

# 学科背景知识在医学文本特征抽取中的应用<sup>\*</sup>

赵迎光 范少萍 安新颖

(中国医学科学院医学信息研究所 北京 100020)

〔摘要〕 分析目前科技文献文本特征抽取方法研究现状,将基于背景知识的 TF-IDF 方法应用到医学文本特征抽取中,在 4 个医学领域中进行试验对比,结果表明该方法在抽取词汇数量较少时可以显著提高抽取效果,在对文本集合中通用词的过滤及重要特征词的识别方面明显优于使用通用 IDF 的 TF-IDF 方法。

〔关键词〕 文本挖掘; TF-IDF; 特征抽取; 知识发现

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2017.04.012

**Application of Disciplinary Background Knowledge in Medical Text Feature Extraction** ZHAO Ying-guang, FAN Shao-ping, AN Xin-ying, Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China

〔Abstract〕 The paper analyzes the conditions of research on the current scientific literature text feature extraction methods, applies the TF-IDF method based on background knowledge in the medical text feature extraction, and conducts experimental comparison in four medical fields. The result indicates that this method can obviously improve the extraction effect when there are few vocabularies to be extracted, and is obviously superior to the IDF based TF-IDF method in the aspects of filtration of commonly-used words in the text set and identification of important feature words.

〔Keywords〕 Text mining; TF-IDF; Feature extraction; Knowledge discovery

## 1 引言

在科学研究中,选定研究对象是最基本的工

作,如果一项科学研究的对象界定存在偏差,那么其结论难免受到影响。以科技文献中的词(关键词/主题词)为对象揭示科研领域的知识特征、知识结构、知识演化以及研究热点等,是情报学与信息计量学研究中的一个重要组成部分。但无论是在对词统计的基础上进行热点分析和趋势分析,还是在词共现关联基础上进行信息的主题聚类 and 特征分析,抑或是在词共现网络基础上进行的知识结构和知识演化分析,都依赖于一个重要的前提:如何设置词的筛选标准以确定待分析对象,即如何从文献文本中抽取特征并依据特征选择所要分析的词汇。

〔修回日期〕 2016-12-23

〔作者简介〕 赵迎光,硕士,助理研究员,发表论文 5 篇。

〔基金项目〕 国家自然科学基金项目“基于语义的医学领域前沿知识发现及演化机制研究”(项目编号:71303259);中央级公益性科研院所基本科研业务费“面向医学科技评价的多源异构数据处理机制研究”(项目编号:2016ZX330027)。

## 2 科技文献文本特征抽取研究现状分析

### 2.1 概述

在情报学与信息计量科学中,文本主题挖掘方法大多以词频为特征,简单对高频词进行人工选取,但这种方法忽略了词本身的特征,而且在分析和挖掘过程中极度依赖于研究者的人工优化和解读,其结果又偏于空泛,未充分发挥基于词的分析在挖掘语义关联、揭示知识结构上的优势。基于此,本研究在对情报领域主题挖掘方法进行调研的基础上,分析各种应用中的数据处理过程及方法特征,对计算机领域的无监督和有监督的特征抽取方法在主题挖掘中的应用现状进行了探讨,下面分别将其作为传统方法和改进方法进行描述。

### 2.2 传统的科技文献特征抽取方法

从特征抽取的流程来看,词汇来源的确定是特征抽取的起点,特征词通常来源于关键词、统一标引主题词、标题及摘要提取词。由于关键词本身具备标引性和便利性,且是文献核心内容的浓缩和提炼,高度概括文献的基本内容,因此较多研究采用关键词作为分析单元。然而,由于文献中关键词的选择存在较强的主观性,使得表达意思完全相同的不同关键词被看作是完全相异的词汇,从而使得最终分析结果出现一定偏差,为此,一些研究者从多个角度进行改进,例如同义词合并<sup>[1]</sup>、通用词及停用词过滤<sup>[2-4]</sup>、关键词扩充<sup>[5-6]</sup>、根据词性限制特定词汇<sup>[7]</sup>等;在词汇差异化处理方面,一些学者根据位置、期刊来源、主要主题词/副主题词等不同方面对词汇进行加权处理<sup>[8-10]</sup>。在词汇来源选择的基础上,有研究者以词频为基础进行改进或优化,如根据齐普夫第二定律确定高频词阈值、抽取中间频次的词<sup>[11]</sup>、最大频繁项集方法<sup>[12]</sup>、多元共线高频词<sup>[13]</sup>等;也有研究者以新的视角进行选取,如将关键词集合转化为网络,基于社会网络分析方法进行词汇网络分析<sup>[14]</sup>。但是,由于关键词网络中上述指标与词频线性相关,因而提取的关键词与高频词差异性并不大。

### 2.3 改进的科技文献特征抽取方法

针对传统方法特征选取的各个环节中存在的问题,一些研究者从新的视角对其进行了探索和改进,包括引入外部数据特征、采用新的文本挖掘技术等。同时随着科技文献数据的快速增长,也为提高文本挖掘效果带来了更多可用的知识源和数据源,其中通过外部知识来改进特征抽取方法就是一个重要的应用,而且在这方面的研究也越来越多。Chen 等<sup>[15]</sup>将科研领域置于其背景学科内,从全局视角考察关键词对领域研究特点的表征能力,结果表明该方法所得到的关键词集能深入揭示领域的研究特点;陈果等<sup>[16]</sup>将该方法进一步应用到了机构层面,提出了研究机构的特色关键词方法,与高频关键词方法相比,特色关键词的语义内涵更为清晰,所揭示的结果比传统高频词方法更深入,并且更符合实际情况;除了使用领域知识外,通过多来源数据补充作者信息、引文信息等<sup>[17]</sup>也可以显著提高文本抽取效果;在医学领域,通过融合外部 MeSH 词表<sup>[18]</sup>的语义信息也可以优化聚类效果;除此以外,越来越多的深度学习,如 word2vec、textrank 等方法不断被应用到科技文本主题挖掘中,为情报学和信息计量的发展带来了新的机遇。

## 3 研究方法

### 3.1 概述

TF-IDF 是一种用于信息检索与数据挖掘的常用加权技术,用以评估特定的词对于文本集合中的一个文档的重要程度。但在之前的应用中发现,当文献集合数量偏小时该方法存在 IDF 值区分度不高,从而导致抽取效果不明显的问题;同时对于通用 IDF 值来说,由于其计算所用的文档集合涉及领域广泛,在具体的学科分析中,文献中重要特征词所获得的权重较低,而学科内的通用词可能获取较高的权重,因此通用 IDF 的方法并不适用与具体学科的主题挖掘。医学领域作为一门成熟的学科,具有丰富的科技文本资源,学科研究内容清晰,学科层次明确,鉴于此,在探讨文本特征抽取研究现状

的基础上，以医学领域科技文献为研究对象，尝试将基于学科背景知识的主题挖掘思路应用到 TF-IDF 方法中，根据所要分析的领域所属学科，将其上一级学科作为背景知识，将背景知识的 IDF 值用于该领域的文本特征抽取，以期探讨文本信息提取可能的改进方向。

### 3.2 学科选择

为了确保方法的适用性，本研究选定 4 个临床学科（眼科学、口腔医学、耳鼻咽喉科学、妇产科学）数据集合作为研究对象，使用中图法分类号在中国生物医学数据库中检索 2015 年数据，删除其中无摘要或者有缺漏的数据，最终获取数据 14 331 条，各集合数据量，见表 1。

表 1 4 个学科数据量

| 序号 | 学科     | 数据量（条） |
|----|--------|--------|
| 1  | 眼科学    | 3 708  |
| 2  | 口腔医学   | 1 762  |
| 3  | 耳鼻咽喉科学 | 2 244  |
| 4  | 妇产科学   | 6 617  |

### 3.3 学科背景知识选取

上述学科的上级学科为临床医学，因此选择临床医学数据作为学科背景知识。以背景知识作为训练数据，对后续的特征抽取影响很大，数据量越充足其训练结果就越接近特征的真实分布情况，同时为了避免和学科选择中的数据重叠，本研究从中国生物医学数据库中选取 2011-2014 年临床医学领域的中文期刊论文，剔除无关键词、标题或摘要的数据，最终共获取 923 819 篇题录信息。

### 3.4 数据预处理

将获取到的文本按照标题、关键词、摘要的格式进行字段转换并存储到本地数据库 MySQL 中。由于医学术语专业性强，而中文医学词表及术语又相对缺乏，因此为了确保分词的准确性，将所有背景知识数据中的关键词都作为词表来源，之后再对词表进行进一步处理，包括删除单个字的词、纯数

字、纯英文字母及明显错误的词，最终得到词表中 231 178 个。使用 Python 分词工具 jieba，采用纯字典分词方法及正向最长匹配算法，对背景知识数据集的标题和摘要分词。为了确保该方法的可扩展性和自动化，在此过程中并未有人工对停用词及同义词做任何处理，全部为自动计算结果。表 2 为分词完成后最终得到的背景知识中词频排在前 20 位的词。从背景知识的词频可以看出，词频较高的词都是医学领域中的常用词，这些词在医学领域中的关键词中没有具体意义，但相对于通用领域，如舆情分析、新闻分析领域来说，“临床”、“患者”、“治疗”、“细胞”等词又具有明确的医学特色，因此在科技文本特征抽取中，其学科背景信息对文本抽取方法及质量评价至关重要；同时，在不同级别及不同领域的知识中，通用词数量也各不相同，某些领域通用词数量庞大，因此仅仅通过人工来筛选停用词或者归并同义词并不现实。

表 2 背景知识中词频 TOP20 的词

| 序号 | 词  | 词频      | 序号 | 词  | 词频      |
|----|----|---------|----|----|---------|
| 1  | 方法 | 408 063 | 11 | 观察 | 165 988 |
| 2  | 结果 | 376 357 | 12 | 影响 | 161 990 |
| 3  | 目的 | 368 486 | 13 | 应用 | 160 229 |
| 4  | 结论 | 362 108 | 14 | 差异 | 131 483 |
| 5  | 治疗 | 296 129 | 15 | 比较 | 122 650 |
| 6  | 患者 | 276 905 | 16 | 作用 | 119 157 |
| 7  | 临床 | 247 212 | 17 | 检测 | 116 818 |
| 8  | 分析 | 233 436 | 18 | 意义 | 113 402 |
| 9  | 研究 | 231 373 | 19 | 疗效 | 106 403 |
| 10 | 探讨 | 212 215 | 20 | 细胞 | 103 885 |

### 3.5 背景学科的 IDF 训练

根据 IDF 的计算公式，计算背景学科中每个词的 IDF 值：

$$IDF(t) = \log \left( \frac{\text{背景学科中所有文档数}}{\text{包含 } t \text{ 的文档数}} \right)$$

式中： $t$  为文档中的词。通过计算得到 231 178 个词的 IDF 值，表 3 所示的为 IDF 值最小的 20 个词，IDF 值越小，表明该词在文本集合中的重要性越低。由表 3 可知，IDF 值和词频个数整体上呈反

比关系，但也不能排除某些热点词汇在一些领域频繁出现的现象。

| 表 3 学科背景中 IDF 值最小的 20 个词 |    |      |    |    |      |
|--------------------------|----|------|----|----|------|
| 序号                       | 词  | IDF  | 序号 | 词  | IDF  |
| 1                        | 方法 | 0.36 | 11 | 影响 | 0.76 |
| 2                        | 结果 | 0.39 | 12 | 应用 | 0.76 |
| 3                        | 目的 | 0.40 | 13 | 差异 | 0.85 |
| 4                        | 结论 | 0.41 | 14 | 比较 | 0.88 |
| 5                        | 治疗 | 0.50 | 15 | 作用 | 0.89 |
| 6                        | 患者 | 0.53 | 16 | 检测 | 0.90 |
| 7                        | 临床 | 0.57 | 17 | 意义 | 0.91 |
| 8                        | 分析 | 0.60 | 18 | 疗效 | 0.94 |
| 9                        | 研究 | 0.60 | 19 | 细胞 | 0.95 |
| 10                       | 探讨 | 0.64 | 20 | 手术 | 0.96 |

3.6 学科特征抽取

使用医学术语词典对 4 个学科数据的标题和摘要进行分词，采用训练好的背景学科 IDF 数据，根据 TF-IDF 方法进行特征词抽取。为了查看抽取不同数量特征词时该方法的抽取效果，分别对 4 个学科抽取 5~40 个词，最终得到 36×4 个抽取结果。

4 结果与分析（图 1 - 图 3）

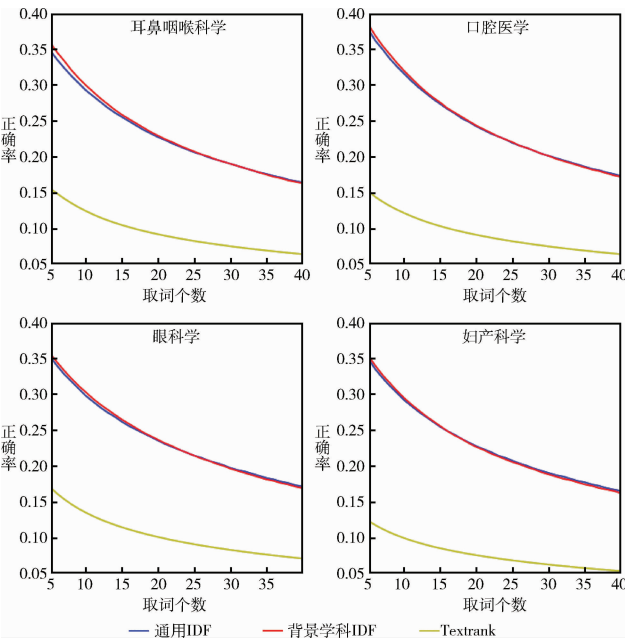


图 1 4 类医学文本特征抽取试验 - 准确率

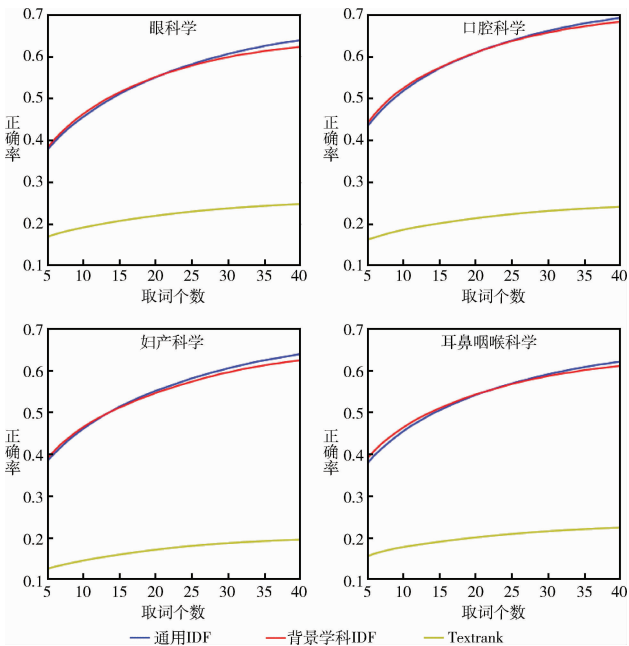


图 2 4 类医学文本抽取试验 - 召回率

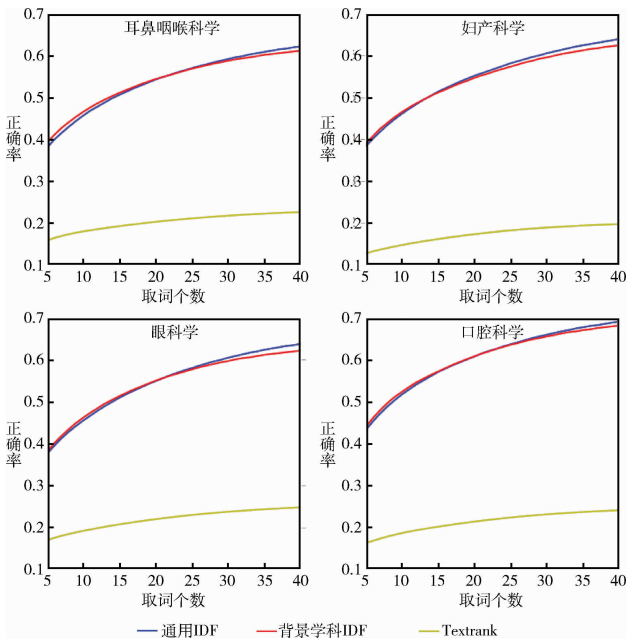


图 3 4 类医学文本抽取试验 - F 值

本研究选取每篇文献中作者标注的关键词作为结果参照数据，同时以通用领域 IDF 的 TF-IDF 方法、无监督学习的 TextRank 方法作为对比，分别计算不同抽取个数时的准确率、召回率和 F 值。可以看到，通用 IDF 方法和背景学科 IDF 方法差距较小，而无监督的 TextRank 方法则效果不明显。从通

用 IDF 和背景学科 IDF 方法的对比来看,在不同学科领域,加入学科背景知识的 TF-IDF 方法在抽取特征词为 20 个左右时,相对于使用通用领域 IDF 的 TF-IDF 方法都有显著提升,但是在抽取特征词大于 20 个时,其准确率又有明显下降。

针对该种情况,通过对两类文本抽取的数据进行详细分析发现,加入学科背景方法对重要词汇的抽取能力要明显强于通用 TF-IDF 方法。表 4 所示为在具体抽取 1 篇文献时两种方法抽取结果的表现。在通用 IDF 方法中,“患者”、“中药”、“我院”、“临床资料”等相对于临床领域并无实际意义的词具有较高的特征值,这是由于这些词相对于通用词汇来说具有明确的医学特征,而在基于学科背景的方法中,这些词的权重明显下降,因而代表文献主题内容的词获得了较高权重。经过大量的实例分析发现,基于学科背景方法在抽取较少词时能够明显过滤掉无领域意义的停用词,而不需要人工参与。在目前的大量抽取工作中,随着分析数据的增多,通过人工来进行停用词表制定已经无法实现,而 1 篇题录信息通常需要抽取的关键词在 10 个以下,已经能够充分表示文章内容,因此该方法可以显著提高文本特征抽取效果,而且在明确的学科领域基础上,具有良好的可扩展性。

表 4 通用 IDF 和学科背景 IDF 方法抽取实例对比

| 顺序 | 通用 IDF 抽取 | 学科背景 IDF 抽取 |
|----|-----------|-------------|
| 1  | 子宫切口      | 子宫切口        |
| 2  | 血肿        | 剖宫产术后       |
| 3  | 剖宫产术后     | 血肿          |
| 4  | B 超检查     | B 超检查       |
| 5  | 患者        | 回声          |
| 6  | 回声        | 治疗后         |
| 7  | 中药        | 2008 年      |
| 8  | 我院        | 使用抗生素       |
| 9  | 临床资料      | 剖宫产后        |
| 10 | 使用抗生素     | 早期诊断        |

5 结语

通过引入学科背景知识,本研究提出了基于学科背景知识的 TF-IDF 方法。结果表明该方法可以在选取少量文本特征时显著提升特征抽取结果,有

效过滤通用停用词;同时,计算得到的 IDF 值可用于辅助人工筛选停用词。但是,本研究只考虑了 IDF 中特征词的分布,在效率上还有一定的改进空间,包括加入语义信息改进同义词的识别效果、加入时间信息通过词的变化趋势揭示词的重要程度等,这些内容都是下一步的研究方向。

参考文献

- 1 李纲,王忠义. 基于语义的共词分析方法研究 [J]. 情报杂志, 2011, 30 (12): 145-149.
- 2 钟伟金. 共词分析法应用的规范化研究——主题词和关键词的聚类效果对比分析 [J]. 图书情报工作, 2011, 55 (6): 114-118.
- 3 赵辉,彭洁,刘润达,等. 基于关键词网络的科技项目多角度演化分析 [J]. 情报学报, 2011, 30 (6): 658-667.
- 4 赵迎光,安新颖,李勇,等. 一种基于生命周期理论的文献热点发现方法——以肿瘤领域为例 [J]. 现代图书情报技术, 2012, (11): 86-91.
- 5 唐晓波,肖璐. 融合关键词增补与领域本体的共词分析方法研究 [J]. 现代图书情报技术, 2013, (11): 60-67.
- 6 邵作运,李秀霞. 共词分析中作者关键词规范化研究——以图书馆个性化信息服务研究为例 [J]. 情报科学, 2012, 30 (5): 731-735.
- 7 胡燕,邱英. 基于改进词共现模型的自动摘要研究 [J]. 计算机与数字工程, 2008, 36 (2): 26-28.
- 8 吴清强,赵亚娟. 基于论文属性的加权共词模型探讨 [J]. 情报学报, 2008, 27 (1): 89-92.
- 9 钟伟金. 基于主要主题词加权的共词聚类分析法效果研究 [J]. 情报学报, 2009, 28 (2): 214-219.
- 10 An XY, Wu QQ. Co-word Analysis of the Trends in Stem Cells Field Based on Subject Heading Weighting [J]. Scientometrics, 2011, 88 (1): 133-144.
- 11 Pinto D, Jiménez-Salazar H, Rosso P. Clustering Abstracts of Scientific Texts Using the Transition Point Technique [C]. International Conference on Intelligent Text Processing and Computational Linguistics, 2006: 536-546.
- 12 徐硕,乔晓东,朱礼军,等. 共现聚类分析的新方法:最大频繁项集挖掘 [J]. 情报学报, 2012, 31 (2): 143-150.
- 13 李树青. 基于三词共现分析的学者主要研究兴趣识别及个性化外文推荐服务的实现 [J]. 情报学报, 2013, 32 (6): 629-639.

(下转第 81 页)

思维方式。信息－知识－智能转化律指导图书馆在开展循证医学信息服务时，明确做什么，以及如何科学地做。医院图书馆必须综合考虑馆藏、人员以及现有服务能力，以临床医护技人员和患者为中心，不断探索和创新，接受现代循证医学模式提出的挑战，以提供文献信息保障服务为基础，以深入临床为切入点，提供临床所需的智能策略。

## 参考文献

- 1 胡大一, 项志敏. 临床医学模式转变—从经验医学到循证医学 [J]. 中国实用内科杂志, 1998, 18 (11): 688 - 689.
- 2 田杰, 姜钦, 陈池梅, 等. 国内外循证医学信息服务研究现状及启示 [J]. 医学信息学杂志, 2016, 37 (4): 74 - 78.
- 3 钟义信. 论“信息－知识－智能转换规律” [J]. 北京邮电大学学报, 2007, 30 (1): 5 - 12.
- 4 钟义信. 知性学引论——信息 知识 智能的统一理论 [J]. 中国工程科学, 2004, 6 (6): 1 - 8.
- 5 Sackett DL, Straus SE, Richardson WS, et al. Evidence

- based Medicine: how to practice and teach EBM. Second Edition [M]. Churchill Livingstone: Edinburgh, 2000.
- 6 张志美, 顾骏, 董建成. 虚拟参考咨询在循证医学信息服务中的应用研究 [J]. 医学信息学杂志, 2006, (3): 223 - 225.
- 7 陈飞霞. 三级医院图书馆构建多元化信息服务模式实践与思考 [J]. 医学信息学杂志, 2013, 34 (1): 82 - 85.
- 8 张萍, 金文力. “提出问题往往比解决问题更重要”——如何在物理教学中培养提问能力 [J]. 学科教育, 2004, (8): 35 - 38.
- 9 崔艳玲. 基于循证医学的信息服务拓展 [D]. 广州: 中山大学, 2004.
- 10 田新玉. 循证医学环境下国外医院图书馆信息服务研究 [J]. 医学情报工作, 2005, (5): 397 - 399.
- 11 陈春英. 医学图书馆与循证医学 [J]. 现代医院, 2007, 7 (4): 138 - 140.
- 12 贺青. 深化医学图书馆信息服务 促进循证医学发展 [J]. 医学信息学杂志, 2010, 31 (6): 62 - 64.
- 13 石志红, 邵杨芳. 图书馆循证医学信息服务模式探析 [J]. 医学信息学杂志, 2014, 35 (5): 76 - 79.

(上接第 54 页)

- 14 Borgatti SP, Everett MG. Models of Core/Periphery Structures [J]. Social networks, 2000, 21 (4): 375 - 395.
- 15 Chen G, Xiao L. Selecting Publication Keywords for Domain Analysis in Bibliometrics: a comparison of three methods [J]. Journal of Informetrics, 2016, 10 (1): 212 - 223.
- 16 陈果. 基于特色关键词的科研机构研究主题揭示: 方法与实证 [J]. 图书情报工作, 2014, (16): 110 - 115.

- 17 Aljaber B, Stokes N, Bailey J, et al. Document Clustering of Scientific Texts Using Citation Contexts [J]. Information Retrieval, 2009, 13 (2): 101 - r131.
- 18 Zhu S, Zeng J, Mamitsuka H. Enhancing Medline Document Clustering by Incorporating MeSH Semantic Similarity [J]. Bioinformatics, 2009, 25 (15): 1944 - 1951.

## 敬告作者

《医学信息学杂志》网站现已开通，投稿作者请登录期刊网站：<http://www.yxxxx.ac.cn>，在线注册并投稿。

《医学信息学杂志》编辑部