情报学报 ISSN 1000-0135
第31卷第10期1052-1061,2012年10月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol. 31 No. 10, 1052-1061 October 2012

doi:10.3772/j.issn.1000-0135.2012.10.006

一种基于引用上下文和引文网络的相关反馈算法[1)]

吴夙慧[1] 成 颖[1] 郑彦宁[2] 潘云涛[2]

(1.南京大学信息管理系,南京 210093;2.中国科学技术信息研究所,北京 100038)

摘要 相关反馈是一种根据用户或系统的相关性判断重构初始检索提问的方法,已被证明可以有效地改进检索效果。具体到学术文献,其引用关系表征了文献内容上的相关性,因而可以为相关反馈提供有价值的辅助信息。本文提出了一种基于引用上下文、文献同被引和文献耦合的相关反馈改进算法。该算法的基本思想包括:利用学术文献的引用上下文信息扩充词包模型(bags of words)进行文本表示;在相关文献判断阶段利用相关文献在引文网络中与其他文献的同被引强度和耦合强度扩充相关文献集合;结合基于聚类的相关反馈思想抽取查询扩展项。实验证明该算法提高了相关反馈效果。此外,相关分析的结果表明文献同被引以及文献耦合强度与文献内容相似度具有显著的相关性。

关键词 相关反馈 引用上下文 同被引 文献耦合 聚类

A Relevance Feedback Algorithm Based on Citation Context and Citation Network

Wu Suhui[1], Cheng Ying[1], Zheng Yanning[2] and PanYuntao[2]

(1. *Department of Information Management, Nanjing University, Nanjing* 210093;
2. *Institute of Scientific and Technical Information of China, Beijing* 100038)

Abstract Relevance feedback is a method for refactoring retrieval query according to the relevance judgment by system or user. It is proved to improve retrieval result effectively. And for the information retrieval on academic literature, the reference relationship characterizes the correlation on content, so the reference relationship can provide supplementary information in relevance feedback. In this paper, a novel relevance feedback algorithm based on citation context, co-citation and bibliographic coupling is proposed. A citation context is the text surrounding the reference markers used to refer to other scientific works. The citation context can provide additive terms to represent the academic literature, this algorithm use citation context to expand the "bags of words" model. In the stage of relevance judgment, we use the relation of co-citation and bibliographic coupling in citation network to expand the set of relevance document. Finally, the algorithm uses the clustering method to extract terms to expand query in relevance document. Experimental results show that the retrieval quality is improved. In addition, we investigate the correlation of co-citation, bibliographic coupling and literature content by correlation analysis in statistics.

Keywords relevance feedback, citation context, co-citation, bibliographic coupling, clustering

收稿日期:2012年2月9日

作者简介:**吴夙慧**,男,1988年生,硕士生,主要研究方向:自然语言处理。E-mail:wush13@126.com。**成颖**,男,1971年生,副教授,主要研究方向:信息检索。**郑彦宁**,男,1965年生,研究员,主要研究方向:情报技术与方法、竞争情报。**潘云涛**,女,1967年生,研究员,主要研究方向:科学计量学。

1) 本文得到国家社科基金项目"中文学术信息检索系统相关性集成研究"(项目批准号10CTQ027)、教育部人文社会科学研究规划基金项目"面向用户的相关性标准及其应用研究"(项目批准号07JA870006)及中国科学技术信息研究所合作研究项目的资助。

1 引 言

在信息检索中,由于检索语言的二义性、用户信息素质的参差不齐以及用户对数据集和检索系统的了解有限等原因,检索系统所反馈的检索结果往往容易偏离用户的信息需求,查全率和查准率都不能令人满意。对于检索用户而言,根据反馈的检索结果来修正自己的检索提问往往是十分困难的,因而就需要检索系统采用各种方法对用户的检索提问进行修正和加工,以期获得更优的检索结果。其中,相关反馈技术是被广泛研究且被证明效果较好的方法。相关反馈技术是指利用用户(用户参与的相关反馈)或系统(伪相关反馈)对前一次检索结果的相关性判断,系统自动重构检索提问,将相关文献中的特征项加入到检索提问中或提高其权重,同时抑制不相关文献中的特征项,反复迭代这一过程直到得到满意的检索结果的技术[1]。

具体到学术信息检索系统,由于学术资源具有更高的专业性,用户对检索结果的甄别也更为困难,因此对于学术检索系统的相关反馈技术相应地提出了更高的要求。同时学术文献也拥有其他类型文本所不具备的传播特征,学术文献之间通过引用关系联系起来构成引文网络,被引文献和施引文献、耦合文献以及同被引文献之间存在内容上的相关性[2],这些引用关系可以为传统的基于文本内容的信息检索研究提供很多的辅助信息。

基于此,本文提出了一种改进的相关反馈算法,利用引文网络中文献间的引用关系、文献耦合和同被引关系对相关反馈中的文本表示和文献相关性判断等步骤进行扩展,并结合文本聚类的思想,对传统的基于文献内容的相关反馈方法进行改进。通过与三种对比算法的比较,证明了算法的优越性。另外,通过对实验数据的相关分析,证实了文献的耦合强度和同被引次数与文本内容的相似度存在显著的相关性,也为本文提出的算法提供了更好的支持。

2 相关研究

关于相关反馈最早的研究可以追溯到 Rocchio[3] 的研究,他通过实验证明了相关反馈能够改进信息检索的质量,随后学界对此进行了一系列的研究和实验。根据文本表示模型的不同,这些研究可以分为基于向量空间模型的相关反馈技术,基于概率模型的相关反馈技术和基于布尔模型的相关反馈技术[1]。

2.1 基于聚类的相关反馈

在对于相关反馈的众多研究中,一种非常巧妙的解决思路是基于聚类思想的相关反馈方法,这些研究将文本聚类思想应用到相关反馈中的相关文档判断上,取得了不俗的实验结果。

Buckley 等[4]的研究认为文献之所以与用户查询相关是由于文献包含了用户所需要的关键概念,他对前一次查询得到的结果集的前 N 篇文献(根据伪相关反馈的思想默认为相关文献)进行聚类,从每一个聚簇中抽取权值较大的特征项,再从与初始查询相似度较大的几个聚簇中抽取出一些特征项,把这些特征项加入查询中,通过在 TREC6 上的实验证明了这样的查询扩展方法可以将一些用户查询提问的超概念加入到查询中,得到了更好的检索结果。Iwayama[5]采用了另一种假设,他认为相关文献之间常常是较为相似的,而不相关文献则是较为松散的,因而他通过对前一次检索得到的前 N 篇文献进行聚类,考察所聚成的若干个聚簇,按照聚簇中所包含的相关文档比例对聚簇排序,从排名较高的几个聚簇中抽取特征项加入用户查询,实验也表明当用户初始检索提问质量较低时,算法可以获得更好的查全率和查准率。Choi[6,7]等利用 Salton 和 Fox[8]提出的扩展布尔检索模型,采用类似 Buckley 的聚类方法进行相关反馈,得到了比经典 DNF 反馈方法更好的检索效果。另外 Lee[9]、钟敏娟等[10]的研究也都采用了基于聚类的相关反馈思想。

2.2 文献同被引和文献耦合

文献同被引概念最早是由 Small[11]提出的,如果两篇文献同时被其他论文引用,则称这两篇文献同被引,而文献耦合是引文网络中另一种文献之间的关系,即如果两篇文献同时引用其他论文,则称这两篇文献互相耦合。同被引和耦合关系都反映了引文网络中两篇文献的联系,两者主要的不同在于两篇文献之间的耦合强度是固定的,而同被引次数是随着时间动态变化的,具有更好的前瞻性[12]。这两种分析方法现在已经被广泛地应用到学术领域热点分析、学术论文聚类等[13,14]方面。

同被引分析和耦合分析方法认为如果两篇文献被大量其他文献共同引用,或拥有较多的共同参考文献,则它们的内容常常具备较强的相关性。用社

会网络分析的观点来解释，即当两篇文献在引文网络中拥有大量的共同邻居节点，则这两篇文献具备内容上的相关性，且随着共同邻居数量的增加，内容相关性也增加。

基于这一假设，近年来的一些研究将文献之间的引用信息应用到文本聚类上，章成志等[15]采用类似于网页 PageRank 值计算的思想计算论文的 PageRank 值，计算中考虑了每篇施引文献的被引量和引文数，以此作为加权值运行“样本加权 K-means 聚类算法”，改善了聚类效果。吴夙慧等[16]采用论文的同被引关系进行先期聚类，得到 K-means 算法的初始聚类中心和 K 值，然后运行 K-means 算法也获得了更优的实验效果。

本文提出的相关反馈算法也将运用这一假设，并在 4.1 节的实验中将通过实验数据来验证这一假设。

2.3 引用上下文

引用上下文（citation context）是指一篇论文在引用其他文献时，引用标识附近的上下文。一般而言，在作者进行引用行为时，会对被引文献的内容进行简单扼要的概括，来自施引文献的这些概括内容是对被引文献文本内容很好的扩展，常常包括了原文所没有的特征项。

由于引用上下文的优点，它在科技文献的信息检索和自然语言处理领域被广泛应用，Nakov 等[17]分析概括了引用上下文的五大研究用途：作为对比语料库的来源、文献的自动文摘研究、一词多义和一义多词的消解、实体识别和关系抽取以及提高引文索引的检索功能。Mercer 和 Marco[18]认为在进行引文索引设计时，除了需要考虑引用链接以外，还需要将引用上下文纳入引文索引中。Bradshaw[19]在对检索结果排序时考虑了从引用上下文中提取的特征项与查询的相似度以及论文的引用量，得到的排序结果准确率更高。Elkiss[20]对引用上下文的文本表示能力进行了详细的定量研究，结果显示引用上下文与文摘相比，提供了更多的额外信息，这些额外的信息对文本检索和自动文摘都有着十分重要的意义。Aljaber 等[21]采用 K-means、层次聚类等多种聚类算法对传统的文本内容表示方法，基于引用上下文的文本表示方法以及文本内容和引用上下文混合表示方法的效果进行了比较，实验结果显示混合表示方法获得了最好的聚类效果。Ritchie 等[22,23]在计算语言学论文集上做了相似的实验，得到了同样的结果。

从以上研究中可以看出，文献耦合、文献同被引和引用上下文在学术文本检索领域在当前还主要应用于文本聚类，它们对文本表示的补充效果得到了以上一系列研究的证实。在此基础上，本文将其应用于相关反馈研究，并结合基于聚类的相关反馈思想[4~10]，以期获得更好的反馈效果。

3 基于引用上下文和引用网络的相关反馈算法

与 Rocchio 提出的经典相关反馈算法相比，本文提出的相关反馈算法的主要改进体现在文本表示和反馈算法两个部分。其中，文本表示综合利用文本内容和引用上下文得到文本向量；反馈算法将文献的引用关系纳入算法之中，引用关系可以很好地反映文献的内容相关性，而其在学术检索系统中是容易获得的；另外，算法还结合了基于聚类的相关反馈思想。

3.1 基于引用上下文的文本表示

本算法采用文本内容和引用上下文相结合的文本表示方法来构建文本向量。对于文本集 D 中的每一篇文献 d_i，其文本内容通过停用词移除、后缀消除、特征降维以及 tfidf 权重计算等步骤可以表示为一个文本向量 $d_i(w_{i1},w_{i2},\cdots,w_{ij})$。而对于文献 d_i 而言，还同时拥有一个引用上下文集合 $C_i\{c_{i1},c_{i2},\cdots,c_{im}\}$，其中 C_i 是引用 d_i 的文献所形成的集合，m 为文献 d_i 的引用上下文的数目，对这些引用上下文也同样可以得到另一个足以反映其文献特征的特征项集合 $CW_i\{cw_{i1},cw_{i2},\cdots,cw_{in}\}$。从向量 CW_i 中选取最优的特征项加入集合 d_i 就可以得到 d_i 新的文本向量 $(w_{i1},w_{i2},\cdots,w_{ik})$。

该步骤存在两个重要的问题，一是如何得到精确的引用上下文，避免损失引用信息或受到噪声信息的干扰。二是如何从引用上下文中选取最佳的特征项加入原始的文本向量中。

第一个问题，Bergmark[24,25]、Powley 和 Dale[26]、Councill 等[27]进行了卓有成效的研究，这些研究的主要思想是首先提取参考文献列表中每条引文的作者、发表年、题名等信息，由于引文标注都拥有统一的规范，因而利用正则表达式匹配就可以获得每一篇引文的作者和发表年等信息。然后在正文中利用自然语言处理中的命名实体识别算法[28~30]识别出

人名、年代这样的命名实体并与引文中提取出的各项信息进行匹配，最终定位到正确的引用上下文。而本文采用的数据集是从 Biomed 数据库下载的 HTML 文件，Biomed 数据库已经人工在施引标识与参考文献列表之间进行了链接，因而可以很方便地定位到引用上下文。

第二个问题，本文采用了 Aljaber 等[21]的做法：

（1）首先从文献的引用上下文中移除停用词，并对英文单词进行后缀消解，本文采用的停用词表是 Salton[31]的停用词表。

（2）计算剩余特征项的 tfidf 值，选取前 30% 的特征项，以去除掉不重要特征项的干扰。

（3）将剩余的特征项加入到被引文献的文本向量中，如果原集合中已经拥有该特征项，则选取较大的 tfidf 值。

Aljaber 等[21]在论文中统计了原始文本向量和加入了引用上下文特征项的文本向量的余弦相似度。发现相似度集中在 0.2 ~ 0.4 之间，相似度并不显著，因而也说明了新的文本表示方法提供了许多新的特征项，扩充了原文的语义特征。

3.2 基于同被引和耦合的相关反馈算法

本文提出的相关反馈算法在相关文献判断阶段引入高同被引文献和耦合强度大的文献扩充相关文献集，并采用先期聚类的方法，从各聚簇中选择特征项进行查询扩展。

算法描述需要使用到文献耦合矩阵和文献同被引矩阵。文献耦合矩阵可以表示为公式(1)：

$$\begin{pmatrix} C_{00} & C_{01} & \cdots & C_{0(m-1)} \\ C_{10} & C_{11} & \cdots & C_{1(m-1)} \\ C_{(m-1)0} & C_{(m-1)1} & \cdots & C_{(m-1)(m-1)} \end{pmatrix} \quad (1)$$

其中，C_{ij} 表示学术文献 i 和 j 的共有参考文献条数，当 $i = j$ 时表示学术文献 i 的参考文献总数，可简写为 C_i。

类似地，同被引矩阵也可以按照同样的公式表示，此时 C_{ij} 表示学术文献 i 和 j 的同被引次数，C_i 表示为文献 i 的总被引次数。

算法基本描述如下：

Input 初始查询向量 $q(q_1, q_2, \cdots, q_n)$，扩展文本向量集合 $D\{d_1, d_2, \cdots, d_m\}$，其中每一个文本向量 d_i 都是由 3.1 节得到的扩展文本向量 $d_i(w_{i1}, w_{i2}, \cdots, w_{in})$，同被引矩阵，耦合矩阵。

Output 检索结果列表 L

Step：

Step 1 计算查询向量 q 与 D 中每一个 d_i 的相似度，得到一个按照相似度降序排列的检索结果列表 L，根据伪相关反馈的思想，默认列表 L 的前 N 篇文献为相关文献，加入相关文献集合 RD 中，再考察 RD 集合中的每篇文献在同被引矩阵和耦合矩阵中与其他文献的耦合强度和同被引次数，选取一个阈值 P，将耦合强度 + 同被引次数（即共同邻居数）最大的 P 篇文献作为相关文献，加入集合 RD 中。

Step 2 对 RD 集合中的文献进行聚类，聚类算法采用 K-means 算法[16]，得到 K 个聚簇，每个聚簇 C_i 的聚簇中心向量记为 $C_i(w_{i1}, w_{i2}, \cdots, w_{in})$。计算 C_i 与查询向量 q 的相似度，得到排在前 $K/3$ 的聚簇。

Step 3 给定用于扩展查询的特征项集合 T，选定阈值 m，从 Step2 中得到的全部 K 个聚簇中抽取权值较大的前 m 个特征项加入集合 T，再从 Step2 中得到的前 $K/3$ 个聚簇中抽取剩余特征项中权值较大的前 m 个特征项也加入集合 T。根据集合 T 中的特征项及其 tfidf 值用 Rocchio 公式修正查询向量 q。

Until：转 Step1，直到到达事先设定的迭代次数。

其中 Step3 的查询扩展特征项抽取方法采用了 Buckley 等[4]提出的方法，在相关研究部分 2.1 节已经叙及。算法的另一个问题是在 Step2 中的 K-means 聚类中如何选取最佳的 K 值，此处的 K 值选取采用了吴夙慧等[16]提出的方法，利用学术信息检索系统中易于获得的同被引矩阵进行先期层次聚类，每一次迭代将最近的两个文本聚在一起，经过 $n-1$ 次迭代后将得到一个逐步增大的聚类距离序列集合 $D = \{D_1, D_2, \cdots, D_{n-1}\}$，计算 $\Delta D_i = D_{i+1} - D_i$，使得 ΔD_i 取得最大值的那一次迭代即为最佳的 K 值。由于进行聚类的数据集较小，因而可以很快地得到 K 值和初始聚类中心。

实验中参数 N 和 m 的选取主要参照了 Buckley[4]的研究，其中 $N = 20, m = 100$。Rocchio 公式的参数选择了较常用的 $\alpha = 1, \beta = 1$[32]，P 值选取 5。

图 1 给出了整个相关反馈算法的步骤。

4 实验与结果分析

本文的实验分为两部分，第一部分实验验证文献耦合强度、文献同被引次数和文献内容相似度之间的相关性假设。第二部分实验通过对比算法的比较，检验本文提出的相关反馈算法的效果。

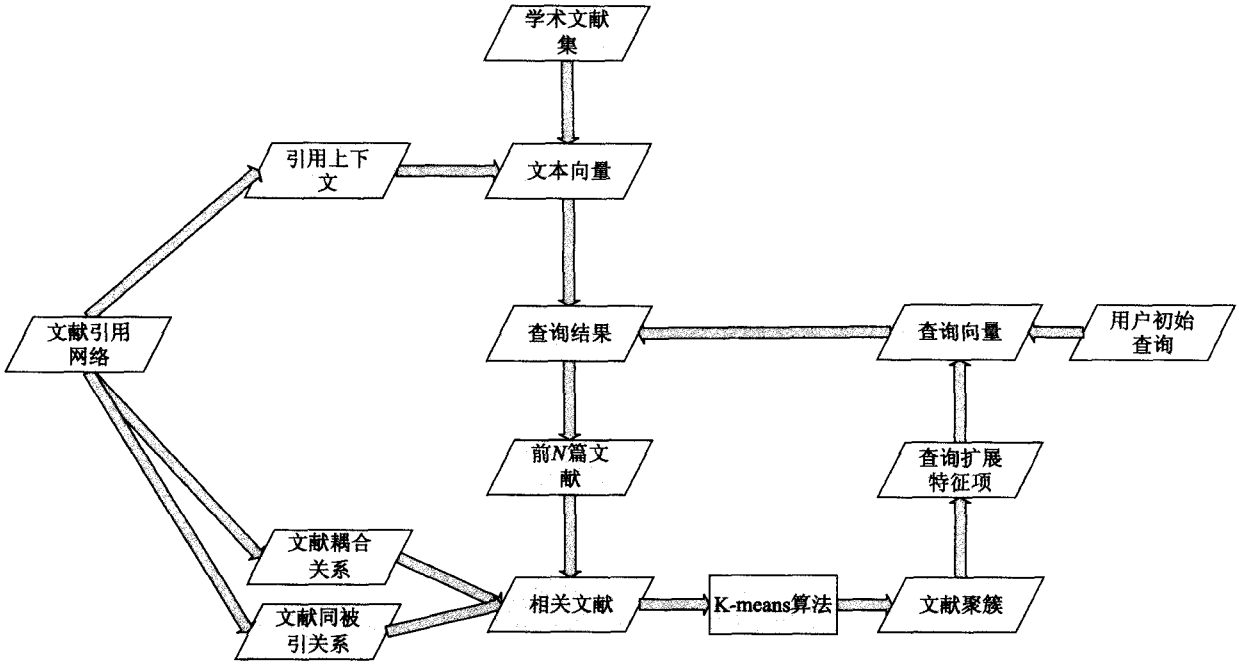

图1 算法基本流程

4.1 耦合强度、同被引次数和内容相似度相关性分析

4.1.1 数据来源和实验步骤

实验选用了两部分数据，第一部分数据来自 Web of Knowledge 数据库，因为 ISI of Knowledge 数据库可以较简单地获得一篇文献的参考文献和施引文献，在该数据库中检索情报学领域的五个研究方向 information retrieval, data mining, knowledge management, citation analysis 和 competitive intelligence，各选取 50 篇高被引论文，共 250 篇论文，套录其题名、文摘、施引文献和参考文献。第二部分数据选择 BioMed 数据库中的论文，该数据库收录有 100 多种期刊，从 BioMed 数据库中选取生物医学领域的一个综合性期刊 Journal of Biomedical Science，下载被浏览最多的 500 篇文献的全文 html 文件，由于该数据库中论文具有统一格式，自编程序获得每篇论文的参考文献、题名、文摘和正文。

按照 3.2 节的方法分别得到两部分数据的耦合矩阵和同被引矩阵，并将第一部分 250 篇论文的题名和文摘表示为文本向量，将第二部分 500 篇论文的题名、文摘和正文表示为文本向量，其中题名和文摘赋予两倍的权值。计算每两篇论文之间的文本相似度，得到文本相似度矩阵。

4.1.2 实验结果及分析

第一部分数据得到的文本相似度矩阵实际上是题名-文摘相似度矩阵，选用的相似度计算方法是余弦相似度。统计得到同被引次数的范围为 0 ~ 72，耦合次数的范围为 0 ~ 18，对同被引次数、耦合次数和题名-文摘相似度分别进行相关性分析，得到分析结果如表 1、表 2 所示。

表1 同被引次数-文摘相似度相关性分析

		同被引次数	题名文摘相似度
同被引次数	Pearson Correlation	1	0.643**
	Sig. (2-tailed)		0.000
	N	31125	31125
题名文摘相似度	Pearson Correlation	0.643**	1
	Sig. (2-tailed)	0.000	
	N	31125	31125

表 2 耦合次数-文摘相似度相关性分析

		耦合次数	题名文摘相似度
耦合次数	Pearson Correlation	1	0.584*
	Sig. (2-tailed)		0.003
	N	31125	31125
题名文摘相似度	Pearson Correlation	0.584*	1
	Sig. (2-tailed)	0.003	
	N	31125	31125

从表 1 和表 2 可以看出同被引次数、耦合次数和题名-文摘相似度存在统计学意义上的相关性。另外，相比耦合次数、同被引次数和题名文摘相似度之间的相关性更加显著。

第二部分数据得到的文本相似度矩阵采用全文进行计算，同样选用余弦相似度计算方法。统计得到同被引次数的范围为 0～70，耦合次数的范围为 0～23，对同被引次数、耦合次数和全文相似度分别进行相关性分析，得到的分析结果如表 3、表 4所示。

由表 3、表 4 可以发现，同被引次数、耦合次数和全文相似度都存在显著的相关性。其中同被引次数和全文相似度的相关性更为显著。

综合两部分实验得到的实验结果，可以认为文献的同被引强度和耦合强度与文献的内容相似度存在统计学意义上的相关性，本文的实验为该假设提供了数据支持。数据表明，相比仅使用题名和文摘的文本表示方法，采用文献全文进行文本表示，其与同被引强度和耦合强度之间的相关性更显著，因而在后文 4.2 节的算法性能分析中将采用文献全文来进行文本向量表示。

4.2 基于引用关系的相关反馈算法性能分析

4.2.1 数据来源和实验步骤

实验数据同样采用 BioMed 数据库中的论文，从期刊 Journal of Biomedical Science 中下载 1000 篇文献全文的 html 文件，使用正则匹配的方法获取每篇论文的参考文献、题名、文摘和正文。利用这些信息生成同被引矩阵、耦合矩阵和文本向量。

实验采用三种对比算法以测试新算法对学术信息检索相关反馈的改进效果，第一种对比算法与本文提出算法的区别是不采用引用上下文进行文本向量的扩展，其他相同（以下简称算法 1）；第二种对比

表 3 同被引次数-全文相似度相关性分析

		同被引次数	全文相似度
同被引次数	Pearson Correlation	1	0.697**
	Sig. (2-tailed)		0.000
	N	124750	124750
全文相似度	Pearson Correlation	0.697**	1
	Sig. (2-tailed)	0.000	
	N	124750	124750

表 4 耦合次数-全文相似度相关性分析

		耦合次数	全文相似度
耦合次数	Pearson Correlation	1	0.661**
	Sig. (2-tailed)		0.000
	N	124750	124750
全文相似度	Pearson Correlation	0.661**	1
	Sig. (2-tailed)	0.000	
	N	124750	124750

算法与本文提出算法的区别是不采用同被引文献和强耦合文献进行相关文献集的扩展，其他相同（以下简称算法2）；第三种对比算法与本文提出算法的区别是不采用聚类的方法对相关文献进行处理，而直接用Roccio公式相关反馈，其他相同（以下简称算法3）。参数的选取已在3.2节说明，选取 $N=20, m=100, P=5$。相关反馈迭代次数为3次，此外，Rocchio公式的参数选择 $\alpha=1$，$\beta=1$。

4.2.2 评价指标

对于信息检索效果的评价，查全率（Recall）和查准率（precision）是被广泛应用的指标，在此基础上也提出了一系列改进的评价方法。本实验将采用P@N和11点平均查准率（11point precision average）两种评价方法对检索结果进行评价。

其中P@N方法是指查询结果中前 N 篇文献的查准率，即排在检索结果前 N 篇的文献中相关文献所占的比例。

11点平均查准率（11point precision average）评价方法由Harman[33]提出，检索系统所返回的检索结果列表是一个按照与查询相似度递减的列表 $D\{d_1, d_2, \cdots, d_n\}$，对于列表中的每一个元素 d_i 都对应一个查全率和查准率，d_i 处的查全率等于排在 d_i 之前的相关文献占整个文档集合中相关文献的比例，d_i 处的查准率等于排在 d_i 之前的文献中相关文献所占的比例，其计算方法如公式（2）、公式（3）所示：

$$Recall_i = \frac{N_{(1\cdots i)rel}}{N_{rel}} \tag{2}$$

$$Precision_i = \frac{N_{(1\cdots i)rel}}{N_{(1\cdots i)}} \tag{3}$$

其中，$N_{(1\cdots i)rel}$ 表示 d_1 到 d_i 中相关文献的数目，$N_{(1\cdots i)}$ 表示 d_1 到 d_i 文献数目，$N_{(1\cdots i)}=i$，N_{rel} 指整个文献集中相关文献的总数。按照 $Recall_i$ 的值将列表分为等距离的10个区间，计算11个分隔点对应的 $Precision_i$ 值，$Precision_i$ 值越大表示检索效果越好。

4.2.3 实验结果分析

实验从Mesh词表中几个不太相关的领域选取10个关键词，它们是sugar alcohol（糖醇），DNA virus（DNA病毒），immune system（免疫系统），toxic actions（毒副作用），dietary carbohydrates（饮食碳水化合物），public health（大众健康），early diagnose（早期诊断），lung cancer（肺癌），behavioral symptoms（行为表征），endophyte（内生菌），用这10个关键词作为初始查询进行检索，分别在四种算法下得到检索结果，并求出它们的P@10、P@20值，如表5、表6所示。

从表5、表6可以发现，相比三种对比算法，新算法取得了更好的查准率，因而新算法的三项改进都对检索效果的提高起到了促进作用。

再采用11point precision average方法进行评价，对十个查询求出11个点的平均查准率，得到四种算法查全率和查准率的二维图如图2所示。从图2中可以发现，新算法在11个点处的平均查准率都优于其他三种对比算法，与上面的P@N指标评价结果吻合。

表5 10组初始查询的P@10值

初始查询	算法1	算法2	算法3	新算法
sugar alcohol	0.5	0.6	0.6	0.7
DNA virus	0.6	0.8	0.7	0.7
immune system	0.7	0.5	0.8	0.7
toxic actions	0.5	0.8	0.6	0.7
dietary carbohydrates	0.6	0.6	0.5	0.9
public health	0.4	0.5	0.4	0.6
early diagnose	0.6	0.7	0.5	0.6
lung cancer	0.5	0.6	0.6	0.5
behavioral symptoms	0.7	0.8	0.6	0.7
endophytes	0.5	0.4	0.5	0.6
均值	0.56	0.63	0.58	0.67

表 6　10 组初始查询的 P@20 值

初始查询	算法 1	算法 2	算法 3	新算法
sugar alcohol	0.7	0.8	0.7	0.85
DNA virus	0.8	0.85	0.9	0.9
immune system	0.75	0.75	0.7	0.8
toxic actions	0.65	0.9	0.75	0.85
dietary carbohydrates	0.7	0.7	0.75	0.8
public health	0.65	0.7	0.6	0.7
early diagnose	0.75	0.7	0.65	0.7
lung cancer	0.65	0.65	0.7	0.9
behavioral symptoms	0.75	0.8	0.7	0.85
endophytes	0.65	0.7	0.75	0.8
均值	0.705	0.755	0.720	0.815

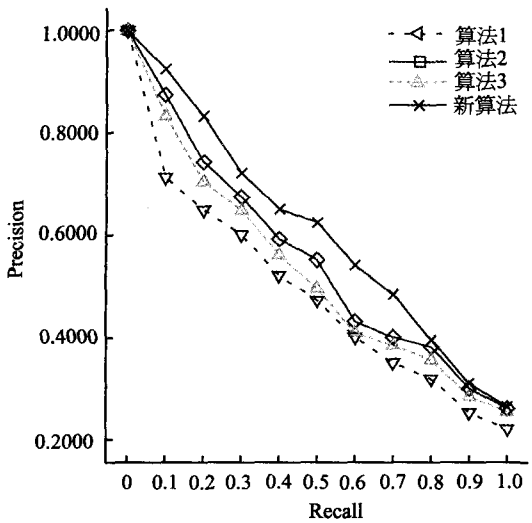

图 2　查全率和查准率二维对应图

4.2.4　实验结果的假设检验

从 4.2.3 节的图表中可以看到新算法的各项指标都优于三种对比算法，但是这些改进是否具有统计学上的显著性还需要通过假设检验来验证。本文对三种对比算法得到的 P@10、P@20 和 11 个点的平均查准率与新算法得到的相应指标进行配对 t 检验，得到的结果如表 7 所示。

从表 7 中可以看到，在 95% 的置信区间下，P 值都等于 0.000、0.000、0.000，通过了检验，因而可以认为相比三种对比算法，本文提出的相关反馈算法在检索效果上的改进具有统计学意义。

5　结　语

本文利用学术文献的引用上下文和同被引、耦合关系，提出了一种扩展的相关反馈算法。算法利用引用上下文扩展学术文献的内容信息，利用同被引、耦合关系扩展相关文献集，并结合基于聚类的相关反馈思想，取得了较好的效果。此外本文还存在一些不足和可扩展之处值得思考。

（1）本文提出的算法在进行相关反馈和查询扩展时利用了学术文献的引用关系和同被引关系。事实上，学术文献的引用关系和同被引关系包括很多不同的类型，如同被引关系可以细分为同一篇文章中被引，同一篇文章的同一章中被引，同一篇文章同

表 7　配对 t 检验结果

		Paired Differences			T	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean			
Pair 1	算法 1 - 新算法	-0.122283	0.048483	0.013996	-8.737	11	0.000
Pair 2	算法 2 - 新算法	-0.053167	0.033166	0.009574	-5.553	11	0.000
Pair 3	算法 3 - 新算法	-0.083478	0.040068	0.011567	-7.217	11	0.000

一段中被引,同一篇文章同一句被引。根据 Elkiss 的研究[20],这些不同类型的同被引关系表征着文本不同的语义相似度。

另外,学术文献之间的引用关系也存在不同的类型,规范的学术文献通常拥有较为固定的格式,一般来说,在论文引言(introduction)、相关研究(related work)和结论(conclusion)部分的施引,引用标识附近的上下文更能反映被引文献的内容,以这部分内容作为查询扩展的来源可以获得更好的检索结果,而在学术论文的数据处理和结果分析部分的施引,其引用标识附近的上下文则往往相似度有限甚至无关。在本文中,没有对这些不同类型的引用和同被引关系进行区分,而如果对这些表征不同语义相似度的关系进行不同赋权处理,应该可以得到更好的结果。

(2) 一篇学术论文拥有各方面的内容特征和外部特征,可以将其表示为一个这样的集合:学术文献{内容,作者,发表期刊(会议),机构,基金,出版时间,引文,被引}。成颖[34]通过对学生用户学术文献检索作业的内容分析,发现以上的文献特征都会对用户的相关性判断产生较大的影响。本文的研究采用了内容、引文和被引三项特征,而其他的特征也提供了许多相关性判断的依据,因而这些特征对检索系统的相关反馈、查询扩展和相关性排序等研究也应该有着很大的意义。

(3) 网页的链接网络将互联网中的网页连接在一起,具有许多与引文网络相似的特征。因而共链信息和链接的锚文本对于网页的检索也应该具有很大的意义。共链分析方法最早由 Larson[35]在 1996 年提出的,随后 Thelwall 和 Wilkinson[36]、Vaughan 和 You[37]、Wang 和 Kitsuregawa[38]等利用这一方法进行了一系列卓有成效的研究,这些研究主要集中网页的聚类上。但是引用行为和链接行为有一个很大的不同,学术文献的引用行为是由受过专业训练的学者按照学术规范进行的学术行为,而链接行为则是一种较为随意和普遍的传播行为,而且由于商业利益的驱使,一些网页会故意地链接一些无关网页,并加上一些吸引眼球而内容无关的锚文本,因此也造成了许多的噪声信息。所以在网页检索、自动文摘、自动聚类等研究中能否考虑锚文本和链接结构信息也是值得研究的课题。

引用信息是学术文献极为重要且易于获得的信息,在引文网络上的挖掘可以获得许多文本内容所不能揭示的辅助信息,因而在这一领域的研究依然大有可为,而如何解决上述的三个问题也将是一个值得思考的命题。

参 考 文 献

[1] 宋玲丽,成颖,单启成. 信息检索系统中的相关反馈技术[J]. 情报学报,2005,24(1):34-41.

[2] 邱均平. 论"引文耦合"与"同被引"[J]. 图书馆,1987,(3):13-19.

[3] Rocchio J. Relevance Feedback In Information Retrieval. The SMART Retrieval System. Englewood Cliffs: Prentice-Hall, Inc, 1971:313-323.

[4] Buckley C, Mitray M, Walz J, et al. Using Clustering and SuperConcepts Within SMART : TREC6[J]. Information Processing & Management, 2000, 36(1):109-131.

[5] Iwayama M. Relevance Feedback with a Small Number of Relevance Judgements: Incremental Relevance Feedback vs. Document Clustering[C]// Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval. USA, New York, 2000:10-16.

[6] Choi J, Kim M, Raghavan V V. Adaptive Feedback Methods in an Extended Boolean Model[C]//Proceedings of ACM SIGIR workshop on mathematical/formal methods in information retrieval, 2001:42-49.

[7] Choi J, Kim M, Raghavan V V. Adaptive relevance feedback method of extended boolean model using hierarchical clustering techniques [J]. Information Processing and Management, 2006, 42:331-349.

[8] Salton G, Fox E A, Voorhees E M. Advanced feedback methods in information retrieval[J]. Journal of the American Society for Information Science, 1985, 36(3):200-210.

[9] Lee K S, Croft W B, Allan J. A Cluster-Based Resampling Method for Pseudo-Relevance Feedback [C]// Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval. USA, New York, 2008:235-242.

[10] 钟敏娟,万常选,焦贤沛. 基于聚类和词组抽取的XML查询扩展[J]. 情报学报,2010,29(4):597-604.

[11] Small H. Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents [J]. Journal of the American Society for Information Science, 1973, 24(4):265-269.

[12] 耿海英. 共引分析方法及其应用研究[D]. 北京:中国科学院国家科学图书馆,2007.

[13] Chen C M, Ibekwe-SanJuan F, Hou J H. The structure and dynamics of co-citation clusters: a multiple perspective co-citation analysis [J]. Journal of the American Society for Information Science and

Technology,2010,61(7):1386-1409.

[14] 赵悦阳,崔雷.专题文献的同被引聚类分析在表现学科专业发展历史的可靠性评价.情报学报,2005,24(4):414-421.

[15] 章成志,师庆辉,薛德军.基于样本加权的文本聚类算法研究[J].情报学报,2008,27(1):42-48.

[16] 吴夙慧,成颖,郑彦宁,等.基于学术文献同被引分析的K-means算法改进研究[J].情报学报,2012,31(1):82-94.

[17] Nakov P I,Schwartz A S,Hearst M A. Citances:Citation Sentences for Semantic Analysis of Bioscience Text [C]//Proceedings of the SIGIR'04 workshop on search and discovery in bioinformatics,2004:81-88.

[18] Mercer R E,Marco C D. A Design Methodology for a Biomedical Literature Indexing Tool Using the Rhetoric of Science[C]//Proceedings of the bioLink workshop in conjunction with human language technology conference/North American chapter of the association for computational linguistics annual meeting (HLT/NAACL),2004:77-84.

[19] Bradshaw S. Reference Directed Indexing:Redeeming Relevance for Subject Search in Citation Indexes[C]// Proceedings of the 7th European conference on research and advanced technology for digital libraries, 2003: 499-510.

[20] Elkiss A, Shen S, Fader A, et al. Blind Men And Elephants:What Do Citation Summaries Tell Us About A Research Article? [J]. Journal Of The American Society For Information Science And Technology,2009, 59(1):51-62.

[21] Aljaber B,Stokes N,Bailey J,et al. Document clustering of scientific texts using citation contexts[J]. Information Retrieval,2010,13(2):101-131.

[22] Ritchie A, Robertson S, Teufel S. Comparing Citation Contexts for Information Retrieval[C]//Proceedings of the 17th ACM conference on information and knowledge management,CIKM 2008. Napa Valley,CA,USA:ACM, 2008 :213-222.

[23] Ritchie A,Teufel S,Robertson S. Using Terms From Citations for Information Retrieval: Some First Results [C]//Proceedings of the 30th European conference on information retrieval (ECIR). 2008:211-221.

[24] Bergmark D. Automatic Extraction Of Reference Linking Information From Online Documents [C]//Technical Report CSTR 2000-1821, Cornell Digital Library Research Group,2000.

[25] Bergmark D,Phempoonpanich P,Zhao S. Scraping the ACM digital library[J]. SIGIR Forum,2000,35(2):1-7.

[26] Powley B,Dale R. Evidence-Based Information Extraction For High-Accuracy Citation Extraction And Author Name Recognition[C]//Proceedings of the 8th RIAO international conference on large-scale semantic access to content,2007.

[27] Councill I G,Giles C L,Kan M Y. Parscit:An open-source crf reference string parsing package [C]// Proceedings of language resources and evaluation conference (LREC 08),2008.

[28] Zhao S J. Named Entity Recognition in Biomedical Texts using an HMM Model[C]// 04 Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications,2004.

[29] Borthwick A. A Maximum Entropy Approach To Named Entity Recognition [D]. New York: New York University,1999.

[30] Kazama J, Makino T, Ohta Y, et al. Tuning Support Vector Machines for Biomedical Named Entity Recognition//BioMed '02 Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain-Volume 3,2002.

[31] a11-smart-stop-list[EB/OL]. http://jmlr. csail. mit. edu/papers/volume5/lewis04a/a11-smart-stop-list/. 2005-12-01/2005-1-8.

[32] 严华云,刘其平,肖良军.信息检索中的相关反馈技术综述.计算机应用研究,2009,26(1):11-14.

[33] Harman D. Evaluation Techniques and Measures[C]// Proceedings 4th Text Retrieval Conference. 1996:6-14.

[34] 成颖.面向学术新人的相关性判据研究——基于本科课程论文的内容分析[J].情报学报,2011,30(9): 522-535.

[35] Larson R R. Bibliometrics of the World Wide Web:An Explo-ratory Analysis of the Intellectual Structure of Cyberspaces[C]//Proceedings of the Annual Meeting of the American Society of Information Science,1996.

[36] Thelwall M,Wilkinson D. Finding similar academic Web sites with links,bibliometric couplings and colinks [J]. Information Processing and Management,2004,40(3): 515-526.

[37] Vaughan L,You J. Mapping business competitive positions using Web co-link analysis[C]//The Proceedings of ISSI,2005.

[38] Wang Y,Kitsuregawa M. Enhancing contents-link coupled Web page clustering and its evaluation [C]// Proceedings of data engineering workshop (DEWS2004), 2004.

(责任编辑 化柏林)