similarity between the users' properties using the Jackard_measure and we select the set of users which has a Jackard_measure≥ 1. For the Epinion dataset, we have computed the similarity between all users couples using the Pearson coefficient correlation.

The performance is assessed by the measures of *Recall*, *Precision* and Fβ_measure, computed over all vertices [9]. These measures attempt to estimate whether the prediction of these vertices in the same community is 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: $Precision = \frac{|C \cap S|}{|C|}$; *Recall* indicates how many vertices are predicted to be in the same community in a retrieved community: $Recall = \frac{|C \cap S|}{|S|}$; F$\beta$_measure is the harmonic mean of *Precision* and *Recall*: $F\beta\_measure = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}, where \beta \in \{1, 2\}$.

Finally, for the performances comparison with the baseline and the Louvain-C approach according to the ground truths, an overall average score of the *Precision*, *Recall*, $F1\_measure$ and $F2\_measure$ is computed over the MovieLens, Epinion and Bibliographic datasets. The results are depicted in Fig. 4. Thus, according to the sketched histograms in Fig. 4, we can point out that the *CoMRCA* approach outperforms the baseline and the Louvain-C over the three datasets. In fact, as expected, for the Bibliographic dataset the *Recall* values of the baseline and Louvain-C are much lower than those achieved by our approach among the two ground truths (GT_1^A and GT_2^A). As it is shown, the average *Recall* of the *CoMRCA* approach achieves 66.85% and 67.05% compared with the baseline which has 28.31% and 14.58% vs. an excess of about 38.54% and 52.47% compared with the baseline among the two ground truths, respectively. The same is for the MovieLens and Epinion datasets, where the *Recall* outperforms the baseline and the Louvain-C. It is observed that average *Recall* of our approach achieves 77.03% and 80.36% vs. an excess of about 46.28% and 45.15% compared with the baseline among the MovieLens and Epinion datasets, respectively. Indeed, in terms of $F1\_measure$

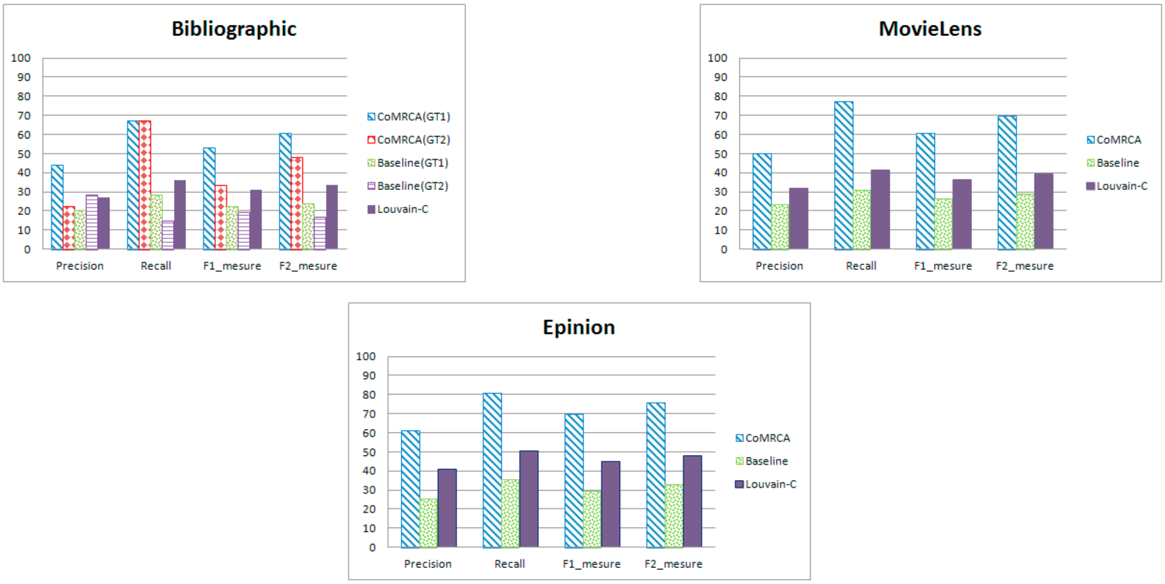

Fig. 4. Average score of the *Precision*, *Recall*, *F*1*_measure* and *F*2*_measure* of *CoMRCA* approach *vs.* those of the Baseline(B) and the *Louvain−C* algorithm.

and *F*2*_measure*, the *CoMRCA* approach outperforms considerably the baseline among all datasets. In this case, we can say that the baseline has only a small number of communities detected fairly well and not many detected communities that reflect to the ground truth communities. However, the percentage of *Precision* for the baseline outperforms slightly our approach according to the Bibliographic dataset among the GT_2^A. Whereas, for the MovieLens and Epinion datasets, our approach has an average of 50.06% and 60.89%, indicating a drop of 23% and 25.3% vs. an excess of about 27.09% and 35.59% against the baseline over the MovieLens and Epinion dataset, respectively. Hence, considering the three datasets, our approach outperforms the Baseline and the Louvain-C method in terms of *Precision*, *Recall*, *F*1*_measure* and *F*2*_measure* often by a large margin in the *Recall* score. The performed experiments demonstrate that the use of multi-relations enables detecting multi-relational communities from three different datasets. In fact, we can explain the low accuracy of the Louvain-C algorithm compared with our approach by the fact that it tends to detect communities in each subnetwork separately. Then they combine the obtained partitions. As a result, this separation induces a loss of knowledge, while it could be used to filter their sources, improve their relevance and accuracy and customize their results. The second reason is that it tries essentially to maximize the ratio of intra-community links to inter-community links without taking into account the semantic relation between users. We can conclude that the relational community improves the community structure and leads to extract relevant communities. An interesting observation is that the richer the dataset of shared relations and entities is, the more the *Recall* rises, hence the higher matching between the ground truth communities and the more important extracted communities. We can conclude that the increase in relations and entities improves communities structurally and semantically. As it can be noted from the three histograms, the obtained results highlight that the proposed approach is more efficient in the bigger dataset in term of *Recall* measure. In fact, as it is mentioned, the *Recall* goes up from 66.85% and 67.05% for the Bibliographic dataset (336 authors) over the GT_1^A and GT_2^A, respectively, to 77.03% for the MovieLens dataset and from 77.03% for the MovieLens dataset (500 users) to 80.36% from the Epinion dataset (720 users). We could explain this by the fact that the increasing of users number leads to extract a set of communities more cohesion which increase the similarity between the ground truth and the extracted communities.

5.3. *Efficiency of CoMRCA approach*

In order to test the efficiency of the *CoMRCA* approach, we carry out experiments that run our community discovery algorithm on three datasets (Bibliographic, MovieLens and Epinion) and compare the community structure obtained to the structure of communities discovered by the Louvain-C algorithm. We have computed the composite modularity [7] to evaluate the partition of the related node sets defined as follows: $Q = \sum_{y=1}^{s} \frac{m^{[y]}}{m} Q^{[y]}$, where $G^{[y]}$ denotes the subnetwork which consists of the set of hyperedges $E^{[y]}$ and the incident nodes, $G = N^{[1]} \bigcup N^{[2]} \bigcup ... \bigcup N^{[s]}$ and $N = (V^{[1]} \bigcup V^{[2]} \bigcup ... \bigcup V^{[r]}, E^{[1]} \bigcup E^{[2]} \bigcup ... \bigcup E^{[s]})$. Here $m^{[y]} = |E^{[y]}|$ is the number of edges in $G^{[y]}$, m is the total number of edges and $Q^{[y]}$ is the modularity in $G^{[y]}$.
As it is illustrated in Fig. 5 our approach outperforms slightly the Louvain-C approach. The modularity of the *CoMRCA* approach

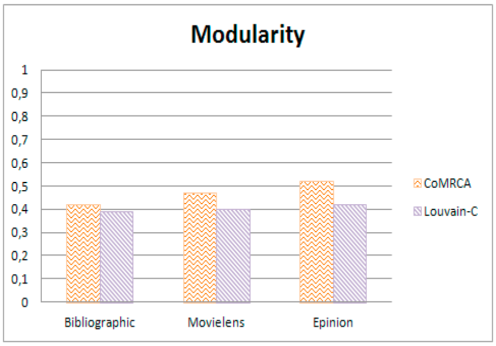

Fig. 5. Modularity Comparison for *CoMRCA* approach vs. Louvain-C on each dataset.

has 0.42, 0.47, and 0.52 compared with Louvain-C which has 0.39, 0.4, and 0.42 over the Bibliographic, MovieLens and Epinion datasets, respectively. We conclude that the *CoMRCA* approach gives a set of communities with a good partition for all datasets.

6. Conclusion

We have suggested a new approach called *CoMRCA* for detecting multi-relational communities from heterogeneous social networks. Our approach is based on two main steps: first, we have proposed to model our social network to a *CLF* based on *RCA* techniques. Second, we have put forward a new algorithm called *SearchCommunity* in order to navigate in such a *CLF* and detect a set of semantically rich multi-relational communities. The theoretical analysis and the empirical evaluation demonstrate a better performance over the modularity compared with the Louvain-C approach. The results show that the *CoMRCA* approach detects a set of semantically rich multi-relational communities with a high accuracy from general networks without requiring a priori knowledge from users. Our future research is to evaluate and test our approach on other real-world multi-relational networks such as the genetic data collection for medical diagnosis and include new selection measures, for the noisy datasets, such as the coverage, the stability or the well known support score that minimise the noise and select relevant concepts from Galois lattices [5]. We also plan to propose a new system recommendation based on our *CoMRCA* approach.

References

[1] Comar, P.M., Tan, P., Jain, A.K., 2012. A framework for joint community detection across multiple related networks. Neurocomputing .

[2] Dolques, X., Le Ber, F., Huchard, M., 2013. Aoc-posets: a scalable alternative to concept lattices for relational concept analysis, in: CLA 2013: 10th International Conference on Concept Lattices and Their Applications, pp. 129–140.

[3] Ganter, B., Wille, R., 1999. Formal Concept Analysis: Mathematical Foundations. Springer, Berlin.

[4] Girvan, M., Newman, M.E.J., 2002. Community structure in social and biological networks. Proceedings of the National Academy of Sciences .

[5] Klimushkin, M., Obiedkov, S.A., Roth, C., 2010. Approaches to the selection of relevant concepts in the case of noisy data, in: Formal Concept Analysis, 8th International Conference, ICFCA 2010, Agadir, Morocco, March 15-18, 2010. Proceedings, pp. 255–266.

[6] Lin, Y.R., Sun, J., Castro, P., Konuru, R., Sundaram, H., Kelliher, A., 2009. Metafac: Community discovery via relational hypergraph factorization, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA.

[7] Liu, X., Liu, W., Murata, T., Wakita, K., 2014. A framework for community detection in heterogeneous multi-relational networks. Advances in Complex Systems .

[8] Rouane-Hacene, M., Huchard, M., Napoli, A., Valtchev, P., 2013. Relational concept analysis: Mining concept lattices from multi-relational data. Annals of Mathematics and Artificial Intelligence .

[9] Song, S., Cheng, H., Yu, J.X., Chen, L., 2014. Repairing vertex labels under neighborhood constraints. PVLDB .

[10] Sun, Y., Aggarwal, C.C., Han, J., 2012. Relation strength-aware clustering of heterogeneous information networks with incomplete attributes. PVLDB .

[11] Sun, Y., Yu, Y., Han, J., 2009. Ranking-based clustering of heterogeneous information networks with star network schema, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, France, June 28 - July.

[12] Yang, J., Leskovec, J., 2012. Defining and evaluating network communities based on ground-truth, in: ICDM, IEEE Computer Society. pp. 745–754.

[13] Zhang, Z., Li, Q., Zeng, D., Gao, H., 2013. User community discovery from multi-relational networks. Decis. Support Syst. 54.