Information Processing and Management xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Expert group formation using facility location analysis

Mahmood Neshati [a,*], Hamid Beigy [a], Djoerd Hiemstra [b]

[a] Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
[b] Database Research Group, Electrical Engineering, Mathematics and Computer Science (EEMCS) Department, University of Twente, The Netherlands

ARTICLE INFO

Article history:
Received 23 February 2013
Received in revised form 30 September 2013
Accepted 8 October 2013
Available online xxxx

Keywords:
Expert group formation
Expert finding
Topic model
Facility location analysis
Greedy approach
Linear programming

ABSTRACT

In this paper, we propose an optimization framework to retrieve an optimal group of experts to perform a multi-aspect task. While a diverse set of skills are needed to perform a multi-aspect task, the group of assigned experts should be able to collectively cover all these required skills. We consider three types of multi-aspect expert group formation problems and propose a unified framework to solve these problems accurately and efficiently. The first problem is concerned with finding the top k experts for a given task, while the required skills of the task are implicitly described. In the second problem, the required skills of the tasks are explicitly described using some keywords but each expert has a limited capacity to perform these tasks and therefore should be assigned to a limited number of them. Finally, the third problem is the combination of the first and the second problems. Our proposed optimization framework is based on the *Facility Location Analysis* which is a well known branch of the *Operation Research*. In our experiments, we compare the accuracy and efficiency of the proposed framework with the state-of-the-art approaches for the group formation problems. The experiment results show the effectiveness of our proposed methods in comparison with state-of-the-art approaches.

1. Introduction

The success of a project directly depends on the expertise of the people who are involved. Since the assignment of experts to a task/project must be based on both the required skills of the project and knowledge about the expertise of all candidate experts, it is not an easy task and it is challenging to optimize the assignment. As a result, expert group formation offers a new direction in research and also poses some brand new challenges. Here, the key problem is how to assign a group of experts to a given set of tasks/projects. The problem has attracted considerable interest from different domains. For example, several works have been made for conference paper-reviewer assignment (Karimzadehgan & Zhai, 2012; Karimzadehgan, Zhai, & Belford, 2008; Mimno & McCallum, 2007; Taylor, 2008), expert group formation in social networks (Kargar & An, 2011; Lappas, Liu, & Terzi, 2009) and optimal team formation in *Operation Research* problems (Wi, Oh, Mun, & Jung, 2009; Baykasoglu, Dereli, & Das, 2007).

In real scenarios, while several and sometimes diverse skills are needed to perform a task successfully and completely, these skills are implicitly expressed in the task descriptions. Besides the required skills of a task, in many cases, the relevant skills of experts are also implicitly reflected in their resume. The implicit notion of required skills/expertise aspects makes it difficult to assign an optimal group of experts to a given multi-aspect task. The main challenges of the expert group formation problem are:

* Corresponding author. Tel.: +98 2166166648.
E-mail addresses: neshati@mehr.sharif.edu (M. Neshati), beigy@sharif.edu (H. Beigy), d.hiemstra@utwente.nl (D. Hiemstra).

0306-4573/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ipm.2013.10.001

Please cite this article in press as: Neshati, M., et al. Expert group formation using facility location analysis. *Information Processing and Management* (2013), http://dx.doi.org/10.1016/j.ipm.2013.10.001

Implicit Notion of Aspects: Textual description of projects can only implicitly express the required skills of them, thus a method is needed to transform the textual description of a project into the set of required skills of that project. Similarly, because of implicit notion of expertise, a method is needed to transform the expertise documents (e.g. resume, professional profile, etc.) of each expert into the set of his/her skills.
Coverage Condition: In an ideal expert matching, all required skills of a project should be covered by the union of the skills of the assigned members in a complementary manner.
Confidence Condition: In an ideal expert matching, besides covering all required skills of a project, it is preferable that each member of the group individually be able to cover as many as possible the required skills of that project.
Load Balancing: In many applications of expert group formation, there is a limit on the number of tasks to be done by each expert. To balance the load and conform to the capacity of each expert, it is necessary to set up the problem as to simultaneously assign tasks to all the available experts with consideration of task-load balancing.

As a case study of the expert group formation problem, we consider the problem of review assignment. Review assignment is a common task that many people such as conference organizers, journal editors, and grant administrators would have to do routinely. In this problem, top k relevant reviewers (i.e. a group of experts with k members) should be assigned to each paper (i.e. task) such that all above mentioned criteria are satisfied. Firstly, the required skills for reviewing a paper can be explicitly determined by some keywords or can be inferred from the abstract/body of the paper. Secondary, the related research areas/skills of each reviewer (i.e. expert) can be explicitly expressed by some keywords or can be inferred from his/here previous publications. Thirdly, in an ideal matching, the assigned group of reviewers for a paper should be able to cover all aspects of that paper in a complementary manner. Fourthly, it is preferable that each assigned reviewer of a paper be able to cover as many as possible aspects of the paper. Finally, each member of the program committee (i.e. each reviewer) can only be involved in the review process of a limited number of papers.

Considering expert group formation as a top k retrieval problem, it can be solved by using a standard retrieval model (e.g. the baseline language model proposed in Karimzadehgan et al. (2008)), which computes the relevance score of each expert *individually* and then returns the k experts with the highest relevance scores. However, as the inter-relationships among the relevant experts are ignored, the top k search results are often quite alike and cannot cover all required aspects of a given task. In order to resolve the query aspects and avoid the information redundancy, it is necessary to optimize the top k search results collectively.

In this paper, we formalize the expert matching problem within the unified framework of *Facility Location Analysis* (FLA) (Gonzalez, 2007) taken from *Operation Research*, as a way to account and optimize the expert assignment. In the facility location problem, given a set of customer "locations" D, we would like to find a subset $S \subset D$ to open k "facilities" there, so as to optimize a graph-theoretic objective function that is dependent on the cost of opening a facility at each location and also the distance between each pair of customer and facilities. Specifically, facilities such as warehouses, hospitals, and fire stations, should be placed as close as possible to their customers. Since a customer would just go to the closest facility, there is a competitive relationship among those k-facilities. Similarly, in our proposed framework, we consider the top-k reviewers of each paper as the desirable facilities to be placed as close as possible to their customers (i.e. aspects of papers).

We show that our proposed method can improve the performance of expert matching in comparison with the state-of-the-art techniques for multi aspect/ skill expert matching such as *Greedy Next Best* (Karimzadehgan et al., 2008) and *Integer linear programming* (Karimzadehgan & Zhai, 2012). According to different conditions of the expert matching problem, we define three problems that can be solved by the proposed frame work of FLA. These problems are modeled using the unified framework of facility location analysis. In these problems, given a set of N papers and M reviewers, each paper should be assigned to a group of exactly k reviewers. The above mentioned three problems are given below.

- *Problem 1 (Implicit Aspects – Unconstraint Matching):* In this problem, we assume the aspects (i.e. required skills) of each paper are implicitly represented in the abstract of the paper and the skills of each reviewer can be inferred from the expertise document of that specific reviewer. Generally, the expertise document of an expert can be his/her resume but in this paper, we consider the concatenation of one's publications as his/here expertise document.
  In this problem, while each paper should be assigned to a group of k reviewers such that the skill coverage and confidence of the assigned group be maximal, there is no limitation on the capacity of reviewers (i.e. arbitrary number of papers can be assigned to a reviewer).
- *Problem 2 (Explicit Aspects – Constraint Matching):* In this problem, we assume that the set of the required skills of each paper and also the set of the relevant skills of each reviewer are explicitly determined (for example by using the ACM Categories and Subject Descriptors[1] keywords for computer science related papers). Given a limited number of reviewers, each paper should be assigned to a group of k reviewers such that: firstly, in an ideal matching, all aspects of *all*papers should be covered by the skills of the assigned groups (i.e. maximal coverage). Secondly, in an ideal matching, each member of the assigned groups for a paper should be able to cover as many as possible required skills of that specific paper (i.e. maximal confidence) and finally each reviewer should only be involved in review process of a limited (predefined) number of papers (i.e. load balancing).

[1] http://www.acm.org/about/class/2012.

Please cite this article in press as: Neshati, M., et al. Expert group formation using facility location analysis. *Information Processing and Management* (2013), http://dx.doi.org/10.1016/j.ipm.2013.10.001

- *Problem 3 (Implicit Aspects – Constraint Matching):* This problem is the combination of the first and the second problems. In this problem, we assume that the underlying aspects/skills of papers and reviewers are implicit and on the other hand, each expert has a limited capacity to review the assigned papers. The goal of this problem is to simultaneously maximize the coverage and the confidence of the assigned groups while the capacity condition is satisfied.

The basic idea of our work has been published in a paper of the ADCS conference (Neshati, Beigy, & Hiemstra, 2012). However, the conference paper does not have a complete description of the proposed algorithms due to the page limit. This paper is a significant expansion of the previous paper to add a new automatically generated test collection, more complete description of the algorithms, a new set of scalability and gap analysis experiments, and two theorems in paper which indicate the optimality of the proposed solution.

The rest of the paper is organized as follows. We first discuss facility location problems in Section 2 and our proposed framework for expert matching in Section 3. We then discuss related work in Section 4. In Section 5, we describe our experimental design and evaluation measures. In Section 6, we present our evaluation results. We conclude the paper in Section 7.

2. Facility location analysis

Facility location analysis is a branch of *Operations Research* (OR) (Gonzalez, 2007) and *Computational Geometry* (CG) (Berg, Cheong, Kreveld, & Overmars, 2008) concerning itself with mathematical modeling and solving problems which find the optimal placement of facilities in order to minimize transportation costs, avoid placing hazardous materials near housing, outperform competitors' facilities, etc. *Desirable k-facility placement* (Gabor & Van, 2005) is a type of facility location problems concerned with the selection of k optimal locations among M candidate locations to build k facilities such that the total cost of setup of these facilities and the transportation cost of the customers would be minimal. The goal of optimization in this problem is twofold:

- To minimize the total cost of opening those k facilities.
- To minimize the weighted distances from the customers locations to their closest facilities.

The set of facility locations or simply, facilities, and the set of customers form a part of the input for facility location problems. We denote the set of facilities by F and the set of customers by C. Typically, we are required to open a subset of the facilities and serve the demands of customers by assigning them to one of the open facilities. The distance between a pair of points $f_i \in F$, $c_j \in C$ is denoted by $D(f_i, c_j)$. If a customer j is served by a facility i in a solution, then the distance $D(f_i, c_j)$ is said to be the communication cost of customer j. The sum of communication costs of all customers is the *Communication Cost* of the solution. In the facility location problems, there is a cost associated with opening each facility $i \in F$, denoted by $cost(f_i)$. For a given solution, the cost of opening all the open facilities is called its *Building Cost*.

Various types of the facility location problems are defined in the literature for different usages (Gonzalez, 2007). Two main types of these problems are *uncapacitated* and *capacitated* facility location problems that can be useful to model the expertise matching problems. In this paper, we formally model the *unconstraint* (i.e. *Problem 1*) and *constraint*[2] expert matching problems (Eqs. (1) and (3)) using uncapacitated and capacitated facility location problems, respectively. In the following subsections, we give these problems as well as their approximate and exact solutions.

2.1. Uncapacitated Facility Location Analysis (UFLA)

In Uncapacitated Facility Location (UFLA) problem, exactly k facility locations should be selected among M available facility locations; such that while each customer is assigned to its nearest facility, the overall building cost of facilities and the communication cost of the solution would be minimal. In this problem, an arbitrary number of customers can be assigned to a facility. In other words, there is no constraint on the assignment of customers to the facilities.

Given the set of facilities F, the set of customers C, the distance between each customer c_j and facility location f_i as $D(f_i, c_j)$, the demand of each customer c_j as $demand(c_j)$, and the building cost of each facility f_i as $cost(f_i)$, the overall cost of selection k facilities can be calculated as follows Gonzalez (2007):

$$cost(S) = \lambda \overbrace{\sum_{i=1}^{k} cost(f_i)}^{BuildingCost} + (1-\lambda) \overbrace{\sum_{j=1}^{|C|} demand(c_j) \min_{f \in S} D(f, c_j)}^{CommunicationCost}. \quad (1)$$

In this equation, $S = \{f_1, \ldots, f_k\}$ indicates the set of selected facilities (i.e. the solution set) and $\lambda \in [0, 1]$ is a parameter. According to the above objective function, the total cost equals the sum of the *Building Cost* (i.e. the first summation) and the *Communication Cost* (the second summation).

Fig. 1 illustrates an instance of uncapacitated facility location problem in which the location of the customers and the facilities are indicated by circles and squares, respectively. Assuming equal building cost for all candidate locations (i.e.

[2] In this paper, by constraint expert matching, we mean expert group formation by considering the load balancing condition.

Please cite this article in press as: Neshati, M., et al. Expert group formation using facility location analysis. *Information Processing and Management* (2013), http://dx.doi.org/10.1016/j.ipm.2013.10.001

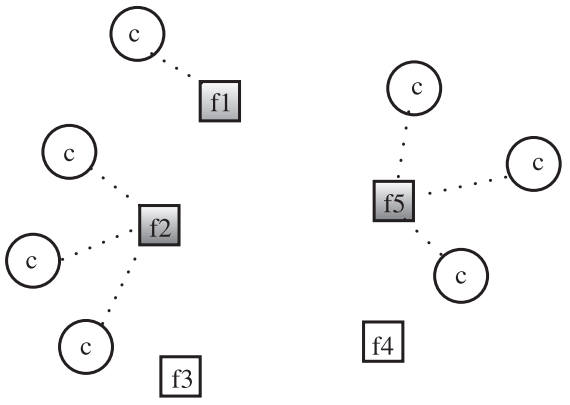

Fig. 1. Uncapacitated facility location problem-optimal facilities are indicated by gray color.

$cost(f_i) = cost(f_j)$, $\forall i, j \in \{1, \ldots, 5\}$), the optimal 3 facility locations and also the assignment of the customers to their nearest location is illustrated in this figure. As indicated in Fig. 1, exactly 3 facilities are selected among 5 available facility locations which minimize the overall communication cost.

As an integer linear programming problem, the uncapacitated facility location problem can be represented by the optimization of the ILP problem indicated in Eq. (2). In this integer linear program, U_i and $X(i,j)$ are the decision variables and $demand(c_j)$ and $D(f_i,c_j)$ are input parameters. U_i is a binary decision variable (according to constraint T_4) which indicates whether facility f_i is a member of final selected facilities (i.e. set S) or not. Constraint T_1 indicates exactly k facilities should be selected in final solution. $X(i,j)$ is also a binary variable (according to constraint T_5) which indicates whether customer c_j is assigned to facility f_i in final solution or not. Constraint T_2 indicates that each customer c_j should be assigned to at least one facility and constraint T_3 indicates that customer c_j can be assigned to facility f_i only if f_i is selected in the final solution.

$$\begin{aligned} &\text{minimize} \quad \lambda \sum_{i=1}^{M} U_i cost(f_i) + (1-\lambda) \sum_{j=1}^{|C|} \sum_{i=1}^{M} demand(c_j) D(f_i, c_j) X(i,j) \\ &\text{subject to} \\ &\qquad T_1 : \sum_{i=1}^{M} U_i = k \\ &\qquad T_2 : \forall j : 1 \leqslant \sum_{i=1}^{M} X(i,j) \\ &\qquad T_3 : \forall i,j : X(i,j) \leqslant U_i \\ &\qquad T_4 : \forall i : U_i \in \{0,1\} \\ &\qquad T_5 : \forall i,j : X(i,j) \in \{0,1\} \end{aligned} \tag{2}$$

The UFLA in general belongs to the class of NP-hard problems, which can be proved by reduction, for example, from the set cover problem (Gonzalez, 2007). Since this problem has an explicit objective function, it is possible to approximately optimize it using Greedy Local Search (GLS), a.k.a. Hill Climbing, as shown in Algorithm 1.

Algorithm 1. Greedy Local Search for UFLA problem

Input: F: (candidate facility locations), k (the cardinality of solution set)
Output: S: top k facility locations
$S \Leftarrow \{f_1, \ldots, f_k\}$
repeat
  for $f \in S$ do
    for$f' \in F - S$ do
      $S' \Leftarrow (S - \{f\}) \cup \{f'\}$
      If$Cost(S') < Cost(S)$ **then**
        $S \Leftarrow S'$
      end If
    end If
  end If
Until S does not change

Please cite this article in press as: Neshati, M., et al. Expert group formation using facility location analysis. *Information Processing and Management* (2013), http://dx.doi.org/10.1016/j.ipm.2013.10.001

The algorithm first initializes set S (i.e. the solution set) with a set of k random facilities and then iteratively refines S by swapping a facility location in S and an available non-selected location in F, until the process converges. Finally, the k selected facilities in S are an approximate solution for the problem.

2.2. Capacitated Facility Location Analysis (CFLA)

Capacitated facility location placement is another type of facility location problem with the same objective function as the uncapacitated facility location problem (i.e. Eq. (1)), but in this problem, each facility has a limited capacity to serve the assigned customers and therefore only a limited (and predefined) number of customers can be assigned to a facility (Gabor & Van, 2005). In this problem, the goal is finding the k facilities with minimal overall cost while the load balancing condition is satisfied. The inputs of this problem are the set of facilities F, the set of customers C, an integer k (i.e. the size of solution), the distance between each customer c_j and facility location f_i as $D(f_i, c_j)$, the demand of each customer c_j as $demand(c_j)$, the building cost of each facility f_i as $cost(f_i)$ and the capacity of each facility f_i (i.e. $capacity(f_i)$).

As an example, Fig. 2 illustrates the same customer and facility locations indicated in Fig. 1. Assuming equal building cost and equal capacity of two for each facility, Fig. 2 indicates the optimal four facilities for this CFLA problem. As indicated in this figure, each facility is responsible to serve to at most two customers and similar to the UFLA problem, each customer is assigned to the nearest opened/selected facility location. Clearly, this problem has no solution for $k = 3$ and $capacity(f_i) = 2$, because the number of customers is more than $3 * 2 = 6$.

While CFLA problem is in general NP-hard, various approximation algorithms (Charikar & Guha, 1999; Chudak & Williamson, 2005) are proposed for this problem. Specifically, Charikar and Guha in Charikar and Guha (1999) proposed an efficient linear programming solution for this problem. Although, we can use these approximated algorithms for modeling the expert matching problem, but because of the small size of the problem in our real applications and efficiency of linear programming, we chose to exactly solve the matching problem following the idea of linear programming. The explanation of our solution for constraint expert matching using the linear programming will be described in Section 3.4.

3. Multi aspect expert matching

In this section, we describe how to model the expert matching problems using the facility location framework. The list of symbols used in this paper is represented in Table 1.

Before describing the facility location framework for expert matching, we describe the author topic modeling method introduced in Karimzadehgan et al. (2008), which is used for implicit topic matching problems (i.e. *Problems* 1 and 3 described in Section 1).

3.1. Expert topic modeling

The notion of related aspects of papers and reviewers in *Problems* 1 and 3 of expert matching (described in Section 1) is implicit, making it difficult to find the optimal group of reviewers for a given paper considering the coverage and the confidence conditions. However, we can assume that the related aspects of a paper can be inferred from its abstract and the skills/aspects of reviewers can also be represented by his/her sample publications. To find the optimal group of reviewers, in this section, we describe a method to find a topic representation for each paper and each reviewer.

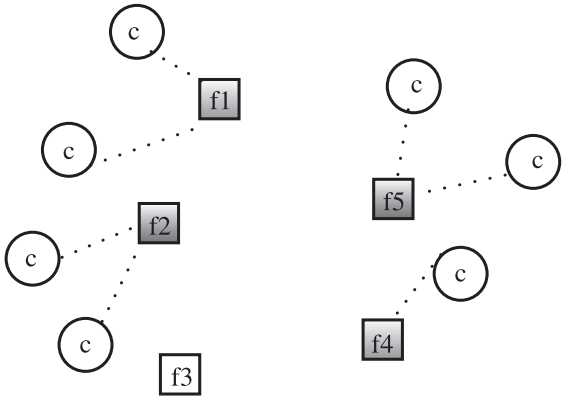

Fig. 2. Capacitated Facility location problem – optimal facilities are indicated by the gray color.

Please cite this article in press as: Neshati, M., et al. Expert group formation using facility location analysis. *Information Processing and Management* (2013), http://dx.doi.org/10.1016/j.ipm.2013.10.001

Table 1
Notations.

Symbol	Description
M	Number of reviewers/experts
N	Number of papers/tasks
T	Number of aspects/topics
e_i	One expert
p_j	One paper
τ_i	One topic
$A_{N\times T}$	Paper-topic matrix
$B_{M\times T}$	Reviewer-topic matrix
r_i	Expertise document of reviewer e_i
k	Number of assigned reviewers for each paper
c_i	Capacity of reviewer e_i
$\|S\|$	Cardinality of set S
$aspect(p_j)$	The set of aspects of paper p_j

Following the idea of reviewer modeling introduced in Karimzadehgan et al. (2008), we can assume that there is a space of T topic aspects, each characterized by a unigram language model such that the papers and the expertise documents can be represented as the mixture of these topics. Let $\tau = (\tau_1, \ldots, \tau_T)$ be a vector of topics where τ_i is a unigram language model and $p(w|\tau_i)$ is the probability of word w for the topic τ_i.

Given M reviewer's expertise documents,[3] arbitrary number of latent topic/aspects can be learned using Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999). Let $R = \{r_1, \ldots, r_M\}$ be the set of expertise documents where document r_i is the expertise document of reviewer e_i, the log likelihood of the expertise document collection according to the PLSA (Hofmann, 1999) equals to:

$$\log(P(R|\tau)) = \sum_{i=1}^{M}\sum_{w\in V} c(w, r_i) \log\left(\sum_{a=1}^{T} p(\tau_a|\theta_i)p(w|\tau_a)\right), \tag{3}$$

where V is the set of all words in the vocabulary, $c(w, r_i)$ is the count of word w in the expertise document r_i and $p(\tau_a|\theta_i)$ is the selection probability of topic τ_a for document r_i.

EM algorithm can be used to compute the maximum likelihood estimation of all parameters including $p(\tau_a|\theta_i)$ and $p(w|\tau_a)$. After learning all the parameters, each expert e_i can be represented using the topic vector τ_i denoted by $(\tau_{1i}, \ldots, \tau_{Ti})$ such that $\tau_{1i} = p(\tau_a|\theta_i)$. Furthermore, using the estimated values of $p(w|\tau_a)$, we can also infer the topic representation for each paper p_j as τ'_j denoted by $(\tau'_{1j}, \ldots, \tau'_{Tj})$ such that $\tau'_{1j} = p(\tau_a|\theta_j)$. Using the above topic modeling method, given N papers, M reviewers, and the number of latent topics T, the set of papers and reviewers can be represented by the paper-topic $A_{N\times T}$ and reviewer-topic $B_{M\times T}$ matrixes. The elements of these matrixes are real numbers in $[0, 1]$, which indicate the relevancy of corresponding paper or reviewer to a topic.

Note that if the relevant topics of papers and reviewers are explicitly given using some keywords (i.e. the *Problem* 2 of expert matching), then the paper-topic and reviewer-topic will be zero-one matrixes which can be defined as follows where $aspect(p_j)$ indicates the set of relevant aspects of paper p_j.

$$A(j,t) = \begin{cases} 0 & \text{if } t \notin aspect(p_j) \\ 1 & \text{if } t \in aspect(p_j) \end{cases} \quad B(i,t) = \begin{cases} 0 & \text{if } t \notin aspect(e_i) \\ 1 & \text{if } t \in aspect(e_i) \end{cases}$$

After representing papers and reviewers using the above matrixes, now we can describe the matching algorithm for retrieving the k-top reviewers for a given paper.

Note that the proposed method for topic learning in this section is applicable, if we have enough papers for a candidate reviewer to infer his/her skills. However, For "cold-start" reviewers (i.e. a reviewer who has not enough papers to infer his/her skills directly), following the idea of Deng, King, and Lyu (2012), it is possible to infer the related topics using the venue (or an specific section of the venue) in which his/her paper is published. For example, for reviewers who previously published a paper in "personalization and user modeling" section of the SIGIR conference, we can easily extract related aspects/skills of them. But in this method, we can only infer the general aspects of a reviewer profile and this method is not able to precisely detect the specific skill aspects of candidate reviewers. As another method, following the idea of relevance propagation (Qin, Liu, Zhang, Chen, & Ma, 2005), it is possible to infer the related aspects of a candidate reviewer using the aspects of his/her co-authors. Temporal topic modeling (AlSumait, Barbará, & Domeniconi, 2008; Daud, 2012) and also temporal expert finding (Hashemi, Neshati, & Beigy, in press) methods can also be used to more precisely infer the related topics of each reviewer.

[3] In this paper, the concatenation of a reviewer's publications is used as his/here expertise document.

Please cite this article in press as: Neshati, M., et al. Expert group formation using facility location analysis. *Information Processing and Management* (2013), http://dx.doi.org/10.1016/j.ipm.2013.10.001

3.2. Modeling expert group formation as facility location placement

The expert group formation problem is concerned with the assignment of N papers to M reviewers such that the following conditions are satisfied:

C_1*(Top-k retrieval)*: Each paper p_j should be assigned to a group of exactly k relevant reviewers.
C_2 *(Maximal Coverage)*: In an ideal matching, the group of reviewers assigned to paper p_j should be able to cover all aspects/topics (i.e. required skills) of paper p_j in a complementary manner.
C_3 *(Maximal Confidence)*: In an ideal matching, each reviewer e_i assigned to paper p_j should be able to cover as many as possible of the topics of paper p_j.

While above conditions are common in all group formation problems, the following condition should also be satisfied in constraint group formation problems (i.e. *Problems* 2 and 3 described in Section 1).

C_4 (*Load Balancing*): Reviewer e_i has a limited capacity and can only be involved in review process of a limited and predefined number of c_i papers.

Following the general idea of facility location analysis, we can reduce each expert group formation problem to a facility location placement problem. These two problems are similar to each other in many aspects:

1. While exactly k facilities should be selected in a k-facility placement problem, in expert group formation, each paper should also be assigned to exactly k reviewers.[4]
2. While each customer in a facility location problem should be assigned to its nearest facility, in expert group formation, each topic of a paper should be assigned to a reviewer who is able to cover that specific topic. If all aspects of a paper p_j is assigned to at least one relevant reviewer, then the maximal coverage condition can be satisfied.
3. While in the facility location problems, each facility has a specific value of opening cost, in expert group formation problem, each reviewer has a specific confidence level for a given paper. So, if we appropriately define the assignment cost of each reviewer for a given paper, then the maximal confidence condition can be satisfied.
4. In capacitated facility location, each facility has a limited capacity to serve customers; similarly, in constraint expert group formation, each expert should be assigned to a limited number of papers. Therefore, using capacitated facility location framework, the load balancing condition can be satisfied.

According to the above similarities between facility location placement and expert group formation problems, we can imagine each reviewer e_i as a facility and each topic/aspect τ of paper p_j as a customer that should be assigned to a facility/reviewer. Given the set of papers, reviewers, paper-topic matrix, reviewer-topic matrix and the number of reviewers for each paper, we should define the building and communication costs to complete the facility location framework. In the following subsections, we will define these costs for each of the group formation problems.

3.3. Implicit Aspects – Unconstraint Matching

The first problem of expert matching is concerned the assignment of N papers to M reviewers such that the conditions C_1,C_2, and C_3 are satisfied. In this problem, we assume that each reviewer has infinite capacity. So, we can assign each of N papers to all available reviewers independent of other papers. In other words, we can independently solve the matching problem for each paper.

Using the topic modeling method described in Section 3.1, we can infer the paper-topic A and reviewer-topic B matrixes. As mentioned in Section 3.2, selection of k reviewers is equivalent to the selection of k facilities in our framework. For a given paper, we have T customers (i.e. T topics) and M candidate facility locations (i.e. M reviewers) and the objective function in UFLA is composed of two parts:

1. *Building Cost*: The building cost indicates the cost of opening a facility at a specific candidate location.
2. *Communication Cost*: The communication cost indicates the access cost of a customer to its nearest facility in the optimal solution.

In expert matching problem, the first part corresponds to the cost of selection of an expert e_i for a paper p_j and the second part indicates the cost of coverage for a topic of a paper p_j by the most relevant assigned reviewer. Therefore, we can define the building cost and communication cost in our framework as follows.

[4] The value of k is not necessarily equal for all papers.

Please cite this article in press as: Neshati, M., et al. Expert group formation using facility location analysis. *Information Processing and Management* (2013), http://dx.doi.org/10.1016/j.ipm.2013.10.001

1. *Building cost* of assignment of reviewer e_i to paper p_j equals to:

$$cost(e_i) = D(\overrightarrow{e_i} \| \overrightarrow{p_j}),$$

where $\overrightarrow{e_i}$ and $\overrightarrow{p_j}$ correspond to the topic vector of reviewer e_i (i.e. the ith row of the reviewer-topic matrix B) and the topic vector of paper p_j (i.e. the jth row of the paper-topic matrix A) and $D(\overrightarrow{e_i} \| \overrightarrow{p_j})$ indicates the Kullback–Leibler (KL)[5]divergence value of these vectors.
Intuitively, if the topic distributions of vectors $\overrightarrow{e_i}$ and $\overrightarrow{p_j}$ are very similar to each other, then we expect that they might be related to the same topics and also their KL divergence value will be very small. As a result, the facility (i.e. the reviewer) e_i will be a very low building cost facility candidate for paper p_j.
2. *Communication cost* of assignment of aspect a of paper p_j to reviewer e_i equals to

$$cost(e_i, \tau_{aj}) = D(\overrightarrow{e_i} \| \overrightarrow{v_a}),$$

where $\overrightarrow{e_i}$ is topic vector of reviewer e_i, τ_{aj} is the weight of topic a in topic vector of paper p_j and $\overrightarrow{v_a}$ indicates the unit vector with all zero elements except for topic a.
In this case, if reviewer e_i is able to cover skill a, then we expect that the weight of topic τ_a in his/her topic vector $\overrightarrow{e_i}$ be higher in comparison with other topics and accordingly the communication cost of customer a (i.e. topic a) and facility e_i (i.e. the reviewer) will be low.

According to the above definitions for building and communication costs, the objective function of unconstraint expert matching problem can be represented as follows:

$$Cost(S, p_j) = \lambda \sum_{i=1}^{k} D(\overrightarrow{e_i} \| \overrightarrow{p_j}) + (1-\lambda) \sum_{a=1}^{T} \tau_{aj} \min_{s \in S} D(\overrightarrow{s} \| \overrightarrow{v_a}), \quad (4)$$

where S is the set of selected reviewers for paper p_j and τ_{aj} indicates the weight of topic a in topic vector of paper p_j. To optimize the above objective function we use the greedy local search method given in Algorithm 1. The output of the algorithm is the top-k reviewers for paper p_j which have not only the maximum confidence but also collectively can cover all aspects of that paper.

3.4. Explicit Aspects – Constraint Matching

The second problem of expert matching (i.e. *Explicit Aspects – Constraint Matching*) is concerned with the assignment of N papers to M reviewers such that in addition to the conditions C_1,C_2 and C_3, the load balancing condition (i.e. C_4) is also satisfied.

In contrast with the first problem of expert matching, in this problem, we assume that the aspects/skills of the papers and experts are explicitly determined. This is a valid assumption because in many applications, the required skills of a project and also the related skills of an expert can be explicitly described by some few keywords. For example, in the paper-reviewer assignment problem, the relevant topics of each paper and reviewer can be described by some few keywords.

The constraint C_4 makes the matching problem very hard, indeed this matching problem belongs to the class of NP-hard problems; furthermore, in this problem the conditions C_2 and C_3 are in competition with each other. Specifically, maximizing the global average confidence of the assigned groups may result in formation of the non-optimal converging groups. In this section, we formally model this problem using the Capacitated Facility Location Analysis (CFLA) (Chudak & Williamson, 2005) and also propose an exact linear programming solution for it.

Similar to Section 3.3, in the CFLA framework of constraint expert matching, each aspect/topic of paper p_j is considered as a customer and each reviewer is considered as a candidate facility location.

Algorithm 2 indicates the linear programming formulation (Gonzalez, 2007) for this CFLA problem. In this linear program, matrix U_{ji} is a $N \times M$ binary decision matrix that indicates the assignment of papers to the reviewers. Specifically, element $u_{ji} = 1$ if and only if, in the final solution, the paper p_j is assigned to the reviewer e_i. As a binary decision variable, $X(i,j,a)$ indicates the assignment of topic a of paper p_j (i.e. a customer) to reviewer e_i (i.e. a facility); specifically, $X(i,j,a) = 1$ if and only if, topic a of paper p_j is assigned to reviewer e_i, and finally, $A_{N \times T}$ is the binary paper-topic matrix; where $A(j,a) = 1$ if and only if, paper p_j is related to topic a.

[5] For discrete probability distributions P and Q, the KL divergence is defined to be

$$D_{KL}(P \| Q) = \sum_i \ln\left(\frac{P(i)}{Q(i)}\right) P(i)$$

It is the expectation of the logarithmic difference between the probabilities P and Q, where the expectation is taken using the probabilities P. The KL divergence is only defined if P and Q both sum to 1 and if $Q(i) = 0$ implies $P(i) = 0$ for all i (absolute continuity). If the quantity 0 ln 0 appears in the formula, it is interpreted as zero because $\lim_{x \to 0} x \ln(x) = 0$ (Kullback & Leibler, 1951). This interpretation makes sense in our problem, because a zero-weight topic of a paper should not affect its distance from reviewers (i.e facilities).

Please cite this article in press as: Neshati, M., et al. Expert group formation using facility location analysis. *Information Processing and Management* (2013), http://dx.doi.org/10.1016/j.ipm.2013.10.001

Algorithm 2. Linear programming formulation of CFLA

$$\text{minimize } \sum_{j=1}^{N}\sum_{i=1}^{M}(\lambda U_{ji} BCost(i,j) + (1-\lambda)\sum_{a=1}^{T} CCost(i,j,a)X(i,j,a))$$

subject to

$$T_1 : \forall j : \sum_{i=1}^{M} U_{ji} = k$$
$$T_2 : \forall i : \sum_{j=1}^{N} U_{ji} \leqslant c_i$$
$$T_3 : \forall j, a : A(j,a) \leqslant \sum_{i=1}^{M} X(i,j,a)$$
$$T_4 : \forall i, j, a : X(i,j,a) \leqslant U_{ji}$$
$$T_5 : \forall i, j : U_{ji} \in \{0,1\}$$
$$T_6 : \forall i, j, a : X(i,j,a) \in \{0,1\}$$

In the linear program given in Algorithm 2, constraint T_1 shows that the sum of elements of each row of matrix U should be equal to k; this means that each paper should be assigned exactly to k reviewers. This constraint can satisfy the top-k retrieval condition (i.e. the condition C_1).

Constraint T_2 indicates that the sum of elements of each column of U should be less than or equal with c_i (i.e. the capacity of the ith reviewer); this means that reviewer e_i can only be assigned to at most c_i papers. This constraint can satisfy the load balancing condition (i.e. the condition C_4).

Constraint T_3 indicates that for each related topic of paper p_j (i.e. for topics that $A(j,a) = 1$) at least one reviewer should be assigned. On the other hand, T_4 indicates that the topic a of the paper p_j can be assigned to the reviewer e_i only if paper p_j is assigned to the reviewer e_i. In other words, if the decision variable $X(i,j,a) = 1$, then the element u_{ji} of the assignment matrix should be equal to 1; As we will see later, these constraints can guarantee the optimal coverage assignment (i.e. the condition C_2).

The objective function in Algorithm 2 is the same as the objective function given in Eq. (4) with this difference that in Algorithm 2, it is defined to globally optimize the matching of all papers simultaneously (i.e. the outer sum is defined on all papers). On the other hand, the minimum distance in Eq. (4) is replaced by decision variable $X(i,j,a)$ which models the same concept (i.e. the best assigned reviewer for topic a).

Intuitively, by minimizing the objective function, both the coverage and the confidence conditions of the matching problem (i.e. C_2 and C_3 conditions) can be satisfied. Firstly, paper p_j is assigned to reviewer e_i (i.e. $u_{ji} = 1$) when the building cost (i.e. $BCost(i,j)$) of this selection is low; this part of objective function can satisfy confidence maximization (i.e. condition C_3). On the other hand, minimization of the communication cost for the relevant topic a of paper p_j (i.e. $CCost(i,j,a)$) results in the assignment of topic a to a reviewer e_i such that e_i is able to cover this topic; therefore this part of the objective function can satisfy the coverage maximization condition (i.e. condition C_4). The parameter λ can be used to make a tradeoff between confidence and coverage conditions. In order to maximize the coverage and confidence of the assigned groups, we can define the building and communication cost as follow.

$$BCost(i,j) = 1 - \frac{\text{aspect}(p_j) \cap \text{aspect}(e_i)}{\text{aspect}(p_j)} \tag{5}$$

$$CCost(i,j,a) = \begin{cases} 0 & \text{if } a \in \text{aspect}(e_i) \\ 1 & \text{if } a \notin \text{aspect}(e_i) \end{cases}. \tag{6}$$

In above equations, aspect (e_i) indicates the set of relevant topics for reviewer e_i and aspect (p_j) is the set of relevant topics of paper p_j. Intuitively, $BCost(i,j)$ indicates the fraction of aspects of paper p_j, which cannot be covered by the reviewer r_i and $CCost(i,j,a)$ indicates whether the assigned reviewer r_i for topic a of paper p_j is able to cover it or not. In next two theorems, we prove that Algorithm 2 can produce the optimal coverage and confidence expert assignment.

Theorem 1 (*Coverage optimality*). *In the second problem of expert matching, if communication cost is defined by Eq. (6) and $\lambda = 0$, then Algorithm 2 will produce the optimal converging groups.*

Proof. Let U be the $M * N$ binary assignment matrix where $u_{ij} = 1$ when paper p_j is assigned to reviewer r_i in the final solution. According to the definition of the coverage measure (Section 5.2), the average value of coverage for N papers according to assignment matrix U is the average value of coverage of each paper and can be computed as follows

$$Coverage(U) = \frac{\sum_{j=1}^{N} Coverage(p_j)}{N} = \frac{1}{N}\sum_{j=1}^{N} \frac{|\text{aspect}(p_j)| - |\text{NCaspect}(p_j)|}{|\text{aspect}(p_j)|}.$$

where |aspect (p_j)| indicates the number of aspects of paper p_j and |NCaspect (p_j)|[6] indicates the number of its aspects, which are not covered by the assigned group of reviewers according matrix U. *Coverage*(U) indicates the average value of coverage for

[6] NCaspect = Not Covered aspects.

Please cite this article in press as: Neshati, M., et al. Expert group formation using facility location analysis. *Information Processing and Management* (2013), http://dx.doi.org/10.1016/j.ipm.2013.10.001

all papers and $Coverage(p_j)$ is the coverage measure for paper p_j. With simple mathematic operations, $Coverage(U)$ can be computed using the following equation.

$$Coverage(U) = 1 - \frac{1}{N}\sum_{j=1}^{N}\frac{|\text{NCaspect}(p_j)|}{|\text{aspect}(p_j)|}. \quad (7)$$

On the other hand, setting $\lambda = 0$ in Algorithm 2 results in the following objective function.

$$OBJ_1 = \sum_{j=1}^{N}\sum_{i=1}^{M}\sum_{a=1}^{T} CCost(i,j,a)X(i,j,a)$$

By minimizing the above objective function, the value of decision variables $X(i,j,a)$ are determined as follows.

1. if $A_{ja} = 0$, then for all values of i the value of $X(i,j,a) = 0$. In other words, if paper p_j is not related to topic a, then this topic will not be assigned to any reviewer.
2. if $A_{ja} = 1$, then only for exactly one value of $i = i_1$ (i.e. the best available reviewer) the value of $X(i_1,j,a) = 1$ and for all other values of i value of $X(i_1,j,a) = 0$. In other words, for each relevant topic a of paper p_j exactly one reviewer will be assigned. This assigned reviewer may be able to cover topic a or not (i.e. $CCost(i_1,j,a)$ can be zero or one). So, if reviewer i_1 is able to cover topic a then the value of objective function does not change; otherwise it increases by 1.

According to the above cases, for each paper p_j the value of

$$OBJ_1(j) = \sum_{i=1}^{M}\sum_{a=1}^{T} CCost(i,j,a)X(i,j,a) = |\text{NCaspect}(p_j)|$$

equals the value of aspects of paper p_j, which is not covered by the assigned group of reviewers (i.e. $|\text{NCaspect}(p_j)|$). Substituting in Eq. (7), we obtain the following equation.

$$Coverage(U) = 1 - \frac{1}{N}\sum_{j=1}^{N}\frac{OBJ_1(j)}{|\text{aspect}(p_j)|}$$

It is obvious by minimizing the values of OBJ_1 ($OBJ_1 = \sum_j OBJ_1(j)$) and normalizing the cost by the number of topics of paper p_j, the coverage measure will be maximized. □

Theorem 2 (*Confidence optimality*). *In the second problem of expert matching, if building cost is defined by Eq. (5) and $\lambda = 1$, then Algorithm 2 will produce the optimal confidence groups.*

Proof. According to the definition of the average confidence (Section 5.2), the average confidence score of an assigned group to paper p_j can be computed as follow

$$\text{Confidence}(p_j) = \sum_{i=1}^{M}\sum_{a=1}^{T}\frac{U_{ji}B_{ia}A_{ja}}{|aspect(p_j)|k}.$$

Therefore, the average value of average confidence for all assignments according to the matrix U is like follows.

$$\text{Confidence}(U) = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{M}\sum_{a=1}^{T}\frac{U_{ji}B_{ia}A_{ja}}{|aspect(p_j)|k},$$

where A, B, U are the paper-topic, the reviewer-topic and the assignment matrixes respectively, and k is the number of reviewers assigned for each paper. The above equation can be re-written as follows.

$$\text{Confidence}(U) = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{M}\frac{U_{ji}}{|aspect(p_j)|k}\left(\sum_{a=1}^{T}B_{ia}A_{ja}\right) \quad (8)$$

On the other hand, by setting $\lambda = 1$, the objective function of Algorithm 2 becomes as follows.

$$OBJ_2 = \sum_{j=1}^{N}\sum_{i=1}^{M}U_{ji}BCost(i,j)$$

According to definition of $BCost(i,j)$ (i.e. Eq. (5)), we have:

$$|aspect(p_j)\bigcap aspect(r_i)| = |aspect(p_j)|(1 - BCost(i,j)).$$

Please cite this article in press as: Neshati, M., et al. Expert group formation using facility location analysis. *Information Processing and Management* (2013), http://dx.doi.org/10.1016/j.ipm.2013.10.001

On the other hand, Eq. (9) is true because the term $B_{ia}A_{ja}$ is equal to one when both B_{ia} and A_{ja} are equal to one (both paper p_j and reviewer r_i are relevant to topic a).

$$\sum_{a=1}^{T} B_{ia}A_{ja} = |aspect(p_j) bigcapaspect(r_i)| \tag{9}$$

So, we have the following equality:

$$\sum_{a=1}^{T} B_{ia}A_{ja} = |aspect(p_j)|(1 - BCost(i,j)) \tag{10}$$

According to Eqs. (8) and (10), we have

$$\text{Confidence}(U) = \frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{M} U_{ji}\left(\frac{1 - BCost(i,j)}{k}\right)$$

With simple modifications, we obtain:

$$\text{Confidence}(U) = \frac{1}{kN}\left[\sum_{j=1}^{N}\sum_{i=1}^{M} U_{ji} - \sum_{j=1}^{N}\sum_{i=1}^{M} U_{ji}BCost(i,j)\right]$$

According to condition T_1 in Algorithm 2, we have the following equation.

$$\sum_{j=1}^{N}\sum_{i=1}^{M} U_{ji} = kN$$

So the following equation is always true:

$$\text{Confidence}(U) = \frac{1}{kN}(kN - OBJ_2) = 1 - \frac{OBJ_2}{kN}.$$

According to the above equation, it is obvious that by minimizing the OBJ_2, the confidence measure will be maximized. □

The above theorems show that the linear programming solution can produce the optimal solution in terms of the coverage and confidence measures. Parameter λ can be used to make a trade-off between these measures.

3.5. Implicit Aspects – Constraint Matching

The constraints in the third problem of expert matching (i.e. *Implicit Aspects – Constraint matching*) are similar to the second problem but in this problem the topics/aspects of papers and reviewers are not predetermined. So, we use the topic modeling method introduced in Section 3.1 to infer the paper-topic and reviewer-topic matrixes. Fortunately, the linear programming method proposed in Algorithm 2 can be used to solve this matching problem. We only need to define the building and communication cost appropriately for implicit aspects. Similar to the method proposed for unconstraint implicit matching in Section 3.3, we define the building and communication cost in the same way described in Eq. (4).

3.6. Solving the integer linear program

Once our problem is formulated, we can use many algorithms to solve it. In our experiments, we use the commercial ILOG CPLEX 12.5 package[7] to solve our reviewer matching problem.

ILOG CPLEX optimizer is able to solve the linear programs with millions of constraints. Specifically, CPLEX uses the *Branch-and-Cut* (Mitchell, 2002) algorithm, which is an exact algorithm consisting of a combination of a cutting plane method with a Branch and-Bound algorithm, to solve the integer linear programs.

The idea of the Branch-and-Bound algorithm is to take a problem and decompose it into smaller problems such that a solution to a smaller problem is also a solution to the given problem. The algorithm recursively decomposes the original problem until it can be solved directly or is proven not to lead to an optimal solution.

In order to solve our integer linear program, using the LP relaxation, the original problem is transformed to a linear programming sub problem which is easy to solve optimality. This linear program without integer constraints is solved using the simplex algorithm. Branch and cut involves running a branch and bound algorithm and using cutting planes to tighten the linear programming relaxations. Specifically, when the optimal solution for the non-integer sub problem is founded, a cutting step tries to add additional constraints. In other words, the cutting step is to generate valid inequalities for the integer hull of the current sub problem and add them to the LP relaxation of the sub problem. At this point, the problem is divided into two sub problems: one is to explore values greater than or equal to the smallest integer greater than the current value, and the other is to explore values less than or equal to the next lesser integer. These

[7] Refer to http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/.

Please cite this article in press as: Neshati, M., et al. Expert group formation using facility location analysis. *Information Processing and Management* (2013), http://dx.doi.org/10.1016/j.ipm.2013.10.001

new linear programs are then solved using the simplex method and the process repeats until a solution satisfying all the integer constraints (or binary constraints) is found. For complete information about the algorithm, a reader is referred to Mitchell (2002).

4. Related work

The problem of expert group formation has recently attracted lots of attention in information retrieval (Karimzadehgan & Zhai, 2012; Karimzadehgan et al., 2008; Tang et al., 2012) and social network communities (Kargar & An, 2011; Lappas et al., 2009). This problem is an extension of the expert finding problem (Balog, Soboroff et al., 2009; Craswell, Vries, & Soboroff, 2005; Soboroff, de Vries, & Craswell, 2006). The inclusion of the expert finding task in TREC Enterprise has attracted lots of attention from 2005 to 2008. Expert finding task is concerned with the retrieval of expert people in a given topic (Balog, Azzopardi, & De Rijke, 2009). Several methods are proposed for expert finding problem such as the language modeling (Balog, Azzopardi et al., 2009), voting model (Macdonald, 2009), and person centric language modeling (Serdyukov & Hiemstra, 2008). While these initial approaches for expert finding were proposed to find the expert persons in an organization (Balog, Azzopardi et al., 2009; Macdonald, 2009; Serdyukov & Hiemstra, 2008), recent methods focused on finding experts in the bibliographic data (Deng, Han, Lyu, & King, 2012; Deng, King et al., 2012; Hashemi et al., in press), social networks (Neshati, Asgari, Hiemstra, & Beigy, 2013; Smirnova, 2011), and question answering forums (Pal & Konstan, 2010).

The problem of multi aspect expert group formation is introduced by Karimzadehgan et al. (2008). Specially, they considered the first problem of expert matching (i.e. *Implicit Aspects – Unconstrained Matching* introduced in Section 3.3) and proposed three different strategies to find a group of experts that maximally covers the required skills of a given query. These methods are (1) the redundancy removal, (2) expert aspect modeling and (3) query aspect modeling (Karimzadehgan et al., 2008). The idea of the redundancy removal strategy is to diversify the set of retrieved experts such that experts with various skills can be selected to maximally cover the required skills of the query. In the query aspect modeling approach, a multi-aspect query is segmented into semantically diverse parts such that each part can be considered as a single aspect query; then, each query part is used to retrieve relevant experts. Finally, the union of retrieved experts for each sub-query is considered as the final answer.

The most effective method proposed in Karimzadehgan et al. (2008) is expert aspect modeling. Similar to our approach for the first problem of expert matching, it is based on learning a topic vector representation for experts (reviewers) and the queries (papers). Using these topic vectors, the *Next Best* greedy approach is utilized to form the optimal skill covering group. In this approach, the members of a group are selected step by step; At step n, a reviewer (i.e. the best candidate in step n) which minimizes the following objective function is selected as the new member of the group.

$$D(\theta_q)\|\left(\frac{\sigma}{n-1}\sum_{i=1}^{n-1}\theta_i + (1-\sigma)\theta_n\right)$$

In this equation, θ_i indicates the topic vector of the ith selected member of the group (specifically, θ_k is the topic vector of the kth candidate reviewer), θ_q is the topic vector of the paper and σ is a parameter, which models the redundancy of skills in selected members. The objective function is the KL-divergence of the topic distribution of the query/paper and the resulting group after selection of the kth expert candidate.

The intuition behind this method is that θ_q gives a measure of which topic aspect is relevant; As a result, if θ_q assigns a high probability to some aspects, then it would prefer to cover these relevant aspects more than other less relevant ones. Thus, the best r_k is the one that works together with $r_1, \ldots, r_{k-1}$ to achieve a coverage distribution most similar to the topic coverage given by the query. So, parameter σ controls how much to rely on the previously picked reviewers $r_1, \ldots, r_{k-1}$ to cover all topic aspects of the query.

In contrast with the *Next Best* greedy approach (Karimzadehgan et al., 2008), our proposed method for implicit aspect matching (illustrated in Algorithm 1), measures at each iteration the ability of the whole k-members of the assigned group to cover the required skills of the query and as a result if a non-appropriate member is selected at the ith step, it can be eliminated at the following steps. However, in the *Next Best* greedy approach, a non-appropriate selected reviewer for a paper cannot be changed. On the other hand, the *Next Best* greedy approach cannot easily be extended for constraint matching problems (i.e. the second and third problems of expert matching). In contrast, our FLA framework for expert matching can be extended for constraint and explicit aspect matching problems.

The problem of constraint expert group formation (i.e. the second and third problems of expert matching) has been recently introduced in Karimzadehgan and Zhai (2012). This problem can be considered as an extension of the paper-review assignment problem. While initial approaches for paper-review assignment (Hettich & Pazzani, 2006; Sun, Ma, Fan, & Wang, 2007) concern only with the assignment of a paper to the relevant reviewers and ignore the coverage, confidence and load balancing conditions, Karimzadehgan and Zhai (2012) introduced the problem of multi-aspect constraint paper-review matching. They proposed an integer linear programming (ILP) method for this problem which is demonstrated in Algorithm 3.

Please cite this article in press as: Neshati, M., et al. Expert group formation using facility location analysis. *Information Processing and Management* (2013), http://dx.doi.org/10.1016/j.ipm.2013.10.001

Algorithm 3. Integer linear programming method for constraint expert matching proposed in Karimzadehgan and Zhai (2012)

$$\begin{aligned}
&\text{maximize } \sum_{j=1}^{N}\sum_{a=1}^{T} t_{ja} \\
&\text{subject to} \\
&\quad T_1 - \forall j: \ \sum_{i=1}^{M} U_{ji} = k \\
&\quad T_2 - \forall i: \ \sum_{j=1}^{N} U_{ji} \leqslant c_i \\
&\quad T_3 - \forall j, a: \ A_{ja} t_{ja} \leqslant \sum_{a=1}^{T} B_{ak} U_{aj} \\
&\quad T_4 - \forall i, j: \ U_{ji} \in \{0, 1\} \\
&\quad T_5 - \forall i, j: \ t_{ji} \in \{0, \ldots, k\}
\end{aligned}$$

In this algorithm, we used the same notation introduced in Table 1 and Algorithm 2. In this algorithm, decision variable $t_{ji} \in [0, k]$ is an integer indicating the number of assigned reviewers of paper p_j that can cover subtopic a of it. Intuitively, this objective function prefers assignments that maximize the coverage of all the subtopics (the inner sum) for all the papers (the outer sum).

In contrast with our proposed framework, this method cannot necessarily optimize the skill coverage of the assigned groups. On the other hand, their method for implicit topics/aspects (i.e. third problem of expert matching) has very low performance in comparison with our CFLA method. We use the methods proposed in Karimzadehgan et al. (2008, 2012) as our baseline algorithms and also use the same dataset to make the results comparable.

Recently, Tang et al. (2012) proposed a general framework based on the convex cost flow optimization for expert matching. Their proposed method mainly focused on authority and soft load balancing constraints and does not solve the coverage and confidence conditions optimally. As another related line of research, authors of Lappas et al. (2009) proposed a method to find a group of experts in a social network. Although this research is closely related to the expert matching problem, their main concern is to find a group of experts in a social network which are able to contribute with each other easily.

5. Experiments

In this section, we present the test data and measures used for evaluating our methods.

5.1. Data set

In our experiments, we use two datasets which are described bellow:

SIGIR dataset: We used the dataset introduced in Karimzadehgan et al. (2008) to evaluate our proposed methods. This dataset is used in several research papers (i.e. Karimzadehgan & Zhai, 2012; Karimzadehgan et al., 2008; Tang et al., 2012) and to the best of our knowledge is the only available dataset for the multi aspect team formation problem. The dataset is crawled from the abstract papers of ACM SIGIR proceedings from years 1971 to 2006. Authors of these papers are considered as the prospective reviewers/experts. For modeling reviewers' expertise, a profile is created for each author by concatenation of all papers written by that specific author. The SIGIR 2007 papers are used to simulate papers that are to be reviewed. In this dataset, there are 73 papers with at least two aspects. A gold standard is created for this dataset by identifying 25 major subtopics for these papers and then assignment of subtopics to all papers and the reviewers by a human expert. In total, there are 73 papers and 189 reviewers in this dataset, which is publicly available at http://timan.cs.uiuc.edu/data/review.html.

PubMed dataset: Because of the limited size of our first dataset (Karimzadehgan et al., 2008), we automatically built a new paper-reviewer dataset. We crawled multi aspect papers from PubMed[8] database. PubMed is a free database accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. Most of the papers in the PubMed database are indexed based on a standard controlled vocabulary. Medical Subject Headings (MeSH) is a comprehensive controlled vocabulary for the purpose of indexing journal articles and books in the life sciences; it can also serve as a thesaurus that facilitates searching. The 2009 version of MeSH contains a total of 25,186 subject headings, also known as descriptors. The descriptors or subject headings are arranged in a hierarchy. In addition to the descriptor hierarchy, MeSH contains a small number of standard qualifiers (also known as subheadings), which can be added to descriptors to narrow down the topic. For example the Breast Neoplasms (i.e. Breast Cancer) is a heading in MeSH tree which has 50 subheading. Each subheading indicates an specific aspect of the subject described by heading. Table 2 indicates some of the subheadings (i.e aspects) of the breast cancer topic. In order to automatically generate our test collection, we crawled a subset of papers form the PubMed database which are related to at least two subheadings of the breast cancer topic. We selected 231 multi-aspect papers related to the breast cancer as the set of given papers for review process. The crawled papers are related to 3.6 aspects (i.e. subheadings) in average. In order to find the candidate reviewers for a given topic (i.e. subheading), we selected authors in PubMed database which have at least

[8] http://www.ncbi.nlm.nih.gov/pubmed.

Please cite this article in press as: Neshati, M., et al. Expert group formation using facility location analysis. *Information Processing and Management* (2013), http://dx.doi.org/10.1016/j.ipm.2013.10.001

Table 2
Some of subheadings of the breast cancer topic in MeSH dataset.

Analysis	Secretion	Pathology	Genetics
Blood	Surgery	Physiology	History
Blood supply	Therapy	Physiopathology	Immunology
Chemistry	Transmission	Psychology	Metabolism
Classification	Ultrasonography	Radiography	Microbiology
Complications	Ultrastructure	Drug therapy	Mortality

10 papers in that subheading as the potential reviewers. In total, we crawled the profile (i.e. title and abstract of the published papers) as well as the related aspects of 98 reviewers. In average, in our crawled dataset, each reviewer is related to 4.2 subheadings of the breast cancer topic. This automatically generated dataset is publicly available at http://isl.ce.sharif.edu/pubmed-dataset/.

5.2. Evaluation measures

While the multi-aspect team formation problem can be cast as a retrieval problem, the traditional relevance-based precision and recall measures cannot be directly applied to measure the matching performance, because they are unable to reflect the coverage and confidence measures in the assigned groups. To measure the performance of our multi-aspect matching algorithms, we used the *Coverage* and *Average Confidence* measures defined in Karimzadehgan et al. (2008).

Coverage score measures the number of different distinct topic aspects of a paper that are covered by the k assigned reviewers to that specific paper. Consider paper p_j with n_j topic aspects $A_1, \ldots, A_{n_j}$ and let n_r denote the number of distinct topic aspects that the k assigned reviewers can cover. Coverage can be defined as the percentage of topic aspects covered by these reviewers:

$$\text{Coverage}(p_j) = \frac{n_r}{n_j}$$

As mentioned before, in addition to maximizing the coverage of topic aspects, the assignments that maximize the confidence of the assigned reviewers are preferable. Specifically, in the same level of coverage, we would prefer an assignment where each reviewer of a paper is able to cover as many aspects as possible. The confidence of reviewer r_i for paper p_j can be defined as the fraction of aspects of paper p_j that reviewer r_i can cover. For a group of reviewers with k members, the average confidence of these assigned reviewers can be used as a quality measure of the assigned group. Using the notations introduced earlier, the *Average Confidence*measure is defined as follow:

$$\text{Average Confidence}(p_j) = \frac{\sum_{i=1}^{k} \frac{n_{r_i}}{n_j}}{k}$$

where k is the number of assigned reviewers and n_{r_i} indicates the number of topics/aspects of the paper that reviewer r_i can cover.

5.3. Baseline methods

In the paper, the FLA framework of multi-aspect expert matching is compared with the methods proposed in Karimzadehgan and Zhai (2012, 2008). As another baseline, we compare the result of FLA for the first problem of expert matching (i.e. *Implicit Aspects – Unconstraint Matching*) with a standard retrieval model (i.e. language modeling with Dirichlet smoothing (Zhai & Lafferty, 2001)). In this method, for each query/paper, expertise documents of reviewers are ranked according to language model score and then the top k reviewers are selected as the assigned expert group for that specific paper. In comparison of the proposed models, statistically significant improvements are measured using a Wilcoxon Signed-Rank test (Wilcoxon, 1945) at the level of 0.05.

6. Experimental results

In this section, an extensive set of experiments were conducted to address the following questions:

- In the first problem of expert matching (i.e. *Implicit Aspects – Unconstraint Matching*), how good is the performance of the UFLA approach? In Section 6.1, we compare the performance of UFLA with various baseline algorithms proposed in Karimzadehgan et al. (2008). In particular, we compare two greedy approaches for expert matching namely, the *Next Best* (Karimzadehgan et al., 2008) search and the UFLA (introduced in Section 3.3) strategies.
- What is the impact of the building and communication cost on the coverage and confidence measures in our FLA framework? How good is the performance of the FLA framework for different values of parameter λ?

Please cite this article in press as: Neshati, M., et al. Expert group formation using facility location analysis. *Information Processing and Management* (2013), http://dx.doi.org/10.1016/j.ipm.2013.10.001

- How good is the performance of the proposed framework for constraint expert matching problems in comparison with the heuristic methods proposed in Karimzadehgan and Zhai (2012)? In Section 6.2, we compare the performance of CFLA with the integer linear programming method proposed in Karimzadehgan and Zhai (2012).
- How scalable is the performance of the proposed framework for real scenarios? In Section 6.4, we investigate the scalability of the proposed framework.

6.1. Implicit Aspects – Unconstraint Matching

In this section, we compare the UFLA method with (1) the language model retrieval model (LM), (2) Redundancy Removal (RR) (Karimzadehgan et al., 2008), and (3) the *Next Best* greedy method which are described in related work section. In order to evaluate the effectiveness of these methods, we compare all well-tuned methods. In these experiments, all papers in each test collection (i.e. 72 papers in SIGIR dataset and 231 papers in PubMed dataset) are assigned to the all available reviewers in corresponding dataset such that each paper gets exactly 3 reviewers. Table 3 indicates the coverage and the average confidence scores and the percentage of improvement for the UFLA method.

According to Table 3, the performance of author topic modeling methods (i.e. *Next Best* and UFLA) are better than other baseline methods (except for Average confidence of *Next Best* on PubMed dataset) and also the coverage and especially the average confidence of the UFLA method is better than the *Next Best* search method.

Since the performance of the *Next Best* and the UFLA methods are dependent on the quality of the topic learning model, in order to fairly compare these methods, we use another model to learn these topics. In this model, we use the skills/aspects associated with each reviewer from the golden set (i.e. in Eq. (3), parameters $p(\tau_a|\theta_i)$ are known) and just learn the word distributions for each topic (i.e. the only unknown parameters in Eq. (3) are $p(w|\tau_a)$). Using the estimated word distribution parameters, we infer the topic vector for each paper and run the *Next Best* and the UFLA algorithms in the same manner using the new topic vectors. Table 4 indicates the coverage and average confidence for this experiment.

According to Table 4, by improving the quality of learned topics, the coverage and average confidence of both FLA and *Next Best* methods is improved. However, the performance of UFLA is again better than the *Next Best*greedy matching. The result of above experiments (i.e. Tables 3 and 4) indicate that independent of the method used for topic learning, the performance of UFLA matching is always better than the *Next Best* method for expert matching. As mentioned before, in contrast with the *Next Best*method, the UFLA measures the quality of the whole assigned group at each iteration and accordingly can improve the performance of matching.

To better understand the behavior of *Next Best* and the UFLA methods, we examine the impact of parameters σ and λ on performance of these methods for both datasets. As mentioned before, parameter σ in *Next Best* method models the skill redundancy in the assigned groups and parameter λ makes the balance between the building cost and commutation cost in the UFLA framework. Fig. 3 indicates the sensitivity of the coverage and the average confidence measures on these parameters for the UFLA and *Next Best* algorithms.

According to this figure, while the coverage score of the *Next Best*method fluctuates for different values of σ, the coverage of UFLA method is stable for $\lambda > 0$. This experiment also shows that eliminating the building cost from the objective function (i.e. $\lambda = 0$ in Eq. (4)) significantly reduces the performance of UFLA matching algorithm. Table 5 indicates some instances of reviewer selection groups using UFLA method on SIGIR dataset.

6.2. Explicit Aspects – Constraint Matching

In this section, we compare our CFLA method with the baseline algorithms for constraint expert matching. The first baseline algorithm is the greedy approach proposed in Karimzadehgan and Zhai (2012). In this method, first, the papers are decreasingly sorted according to the number of their subtopics, i.e., the paper with the largest number of subtopics is ranked first. Then start off with this ranked list of the papers. At each assignment stage, the best reviewer that can cover most subtopics of the paper is assigned. In addition, the review quota and paper quota are checked, i.e., the number of papers assigned to each reviewer and the number of reviewers assigned to each paper. If the review quota is reached, that reviewer is

Table 3
Comparison of the UFLA method with baseline algorithms for the first problem of expert matching – statistically significant improvement is shown by $*$ symbol.

DataSet	Measure	Coverage		Average confidence	
	Method	Score	UFLA improvement (%)	Score	UFLA improvement (%)
SIGIR	Baseline-LM	0.750	20.0*	0.420	34.3*
	Baseline-RR	0.770	16.9*	0.450	25.3*
	Baseline-*Next Best*	0.869	3.6*	0.501	12.6*
	UFLA	**0.900**	–	**0.564**	–
PubMed	Baseline-LM	0.543	52.1*	0.356	91.9*
	Baseline-RR	0.671	23.1*	0.613	11.4*
	Baseline-*Next Best*	0.798	3.5*	0.424	61.1*
	UFLA	**0.826**	–	**0.683**	–

Please cite this article in press as: Neshati, M., et al. Expert group formation using facility location analysis. *Information Processing and Management* (2013), http://dx.doi.org/10.1016/j.ipm.2013.10.001

Table 4
Comparison of the UFLA and *Next Best* methods – for the improved topic learning model – statistically significant improvement is shown by $*$ symbol.

DataSet	Measure	Coverage		Average confidence	
	Method	Score	UFLA improvement (%)	Score	UFLA improvement (%)
SIGIR	Baseline-*Next Best*	0.890	7.1*	0.660	3.0*
	UFLA	**0.953**	–	**0.680**	–
PubMed	Baseline-*Next Best*	0.832	6.9*	0.579	28.5*
	UFLA	**0.889**	–	**0.744**	–

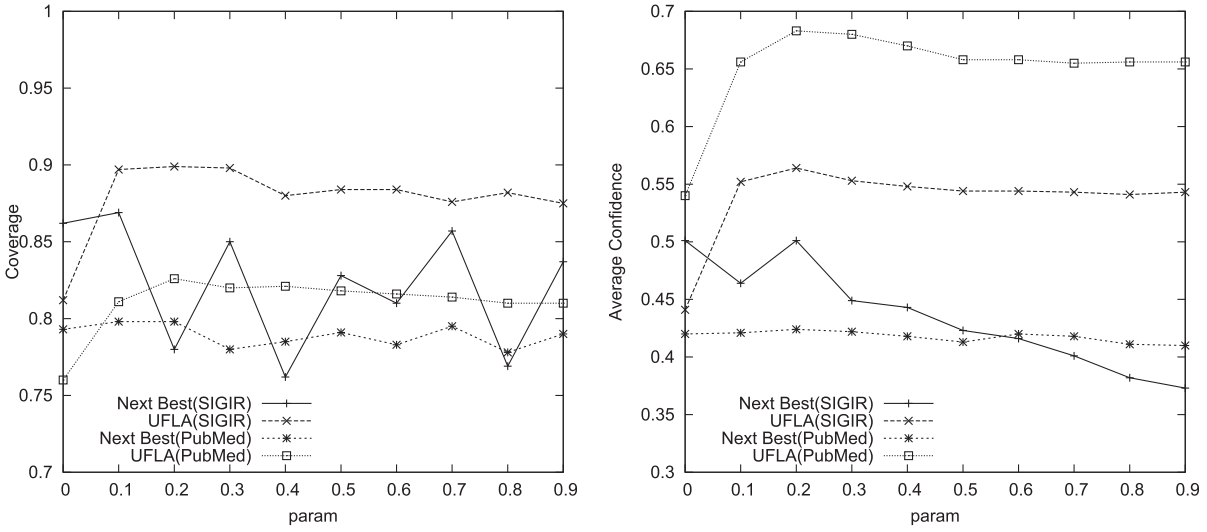

Fig. 3. The sensitivity of UFLA and *Next Best* methods to the parameters λ and σ. Note that parameter σ can be defined in interval [0,0.9].

Table 5
Some UFLA assignments for SIGIR dataset.

Paper	Selected reviewers	Coverage	AVG confidence
Personalized query expansion for the web	Carlos Castillo Marc Najork Udo Hahn	0.33	0.33
Using query contexts in information retrieval	Alistair Moffat Diane Kelly Chengxiang Zhai	0.50	0.50
Towards task-based personal-information management evaluations	Ian Soboroff Saikat Mukherjee Guizhen Yang	1.00	0.66
Utility-based information distillation over temporally sequenced documents	Diane Kelly Vo Ngoc Anh Wessel Kraaij	1.00	0.66
Efficient bayesian hierarchical user modeling for recommendation system	Oren Kurland Luo Si Qing Li	0.66	0.33

removed from our reviewer pool; the same is done when the paper quota is satisfied. This process is repeated until reviewers are assigned to all the papers. The second baseline algorithm is the integer linear programming method introduced in Algorithm 3 to match papers with reviewers. This method tries to globally maximize the number of covered aspects of the assigned groups.

Before comparison with the baseline algorithms, we examine the effect of the building and the communication cost on coverage and average confidence measures. Fig. 4 indicates the sensitivity of coverage for different size of the program committee (i.e. number of available reviewers) on SIGIR dataset. In these experiments, each paper is assigned to 3 reviewers and the capacity of each reviewer is equal to 5. For each program committee size, we randomly select the specified number of

Please cite this article in press as: Neshati, M., et al. Expert group formation using facility location analysis. *Information Processing and Management* (2013), http://dx.doi.org/10.1016/j.ipm.2013.10.001

experts from all available experts (i.e. 189 experts) and repeat each experiment 10 times and report the coverage and the average of confidence in Fig. 4.

According to Fig. 4, increasing the size of committee (i.e. available experts) improves the coverage and average confidence scores and on the other hand, increasing parameter λ (i.e. increasing the building cost and decreasing the communication cost in objective function of the CFLA) decreases the coverage score of the assigned groups. This experiment shows that by emphasizing communication cost in the objective function of the CFLA, the coverage score of assigned groups can be increased. According to this figure, while the average confidence is stable for $\lambda > 0$, the maximum and minimum values for average confidence is reported at $\lambda = 1$ and $\lambda = 0$, respectively. Specifically, for $\lambda = 0$, the value of average confidence score is reduced substantially, which means that by ignoring the building cost, the average confidence score reduces. This experiment confirms the result of Theorems 1 and 2 mentioned in Section 3.4. This experiment also shows that the coverage and the average confidence scores are contradicting constraints and parameter λ can be used to make a trade off between these constraints. We found the same behavior on PubMed dataset and therefore we did not report the sensitivity results on PubMed dataset. In experiments of this section, we use $\lambda = 0.5$ to make a balance between the coverage and the average confidence measures.

In out first experiment, we compare the proposed CFLA method with the integer linear programming method (Karimzadehgan & Zhai, 2012) and the greedy matching algorithms (Karimzadehgan & Zhai, 2012) for different program committee sizes (i.e. number of available reviewers). Each experiment is repeated 30 times and the average of scores are reported. In this experiment, for the SIGIR dataset each paper is assigned to 3 reviewers and the capacity of each reviewer is 5 and for PubMed dataset each paper is assigned to 3 reviewers while the capacity of each reviewer is 10 (because of the limited number of reviewers in PubMed dataset). Fig. 5 indicates the coverage score of CFLA (i.e. proposed model), ILP and the greedy approach for different sizes of the program committee for both datasets.

According to Fig. 5, by increasing the size of the program committee the coverage score is increasing for all methods. In addition, for all committee program sizes, the coverage score of the CFLA method is always better than the greedy and ILP methods. Specifically, the CFLA method can significantly improve the coverage score for small program committee sizes (i.e. less than 120 available reviewers for SIGIR dataset and less than 90 reviewers for PubMed dataset).

Table 6 indicates the average confidence score of the CFLA, ILP and the greedy approach for this experiment. According to this table, for all matching methods, by increasing the size of committee (i.e. increasing the size of available experts), the average confidence score is improved. The performance of ILP and CFLA are almost the same and both are better than the greedy method. According to this experiment, the CFLA method can make expert groups with significantly better coverage score in comparison with the ILP method without reduction of the average confidence score. Specifically, it can improve the coverage score up to 8.90% for small committee sizes in SIGIR dataset and 6.60% for PubMed dataset, while the variation of the average confidence score is negligible.

In the next experiment, we fix the number of reviewers to 30 for SIGIR dataset and 45 for PubMed dataset, and vary the number of papers each reviewer can review while each paper is assigned to 3 reviewers. In this experiment, minimum capacity of SIGIR dataset reviewers equals to $\lceil \frac{73*3}{30} \rceil = 8$ and minimum capacity of PubMed dataset reviewers equals to $\lceil \frac{231*3}{45} \rceil = 16$. In order to avoid bias, we repeat the sampling process (reviewers selection) for 10 times and get the average. The coverage scores are shown in Fig. 6.

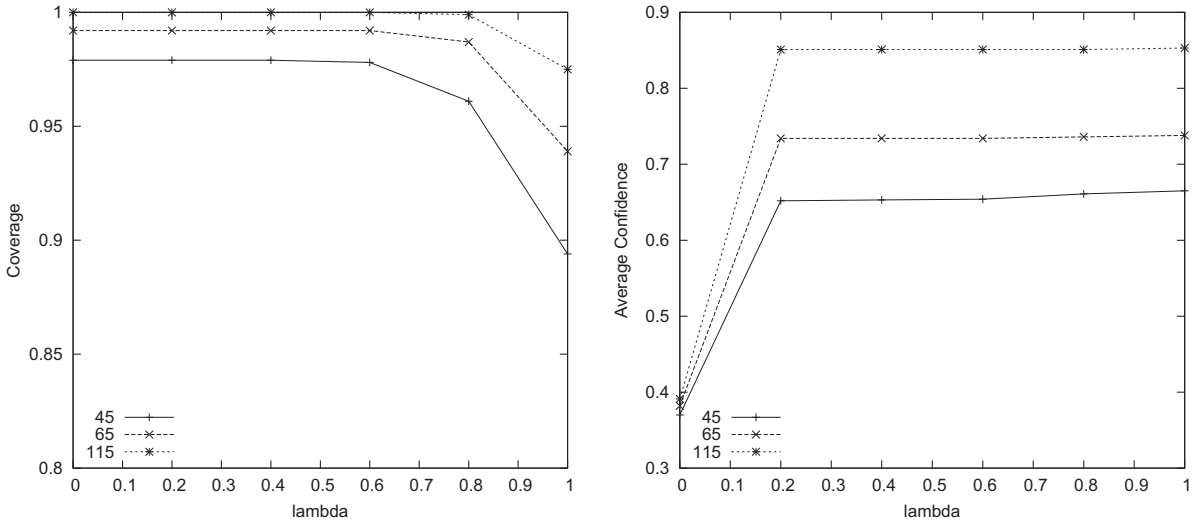

Fig. 4. The sensitivity of CFLA method to the parameter λ for the SIGIR dataset. Each data series indicates scores for a program committee size.

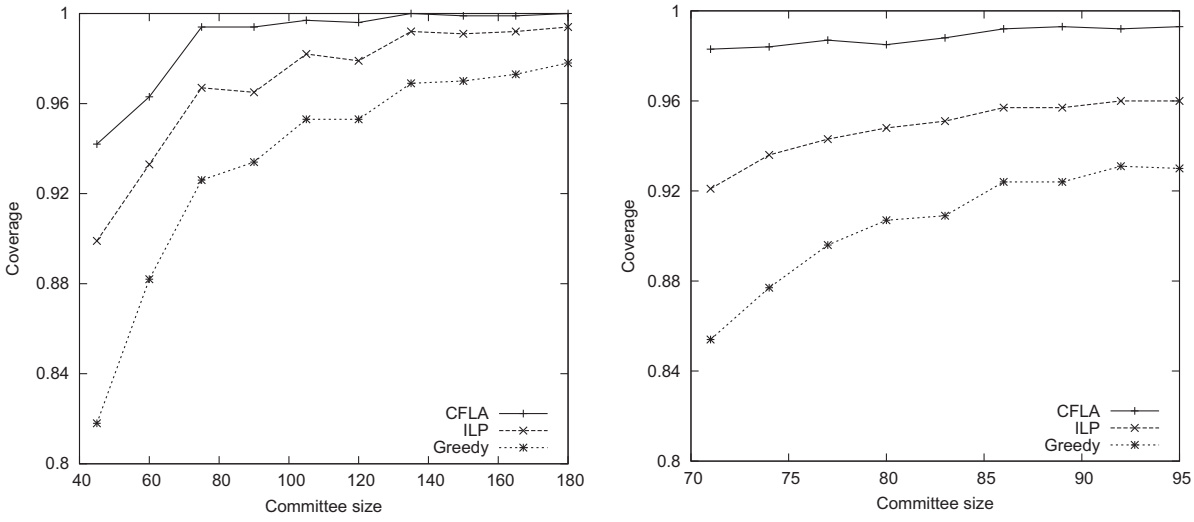

Fig. 5. Coverage score of Greedy, ILP and CFLA for different committee program sizes.

Table 6
Comparison of average confidence for greedy, ILP and CFLA for different committee sizes. Differences between ILP and CFLA are not significant.

Dataset/committee size		45	55	65	105	185
SIGIR	Greedy	0.550	0.634	0.665	0.798	0.882
	ILP	**0.651**	0.708	0.724	0.831	0.914
	CFLA	0.647	**0.710**	**0.729**	**0.837**	**0.916**
		71	74	77	92	95
PubMed	Greedy	0.740	0.768	0.789	0.852	0.856
	ILP	0.800	0.826	0.847	**0.905**	0.907
	CFLA	**0.809**	**0.833**	**0.851**	0.904	**0.909**

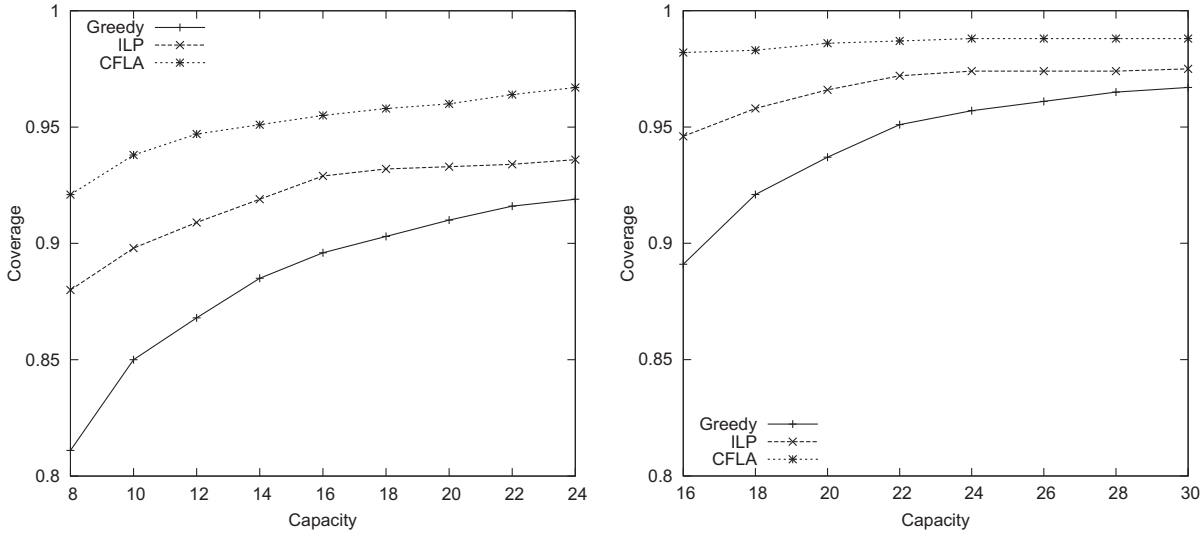

Fig. 6. Coverage score of Greedy, ILP and CFLA for different capacities of reviewers.

Please cite this article in press as: Neshati, M., et al. Expert group formation using facility location analysis. *Information Processing and Management* (2013), http://dx.doi.org/10.1016/j.ipm.2013.10.001

As we increase the number of papers that each reviewer can review, we are also increasing the resources, and as a result, the performance of all algorithms becomes better. Also, comparing the CFLA method with the ILP and greedy approach, the performance of the CFLA method is significantly better than the greedy and ILP methods for all values of capacity of reviewers.

Table 7 indicates the average confidence scores for this experiment. The average confidence score of the CFLA and the ILP methods are almost the same but both are better than the greedy algorithm. This experiment also indicates that the CFLA algorithm can improve the coverage measure while retain the average confidence in the same level. The CFLA can better distribute papers among available reviewers.

In the last experiment, we compare the performance of CFLA and ILP when very limited recourse (i.e. reviewers) are available. In this experiment, for SIGIR dataset, the maximum number of reviewers is 10 for 73 papers and for PubMed dataset, the maximum number of reviewers is 22 for 231 papers. Again we randomly select reviewers and we repeat the sampling process for 30 times and get the average. Each paper gets three reviewers and the number of papers that each reviewer can get is calculated according to the number of reviewers that we have. For example, for SIGIR dataset, if we have five reviewers, each should get 44 papers. Fig. 7 indicates the coverage measure of CFLA and ILP methods.

The result of average confidence is also reported in Table 8. These experiments shows that for very limited recourses the quality of matching for CFLA is always better than the ILP in terms of both the coverage and average confidence.

6.3. Implicit Aspects – Constraint Matching

In this section, we examine the quality of matching experts for the third problem of expert matching. In this case, the aspects/skills of papers and reviewers are implicitly given in abstract and expertise documents. Similar to previous experiments, we use $\lambda = 0.5$ to make a balance between building and communication cost.

While the CFLA method can be directly applied to the probabilistic assignments of subtopics given by PLSA, intuitively, not all the predictions are reliable, especially the low-probability ones. Thus we experimented with pruning low probability values learned with PLSA (i.e., setting low topic probability elements to zero). The greedy approach is not applicable for this matching problem because the aspects/topics of papers and reviewers are not presanctified. So, we use the ILP method introduced in Karimzadehgan and Zhai (2012) as our baseline model. Table 9 indicates the result of matching for the best parameters (i.e. cut-off = 3). In this experiment, for SIGIR dataset, the size of the program committee size is 189 and the capacity of each reviewer is 5 and for PubMed dataset, the size of the program committee size is 98 and the capacity of each reviewer is 10.

Fig. 8 indicates the effect of different values for cut-off and sensitivity of algorithms to parameter λ on coverage score. In this figure, each data series indicate a method and a value of cut-off for example, CFLA (5) indicate expert matching using CFLA method and setting the cut-off value equals to 5. i.e. only top 5 topics are used in topic vector of reviewers and papers.

According to this figure, setting the cut-off equals 1 reduces the performance of both algorithms for SIGIR and PubMed datasets. Also, setting the cut-off more than five increases the noise and as a result the coverage reduces. The best value for cut-off is near to the average number of required skills for each paper in the golden measure. Although, the coverage of the CFLA method fluctuates for different values of λ, for all values the coverage is significantly better than the ILP model. To sum up, according to this experiment, the performance of the CFLA method is significantly better than the ILP method in both coverage and average confidence measures for Implicit Aspect – Constraint Matching problem.

6.4. Scalability of FLA framework

Finally, we study the scalability of our CFLA framework. In the experiments reported so far, we have only evaluated our algorithms using small test collections. In reality, the number of submitted papers to a conference is much larger than SIGIR and PubMed test collections. So, we opt to study the scalability by generating synthetic data to simulate the scenarios of conference review assignments with larger number of submissions.

Specifically, we first select the number of topic areas be 25 (the same number of topics as in the gold standard data), we then get the average number of expertise areas for reviewers from the gold standard data, which is 5, and the average number of topics in query papers (73 papers), which is three. These are used to specify how many topics out of 25 should be one

Table 7
Comparison of average confidence for greedy, ILP and CFLA for different capacities. Differences between ILP and CFLA are not significant.

Dataset/capacity		8	12	16	20	24
SIGIR	Greedy	0.526	0.600	0.629	0.647	0.658
	ILP	0.614	0.640	0.654	0.664	0.668
	CFLA	**0.615**	**0.640**	**0.655**	**0.664**	**0.668**
		16	18	20	26	30
PubMed	Greedy	0.763	0.806	0.833	0.874	0.885
	ILP	0.829	0.861	0.884	0.911	**0.918**
	CFLA	**0.834**	**0.866**	**0.887**	**0.912**	0.917

Please cite this article in press as: Neshati, M., et al. Expert group formation using facility location analysis. *Information Processing and Management* (2013), http://dx.doi.org/10.1016/j.ipm.2013.10.001

Table 8
Comparison of average confidence for ILP and CFLA for very limited resources – statistically significant improvement is shown by $*$ symbol.

Dataset/capacity	4	6	8	10	
SIGIR	ILP	0.243	0.274	0.289	0.318
	CFLA	**0.270***	**0.307***	**0.355***	**0.441***
		13	14	16	17
PubMed	ILP	0.748	0.711	0.754	0.740
	CFLA	**0.754***	**0.726***	**0.759**	**0.745**

Table 9
Comparison of ILP and CFLA for constraint implicit aspect matching problem – statistically significant improvement is shown by $*$ symbol.

Dataset	Method/measure	Coverage		Average confidence	
		Score	CFLA imp. (%)	Score	CFLA imp. (%)
SIGIR	ILP	0.715	15.80	0.347	25.60
	CFLA	**0.828***	–	**0.436***	–
PubMed	ILP	0.659	12.1	0.383	37.9
	CFLA	**0.739***	–	**0.528***	–

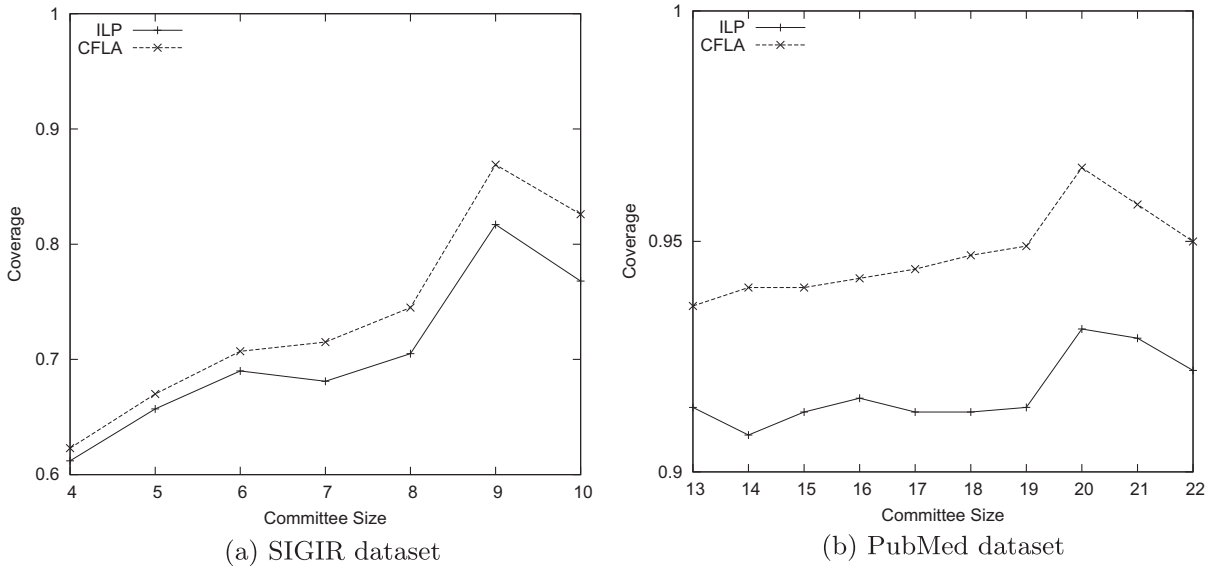

Fig. 7. Coverage score of ILP and CFLA for very limited resources.

for each reviewer and paper. So, we randomly choose five and three topics out of 25 to be the ones assigned to each reviewer and paper, respectively. In our experiments, we also vary the number of topics, (i.e., 50 and 100). When we have 50 topics, we randomly select 10 and 6 out of 50 topics to be one for each reviewer and paper, respectively. Since the number of topics is doubled, i.e., from 25 to 50, the number of expertise topics for reviewers and papers will be doubled too, (i.e., from 5 to 10 for reviewers and 3 to 6 for papers). The same is done when we have 100 topics. The number of reviewers to be assigned to each paper is three and each reviewer can get up to five papers. Then, we vary the number of papers and reviewers. Consider that n is the number of papers, the minimum number of reviewers which are needed is $n * \frac{3}{5}$ (three is the number of reviewers assigned to each paper and five is the capacity of each reviewer).

Fig. 9a shows the runtime of the CFLA algorithm as the number of papers increases. The algorithm run on Intel CPU, 2.8 GHZ, with 8 GB memory. The runtime of the algorithm increases when we have a large number of papers and topics as expected. Given the computational complexity of the CFLA algorithm, this observation is intuitively expected; indeed, as we increase the number of papers, more time is needed to find the optimal assignment, because the number of variables is increased, as a result, the algorithm behaves exponentially in the number of variables.

The CPLEX optimizer for integer programming returns an indication of the quality of the solution that has been found (i.e. the gap) for each iteration of algorithm. Fig. 9b indicates the relation between solving time and the gap of the solution for

Please cite this article in press as: Neshati, M., et al. Expert group formation using facility location analysis. *Information Processing and Management* (2013), http://dx.doi.org/10.1016/j.ipm.2013.10.001

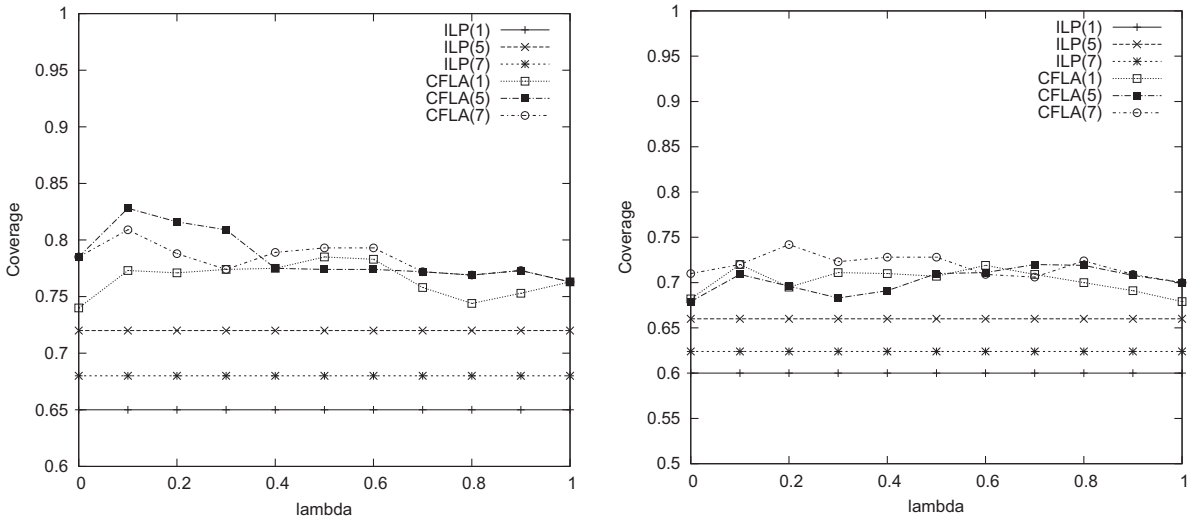

Fig. 8. Coverage score of ILP and CFLA for constraint implicit aspect matching problem. The number in parenthesis indicate the cut-off used for the matching algorithms.

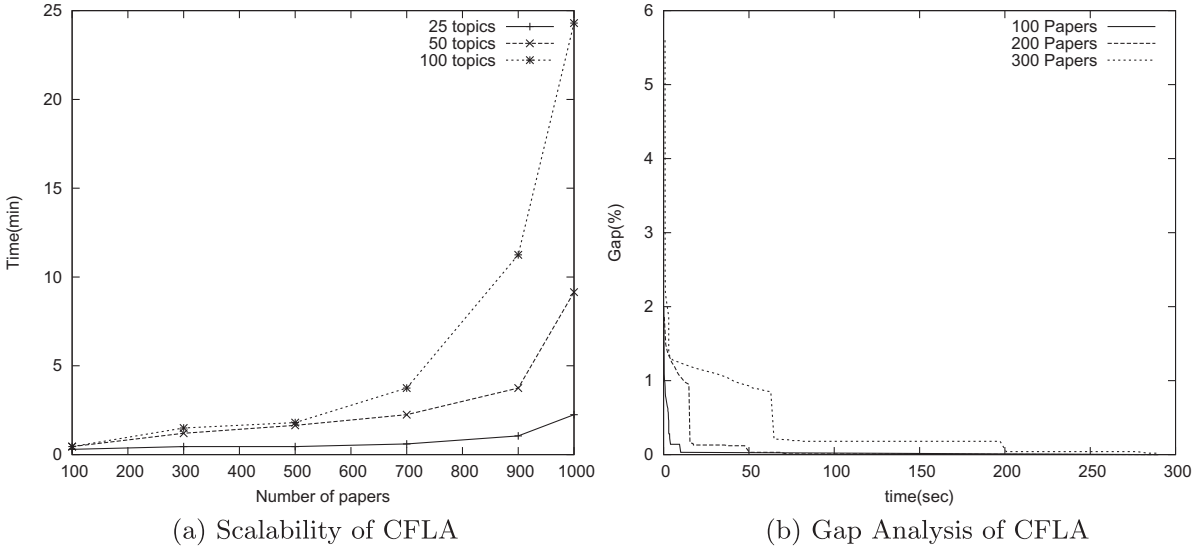

Fig. 9. (a) Runtime of the CFLA algorithm for different size of problems (b) Relation between gap (%) and solving time (sec) for the CFLA algorithms.

different problem sizes. This figure shows that firstly, by increasing the size of problem more time is needed to find the optimal solution (i.e. gap = 0%) and secondly, the longer the program runs the smaller the gap becomes.

Since in most real conferences, the keyword list used for authors and reviewers usually does not have more than 50 keywords, the CFLA algorithm is sufficiently efficient for use in real conferences with large number of submissions, e.g., 1000. Since assignment of reviewers in a conference management system is usually not required to be run in real time, spending more time to get an optimal solution is worthwhile. Thus we can expect CFLA to be useful in a real application.

7. Conclusion

In real scenarios, several and sometimes diverse skills are needed to perform a project successfully and completely. In this paper, we considered the problem of expert group formation (i.e. expert matching) to optimally assign a set of available experts to a project. Three types of the group formation problems are considered in this paper. In the first problem, we assumed that the required skills of a project and also the relevant skills of experts are implicitly expressed by the text documents. The

second problem is concerned with the assignment of experts to multiple projects such that each expert should be involved in a limited number of projects and the third problem is the combination of the first and the second problems. The assigned group of experts to each project should be able to cover all required skills of that project and preferably, each member of an assigned group should also be able to cover all the these aspects. A unified framework based on the facility location analysis is proposed in this paper to address these problems. As a case study, we consider the problem of multi-aspect review assignment which is a common task in conference and journal organizations. The optimality of the proposed method for reviewer matching problem is investigated and several experiments are conducted on a real dataset as well as an automatically generated dataset to compare the performance of the proposed framework with the state-of-the-art methods. Our experiments show that the FLA framework can significantly improve the performance of expert matching in terms of two performance measures.

Acknowledgment

The authors would like to thank Maryam Karimzadegan for providing us the test collection and the anonymous reviewers for their helpful comments that help to improve the paper.

References

AlSumait, L., Barbará, D., & Domeniconi, C. (2008). On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, ICDM '08* (pp. 3–12). Washington, DC, USA: IEEE Computer Society (ISBN 978-0-7695-3502-9).

Balog, K., Soboroff, I., Thomas, P., Craswell, N., de Vries, A. P., & Bailey, P. (2009). Overview of the TREC 2008 enterprise track. In *Proceedings of the seventeenth Text Retrieval Conference (TREC 2008)*.

Balog, K., Azzopardi, L., & De Rijke, M. (2009). A language modeling framework for expert finding. *Information Processing and Management, 45*(1), 1–19.

Baykasoglu, A., Dereli, T., & Das, S. (2007). Project team selection using fuzzy optimization approach. *Cybernetics and Systems, 38*(2), 155–185.

Berg, M. D., Cheong, O., Kreveld, M. V., & Overmars, M. (2008). *Computational geometry: Algorithms and applications* (3rd ed.). Santa Clara, CA, USA: Springer-Verlag TELOS (ISBN 3540779736, 9783540779735).

Charikar, M., & Guha, S. (1999). Improved combinatorial algorithms for the facility location and k-median problems. In *Proceedings of the 40th annual symposium on Foundations of Computer Science, FOCS '99* (pp. 378–388). Washington, DC, USA: IEEE Computer Society.

Chudak, F. A., & Williamson, D. P. (2005). Improved approximation algorithms for capacitated facility location problems. *Mathematical programming, 102*(2), 207–222 (ISSN 0025-5610).

Craswell, N., Vries, A. P. D., & Soboroff, I. (2005). Overview of the TREC 2005 enterprise track. In *Proceedings of the fourteenth Text Retrieval Conference (TREC 2005)*.

Daud, A. (2012). Using time topic modeling for semantics-based dynamic research interest finding. *Knowledge-Based Systems, 26*, 154–163 (ISSN 0950-7051).

Deng, H., Han, J., Lyu, M. R., & King, I. (2012). *Modeling and exploiting heterogeneous bibliographic networks for expertise ranking. Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries, JCDL '12*. New York, NY, USA: ACM (ISBN 978-1-4503-1154-0).

Deng, H., King, I., & Lyu, M. R. (2012). Enhanced models for expertise retrieval using community-aware strategies. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 42*(1), 93–106 (ISSN 1083-4419).

Gabor, A., & Van, J. O. (2005). Approximation algorithms for facility location problems with discrete subadditive cost functions, Ph.D. thesis. Department of Applied Mathematics, University of Twente.

Gonzalez, T. F. (2007). *Handbook of approximation algorithms and metaheuristics (Chapman & Hall/Crc Computer & Information Science Series)*. Chapman & Hall/CRC.

Hashemi, S. H., Neshati, M., & Beigy, H. (2013). Expertise retrieval in bibliographic network: A topic dominance learning approach. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management (CIKM 2013)* (in press).

Hettich, S., & Pazzani, M. J. (2006). Mining for proposal reviewers: Lessons learned at the national science foundation. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '06* (pp. 862–871). New York, NY, USA: ACM (ISBN 1-59593-339-5).

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '99* (pp. 50–57). New York, NY, USA: ACM (ISBN 1-58113-096-1).

Kargar, M., & An, A. (2011). Discovering top-k teams of experts with/without a leader in social networks. In *Proceedings of the 20th ACM international conference on Information and knowledge management, CIKM '11* (pp. 985–994). New York, NY, USA: ACM (ISBN 978-1-4503-0717-8).

Karimzadehgan, M., & Zhai, C. (2012). Integer linear programming for constrained multi-aspect committee review assignment. *Information Processing and Management, 48*(4), 725–740 (ISSN 0306-4573).

Karimzadehgan, M., Zhai, C., & Belford, G. (2008). Multi-aspect expertise matching for review assignment. In *Proceedings of the 17th ACM conference on information and knowledge management, CIKM '08* (pp. 1113–1122). New York, NY, USA: ACM (ISBN 978-1-59593-991-3).

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*(1), 79–86.

Lappas, T., Liu, K., & Terzi, E. (2009). Finding a team of experts in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '09* (pp. 467–476). New York, NY, USA: ACM (ISBN 978-1-60558-495-9).

Macdonald, C. (2009). The Voting Model for People Search, Ph.D. thesis, Department of Computing Science. Univeristy of Glasgow.

Mimno, D., & McCallum, A. (2007). Expertise modeling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '07* (pp. 500–509). New York, NY, USA: ACM (ISBN 978-1-59593-609-7).

Mitchell, J. E. (2002). Branch-and-cut algorithms for combinatorial optimization problems. In *Handbook of applied optimization* (pp. 65–77). Oxford University Press.

Neshati, M., Asgari, E., Hiemstra, D., & Beigy, H. (2013). A joint classification method to integrate scientific and social networks. In *Proceedings of the IR research, 35th European conference on Advances in information retrieval, ECIR'2013* (pp. 122–133). Berlin, Heidelberg: Springer-Verlag.

Neshati, M., Beigy, H., & Hiemstra, D. (2012). Multi-aspect group formation using facility location analysis. In *Proceedings of the seventeenth Australasian Document Computing Symposium, ADCS '12* (pp. 62–71). New York, NY, USA: ACM (ISBN 978-1-4503-1411-4).

Pal, A., & Konstan, J. A. (2010). Expert identification in community question answering: Exploring question selection bias. In *Proceedings of the 19th ACM international conference on Information and knowledge management, CIKM '10* (pp. 1505–1508). New York, NY, USA: ACM (ISBN 978-1-4503-0099-5).

Qin, T., Liu, T.-Y., Zhang, X.-D., Chen, Z., & Ma, W.-Y. (2005). A study of relevance propagation for web search. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '05* (pp. 408–415). New York, NY, USA: ACM (ISBN 1-59593-034-5).

Serdyukov, P., & Hiemstra, D. (2008). Modeling documents as mixtures of persons for expert finding. In *Proceedings of the IR research, 30th European Conference on Advances in Information Retrieval, ECIR'08* (pp. 309–320). Berlin, Heidelberg: Springer-Verlag (ISBN 3-540-78645-7, 978-3-540-78645-0).

Please cite this article in press as: Neshati, M., et al. Expert group formation using facility location analysis. *Information Processing and Management* (2013), http://dx.doi.org/10.1016/j.ipm.2013.10.001

Smirnova, E. (2011). A model for expert finding in social networks. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, SIGIR '11* (pp. 1191–1192). New York, NY, USA: ACM (ISBN 978-1-4503-0757-4).

Soboroff, I., de Vries, A. P., & Craswell, N. (2006). Overview of the TREC 2006 enterprise track. In *Proceedings of the fifteenth Text Retrieval Conference (TREC 2006)*.

Sun, Y.-H., Ma, J., Fan, Z.-P., & Wang, J. (2007). A hybrid knowledge and model approach for reviewer assignment. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences, HICSS '07* (pp. 47–57). Washington, DC, USA: IEEE Computer Society (ISBN 0-7695-2755-8).

Tang, W., Tang, J., Lei, T., Tan, C., Gao, B., & Li, T. (2012). On optimization of expertise matching with various constraints. *Neurocomputing, 76*(1), 71–83 (ISSN 0925-2312).

Taylor, C. J. (2008). On the optimal assignment of conference papers to reviewers. Tech. Rep. MS-CIS-08-30. University of Pennsylvania.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1*, 80–83.

Wi, H., Oh, S., Mun, J., & Jung, M. (2009). A team formation model based on knowledge and collaboration. *Expert Systems with Applications, 36*(5), 9121–9134.

Zhai, C., & Lafferty, J. (2001). A study of smoothing methods for language models applied to Ad Hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '01* (pp. 334–342). New York, NY, USA: ACM (ISBN 1-58113-331-6).

Please cite this article in press as: Neshati, M., et al. Expert group formation using facility location analysis. *Information Processing and Management* (2013), http://dx.doi.org/10.1016/j.ipm.2013.10.001