基于文献耦合的相似文献推荐算法实现*

范云满 方 安 陈凌云

（中国医学科学院医学信息研究所 北京 100020）

〔摘要〕 在调研并总结相似文献推荐算法的基础上，提出一种基于文献耦合的相似文献推荐算法。分别采用离线式算法和在线式算法进行实现，从算法难度、复杂度、所需计算资源等方面进行对比分析，指出在线式算法更具有优势。

〔关键词〕 综述文献；文献推荐；文献耦合；离线式算法；在线式算法

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2018.04.014

Implementation of Similar Literature Recommendation Algorithm Based on Bibliographic Coupling *FAN Yun－man, FANG An, CHEN Ling－yun, Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China*

〔**Abstract**〕 Based on survey and summarization of similar literature recommendation algorithm, the paper sets forward a similar literature recommendation algorithm on the basis of bibliographic coupling, which adopts off－line algorithm and on－line algorithm separately in the implementation, carries out comparative analysis in aspects like difficulty and complexity of the algorithm and the needed computing resources, and points out that on－line algorithm has an edge.

〔**Keywords**〕 Review literature; Literature recommendation; Bibliographic coupling; Off－line algorithm; On－line algorithm

1 引言

科研工作者在研究一个新的领域时通常需要阅读该领域内比较权威的综述文献（Review Literature），会以某一篇感兴趣的综述文献进行扩展，进一步阅读与该文献类似、相关的文献，从而深入了解该领域的研究现状。科研工作者的这种需求反映到内容分析系统中，即根据用户当前的阅读内容为其推荐合适的文献，使其在较短时间内根据一个兴趣点快速发现与当前内容相似的其他文献，称之为资源推荐服务。

综述文献是对某个领域或某个研究点进行深度且有广度的调研，与一般文献相比较其引文（参考文献）的数量较多。针对综述文献的这一特征，本研究提出一种利用文献耦合的文献推荐算法，为克服文献推荐算法的离线式处理耗时较长、不能随文献的数量线性自适应的缺陷，在离线式算法的基础上进一步优化提出一种在线式的文献综述推荐算法。

2 相关研究

2.1 资源推荐服务

为用户提供资源服务推荐，起源于 20 世纪 90

〔修回日期〕 2017－11－10

〔作者简介〕 范云满，硕士，发表论文 12 篇；通讯作者：方安，副研究馆员。

〔修回日期〕 中央级公益性科研院所基本科研业务费项目“医学数字资源长期保存策略研究”（项目编号：2016ZX330022）。

年代 Goldberg 在构建一个小型系统时提出的为用户提供内容过滤的需求[1]。目前资源推荐服务的方法可以分为 3 种，即基于内容、协同和混合方法。资源推荐服务的方法对比，见表 1[2]。这些方法大都是基于词、文章的内容相关，需利用文章的全文内容才能达到较好的效果。本研究需构建一个领域文献发现系统，当在某个领域中输入关键词时，找到与该关键词相关的综述文献，同时需要对每篇文献推荐与其相似、相关的综述文献。而这些综述文献都是题录信息，缺乏文献的全文信息，采用上述方法效果不好。经分析提出一种利用文献计量学方法进行相似文献推荐的算法。

表 1 资源推荐服务方法对比

推荐方法	基于启发式	基于模型
基于内容	TF - IDF	贝叶斯分类器
	内容聚类	内容聚类
		决策树
		人工神经网络
基于协同	最近邻	贝叶斯网络
	相似点聚类	内容聚类
	图理论	人工神经网络
		线性回归
		概率模型
混合方法	预测评级的线性组合	多种方法的组合
	多种投票方案	构建统一的模型
	多种启发式方法组合	

2.2 文献耦合

文献耦合（Bibliographic Coupling）是一种文献计量学的方法，由美国人开斯勒于 1963 年正式提出，他发现越是学科或专业内容相近的论文，其参考文献中包含的相同参考文献数量就越多。同时引用了一篇论文的两篇论文称为耦合论文，这种关系称为文献耦合[3-4]。文献耦合除文献之间耦合[5]的研究外，还有作者耦合研究（作者文献耦合[6-7]、作者关键词耦合[8]）、期刊耦合研究等[9]。文献耦合由于其意义直观、表达内容具有确定性，能够利用文献之间的引文关系实现文献之间的耦合，从而对文献聚类。

3 基于引文聚类的文献推荐算法

3.1 数据集及软件

3.1.1 数据集 由于综述类文献的参考文献数量较多，因此在为其推荐类似文献时利用文献耦合能够取得较好的结果。本研究利用的数据集为发布于 2015 年 Web of Science 的综述类文献共 81 773 篇，该数据集的参考文献数量为 7 719 894 篇，篇均引用的参考文献约 94 篇。

3.1.2 软件 Elasticsearch 是一个高度可扩展的开源的全文搜索检索和分析引擎。它支持用户存储数据并能在近实时的情况下支持检索和分析海量数据的需求。该引擎中的一个索引可以包含多个文档，每个文档中包含多个字段，可根据文档（数据）自身的层级结构进行文档索引的构建。Elasticsearch 支持按照字段匹配、布尔查询，全文检索、汇总等功能。布尔查询是通过 And、Or、Not 等布尔词对多个字段匹配查询进行组合查询。汇总功能是按照某个字段的值进行分组并分别统计各分组中的值。

3.2 公式定义（表 2）

表 2 公式符号及定义

符号	说明
Ω	文献集
D_k	文献集中的第 k 篇文献
Ω_r	Ω 中所有文献的参考文献的集合
S_k	D_k 的参考文献集合
$S_\cap$	$S_\cap = S_k \cap S_n$，D_k 和 D_n 耦合的参考文献
$\lvert S_\cap \rvert$	耦合强度
r_{ik}	S_k 中的第 i 篇参考文献
R_k	Ω_r 中与 Query 匹配的文档集

文献耦合研究依赖于每篇文献后所附的参考文献。对于一篇文献 D_k，有参考文献集合 S_k；对于文献 D_n，有参考文献集合 S_n。按照文献耦合的定义，取 $S_\cap = S_k \cap S_n$。耦合强度 $\lvert S_\cap \rvert$ 为 $S_\cap$ 的大小，$\lvert S_\cap \rvert$ 越大反映两篇文献的耦合度越大。从而将寻找某一篇文献的相似度文献转化成为文献 D_k 推荐

$|S_{\cap}|$ 排名前 M 篇文献（按照 $|S_{\cap}|$ 由大到小排序）。

3.3 离线式算法

离线式算法即在线下预先计算每篇文献与其他文献的耦合强度，其算法如下所示：

for D_K in 文献集：

for D_n in 文献集 且 D_n ! = D_K：

compute $|S_{\cap k}|$ = $|S_k \cap S_n|$

for D_K in 文献集：

//对 文献集的所有的文献 D_n 的 $|S_{\cap k}|$ 倒排序

$S'_{\cap}$ = Sort（$|S_{\cap k}|$，Desc）

For in in（1，$|S'_{\cap}|$）

记录 D_K 对应的 K

则 $S'_{\cap}$ 中对应的前 M 篇文献即为文献 D_K 的推荐文献集

处理时间总计 18 152 秒（5 小时 2 分 32 秒），平均每篇处理时长 0.22 秒，见表 3。

表 3 离线式算法的处理时间

处理项目	处理结果
数据总量	81 773 篇
比例	100%
总处理时间	18 152 秒
平均每篇处理时间	0.22 秒

离线式处理的优点在于算法简单明了，是最容易想到的解决方案；将需要处理的文献集提前一次性处理完毕，将每篇文献耦合度较高的相似文献按照从高到低的顺序保存，如查询某篇文献的相似文献，只需查询 1 次即可得到结果。而离线式处理的不足在于处理时间较长；当文献集有新加入的文献时，需要对整个文献集的所有文献重新处理，时效性不够好；每次离线处理需要大量的计算资源（CPU、内存），随着文献集的扩大、文献数量的增多，计算资源的需求会出现线性增长，导致单机无法处理，甚至需要服务器集群做分布式处理。鉴于离线式处理的特点，为克服其不足，在离线式算法的基础上提出一种在线式计算文献相似度的算法。

3.4 在线式算法

为实现在线式的算法，首先需要设计一种方便处理的数据结构，见图 1。其中 accessionNumber 字段表示某一篇文献的访问号，其余 author、abbreJournal、fullJournal、iSSN、volume、pageStart、doi、year 为该 accessionNumber 文献的某一篇参考文献信息。

```
"accessionNumber": "WOS:000364608300004",
"author": "Bratton DL",
"abbreJournal": "J BIOL CHEM",
"fullJournal": "JOURNAL OF BIOLOGICAL CHEMISTRY",
"iSSN": "0021-9258",
"volume": "272",
"pageStart": "26159",
"doi": "10.1074/jbc.272.42.26159",
"year": 1997
```

图 1 在线式处理的数据结构

其算法的伪代码如下：

With $D_{K:}$

取得 $D_{K:}$ 的参考文献集合 S_k ：

for i in（1，$|S_k|$）：

query = query + “doi = r_{ik} or”

R_k = Ω_r 中与 query 匹配的文档集

R'_k = GroupBy（R_k，accessionNumber）

for i in（1，$|R'_k|$）

记录 r'_k 对应的 k

则 R'_k 中对应的前 M 篇文献即为文献 D_K 的相似文献集

在线式算法利用 Elasticsearch 的检索分组功能，当检索上述构建的索引时，通过检索 doi 与 S_k 中的 doi 匹配的 accessionNumber，对 accessionNumber 进行分组汇总，得到排名前 N 的 accessionNumber 并生成 D_k 耦合文献，即 D_k 的相似文献。随机抽取 10% 的文献利用在线式算法计算每篇文献的处理时间。

表 4 在线式算法的处理时间

处理项目	处理结果
数据总量	81 773 篇
比例	10%（8 177 篇）
总处理时间	815 秒
平均每篇处理时间	0.1 秒

在线式处理算法是克服了离线式处理算法的缺点，不受文档数量的限制，能够实现在线式的文档

增量更新，无需对全部文档重新计算；查询文档的算法复杂度较离线式算法低一个量级，因此需要的计算资源更少，文档数量增长速度是线性的，而不是幂级；篇均处理时间小于离线式算法。不足之处在于该算法的实现需借助第 3 方软件 Elasticsearch，如没有该工具的支持其实现难度（对每篇文档布尔查询的结果进行分组汇总的算法实现）较大。

3.5 算法对比

经过对离线式算法和在线式算法的分别实现并利用数据集进行实际处理，对处理结果进行分析，将两种算法的优缺点进行对比，见表 5。

表 5 离线式算法和在线式算法对比

对比项目	离线式算法	在线式算法
算法难度	简单	复杂
算法复杂度	$N*N$	N
需要计算资源	大	小
篇均处理时间	0.22 秒	0.1 秒
单次处理时间	数小时	0.1 秒（篇均处理时间）
文档增量更新	需要重新计算	适用
第 3 方工具支持	不需要	需要
横向扩展	难度大	容易
纵向扩展	容易	容易

4 结语

本文首先回顾文献相似度的发展历史并调研文献相似度推荐的各种算法。针对综述文献的参考文献数目多的特点，提出利用综述文献的文献耦合提供相似文献的算法。在具体实现上分别采用离线式算法和在线式算法，从处理时间、算法复杂度、算法难度、所需计算资源等方面进行比较。经对比在线式算法虽然在算法的难度上较离线式算法复杂，但其他各方面都优于离线式算法，因此是一种更值得采用的算法。本文利用文献耦合进行相似文献推荐的算法没有对文献耦合的数量进行归一化处理，即如果两篇文献 A、B 的参考文献数量分别是 10 篇、20 篇，文献耦合的数量 9 篇；文献 C 的参考文献数量是 100 篇，与 B 的文献耦合数量 10 篇，会导致 A 与 B 的相似度小于 C 与 B 的相似度。在下一步的研究中应考虑将算法改进，使 A 与 B 的相似度大于 C 与 B 的相似度。

参考文献

1 Goldberg D, Nichols D, Oki B M, et al. Using Collaborative Filtering to Weave an Information Tapestry [J]. Commun, 1992, 35 (12): 61 - 70.

2 Adomavicius G, Tuzhilin A. Toward the Next Generation of Recommender Systems: a survey of the state - of - the - art and possible extensions [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17 (6): 734 - 749.

3 邱均平, 刘国徽. 国内耦合分析方法研究现状与展望 [J]. 图书情报工作, 2014, 58 (7): 131 - 136.

4 Kessler M M. Bibliographic Coupling Between Scientific Papers [J]. Journal of the Association for Information Science and Technology, 1963, 14 (1): 10 - 25.

5 王立学, 冷伏海. 简论研究前沿及其文献计量识别方法 [J]. 情报理论与实践, 2010, 33 (3): 54 - 58.

6 Zhao D, Strotmann A. Evolution of Research Activities and Intellectual Influences in Information Science 1996 - 2005: introducing author bibliographic - coupling analysis [J]. Journal of the Association for Information Science and Technology, 2008, 59 (13): 2070 - 2086.

7 陈远, 王菲菲. 基于 CSSCI 的国内情报学领域作者文献耦合分析 [J]. 情报资料工作, 2011, (5): 6 - 12.

8 刘志辉, 郑彦宁. 基于作者关键词耦合分析的研究专业识别方法研究 [J]. 情报学报, 2013, 32 (8): 788 - 796.

9 李秀霞, 马秀峰, 程结晶. 融入引文内容的期刊耦合分析 [J]. 图书情报工作, 2016, 60 (11): 100 - 106.