Team Formation with Relationship Strength Based on Meta Path in Heterogeneous Network

Xinjun Yang
School 0/ Computer Science
Beijing University ofPosts and Telecommunications Beijing, China
1127078848@qq.com

Pengfei Wang, Wei Fang
School ofComputer Science
Beijing University ofPosts and Telecommunications Beijing, China
{wangpengfei&Fang, wei}@buptedu.cn

Abstract-the goal of the team formation is to discover a collaborative team which contains all the requested skills and minimizes the communication cost between members. In the previous work, the team formation problem is solved in a homogeneous social network, which would waste a lot of computation time. In this paper, we solve the team formation problem in a heterogeneous social network, it will reduce the computation time .In the team formation problem, the communication cost depends on the relationship strength between the experts. Therefore, the estimation of the relationship strength is pretty important. In the previous work, the method of calculating the relationship strength is relatively simple, and cannot well represent the strength of the relationship between experts. According to the heterogeneous network, in this paper, we present an effective method for estimate the strength of relationships between experts based on meta paths. We call it RSMP. Experiments on the DBLP dataset show the effectiveness of the proposed methods.

Keywords; team/ormation, Relationship Strength, meta path, heterogeneous network,

I. INTRODUCTION *(HEADING 1)*

In recent years, team formation problems in social network have gradually become popular research directions at home and abroad. To complete a task with requested skills, it needs to form a team with all the skills. It is important However, what is more important for a task is that members of the team can communicate and cooperate effectivelyj1], Therefore, under a given task, it needs to find a team that can meet the requested skills for the task and with minimal communication cost.

Lappas et al. [2] first used social networking to solve team formation problems, and proposed to consider the cost of communication. They proposed two communication cost method, one is diameter communication cost other is minimum spanning tree cost, to assess team communication efficiency. In most existing team formation jobs, team formation is resolved in a homogeneous network. However, the ownership relationship between individuals and skills cannot be represented in a homogeneous social network and cannot be calculated with social networks. Therefore, in the previous work, the first step of these algorithms was to find candidates in the social network according to the skills which are required by the task Then, Find the team with the least cost to communicate based on the candidate. In this case, a lot of calculation time will be added. In addition, the existing quantification of the relationship strength between individuals in a homogenous network is insufficient. It is only the number of items that have been cooperated between individuals as the sole criterion for calculating the relationship strength[14J. This approach is clearly one-sided, and the relationship strength between individuals should also be influenced by other factors between them.

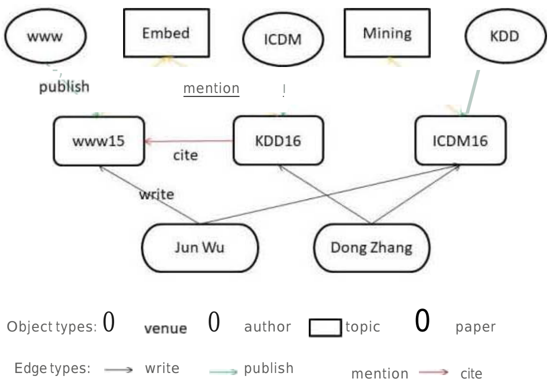
www
Embed
ICDM
Mining
KDD
publish
mention
!
www15
KDD16
ICDM16
cite
write
Jun Wu
Dong Zhang
Object types: 0 venue 0 author topic 0 paper
Edge types: write publish mention cite

Figure 1: An example of heterogeneous networks

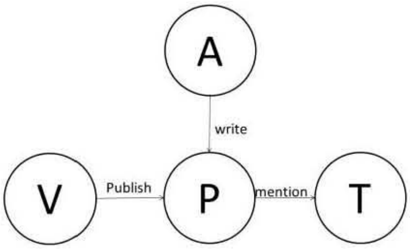

Figure 2: Schemas of a heterogeneous network

In this paper, we solve team formation problems in the heterogeneous social network. Heterogeneous social network, such as DBLP [4], YAGO [5], DBpedi [6] and Freebase [7],

978-1-5386-6565-7118/$31.00 ©2018 IEEE

are networks with more than one type of node and edge. These Heterogeneous networks contain a large number of interrelated facts that can discover more knowledge. Figure 1 illustrates a heterogeneous social network; describe the relationship between different nodes types (e.g., topic, venue, paper and author). For example, Dong Zhang has written an ICDM16 paper, which the topic it mentions is machine "Ming". At the same time, for heterogeneous network, we propose a way of calculating the relationship strength based on the meta path[8]. The meta path is a series of nodes and the relationship between nodes. For instance, $A \rightarrow P \rightarrow V \rightarrow P \rightarrow A$ is a meta path, which presents that two authors (A) are related by publishing papers (P) at the same venue (Y). For estimating the relationship strength based on meta paths, we use the Path Constrained Random Walk (pCRW) which computes the relationship strength by a random walk which start from one object would arrive the other by a meta path instance. The contributions of our paper are summarized as follows:

1、In this paper, the team formation problem is formulated in a heterogeneous social network. It can find more relationship between in individuals and reduce the computation time.

2、we use a new way to compute the relationship strength between individuals. This method can get more reasonable relationship strength between nodes.

The rest of the paper is organized as follows. Section 2 discusses related work. In Section 3, we formally define the problem. Section 4, we presents the algorithms for relationship strength and team formation project. In Section 5, we report our experimental results. Section 6 concludes the paper.

II. RELATED WORK

There are already a lot of researches based on operations research (OR) for team formation problems. In these studies, team formation problems are reduced to integer linear program problem (ILP) [18], focusing on finding the match between team members and skills. The problem is often solved using techniques such as, branch-and-cut [10], genetic algorithms [11] or simulated annealing [9]. However, these studies only focus on finding the matching of team members and skills, and neglecting the interrelationship between team members in the team is the key factor determining team efficiency.

Lappas et al. [2] first used the social network to solve team formation problem, which is a milestone. They proposed that the communication cost among team members should be considered in the process of team building. Evaluating the efficiency of a team's work through the communication cost, and find a skill that meets the task and the cost of communication should be minimal. They define each node on the social network to represent an expert with skills, and the weight of the edges between the expert nodes represents the cost of communication between them. The cost of communication indicates the efficiency of cooperation between experts, and the lower the cost of communication, the lower the cost of cooperation between the two experts. The communication cost between social network nodes is determined by the strength of the relationship between nodes .The stronger the relationship strength, the smaller the communication cost between nodes. They recommend using the minimum spanning tree and diameter to measure the team's communication cost. They also proved that the team formation problem based on the two communication cost metrics is NP-Hard problem. Kargar and An [12] proposed another communication cost metric to compute communication cost between team members which is sum of distance.

The algorithm for solving team-generated problems has been continuously improved, but the strength of the relationship between social network nodes has not received enough attention. In the beginning, they proposed the strength of the relationship between the experts by the papers they co-published accounted for the proportion of all papers they published[14]. Rongjing Xiang [3] proposed that the strength of the relationship should also be determined by the similarity between individuals and the relationship between individuals. However, these methods do not make good use of the interrelationship between individuals, so they cannot get good relationship strength between individuals.

III. PROBLEM DEFINITION

In this part, we introduce some concepts and define team formation problem in heterogeneous networks.

DEFINITION 1. *(Heterogeneous Social Network)A heterogeneous social network is defined as a directed graph* G = (V, E) *with an object type mapping function* $\phi: V \rightarrow L$ *and a link type mapping function* $\psi: E \rightarrow R$*, where each object* v E V *belongs to an object type.* $\Phi(v)$ E L*and each link* e E E *belongs to a link type* $\psi(e)$ E R

Figure 1 shows a small bibliographic heterogeneous network. We can find that it contains four types of objects and four types of links are used to connect objects.

DEFINITION 2. *Heterogeneous Social Network Schema. Given a* G $=(V,E)$ *with mappings* $\phi : V \rightarrow L$ *and* $V : E \rightarrow R$*, the schema* T G *of* G *is a directed graph defined over object types* L *and link types* R*, i.e.* G = (L, R).

The heterogeneous network schema represents the type of all edges between object types. Figure 2 shows the schema of the heterogeneous network in Figure 1, where nodes V, T, P and A correspond to venue, topic, paper and author, respectively. It also contains different types of edges, such as 'write' and 'cite.

DEFINITION 3.*meta path. A relevance path* P *is a path defined on a schema* S $=(V, R)$*, and is expressed as*

$$P = V_1 \overset{R_1}{\rightarrow} V_2 \dots V_{n-1} \xrightarrow{R_{n-1}} Vn$$

Each different type of Meta path represents different type of relationship. For example, a meta path $A \xrightarrow{\text{write}} P \xrightarrow{write\text{-}1} A$ indicates that the two authors have a coauthor relationship. An instance of the meta path P that is corresponds to the pattern of P. For example, path [un Wu → *ICDM16* → *Dong Zhang* in Figurel is an instance ofrneta path $A \xrightarrow{\text{write}} P \xrightarrow{write\text{-}1} A.$

DEFINITION4. *relationship strength based meta palh(RSMP) Given a HING = (V,E) and a meta patti P, the relationship strength oftwo nodes* o s ,o t E *V wah respect to P* is *defined as:*

$$S(Oi' OJ) = \sum_{p_{o_i \to o_j} \in P} s\left(o_i, o_j \middle| p_{o_i \to o_j}\right) \qquad (1)$$

Where $p_{o_i \to o_j}$ is a meta path instance ofP linking from *o,* to OJ , and s(*o.,* OJ $|p_{o_i \to o_j})$ is a relationship strength score. There different way to define $S(Oi, o_j | p_{o_i \to o_j})$. Here, we use Path Constrained Random Walk (PCRW).PCRW is a more sophisticated way to define the $s(o_i, o_j | p_{o_i \to o_j})$ based on an instance $p_{o_i \to o_j}$. According to this definition, $s(o_i, o_j | p_{o_i \to o_j})$ is the probability that a random walk restricted on P would follow the instance $p_{o_i \to o_y}$.We show itin table 1.

Problem Definition6. *For formation problem, given a graph G(V,E) and task T, find a group of experts V ′* C *V,* so *that experts in the V ′ contains the requested skills with T, and the communication cost* is *minimal.*

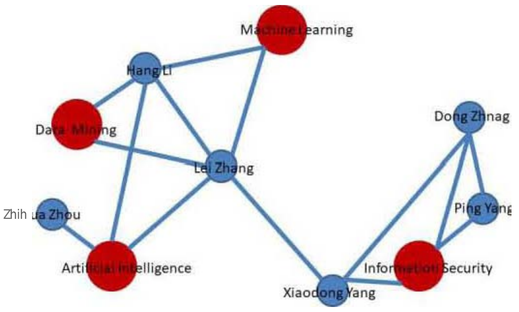

Figure.3. A example of heterogeneous network .Red nodes represent skills and blue nodes represent experts

In this paper, we will construct heterogeneous social network like Figure 3. The weight of the edge between experts represents the strength of the relationship. The greater the weight, the greater of relationship strength. Firstly, we compute the relationship strength between nodes by meta path.Then we find an optimum team in the network We use the Minimum Spanning Tree (Mst)[2] as communication cost function. We call The Mst communication cost of *V ı* as Cc-Mst(V'), which is defined as the cost of the minimum spanning tree that connects all the experts in *V ′* . We call the team fonnation problem using the Cc-Mst as the Mst-If problem.

IV. ALGORITHMS

In this section, we show the algorithin to compute the relationship strength based on meta path and the algorithm to compute optimum team for Mst-If problem in the heterogeneous social Network.

length	P	$p_{o_i \to o_j}$	SPCRW
2	APA	$a_1 \to p_3 \to a_2$	0.25
3	APPA	$a_i \to Pi \to P2 \to a2$	0.5
4	APTPA	$a_i \to Pi \to t_i \to P2 \to a2$	0.25
		$a_i \to P3 \to t3 \to P3 \to a2$	0.25
	APVPA	$a_i \to P3 \to v3 \to P3 \to a2$	0.25
……	……	……	……

Table 1. Meta paths and Instances connectmg al and a2.Length is length of Meta path, p is the pattern of meta path, $p_{o_i \to o_y}$ is instance ofP, SPCRW is probability.

A. *Relationship strength based meta polh(RSMP)*

According to Definition 4, in order to compute the relationship strength of objects OJ, OJ, we need to compute the corresponding relationship strength based each meta path P. For example, in Table 1, we list some meta paths that have instances connecting in Figure 1. We can see that the length of the meta path connecting the two authors has various lengths. In general, as the length of the meta path grows, the number of possible meta paths will increase exponentially. Too many meta paths are not conducive to our calculation of relationship strength.

In order to solve the above problem, in this paper, we use truncated estimation of proximity, which only consider meta paths under a length threshold *l*. According to our research, shorter paths have more information than longer ones, because those long meta paths contain some remote objects, these objects have less semantic associations. Therefore, we define:

$$s_l(o_i, o_j) = \sum_{len(P)<1} s(oi, ojIP) \qquad (2)$$

According to PCRW, we have following definition:

$$S,(Oi,Oj) = \sum_{(oi,o')EE} p\underset{i \to o}{:}(Oi;O')XSl\text{-}1(o',Oj) \qquad (3)$$

Based on the Equations, we develop a method(Alrorithin l)to compute the truncated relationship strength. On the whole, we compute the relationship strength matrix M_k for each k from 1 to I. Firstly, we initialize the matrix M_0 in lines 1-3.Then, we use the Equation2 to update the relationship strength matrix for each k in line 4-9.Finally, we can get truncated relationship strength Matrix M_I. In this part, we get the relationship strength between nodes in the Heterogeneous social Network. So, in section B, we can solve Mst-Tf problem by using it

```
Algorithin I: Iruncated Relationship Strength
Input: G= (V, E), length threshold I
Output: Iruncated Relationship Strength Matrix M,
1 $M_0 \leftarrow \emptyset$
2 for 0, E V do
3     $M_0[0'' \ o_i] \leftarrow 1.0$
4 for K E [1 ..·I]do
5     M, $\leftarrow \emptyset$
6     for 0, E V do
7         for 0' E neighbor($o_i$) do
8             for (O',Oj)EMk_,do
9                 $M_k[o.,0J] \leftarrow$
10                $M_k[o_i, o_j] + p_{o_{i\to o'}}^{\varphi(o_i,o')}$ X $M_{1-1}$(O', Oj)
```

B. Algorithms for the MST -Tfproblem

In this section we discuss algorithm 2 to solve the Mst-Tf problem: the SkillsFirst algorithin

```
Algorithm 2 Ihe SkillsFirst algorithin for the Mst-Tf problem
Input: Graph G (V, E); task I
Output; Ieam V' E V and subgraph G[V'] and communication cost Cc -Mst(V')
1 G, (I, $E_1$)← CompleteGraph(G, I)
2 $G_2$ (I, $E_2$)← MinSpaningIree(G,)
3 $G_3$← $G_3$ (T, ∅)
4 for e E E, do
5     (n, n2)← node(e)
6     $G_{3(V,}$ $E_{3)}$ ← $G_3$ U Path(n" n" G)
7 end
8 $G_4$ ← MinSpaningIree($G_3$)
9 Ieam V' ← Steiner'Ireet Cu,T)
10 Cc-Mst(V') ← $\sum_{e \in V'}$ dee)
```

Algorithin 2 shows the pseudocode of the SkillsFirst .For every requested skills, we get the complete graph G, (I, E,) by CompleteGraph(G, I). In complete graph G_1, for every edge (ei' ej)E E'' the distance between ei and ej is the same as the shortest path distance which is from ei and ej in the G. In line 2, we discover a minimum spanning tree of G_1 by MinSpaningIree (G,) which is based on Prim algorithm [15][16]. In line 3, we initialized G_3 which is with node I and no edge. In line 4-7, we build G_3 by replacing every edge in G_2 with its corresponding shortest path in G. In line 9, we get a Steiner tree[17] V' by removing edges in $G_{4.In}$ line 10, we get Cc-r-Mst (V') which is equal to the distance of its edges in V'.

In algorithm 2, we take O(IV + 'I'[Ioglt" + Tl + lEI) time to construct Gz.In this part, it consuming most time. In the rest of the algorithin, it probably spends O(IIllogIII + IE,I). SO, the time complexity of algorithin 2 is O(IV + 'I'[Ioglt" + Tl + lEI)

5、EXPERIMENT

In this part, we compare our algorithin with other algorithins through experiments. Ihrough experiments, we can verify that our algorithin is more efficient than other algorithin

A. Dataset

In the experiment, we used the DBLP data set Ihe data set contains a series of papers and their author information. For each paper, the title of the paper, the author of the paper, and the conference on the publication are known. This information helps us build heterogeneous social networks and discover meta paths very well. We only keep some major conferences related to computer science in the dataset, such as PDD, PKDD, SIGMOD, WWW and other conferences. We only keep a collection of experts who have published at least two papers in DBLP. An expert's skill is a set of terms that appear at least in the subject of two papers by experts. If at least two papers have been collaborated between the two experts, there will be an edge between the experts. Ihrough the above settings, according to the heterogeneous network construction method, the heterogeneous network required by this paper is constructed. Finally, the heterogeneous network based on DBLP contains 16523 nodes and 61742 edges.

B. Realtionship Strength Based Meta Palh(RSMP)

Before we complete the team formation problem, we need to calculate the strength of the relationship between the expert nodes in the heterogeneous network beforehand. In this paper, we use the relationship strength calculation method based on the meta path. According to the appeal algorithin, we need to limit the length of the meta path I. In this experiment, we limit the length of the meta path to 4. Ihen according to the algorithin 1 proposed in this paper, we calculate the strength of the relationship between heterogeneous network experts. Ihe weight of heterogeneous network edges is the communication cost between experts. According to the greater the strength of the relationship, the smaller the communication cost. We use the reciprocal of the relationship strength to represent the weight of the edge. In the end, we got a heterogeneous network with communication cost. Then we can handle the team formation problem on this heterogeneous network.

C. Result

In this paper, we compare our proposed team formation algorithm SkillsFirst with previous team formation algorithms. The comparison algorithms are EnSeniner, MinAggrSol and RaresFirst. In the team formation task, the skill I is composed of skills. In this paper, we separately experimented with four algorithms on the tasks with the number of tasks 3, 5, 7and 9. In order to highlight the role of RSPM, we need to do experiments in SkillsFirst based on

RSPM and SkillsFirst not based on RSPM. Then based on the experimental results, we can get their corresponded the time consumption and the communication cost of the team.

1) The communication cost

From Table 2, it can be concluded that the communication cost of SkillsFirst not based on RSMP is lower than that of the previous three algorithms, so it can be concluded that the team that generates in the heterogeneous network solution team can get a less cost. It shows that heterogeneous networks are more suitable for team formation than homogeneous networks. At the same time, the communication cost of the SkillsFirst formation team based on RSMP is smaller than not based on RSMP, which indicates that the RSMP proposed in this paper can better express the relationship strength between experts and reduce the communication cost of the formation team.

Number of skills	EnSeniner	MinAggrSol	RaresFirst	SkillsFirst without RSMP	SkillsFlrst **with** RSMP
3	4.69	5.37	3.16	2.93	1.87
5	6.14	7.36	4.03	3.78	2.45
7	9.97	9.52	7.19	6.23	4.56
9	14.74	14.11	10.78	8.96	5.13

Table2: The commumcation cost of EnSemner, MinAggrSol, RaresFirst, SkillsFirst not based on RSMP and SkillsFirst based on RSMP

2) The time consumption

From Table 3 we can see that the time consumption of SkillsFirst not based on RSMP is lower than the previous three, indicating that solving the team formation problem on the heterogeneous network can reduce the time consumption. At the same time, the time consumption of SkillsFirst based on RSMP is higher than that of SkillsFirst which is not based on RSMP. The extra time is mainly used to calculate RSMP. Because the team formation problem is more focused on the communication cost of the formation team, the SkillsFirst based on RSMP increases the time consumption, but reduces the communication cost, so it is completely acceptable.

Number of skills	EnSeniner	MinAggrSol	RaresFirst	SkillsFirst **without** RSMP	SkillsFlrst **with** RSMP
3	183	61	89	47	55
5	227	187	164	84	90
7	264	271	197	129	134
9	335	342	231	172	176

Table3: The time consumption of EnSemner, MmAggrSol, RaresFirst, SkillsFirst not based on RSMP and SkillsFirst based on RSMP

V. CONCLUSION

In this paper, we solve team formation problems in heterogeneous networks. We calculate the strength of the relationship between nodes based on meta path, which makes the relationship strength between nodes more reasonable and comprehensive. Through experiments in the DBLP dataset, we can find that the proposed algorithm has lower time complexity and communication cost than existing algorithms.

ACKNOWLEDGMENT

Sincerely thank my teachers, Professor Fang Wei and teacher Pengfei Wang. They are knowledgeable, academically rigorous, and meticulous in their work. In the past two and a half years, they have meticulously guided me not only in academics, but also gave me meticulous care in life.

REFERENCES

[1] Yan Jian - Ping.Team Management.Beijing : China Textile&Apparel Press , 2005 (in Chinese).

[2] Lappas, T., Liu, K., Terzi, E.: Finding a team of experts in social networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2009) .

[3] Xiang R , Neville J , Rogati M. Modeling relationship strength in online social networks // Proceedings of the 19th International Conference on World Wide Web.Raleigh ,USA, 2010 : 981 - 990.

[4] M. Ley. The dblp computer science bibliography: Evolution, research issues, perspectives. In SPIRE, pages 1-1 0. Springer, 2002.

[5] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In WWW, pages 697-706, 2007.

[6] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. In The Semantic Web, pages 722-735. Springer, 2007.

[7] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In SIGMOD, pages 1247-1250, 2008.

[8] Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, and X. Yu. Pathselclus: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. TKDD, 7(3):11, 2013.

[9] Baykasoglu A , Dereli T , Das S.Project team selection using fuzzy optimization approach.Cybernetics and Systems: An International Joumal , 2007 ,38 (2): 155 - 185

[10] Zzkarian A , Kusiak A.Forming teams : An analytical approach.IIE Transactions , 1999 , 3I (1): 85 - 97.

[11] Wi H , Oh S , Mun J , et al.A team formation model based on knowledge and collaboration. Expert Systems with Applications, 2009 , 36 (5): 9121 - 9134

[12] Kargar, M., An, A.: Discovering top-k teams of experts with/without a leader in social networks. In: Proceedings of the 20th ACM Intemational Conference on Information and Knowledge Management (201I)

[13] Mehdi Kargar, Morteza Zihayat, Aijun An, "Finding affordable and collaborative teams from a network of experts," SDM. Texas, pp. 587595, May 2013.

[14] Datta S, Majumder A, Naidu K, "Capacitated Team Formation Problem on Social Networks," KDD. Beijing, pp. 1005-1013, August 2012.

[15] Li C, Shan M, "Team formation for generalized tasks in expertise social networeks," Proc of IEEE Int Conf on Social Computing. Piscataway, NJ, pp. 2-9, March 2010.

[16] David R C, Robert E T, "Finding minimum spanning trees," in Siam Joumal on Computing, vol. 5, pp. 724-742, April 1976.

[17] AZ Zelikovsky, “A faster approximation algorithm for the steiner tree problem in graphs," Information Processing Letter, vol. 46, pp. 79-83, 1993.

[18] S.-J. Chen and L. Lin, "Modeling team member characteristics for the formation of a multifunctional team in concurrent engineering," IEEE Transactions on Engineering Management, vol. 51, pp. 111-124,2004.