doi:10.3772/j.issn.2095-915x.2018.03.008

# 基于术语抽取与分级匹配的项目指南推荐方法

古迎志[1] 董诚[2,3] 裴兵兵[1] 杜永萍[1]

1. 北京工业大学信息学部 北京 100124;
2. 中国科学技术信息研究所 北京 100038;
3. 富媒体数字出版内容组织与知识服务重点实验室 北京 100038

**摘 要** 信息推荐是自然语言处理领域的重要技术，为进一步向科研人员进行有效的项目指南推荐，本文采用术语词表征文本特征的方式，进行分级匹配推送。通过基于词性规则和句法信息相结合的方法抽取候选术语词，并利用基于统计的方法如 C-value、SCP(Symmetrical Conditional Probability)等进行术语词过滤，提高抽取质量。由指南和科研人员术语词进行分级匹配来表征二者之间的相似度，进而实现对科研人员的个性化指南推荐。对国家科技管理信息系统公共服务平台 2017 年发布的 42 篇指南设计实验进行验证，分析术语抽取结果，评价指南推荐的准确率，结果表明基于 C-value+SCP 的方法取得了更优的术语抽取质量，指南的个性化推荐准确率最高达到 80%。

**关键词：** 术语抽取；推荐技术；科技文献

**中图分类号：** TP391 G35

开放科学（资源服务）标识码（OSID）

# The Recommendation Approach Based on Term Extraction and Graduation Matching

GU Yingzhi[1] DONG Cheng[2,3] PEI Bingbing[1] DU Yongping[1]

1. Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China;
2. Institute of Scientific and Technical Information of China, Beijing 100038, China;
3. Key Laboratory of Rich-media Knowledge Organization and Service of Digital Publishing Content, SAPPRFT, Beijing 100038, China

**基金项目：** 科技部创新方法工作专项（2015IM020500）；北京市自然科学基金资助项目(4153058)。

**作者简介：** 古迎志（1992–），硕士研究生，研究方向：信息推荐，自然语言处理；董诚（1970–），通讯作者，研究员，研究方向：科技管理与科技创新，信息挖掘与服务，E-mail：dongc@istic.ac.cn；裴兵兵（1992–），硕士研究生，研究方向：信息检索，自然语言处理；杜永萍（1977–），博士，副教授，研究方向：信息检索，自然语言处理。

**Abstract** Text recommendation is an important technology in the field of Natural Language Processing. In order to recommend the project guidelines to the researcher, this paper uses terminology to represent the text features and gives the recommendation based on the graduation matching. The rule of part of speech and syntactic information are used for term extraction and the candidate terms are filtered by statistical methods, such as C-value, SCP (Symmetrical Conditional Probability) and so on, so as to improve the extraction quality. The graduation based matching between the guidelines and researchers' terminology is used to measure the similarity, and then to achieve the personalized recommendation. The experiments are implemented on the 42 project guidelines published by the public service platform in 2017. The results show that the C-value+SCP based method achieves better term extraction quality and the personalized recommendation precision is up to 80%.

**Keywords**: Term extraction; recommendation technology; scientific literature

## 1 引言

信息化时代，互联网高速发展，海量信息随之产生，针对不同用户的兴趣爱好进行个性化智能推荐具有重要的现实意义。术语是专业领域中概念的语言指称,能够体现文献的研究领域、反映领域发展趋势。基于术语抽取的个性化文本推荐不仅在自然语言处理领域有重要的研究意义，在数据挖掘、信息推荐等领域更有广泛商业用途。

目前，常用的知识推送[1]技术主要有：(1)基于内容的推送；(2)基于协同过滤的推送；(3)基于数据挖掘的推送；(4)混合推送技术。基于内容过滤推送[2]常常应用于文本信息领域，主要通过将用户感兴趣信息与资源特征进行匹配，来快速找到用户需要的资源，如Diligenti等[3]提出的基于聚类模型、Basile等[4]提出的基于语义贝叶斯以及Wu等[5]提出的基于概率潜在语义分析的推荐方法。针对用户兴趣随时间变化问题，Somlo等[6]和 Zhang等[7]提出了更新用户模型的自适应过滤方法，使用相似度高的推荐对象来更新用户模型；除此之外，还有学者提出使用决策树[8]、人工神经网络[9]等机器学习方法来建立和更新更为复杂的用户模型。基于协同过滤推送[10]一般采用K近邻算法，通过计算用户历史偏好的相似度来得到相似爱好的相邻用户群，根据相似用户对知识的评价来预测目标用户对特定知识的喜好程度，即将相似用户最感兴趣的知识推送给目标用户，如王迪等[11]将改进的协同过滤算法应用于微博平台的信息推荐。协同过滤推送依赖用户偏好数据，存在“冷启动”、稀疏性等问题。Gantner等[12]通过学习属性特征映射来解决冷启动问题；Zhang等[13]利用社会化标签来缓解冷启动问题；Huang等[14]尝试利用关联规则挖掘来解决数据稀疏性问题；Zhang等[15]则提出了矩阵的块对角结构，通过矩阵的块对角变换增加局部密度从而直接缓解稀疏性问题。基于数据挖掘推送[16]主要利用数据挖掘中关联规则和分类的挖掘技术，按照关联规则的重要程度或

doi:10.3772/j.issn.2095-915x.2018.03.008

用户所需知识类型对用户进行知识推送，如曹雪等[17]提出了基于关联规则挖掘的领域知识推荐方法，黎楠等[18]提出了基于主题发现的专利发明人推荐方法。混合推送技术[19]则是多种推送技术进行有机集合，来达到高效准确推送的目的，利用每一种推送技术的优点，同时也对每一种方法所存在的不足之处做了进一步的优化，该方法可以提高推送的质量，但是由于该方法采用了多种技术的结合，在一定程度上会受到约束。

本文主要针对指南文本的个性化推荐技术展开研究，利用基于词性规则和句法信息的方式抽取出候选术语词，并使用C-value[20-22]、SCP[23]、PMI (Pointwise Mutual Information)等基于统计的方法过滤并选择质量较高的术语。以此为基础，对科技指南和科研人员信息进行分级匹配，进而实现个性化推荐。本文第2节阐述了基于多策略的术语提取模型并实现科技指南个性化推荐。第3节通过实验比较验证了术语抽取和指南推荐效果。最后为结论。

## 2 基于术语抽取的推荐模型

本文所使用的基于术语词的文本推荐模型主要分为三部分：术语抽取模型、术语过滤模型和推荐匹配模型。通过术语提取模型对指南信息和科技人员信息进行术语词抽取并由过滤模型过滤质量较差的术语词，最终通过推荐匹配模型计算相关性，实现指南的个性化推荐。具体过程如图1。

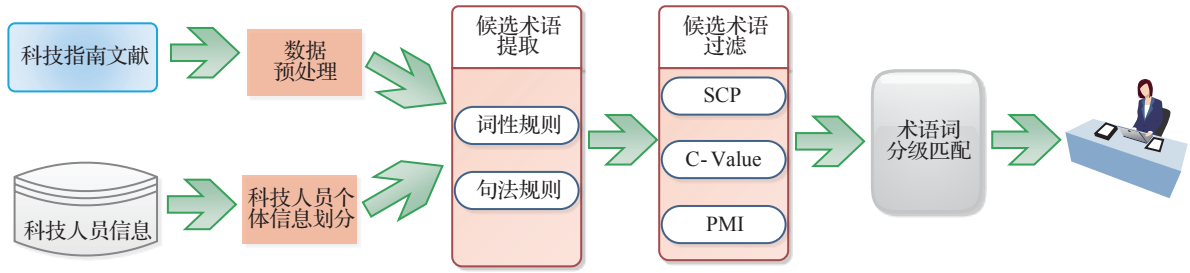


图1 基于术语抽取的项目指南推荐模型结构

### 2.1 术语抽取模型

术语体现和负载了一个学科领域的核心知识，一篇文章中的术语词很大程度上表征了此篇文章的领域性。所以可以通过文本中抽取的术语词表征文本之间的相似度。本文主要对原始指南文本使用基于词性和句法结构的方式抽取出候选术语词。

（1）基于词性规则的术语抽取

从语言学系统的观点和术语内部结构发现大多数术语具体有以下特点[24]：

- 术语的长度特点：中文术语长度主要是2到6个字；
- 术语大多是名词性的短语；
- 术语形成模式特点，如Noun+Noun, Adj+Noun+Noun 等；
- 有些字几乎不可能出现在术语中，如“

的”、“是”、“些”。

所以本文使用如表1所示规则的方式抽取出候选术语，除此之外对于词性为学术相关词性（如gi、gm）时，也会加入到候选术语集中。

表1 基于词性规则术语抽取

| 编号 | 方式 | 实例词语 |
|---|---|---|
| 1 | Noun $^{+}$ Noun | 数据环境、内存模型 |
| 2 | (Adj \| Noun) $^{+}$ Noun | 智能家居、数据环境 |
| 3 | ((Adj \| Noun) $^{+}$ \| ((Adj \| Noun) * (Noun Prep)$^{?}$)(Adj \| Noun) *)Noun | 云计算基础理论 |
| 4 | (Adj\|Noun\|Verb)$^{+}$(Verb\|-Noun)$^{?}$(Noun\|Verb) | 数据中心调度 |

（2）基于句法结构的术语提取

本文利用Hanlp开源工具[①]内置的句法分析器对语句进行句法分析及词性标注，根据依存句法标注关系找到当前词语对象与中心词的依存关系，只有当前词语与中心词满足以下条件时进行抽取。

- 当前词语与中心词的依存关系为限定；
- 中心词与当前词语的依存关系为内容或受事；
- 当前词语与中心词距离为1；
- 当前词语与中心词粗粒度词性皆为名词。

例如：“面向工业生产环境的决策技术”中“决策技术”一词，“决策”与中心词的依存关系为限定，“技术”为中心词且依存关系为内容，二者相邻且粗粒度词性均为名词，故将其抽取，具体示例如图2。

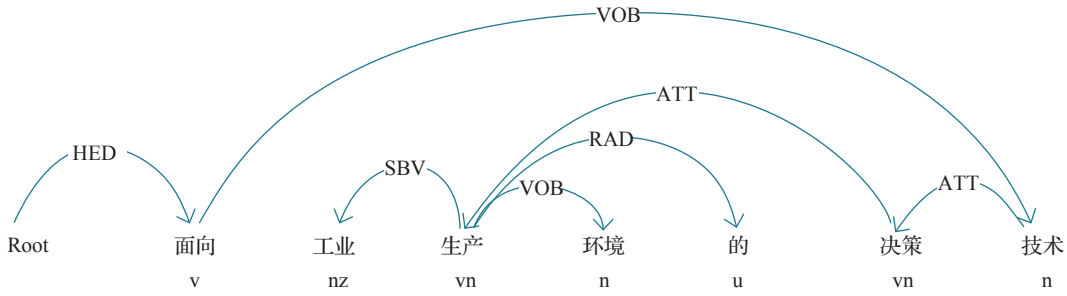


图2 基于句法结构术语抽取

## 2.2 术语过滤模型

基于词性和句法结构抽取出来的词语，可能含有噪音数据，需要采用一定的方法将其过滤。本文将候选术语词分别经过通用词、C-value、SCP等过滤器，最终保留下来的即为既定的术语词。其中，通用词典为在不同领域均可能出现的术语词，不具有代表性，如：“研究”、“方法”、“理论”并人工进行了校验，增强了准确性。同时，为了给术语词增加其在文中不同位置(标题、敏感区、其它)的属性，本文也设置了Position方法，方便后期权值计算。最后，采用基于统计的方法衡量候选术语词的术语性、单元性、互信息等多方面的属性值，具体计算模型如下。

(1) 为验证候选术语词作为术语出现的可能性本文使用了C-value方法，其综合利用了语言学知识和统计信息，适用于抽取复杂术语（通常由两个或两个以上的词语组成），特别是在

① http://hanlp.linrunsoft.com

doi:10.3772/j.issn.2095-915x.2018.03.008

抽取网状术语时取得了较好的效果。具体计算方法如下：

$$\text{C-value}(a)=\begin{cases}\log_2|a|f(a), & \text{如果}a\text{的父串只有其本身}\\ \log_2|a|\left(f(a)-\dfrac{\sum_{b\in T_a}f(b)}{p(T_a)}\right), & \text{其它情况}\end{cases}$$

公式（1）

其中，$|a|$表示词长，$f(a)$表示词频，$T_a$表示包含$a$的词语的集合，$p(T_a)$是$a$所有父串的个数。

(2) 为验证候选术语词作为一个整体词出现的可能性，本文使用改进的SCP方法MSCP[25]衡量词串中各词之间的紧密性。具体计算方法如下：

$$MSCP(w_1\ldots\ldots w_n)=\begin{cases}\dfrac{F(w_1\ldots\ldots w_n)^2}{Avp}, & n\geq 2\\ 1, & n=1\end{cases}$$

公式（2）

$$\text{Avp}=\frac{1}{n-1}\sum_{i=1}^{n-1}F(w_1\ldots\ldots w_i)*F(w_{i+1}\ldots\ldots w_n)$$

公式（3）

其中$w_1\ldots\ldots w_n$是指候选术语词串，$w_i$是组成该候选术语的词语。$F(w_1\ldots\ldots w_n)$是指候选术语的加权词频。词串的加权词频计算公式为：

$$\text{F}(a)=\sum_{b\in S_a}(f_b(a)*weight(b))$$

公式（4）

其中。$S_a$是指术语出现的区域类别集合。b是某特定区域。$f_b(a)$是指词串$a$在区域b中出现的词频。$weight(a)$是指候选术语在文本区域中被赋予的权值。

(3) 考虑到术语词为新词的可能性，本文使用了点间互信息(PMI)方法衡量新词字符之间的结合紧密程度，对于候选词$S=S_1S_2\ldots\ldots S_n$的点间互信息计算：

$$PMI(S)=\log\frac{F(S)}{\prod_{i=1}^{n}(F(S_i)-F(S))}$$

公式（5）

其中，$F(S)$和$F(S_i)$分别表示词语$S$和字符$S_i$的出现频率。

## 2.3 分级匹配推荐

本文将科研人员数据划分为以下三个区域：科研人员文献数据关键词、长串关键词分词结果、科研人员文献数据摘要，三个区域抽取的术语词集合分别用$R_1$，$R_2$，$R_3$表示。指南文本格式较为统一分为题目、标题、正文，且在正文部分常常出现如“研究”、“应用”、“构建”、“探索”、“提出”等触发词语，这些词语上下文通常出现专业术语词，我们将这样的上下文区域标记为敏感区。本文将指南文本划分成以下两个区域：标题（包括一级标题、二级标题）和敏感区、其它区域，两个区域抽取的术语词集合分别用$G_1$，$G_2$表示。

当抽取出指南和科研人员术语词后，需要进行分级匹配推荐，具体流程如图3，由科研人员不同级别的术语词$R_1$，$R_2$和$R_3$，以及科技项目指南的术语词$G_1$和$G_2$各自交叉进行分级匹配，匹配度计算如公式(6)，这里α、β、X、Y、Z为权重参数。匹配度计算过程中，不同权重对推荐结果有较大影响。本文实验令α+β=1，α的权重按0.1的梯度递增，令10≥X>Z>Y>0，α权重按1的梯度递增，采用分层抽样的方式选择权重组合，经多次试验取值后设定令α=0.6，β=0.4，X=10，Y=4，Z=6。

$$MatchingDegree = X*\left(\alpha\frac{|R_1\cap G_1|}{|G_1|}+\beta\frac{|R_1\cap G_2|}{|G_2|}\right)+Y*\left(\alpha\frac{|R_2\cap G_1|}{|G_1|}+\beta\frac{|R_2\cap G_2|}{|G_2|}\right)+Z*\left(\alpha\frac{|R_3\cap G_1|}{|G_1|}+\beta\frac{|R_3\cap G_2|}{|G_2|}\right)$$

公式（5）

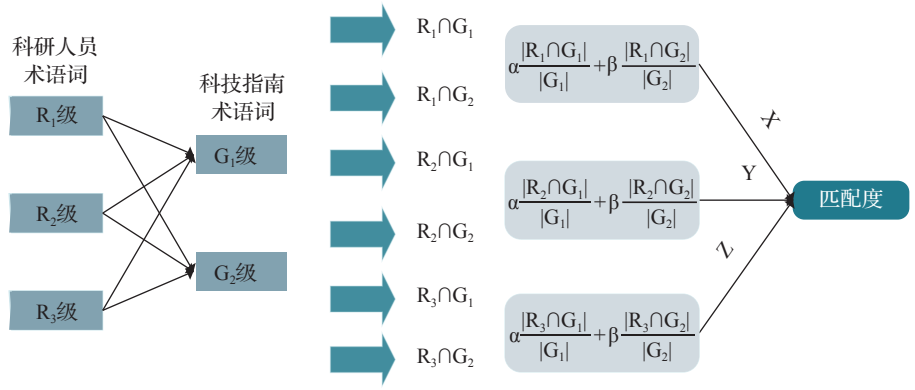


图3 指南与科研人员分级匹配推荐模型

## 3 实验与分析

### 3.1 实验设置

本文实验数据主要来自于国家科技管理信息系统公共服务平台2017年发布的42项指南文献，2000名科研人员发表的文献，其中科研人员发表的文献主要数据格式如表2。

表2 科研人员数据

| ID | 属性字段 | 释义 |
|---|---|---|
| 1 | person_id | 科研人员编号 |
| 2 | research_direction | 科研人员研究方向 |
| 3 | result_title | 科研人员文献数据标题 |
| 4 | keyword | 科研人员文献数据关键词 |
| 5 | result_abstract | 科研人员文献数据摘要 |

为验证术语抽取方法的有效性，随机选取十篇指南基于词性和句法的方式抽取术语词，并分别选取C-value、SCP、C-value+SCP、PMI进行噪音过滤。随机选取的10篇指南信息分布如表3。同时，为验证指南推荐的效果，对其进行了人工评测，使用准确率Precision进行评价。

$$准确率=\frac{模型正确推荐的科研人员数量}{模型推荐的科研人员数量}\times 100\%$$

（公式 7）

表3 10项项目指南信息分布

| 指南 ID | 名称 |
|---|---|
| 1 | “云计算和大数据”重点专项 2017 年度项目申报指南 |
| 2 | “典型脆弱生态修复与保护研究”重点专项 2017 年度项目申报指南 |
| 3 | “化学肥料和农药减施增效综合技术研发”试点专项 2017 年度项目申报指南 |
| 4 | “大气污染成因与控制技术研究”试点专项 2017 年度项目申报指南 |
| 5 | “干细胞及转化研究”试点专项 2017 年度项目申报指南 |
| 6 | “材料基因工程关键技术与支撑平台”重点专项 2017 年度项目申报指南 |
| 7 | “生殖健康及重大出生缺陷防控研究”重点专项 2017 年度项目申报指南 |
| 8 | “精准医学研究”重点专项 2017 年度项目申报指南 |
| 9 | “七大农作物育种”试点专项 2017 年度项目申报指南 |
| 10 | “蛋白质机器与生命过程调控”重点专项 2017 年度项目申报指南 |

doi:10.3772/j.issn.2095-915x.2018.03.008

### 3.2 评价结果

我们采用基于词性规则和句法分析的方式进行术语抽取，并采用C-Value，SCP和PMI进行术语过滤，在如表3所示的10篇指南上进行实验，得到的术语词数量分布如图4所示。

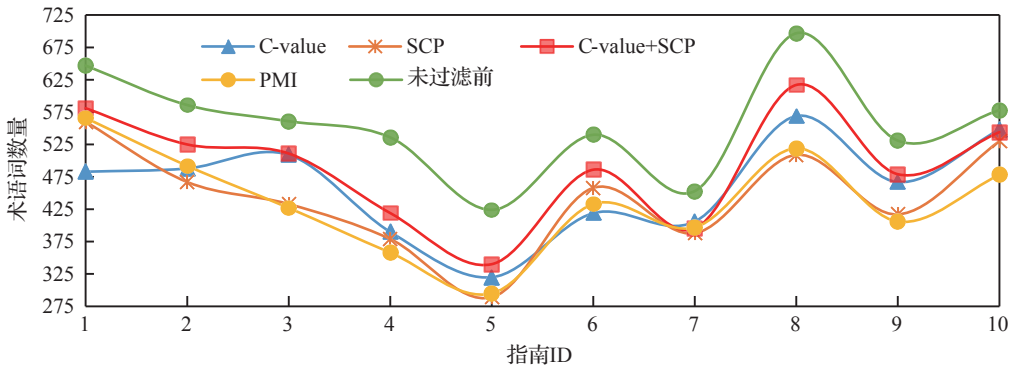


图4 不同方式抽取术语词数量

由图4可以看出，使用C-value+SCP方式能够保留较高数量的术语词。且该方法的优势在于兼顾了术语词的术语性、单元性两方面的属性，引入的噪音数据较少，术语词质量较好。其过滤掉的术语词如“仪测定”、“化学气”、“所得氧物”、“形成无限”、“水为基本原料”等，不同指南抽取的术语示例如表4。

表4 指南抽取术语实例

| 指南 ID | 名称 |
|---|---|
| 1 | 节点技术；数据资源库；海量数据；孤岛数据；深度神经网络；视频编码；数据驱动 |
| 2 | 海岛生态；河谷土壤；生态治理；全国生态；生态安全格局；森林生态系统；干旱河谷区 |
| 3 | 长江流域；先导多样性；土壤处理；土壤酸化；生物生态；农作物识别；生物除草剂 |
| 4 | 污染物来源；空气污染；硫氮资源；汽油机颗粒物；污染源控制；污染控制技术；大气化学反应 |
| 5 | 细胞异质性；移植免疫；组织微环境；细胞周期；小分子调控；遗传调控；肝脏干细胞 |
| 6 | 混合电荷；流程材料；物理模拟；动力学；流程材料；预测模型；规模组合式；工艺数据库 |
| 7 | 细胞谱系；卵巢综合征；卵母细胞；人类胚胎；免疫调节；出生缺陷疾病；生殖细胞 |
| 8 | 蛋白质鉴定；临床医学；生物医学；抗肿瘤药物；免疫细胞；患者临床；蛋白质变异体 |
| 9 | 黄淮冬麦区；小麦品种；抗病性状；抗旱能力；生长调节剂；品质遗传改良；分子标记 |
| 10 | 遗传信息；冷冻电镜；生物成像技术；神经疾病；组织器官；免疫调控机制；辐射光源 |

为进一步确定科技指南抽取的术语词质量，本文对随机选取的这十篇指南术语词，进行了人工评测来计算术语抽取的准确率。计算公式如下：

$$准确率=\frac{正确的术语词数量}{实验抽取的术语词数量}\times 100\% \quad （公式 8）$$

实验结果表明实验抽取的术语词平均准确率保持在78%左右。

对于指南向科研人员的推荐结果，我们对42篇指南推荐准确率进行人工评测，具体评价如图5所示，其中推荐准确率均值约为72.95%，其中最高的达到80%，仅第18个指南准确率较低，其主要原因是发布的指南领域特指为新能源汽车，该领域是比较新兴的方向，研究此方向的科研人员数据较少，致使抽取的新能源汽车相关术语词数量变少，无法较为全面地表示科研人员的真实研究兴趣等信息，同时，匹配成功的术语词变少，不具备更高的可靠性，使得正确率降低。

本文采用的方法适用于科研人员文献数据较为充分的情况，在此基础上，算法抽取到的术语词数量规模越大，覆盖的领域术语更全面，推荐结果也更加准确。

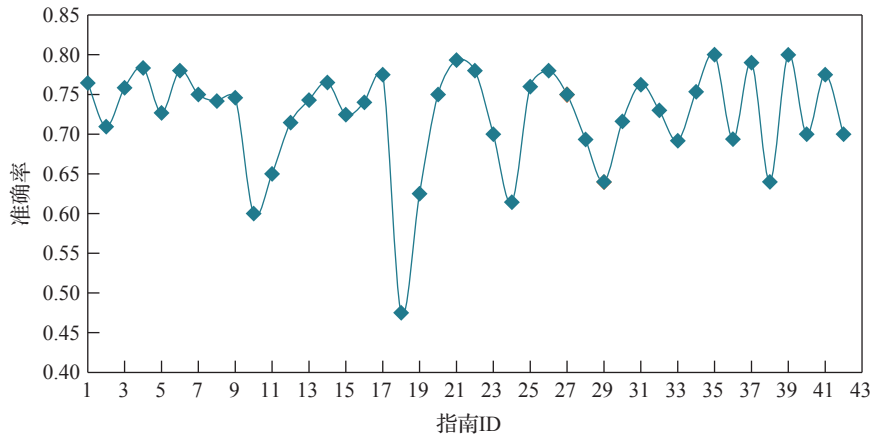


图5 指南推送正确率评价结果

## 4 总结

本文使用基于词性规则和句法信息的方法对已经发布的项目指南及科研人员数据进行术语抽取，比较了不同过滤方式后术语词抽取分布，由术语词来表征文本之间的相似度，进而实现对科研人员的个性化指南推送。实验结果表明，本文工作对于形如指南这种语言描述精炼的信息推送，使用基于词性规则和句法分析的方式来抽取术语，基于统计的方法来有效过滤，较好的表征了指南和科研人员的领域特征和研究方向，对相似度计算提供了更多的信息，有效保证了推荐的效果。

本文目前的研究为使用术语词表征科研人员特征，对如何使用用户画像的相关技术来描述科研人员特征，并进行更有效指南推荐将是后续工作研究的重点。

## 参考文献

[1] 周明建, 陶俊才. 知识管理系统中的知识推送[J]. 计算机辅助设计与图形学学报, 2006, 18(8):1218-1223.

[2] Aggarwal C C, Wolf J L, Wu K L, et al. Horting hatches an egg: a new graph-theoretic approach to collaborative filtering[C]. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 1999:201-212.

[3] Diligenti M, Gori M, Maggini M. Users, Queries and Documents:A Unified Representation for Web Mining[C]. Ieee/wic/acm International Joint Conferences on Web Intelligence and Intelligent Agent Technologies, 2009. Wi-Iat. IEEE, 2009:238-244.

[4] Basile P, Tinelli E, Degemmis M, et al. Semantic Bayesian Profiling Services for Information Recommendation[C]. International Conference, Kes 2007 and Xvii Italian Workshop on Neural Networks Conference on Knowledge-Based Intelligent Information and Engineering Systems. Springer-Verlag, 2007:711-719.

[5] Wu H, Wang Y, Cheng X. Incremental

doi:10.3772/j.issn.2095-915x.2018.03.008

probabilistic latent semantic analysis for automatic question recommendation[C]. ACM Conference on Recommender Systems. ACM, 2008:99-106.

[6] Somlo G L, Howe A E. Adaptive Lightweight Text Filtering[C]. International Conference on Advances in Intelligent Data Analysis. Springer-Verlag, 2001:319-329.

[7] Zhang Y, Callan J, Minka T. Novelty and redundancy detection in adaptive filtering[C]. ACM, 2002:81-88.

[8] Nikovski D, Kulev V. Induction of compact decision trees for personalized recommendation[C]. ACM Symposium on Applied Computing. ACM, 2006:575-581.

[9] Sharma A, Dey S. A document-level sentiment analysis approach using artificial neural network and sentiment lexicons[J]. Acm Sigapp Applied Computing Review, 2012, 12(4):67-75.

[10] 刘枚莲, 丛晓琪, 杨怀珍. 改进邻居集合的个性化推荐算法[J]. 计算机工程, 2009, 35(11):196-198.

[11] 王迪, 王东雨. 基于协同过滤算法的微博平台的信息推荐研究[J]. 情报工程, 2016, 2(2):81-87.

[12] Gantner Z, Drumond L, Freudenthaler C, et al. Learning Attribute-to-Feature Mappings for Cold-Start Recommendations[C]. IEEE, International Conference on Data Mining. IEEE, 2011:176-185.

[13] Zhang Z K, Liu C, Zhang Y C, et al. Solving the Cold-Start Problem in Recommender Systems with Social Tags[J]. 2010, 92(2):28002-28007.

[14] Huang Z. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering[J]. Acm Transactions on Information Systems, 2015, 22(1):116-142.

[15] Zhang Y, Zhang M, Liu Y, et al. Improve collaborative filtering through bordered block diagonal form matrices[C]. International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2013:313-322.

[16] Sudha V, Lakshmi G, Shalabh B, et al. An efficient and recommendation system for TV programs[J]. Multimedia System, 2008, 14(2):73-87.

[17] 雷雪, 侯人华, 曾建勋. 关联规则在领域知识推荐中的应用研究[J]. 情报理论与实践, 2014(12):67-70.

[18] 黎楠, 杜永萍, 何明. 基于主题发现的专利发明人推荐方法[J]. 情报工程, 2015, 1(3):90-97.

[19] 朱岩, 林泽楠. 电子商务中的个性化推荐方法评述[J]. 中国软科学, 2009(2):183-192.

[20] Frantzi K T, Ananiadou S, Tsujii J I. The C-value/NC-value Method of Automatic Recognition for Multi-Word Terms[C]. European Conference on Research and Advanced Technology for Digital Libraries. Springer-Verlag, 1998:585-604.

[21] 杨雅娜, 刘胜奇. 基于TValue融合领域度的术语抽取法[J]. 情报工程, 2015, 1(5):25-31.

[22] Maynard D, Ananiadou S. TRUCKS: a model for automatic multi-word term recognition[J]. Journal of Natural Language Processing, 2000, 8.

[23] Silva J F, Lopes C P. A Local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units[C]. Proceedings of the 6th Meeting on the Mathematics of Language, 1999.

[24] 邢红兵. 信息领域汉英术语的特征及其在语料中的分布规律[J]. 产品安全与召回, 2000(3):17-21.

[25] 史东娜. 基于半监督学习的特点通过领域术语抽取算法研究[D]. 北京: 北京邮电大学, 2009:32-33.