Neurocomputing 76 (2012) 71–83

Contents lists available at ScienceDirect

Neurocomputing

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

# On optimization of expertise matching with various constraints

Wenbin Tang [a,*], Jie Tang [a], Tao Lei [a], Chenhao Tan [b], Bo Gao [a], Tian Li [c]

[a] Department of Computer Science and Technology, Tsinghua University, Beijing 10084, China
[b] Department of Computer Science, Cornell University, Ithaca, NY 14850, USA
[c] Department of Computer Science, Beijing University of Aeronautics and Astronautics, Beijing 100084, China

ARTICLE INFO

Available online 31 July 2011

*Keywords:*
Expertise matching
Constrained optimization
Paper–reviewer assignment
Topic models

ABSTRACT

This paper studies the problem of expertise matching with various constraints. Expertise matching, which aims to find the alignment between experts and queries, is a common problem in many applications such as conference paper–reviewer assignment, product–reviewer alignment, and product-endorser matching. Most existing methods formalize this problem as an information-retrieval problem and focus on finding a set of experts for each query independently. However, in real-world systems, various constraints are often needed to be considered. For example, in order to review a paper, it is desirable that there is at least one senior reviewer to guide the reviewing process. An important question is: "Can we design a framework to *efficiently* find the *optimal solution* for expertise matching under various constraints?" This paper explores such an approach by formulating the expertise matching problem in a constraint-based optimization framework. In the proposed framework, the problem of expertise matching is linked to a convex cost flow problem, which guarantees an optimal solution under various constraints. We also present an online matching algorithm to support incorporating user feedbacks in real time. The proposed approach has been evaluated on two different genres of expertise matching problems, namely conference paper–reviewer assignment and teacher–course assignment. Experimental results validate the effectiveness of the proposed approach. Based on the proposed method, we have also developed an online system for paper–reviewer suggestions, which has been used for paper–reviewer assignment in a top conference and feedbacks from the conference organizers are very positive.



## 1. Introduction

The fusion of computer technology and human collective intelligence has recently emerged as a popular way for users to find and share information on the internet. For example, ChaCha.com, one of the largest mobile search engines, has already attracted users to answer over 300 million questions; Epinions.com, a consumer review site, has already collected thousands of millions of reviews for products. The human-based computation offers a new direction in search with its unique use of human intelligence; however, it also poses some brand new challenges. One key problem, referred to as expertise matching, is how to align human experts with questions (queries). Straightforwardly, we hope that the human experts who are assigned to answer a question have the specific expertise related to the question. But it is obviously insufficient. An ideal matching system should also consider various constraints in the real world, for example, an expert can only answer a certain number of questions (load balance); as the authoritative degree of different experts may vary largely, it is desirable that each question can be answered/reviewed by at least one senior expert (authority balance); a question may be related to multiple different aspects (topics), thus it is expected that the combined expertise of all assigned experts could cover all aspects of questions (topic coverage).

The problem has attracted considerable interests from different domains. For example, several works have been made for conference paper–reviewer assignment by using methods such as mining the web [1], latent semantic indexing [2], probabilistic topic modeling [3,4], integer linear programming [5], minimum cost flow [6] and hybrid approach of domain knowledge and matching model [7]. A few systems such as [8–13] have also been developed to help proposal–reviewer and paper–reviewer assignments. However, most existing methods mainly focus on improving the accuracy of measuring the relevance between queries and experts, i.e., how to find (or rank) relevant experts for each query, but ignore the different constraints or tackle the constraints using heuristics. Moreover, these methods usually do not consider user feedbacks. On the other hand, there are some methods focusing on expert finding. For example, Fang et al. [14] proposed a probabilistic model for expert finding, and Petkova et al. [15] employed

* Corresponding author.
*E-mail address:* tangwb06@gmail.com (W. Tang).

0925-2312/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2011.04.039

Neurocomputing 76 (2012) 71–83

SCIENCERDIRECT可用的内容列表

Neurocomputing

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

# 以各种制约因素优化匹配

Wenbin Tang a, , Jie Tang a, Tao Lei a, Chenhao Tan b, Bo Gao a, Tian Li c

清华大学计算机科学与技术系，北京10084，中国康奈尔大学计算机科学系，北京
北京北京北京市电脑科学科技科学科学科学科学科学科学科学科学科学科学科学

ARTICLE INFO

Available online 31 July 2011

关键词：专业知识匹
配约束优化纸张 –
审阅者分配主题模型

ABSTRACT

本文研究了与各种限制的专业知识问题。专业匹配，旨在找到专家和查询之间的对齐，是许多应用程序中的常见问题，如会议纸张 – 审阅者分配，产品 – 评论员对齐和工具 – 内匹配。大多数现有方法将此问题正式化为信息 – 检索问题，并专注于独立查找每个查询的一组专家。然而，在现实世界的系统中，通常需要考虑各种约束。例如，为了审查一篇论文，希望至少有一个高级审查员指导审查过程。一个重要问题是："我们可以设计一个框架，以便有效地找到各种约束下的专业知识匹配的最佳解决方案吗？ "" 本文通过在基于约束的优化框架中制定专业知识匹配问题来探讨这种方法。在拟议的框架中，专业知识匹配问题与凸起成本流量问题相关联，这保证了在各种约束下的最佳解决方案。我们还介绍了一个在线匹配算法，以支持实时结合用户反馈。拟议的方法已经在两种不同类型的专业知识匹配问题上进行了评估，即会议纸张 – 审阅者分配和教师课程任务。实验结果验证了拟议方法的有效性。根据提出的方法，我们还开发了一个纸张审阅者建议的在线系统，这些建议已被用于纸张 – 审阅者在大会上的审阅者分配，从会议组织者的反馈是非常积极的。 

## 1. Introduction

电脑技术和人类集体智能融合最近被出现为用户在互联网上查找和分享信息的流行方式。例如，Chacha.com是最大的移动搜索引擎之一，已经吸引了用户以回答超过3亿个问题;
Epinions.com，消费者审查网站，已经收集了数千千万的产品评论。以其独特的人类智能的独特使用，基于人的计算提供了新的方向;但是，它也会带来一些全新的挑战。称为专业知识匹配的一个关键问题是如何用问题（查询）对齐人类专家。直截了当，我们希望被分配回答问题的人体专家具有与问题相关的具体专业知识。但它显然是不够的。理想的匹配系统还应该考虑现实世界中的各种限制，例如，专家只能回答一定数量的问题（负载余额）;作为不同的权威程度专家可能很大程度上有所不同，希望每个问题都可以通过至少一个高级专家（权威BAL）回答/审查;一个问题可能与多个不同方面（主题）有关，因此预计所有指定专家的合并专业知识可能会涵盖问题的所有方面（主题覆盖）。问题引起了不同域名的相当兴趣。例如，已经使用诸如挖掘Web
[1]，潜在语义索引[2]，概率主题建模[3,4]，整数线性编程[5]来对诸如诸如挖掘纸张审阅者分配进行若干工程，最小成本流动[6]和域知识和匹配模型的混合方法[7]。还制定了一些诸如[8–13]的系统，以帮助提案 –
审阅者和纸张审阅者分配。然而，大多数现有方法主要关注提高测量查询和专家之间相关性的准确性，即如何查找每个查询的（或等级）相关专家，但忽略不同的约束或使用启发式解决约束。此外，这些方法通常不考虑用户反馈。另一方面，有一些方法专注于专家发现。例如，方等人。
[14]提出了一个专家查找的概率模型，佩特科维等人。 [15]雇用

* Corresponding author.
*E-mail address:* tangwb06@gmail.com (W. Tang).

0925–2312 / $ – 查看Front Mark & 2011 Elsevier
B.v.保留所有权利。DOI：10.1016 / J.neucom.2011.04.039

a hierarchical language model in enterprise corpora. Balog et al. [16] employed probabilistic models to study the problem of expert finding, which try to identify a list of experts for a query. However, these methods retrieve experts for each query independently, and cannot be directly used to deal with the expertise matching problem. Thus, several key questions arise for expertise matching, i.e., how to design a framework for expertise matching to guarantee an optimal solution under various constraints? How to develop an online algorithm so that it can incorporate user feedbacks in real time?

Fig. 1 shows an example of paper–reviewer matching problem (assigning reviewers to each paper). In the problem, the topics corresponding to a reviewer (i.e., the expertise of the reviewer) can be "machine learning", "data mining", "computational theory", etc. Also, each paper has a distribution on different topics. There are some requirements to be a good assignment. First, for each paper, the assigned reviewers' expertise should cover the topics of the paper, and all the reviewers should have a load balance (each reviewer can only review a certain number of papers). In addition, some reviewers might be senior and some might be average. We always hope that the review process of each paper can be "supervised" by at least one senior reviewer. Another example is the patient–doctor matching case, the topics corresponding to the doctor's expertise include "pediatrics", "rheumatology", "neurology", etc. Each doctor has different expertise degrees on different topics, while the disease of a patient also has a relevance distribution on the topics. Ideally, when arranging a consultation for a patient, the topics of the assigned doctors should contain the potential causes of the patient's disease (e.g, a consultation for an SLE patient requires rheumatologist, nephrologist, cardiologist and neurologist), and all the doctors should have a load balance so that no doctor overstrains.

*Contributions*. In this paper, we formally define the problem of expertise matching and propose a constraint-based optimization framework to solve the problem. Specifically, the expertise matching problem is cast as a convex cost flow problem and the objective is then to find a feasible flow with minimum cost under certain constraints. We theoretically prove that the proposed framework can achieve an optimal solution under various constraints and develop an efficient algorithm to solve it. This paper is an extension and refinement of our previous conference paper [17], and differs the previous work in two aspects. (1) We reformalize our framework considering "multi-topic coverage" matching, which is very important to paper–reviewer assignment problem. We show that topic coverage measures (e.g. *Coverage* and *Confidence*) can also be incorporated into our framework. (2) An additional dataset is introduced to evaluate the performance of our approach on multi-topic coverage. Experimental results substantiate the effectiveness and efficiency of the proposed approach. We have applied the proposed method to help assign reviewers to papers for a top conference. Feedbacks from the conference organizers confirm the usefulness of the proposed approach.

The rest of the paper is organized as follows: Section 2 reviews the relevant literatures. Section 3 formally formulates the problem. Section 4 explains the proposed optimization framework. Section 5 gives experimental results that validate the effectiveness and the computational efficiency of our methodology. Finally, Section 6 concludes.

## 2. Related work

In general, existing methods for expertise matching mainly fall into two categories: probabilistic model and optimization model. The probabilistic model tries to improve the matching accuracy between experts and queries based on different probabilistic models such as keyword matching [1], latent semantic indexing [2], probabilistic topic modeling [3,4]. However, most of these methods do not consider the various constraints or simply consider the constraints by heuristics. The optimization model tries to incorporate the constraints as a component in an optimization framework such as integer linear programming [5] and minimum cost flow [6].

Most previous works cast expert matching or expert finding as an information-retrieval problem, in which every expert is represented as a "expertise" document and given a query, the goal is to retrieve most relevant experts. As a result, these methods mainly focus on two points: how to define the matching score between a query and a document; and how to represent each expert [18,19]. For example, Dumais and Nielsen [2] use latent semantic indexing (LSI) as the retrieval method and the abstracts provided by reviewer as expertise documents. Yu et al. [20] represent experts by analyzing text content and extracting related information. Basu et al. [1,21,22] integrate different sources of information for recommendation (e.g. publications, research interests, etc.). Yarowsky and Florian [23] assign a paper by computing its cosine similarity with a reviewer and choosing the one with the highest rank. Other expert finding work include [24,25].

In addition, different language models [14–16,26,27] and topic models [28] are used for expert matching/finding problem. In all of the language models, the matching score is the probability of a query given an expertise document, i.e., $p(q|d)$, but its definition varies. Mimno and McCallum [4] improve the matching accuracy by proposing a novel topic model Author-Persona-Topic (APT), in which experts are represented as independent distributions over topics. Karimzadehgan et al. also consider matching experts on multiple aspects of expertise [3]. Unlike the previous probabilistic

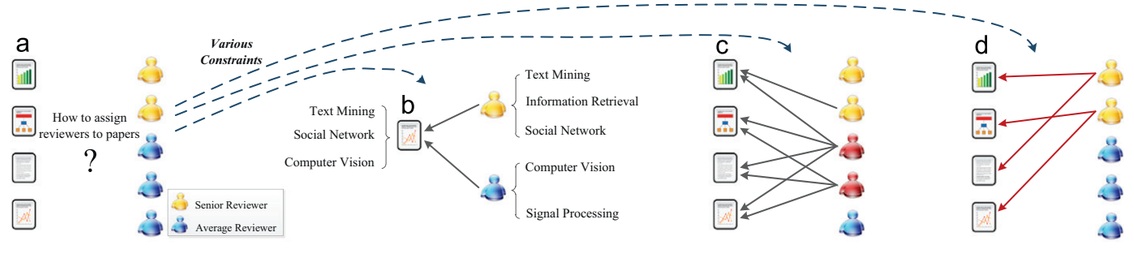



**Fig. 1.** (a) An illustration of the paper–reviewer assignment problem. To make an ideal assignment, some constraints should be considered. For example: (b) Each paper is related to one or several topics, each reviewer also has expertise on different topics. For a paper, the combination of the expertise of all assigned reviewers should cover all topics of the paper. (c) The assignment should be "load balanced". A bad example is shown, where the two reviewers marked with red are assigned with too many works, meanwhile the blue one does not get any assignment. (d) The authoritative degree of reviewers may vary largely, it is also desirable that each paper is assigned with at least one senior reviewer, so that she/he can guide the review process.



企业语料库中的分层语言模型。Balog等人。[16]采用概率模型来研究专家发现的问题，该问题试图确定查询专家名单。但是，这些方法独立检索每个查询的专家，不能直接用于处理专业匹配问题。因此，有关专业知识匹配的几个关键问题，即如何设计专业知识匹配的框架，以保证在各种限制下的最佳解决方案？如何开发在线算法，以便它可以实时合并用户反馈？

图。图1示出了纸张审阅者匹配问题的示例（将审阅者分配给每份纸张）。在问题中，对应于审阅者（即，审阅者的专业知识）对应的主题可以是'机器学习"，'数据挖掘"，'计算-ORY"等。也，每篇论文也是如此在不同的主题上发行。有一些要求是一个好的任务。首先，对于每篇论文，指定的审稿人的专业知识应该涵盖论文的主题，所有审稿人都应该有负载余额（每个审阅者只能审查一定数量的论文）。此外，一些审稿人可能是高级，有些人可能是平均的。我们始终希望每份纸张的审查过程可以由至少一个高级评价者""监督"。另一个例子是患者 -
医生匹配案例，主题追加到医生的专业知识包括"儿科"，"Rheuma-Tological" '，'神经病学'等。每位医生都有不同的专业知识学位主题，而患者的疾病也有关于主题的相关分布。理想情况下，在安排患者的咨询时，指定医生的主题应包含患者疾病的潜在原因（例如，SLE患者的咨询需要风湿病学家，肾病学家，心理学家和神经病学家），以及所有医生应该有一个负载余额，以便没有医生过剩。贡献。在本文中，我们正式定义了专业知识匹配问题并提出了基于约束的优化框架来解决问题。具体地，专业知识匹配问题作为凸起成本流问题，目标是在某些约束下找到具有最小成本的可行流量。理论上我们证明，所提出的框架可以在各种结构下实现最佳解决方案，并开发一种有效的算法来解决它。本文是我们之前的会议论文的延伸和改进[17]，与以前的工作有所不同。
（1）我们考虑匹配的"多主题覆盖"匹配来重新正式化我们的框架，这对纸张审阅者分配问题非常重要。我们展示了主题覆盖措施（例如，覆盖范围和信心）也可以纳入我们的框架。
（2）引入了一个额外的数据集以评估我们对多主题覆盖的方法的性能。实验结果证实了效果和效率

提出的方法。我们已应用提出的方法，以帮助将审阅者分配给首脑会议的论文。会议组织者的反馈确认了拟议方法的有用性。其余部分组织如下：第2节审查相关文献。第3节正式制定了漂白剂。第4节解释了所提出的优化框架。第5节给出了验证了我们方法的有效性和计算效率的实验结果。最后，第6节结束了。

## 2. Related work

一般来说，现有的专业知识匹配方法主要分为两类：概率模型和优化模型。概率模型试图根据关键字匹配[1]，潜在语义索引[2]，概率主题建模[3,4]等不同概率模型来提高专家和查询之间的匹配准确性。然而，这些方法中的大多数不考虑各种限制或简单地通过启发式来控制约束。优化模型试图将约束包含在优化框架中的组件，例如整数线性编程[5]和最小成本流程[6]。最先前的工程投射专家匹配或专家查找作为信息检索问题，其中每个专家都被称为"专业知识"文件，并给出了查询，目标是检索大多数相关专家。结果，这些方法主要关注两点：如何在查询和文档之间定义匹配分数;以及如何代表每个专家[18,19]。例如，Dumais和Nielsen
[2]使用潜在语义索引（LSI）作为审阅者提供的检索方法和摘要作为专业知识文件。yu等人。
[20]通过分析文本内容和提取相关信息来代表专家。Basu等人。[1,21,22]整合不同的信息来源以供建议（例如出版物，研究兴趣等）。Yarowsky和Florian
[23]通过将其与审阅者的余弦相似度计算并选择具有最高等级的余弦相似性分配纸张。其他专家寻找工作包括[24,25]。此外，不同的语言模型[14-16,26,27]和主题模型[28]用于专家匹配/查找问题。在所有语言模型中，匹配分数是给定专业知识文件的查询的概率，即P Q9D ‖，但其定义变化。MIMNO和MCCALLUM
[4]通过提出一个新颖的主题模型作者 -
Persons-pompulea-主题（APT）来提高匹配的准确性，其中专家表示为与主题的独立发行版。
Karimzadehgan等。还要考虑匹配专门知识的多个方面的专家[3]。与以前的概率不同

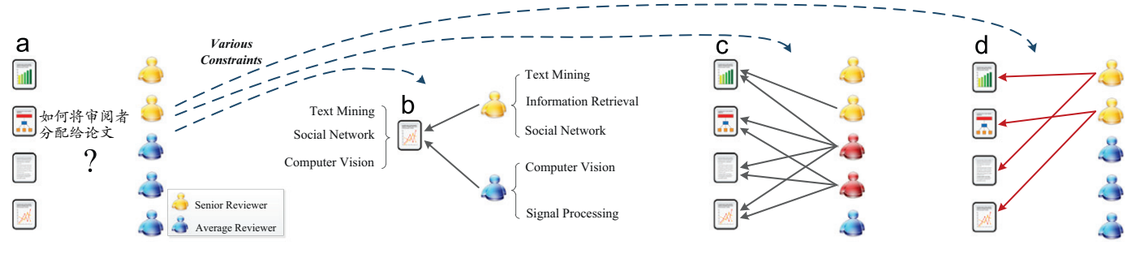



图1.（a）纸张审阅者分配问题的图示。为了做出理想的作业，应考虑一些约束。例如：（b）每篇论文与一个或多个主题有关，每个审稿人也具有不同主题的专业知识。对于一篇论文，所有指定审核人员的专业知识的组合应涵盖本文的所有主题。（c）作业应该是"负载平衡"。显示了一个糟糕的例子，其中标有红色的两个评论员被分配了太多的作品，同时蓝色没有得到任何分配。（d）审阅者的权威程度可能很大程度上，也有意义上，每篇论文至少有一个高级评论员分配，以便她/他可以指导审查流程。

models in which a query is matched as a whole unit, it tries to find "comprehensive" matchings to cover all subtopics of a query. New measures *Coverage* and *Confidence* are defined to evaluate multi-aspect expertise matching results. Several matching methods are proposed and show better multi-aspect performance than traditional ones.

However, most of the aforementioned works treat each query independently and ignore the certain constraints (e.g. load balance), thus they cannot be directly adapted to an expertise matching system. In the real world, expertise matching is a highly constrained problem, and some existing works study this constrained optimization problem using various methods. For example, Guervós et al. [29] combine a greedy and an evolutionary algorithm [30,31] to assign papers to reviewers. Karimzadehgan et al. [5] and Taylor [32] cast it as an integer linear programming (ILP) problem so approximate solutions can be found by any ILP solver. Sun et al. [7] solves the reviewer assignment by hybrid approach of domain knowledge. Recently, a few systems [8–10,33,34,11–13] have also been developed to help proposal–reviewer and paper–reviewer assignments. However, the expertise matching problem is still treated as an information-retrieval problem, which obviously cannot result in an optimal solution.

In this paper, we aim to formalize the problem of expertise matching in a constraint-based optimization framework and propose an efficient algorithm to solve the framework. The differences of our work from existing work are: (a) we offer an optimization framework that incorporates the expertise matching and various constraints together; (b) the framework can be easily extended since new constraint can be combined into the optimization framework by simply defining a new (hard or soft) constraint; and (c) the framework can guarantee an optimal solution.

## 3. Problem formulation

In this section, we first give several necessary definitions and then present a formal definition of the problem.

Given a set of experts $V=\{v_i\}$, each expert has different expertise over all topics. Formally, we assume that there are $T$ aspects of expertise (called topics) and each expert $v_i$ has different expertise degrees on different topics. Further, given a set of queries $Q=\{q_j\}$, each query is also related to multiple topics. Given this, we first define the concept of topic model.

**Definition 1** (*Topic model*). A topic model $\theta$ of an expert (or a query) is a multinomial distribution of words $\{p(w|\theta)\}$. Each expert (query) is considered as a mixture of multiple topic models. The assumption of this model is that words associated with the expert (query) are sampled according to the word distributions corresponding to each topic, i.e., $p(w|\theta)$. Therefore, words with the highest probability in the distribution would suggest the semantics represented by the topic.

Assuming we have $T$ topics, the expertise degree of expert $v_i$ on topic $z\in\{1\cdots T\}$ is represented as a probability $\theta_{v_i z}$ with $\sum_z \theta_{v_i z}=1$. Similarly, for each query, we also have a $T$-dimensional topic distribution with $\sum_z \theta_{q_j z}=1$. Notations are summarized in Table 1.

It is easy to understand that each query $q_j$ can be represented as a sequence of words, i.e., $d_{q_j}$. To represent every expert $v_i$, without loss of generality, we also consider it as a sequence of words, i.e., $d_{v_i}$. Based on this representation, we can calculate the similarity (or relevance score) between each query and every expert using measures such as cosine similarity or language model. Given this, we can define our problem of expertise matching with various constraints.

**Table 1**
Notations.

| Symbol | Description |
|---|---|
| $M$ | Number of experts |
| $N$ | Number of queries |
| $T$ | Number of topics |
| $V$ | The set of candidate experts |
| $Q$ | The set of queries |
| $v_i$ | One expert |
| $q_j$ | One query |
| $\theta_{v_i z}$ | The probability of topic $z$ given expert $v_i$ |
| $\theta_{q_j z}$ | The probability of topic $z$ given query $q_j$ |
| $T(v_i)$ | The set of major related topics of expert $v_i$ |
| $T(q_j)$ | The set of major related topics of query $q_j$ |

**Problem 1** (*Expertise matching with constraints*). Given a set of experts $V$ and a set of queries $Q$, the objective is to assign $m$ experts to each query by satisfying certain constraints, such as (1) the number of assigned queries with each expert should be in a range $[n_1,n_2]$, where $n_1\le n_2$; (2) the experts' major topics should cover the query's related topics; (3) the assignment should avoid some conflict-of-interest (COI).

Actually in some applications, satisfying the constraints is more important than matching expertise with the queries. For example, in the conference paper–reviewer assignment, the authors of a paper should not be assigned to review their own papers. This must be a hard constraint. While in some other scenario, the constraint is relatively soft, for example the load balance among experts. The number of assigned queries to each expert can be in a range between $n_1$ and $n_2$. In existing works, Dumais et al. [2] and Mimno et al. [4] mainly focus on improving the accuracy of expertise matching, but ignore how to obtain an optimal matching satisfying the various constraints. Karimzadehgan et al. [5] use integer linear programming to find the solution for expertise matching with constraints. However, the proposed model cannot guarantee an optimal solution. In this work, we propose a generalizable optimization framework to solve this problem. Various constraints can also be incorporated in the framework.

## 4. The constrained optimization framework

In this section, we propose a constraint-based optimization framework for expertise matching. We develop an efficient algorithm to solve the optimization framework based on the theory of convex cost flow, and also present an online matching algorithm to incorporate user feedbacks in real time.

*Basic idea.* The basic idea of our approach is to formulate this problem in a constrained optimization framework. Different constraints can be formalized as penalty in the objective function or be directly taken as the constraints in the optimization solving process. For solving the optimization framework, we transform the problem to a convex cost network flow problem, and present an efficient algorithm which guarantees an optimal solution.

### 4.1. The framework

Now, we explain the proposed approach in detail. In general, our objective can be viewed from two perspectives: maximizing the matching score between experts and queries and satisfying



查询作为整个单位匹配的模型，它试图找到“”全面“匹配以覆盖查询的所有子主题。新措施覆盖和信心被定义为评估多方面专业知识匹配结果。提出了几种匹配方法，并显示出比传统的更好的多方面性能。然而，大多数上述作品独立地处理每个查询并忽略某些约束（例如，负载平衡），因此它们不能直接适应专业匹配系统。在现实世界中，专业匹配是一个高度约束的问题，一些现有的作品使用各种方法研究了这种相应的优化问题。对于考试，Guervo'S等人。
[29]结合了贪婪和进化算法[30,31]将文件分配给审阅者。
Karimzadehgan等。 [5]和Taylor
[32]将其投射为整数线性编程（ILP）问题，因此任何ILP求解器都可以找到近似解决方案。太阳等。
[7]通过域知识的混合方法解决审阅者分配。最近，一些系统

还制定了[8−10,33,34,11−13]，以帮助提案 −
审阅者和纸张审阅者任务。但是，匹配问题仍然被视为信息检索问题，这显然无法导致最佳解决方案。在本文中，我们的目标是在基于约束的优化框架中模拟专业知识匹配问题，并提出了一种努力解决框架的算法。我们对现有工作的工作的差异是：（a）我们提供优化框架，将专业知识匹配和各种限制集合在一起;（b）由于简单地定义新的（硬或软）约束，可以轻松扩展框架，因为新约束可以将新约束组合到优化框架中;（c）框架可以保证最佳解决方案。

## 3. Problem formulation

在本节中，我们首先给出几个必要的定义，然后提出问题的正式定义。鉴于一组专家
V1/4[vi]，每个专家对所有主题都有不同的专业知识。从形式上讲，我们假设有T 方面的专业知识（称为主题），每个专家 vi 在不同主题上具有不同的专业知识学位。此外，鉴于第 1/4 季度/qj}的一组查询，每个查询还与多个主题相关。鉴于此，我们首先定义了主题模型的概念。

定义1（主题模型）。专家（或查询）的主题模型Y是单词FP w9y g的多项分布。每个专家（查询）被视为多主题模型的混合。该模型的假设是根据与专家（查询）相关联的单词根据关于每个主题的词分布，即P w9y 。因此，分布中具有最高概率的单词将建议主题所代表的神学。

假设我们有T主题，主题ZAF1 TG专业vi的专业知识学位zaf1示为具有p
zyviz 1的概率YVIZ。类似地，对于每个查询，我们也具有与p zyqjz 的t维主题分布。1.总结了符号

表1.很容易理解，每个查询Qj都可以表示为单词序列，即dqj。要代表每个专家VI，没有普遍性，我们也将其视为一系列单词，即DVI。基于此表示，我们可以计算每个查询和每个专家之间的相似性（或相关性分数）使用余弦相似性或语言模型等措施。鉴于此，我们可以定义我们与各种限制的专业知识问题。

**Table 1**
Notations.

| Symbol | Description |
|---|---|
| | m次数n次疑问t题目v候选专家q集查询vi一个专家qj一个查询yviz主题z给定专题vi yqjz给定查询qj<br>t的概率z给定查询qj<br>t（vi）关于专家VI<br>T（QJ）的主要相关主题集的一组主要相关主题QJ |

问题1（与约束匹配的专业知识）。给定一组专家v和一组查询Q，目标是通过满足某些约束来将M专家分配给每个查询，例如（1）每个专家的分配查询的数量应该在一个范围内[n1，n2]，其中n1
rn2;（2）专家的主要主题应涵盖查询的相关主题;（3）任务应避免某种利益冲突（COI）。

实际上在某些应用中，满足约束比与查询的匹配专业知识更重要。例如，在会议论文审阅者分配中，不应分配文件的作者来审查自己的论文。这必须是一个艰难的约束。虽然在一些其他场景中，约束相对较软，例如专家之间的负载平衡。对每个专家的分配查询的数量可以在N1和N2之间的范围内。在现有的作品中，Dumais等人。 [2]和mimno等人。
[4]主要专注于提高专业知识匹配的准确性，但忽略了如何获得满足各种约束的最佳匹配。 Karimzadehgan等。
[5]使用整数线性编程来找到与约束匹配的专业知识的解决方案。但是，所提出的模型无法保证最佳解决方案。在这项工作中，我们提出了一个普遍的优化框架来解决这个问题。各种约束也可以包含在框架中。

4.受限制的优化框架

在本节中，我们提出了一种基于约束的优化框架，用于专业匹配。我们开发了一种高效的算法，基于凸成本流理论来解决优化框架，并呈现一个在线匹配算法，以实时合并用户反馈。基本想法。我们方法的基本思想是在约束优化框架中制定这个问题。不同的共同之躯可以在目标函数中正式化，或者直接被视为优化解决过程中的约束。为了解决优化框架，我们将问题转换为凸起成本网络流量问题，并提出了一种保证最佳解决方案的有效算法。

### 4.1. The framework

现在，我们详细解释了提出的方法。一般来说，我们的目标可以从两个角度看，最大化专家和疑问之间的匹配分数和令人满意

the given constraints. Formally, we denote the set of experts to answer query $q_j$ as $V(q_j)$, and the set of queries assigned to expert $v_i$ as $Q(v_i)$. Further, we denote the matching score between expert $v_i$ and query $q_j$ as $R_{ij}$. Therefore, a basic objective function can be defined as follows:

$$\text{Max} \sum_{v_i \in V} \sum_{q_j \in Q(v_i)} R_{ij} \tag{1}$$

The objective function can be equivalently defined as $\sum_{q_j \in Q} \sum_{v_i \in V(q_j)} R_{ij}$. In different applications, the constraints will be different. Here we use several general constraints to explain how the proposed framework can incorporate different constraints.

The first constraint is that each query should be assigned with exactly $m$ experts. For example, in the paper–reviewer assignment task, each paper should be assigned with 3 or 5 reviewers. This constraint can be directly added into the optimization problem. Formally, we have

$$\textbf{ST1}: \quad \forall q_j \in Q, \ |V(q_j)| = m \tag{2}$$

The second constraint is called as *expert load balance*, indicating that each expert can only answer a limited number of queries. There are two ways to achieve this purpose: define a *strict* constraint or add a *soft penalty* to the objective function.

For *strict*, we add a constraint indicating that the number of assigned queries to every expert $v_i$ should be equal or larger than a minimum number $n_1$, but be equal or smaller than a maximum number $n_2$. The *strict* constraint can be written as

$$\textbf{ST2}\ (\textit{strict}): \quad \forall v_i \in V, \quad n_1 \le |Q(v_i)| \le n_2 \tag{3}$$

The other way is to add a soft penalty to the objective function (Eq. (1)). For example, we can define a square penalty as $|Q(v_i)|^2$. By minimizing the sum of the penalty $\sum_i |Q(v_i)|^2$, we can achieve a *soft* load balance among all experts, i.e.:

$$\text{softpenalty}: \quad \text{Min} \sum_{v_i \in V} |Q(v_i)|^2 \tag{4}$$

These two methods can be also used together. Actually, in our experiments, soft penalty method gives better results than strict constraint. Combining them together can yield a further improvement.

The third constraint is called *authority balance*. In real application, experts have different expertise level (authoritative level). Take the paper–reviewer assignment problem as an example. Reviewers may be divided into two levels: senior reviewers and average reviewers. Intuitively, we do not hope that the assigned reviewers to a paper are all average reviewers. It is desirable that the senior reviewers can cover all papers to guide (or supervise) the review process. Without loss of generality, we divide all experts into $K$ levels, i.e., $V^1 \cup V^2 \cup \cdots \cup V^k = V$, with $V^1$ representing experts of the highest authoritative level. Similar to *expert load balance*, we can define a strict constraint like $|V^1 \cap V(q_j)| \ge 1$, and also add a penalty function to each query $q_j$ over the $k$-level experts. Following, we give a simple method to instantiate the penalty function:

$$\textbf{ST3}: \quad \text{Min} \sum_{k=1}^{K} \sum_{j=1}^{N} |V^k \cap V(q_j)|^2 \tag{5}$$

Besides the above constraints, we also wish to assign experts that can cover all topics in the query. In the paper–reviewer assignment problem, for example, a paper may be related to several research areas, thus ideally in a comprehensive assignment, the paper is reviewed by a group of experts which cover all of the topics.

We introduce *related topics* of queries/experts. The related topics of query $q_j$ and expert $v_i$ are denoted as $T(q_j)$ and $T(v_i)$, indicating the most relevant aspects to the query and expert, respectively. $T(q_j)$ and $T(v_i)$ can be determined in different ways. Again in the paper–reviewer assignment problem, authors may be required to select the related topics of their papers from predefined categories in the submission, and reviewers can also select their expertise topics. In addition, it is also possible to estimate related topics from the learned topic distributions: (a) select top-$k$ topics in $\theta_{v_i}$, $\theta_{q_j}$ as the corresponding related topics; or (b) use thresholds $\tau_v$ and $\tau_q$ to prune topics, i.e., related topics are determined by $\theta_{v_i,z} > \tau_v$ and $\theta_{q_j,z} > \tau_q$.

Follow the work [3], we can incorporate different evaluation measures for topic covering. One measure is called *Coverage*, as we hope that assigned experts can cover different topics of a given query, i.e.,

$$\textit{Coverage}(q_j) = \frac{|T(q_j) \cap \bigcup_{v_i \in V(q_j)} T(v_i)|}{|T(q_j)|} \tag{6}$$

In this way, an optimal assignment should maximize the *Coverage*, e.g., $T(q_j) \subseteq \bigcup_{v_i \in V(q_j)} T(v_i)$. But this becomes the NP-hard set cover problem which is intractable. Consequently, we choose to find less optimal solutions by making further assumptions: for a specific query $q_j$, each assigned expert $v_i \in V(q_j)$ can select only one *responsible topic* $\hat{T}_{q_j}(v_i) \in T(v_i)$, and covers this topic for the query. The optimal solution under the assumption provides a lower bound of the original problem. Finding the matching maximizing the coverage with responsible topics actually opens another optimization problem, but fortunately this can be incorporated into our framework. We leave the discussion to Section 5.2.

Another measure proposed is called *Confidence*, as we prefer the related topics to be covered by as many experts as possible, i.e.,

$$\textit{Confidence}(q_j) = \frac{1}{m} \sum_{v_i \in V(q_j)} \frac{|T(q_j) \cap T(v_i)|}{|T(q_j) \cap \bigcup_{v_i \in V(q_j)} T(v_i)|} \tag{7}$$

Generally there is a *Coverage-Confidence* tradeoff. To achieve both high coverage and high confidence, a measure *Average Confidence* is accordingly defined as:

$$\textit{AverageConfidence}(q_j) = \frac{1}{m} \sum_{v_i \in V(q_j)} \frac{|T(q_j) \cap T(v_i)|}{|T(q_j)|} \tag{8}$$

We finally choose the average confidence as the fourth constraint in our optimization framework.

$$\textbf{ST4}: \quad \text{Max} \sum_{q_j \in Q} \sum_{v_i \in V(q_j)} \frac{|T(q_j) \cap T(v_i)|}{|T(q_j)|} \tag{9}$$

The last constraint is called *COI avoidance*. In many cases, we need to consider the conflict-of-interest (COI) problem. For example, an author, of course, should not review his own or his coauthors' paper. This can be accomplished through employing a binary $M \times N$ matrix $U$. An element with value of 0, i.e., $U_{ij}=0$, represents expert $v_i$ has the conflict-of-interest with query $q_j$. A simple way is to multiply the matrix $U$ with the matching score $R$ in (Eq. (1)).

Finally, by incorporating different constraints in Eqs. (4)–(9) and the COI matrix $U$ into the basic objective function (Eq. (1)), we can result in a constraint-based optimization framework, e.g.:

$$\begin{aligned} \text{Max} \ & \sum_{v_i \in V} \sum_{q_j \in Q(v_i)} U_{ij} R_{ij} - \sum_{k=1}^{K} \left( \mu_k \sum_{j=1}^{N} |V^k \cap V(q_j)|^2 \right) \\ & - \beta \sum_{v_i \in V} |Q(v_i)|^2 + \lambda \sum_{q_j \in Q} \sum_{v_i \in V(q_j)} \frac{|T(q_j) \cap T(v_i)|}{|T(q_j)|} \\ \text{s.t.} \ & \forall q_j \in Q, \ |V(q_j)| = m \\ & \forall v_i \in V, \ n_1 \le |Q(v_i)| \le n_2 \end{aligned} \tag{10}$$



给定的约束。正式，我们表示，将查询QJ作为v（QJ）以及分配给Expert
VI的查询集作为Q（vi）。此外，我们表示专家VI和查询QJ之间的匹配分数作为RIJ。因此，基本目标函数可以定义如下：

$$\text{Max} \sum_{v_i \in V} \sum_{q_j \in Q(v_i)} R_{ij} \tag{1}$$

客观函数可以等同地定义为P QJ AQ P
viav qj rij。在不同的应用中，约束

会是不同的。在这里，我们使用几个一般约束来解释所提出的框架如何包含不同的约束。第一个约束是每个查询都应以完整的专家分配。例如，在纸张审阅者分配任务中，应使用3或5名审核员分配每份纸张。可以直接添加该约束在优化问题中。正式，我们有

$$\textbf{ST1}: \quad \forall q_j \in Q, \ |V(q_j)| = m \tag{2}$$

第二个约束被称为专家负载余额，表明每个专家只能应答有限数量的查询。有两种方法可以实现此目的：定义严格的约束或向目标函数添加软罚款。对于严格，我们添加了一个约束，指示每个专家VI的分配查询的数量应该等于或大于最小数字N1，但是等于或小于最大数量N2。严格的约束可以写成

$$\textbf{ST2}\ (\textit{strict}): \quad \forall v_i \in V, \quad n_1 \le |Q(v_i)| \le n_2 \tag{3}$$

另一种方法是向目标函数（EQ.（1））添加软罚款。例如，我们可以定义AS9Q VI 92.通过最小化罚款Pi9q vi 92的总和，我以在所有专家中实现软载余额，即：

$$\text{softpenalty}: \quad \text{Min} \sum_{v_i \in V} |Q(v_i)|^2 \tag{4}$$

这两种方法也可以一起使用。实际上，在我们的实验中，软惩罚方法提供比严格的限制更好的结果。将它们结合在一起可以产生进一步的改进。第三个约束被称为权限余额。在实际应用中，专家有不同的专业知识水平（权威）。以纸张审阅者分配问题为例。审核人员可分为两个级别：高级审查员和平均审稿人。直观地，我们不希望将分配的审稿人员造成一份纸质是平均审稿人。希望高级审查人员可以涵盖所有文件来指导（或监督）审查过程。不损失普遍性，我们将所有专家划分为K水平，即V1 [v2 [v2
[v1表示最高权威级别的专家。类似于专家负载余额，我们可以定义一个严格的约束，如9v1
\v qj 9z1，并在K级专家上为每个查询Qj添加惩罚功能。以下情下，我们提供了一种简单的方法来实例化惩罚功能：

$$\textbf{ST3}: \quad \text{Min} \sum_{k=1}^{K} \sum_{j=1}^{N} |V^k \cap V(q_j)|^2 \tag{5}$$

除了上述约束之外，我们还希望分配可以涵盖查询中所有主题的专家。例如，在纸张审阅者分配问题中，一篇论文可能与若干研究领域有关，因此理想情况下，该论文由一组专家审查，涵盖了所有主题。

我们介绍查询/专家的相关主题。查询QJ和专家VI的相关主题表示为T（QJ）和T（VI），分别表示查询和专家的最相关方面。T（QJ）和T（VI）可以以不同的方式确定。再次在纸张审阅者分配问题中，可能需要提交人从提交中预定义的类别选择文件的相关主题，审阅者也可以选择他们的专业知识。此外，还可以从学习主题分发中估计相关主题：（a）在YVI，YQJ中选择Top−K主题作为相应的相关主题;或（b）使用阈值电视和tq to
prune主题，即相关主题由YVIZ 4TV和YQJZ
4TQ确定。按照工作[3]，我们可以纳入不同的评估措施，以了解主题覆盖。一项措施称为覆盖，我们希望指定的专家可以涵盖给定查询的不同主题，即，

$$\textit{Coverage}(q_j) = \frac{|T(q_j) \cap \bigcup_{v_i \in V(q_j)} T(v_i)|}{|T(q_j)|} \tag{6}$$

通过这种方式，最佳任务应最大化覆盖率，例如，t qj dsviav qj t vi。但这成为了难以相容的NP−Hard
Set涵盖问题。因此，我们选择通过进一步假设找到更少的最佳解决方案：对于特定的查询QJ，每个分配的专家VI Qj 只能选择一个响应的主题ˆ
tqj vi at vi ，并涵盖查询的本主题。假设下的最佳解决方案提供了原始问题的下限。找到匹配最大化的覆盖范围与负责任主题实际上打开另一个优化产品，但幸运的是，这可以纳入我们的框架。我们将讨论留给5.2节。提出的另一项措施被称为信心，因为我们更喜欢与尽可能多的专家涵盖的相关主题，即，

$$\textit{Confidence}(q_j) = \frac{1}{m} \sum_{v_i \in V(q_j)} \frac{|T(q_j) \cap T(v_i)|}{|T(q_j) \cap \quad _{v_i \in V(q_j)} T(v_i)|} \tag{7}$$

通常存在覆盖置信措施。为实现高覆盖率和高信任，因此衡量平均置信度被定义为：

$$\textit{AverageConfidence}(q_j) = \frac{1}{m} \sum_{v_i \in V(q_j)} \frac{|T(q_j) \cap T(v_i)|}{|T(q_j)|} \tag{8}$$

我们终于选择了我们优化框架中的第四个限制的平均信心。

$$\textbf{ST4}: \quad \text{Max} \sum_{q_j \in Q} \sum_{v_i \in V(q_j)} \frac{|T(q_j) \cap T(v_i)|}{|T(q_j)|} \tag{9}$$

最后的约束被称为COI避免。在许多情况下，我们需要考虑利益冲突（COI）问题。例如，当然，作者不应该审查自己或他的共同咨询纸。这可以通过采用二进制M n矩阵U来完成。具有0，i.，Uij 0的元素，表示专家VI对查询QJ具有兴趣冲突。一种简单的方法是将矩阵U乘以匹配得分r（eq。（1））。最后，通过在EQ S中纳入不同的约束。（4）−
（9）和COI矩阵U进入基本目标函数（EQ。（1）），我们可以导致基于约束的优化框架，例如：

$$\begin{aligned} \text{Max} \ & \sum_{v_i \in V} \sum_{q_j \in Q(v_i)} U_{ij} R_{ij} - \sum_{k=1}^{K} \left( \mu_k \sum_{j=1}^{N} |V^k \cap V(q_j)|^2 \right) \\ & - \beta \sum_{v_i \in V} |Q(v_i)|^2 + \lambda \sum_{q_j \in Q} \sum_{v_i \in V(q_j)} \frac{|T(q_j) \cap T(v_i)|}{|T(q_j)|} \\ \text{s.t.} \ & \forall q_j \in Q, \ |V(q_j)| = m \\ & \forall v_i \in V, \ n_1 \le |Q(v_i)| \le n_2 \end{aligned} \tag{10}$$

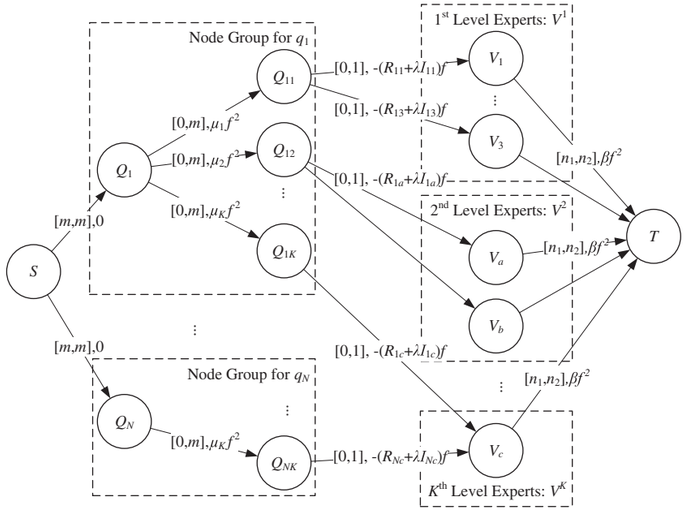



**Fig. 2.** The construction of convex cost network flow according to objective function (10).

where $\lambda$, $\beta$ and $\mu_k$ are lagrangian multipliers, used to tradeoff the importance of different components in the objective function.

Now the problem is how to define the topic distribution $\theta$, how to calculate the pairwise matching score $R_{ij}$, and how to optimize the framework.

### 4.2. Modeling multiple topics

The goal of topic modeling is to associate each expert $v_i$ with a vector $\theta_{v_i} \in \mathbb{R}^T$ of $T$-dimensional topic distribution, and as well to associate each query $q_j$ with a vector $\theta_{q_j} \in \mathbb{R}^T$. The topic distribution can be obtained in many different ways. For example, in the paper–reviewer assignment problem, each reviewer can select their expertise topics from a predefined categories. In addition, we can use statistical topic modeling [35,36] to automatically extract topics from the input data. In this paper, we use the topic modeling approach to initialize the topic distribution of each expert and each query.

To extract the topic distribution, we can consider that we have a set of $M$ expert documents and $N$ query documents (each representing an expert or a query). An expert's document can be obtained by accumulating the content information related to the expert. For example, we combine all publication papers as the expert document of a reviewer, thus expert $v_i$'s document can be represented as $d_{v_i} = \{w_{ij}\}$. Each query can also be viewed as a document. Then we can learn these $T$ topic aspects from the collection of expert documents and query documents using a topic model such as LDA [36]. Specifically, let $D = \{d_{v_1}, \ldots, d_{v_M}\}$ be the set of experts' documents. The log-likelihood of the whole collection according to LDA is

$$\log p(D|\theta,\phi) = \sum_{d \in D} \sum_{w \in d} c(w,d) \log \left( \sum_{z=1}^{T} p(w|z,\phi_z) p(z|d,\theta_d) \right) \tag{11}$$

where $c(w,d)$ is the count of word $w$ in document $d$, $p(w|z,\phi_z)$ is the probability of topic $z$ generating word $w$, and $p(z|d,\theta_d)$ is the probability of document $d$ containing topic $z$.

We use the Gibbs sampling algorithm [37,38] to learn the topic distribution $\theta_{v_i}$ for each expert. The topic distribution of query $\theta_{q_j}$ can be accordingly inferred from the produced $\theta_{v_i}$.

### 4.3. Pairwise matching score

We employ a language model-based retrieval method to calculate the pairwise matching score. With language model, the matching score $R_{ij}$ between expert $v_i$ and query $q_j$ is interpreted as a probability $R_{ij}^{LM} = p(q_j|d_i) = \prod_{w \in q_j} p(w|d_i)$, where

$$p(w|d_i) = \frac{N_{d_i}}{N_{d_i}+\lambda_D} \cdot \frac{tf(w,d_i)}{N_{d_i}} + \left(1 - \frac{N_{d_i}}{N_{d_i}+\lambda_D}\right) \cdot \frac{tf(w,\mathbf{D})}{N_{\mathbf{D}}} \tag{12}$$

where $N_{d_i}$ is the number of word tokens in document $d_i$, $tf(w,d_i)$ is the number of occurring times of word $w$ in $d_i$, $N_{\mathbf{D}}$ is the number of word tokens in the entire collection, and $tf(w,\mathbf{D})$ is the number of occurring times of word $w$ in the collection $\mathbf{D}$. $\lambda_D$ is the Dirichlet smoothing factor and is commonly set according to the average document length in the collection [26].

Our previous work extended LDA and proposed the ACT model [39] to generate a topic distribution. By considering the learned topic model, we can define another matching score as

$$R_{ij}^{ACT} = p(q_j|d_i) = \prod_{w \in q_j} \sum_{z=1}^{T} p(w|z,\phi_z) p(z|d,\theta_{d_i}) \tag{13}$$

Further, we can define a hybrid matching score by combining the two probabilities together

$$R_{ij}^{H} = R_{ij}^{LM} \times R_{ij}^{ACT} \tag{14}$$

### 4.4. Optimization solving

In order to maximize the objective function (Eq. (10)), we construct a convex cost network with lower and upper bounds imposed on the arc flows. Fig. 2 illustrates the constructing process as described in Algorithm 1.[1] Convex cost flow problem can be solved by transforming to an equivalent minimum cost flow problem [40]. The minimum cost flow of the network gives an optimal assignment with respect to (Eq. (10)).

[1] Every arc in the network is associated with lower and upper bound denoted as [$l$, $u$] and a convex function of the arc flow $f$.



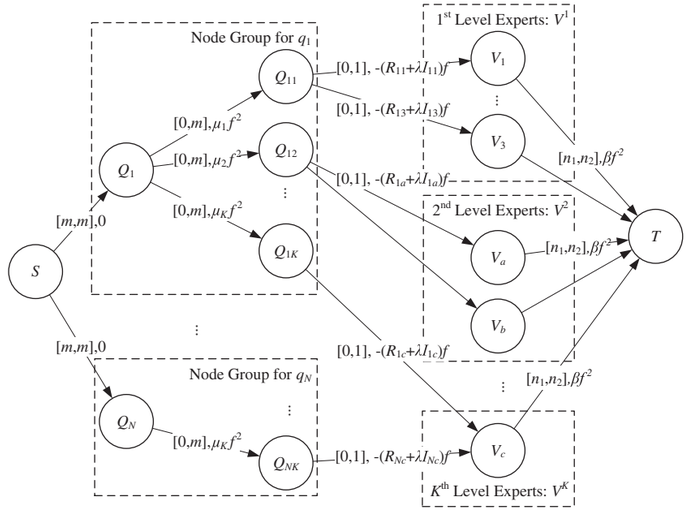



图2.根据客观函数（10）的凸起成本网络流的构造。

其中L，B和MK是Lagrangian乘法器，用于在目标函数中对不同组件的重视进行权衡。现在问题是如何定义主题分发y，如何计算成对匹配分数RIJ，以及如何优化框架。

### 4.2. Modeling multiple topics

主题建模的目标是将每个专家VI与T维主题分发的矢量YVI艺术相关联，也将每个查询QJ与矢量YQJ艺术相关联。该主题分布可以以多种不同的方式获得。例如，在纸张审阅者分配问题中，每个审阅者都可以从预定义的类别中选择他们的专业知识。此外，我们可以使用统计主题建模[35,36]来自动从输入数据中提取主题。在本文中，我们使用主题建模方法初始化每个专家和每个查询的主题分布。要提取主题分发，我们可以考虑我们有一组M个专家文档和N查询文档（每个代表专家或查询）。通过累积与专家有关的内容信息，可以获得专家的文件。例如，我们将所有出版物文件与审阅者的专家文件相结合，因此专家VI的文件可以代表DVI FWIJG。每个查询也可以被视为文档。然后，我们可以使用诸如LDA
[36]的主题模型来了解这些T主题方面。具体而言，让D FDV1，。。。，DVMG是一组专家文件。根据LDA的整个系列的日志可能性是

$$\log p(D|\theta,\phi) = \sum_{d \in D} \sum_{w \in d} c(w,d) \log \left( \sum_{z=1}^{T} p(w|z,\phi_z) p(z|d,\theta_d) \right) \tag{11}$$

其中C（w，d）是文档D中的单词W的计数，p w9z，fz 是主题z成字w的概率，并且p z9d，yd 是包含主题z的文档d的概率。我们使用GIBBS采样算法[37,38]以了解每个专家的主题分发YVI。查询YQJ的主题分布可以从生产的YVI推断出来。

### 4.3. Pairwise matching score

我们采用基于语言模型的检索方法来计算成对匹配分数。通过语言模型，Expert
VI和查询QJ之间的匹配分数RIJ被解释为概率RLM
IJ P qj9di qwaqjp w9di

$d_i$ $D$ $d_i$ $d_i$ $D$ $\mathbf{D}$

其中NDI是DID
DI，TF（W，DI）中的单词令牌的数量是DI中的单词W的发生时间，ND是整个集合中的单词令牌的数量，而TF（W，D）是收集D.
LD中的单词W的发生次数是Dirichlet平滑因子，并且通常根据集合中的平均文档长度进行设置[26]。我们以前的工作扩展了LDA并提出了该法案

[39]要生成主题分发。通过考虑学习的主题模型，我们可以将另一个匹配分数定义为

$$R_{ij}^{ACT} = p(q_j|d_i) = \prod_{w \in q_j} \sum_{z=1}^{T} p(w|z,\phi_z) p(z|d,\theta_{d_i}) \tag{13}$$

此外，我们可以通过将这两个概率组合在一起来定义混合匹配分数

$$R_{ij}^{H} = R_{ij}^{LM} \times R_{ij}^{ACT} \tag{14}$$

### 4.4. Optimization solving

为了最大化目标函数（EQ。（10）），我们构建一个凸起成本网络，其施加在电弧流上的下限和上限。图。图2示出了如算法1.1中所述构建过程1.1凸起成本流程问题可以通过转换为等同的最小成本流动问题来解决[40]。网络的最小成本流程对（EQ（10））提供了最佳分配。

1网络中的每个弧与弧形和上束相关联，表示为[l，u]和电弧流f的凸起函数。

**Algorithm 1.** Optimization solving algorithm.

**Input**: The set of experts $V$; the set of queries $Q$; the matching score matrix $R_{M\times N}$; the COI matrix $U_{M\times N}$; Number of expertise level $K$; $m$, $n_1$, $n_2$ as described above.
**Output**: An assignment of experts to queries maximizing objective function (10).

**1.1** Create a network $G$ with source node $S$ and sink node $T$;
**1.2** **foreach** $q_j \in Q$ **do**
**1.3** | Create $K+1$ nodes, denoted as $Q_j, Q_{j1}, \ldots, Q_{jK}$
**1.4** | respectively;
**1.5** | Add an arc from source node $S$ to node $Q_j$, with zero cost and flow constraint $[m,m]$; Add an arc from node $Q_j$ to $Q_{jk}$, with square cost function $\mu_k f^2$ and flow constraint $[0,m]$;
**1.6** **foreach** $v_i \in V$ **do**
**1.7** | Create a node $V_i$;
**1.8** | Add an arc from $V_i$ to sink node $T$, with square cost function $\beta f^2$ and flow constraint $[n_1,n_2]$;
**1.9** **foreach** $v_i \in V, q_j \in Q$, s.t. $U_{ij}=1$ **do**
**1.10** | $k=$ expert level of $v_i$;
**1.11** | Add an arc from $Q_{jk}$ to $V_i$, with linear cost function $-(R_{ij}-\lambda I_{ij})f$ and flow constraint $[0,1]$;
**1.12** Compute the minimum cost flow on $G$;
**1.13** **foreach** $v_i \in V, q_j \in Q$, s.t. $U_{ij}=1$ **do**
**1.14** | $k=$ expert level of $v_i$;
**1.15** | **if** *flow* $f(Q_{jk},V_i)=1$ **then** Assign query $q_j$ to expert $v_i$;

**Theorem 1.** *Algorithm 1 based on minimum convex cost flow gives an optimal solution.*

**Proof.** First the minimum convex cost flow problem (MCCF) can be formulated as the following optimization problem:

$$\text{Min} \sum_{(a,b)\in E(G)} C_{ab}(f(a,b))$$
$$\text{s.t.} \quad \forall a \in V(G), \sum_{b:(a,b)\in E(G)} f(a,b) = \sum_{b:(b,a)\in E(G)} f(b,a)$$
$$\forall (a,b) \in E(G), \quad l_{ab} \le f(a,b) \le u_{ab} \tag{15}$$

The model is defined on a directed network $G=(V(G),E(G))$ with lower bound $l_{ab}$, upper bound $u_{ab}$ and a convex cost function $C_{ab}(f(a,b))$ associated with every arc $(a,b)$.

Now we prove that minimizing (Eq. (15)) on the graph $G$ constructed in Algorithm 1 is equivalent to maximizing (Eq. (10)). For simplicity, we use $I_{ij}$ to denote $|T(q_j)\cap T(v_i)|/|T(q_j)|$. For the constructing process, we see a feasible flow on $G$ is mapping to a query–expert assignment. The flow from $S$ to $Q_j$ indicates the number of experts assigned with query $q_j$, and the flow from $V_i$ to $T$ indicates the number of queries assigned to expert $v_i$. And the cost between $V_i$ and $T$ is corresponding to the *load balance* soft penalty function (Eq. (4)). The meaning of the flow from $Q_j$ to $Q_{jk}$ is the number of $k$th-level experts assigned to $q_j$, thus we impose a square cost function $\mu_k \cdot f^2$ on the arcs which is equivalent to the negative of the *authority balance* penalty. The flow from $Q_{jk}$ to $V_i$ means we assign query $q_j$ to expert $v_i$, it is easy to find that no query will be assigned to the same expert twice since we give an upper bound of 1 on the arc, while the cost is equivalent to the negative of matching score and topic average confidence score. Therefore, our problem can be reduced to an equivalent MCCF problem, where the objective function of MCCF problem (Eq. (15)) is the negative form of (Eq. (10)).

In practice, it is not necessary to add all $(Q_{jk},V_i)$ arcs. To further reduce the complexity of the algorithm, we first greedily generate an assignment and preserve corresponding arcs, then keep only $c\cdot m$ arcs for $Q_{jk}$ and $c\cdot n_2$ arcs for $V_i$ which have highest matching score ($c$ is a fixed constant). We call this process *Arc-Reduction*, which will reduce the number of arcs in the network without influencing the performance too much. To process large scale data, we can leverage the parallel implementation of convex cost flow [41].

### 4.5. Online matching

After an automatic expertise matching process, the user may provide feedbacks. Typically, there are two types of user feedbacks: (1) pointing out a false match; (2) specifying a new match. Online matching aims to adjust the matching result according to the user feedback. One important requirement is how to perform the adjustment at real time. In our framework, we provide online interactive adjustment without recalculating the whole cost flow. For both types of feedbacks, we can easily accomplish online adjustment by canceling some flows and augmenting new assignments in our framework. We give Algorithm 2 to consider the first type of feedback, which still produces an optimal solution.

**Algorithm 2.** Online matching algorithm.

**Input**: A minimum cost network flow $f$ on $G$ corresponding to the current assignment; an inappropriate match $(v_i,q_j)$.
**Output**: A new assignment.

**2.1** $k=$expert level of $v_i$;
**2.2** **if** $f(Q_{jk},V_i)=1$ **then**
**2.3** | Construct the residual network $G(f)$;
**2.4** | Compute the shortest path $P_{back}$ from $T$ to $S$ on $G(f)$ which contains backward arc $(V_i,Q_{jk})$;
**2.5** | Cancel(rollback) 1 unit of flow along $P_{back}$ and update $G(f)$;
**2.6** | Remove arc $(Q_{jk},V_i)$ from $G$ and update $G(f)$;
**2.7** | Compute shortest augmenting path $P_{aug}$ from $S$ to $T$;
**2.8** | Augment 1unit of flow along $P_{aug}$;

**Lemma 1** (*Negative cycle optimality conditions*). *Ahuja et al. [40] A feasible solution $f^*$ is an optimal solution of the minimum cost flow problem if and only if it satisfies the negative cycle optimality conditions*: *namely, the residual network $G(f^*)$ contains no negative cost cycle.*

**Theorem 2.** *Algorithm 2 produces an optimal solution in the network without assignment* $(q_j, v_i)$.

**Proof.** According to Lemma 1, the residual network $G(f)$ contains no negative cost cycle since the given flow $f$ has the minimum cost. In Algorithm 2, we remove the inappropriate match $(v_i,q_j)$ and adjust the network flow in line (2.3)–(2.5). Denote the feasible flow in the network after line (2.5) as $f'$. According to the SAP (short augmenting path) algorithm of cost flow, if $f'$ has the minimum cost(i.e., $G(f')$ contains no negative cycle), the algorithm will give the optimal solution. We show the optimality of $f'$ by contradiction. Assume $G(f')$ contains a negative cycle $C$, $C$ must intersect with the shortest path $P_{back}$ computed online (2.3), since the original $G(f)$ contains no negative cycle. Thus merging $C$ into path $P_{back}$ will generate a shorter path, which contradicts with the assumption that $P_{back}$ is shortest. Therefore, $f'$ has the minimum



算法1.优化求解算法。

输入：专家v;查询Q;匹配得分矩阵Rm n;COI矩阵UM n;专业知识级别k;如上所述的M，N1，N2。产出：向询问专家的分配最大化客观函数（10）。1.1使用源节点S和宿节点T创建网络G;1.2 foreach qj aq做1：3 1：4 1：5

创建k 1节点，表示为qj，qj1，。。。，QJK分别;从源节点S向节点Qj添加一个弧，零成本和流量约束 m，m ;从节点QJ到QJK的电弧，方形成本函数MKF 2和流量约束 0，m ;

1.6 FOREACH VI AV DO 1：7 1：8创建节点VI;将VI的电弧添加到沉积节点T，方形成本函数BF 2和流量约束 n1，n2 ;

1.9 FOREACH VI AV，QJ AQ，S：T：UIJ 1做1:10 1:11 K vi的专家水平;从QJK向VI添加电弧，线性成本函数 rij liij和流量约束 0,1 ;

1.12计算g的最小成本流量;1.13 FOREACH VI AV，QJ AQ，S：T：UIJ 1做1:14 1:15 K vi的专家水平;如果流f qjk，则vi 1然后将查询qj分配给专;

定理1.基于最小凸起成本流的算法1提供了最佳解决方案。

MCCF问题的功能（EQ。（15））是（EQ.（10））的负形式。在实践中，没有必要添加所有（QJK，VI）弧。为了进一步降低算法的复杂性，我们首先贪婪地生成分配并保留相应的弧，然后只能保持QJK的C M弧和用于VI的弧度，具有最高匹配得分（C是固定常数）。我们称此过程还原，这将减少网络中的弧数而不会影响性能太大。为了处理大规模数据，我们可以利用凸起成本流的并行实现[41]。

### 4.5. Online matching

在自动专业知识匹配过程之后，用户可以提供反馈。通常，有两种类型的用户馈回：（1）指出假匹配;（2）指定新匹配。在线匹配旨在根据用户反馈调整匹配结果。一个重要要求是如何实时进行调整。在我们的框架中，我们提供在线交互式调整，而无需重新计算整个成本流程。对于这两种类型的反馈，我们可以通过取消某些流程并在框架中增强新分配来轻松完成在线调整。我们给予

算法2考虑第一种类型的反馈，它仍然产生最佳解决方案。

算法2.在线匹配算法。

输入：G上的最小成本网络流量F对应于当前分配;不当匹配（vi，qj）。输出：新分配。2.1k pertvi水平;2.2如果f qjk，vi 1那么2：3 2：4

构建残余网络G F ;计算来自T到S的最短路径PBack，其中包含后向弧 vi，qjk ;CANCELINGRALLBACK 1沿迎战的流量单位，更新G F ;从g中删除arc qjk，vi 和更新g f ;计算到t的最短增强路径波动;沿着Phug的流量增加1个;

该模型在具有下限实验室，上限UAB和凸起成本函数Cab f a的定向网络G （V（g），e（g））上定义，与每个弧（a，b）相关联。现在我们证明，在算法1中构造的图表G上最小化（EQ（15））相当于最大化（EQ。（10））。为简单起见，我们使用IIJ表示9t qj \t vi 9=9t qj 9。对于构建过程，我们看到G上的可射到查询专家分配。从S到QJ的流量表示查询Qj分配的专家数量，并且从VI到T的流量表示分配给专家VI的查询数。并且VI和T之间的成本与负载平衡软罚函数（Eq。（4））相对应。从QJ到QJK的流量的含义是分配给QJ的kth级专家的数量，因此我们在弧上强加一个平方成本函数mk f 2，相当于权威余额惩罚的负面。从QJK到VI的流程意味着我们将查询QJ分配给Expert VI，很容易发现没有查询将被分配给同一专家两次，因为我们在弧上提供1的上限，而成本相当于匹配分数和主题平均置信度得分。因此，我们的问题可以减少到相同的MCCF问题，其中目标

引理1（负周期最优性条件）。Ahuja等。一种可行的解决方案Fn是且仅当它满足负周期最优性条件时，最小成本流动问题的最佳解决方案是：即，残差网络G（Fn）不包含负成本循环。

定理2.算法2在没有分配的情况下在网络中产生最佳解决方案（QJ，VI）。

证明。根据LEMMA 1，由于给定的流F具有最小成本，因此残余网络G（F）不包含负成本循环。在算法2中，我们删除不当匹配（VI，QJ），并在线（2.3）－（2.5）调整网络流量。在线（2.5）作为F 0.根据成本流量的SAP（短增强路径）算法，如果F 0具有最小成本（即，G F0 不包含负周期），则该算法将提供最佳解决方案。我们通过矛盾显示F 0的最优性。假设G F01包含负周期C，C必须与在线计算的最短路径送（2.3）相交，因为原始g（f）不包含负周期。因此，将C合并到路径送方案将产生较短的路径，这与假设PBACK最短的假设相矛盾。因此，F 0具有最小值

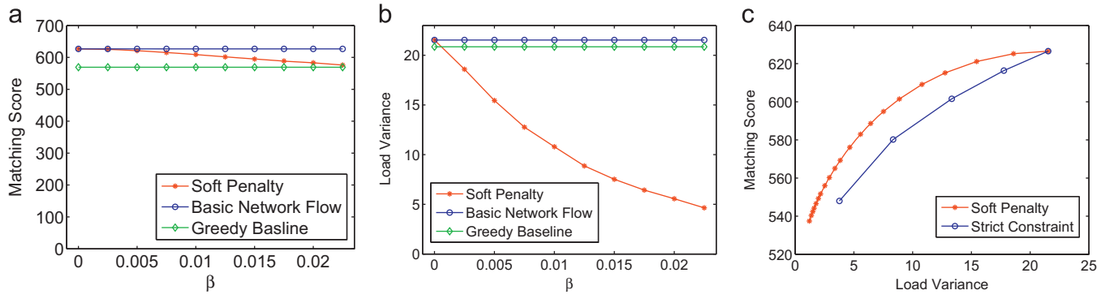



**Fig. 3.** (a) and (b) illustrate how soft penalty function influences the matching score (MS) and load variance with different $\beta$ respectively. (c) gives a comparison between soft penalty function and strict constraint methods towards load balance.

cost. Accordingly, Algorithm 2 gives the optimal solution after augmenting the new assignment. □

## 5. Experimental results

The proposed approach for expertise matching is very general and can be applied to many application to align experts and queries. We evaluate the proposed framework on two different genres of expertise matching problems: paper–reviewer assignment and course–teacher assignment. Three experiments on different datasets are conducted to show the effectiveness of the proposed method. All datasets, code, and detailed results are publicly available.[2] All the experiments are carried out on a PC running Windows XP with Intel Core2 Quad CPU Q9550 (2.83 GHz), 4G RAM.

### 5.1. Paper–reviewer assignment experiment

*Dataset*. The paper–reviewer dataset consists of 338 papers and 354 reviewers. The reviewers are the program committee members of KDD'09 and the 338 papers are those published on KDD'08, KDD'09, and ICDM'09. For each reviewer, we collect her/his all publications from academic search system Arnetminer[3] [42] to generate the expertise document. As for the COI problem, we generate the COI matrix $U$ according to the coauthor relationship in the last five years and the organization they belong to. Finally, we set that a paper should be reviewed by $m=5$ experts, and an expert at most reviews $n_2=10$ papers.

*Baseline methods and evaluation metrics*. We employ a greedy algorithm as the baseline. The greedy algorithm assigns experts with highest matching score to each query, while keeping the load balance for each expert (i.e., $|Q(v_i)| \le n_2$) and avoiding the conflict-of-interest.

As there are no standard answers, in order to quantitatively evaluate our method, we define the following metrics:

*Matching score* (*MS*): It is defined as the accumulative matching score.

$$MS = \sum_{v_i \in V} \sum_{q_j \in Q(v_i)} U_{ij} R_{ij}$$

*Load variance* (*LV*): It is defined as the variance of the number of papers assigned to different reviewers.

$$LV = \sum_{i=1}^{M} \left( |Q(v_i)| - \frac{\sum_{i=1}^{M} |Q(v_i)|}{M} \right)^2$$

*Expertise variance* (*EV*): It is defined as the variance of the number of top level reviewers assigned to different papers.

$$EV = \sum_{j=1}^{N} \left( |V(q_j) \cap V^1| - \frac{\sum_{j=1}^{N} |V(q_j) \cap V^1|}{N} \right)^2$$

*Results*. In this experiment, we tune different parameters to analyze the influence on the accumulative matching score. We also evaluate the efficiency of our proposed approach.

We first set $\mu = \mathbf{0}$ and tune the parameter $\beta$ to find out the effects of soft penalty function. Fig. 3(a) illustrates how soft penalty function influences the matching score with different $\beta$. We see that the matching score decreases slightly with $\beta$ increasing. Fig. 3(b) shows the effects of load variance with $\beta$ varied. We see that the load variance changes very fast toward balance.

In Fig. 3(c), we compare the two different methods to achieve load balance, namely, strict constraint and soft penalty. The two LV–MS curves are respectively generated by setting different minimum numbers $n_1$ for strict constraint and varying the weight parameter $\beta$ for soft load balance penalty. The curves show that soft penalty outperforms strict constraint towards load balance.

Then we set $\beta$ to 0 to test the effects of authority balance. Experts are divided into two levels base on their *H*-index, and we set $\mu_2 = 0$ to consider the balance of the senior reviewers only. Fig. 4 presents the accumulative matching score (a) and expertise variance (b) with $\mu_1$ varied.

Further, we analyze the effects of different constraints. Specifically, we first remove all constraints (using Eq. (1) only), and then add the constraints one by one in the order (load balance, authority balance, and COI). In each step, we perform expertise matching using our approach. Table 2 lists the accumulative matching score obtained in each step. We see that the load balance constraint will reduce the expertise matching score, while the other constraints have little negative effect. This is because senior experts are often good at many aspects (the matching score between them and many queries are large), thus assigned with heavy load in traditional matching. In our method, we try to get a more reasonable assignment by adding load balance constraint, which will restrict the work load of those senior experts. As a result, the matching score decreases.

To clearly illustrate the effect of load-balance constraint, we present Fig. 5, in which we see that traditional information-retrieval based method assigns many papers to senior reviewers, while some reviewers do not get any work. The load-balance constraint is necessary to generate a reasonable matching.

Finally, we evaluate the efficiency performance of the proposed algorithm. We compare the CPU time of the original optimal algorithm and the version with *Arc-Reduction*. As shown in Fig. 6, the *Arc-Reduction* process can significantly reduce the time consumption.

[2] http://www.arnetminer.com/expertisematching
[3] http://arnetminer.org

W. Tang et al. / Neurocomputing 76 (2012) 71–83 77

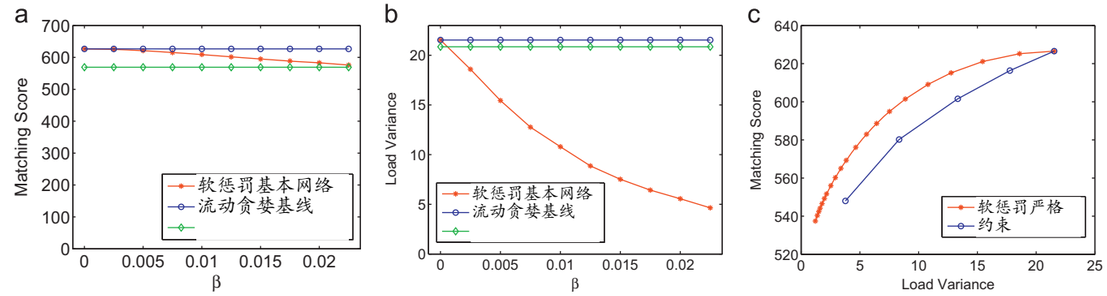



图3.（a）和（b）示出了软惩罚功能如何影响匹配分数（MS）和与不同B的负载方差。（c）在软罚函数和严格的负载平衡方面进行了比较。

成本。因此，算法2在增强新分配后给出最佳解决方案。&

## 5. Experimental results

所提出的专业匹配方法非常一般，可以应用于许多应用程序来对准专家和查询。我们评估了两种不同类型的专业知识匹配问题的拟议框架：纸张－审阅者分配和课程－教师分配。进行了在不同数据集上进行的三个实验以显示所提出的方法的有效性。所有数据集，代码和详细结果都是公开可用的。

### 5.1. Paper–reviewer assignment experiment

数据集。纸张审阅者数据集由338篇论文和354名审稿人组成。审阅者是KDD'09的计划委员会，338篇论文是KDD'08，KDD'09和ICDM'09发表的文件。对于每个审阅者，我们从学术搜索系统ARNetminer3
[42]中收集她/他的所有出版物，以产生专业知识文件。至于COI问题，我们根据过去五年的同志关系和他们所属的组织生成COI矩阵U.最后，我们设定了一篇论文，由M 5专家审查，以及大多数评论的专家N2 10论文。基线方法和评估指标。我们使用贪婪的算法作为基线。贪婪算法将具有最高匹配分数的专家分配给每个查询，同时保持每个专家的负载余额（即，9q vi 9rn2），并避免兴趣的混合。由于没有标准答案，为了定量评估我们的方法，我们定义了以下度量：匹配分数（MS）：它被定义为累积匹配分数。

$$MS = \sum_{v_i \in V} \sum_{q_j \in Q(v_i)} U_{ij} R_{ij}$$

加载方差（LV）：它被定义为分配给不同审阅者的论文数量的方差。

$$LV = \sum_{i=1}^{M} \left( |Q(v_i)| - \frac{\sum_{i=1}^{M} |Q(v_i)|}{M} \right)^2$$

专业方案（EV）：它被定义为分配给不同论文的顶级审稿人数的差异。

$$EV = \sum_{j=1}^{N} \left( |V(q_j) \cap V^1| - \frac{\sum_{j=1}^{N} |V(q_j) \cap V^1|}{N} \right)^2$$

结果。在这个实验中，我们调整了不同的参数来分析对累积匹配分数的影响。我们还评估了我们提出的方法的效率。我们首先设置m∈0并调整参数b以找出软罚函数的效果。图。图3（a）示出了软惩罚功能如何影响匹配分数与不同的b。我们看到匹配分数略微减少，B增加。图。图3（b）显示了与B变化的负载方差的影响。我们看到负载方差变化非常快，平衡。在图3（c）中，我们比较两种不同的方法来实现负载平衡，即严格的限制和软罚性。两个LV−MS曲线分别通过设置不同的最小数量N1来产生用于严格约束并改变重量参数B进行软载余额惩罚。曲线表明，软惩罚优于负载平衡的严格限制。然后我们将b设置为0以测试权威余额的影响。专家分为他们的H−Index的两个层面，我们设置了M 2 0，以考虑仅考虑高级审阅者的余额。图4显示了M1变化的累积匹配得分（A）和专业知识方差（B）。此外，我们分析了不同约束的影响。具体而言，我们首先删除所有约束（仅使用Eq。（1）），然后按顺序（负载余额，权限余额和COI）逐个添加约束。在每一步中，我们使用我们的方法执行专业知识匹配。表2列出了每个步骤中获得的累积匹配分数。我们看到负载余额约束将减少专业匹配分数，而其他约束则具有很小的负面影响。这是因为在许多方面（它们之间的匹配分数和许多查询都是很大的），因此在传统匹配中分配了重负荷。在我们的方法中，我们尝试通过添加负载余额约束来获得更合理的分配，这将限制这些高级专家的工作负荷。结果，匹配得分降低。为了清楚地说明负载平衡约束的影响，我们存在图5，我们看到传统信息检索的方法为高级审阅者分配许多文件，而一些审稿人则没有任何工作。负载余额约束是生成合理匹配的必要条件。最后，我们评估了所提出的算法的效率性能。我们比较原始最佳算法的CPU时间和带有弧度的版本。如图1所示。如图6所示，弧度减少过程可以显着降低时间消耗。

[2] http://www.arnetminer.com/expertisematching
[3] http://arnetminer.org

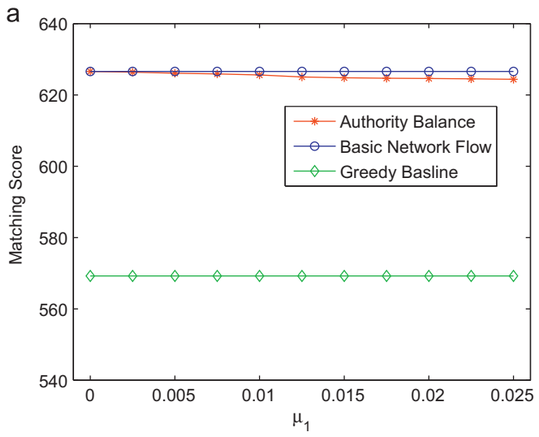



**Fig. 4.** Matching score (MS) and expertise variance (EV) with $\mu_1$ varied.

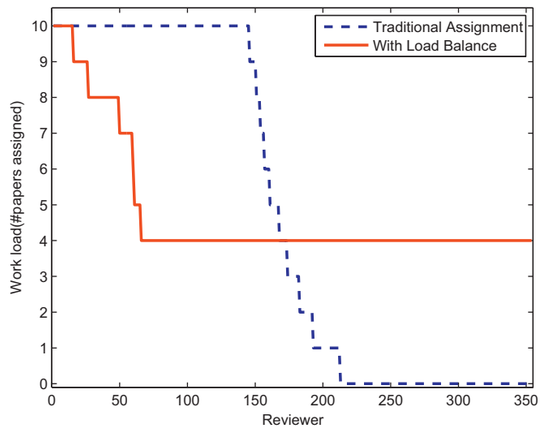



**Fig. 5.** The work load (number of paper assigned) of every reviewer.

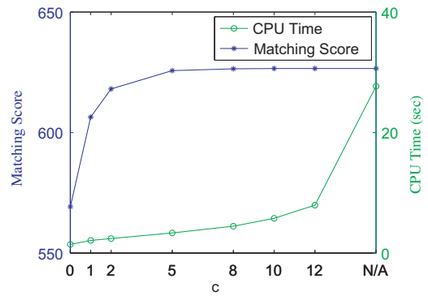



**Fig. 6.** Efficiency performance (s).

**Table 2**
Effects of different constraints on matching score.

| Constraint | Matching score |
|---|---|
| Basic objective function (Eq. (1)) | 635.51 |
| + Load balance *soft penalty* with $\beta = 0.02$ | 592.83 |
| + Authority balance with $\mu = (0.02,0)^T$ | 599.37 |
| + COI | 590.14 |

For example, when setting $c=12$ in this problem, we can achieve a $>3\times$ speedup without any loss in matching score.

We further use a case study (as shown in Tables 3 and 4) to demonstrate the effectiveness of our approach. We see that the result is reasonable. For example, Lise Getoor, whose research interests include relational learning, is assigned with a lot of papers about social network.

### 5.2. Multi-topic paper–reviewer assignment experiment

*Dataset*. We use another dataset (D2) to verify the performance on "topic coverage". The dataset D2 is provided by [3],[4] consisting of 73 queries and 189 reviewers. The 73 queries are paper abstracts from SIGIR'07, where each of them is related to at least two topics. The documents of reviewers are the combination of all

[4] http://timan.cs.uiuc.edu/data/review.html

**Table 3**
Example assigned papers to three reviewers.

| Reviewer | Assigned papers |
|---|---|
| Lise Getoor | Evaluating statistical tests for within-network classifiers of…<br>Discovering organizational structure in dynamic social network<br>Connections between the lines: augmenting social networks with text<br>MetaFac: community discovery via relational hypergraph factorization<br>Relational learning via latent social dimensions<br>Influence and correlation in social networks |
| Wei Fan | Mining data streams with labeled and unlabeled training examples<br>Vague one-class learning for data streams<br>Active selection of sensor sites in remote sensing applications<br>Name-ethnicity classification from open sources<br>Consensus group stable feature selection<br>Categorizing and mining concept drifting data streams |
| Jie Tang | Co-evolution of social and affiliation networks<br>Influence and correlation in social networks<br>Feedback effects between similarity and social influence…<br>Mobile call graphs: beyond power-law and lognormal distributions<br>Audience selection for online brand advertising: privacy-friendly… |

78 W. Tang et al. / Neurocomputing 76 (2012) 71–83

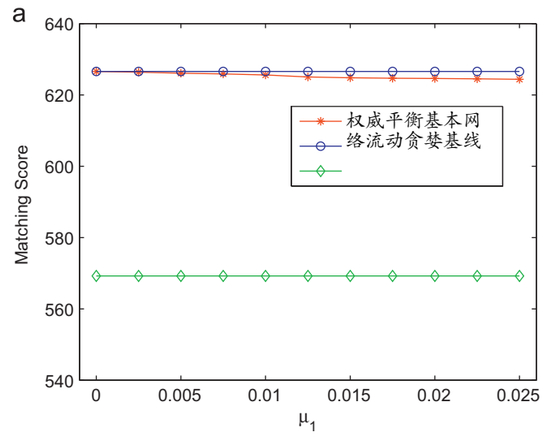



图4.匹配分数（MS）和专业知识方差（EV）各种不同。

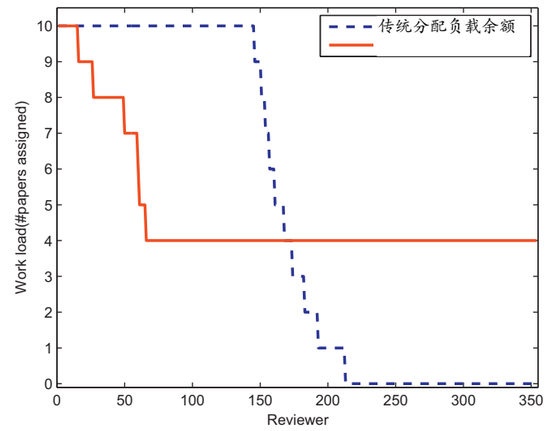



**Fig. 5.** The work load (number of paper assigned) of every reviewer.

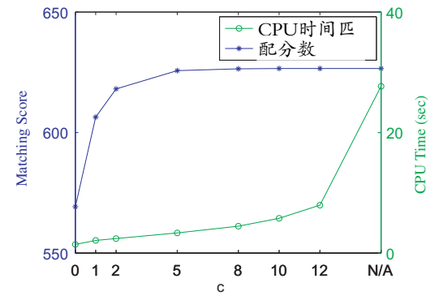



**Fig. 6.** Efficiency performance (s).

表2不同约束对匹配分数的影响。

| Constraint | Matching score |
|---|---|
| 基本目标函数（eq。（1））635.51　负载平衡软惩罚与b<br>0:02 592.83<br>+　　=(　　) | |
| + COI | 590.14 |

例如，当设置C　12在这个问题中，我们可以在没有任何匹配分数的情况下实现43° 的加速。我们进一步使用案例研究（如表3和4所示）来证明我们方法的有效性。我们看到结果是合理的。例如，Lise
Getoor，其研究兴趣包括关系学习，被分配有很多关于社交网络的论文。

表3示例将文件分配给三位审阅者。

| Reviewer | Assigned papers |
|---|---|
| Lise评估网络内的统计学测试，以便在线的动态社交网络连接中发现组织结构：通过文本METAFAC的增强社交网络：通过关系超照片分解关系学习通过潜在社会网络的影响和社交网络中的相关性 | |
| Wei粉丝挖掘数据流具有标记和未标记的培训示例模糊单级学习数据流中的遥感应用中的传感器站点的活动选择来自开放消息来源的种族分类，共识组稳定特征选择分类和采矿概念漂移数据流 | |
| 杰唐社会和隶属网络的共同演变在社交网络中的影响和相关性，相似性和社会影响力的反馈效应：超越幂律和逻辑正式发行版在线品牌广告的观众选择：隐私友好 | |

### 5.2. Multi-topic paper–reviewer assignment experiment

数据集。我们使用另一个数据集（D2）来验证“主题覆盖”上的性能。数据集D2由[3]，其中4个由73个查询和189名审阅者组成。73个查询是Sigir'07的纸质摘要，其中每个主题与至少两个主题有关。审稿人的文件是所有人的结合

abstracts published in SIGIR1971-2006 by the corresponding reviewers. In this dataset, 25 topics such as "text mining", "clustering" and "language models" are identified by human experts, and the related topics of each query and each reviewer are manually labeled as a gold standard for evaluation. We use the Lemur toolkit [43] for pre-processing to tokenize each query and document, removing common stop words.

*Settings.* Since the authors and reviewers in the dataset are anonymous, we do not consider COI and authoritative balance in this experiment. In order to be consistent with the setting in [3], we set a paper to be reviewed by $m=3$ experts, and use *Coverage* (Eq. (6)), *Confidence* (Eq. (7)), *AverageConfidence* (Eq. (8)), and $F_{score}$ as the measures. In all the methods, the top topics with largest $\theta$ value are selected as the related topics for each query and expert, respectively.

*Comparison methods.* We implement three methods in [3] as comparison methods. The first baseline approach use a language model to retrieve reviewer documents for each query. Specifically, given a query $q_j$, one way to rank reviewers is to use the probability of query $q_j$ given reviewer document $d_i$, i.e.,

$$p(q_j|d_{v_i}) = \prod_{w \in q_j} \frac{tf(w,d_{v_i}) + \mu p(w|D)}{N_{d_{v_i}} + \mu} \tag{16}$$

where $N_{d_{v_i}}$ is the document size of $d_{v_i}$, $tf(w,d_{v_i})$ is the number of occurring times of word $w$ in $d_{v_i}$, $p(w|D)$ is the unigram language model of the entire reviewer document collection and $\mu$ is a Dirichlet smoothing factor which empirically set to 1000. In addition, we use a smoothed version of $p(w|D)$ so it would not return zero when word $w$ is unseen in the entire document collection, i.e.,

$$p(w|D) = \frac{tf(w,D)+1}{N_D + V} \tag{17}$$

where $V$ is the estimated vocabulary size (i.e., the number of distinct words).

The second baseline method uses KL-divergence for reviewer retrieval [26]:

$$\sum_{w \in q_j, tf(w,d_i)>0} p(w|q_j) \log\left(1 + \frac{tf(w,d_i)}{\mu p(w|D)}\right) + \log \frac{\mu}{\mu + N_{d_i}} \tag{18}$$

We select the best method proposed in [3] as the third comparison method. The intuition of this method is to select a group of reviewers that work together to achieve a most similar topic distribution to the topics of a query. Specifically, given a paper query, this method picks reviewers one by one by choosing new reviewer with minimum KL-divergence value:

$$D(\theta_q || \theta^{r_k}_{r_1,\ldots,r_{k-1}}) \tag{19}$$

where $\theta_q$ is the topic distribution of the query and $\theta^{r_k}_{r_1,\ldots,r_{k-1}}$ is an averaged topic distribution of previously picked reviewers $r_1,\ldots,r_k$, i.e.,

$$p(z|\theta^{r_k}_{r_1,\ldots,r_{k-1}}) = \frac{\sigma}{k-1} \sum_{i=1}^{k-1} p(z|\theta_{r_i}) + (1-\sigma) p(z|\theta_{r_k}) \tag{20}$$

where $\sigma$ indicates how much to rely on the previously picked reviewers $r_1,\ldots,r_{k-1}$.

The three methods will be referred as Baseline-Pr, Baseline-KL and Topic-KL in following discussion.

*Matching with responsible topics.* Maximizing the global *Coverage* actually is an NP-hard set cover problem. Thus we add an assumption to make it tractable, in which for a specific query $q_j$, each assigned expert $v_i \in V(q_j)$ can select only one *responsible topic* $\hat{T}_{q_j}(v_i) \in T(v_i)$, and covers this topic for the query. The optimization problem with the assumption provides a lower bound of the original one, and can be easily incorporated into our framework. As shown in Fig. 7, we create $|T_{q_j}|+1$ nodes for a query $q_j$, where $Q_j$ is corresponding to the query, $Q_{jk}$ represents the $k$-th related topic of $q_j$. We add two arcs from $Q_j$ to $Q_{jk}$ for every $k$, with costs of $-\sigma\theta_{jk}$ and 0 respectively. The costs of the arcs is mapping to the topic coverage, where $\sigma$ is a multiplier used to adjust the importance. Then, for each expert $v_i$ whose related topics include $k$, we add an arc from $Q_{jk}$ to node $V_i^j$, which means the expert is able to be responsible for the topic. $V_i^j$ is then connected to the expert node $V_i$, where node $V_i^j$ is used to avoid the query to be assigned to the same expert twice. Combining with the matching

**Table 4**
List of reviewers for five random papers.

| Paper | Assigned reviewers |
|---|---|
| Audience selection for online brand advertising: privacy-friendly social network targeting | C. Lee Giles, Jie Tang, Matthew Richardson, Hady Wirawan Lauw, Elena Zheleva |
| Partitioned logistic regression for spam filtering | Rong Jin, Chengxiang Zhai, Saharon Rosset, Masashi Sugiyama, Annalisa Appice |
| Structured learning for non-smooth ranking losses | Xian-sheng Hua, Tie-yan Liu, Hang Li, Yunbo Cao, Lorenza Saitta |
| Unsupervised deduplication using cross-field dependencies | Chengxiang Zhai, Deepak Agarwal, Max Welling, Donald Metzler, Oren Kurland |
| The structure of information pathways in a social communication network | C. Lee Giles, Wolfgang Nejdl, Melanie Gnasa, Michalis Faloutsos, Cameron Marlow |

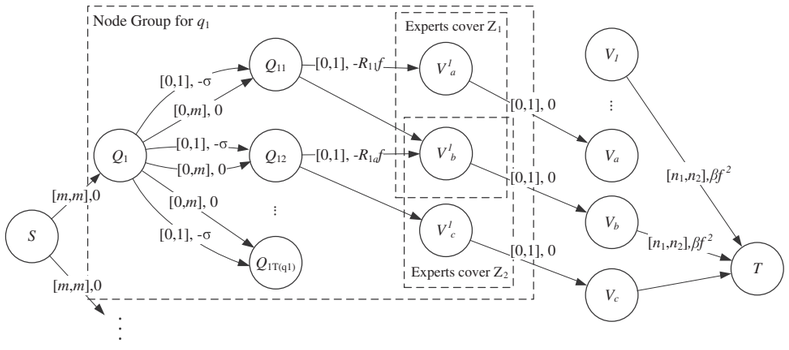



**Fig. 7.** The configuration of the network for matching with responsible topics.

W. Tang et al. / Neurocomputing 76 (2012) 71–83 79

摘要由相应的审稿人在Sigir1971−2006发表。在这个数据集中，25个主题，如“文本挖掘”，“群集”和“语言模型”的人员专家，以及每个查询的相关主题和每个评论员都被手动标记为金标准进行评估。我们使用Lemur Toolkit [43]进行预处理以授权每个查询和文档，删除常见的停止单词。设置。由于数据集中的作者和审核人员是匿名的，因此我们在本实验中不考虑COI和权威余额。为了与[3]中的设置一致，我们设定了由M 3专家审查的纸张，并使用覆盖范围（EQ。（6））， 信心（EQ。（7）），AverageConfence（EQ（8））和FScore作为措施。在所有方法中，分别选择具有最大Y值的顶主题作为每个查询和专家的相关主题。比较方法。我们在[3]中实施三种方法作为比较方法。第一个基线方法使用语言模型来检索每个查询的审阅者文档。具体地，给定查询QJ，排名审阅者的一种方式是使用查询QJ给定审阅者文档di的概率，即，

$$p(q_j|d_{v_i}) = \prod_{w \in q_j} \frac{tf(w,d_{v_i}) + \mu p(w|D)}{N_{d_{v_i}} + \mu} \tag{16}$$

如果NDVI是DVI的文档大小，TF W，DVI 是DVI中WORD W的发生时间的数量，P w9d 是整个评论者文档收集的UNIGRAM语言模型，M是一个虚拟设置到1000的Dirichlet平滑因子。在另外，我们使用P w9d的平滑版本，所以当Word W在整个文档中看不见时，它不会返回零

表4五个随机纸的审稿人列表。

| Paper | Assigned reviewers |
|---|---|
| C. Lee Giles，Jie Tang，Matthew Richardson，Hady Wiawan Lauw，Elena Zheleva垃圾邮件过滤后勤回归Rong Jin，Chengxiang Zhai，Saharon Rosset，Masashi Sugiyama，Annalisa Appice | |
| Surtucture学习非平滑排名亏损xian−sheng hua，刘铁燕刘，云博曹，洛伦加塞纳无监督重复数据删除采用跨场依赖性诚信翟，深斋瓦尔·瓦尔，Max康策，唐纳德·梅兹勒，奥伦库尔兰社交通信网络中的信息途径结构。李吉尔德，沃尔夫冈Nejdl，Melanie Gnasa，Michalis Faloutsos，Cameron Marlow | |

collection, i.e.,

$$p(w|D) = \frac{tf(w,D)+1}{N_D + V} \tag{17}$$

其中V是估计的词汇大小（即，不同词语的数量）。第二个基线方法使用KL分歧进行审阅者检索[26]：

$$\sum_{w \in q_j, tf(w,d_i)>0} p(w|q_j) \log\left(1 + \frac{tf(w,d_i)}{\mu p(w|D)}\right) + \log \frac{\mu}{\mu + N_{d_i}} \tag{18}$$

我们选择[3]中提出的最佳方法作为第三比较方法。这种方法的直觉是选择一组审阅者，该审阅者共同努力实现对查询主题的最相似的主题分布。具体地，给定纸质查询，该方法通过选择具有最小kl分歧值的新审阅者选择审阅者：

$$D(\theta_q || \theta^{r_k}_{r_1,\ldots,r_{k-1}}) \tag{19}$$

其中YQ是查询和YRK R1的主题分布，.....，RK 1是先前选择的评论者R1，Y，RK，即，即，

$$p(z|\theta^{r_k}_{r_1,\ldots,r_{k-1}}) = \frac{\sigma}{k-1} \sum_{i=1}^{k-1} p(z|\theta_{r_i}) + (1-\sigma) p(z|\theta_{r_k}) \tag{20}$$

其中S表示依赖先前选择的审稿人R1，Y，RK 1。在下面的讨论中，这三种方法将被称为基线−PR，基线−KL和主题−KL。与负责任主题相匹配。最大化全球覆盖率实际上是一个np−hard set涵盖问题。因此，我们添加了一个假设使其易于，其中对于特定的查询QJ，每个分配的专家vi qj 只能选择一个负责主题` tqj vi at vi ，并介绍了查询的本主题。假设的优化问题提供了原始的优化问题，并且可以轻松地结合到我们的框架中。如图7所示，我们为查询QJ创建9TQJ9 1节点，其中QJ对应于查询，QJK表示QJ的第k相关主题。我们将两个弧从QJ添加到QJK，每k，分别具有 syjk和0的成本。弧的成本正在映射到主题覆盖范围，其中S是用于调整重要性的乘数。然后，对于相关主题包括k的每个专家VI，我们将QJK添加到节点v J i的弧，这意味着专家能够负责主题。 V J I连接到专家节点VI，其中节点v j i用于避免查询分配给同一专家两次。结合匹配

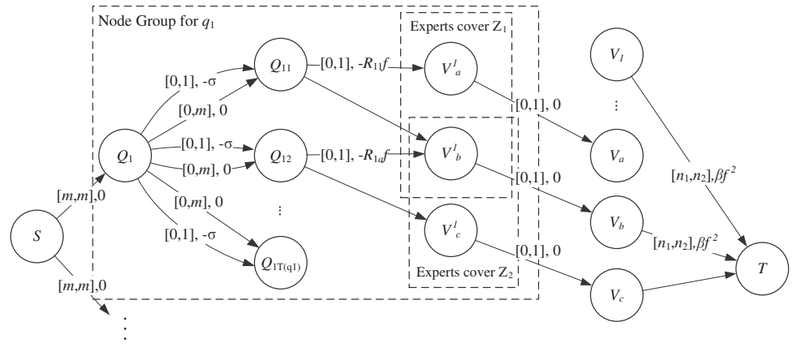



图7.与负责主题匹配的网络配置。

score and the load balance constraint, we get the final configuration as shown Fig. 7.

*Results*. We exam our approach both with and without load balance constraint. The results of all methods on dataset D2 is presented in Table 5, from which we see that both our approach and Topic-KL outperform the two baseline methods on *Coverage*, *Confidence*, *AverageConfidence* and $F_{score}$. Interestingly, the baseline methods achieve better *Confidence*. This may be because the assigned reviewers of a paper are more likely to cover the same topic many times. Our approach without load balance achieves comparable performance with Topic-KL. It's noticeable when adding load balance constraint, the proposed method still produces good results, which validates the usefulness of our approach. In Table 5, we use our approach-L10, our approach-L5 to represent the proposed method with different load balance settings, i.e., $n_2=10$ and $n_2=5$.

*Parameter sensitivity*. Now we investigate the influences of the parameters in the framework. We first consider the multiplier $\sigma$, which controls the importance of *Coverage* score. When setting $\sigma=0$, the approach degenerates into the greedy baseline with poor performance. With a larger $\sigma$, both *Coverage* and *Confidence* increase and archive the best result at $\sigma=0.2$. It is also noticeable that the approach is not very sensitive to the parameter, as both *Coverage* and *Confidence* become stable with a relatively large $\sigma$ (Fig. 8).

The number of topics also affects the performance. We employ the Gibbs sampling algorithm to learn topic models for different number of topics (e.g. 10, 30, 50 topics). In addition, we vary the number of related topics (i.e., $|T(q_j)|$ and $|T(v_i)|$) from 1 to 7, and $T(q_j)$, $T(v_i)$ are determined by selecting top-k topics in $\theta_{q_j}$ and $\theta_{v_i}$ as we have discussed in Section 4.1. The sensitivity curves of *Coverage* and *Confidence* are plotted in Fig. 8 and Fig. 9. With the help of pairwise matching scores, *Coverage* and *Confidence* remain $>80\%$ and $>54\%$ even we set $T(q_j)=T(v_i)=1$ and topic number to be 10. Moreover, we see that too large or too small number of related topics do not produce good results. The appropriate number of related topics is about 3 or 4, which is fairly near the ground truth. Setting too few topics (e.g. 10) may hurt the final performance, but using enough topics (20, 30, 50) would not make big differences. One explantation is, small number of topics may limit the discriminative power of topics.

**Table 5**
Comparison of all methods on four different measures: *Coverage*, *Confidence*, *AverageCoverage* and $F_{Score}$. The number of topics is set to 20 when learning topic model.

| Methods | Coverage (%) | Confidence (%) | AverageConfidence (%) | $F_{score}$ (%) |
|---|---|---|---|---|
| Baseline-Pr | 74 | 62 | 46 | 63 |
| Baseline-KL | 75 | 62 | 45 | 63 |
| Topic-KL | 87 | 58 | 53 | 67 |
| Our approach | 87 | 60 | 51 | 67 |
| Our approach-L10 | 86 | 58 | 49 | 66 |
| Our approach-L5 | 80 | 59 | 48 | 65 |

### 5.3. Course–teacher assignment experiment

*Dataset*. In the course–teacher assignment experiment, we manually crawled graduate courses from the department of Computer Science (CS) of four top universities, namely CMU, UIUC, Stanford, and MIT. In total, there are 609 graduate courses from the fall semester in 2008 to 2010 spring, and each course is instructed by 1–3 teachers. Our intuition is that teachers' research interest often match the graduate courses he/she is teaching. Thus we still use the teachers' recent (five years) publications as their

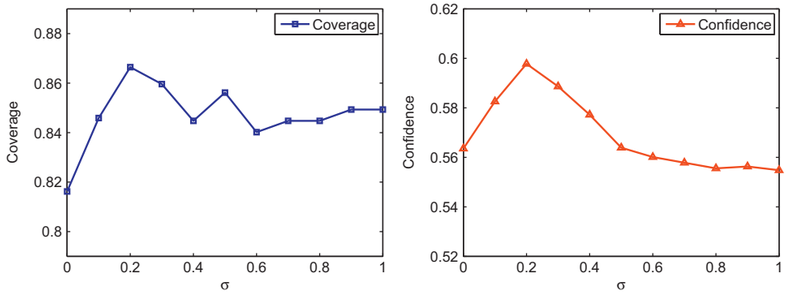



**Fig. 8.** Sensitivity of *Coverage* and *Confidence* to different Lagrangian multiplier $\sigma$.

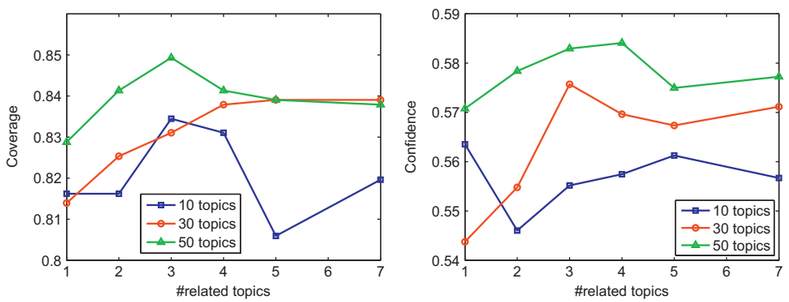



**Fig. 9.** Sensitivity of *Coverage* and *Confidence* to the number of related topics for different topic models, with $\sigma=0.3$.



得分和负载余额约束，我们得到了如图7的最终配置。7.结果。我们在没有负载余额约束的情况下考试我们的方法。 DataSet D2上的所有方法的结果呈现在表5中，我们看到我们的方法和主题−KL均优于覆盖，信心，平均值和FScore的两种基线方法。有趣的是，基线方法实现更好的信心。这可能是因为纸张的指定审阅者更有可能多次涵盖相同的主题。我们没有负载余额的方法实现了主题−KL的可比性。在添加负载余额约束时，它是明显的，所提出的方法仍然可以提供良好的效果，这验证了我们方法的有用性。在表5中，我们使用我们的方法−L10，我们的方法−15表示具有不同负载平衡设置的提出方法，即N2 10和N2 5。参数灵度。现在我们研究了框架中参数的影响。我们首先考虑乘法器S，控制覆盖率得分的重要性。设置时

S≠0，方法退化为贪婪的基线，性能差。具有较大的S，覆盖率和置信度增加并归档S 0：2的最佳结果。对于参数来说，方法也不是非常敏感的，因为覆盖率和置信度都具有相对较大的S（图8）。主题的数量也会影响性能。我们使用GIBBS采样算法来学习不同数量的主题（例如10,30,50主题）的主题模型。此外，我们改变了从1到7的相关主题（即，9t qj 9和9t vi 9）的数量，而T（QJ），T（VI）是通过选择YQJ和YVI中的Top−K主题来确定，因为我们在部分中讨论4.1。覆盖的遮盖力和置信度的灵敏度曲线在图4中绘制。参照图8和图9.在成对匹配分数的帮助下，覆盖率和置信度仍然是480％和454％，即使我们设置为T（QJ） 1（vi）1和主题号是10.此外，我们看到的相关主题数量太大或太多的相关主题不会产生良好的结果。适当数量的相关主题约为3或4，这在实地真理附近。设置太少的主题（例如，10）可能会损害最终表现，但使用足够的主题（20,30,50）不会产生大差异。一个渐次进入是，少数主题可能会限制主题的辨别力量。

表5对四种不同措施的所有方法的比较：覆盖，信心，verageCoverage和fscore。学习主题模型时，主题的数量被设置为20。

| Methods | Coverage (%) | Confidence (%) | AverageConfidence (%) | $F_{score}$ (%) |
|---|---|---|---|---|
| Baseline-Pr | 74 | 62 | 46 | 63 |
| Baseline-KL | 75 | 62 | 45 | 63 |
| Topic-KL | 87 | 58 | 53 | 67 |

我们的方法87 60 51 67我们的方法−L10 86 58 49 66我们的方法−15 80 59 48 65

### 5.3. Course–teacher assignment experiment

数据集。在课程作业实验中，我们手动爬出了四所顶级大学计算机科学（CS）的研究生课程，即CMU，UIUC，Stanford和MIT。总共有2008年秋季学期的609个研究生课程春季，每门课程都被1−3名教师指示。我们的直觉是，教师的研究兴趣往往与他/她正在教学的研究生课程相匹配。因此，我们仍然使用教师最近（五年）出版物作为他们的出版物

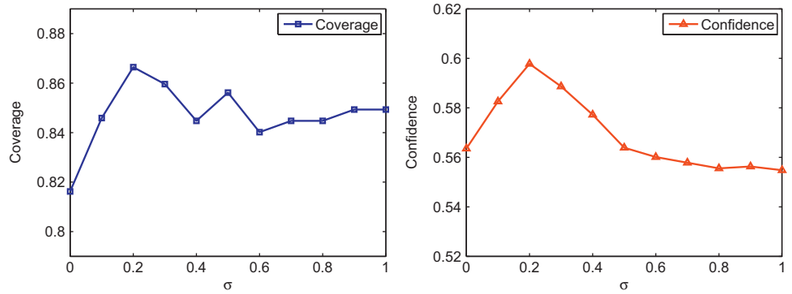



图8.不同拉格朗日乘法器的覆盖和信心的敏感性。

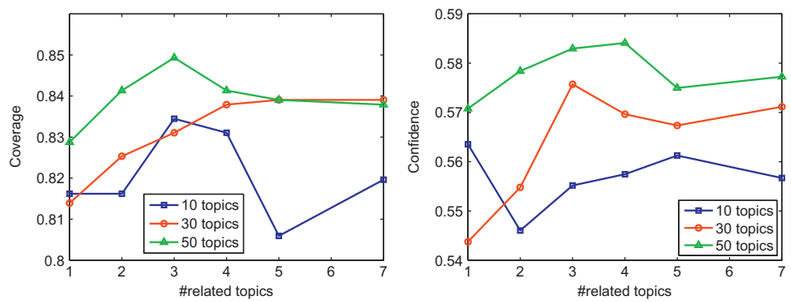



图9.不同主题模型的覆盖率和相关主题次数的敏感性，S 0：3。

expertise documents, while the course description and course name are taken as the query.

*Baseline methods and evaluation metrics*. We employ the same greedy method used in experiment 5.1 as baseline. The real assignment is extracted as the ground truth. Thus, we perform the evaluation in terms of precision.

*Results*. Fig. 10(a) shows the assignment precision in the course–teacher assignment task by our approach and the baseline method, and (b) shows the effects of the parameter $\beta$ on the precision on UIUC data. The precision is defined as the ratio of the number of correct assignments(consistent with the ground truth data) over total number of assignments. As Fig. 10(a) shows, in all the datasets we collect from top universities, our algorithm outperforms the greedy method greatly. And in Fig. 10(b), as the $\beta$ increases, the precision of our approach increases in general and decreases slowly after it exceeds the peak value. The peak value is more than 50% larger than the initial precision, which validates the effectiveness of the soft penalty approach.

We conduct a further analysis on the UIUC dataset. As Table 6 shows, some professors with publications in various domains, are likely to be assigned with many courses in the baseline algorithm. But in real situation, most professors, though with various background, want to focus on several directions. Thus some courses should be assigned to younger teachers. While in our algorithm, the situation is much better. And we can see that each teacher is assigned with a reasonable load as well as a centralized interest.

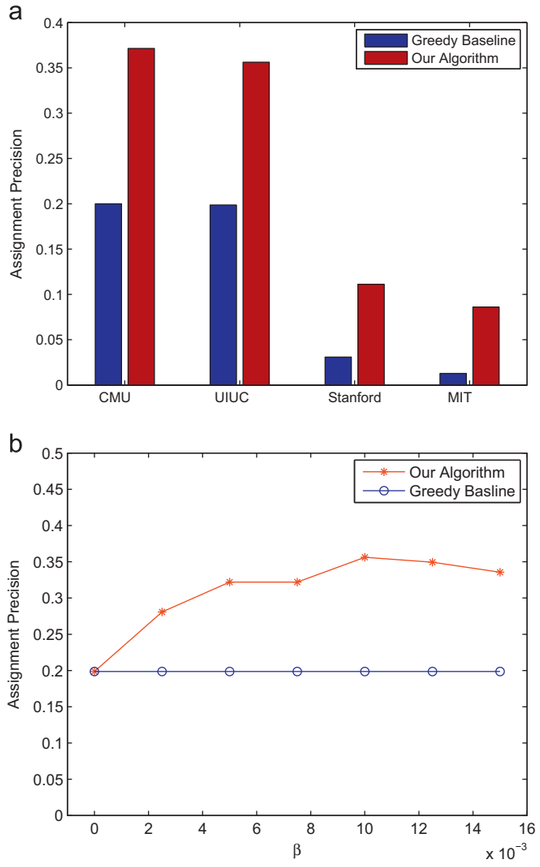



**Fig. 10.** Course–teacher assignment performance (%). (a) Course assignment results and (b) precision vs. $\beta$ on UIUC data.

### 5.4. Online system

Based on the proposed method, we have developed an online system for paper–reviewer suggestions, which is available at http://review.arnetminer.org/. Fig. 11 shows a screenshot of the system. The input is a list of papers (with titles, abstracts, authors, and organization of each author) and a list of conference program committee (PC) members. We use the academic information stored in ArnetMiner to find the topic distribution for each paper and each PC member [42]. With the two input lists and the topic distribution, the system automatically finds the match between papers and authors. As shown in Fig. 11, there are 5–7 papers assigned to each PC member and the number of reviewers for each paper is set as 3. The system will also avoid the conflict-of-interest (COI) according to the coauthorship and co-organization relationship. In addition, users can provide feedbacks for online adjustment, by removing or confirm (fix) an assignment.

## 6. Conclusion and future work

In this paper, we study the problem of expertise matching in a constraint-based framework. We formalize the problem as a minimum convex cost flow problem. We theoretically prove that the proposed approach can achieve an optimal solution and develop an efficient algorithm to solve it. Experimental results on two different types of datasets demonstrate that the proposed approach can effectively and efficiently match experts with the queries. Also we provide an algorithm to consider user feedbacks in real time. We are now applying the proposed method to several real-world applications. Feedbacks from the users are very positive.

The general problem of expertise matching represents a new and an interesting research direction. There are many potential

**Table 6**
Case study: professors with many courses assigned in UIUC (2008, fall - 2010, spring).

| Professor | Pub papers | Courses assigned (baseline) | Courses assigned (our approach) |
|---|---|---|---|
| Jose Meseguer | 237 | 23 courses<br>Database systems (2008, spring)<br>Programming languages and compilers (2008, spring)<br>Iterative and multigrid methods (2009, spring)<br>Programming languages and compilers (2009, spring) | 7 courses<br>Programming languages and compilers (2008, spring)<br>Programming language semantics (2008, spring)<br>Programming languages and compilers (2008, fall)<br>Programming languages and compilers (2009, spring) |
| ChengXiang Zhai | 117 | 18 courses<br>Computer vision (2009, spring)<br>Text information systems (2009, spring)<br>Stochastic processes and applic (2009, fall)<br>Computer vision (2008, spring) | 7 courses<br>Text information systems (2008, spring)<br>Stochastic processes and applic (2008, fall)<br>Text information systems (2009, spring)<br>Stochastic processes and applic (2009, fall) |



专业文件，而课程描述和课程名称被视为查询。基线方法和评估指标。我们使用实验5.1中使用的相同的贪婪方法作为基线。真正的分配是作为地面真理提取的。因此，我们在精度方面进行评估。结果。图。图10（a）通过我们的方法和基线方法示出了课程－
教师分配任务中的分配精度，并且（b）显示了参数B对的效果

UIUC数据的精度。精度被定义为在总分配总数上的正确分配数量（与地面真实数据一致）的比率。如图10（a）所示，我们从顶部大学收集的所有数据集中，我们的算法大大大大地表现了贪婪的方法。并且在图1中。如图10（b）所示，随着B的增加，我们的方法的精度一般增加并且在超过峰值后缓慢降低。峰值比初始精度大超过50%，验证了软罚球方法的有效性。我们对UIIC数据集进行进一步的分析。作为表6.

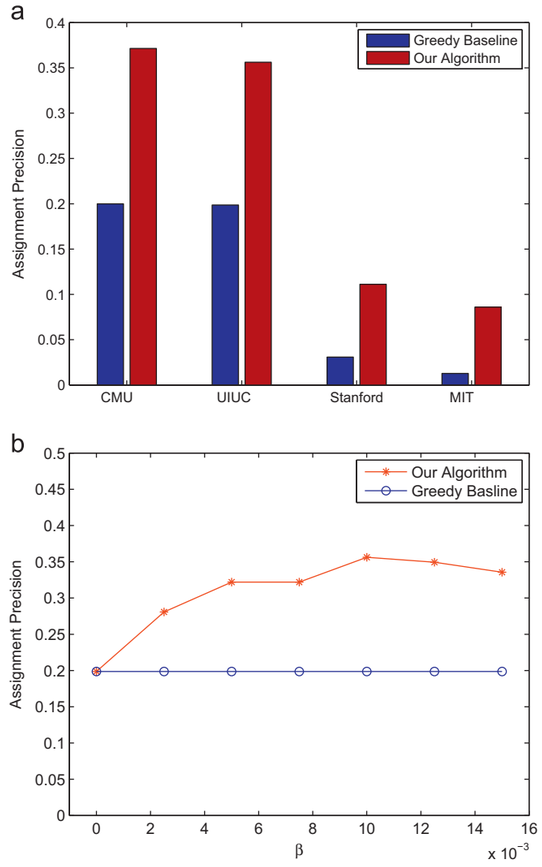



图10.课程教师分配性能（%）。（a）课程分配结果和（b）对UIUC数据的精度与B.

shows, some professors with publications in various domains, are likely to be assigned with many courses in the baseline algorithm. But in real situation, most professors, though with various background, want to focus on several directions. Thus some courses should be assigned to younger teachers. While in our algorithm, the situation is much better. And we can see that each teacher is assigned with a reasonable load as well as a centralized interest.

### 5.4. Online system

Based on the proposed method, we have developed an online system for paper–reviewer suggestions, which is available at http://review.arnetminer.org/. Fig. 11 shows a screenshot of the system. The input is a list of papers (with titles, abstracts, authors, and organization of each author) and a list of conference program committee (PC) members. We use the academic information stored in ArnetMiner to find the topic distribution for each paper and each PC member [42]. With the two input lists and the topic distribution, the system automatically finds the match between papers and authors. As shown in Fig. 11, there are 5–7 papers assigned to each PC member and the number of reviewers for each paper is set as 3. The system will also avoid the conflict-of-interest (COI) according to the coauthorship and co-organization relationship. In addition, users can provide feedbacks for online adjustment, by removing or confirm (fix) an assignment.

6.结论和未来的工作

在本文中，我们研究了基于约束框架的专业知识问题。我们将问题形式形式化为最小凸起成本流动问题。理论上，我们证明了所提出的方法可以实现最佳解决方案并开发一种有效的算法来解决它。两种不同类型的数据集上的实验结果表明，所提出的方法可以有效且有效地将专家与查询匹配。我们还提供了一种算法，可以实时考虑用户反馈。我们现在将提议的方法应用于几个现实世界的应用程序。来自用户的反馈非常积极。专业知识匹配的一般问题代表了一种新的和有趣的研究方向。有很多潜力

表6案例研究：UIUC中分配了许多课程的教授（2008年，2010年秋季，春季）。

| Pub Papers教授分配（基线）课程（我们的方法） |
|---|
| Jose Meseguer 237<br>23课程7课程数据库系统（2008年，春季）编程语言和编译器（2008，Spring）编程语言和编译器（2008，Spring）编程语言语言语言语言语言（2008，Spring）迭代和多重资源方法（2009，Spring）编程语言和编译器（2008年，秋季）编程语言和编译器（2009年，Spring）编程语言和编译器（2009年，春天） |
| 成翔翟117课程7课程电脑愿景（2009年，春季）文本信息系统（2008年，春季）文本信息系统（2009年，春季）随机流程和亚申请（2008，秋季）随机流程和亚申请（2009，秋季）文本信息系统（2009年，Spring）计算机视觉（2008，Spring）随机流程和申请（2009年，秋季） |

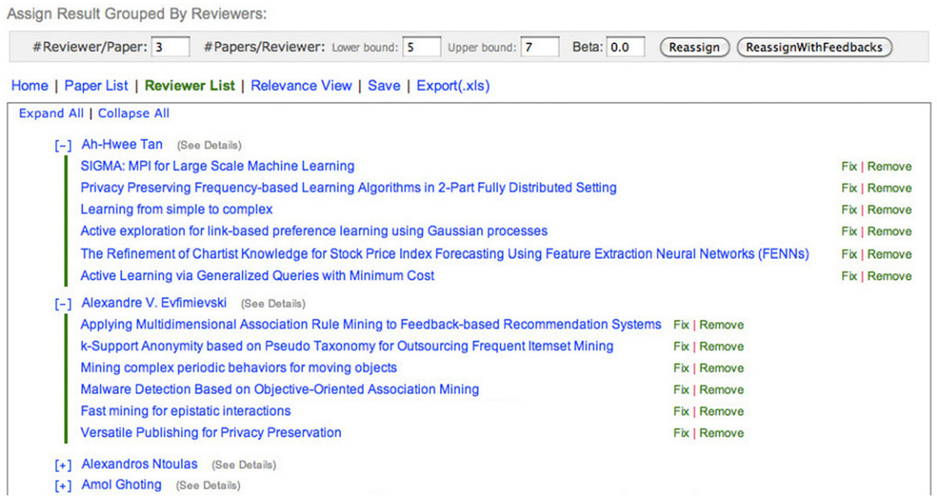



**Fig. 11.** Screenshot of the online system.

future directions of this work. One interesting issue is to apply the proposed framework to question answer (e.g., Yahoo! Answer), where one of the most important issues is how to identify who can answer a new question. Another interesting issue is to incorporate some supervised information into our framework to further improve the performance of expertise matching. Finally, it is important to consider the influence between users when extending expertise matching to the social network.

## Acknowledgments

The work is supported by the Natural Science Foundation of China (Nos. 61073073, 60703059, and 60973102), Chinese National Key Foundation Research (No. 60933013), and National High-tech R&D Program (No. 2009AA01Z138). The work was done when the last author was visiting Tsinghua University.

## References

[1] C.B. Haym, H. Hirsh, W.W. Cohen, C. Nevill-manning, Recommending papers by mining the web, in: Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI'99), 1999, pp. 1–11.
[2] S.T. Dumais, J. Nielsen, Automating the assignment of submitted manuscripts to reviewers, in: SIGIR'92: Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 1992, pp. 233–244.
[3] M. Karimzadehgan, C. Zhai, G. Belford, Multi-aspect expertise matching for review assignment, in: Proceedings of the 17th ACM International Conference on Information and Knowledge Management (CIKM'08), 2008, pp. 1113–1122.
[4] D. Mimno, A. McCallum, Expertise modeling for matching papers with reviewers, in: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'07), 2007, pp. 500–509.
[5] M. Karimzadehgan, C. Zhai, Constrained multi-aspect expertise matching for committee review assignment, in: Proceedings of the 17th ACM International Conference on Information and Knowledge Management (CIKM'09), 2009, pp. 1697–1700.
[6] D. Hartvigsen, J.C. Wei, R. Czuchlewski, The conference paper–reviewer assignment problem, Decision Sciences 30 (3) (1999) 865–876.
[7] Y.-H. Sun, J. Ma, Z.-P. Fan, J. Wang, A hybrid knowledge and model approach for reviewer assignment, in: Proceedings of the 40th Hawaii International Conference on Systems Science (HICSS-40 2007), 2007, p. 47.
[8] S. Hettich, M.J. Pazzani, Mining for proposal reviewers: lessons learned at the national science foundation, in: Proceeding of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06), 2006, pp. 862–871.
[9] D. Conry, Y. Koren, N. Ramakrishnan, Recommender systems for the conference paper assignment problem, in: RecSys'09: Proceedings of the Third ACM Conference on Recommender Systems, ACM, New York, NY, USA, 2009, pp. 357–360.
[10] N.D. Mauro, T.M.A. Basile, S. Ferilli, Grape: an expert review assignment component for scientific conference management systems, in: IEA/AIE'05: Proceedings of the 18th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, 2005, pp. 789–798.
[11] R. Van De Stadt, Cyberchair: A Web-Based Groupware Application To Facilitate The Paper Reviewing Process, URL 〈http://borbala.com/cyberchair/wbgafprp.pdf〉.
[12] Microsoft Conference Management Toolkit (cmt), URL 〈http://cmt.research.microsoft.com/cmt/〉.
[13] The Easychair Software, URL 〈http://www.easychair.org/〉.
[14] H. Fang, C. Zhai, Probabilistic models for expert finding, in: Proceedings of the 29th European Conference on Information Retrieval Research ECIR'07, 2007, pp. 418–430.
[15] D. Petkova, W.B. Croft, Hierarchical language models for expert finding in enterprise corpora, International Journal on Artificial Intelligence Tools (2008) 5–18.
[16] K. Balog, L. Azzopardi, M. de Rijke, Formal models for expert finding in enterprise corpora, in: Proceedings of the 29th ACM SIGIR International Conference on Information Retrieval (SIGIR'2006), 2006, pp. 43–55.
[17] W. Tang, J. Tang, C. Tan, Expertise matching via cosntraint-based optimization, in: Proceedings of 2010 IEEE/WIC/ACM International Conferences on Web Intelligence (WI'2010), 2010.
[18] D. Yimam, A. Kobsa, Demoir: a hybrid architecture for expertise modeling and recommender systems, in: Proceedings of the Ninth IEEE International Workshops on Enabling Technologies: Infrastructure for Colaborative Enterprises, 2000, pp. 67–74.
[19] Y. Cao, J. Liu, S. Bao, H. Li, Research on expert search at enterprise track of trec 2005, in: TREC, 2005.
[20] Y. Fu, W. Yu, Y. Li, Y. Liu, M. Zhang, S. Ma, Thuir At Trec 2005: Enterprise Track, 2005.
[21] C. Basu, H. Hirsh, W.W. Cohen, C. Nevill-Manning, Technical paper recommendation: a study in combining multiple information sources, Journal of Artificial Intelligence Research 14 (2001) 231–252.
[22] C. Basu, H. Hirsh, W. Cohen, Recommendation as classification: using social and content-based information in recommendation, in: Proceedings of the Fifteenth National Conference on Artificial Intelligence, AAAI Press, 1998, pp. 714–720.
[23] D. Yarowsky, R. Florian, Taking the Load Off The Conference Chairs: Towards A Digital Paper–Routing Assistant, 1999.
[24] J. Zhang, J. Tang, J. Li, Expert finding in a social network, Advances in Databases: Concepts Systems and Applications 23 (2010) 1066–1069.
[25] D. Yimam-Seid, A. Kobsa, Expert finding systems for organizations: problem and domain analysis and the demoir approach, Sharing Expertise: Beyond Knowledge Management 23 (2003) 327–358.
[26] C. Zhai, J. Lafferty, A study of smoothing methods for language models applied to ad hoc information retrieval, in: Proceedings of the 24th ACM SIGIR International Conference on Information Retrieval (SIGIR'01), 2001, pp. 334–342.



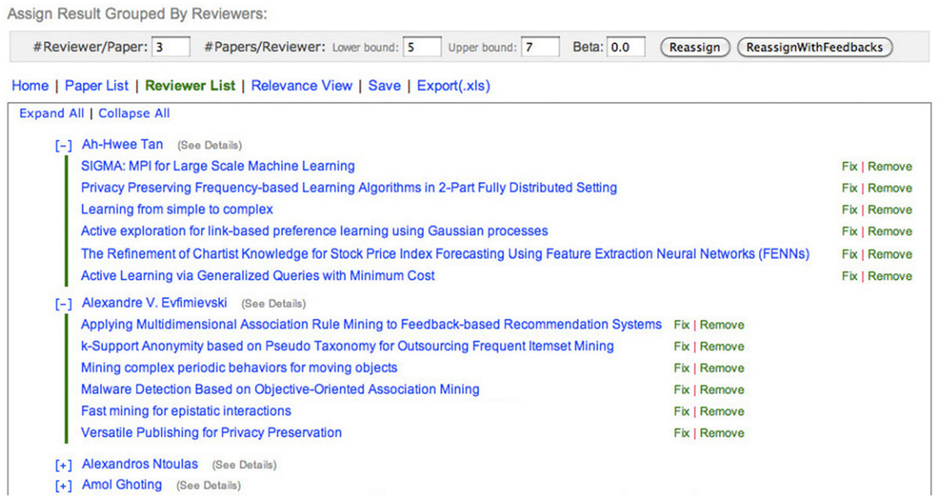



图11。在线系统的屏幕截图。

这项工作的未来方向。一个有趣的问题是将拟议的框架应用于质疑答案（例如，雅虎！答案），其中最重要的问题之一是如何识别谁可以回答一个新问题。另一个有趣的问题是将一些受监管信息纳入我们的框架，以进一步提高专业匹配的绩效。最后，重要的是考虑用户在与社交网络匹配的专业知识时的影响。

## Acknowledgments

该工作得到了中国自然科学基金（No.61073073,60703059和60973102）的支持，中国国家重点基础研究（60933013号）和国家高科技研发计划（No.2009AA8Z138）。最后一位作者访问清华大学时，这项工作是完成的。

## References

[1] C.B.Haym，H. Hirsh，W.W. Cohen，C. Nevill-Manning，通过开采网络推荐论文，在：第20届国际人工智能联席会议（IJCAI'99），1999，第1–11页的第20届国际联席会议。[2] S.T. Dumais，J.Nielsen，自动向审稿人员分配提交的稿件， Sigir'92：第15届年度国际ACM
Sigir会议的关于信息检索，ACM，纽约，纽约，美国，1992，PP的第15届国际ACM Sigir会议。233–244。[3] M. Karimzadehgan，C.
Zhai，G.Belford，多个方面专业匹配审查任务，为：第17届ACM信息和知识管理国际会议（CIKM'08），2008，PP。1113– 1122。[4] D. MIMNO，A.
McCallum，与审稿人员匹配论文的专业知识建模，包括：第13届ACM
SIGKDD国际知识发现和数据挖掘国际会议（SIGKDD'07），2007，PP。500–509。[5] M. Karimzadehgan，C.
Zhai，委员会审查任务的受限多方面专业匹配，适用于：第17届ACM信息和知识管理国际会议（CIKM'09），2009，PP。1697–1700。[6] D.
Hartvigsen，J.C.Wei，R.Czuchlewski，会议论文 –
审阅者分配问题，决策科学版30（3）（1999）865–876。[7]
y.–h.太阳，J.Ma，Z.–P. FAN，J.
WANG，审查员作业的混合知识和模型方法，在：40夏威夷系统科学国际会议（HICSS–40 2007），2007，p。47. [8] S. Hettich，MJ
Pazzani，提案审查员挖掘：在国家科学基金会的经验教训，在：第15届ACM
SIGKDD国际知识发现和数据挖掘会议上（KDD'06），2006年，pp.862–871。
[9] D.
Conry，Y.Koren，N.Ramakrishnan，Constence纸张分配问题的推荐系统，IN：Recsys'09：第三个ACM会议的会议员工，ACM会议推荐者系统，ACM，纽约，NY，美国，2009年，第357–360页。[10]
N.D.Mauro，T.M.A.巴斯蒂尔，S.Ferilli，葡萄：科学会议管理系统的专家审查任务组成部分，IEA /
AIE'05：第18届工业和工程应用程序的贸易委员会的人工智能和专家系统，2005年，PP的诉讼程序。789–798。[11] R. Van de Stadt，Cyber Chair：基于Web的群件应用程序，用于遵守纸质评论过程，URL /
HTTP://borbala.com/cyberchair/wbgafprp.pdf§。[12]
Microsoft会议管理工具包（CMT），URL / http://cmt.Research。
microsoft.com/cmt/s。[13] EasyChair软件，URL /
http://www.easychair.org/s。[14] H. Fang，C.
Zhai，专家查找的概率模型，在：第29届欧洲信息检索研究会议核发会议核发生组织07,2007，PP。418–430。[15] D. Petkova，W.B.
Croft，Enterprise
Corpora中专家查找的分层语言模型，国际人工智能工具（2008）5–18。[16] K. Balog，L.Azzopardi，M. de
Rijke，企业集团专家的正式模型，包括：第29届ACM
Sigir国际信息检索国际会议（Sigir'2006），2006，PP。43– 55。[17] W. Tang，J. Tang，C.
TAN，通过基于COSNTRAINT的优化匹配的专业知识，IN：2010年IEEE / WIC / ACM国际会议关于WEB
Intelligence（WI'2010），2010年的课程。[18 ] D. YIMAM，A.
Kobsa，Demoir：用于专业知识建模和推荐系统的混合架构，在：第九IEEE国际研讨会上的培训技术的程序：叠层进入的基础设施，2000，PP。67–74。[19] Y.Cao，J. Liu，S. Bao，H李，TREC
2005企业轨道专家搜索研究，IN：TREC，2005. [20] Y.Fu，W. Yu，Y.
Li，Y.刘，M. Zhang，S. Ma，Thuir在TREC 2005：Enterprise
Track，2005. [21] C. Basu，H.
Hirsh，WW科恩，C.尼维曼宁，技术论文建议：结合多种信息来源的研究，人工智能研究杂志14（2001）231–252。[22] C.
Basu，H.Hirsh，W.Cohen，分类建议：使用基于社会和基于内容的信息的建议书，在：第五十届全国人工智能大会上的诉讼程序，Aaai
Press，1998，PP。714– 720。[23] D. Yarowsky，R.
Florian，将负担从会议椅上取消；走向数字纸张路由助理，1999年。[24]
J. Zhang，J. Tang，J. Li，社交网络中的专家查找。
，数据库的进步：概念系统和应用程序23（2010）1066–1069。[25] D.
Yimam–Seid，A.
Kobsa，组织的专家查找系统：问题和域分析以及Demoir方法，共享专业知识：超越知识管理23（2003）327–358。[26] C.
Zhai，J.Lafferty，用于适用于临时信息检索的语言模型的平滑方法的研究，IN：24th ACM
Sigir国际信息检索国际会议（Sigir'01），2001，PP。334 –342。

[27] K. Balog, L. Azzopardi, M. de Rijke, A Language Modeling Framework For Expert Finding, 2008.
[28] X. Wei, W.B. Croft, Lda-based document models for ad-hoc retrieval, in: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval SIGIR'06, ACM, New York, NY, USA, 2006, pp. 178–185.
[29] J.J.M. Guervós, P.A.C. Valdivieso, Conference paper assignment using a combined greedy/evolutionary algorithm, in: PPSN, 2004, pp. 602–611.
[30] E. Zitzler, L. Thiele, Multiobjective Evolutionary Algorithms: A Comparative Case Study and the Strength Pareto Approach, 1999.
[31] M. Cámara, J. Ortega, F. de Toro, A single front genetic algorithm for parallel multi-objective optimization in dynamic environments, Neurocomputing 72 (2009) 3570–3579.
[32] C.J. Taylor, On the Optimal Assignment of Conference Papers to Reviewers, Technical Report, MS-CIS-08-30, Computer and Information Science Department, University of Pennsylvania, 2008.
[33] M.A. Rodriguez, J. Bollen, An algorithm to determine peer–reviewers, in: Paroceedings of the Conference on Information and Knowledge Management, ACM Press, Napa, California, 2008, pp. 319–328 doi: 10.1145/1458082.1458127.
[34] S. Benferhat, J. Lang, Conference paper assignment, International Journal of Computational Intelligence Systems 16 (10) (2001) 1183–1192.
[35] T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), 1999, pp. 50–57.
[36] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993–1022.
[37] T.L. Griffiths, M. Steyvers, Finding scientific topics, Proceedings of the National Academy of Sciences PNAS'04 (2004) 5228–5235.
[38] J. Cao, T. Xia, J. Li, Y. Zhang, S. Tang, A density-based method for adaptive lda model selection, Neurocomputing 72 (2009) 1775–1781.
[39] J. Tang, R. Jin, J. Zhang, A topic modeling approach and its integration into the random walk framework for academic search, in: Proceedings of 2008 IEEE International Conference on Data Mining (ICDM'08), 2008, pp. 1055–1060.
[40] R.K. Ahuja, T.L. Magnanti, J.B. Orlin, Network Flows: Theory, Algorithms, and Applications, Prentice Hall, 1993.
[41] P. Beraldi, F. Guerriero, R. Musmanno, Parallel algorithms for solving the convex minimum cost flow problem, Computational Optimization and Applications 18 (2) (2001) 175–190.
[42] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, Z. Su, Arnetminer: Extraction and mining of academic social networks, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'08), 2008, pp. 990–998.
[43] The Lemur Toolkit For Language Modeling And Information Retrieval, URL 〈http://lemurproject.org/〉.

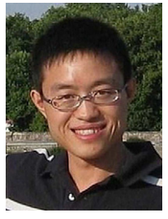
**Tao Lei** is currently a research assistant in KEG Group, Tsinghua University. He obtained a BS degree in Peking University. His research interests focus on machine learning and text mining.

**Chenhao Tan** is a Ph.D. student in Cornell University. His research interests are machine learning and social network.

**Bo Gao** is a software engineer in KEG Group, Tsinghua University. He is currently in charge of the development and maintenance of the academic social network ArnetMiner.

**Wenbin Tang** is a master student in Tsinghua University, supervised by Prof. Jie Tang. His research interests are text mining, social network and computer vision.

**Tian Li** is an undergraduate student in Beijing University of Aeronautics and Astronautics.



**Jie Tang** is an associate professor in Tsinghua University. His research interests are machine learning and text mining.

W. Tang et al. / Neurocomputing 76 (2012) 71–83 83

[27] K. Balog, L.Azzopardi, M. de Rijke, 专家查找的语言建模框架, 2008年。[28] X.
Wei, W.B.克罗夫特, 基于LDA的文档模型, 适用于Ad-hoc检索, In: 第29届国际ACM Sigir会议的关于信息检索和开发的ACM
Sigir'06, ACM, 纽约, 纽约, 美国, 2006, PP。178– 185。[29] J.J.M.
guervo的, p.a.c。
Valdivieso, 使用组合禽棲/进化算法的会议纸张分配, IN: PPSN, 2004, PP。602–611。[30] E. Zitzler, L.
Thiele, 多目标进化算法: 一种比较案例研究和强度Pareto方法, 1999。[31] M.Ca'Mara, J.Ortega, F. de
Toro, 单一前遣传算法用于动态环境中的并行多目标优化, 神经会计机72 (2009) 3570–3579。[32] CJ
Taylor在审查员, 技术报告, MS-CIS-08-30, 宾夕法尼亚大学, 2008年培养师, 技术报告, MS-CIS-08-30, 计算机和信息科学事件的最佳分配, 一种确定同行评审员的算法, IN: 信息和知识管理会议, ACM
Press, Napa, California, 2008, PP。319–328 Doi: 10.1145 / 1458082.1458127。
[34] BENFERHAT, J.
LANG, 会议论文任务, 国际计算智能系统16 (10) (2001) 1183–1192。[35] T.
Hofmann, 概率潜在语义索引, 在: 31年度国际ACM
Sigir会议上的研究和开发信息检索 (Sigir'99), 1999, PP。50–57。[36] D.M.
Blei, A.Y.
ng, m.i.约旦, 潜在的Dirichlet分配, 机器学习研究学报3 (2003) 993–1022。[37] T.L. Griffiths, M.
Steyvers, 寻找科学主题, 国家科学院的诉讼程序PNAS'04 (2004) 5228–5235。
[38]
J.Cao, T.Xia, J.Li, Y.张, S.唐, 一种基于密度的自适应LDA模型选择方法, 神经计算器72 (2009) 1775–1781。[39] J. Tang, R. Jin, J.
Zhang, 一个主题建模方法及其进入学术搜索随机步行框架的融合, 共度: 2008年IEEE数据挖掘国际会议 (ICDM'08), 2008年, PP。1055–1060。[40] R.K.
Ahuja, T.L. Magnanti, JB orin, 网络流量: 理论, 算法和应用, Prentice
Hall, 1993. [41] P.Bersdi, F. Guerriero, R.
Musmanno, 用于解决凸起最小成本流量问题的并行算法, 计算优化和应用 – 阳离子18 (2) (2001) 175–190。[42] J. Tang, J. Zhang, L. Yao, J.Li, L.
Zhang, Z. Su, Arnetminer: 学术社交网络的提取和采矿, 在: 第14届ACM
SIGKDD国际知识发现国际会议上和数据挖掘 (SIGKDD'08), 2008, PP。990–998。[43]语言建模和信息检索的狐猴工具包, URL / http://lemurproject.org/s。

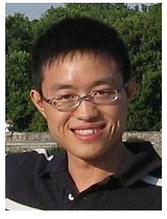
Tao
Lei目前是清华大学柯格集团的一名研究助理。他在北京大学获得了BS学位。他的研究兴趣专注于机器学习和文本挖掘。

Chenhao
Tan是一个博士学位。康奈尔大学的学生。他的研究兴趣是机器学习和社交网络。

BO高是清华大学柯格集团的软件工程师。他目前负责学术社交网络ARNetminer的发展和维护。

温斌唐是清华大学的一名硕士学位, 由杰唐教授监督。他的研究兴趣是文本挖掘, 社交网络和计算机愿景。

天丽是北京航空航天大学的本科生。



杰唐是清华大学的副教授。他的研究兴趣是机器学习和文本挖掘。