DOI:10.3969/j.issn.1673-4785.201205041
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20120720.1005.001.html

面向学术社区的专家推荐模型

李春英[1],汤庸[2],陈国华[2],汤志康[3]
(1.肇庆学院 计算机学院,广东 肇庆 526061;2.华南师范大学 计算机学院,广东 广州 510631;3.广东技术师范学院 计算机学院,广东 广州 510665)

摘 要:在学术社区提供的服务中,对于研究者特别是青年研究者来说,专家推荐是一个必不可少的部分.目前提供学术信息服务的所有中文搜索引擎中,都没有提供用户感兴趣的专家推荐服务.因此,提出了一个面向学术社区的专家推荐模型.使用改进的H参数对学者n年时间内发表的论文成果进行量化,获取专家列表;使用概率主题模型从作者发表的论文中提取主题向量作为学者的研究方向;根据矩阵奇异值分解对构建的词项-文档矩阵进行降维,进而生成词项-词项关系矩阵,实现对搜索关键词的查询扩展,并计算查询扩展向量与作者主题向量之间的相关度,根据相关度大小进行排序推荐.在SCHOLAT(学者网)数据集上验证模型的有效性,实验结果表明提出的模型达到了预期的效果.

关键词:学术专家推荐;H参数;概率主题模型;查询扩展;
中图分类号:TP393 **文献标志码**:A **文章编号**:1673-4785(2012)04-0365-05

Research on an expert recommendation model based on the scholar community SCHOLAT

LI Chunying[1], TANG Yong[2], CHEN Guohua[2], TANG Zhikang[3]
(1.School of Computer, Zhaoqing University, Zhaoqing 526061, China; 2.School of Computer Science, South China Normal University, Guangzhou 510631, China; 3.School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou 510665, China)

Abstract: Among the services offered by the academic community, expert recommendation is an indispensable component for researchers, especially young researchers. At present, expert recommendation services have not been offered to users on all of the Chinese search engines offering academic information services. Thus, a scholar community oriented expert recommendation model was proposed. The H-index was improved to quantify the achievements of a scholar based on the published papers in the last n years, and then the expert list was given based on the improved H-index. The research interests of a researcher were obtained based on the topics extracted by the probabilistic topic model. In order to carry out high recall retrieval, a query expansion strategy was used: the singular value decomposition step was applied to the term-document matrix to reduce the dimensionality of the matrix and obtain the term-term relationship matrix, and then the highly related terms were selected to make up the expanded query. Finally, the relevance between the expanded query and the scholar's topic vectors was calculated and the results were represented in a descending order. An experiment was conducted on the dataset collected from an existing scholar community, SCHOLAT, to verify the effectiveness of the proposed model. The experimental results demonstrate that the proposed model produces the expected results.

Keywords: expert recommendation; H index; probabilistic topic model; query expansion

学术合作研究越来越受到人们的重视[1].在学术研究领域有相同研究兴趣或者工作在不同学科、领域的科研人员常常组成一个团队进行合作,显然这使得更多的问题得到了解决.比如一篇电子商务方面的论

收稿日期:2012-05-24. **网络出版日期**:2012-07-20.
基金项目:国家自然科学基金资助项目(60970044);广东省科技计划资助项目(2010B010600031);广州市科技计划资助项目(2010J-D00511).
通信作者:李春英.E-mail:zqxylcy@163.com.

文,可能是从事计算机研究、经济学研究和管理学研究的学者共同的智慧结晶. 实际上,学者间高水平的合作具备更强的生产力. 因此,找到潜在的成功合作者对于研究者特别是青年研究者来说是倍受欢迎的. 然而,团队合作常常局限于同一学科、同一科研院所的内部. 对于我国大多数二三类院校的研究者而言,因缺乏学术带头人导致众多研究者特别是青年研究者无法超越现实的距离而徘徊不前. 而国内外大多数学术搜索引擎,如中国知网、万方数据知识服务平台、维普资迅、Scirus、Google Scholar、CiteSeer、CiteULike、DBLP、C-DBLP 等都具备了文献检索的功能,但他们都没有对有着相似研究兴趣和潜在合作关系的学者进行有效地挖掘和推荐.

为了有效地挖掘潜在的合作者并进行推荐,本文提出了一个面向学术社区的专家推荐系统模型. 在这个学术社区内,用户可以按照关键字搜索相关论文;系统可以根据用户的研究兴趣,为用户自动推荐最新的相关论文;另外还可以自动管理用户的学术资料,如果用户有新论文发表出来,当用户登录时, 系统将会提醒用户将该论文收藏到自己的主页中,这样就极大地方便了用户对自己资料的管理,同时可以让其他用户及时地了解到自己的最新工作进展[2]. 除常规功能外,本文着重论述学术专家推荐模型的设计及实验测评. 该模型包括 3 个部分:1)通过分析学者公开发表的论文被引用的次数、录用期刊的影响因子以及发表论文的数量 3 个方面对学者的学术价值进行量化;2)利用主题模型提取学者的研究方向;3)对搜索关键词进行查询扩展,并计算其与作者主题词之间的相关度,按相关度排序在推荐系统中给出用户需要的学者专家列表,其中用户可按影响力进行排序.

1 专家推荐模型的相关工作

专家推荐模型是一种面向学术领域的学术推荐搜索引擎. 对于学术推荐,近年来人们开展了大量的研究工作并取得了丰硕的研究成果. 文献[3]在 CiteULike 社区结合了传统的协同过滤的优点和概率主题模型进行建模,为用户推荐论文. 文献[4]提出一个基于合作发现的搜索引擎,为学者推荐潜在的学术研究合作伙伴. 对于学术搜索引擎,文献[2]已经做了非常详尽的阐述,在此不再赘述.

总之,在目前提供学术信息服务的所有中文搜索引擎中,都没有提供推荐用户感兴趣的领域专家服务. 如果结合学术社区提供一个易于使用的专家推荐服务,一定能使科研工作者特别是青年科研工作者感兴趣,并可能给他们未来的研究工作带来极大方便. 下面将分别阐述面向学术社区的专家推荐模型的详细设计和实验评估,系统的整体架构如图 1 所示.

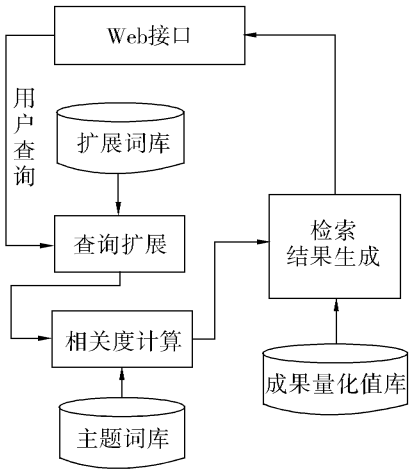


图 1 系统的整体架构

Fig. 1 The system architecture

2 模型的详细设计

2.1 成果量化

对于获得诺贝尔奖的科学家而言,他们研究工作的影响和主题相关性是毋庸置疑的. 但对于大多数的研究人员而言,该如何量化个人科研成果累积的影响和相关性? 科研人员公开发表的论文记录显然是对量化有用的信息. 各个科研机构往往是利用科研人员有限的成果资源,进行比较和评价. 虽然这种量化可能使人反感,但在高校、科研院所,它是科研人员职务招聘、晋升职称和补助奖励的有效依据. J. E. Hirsch 在文献[5]中提出利用个人在过去 n 年时间内发表的论文数量 p、论文被引用的次数 c 和录用期刊的影响因子去评估个人的科研成果,即所谓的 H 参数. J. E. Hirsch 的 H 参数在物理学科[6-7]和在科学计量学[8]方面得到了验证并获得了广泛的认同,表明该参数对于量化个人的科研成果是有效的. J. E. Hirsch 提出 H 参数的具体量化公式[5]如式(1)所示.

$$H = \frac{c}{1 + \frac{c}{p}}n. \tag{1}$$

式(1)中未直接考虑期刊的影响因子对 H 参数的影响. 因此对其加以改进,使其能够更加准确地量化个人的科研成果.

令 A 表示某期刊在第 $y-2$ 年和第 $y-1$ 年出版的所有文章在第 y 年被引用的次数之和,B 表示该期刊在第 $y-2$ 年和第 $y-1$ 年所发表的文章篇数之

和,则该期刊在第 y 年的影响因子 $\mathrm{IF}_y = \frac{A}{B}$. 假设一个人在过去 n 年时间内发表论文 p 篇,C_i 表示第 i 篇论文被引用的次数,IF_y 表示收录第 i 篇论文的期刊当年的影响因子,改进的 H 参数如式(2)所示.

$$\mathrm{H} = \frac{\sum_{i=1}^{p} C_i \times \mathrm{IF}_y}{1 + \frac{\sum_{i=1}^{p} C_i \times \mathrm{IF}_y}{\sum_{i=1}^{p} \mathrm{IF}_y}}. \tag{2}$$

2.2　概率主题模型

概率主题模型越来越多地应用于图像处理和自然语言处理领域. 在自然语言处理领域中,主题可以看成是词项的概率分布. 主题模型通过词项在文档级的共现信息抽取出语义相关的主题集合,并能够将词项空间中的文档变换到主题空间,得到文档在低维空间中的表达. 这为语料库挖掘、文档分类和信息检索工作提供了极大的便利. 本文将使用主题模型抽取作者全部文章的主题信息,进而形成作者研究方向的主题集合.

使用主题模型对文档的生成过程进行模拟,再通过参数估计得到各个主题. 最简单的主题模型是 LDA(latent Dirichlet allocation)[9]. 假定 φ_t 表示主题 t 中的词项概率分布;θ_j 表示第 j 篇文档的主题概率分布;φ_t、θ_j 又作为多项式分布的参数分别用于生成单词和主题,服从 Dirichlet 分布;T 代表主题数目;M 代表文档数目;N_j 表示第 j 篇文档的长度;ω_{j_n} 和 Z_{j_n} 分别表示第 j 篇文档中第 n 个单词及其主题;α 和 β 是 Dirichlet 分布的参数,通常是固定值且是对称分布的[10].

则对于语料库中的每一篇文档 ω_j,LDA 的生成过程如下:

1)对主题采样 $\varphi_t \sim \mathrm{Dir}(\beta)$,$t \in [1,T]$;

2)采样主题概率分布 $\theta_j \sim \mathrm{Dir}(\alpha)$;

3)采样文档的单词数目 $N \sim \mathrm{Poiss}(\xi)$;

4)对文档 j 中的每个单词 n:

①选择隐含主题 $Z_{j_n} \sim \mathrm{Multinomial}(\theta_j)$;

②生成一个单词 $\omega_{j_n} \sim \mathrm{Multinomial}(\varphi_{Z_{j_n}})$.

这个过程表明了从每一篇文档中提取主题词的过程. 对于给定的语料库,根据给定的最优化目标函数,使用 Gibbs 参数估计方法得到对参数的估计值. 利用训练好的模型对新文档进行推断,发现 T 个主题,进而将指定的词项空间表达的文档分解降维,得到所需要的主题集合.

2.3　查询扩展

查询扩展是查询优化的一个分支研究方向,也是目前改善信息检索中查全率和查准率的关键技术之一. 查询扩展是指为了保证用户搜索时使用的关键词和作者主题词相关,需将用户搜索时使用的关键词进行语义扩展,把与原关键词语义相关的词或词组添加到原查询中,得到比原查询更长的新查询,以便更完整、更准确地描述原查询所隐含的语义,帮助其提供更多有利于判断文档相关性的信息,提高检索的查全率和查准率.

隐性语义索引(latent semantic indexing, LSI)[11]用于发现文本中词项-文档之间的语义关系. 在 LSI 模型中,词项-文档矩阵 $\boldsymbol{C}$ 用于表示词项和文档之间的关系,$\boldsymbol{C} = (C_{ij})$,其中 C_{ij} 表示第 i 个词项在第 j 篇文档中的权重值,即第 i 个词项在第 j 篇文档中出现的次数.

LSI 通过奇异值分解对高维稀疏的词项-文档矩阵构造低阶最佳近似,以减轻计算的复杂度. 适用奇异值分解降维的基本思想为:假设 $\boldsymbol{C}_{m\times n}$ 是词项-文档矩阵;m 是词项空间的维度,n 是文档个数,则 $\boldsymbol{C}\boldsymbol{C}^{\mathrm{T}}$ 是 m 阶对称方阵,其元素 (i,j) 代表了词项 i 和词项 j 的共现次数,反映了任意 2 个词项 (i,j) 之间的相似度. 则

1)令 $\mathrm{rank}(\boldsymbol{C}) = r$,则 $\boldsymbol{C}\boldsymbol{C}^{\mathrm{T}}$ 的特征值为 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_r > 0$,$\lambda_{r+1} = \lambda_{r+2} = \cdots = \lambda_m = 0$,令 $\sigma_i = \sqrt{\lambda_i}(i = 1,2,\cdots,r)$,$\sigma_i$ 称为 $\boldsymbol{C}$ 的奇异值.

2)存在正交矩阵 $\boldsymbol{U}_{m\times r}$、$\boldsymbol{V}_{m\times r}$和广义对角阵 $\boldsymbol{\Sigma}_{r\times r}$(其中 $\boldsymbol{\Sigma}_{ii} = \sigma_i$)使得 $\boldsymbol{C} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$,则 $\boldsymbol{C}\boldsymbol{C}^{\mathrm{T}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}} = \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^{\mathrm{T}}$.

对于给定的矩阵 $\boldsymbol{C}$ 和正整数 $k(k \leqslant r)$,找到一个矩阵 $\boldsymbol{C}_k$,使用 Frobenius 范数来估算 $\boldsymbol{C} - \boldsymbol{C}_k$ 的误差,即 $\| \boldsymbol{C} - \boldsymbol{C}_k \|_F = \sqrt{\sigma_{k+1}^2 + \sigma_{k+2}^2 \cdots + \sigma_r^2}$,计算并且输出 $(\boldsymbol{C}\boldsymbol{C}^{\mathrm{T}})_{k\times m} = \boldsymbol{U}\boldsymbol{\Sigma}_k^2\boldsymbol{U}^{\mathrm{T}}$,作为 $\boldsymbol{C}\boldsymbol{C}^{\mathrm{T}}$ 降维后的最佳近似,其中的行向量显示了某一词项与所有其他词项的相关程度.

2.4　相关度计算

当用户进行查询请求时,首先将关键词经过中文分词处理,然后对其分词结果进行查询扩展,并将所有结果作为查询关键词向量 $\boldsymbol{U}$ 的分量,个数作为关键词向量空间的维数. 最后使用 Salton 的 TF-IDF 公式计算向量 $\boldsymbol{U}$ 中每个关键词分量的权值,详见式(3).

$$\omega_{ik} = \frac{t_{ik}\lg\left(\frac{N}{n_k} + 0.1\right)}{\sqrt{\sum_{k=1}^{n} (t_{ij})^2 \times \lg^2\left(\frac{N}{n_k} + 0.1\right)}}. \tag{3}$$

式中：t_{ik}表示关键词分量 U_k 在主题向量 $\boldsymbol{T}_{id}$ 中出现的次数，N 表示主题词库中主题向量的总数，n_k 表示主题词库中含有关键词 U_k 的主题向量数. 因此，查询关键词被扩展为一个查询关键词向量：向量的维数就是分词后的中文词语个数，向量每一维分量的大小就是每个分量的权值. 对每一个主题向量 $\boldsymbol{T}_{id}$，每个主题分量的权值取文档主题的概率分布值. 因此主题向量分量的权值可用其对应的概率分布值表示即 $\boldsymbol{P}_{id}=[P_{id_1}\ P_{id_2}\cdots P_{id_n}]$. 此时，要计算关键词和主题向量的相关度，可以认为是向量 $\boldsymbol{U}$ 和向量 $\boldsymbol{T}_{id}$ 之间的相关度，而计算向量之间的相关度，可以使用向量夹角余弦系数进行衡量，如式(4)所示，最后按相关度大小进行排序并将结果页面推荐给用户.

$$\text{similarity}(\boldsymbol{U},\boldsymbol{T}_{id}) = \cos\theta = \frac{\sum_{k=1}^{n}(U_k \times P_{id_k})}{\sqrt{\sum_{k=1}^{n}U_k^2 \times \sum_{k=1}^{n}P_{id_k}^2}}. \tag{4}$$

3 实验评估

3.1 成果量化实验

量化计算关键在于数据库的设计和查询算法. 论文和作者是多对多的关系，为了分担部分计算压力、提高查询性能，需提前计算好部分数据结果. 因此，需将 SCHOLAT 数据集中的数据分成 3 个部分.

1)论文信息表：论文 ID(主键)、论文名称、作者、作者单位、发表刊物、影响因子、出版年份、参考文献、引用次数.

2)论文作者关系表：论文 ID(外键)、作者.

3)作者信息表：作者 ID(主键)、作者名称、作者单位、研究方向、$\boldsymbol{C}$ 值(取自 $\sum_{i=1}^{p}C_i \times \text{IF}_y$)、IF 值(取自 $\sum_{i=1}^{p}\text{IF}_y$)、H 值.

将论文信息表中的作者(合作者)、作者单位进行分词处理并将结果存入论文作者关系表、作者信息表. 当有信息更新时，系统将论文信息处理后分别存入论文信息表、论文作者关系表和作者信息表，并更新 $\boldsymbol{C}$ 值和 IF 值字段，进而更新作者信息表的 H 值.

从学者网(SCHOLAT)数据集中选取汤庸等 100 位学者在 2006 年 1 月 1 日—2010 年 12 月 31 日 5 年共 2 513 篇论文进行量化，从中剔除了引用次数为 0 的论文 373 篇，实际参加测试的论文数目为 2 140 篇，实验所需期刊影响因子数据来源于中国科技期刊引证报告(核心版)和维普资讯网. 实验结果显示量化模型有效. 因 SCHOLAT 数据集目前不包括英文文献，导致总体量化值偏低，但与期望值相似. 按职务量化求均值后的结果如表 1 所示.

表 1 学者成果量化值

Table 1 Quantify the achievements of a scholar

ID	职务	H 参数平均值
1	教授/博士生导师	20.32
2	教授/硕士生导师	13.63
3	副教授	3.97
4	讲师	0.98

3.2 概率主题模型实验

实验中，设定 Gibbs 算法的迭代次数是 1 000 次，经多次实验，Dirichlet 的先验参数 α 和 β 取值为 $\alpha=20/\text{K}$，$\beta=0.01$，起到了平滑数据的作用. 采用 Perplexity 评估方法(如式(5)所示)确定最佳主题个数 $T=200$，如图 2 所示. 在 SCHOLAT 数据集上将作者全部文章的标题和摘要合并后分词，将分词后的词项集合和最佳 T 值作为 LDA 算法的输入项，得到每个作者论文的潜在主题集合，将每个潜在主题下概率最大的词项提取出来构成每个作者研究方向向量 $\boldsymbol{T}_{id}$，并将每个主题的概率分布值作为其在向量 $\boldsymbol{P}_{id}$ 中的权值.

$$\text{perplexity}(D_{\text{test}}) = \exp\left\{\frac{-\sum_{d=1}^{M}p(d_d)}{\sum_{d=1}^{M}N_d}\right\}. \tag{5}$$

式中：N_d 为文本 d 的长度，$p(d_d)$ 是待测试模型产生文档 d_d 的概率.

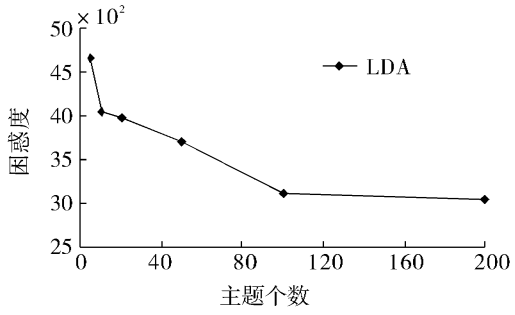


图 2 主题个数与困惑度的变化关系

Fig.2 Relation of number of topics and perplexity

3.3 查询扩展及相关度计算实验

从论文库中抽取成果量化值大于 5 的计算机相关研究方向的 100 个作者的 2 513 篇论文，将每一个作者名下的文章标题和摘要合并成一篇文档，则参与测试的文档数目为 100，然后对 100 篇文档进行分词，去掉停用词等没有实际意义的信息后，共 175 910个词项参与实验，编写实验源程序建立词项-文档共生矩阵 $\boldsymbol{C}$，使用 Lanczos 算法计算 SVD，对

所建立的高维稀疏的词项-文档矩阵分解降维. 实验取得最佳 K 值,$K=53\ 853$,计算并输出 $\boldsymbol{C}_k$,进而输出 $\boldsymbol{A}=\boldsymbol{C}_k\boldsymbol{C}_k^{\mathrm{T}}$,则矩阵 $\boldsymbol{A}$ 为词项-词项的相关度矩阵,A_{uv} 表示词项 u 和词项 v 的相关度权值. 查询时将与用户关键词相关度最大的前 200 个词项作为扩展词项加入到用户的查询中,其中用户的原始查询词项最能直接反映用户查询意图,其权值置为最大. 在相关度计算方面,以单个词项查询作为测试条件,选择查询扩展向量与作者主题向量进行向量夹角余弦系数计算时,系统的响应时间为 132 ms. 这显然比设计成与矩阵 $\boldsymbol{C}_k$ 中的每一列列向量进行向量夹角余弦系数计算的方案的系统响应时间要少很多. 因此,在查全率和查准率近似的情况下,前者大大降低了计算的复杂度,提高了系统的响应时间.

4　结束语

提出了一个面向学术社区的专家推荐系统模型,给出了系统的总体架构及各个部分的详细设计方案,在 SCHOLAT 数据集上做实验验证了模型的有效性. 其中,成果量化模型和概率主题模型部分均为离线运算,降低了系统的压力. 不足之处是成果量化模型中选择参与计算的成果时间跨度较小且没有考虑合作者的权重问题,主要原因是目前 SCHOLAT 数据集有些数据不够充分,以及无法批量获得论文的通信作者信息,下一步应用时将主要解决这些问题.

参考文献:

[1] HUANG J, ZHUANG Z, LI J, et al. Collaboration over time: characterizing and modeling network evolution[C]// Proceedings of the International Conference on Web Search and Web Data Mining. Palo Alto, USA, 2008: 107-116.

[2] 陈国华, 汤庸, 彭泽武,等. 基于学术社区的学术搜索引擎设计[J]. 计算机科学, 2011, 38(8): 171-175.
CHEN Guohua, TANG Yong, PENG Zewu, et al. Design of an academic search engine based on the scholar community [J]. Computer Science, 2011, 38(8): 171-175.

[3] WANG Chong, BLEI D M. Collaborative topic modeling for recommending scientific articles[C]//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2011: 448-456.

[4] CHEN H H, GOU Liang, ZHANG Xiaolong, et al. Collabseer: a search engine for collaboration discovery[C]//Proceedings of JCDL. Ottawa, Canada, 2011: 231-240.

[5] HIRSCH J E. An index to quantify an individual's scientific research output[J]. The National Academy of Sciences of the USA, 2005, 102(46): 16569-16572.

[6] POPOV S B. A parameter to quantify dynamics of a researcher's scientific activity[EB/OL]. [2011-11-03]. http://arxiv.org/abs/physics/0508113.

[7] BATISTA P D, CAMPITELI M G, KINOUCHI O, et al. A complementary index to quantify an individual's scientific research output[J]. Scientometrics, 2006, 68 (1): 179-189.

[8] BORNMANN L, DANIEL H D. Does the h-index for ranking of scientists really work?[J]. Scientometrics, 2005, 65 (3): 391-392.

[9] BLEI D, NG A, JORDAN M. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

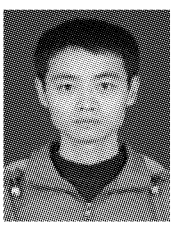
[10] 徐戈, 王厚峰. 自然语言处理中主题模型的发展[J]. 计算机学报, 2011, 34(8): 1423-1436.
XU Ge, WANG Houfeng. The development of topic models in natural language processing[J]. Chinese Journal of Computers, 2011, 34(8): 1423-1436.

[11] DEERWESTER S, DUMAIS S T, LANDAUER T K, et al. Indexing by latent semantic analysis[J]. Journal of The American Society for Information Science, 1990, 41 (6): 391-407.

作者简介:

李春英,女,1978 年生,讲师,CCF 会员(E200019159M),主要研究方向为学术信息检索与推荐、人工智能.

汤庸,男,1964 年生,教授,博士生导师,博士,中国计算机学会协同计算专委会副主任,中国人工智能学会网络专委会副主任,广东省计算机学会常务副理事长,广东省网络文化协会副会长. 主要研究方向为数据库、协同计算、云服务软件,发表学术论文多篇.



陈国华,男,1984 年生,讲师,博士,主要研究方向为学术信息检索、机器学习.