Multi-Aspect Group Formation using Facility Location Analysis

Mahmood Neshati, Hamid Beigy
Computer Engineering Department
Sharif University of Technology
{neshati, beigy}@ce.sharif.edu

Djoerd Hiemstra
Computer Science Department
University of Twente
d.hiemstra@utwente.nl

ABSTRACT
In this paper, we propose an optimization framework to retrieve an optimal group of experts to perform a given multi-aspect task/project. Each task needs a diverse set of skills and the group of assigned experts should be able to collectively cover all required aspects of the task. We consider three types of multi-aspect team formation problems and propose a unified framework to solve these problems accurately and efficiently. Our proposed framework is based on Facility Location Analysis (FLA) which is a well known branch of the Operation Research (OR). Our experiments on a real dataset show significant improvement in comparison with the state-of-the art approaches for the team formation problem.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Retrieval models

General Terms
Algorithms, Experimentation

Keywords
Multi aspect Team Formation, Expert Matching, Expert Finding, Facility Location Analysis

1. INTRODUCTION
Expert group formation has recently attracted a lot of attention in Information Retrieval and Data management communities. 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. In real scenarios, several and sometimes diverse skills are needed to perform a project successfully and completely. Generally, these required skills are implicitly expressed in project descriptions. Besides the required skills of a project, in many cases, the relevant skills of experts are also implicitly reflected in their resume. The main challenges of the expert matching problem are:

1) 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.
2) 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.
3) In an ideal expert group formation, all required skills of a project should be covered by the union of the skills of the assigned group members in a complementary manner (i.e. *Coverage condition*).
4) In an ideal expert group formation, besides covering all required skills of a project, it is preferable that each member of the assigned group individually be able to cover as many as possible the required skills of the project (i.e. *Confidence condition*)
5) Forming multiple dependent expert groups, each expert can only be involved in a limited number of projects. In other words, in formation of multiple expert groups with limited recourses (i.e. experts), the *load balancing* condition should be considered.

While all above conditions are natural and practical, the combination of these conditions in a real application can be challenging. Specifically, with a limited number of available experts, simultaneously maximizing the confidence and coverage of the assigned groups is an interesting and also a non-trivial problem.

As a case study for 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 such that all above mentioned criteria are satisfied. Specifically, 1) the required skills for reviewing a paper can be explicitly determined by some keywords or should be inferred from the abstract/body of the paper. 2) The related research areas/skills of each reviewer (i.e. expert) can be expressed explicitly by some keywords or be inferred from his/here previous papers. 3) Ideally the assigned group of reviewers for each paper should be able to cover all required aspects of that paper. 4) It is preferable that each assigned reviewer of a paper be able to cover all aspects/topics of the paper. 5) In a real conference, each member of program committee (i.e. expert) can only be involved in the review process of a limited number of papers.

In this paper, we formalize the expert matching problem within the unified framework of *Facility Location Analysis (FLA)* taken from Operation Research [1], as a way to account and optimize the expert assignment. 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* [2] and *Integer linear programming* [3].

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. topics). According to different conditions of the expert matching problem, we define three

ADCS '12, December 5-6, 2012, Dunedin, New Zealand. Copyright 2012 ACM 978-1-4503-1411-4/12/2012 …$15.00.

problems that all can be solved by the proposed frame work of FLA.

In these problems, given a set of *N* papers and *M* reviewers, each paper should be assigned to a group of exactly *k* members.

1- *Implicit Aspects-Unconstraint Matching (Problem 1):* 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, each paper should be assigned to a group of *k* reviewers such that the skill coverage and confidence of the assigned group be maximal. However, in this problem, there is no limitation on the capacity of reviewers (i.e. arbitrary number of papers can be assigned to a reviewer).

2- *Explicit Aspects-Constraint Matching (Problem 2)*: In this problem, we assume that the set of the required skills of a paper and also the set of the relevant skills of a reviewer are explicitly determined *(*for example by a set of predefined keywords). In this problem, each paper should be assigned to a team of *k* reviewers such that: 1) in an ideal matching, all aspects of all papers should be covered by the skills of the assigned groups. 2) In an ideal matching, each member of the assigned groups for a paper should be able to cover all required skills of that specific paper and finally 3) each reviewer should get only a limited and predefined number of papers to review.

3- *Implicit Aspects- Constraint matching (Problem 3)*: This problem is the combination of the first and the second problems. In this problem, we assume that the notion of aspects/skills of papers and experts 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 maximize the coverage and the confidence of the assigned groups while the load balancing condition is satisfied.

All above mentioned problems are modeled using the unified framework of facility location analysis. Our experiments demonstrate that this framework outperforms the current state of the art algorithms of expert matching.

2. Facility Location Analysis

Facility location analysis is a branch of operations research[1] and computational geometry concerning itself with mathematical modeling and solution of problems concerning optimal placement of facilities in order to minimize transportation costs, avoid placing hazardous materials near housing, outperform competitors' facilities, etc. *Desirable k-facility placement* [4] is a type of facility location problems which concerns with the selection of the *k* optimal locations among *P* 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 two-fold:

1. To minimize the total cost of opening those facilities, and,

2. To minimize the weighted distances from the customers locations to their closest facilities.

Various types of the facility location problems are defined in the literature [1] for different usages. Two main types of these problems are *uncapacitated* and *capacitated* facility location placements that can be useful to model the expertise matching problem. In this paper, we formally model the *unconstraint* (i.e. problem 1) and *constraint* multi-aspect/skill expertise matching (problem 2 and 3) using uncapacitated and capacitated facility location placement respectively. In the following subsections, we introduce these problems as well as their approximate and exact solutions.

2.1 Uncapacitated Facility Location Analysis (UFLA)

In Uncapacitated Facility Location (UFLA) problem, *k* facility locations should be selected among *N* available facility locations; such that while each customer is assigned to its nearest facility, the overall cost of building all facilities is minimal. Considering the general definition of facility location problem, an arbitrary number of customers can be assigned to a facility. In other words, there is no constraint on the assignment of the customers to the facilities. The overall cost of building *k* facilities can be defined as follows:

$$cost(S) = \overbrace{\lambda \sum_{i=1}^{k} cost(f_i)}^{Building\ Cost} + \overbrace{(1-\lambda) \sum_{j=1}^{m} demand(c_j) \min_{f \in S} D(f, c_j)}^{Communication\ Cost} \quad (1)$$

In this equation, $S = \{f_1, \dots f_k\}$ indicates the set of selected facilities, $cost(f_i)$ is the opening cost of facility f_i, $D(f, c_j)$ indicates the distance between the customer c_j and facility f, $demand(c_j)$ indicates the demand of customer c_j and λ is a parameter in [0,1]. According to the above objective function, the total cost of opening *k* facilities equals the sum of the *Building Cost* (i.e. the first summation) and the *Communication Cost* (the second summation). Figure 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. $cost(f_i) = cost(f_j), i, j \in \{1,2, \dots 5\}$), the optimal 3 facility locations (among 5 available candidate locations) and also the assignment of the customers to their nearest location is illustrated in Figure 1. Please note that only 3 facilities can be selected in the optimal solution of Figure 1.

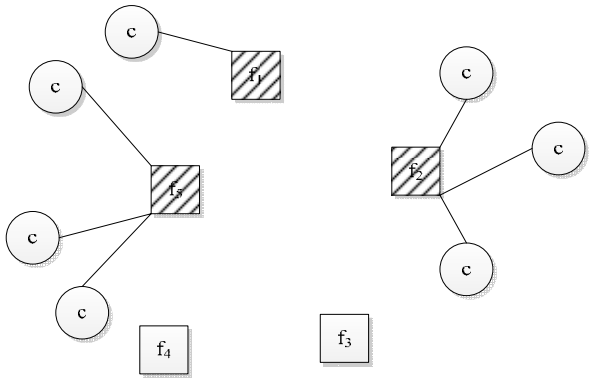

Figure 1. Uncapacitated Facility location problem- Optimal facilities are indicated by the dashed texture

The UFLA is in general NP-hard, which can be proved by reduction, for example, from the set cover problem [1]. Since this problem has an explicit objective function, it is possible approximately optimize it using *Greedy Local Search* (GLS), a.k.a. Hill Climbing, as shown in Algorithm 1. 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 D (D indicates the set of candidate facility locations), until the process

converges. Finally, the *k* facility in S is an approximate solution for the problem.

Algorithm 1. Greedy Local Search for UFLA problem

Input: D (set of candidate facility locations),
k (the cardinality of solution set)
Output: S → top-k facility locations

```
1-  S ← {d1, ... dk}
2-  repeat
3-    for d ∈ S do
4-      for d′ ∈ D\S do
5-        S′ ← (S\{d}) ∪ {d′}
6-        if Cost(S′) < Cost(S) then
7-          S ← S′
8-        end
9-      end
10-   end
11- until S does not change;
```

2.2 Capacitated Facility Location Analysis

Capacitated Facility location analysis (CFLA) is another type of facility location problems that has the same objective function similar to the UFLA. However, in CFLA, 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. Figure 2 illustrates the same customer and facility locations indicated in Figure 1. Assuming equal building cost and equal capacity of 2 for each facility, Figure 2 indicates the optimal top 3 facilities for this CFLA. As indicated in this figure, each facility is responsible to serve to *at most* 2 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 =2 because the number of customers is bigger than 3*2=6.

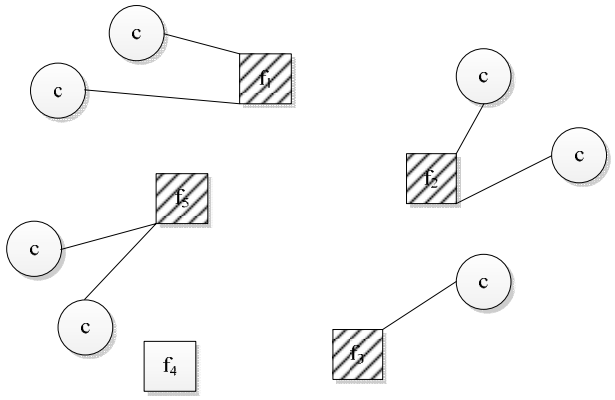

Figure 2. Capacitated Facility location problem- Optimal facilities are indicated by the dashed texture

While CFLA problem is in general NP-hard, various approximation algorithms [5][6] are proposed for this problem. Specifically, [6] proposed an efficient linear programming solution for this problem. We can use the approximation algorithms for modeling the expert matching problem, but because of the small size of the problem in our real applications, we chose to exactly solve the matching problem following the idea of linear programming. The explanation of our solution for constraint expert matching based on linear programming is described in section 3.2.

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.

Table 1. Notations

Symbol	Description
M	number of reviewers/experts
N	number of papers/projects
T	number of aspects/topics
e_i	one expert
p_j	one paper
k	Size of assigned group
c	Capacity of each reviewer

3.1 Implicit Aspects- Unconstraint matching

The first problem of expert matching (i.e. *Implicit Aspects-Unconstraint matching)* concerns with the assignment of *N* papers to *M* reviewers such that the following conditions are satisfied:

C_1- Each paper should be assigned to a group of exactly k reviewers.

C_2- (*Maximal Coverage*) - In an ideal matching, all aspects/topics (i.e. required skills) of each paper p_j should be covered by the assigned group of experts to that paper.

C_3- (*Maximal* Confidence) - In an ideal matching, each assigned reviewer e_i to paper p_j should have all the required skills of paper p_j.

In this problem, the notion of related aspects of papers and reviewers is implicit, making it difficult to maximize the aspect coverage. We assume that the related aspects of a paper can be inferred from its abstract and the skills/aspects of reviewers can be represented by his/her sample publications. We use the concatenation of a reviewer's publications as his/here expertise document. To maximally cover multiple aspects of papers, we try to find a topic representation for each paper and reviewer. Following the idea of reviewer modeling introduced in[2], 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 $\tau = (\tau_1, \dots, \tau_k)$ be a vector of topics. τ_i is a unigram language model and $p(w|\tau_i)$ is the probability of word w according to the topic τ_i. Given M reviewer's expertise documents, we can learn arbitrary number of latent topic/aspects using Probabilistic Latent Semantic Analysis (PLSA) [7]. Let $R = \{r_1, \dots r_n\}$ be the set of expertise documents (i.e. document r_i is the expertise document of reviewer e_i), the log likelihood of the expertise document collection according to the PLSA is:

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

In this equation V is the set of all the 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 probability of selection of the topic τ_a for document r_i. We can use the EM algorithm to compute the maximum likelihood estimate of all parameters including $p(a|\theta_i)$ and $p(w|\tau_a)$. After learning all the parameters, we can represent each expert e_i using the topic vector $\tau_i(\tau_{1i}, \dots \tau_{Ti})$. Furthermore, using the estimated values of $p(w|\tau_a)$, we can infer the topic representation for each paper p_j as $\tau'_j(\tau_{1j}, \dots \tau_{Ti})$. After representing papers and reviewers using the above topic model, now we can describe the matching algorithm for retrieving the *k*-top reviewers for each paper.

Following the idea of uncapacitated facility location analysis (UFLA), our intuition to match each paper p_j with a set of k reviewers can be explained like follows. We can imagine each relevant topic t_a of paper p_j as a consumer with demand of τ_{aj} and each available expert e_i as a candidate facility location. According to the condition C_1, We should open k facilities (i.e. select k reviewers) among M candidate facility locations (i.e. the number of all available reviewers) to serve the T customers (i.e. the number of aspects of paper p_j) such that the overall cost of opening those connection be minimal.

In order to match the paper p_j with M available reviewers, there are T customers and M candidate facility locations. As mentioned before, the objective function in UFLA is composed of two parts: 1) *Building Cost:* which indicates the cost of opening the facility at a specific location (i.e. the cost of selection of expert e_i for paper p_j) and 2) *Communication Cost:* which indicates the access cost of a customer (i.e. an aspect/topic) to the nearest facility location (i.e. the best reviewer for that specific aspect).

To maximize the aspect coverage of the assigned group (i.e. to satisfy the condition C_2), the assigned group should be selected such that each topic τ_a(i.e. the ath customer) of paper p_j can be assigned to a near facility location (i.e. to a reviewer who is able to cover topic a). On the other hand, to maximize the confidence of each assigned reviewers (i.e. to satisfy the condition C_3); we should select low-cost facility locations (i.e. reviewers with maximum confidence).

Therefore, we can define the building cost and communication cost in our framework as follows:

1- *Building cost* of assignment of reviewer e_i to paper p_j: $cost(e_i) = D(\overrightarrow{e_i}||\overrightarrow{p_j})$, where $\overrightarrow{e_i}$ and $\overrightarrow{p_j}$ are the topic vectors of reviewer e_i and paper p_j and $D(\overrightarrow{e_i}||\overrightarrow{p_j})$ indicates the Kullback-Leibler (KL) divergence value of these vectors.

2- *Communication cost* of assignment of aspect a of paper p_j to reviewer e_i: $cost(e_i, \tau_{aj}) = D(\overrightarrow{e_i}||\overrightarrow{v_a})$, where $\overrightarrow{e_i}$ is the topic vector of reviewer e_i and $\overrightarrow{v_a}$ indicated the unit vector with all zero elements except for topic a.

Intuitively, if the distribution of vectors $\overrightarrow{e_i}$ and $\overrightarrow{p_j}$ is 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. reviewer) e_i will be a very low building cost facility for paper p_j. On the other hand, if reviewer e_i has the skill a, we expect that the weight of his/her corresponding element in vector $\overrightarrow{e_i}$ might be higher in comparison with other elements and accordingly the communication cost of customer a (i.e. topic a) and facility r_i (i.e. reviewer) will be low.

To sum up, the objective function of unconstraint expert matching problem can be represented as follows:

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

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 local greedy search method introduced in algorithm 1. The output of the algorithm 1 is the *k*-best reviewers for paper p_j which have not only the maximum confidence but also collectively can cover all aspects of that paper.

3.2 Explicit Aspects- Constraint matching

The second problem of expert matching (i.e. *Explicit Aspects-Constraint matching)* concerns with the assignment of *N* papers to *M* reviewers such that in addition to the conditions C_1, C_2 and C_3, the following condition is satisfied:

C_4–(Capacity Condition) each reviewer has a limited capacity and can only be assigned to a limited and predefined number of papers.

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.

The constraint **C_4** makes the matching problem very hard, indeed this matching problem is also NP-hard; furthermore, in this problem the conditions **C_2** and **C_3** are in competition with each other. In other words, maximizing the global average confidence of the assigned group may result in formation of the non-optimal converging groups. In this section, we formally model this problem using Capacitated Facility location analysis (CFLA) and also propose an exact linear programming solution for it.

Similar to the UFLA method, in the CFLA framework of constraint expert matching, each aspect/topic of paper p_j is considered as a customer and each reviewer/expert is considered as a candidate facility location. In order to form optimal aspect covering groups, each required aspect of a paper should be assigned to reviewer who is able to cover that topic. On the other hand, it is preferable that the assigned reviewers of a paper be able to cover all required skills of that paper. In CFLA framework, the first condition can be satisfied by modeling the communication cost between aspects and reviewers and the second condition can be satisfied by modeling the building cost of each reviewer.

Algorithm 2 indicates the linear programming solution[1] for this CFLA problem. In this linear program, matrix M_{ij} is a $(N \times M)$ binary decision matrix that its element indicates the assignment of papers to the reviewers. Specifically, element $m_{ij} = 1$ if and only if in the final solution, the paper p_i is assigned to the reviewer r_j. As a binary decision variable, $X(i, j, t)$ indicates the assignment of topic t of paper p_i (i.e. a customer) to the reviewer e_j (i.e. a facility). Finally, $A(N \times T)$ is the paper-topic association matrix; where $A(i, t) = 1$ if and only if paper p_i is related to the topic t.

In this linear program, constraint T_1 shows that sum of elements of each row in M_{ij} should be equal to k; this means that each paper should be assigned exactly to k reviewers. Constraint T_2 indicates that the sum of elements of each column of M_{ij} should be less than c (i.e. capacity of reviewers); this means that each

reviewer can only be assigned to at most c papers. Constraint T_3 indicates that for each related topics of paper p_i (i.e. for topics that $A(i,t)=1$) at least one reviewer should be assigned. As an important constraint T_4 indicates that topic t of paper p_i can be assigned to the reviewer r_j only if paper p_i is assigned to the reviewer r_j. In other words, if decision variable $X(i,j,t)=1$ then the value of M_{ij} should be equal to 1; this means we can assign topic t of paper p_i to reviewer r_j only if r_j is selected for paper p_i.

It is worth mentioning that the objective function in Algorithm 2 is same as the objective function of equation (1) with this difference that in algorithm 2, it is defined to globally optimize the matching of all papers (i.e. the outer sum is defined on all papers). As another point, the minimum distance in equation (1) is replaced by the decision variable $X(i,j,k)$. Intuitively, by minimizing the objective function two conditions of the problem can be satisfied. Firstly, paper p_i is assigned to reviewer r_j (i.e. $M_{ij}=1$) when the building cost (i.e. $BCost(i,j)$) of this selection is low; this part of objective function can satisfy confidence maximization. On the other hand, minimization of the communication cost of topic t (i.e. $CCost(j,t)$) results in the assignment of topic t to reviewer r_j such that r_j is able to cover this topic; this part of the objective function can satisfy the coverage maximization condition and parameter λ can be used to make the tradeoff between confidence and coverage conditions.

Algorithm 2. Linear programming solution of CFLA

$$Opt = \min\left(\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\lambda\, M_{ij}BCost(i,j) + (1-\lambda)\sum_{t=1}^{l} CCost(j,t)X(i,j,t)\right)\right)$$

$$T_1 - \forall i: \sum_{j=1}^{M} M_{ij} = k \qquad T_4 - \forall i,j,t: X(i,j,t) \le M_{ij}$$

$$T_2 - \forall j: \sum_{i=1}^{N} M_{ij} \le c \qquad T_5 - \forall i,j: M_{ij} \in \{0,1\}$$

$$T_3 - \forall i,t: A(i,t) \le \sum_{j=1}^{m} X(i,j,t) \qquad T_6 - \forall i,j,k: X(i,j,k) \in \{0,1\}$$

To optimize the coverage and confidence of the assigned groups, we can define the building and communication cost as follow:

$$BCost(i,j) = \frac{aspect(p_i) \cap aspect(r_j)}{aspect(p_i)}$$

$$CCost(j,t) = \begin{cases} 0 \; if \; t \in aspect(r_j) \\ 1 \; otherwise \end{cases}$$

3.3 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. We use the PLSA topic modeling to infer the topic representation for each paper and reviewer. Utilizing the inferred topic vectors, we can use algorithm 2 for expert matching. Similar to the proposed method for estimation of Building and Communication cost in problem 1, we define the building and communication cost in the same way.

4. Related work

The problem of expert group formation has recently attracted a lot of attention in information retrieval [2] [3] [8] and social network communities [9]. This problem can be considered as an extension of the expert finding[10] problem. Expert finding is a well studied problem in IR community which concerns itself with the finding of knowledgeable people in a given topic[11]. Several algorithms are proposed for expert finding problem including the language modeling[11], voting model, and person centric language modeling [12]. While initial approaches for expert finding concern with finding the knowledgeable persons in an organization[11] , recent methods focused on finding experts in the bibliographic data[13].

The problem of multi aspect expert group formation is initially introduced by Karimzadegan and Zhai[2]. Specially, they considered the first problem of expert matching (i.e. Implicit Aspects- unconstrained matching) and proposed three different strategies to find a group of experts that maximally cover all required skills of a given query. The proposed methods [2] are redundancy removal, expert aspect modeling and query aspect modeling.

The idea of the redundancy removal method is to diversify the set of retrieved experts such that experts with various skills can be selected to cover all required skill of the query. In the query aspect modeling method, 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. Then, the union of retrieved experts is considered as the final answer.

The most effective proposed method in[2] is the 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 and queries (i.e. in paper-review assignment problem, each query is equivalent to a paper). Using these vector topics, 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 k, an expert (i.e. *the best candidate in step k*) which minimizes the following objective function is selected as a new member of the group:

$$D(\theta_q || \left(\frac{\sigma}{k-1}\sum_{i=1}^{k-1} p(a|\theta_i) + (1-\sigma)p(a|\theta_k)\right))$$

In this equation, $p(a|\theta_i)$ indicates the topic distribution of the i^{th} selected member of the group, the probability $p(a|\theta_k)$ is the topic distribution of the k^{th} candidate expert and σ is a parameter to model the redundancy of skills in selected members. Thus, the above objective function is the KL divergence of the topic distribution of the query/paper and the resulting group after selection of the k-th expert candidate.

In contrast with the *Next Best* greedy approach [2], our proposed method for implicit aspect matching problem (illustrated in Algorithm 1), measures at each iteration the ability of whole *k*-members of a group to cover the required skills of the query and as a result, if a non-appropriate member is selected at step k it can be eliminated at next steps. However, in the *Next Best* greedy approach, a non-appropriate selected member at step k cannot be changed. On the other hand, the *Next Best* greedy approach cannot easily be extended for constraint matching problems (i.e. problem 2 and 3). In contrast, our FLA framework for expert matching can be easily extended for constraint and explicit aspect matching problems.

The problem of constraint expert group formation (i.e. Problem 2 and 3) recently introduced in [3], can be considered as an extension of the paper-review assignment problem [14][15]. While these initial approaches for this problem concern with the

assignment of papers to relevant reviewers, Karmizadegan and Zhai [3], introduced the problem of multi-aspect constraint paper review matching (i.e. problem 2 and 3). They proposed a heuristic method based on the integer linear programming (ILP)
measure, it is not able to optimize the coverage measure. On the other hand, their proposed method for implicit topics/aspects (problem 3) has very low coverage and confidence in comparison with our CFLA method. We use the methods proposed in [3] and [2] as our baseline model and also use the same dataset to make the results comparable.

Recently, Tang et al. [8] 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 [9] proposed a method to find a group of experts in a social network. Although this problem 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 to each other easily.

5. Experiments

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

5.1 Data set

We used the dataset introduced in [2] to evaluate our proposed methods. This dataset is used in several research papers ([2] [3] [8]) and to the best of our knowledge is the only available dataset for multi aspect team formation problem. The dataset is crawled from the abstract papers of ACM SIGIR proceedings from years 1971-2006. 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.

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 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 proposed in [2].

Coverage score measures the number of different distinct topic aspects that are covered by the k assigned reviewers as a function of aspects in the query. Consider a paper with n_A topic aspects $A_1, \ldots, A_{n_A}$ 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:

$$\text{Coverage} = \frac{n_r}{n_A}$$

As mentioned before, in addition to the maximizing the coverage of topic aspects, the assignments that maximize the confidence of the assigned reviewers are more preferable. Specifically, in the same level of coverage, we would prefer an assignment where each reviewer is able to cover as many aspects as possible. Using the notations introduced earlier, the Average Confidence measure is defined as follows:

$$\text{Average Confidence} = \frac{\sum_{i=1}^{n_A} \frac{n_{A_i}}{k}}{n_A}$$

In this equation, k is the number of assigned reviewers and n_{A_i} indicates the number of assigned reviewers that can cover the i^{th} topic/aspect.

5.3 Baseline Methods

In the experimental result section, the FLA framework of multi-aspect expert matching is compared with the methods proposed in [2] and [3]. 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). 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-Ranktest at the level of 0.05.

6. Experimental Results

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

1) 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 baselines proposed in [2]. In particular, we compare two greedy approaches for expert matching namely, *Next Best* search and *Local Search* strategies.
2) What is the impact of 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 the parameter λ?
3) How good is the performance of the proposed framework for constraint expert matching problems in comparison with the heuristic methods proposed in [3] ?

6.1 Implicit Aspects- Unconstraint matching

In this section, we compare the UFLA method described in section 3.1 with the language Model (LM), Redundancy Removal (RR), and the *Next Best* greedy method described in related work section. In order to evaluate the effectiveness of our methods, we compare all well-tuned methods. In these experiments, 73 papers are assigned to the 189 available reviewers such that each paper gets exactly 3 reviewers. Table 2 indicates the coverage and the average confidence scores and the percentage of improvement for UFLA method.

Table 2. Comparison of the UFLA method with baseline algorithms for the first problem of expert matching- statically significant improvement is shown by * symbol.

Measure	Coverage		Average Confidence	
	result	%Δ v.s. UFLA	result	%Δ v.s. UFLA
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***	-

According to the Table 2, the performance of author topic modeling methods (i.e. *Next Best* and UFLA) are better than other baseline methods 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 equation (1), the parameters $p(a|\theta_i)$ are known) and just learn the word distributions for each topic (i.e. the only unknown parameters in equation (1) are the $p(w|a)$). 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 3 indicates the coverage and average confidence for this experiment.

Table 3. Comparison of the UFLA and *Next Best* methods- for the improved topic learning model- statically significant improvement is shown by * symbol.

Measure	Coverage		Average Confidence	
	result	%Δ v.s. UFLA	result	%Δ v.s. UFLA
Baseline-*Next Best*	0.890	+7.1%	0.660	+3.0%
UFLA	**0.953***	-	**0.680**	-

According to tabel3, by improving the topic modeling, the coverage and average confidence of both FLA and *Next Best* methods are improved. However, the performance of UFLA is again better than the *Next Best* greedy matching. The result of these experiments (i.e. Table 2 and Table 3) 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.

To better understand the behavior of *Next Best* and the UFLA methods, we examine the impact of σ and λ on performance of these methods. As mentioned before, the parameter σ in *Next Best* method models the skill redundancy in the assigned groups and the parameter λ makes the balance between building cost and commutation cost in the UFLA framework. Figure 3 and Figure 4, indicates the sensitivity of the coverage and the average confidence measures on these parameters for the UFLA and *Next Best* algorithms.

According to Figure 3, 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 equation (1)) significantly reduces the performance of UFLA matching algorithm. The same pattern is observable in Figure 4.

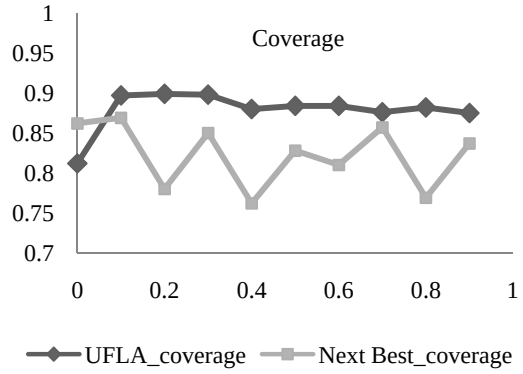

Figure 3. The sensitivity of coverage measure on parameter λ and σ

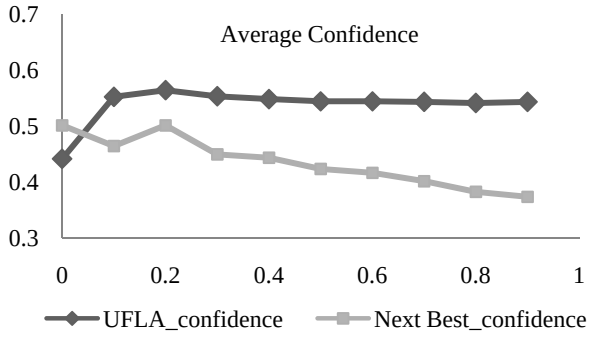

Figure 4. The sensitivity of confidence measure on parameter λ and σ

6.2 Explicit Aspects- Constraint Matching

In this section, we compare our CFLA method for constraint matching with the baseline methods proposed in [3]. The first baseline algorithm is the greedy approach proposed for constraint expert matching proposed in [3]. In this method, First, the papers are decreasingly sorted according to the number of subtopics they contain, 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 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 proposed in [3] to match papers with reviewer. This method tries to globally maximize the number of covered aspects of the assigned groups.

Before comparison with the baseline algorithms, we examine the impact of building and communication cost in CFLA model on coverage and average confidence measures. Figure 5, indicates the sensitivity of coverage for different size of program committee

(i.e. number of available reviewers). 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 specified number of experts from all available experts (i.e. 189 experts) and repeat each experiment 10 times and report the average of coverage in Figure 5.

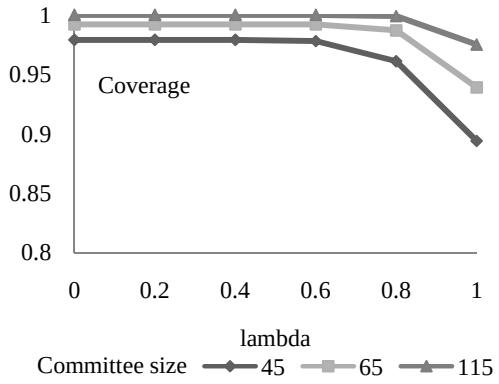

Figure 5. The sensitivity of average confidence on parameter λ CFLA matching- each date series indicates coverage score for different program committee size.

According to the Figure 5, we can see that increasing the size of committee (i.e. available experts) improves the coverage score and on the other hand increasing the 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 on communication cost in the objective function of CFLA, the coverage score of assigned groups can be increased.

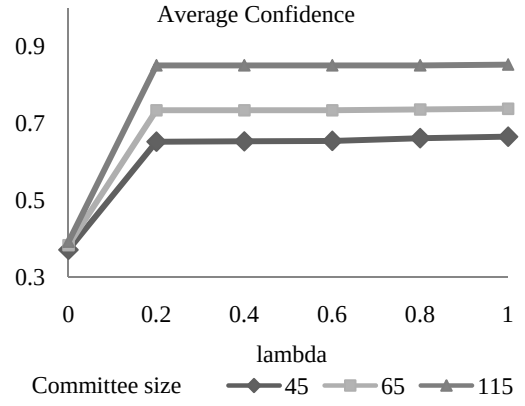

Figure 6. The sensitivity of average confidence on parameter λ CFLA matching- each date series indicates coverage score for different program committee size.

Figure 6 indicates the result of above mentioned experiment in terms of the average confidence score. While the average confidence is stable for $\lambda > 0$, the maximum and minimum values for λ is occurred at $\lambda = 1$ and $\lambda = 0$ respectivle. 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 shows that the coverage and the average confidence scores are contradicting constraints. In all other experiments of this section, we use $\lambda = 0.5$ to make a balance between the coverage and the average confidence measures.

In the next experiment, we compare the proposed CFLA method with the integer linear programming method [3] and the greedy matching algorithms for different program committee sizes (i.e. number of available reviewers). Each experiment is repeated 10 times and the average of scores are reported. In this experiment, each paper is assigned to 3 reviewers and the capacity of each reviewer is 5. Figure 7 indicates the coverage score of CFLA (i.e. proposed model), ILP and the greedy approach for different sizes of program committee.

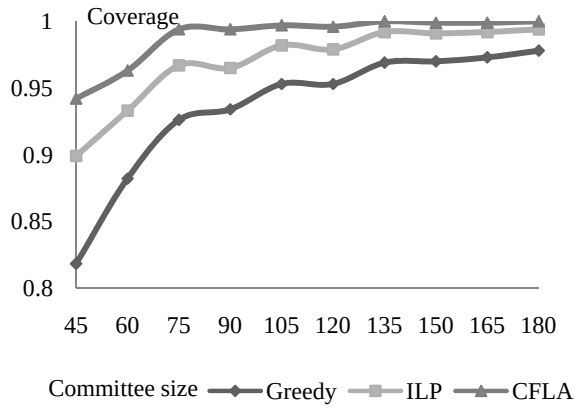

Figure 7. Coverage score of Greedy, ILP and CFLA for various committee program sizes.

According to Figure 7, by increasing the size of 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). Table 4 indicates the average confidence score of the CFLA (i.e. proposed model), ILP and the greedy approach for this experiment.

Table 4. Comparison of all method based on the Average Confidence

Committee size	45	55	65	105	185
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**

According to Table 4, 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 tge greedy method. According to this experiment, the CFLA method can detect 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, while the variation of the average confidence score is negligible.

In the next experiment, we fix the number of reviewers to 30, and vary the number of papers each reviewer can review. In order to avoid bias, we repeat the sampling process (selection 30 reviewers) for 10 times and get the average. The coverage scores are shown in Figure 8.

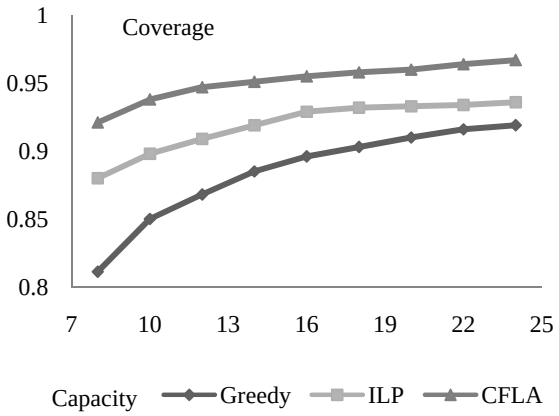

Figure 8. Coverage score of Greedy, ILP and CFLA for different capacity of reviewers.

As we increase the number of papers that each reviewer can get, 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 method for all values of capacity of reviewers. Table 5 indicates the average confidence scores for this experiment.

Table 5. Comparison of all method based on the Coverage and Average Confidence

Capacity/Size	8	12	16	20	24
Greedy	0.526	0.6	0.629	0.647	0.658
ILP	0.614	0.64	0.654	0.664	0.668
CFLA	0.615	0.64	0.655	0.664	0.668

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. It means that 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) is available. In this experiment, the maximum number of reviewers is 10 for 73 papers. Again we randomly select 10 reviewers and we repeat the sampling process for 10 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, if we have five reviewers, each should get 44 papers. Figure 9 indicates the coverage measure of CFLA and ILP methods.

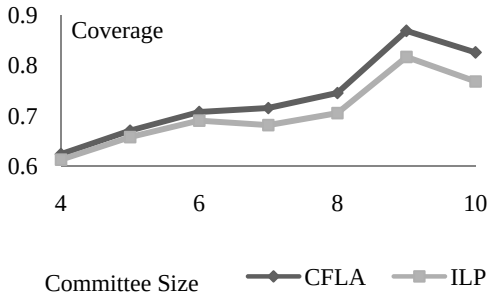

Figure 9. Coverage score of ILP and CFLA for very limited resources

The result of average confidence is also reported in Table 6. This experiments shows that for very limited recourses the quality of matching for CFLA is better than the ILP in terms of both the coverage and average confidence.

Table 6. Average Confidence score of ILP and CFLA for very limited resources

Capacity/Size	4	6	8	10
ILP	0.243	0.274	0.289	0.318
CFLA	**0.270***	**0.307***	**0.355***	**0.441***

6.3 Implicit Aspects- constraint Matching

In this section, we examine the quality of matching experts for 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 our 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 predetermined. So, we use the ILP method introduced in[3] as our baseline model. Table 7 indicates the result of matching for the best parameters (i.e. cut-off =3). In this experiment, the size of the program committee size is 189 and the capacity of each reviewer is 5.

Table 7. Comparison of ILP and CFLA for implicit Aspect

	Coverage		Average Confidence	
	Value	%**Δ** v.s ILP	Value	%**Δ** v.s ILP
ILP	0.715	-	0.347	-
CFLA	**0.828***	+15.8%	**0.436***	+25.6%

Figure 10 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 topic is used in topic vector of reviewers and papers.

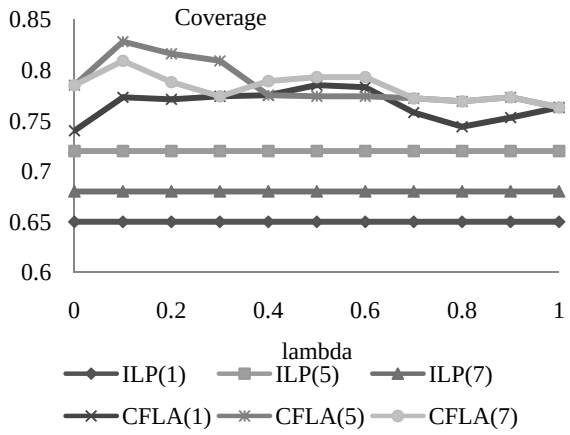

Figure 10. Coverage score of ILP and CFLA for implicit aspects

According to Figure 10, setting the cut-off equals 1 reduces the performance of both algorithms. Also, setting the cut-off more than five increases the noise and as a result the coverage reduces. The optimal value for cut-off is near to average number of required skills for each paper in the golden measure (i.e. 5 aspects for each paper). 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.

7. Conclusion

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

This work was in part supported by a grant from *Iran telecommunication research center* (ITRC). We also would like to thank the authors of [3] for making their test collection publicly available.

8. References

1 Gonzalez, Teofilo F. *Handbook of Approximation Algorithms and Metaheuristics*. Chapman & Hall/Crc Computer & Information Science Series, 2007.

2 Karimzadehgan, Maryam, Zhai, ChengXiang, and Belford, Geneva. Multi-aspect expertise matching for review assignment. In *Proceedings of the 17th ACM conference on Information and knowledge management* (2008), 1113-1122.

3 Karimzadehgan, Maryam and Zhai, ChengXiang. Integer linear programming for Constrained Multi-Aspect Committee Review Assignment. *Inf. Process. Manage.*, 48 (2012), 725--740.

4 Gabor, A.F. and Ommeren van, J.C.W. *Approximation algorithms for facility location problems with discrete subadditive cost functions*. Department of Applied Mathematics, University of Twente, Enschede, 2005.

5 Chudak, Fabián A. and Williamson, David P. Improved approximation algorithms for capacitated facility location problems. *Mathematical Programming: Series A and B*, 102, 2 (2005), 207 - 222.

6 Charikar, Moses and Guha, Sudipto. Improved Combinatorial Algorithms for the Facility Location and k-Median Problems. In *FOCS '99 Proceedings of the 40th Annual Symposium on Foundations of Computer Science* (1999), 378.

7 Hofmann, Thomas. Probabilistic latent semantic indexing. In *Proceedings of ACMSIGIR'99* (1999), 50-57.

8 Tang, Wenbin, Tang, Jie, Lei, Tao, Tan, Chenhao Gao, Bo, and Li, Tian. On optimization of expertise matching with various constraints. *Neurocomput.*, 76, 1 (2012), 71-83.

9 Lappas, Theodoros, Liu, Kun, and Terzi, Evimaria. Finding a team of experts in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data minig* (2009), 467-476.

10 Balog, K, Soboroff, I, Thomas, P, Craswell, N, and Bailey, P. The Seventeenth Text Retrieval Conference Proceedings (TREC 2008). In *NIST* (2009).

11 Balog, Krisztian, Azzopardi, Leif, and de Rijke, Maarten. A language modeling framework for expert finding. *Inf. Process. Manage.*, 45, 1 (2009), 1-19.

12 Serdyukov, Pavel and Hiemstra, Djoerd. Modeling documents as mixtures of persons for expert finding. In *Proceedings of the IR research, 30th European conference on Advances in information retrieval* (2008), 309-320.

13 Deng, Hongbo, Han, Jiawei, Michael, Lyu, and Irwin, King. Modeling and Exploiting Heterogeneous Bibliographic Networks for Expertise Ranking. In *2012 ACM/IEEE Joint Conference on Digital Libraries (JCDL 2012)* (2012).

14 Taylor, Camillo J. *On the optimal assignment of conference papers to reviewers*. University of Pennsylvania.Technical Reports, 2009.

15 Mimno, David and McCallum, Andrew. Expertise modeling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (2007), 500-509.