图书情报工作
第 59 卷 第 1 期 2015 年 1 月

联邦检索研究综述

杨海锋[1,2] 陆伟[1]
[1] 武汉大学信息管理学院 武汉 430072 [2] 江西理工大学应用科学学院 赣州 341000

摘要: [目的/意义] 对联邦检索研究进行梳理,总结发展现状,明确发展方向。 [方法/过程] 在大量文献调研的基础上对联邦检索研究进行总结和评述。 [结果/结论] 联邦检索包括数据集描述、数据集选择和结果合并 3 个阶段,各阶段从不同研究角度提出了较多算法,但缺乏权威的测试数据集和统一的评价标准。联邦检索理论和技术被广泛应用,但大数据环境下也为其提出了许多新的研究课题。

关键词: 联邦检索 数据集合描述 数据集合选择 结果合并
分类号: G253
DOI: 10.13266/j.issn.0252-3116.2015.01.018

1 引言

联邦检索是信息检索研究的重要组成部分,也是检索国际会议 TREC(文本检索会议)的任务之一。在国内,联邦检索也称作分布式检索、跨库检索、集成检索等。在全文检索数据库中国知网中,笔者通过关键词“联邦检索”进行检索,得到 4 篇核心期刊的论文。4 篇论文都发表在图书情报类杂志上,主要以对联邦检索系统的分析和评述为主。同样,笔者分别使用关键词“分布式检索”、“跨库检索”和“集成检索”进行检索,并人工统计与图书情报主题相关的论文数量,结果分别是 7 篇、148 篇和 7 篇。从检索过程可以看到,联邦检索(分布式检索、跨库检索、集成检索)的研究基本上是从计算机学科和图书情报学科两方面进行的,计算机学科更侧重于技术方面的研究,而图书情报学科除了技术研究外,更多地涉及到检索系统的应用、评价和比较。比如学者陈朋对 ISI Web of Knowledge 集成检索平台进行了评述,学者张云秋对国内外 6 个跨库检索系统的功能进行了比较研究,学者陈家翠探讨了联邦检索的机制及其存在的问题。上述研究对联邦检索的探讨比较分散,未能从整体上把握联邦检索的环节和涉及的主要算法,且目前国内研究发文量较少,研究深度较浅。本文将全面介绍联邦检索的运行机理、主要算法,结合国内外学者的研究成果,尽可能反映出联邦检索最新的研究状况和成果。

元搜索(metasearch)、联邦检索(federated search)、聚合检索(aggregated search)和垂直检索(vertical search)是几个日常应用较多的概念,但容易被混淆。元搜索可以被看成是联邦检索在网络文本检索中的一种形式[1],元搜索引擎隐藏了背后为其并行工作的多个独立的搜索引擎(“黑盒搜索引擎”),但元搜索引擎不保存相关文档索引,用户从一个入口进入检索,并得到召回率较高的结果。而联邦检索的不同在于其在检索过程中涉及到信息源的描述、与查询相关信息源的选择,这和元搜索背后的“黑盒搜索引擎”形成了鲜明的对比,表明其过程的处理要更复杂一些[2-3]。和传统集中式的搜索引擎相比较,联邦检索对不容易爬取的隐藏网页内容(比如百度、谷歌等搜索引擎)也能进行有效的处理[4]。聚合检索不是一个纯粹的新研究领域,它是从联邦检索、元搜索、语义检索以及实体检索等发展而来的[5],和传统信息检索结果序列不同,它更加关注返回结果的聚合,其中包括图片、视频和新闻等,这样能更好地满足需求多样且需求表述不明确的用户的要求。垂直检索通常是针对通用搜索引擎信息量大、结果不准确和不深入等提出的搜索引擎的细分,主要面向特定用户、特定领域和特定学科等的“专”、“精”、“深”的搜索引擎。M. Shokouhi 等[4]认为,聚合检索就是垂直检索,包括数据库的垂直选择和结果的合并。对于上述概念也不必过于纠结于其细节,因为网络化环境下,各类型检索之间的相互融合越来越紧密,你中有我,我中有你。

作者简介: 杨海锋(ORCID: 0000-0002-7344-7394),讲师,博士研究生,E-mail: yanghfky2007@163.com; 陆伟(ORCID: 0000-0002-0929-7416),教授,博士,博士生导师。
收稿日期: 2014-11-25 修回日期: 2014-12-20 本文起止页码: 134-143 本文责任编辑: 易飞

图1描述了联邦检索系统的典型结构。用户将自己的查询提交给检索代理(the broker)，检索代理与查询相关的独立的数据集合进行关联，并将从中返回的结果合并处理后发送给用户。图1中阴影颜色较浅的部分就是与查询相关的数据集合，同时检索代理中包含了对各独立数据集合的信息描述。从图1可以看出联邦检索的三大主要问题[4]：①怎样描述数据集合；②怎样选择与查询相关的数据集合；③如何对返回的结果进行合并处理。需要说明的是，由于不同文献对相关术语的描述不同，本文中使用的文档数据库、资源数据库、数据集合、信息源等表示相同的概念。T. T. Avrahami等开发的美国联邦政府统计信息门户网站的联邦搜索引擎是一个典型的联邦检索系统，它建立在美国100多个联邦部门公布的统计数据之上，当时关于联邦搜索引擎的实际应用还寥寥无几，大多还停留在理论研究层面[5]。

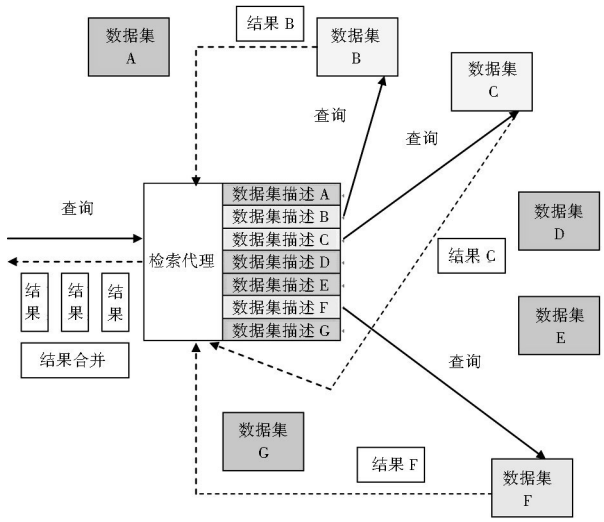

图1 联邦检索系统典型结构[4]

联邦检索技术已被许多检索任务所使用。如博客精选使用户能找到志同道合的博友，该问题可以被看做是一个资源选择问题，对于每个博客站点，即使该站点包含的发帖数量差异较大，现有的资源选择方法都能很好地解决[6-7]。类似地，在专家检索中可将每个专家看成是自己所产生的文献的集合。聚合检索使用多个垂直检索的结果来丰富自己的检索结果，也是联邦检索的一种[8]。同时，多媒体数据的检索也大量使用了联邦检索技术，J. Callan等[9]在欧盟资助的MIND项目中研究了多媒体分布式数字图书馆资源选择与数据的融合；Si Luo等[10]将结果合并算法应用于多语言文档检索合并，取得了一定的效果。

在联邦检索系统的使用中，国外常用的可以在网上获得的搜索引擎如表1所示：

表1 国外常用联邦检索[11-14]

免费联邦检索入口	商业联邦搜索引擎
Science Accelerator	WebFeat
Science. gov	Exlibris metalib
WorldWideScience. org	360 search-(used by IIM Ahmedabad)
Scitopia	EBSCO Host integrated search

研究发现，和国内图书馆热衷于购买现成的商业联邦系统(比如Metallib系统)不同，国外许多数字图书馆都采用了自行开发的方式，根据用户的特点和需求量身打造[15-16]。比如新墨西哥州洛斯阿拉莫斯国家实验室研究图书馆项目FlashPoint、国立墨西哥大学细胞生理学研究所开发的Hermes等[17]。国内成熟的联邦检索系统有清华同方的异构统一检索平台USP、中国高等教育文献信息保障系统统一检索平台CALIS、中国国家数字图书馆(CSDL)跨库检索平台等。在相关系统的使用过程中，笔者发现国内联邦检索系统中查询词的规范化、查询细节的显示、个性化功能的改善等都越来越受到了重视。

2 数据集合的描述

为了能找到和查询相关的数据集合，检索代理必须存储每一个数据集合的相关描述信息，这些统计信息可以是人工产生的一些描述信息，但由于数据集合数量比较多，人工描述效率低且过于简单，为了克服人工完成数据集描述的不充分性，这部分工作主要依靠程序自动生成[4,18-19]。这就是对数据集合的描述。检索代理能否自动获得数据集合的描述信息，与联邦检索所处的环境有关。

2.1 协同环境下文档集合描述

在协同环境下的联邦检索中，检索代理能自动获得文档集合的词典统计信息和有用的元数据，并能有效地计算和查询相匹配的数据集合的得分[20]。由超过11个公司和组织联合开发的STARTS协议[21]提供了检索代理，用于数据集合选择、查询映射和结果合并等多种类型的元数据。该协议也是该领域使用最广泛的标准之一。这些元数据的属性主要包括文档评分范围、停用词序列、支持查询的域和抽样结果。同时每个搜索引擎支持的查询语言也包括在数据集合描述中。由协议STARTS定义的查询语言包括两个主要部分：过滤器表达式(布尔组件)和排序表达式(向量空间组件)。过滤器表达式指出了查询结果中文档必须满足的条件，而排序表达式则硬性约束查询结果中必须出现的单词，并在该单词上施加某种规则来对查询结果

中的文档进行排序。因此,这两个表达式缩小了文档查询的范围和规模[22]。

当然,数据集合描述的信息是否全面还取决于各数据集合之间的协作程度和搜索协议的复杂程度。比如有的数据集合描述信息中存储了每个数据集合的文档词频、词项权重和最大词频等[23-24]。B. Yuwono[25]等在 D-WISE 系统的数据集合描述中包含了每个数据集文档的数量和词项在数据集中的分布频率。D. D'Souza 和 L. Gravano 等[20,26]提出了 n 个词项的索引,也就是对集合中每个文档来说,数据集描述集合中只存放词频最高的前 n 个词项的索引信息。J. Arguello 等[27-28]尝试在垂直搜索中将查询日志存储到描述集合中。J. Kim 等[29]在邮件数据集合中也进行了相关的研究。

2.2 QBS

在非协同联邦检索环境中,数据集合不能给检索代理自动提供词项统计等信息,检索代理需要从数据集合中进行文档抽样来对数据集合进行描述(比如词项统计数据、文档频率等)。J. Callan 等[21]认为对于不同部门掌控的多领域网络资源,现有资源描述方法的效果非常有限,为提高集合的描述准确度,他们提出了基于查询的抽样方法(Query-based Sampling)。其算法描述如下[4,21]:

(1) 选择初始查询词并提交到数据集合,通常查询词是一个词项,这个词项应该是与很多文档相关的查询;

(2) 执行查询,和查询匹配的前 n 个结果返回;

(3) 基于检索结果来更新数据集合的描述:

(a) 抽取前 n 篇返回文档中的词项和出现的频率;

(b) 将抽取的词项及其频率添加到资源描述集合中;

(4) 如果抽样停止的条件还没有达到,则从最新的描述中选择查询词,并执行步骤(2)。

上述算法描述中,依据经验 n 通常取值为 4 比较合适,抽样停止的标准是依据抽样文档的数量或者抽样查询的数量而定义的。J. Callan 等[21]提出如果下载达到 300 - 500 篇不同的文档,即可停止;M. Shokouhi 等[30]对其 300 个抽样文档能否充分覆盖数据集合的词项提出了质疑,提出了一种新的抽样策略并加以验证,并指出抽样文档的多少要根据文档集合的不同而变化,实验结果表明其策略在动态变化的样本中达到很好的词项覆盖。

传统的基于查询的抽样(QBS)方法存在一些缺陷。随机选择查询可能无法返回充足的结果,导致抽样过程的不充分。同时,抽样文档数量的多少可能会影响到文档集描述的效果。因此,M. Shokouhi 等[21]提出了自适应抽样的方法(adaptive sampling)。和传统方法比较,自适应方法能根据抽样中新词出现的比例自适应地来决定什么时候停止探测,在 2 个测试平台上进行的实验发现,新方法不仅比样本数量固定的方法效率高,而且在一些情况下取得了比集中式检索更好的效果。J. Caverlee 等[31]提出了自适应抽样结束的 3 个标准,分别为同比例的文档率(proportional document Ratio)、同比例的词汇率(proportional vocabulary Ratio)和词汇增长(vocabulary growth)。同时,还将自适应抽样方法的执行过程分解为 3 个步骤:

(1) 种子样本:为了引导抽样查询迭代的进行,首先从每个数据库中选择一个最初的样本;

(2) 动态抽样分配:使用数据库的种子样本来估计参加检索的所有数据库的规模大小和质量参数;

(3) 动态抽样执行:对上一步基于 QBS 动态分配的文档,这一步中 n 个数据库按照此文档进行抽样。

为了避免抽样的不充分性,学者们也提出了其他改进的方法。为了克服低频词不能在样本中很好体现的问题,P. G. Ipeirotis 等提出了一种收缩技术(shrinkage technique)[32],该方法假设相关主题的数据集合是共同分享其词项的,据此将数据集合分成不同的主题,并用近似的内容来描述同一类中的数据集合。通过在 315 个真实网络数据库上的评估表明,与没采用该技术的方法相比,采用收缩技术的数据集合描述更充分。P. G. Ipeirotis 等[33]还提出了一种聚焦查询探测方法(focused query probe),该方法认为和主题类别相关的查询能检索到该类别中的文档。该方法采用了一个被训练的文档分类器,同时参与训练的文档遵守一定的规则,抽样的探测查询从分类规则中被抽取。比如分类器定义了 Cancer - > Health,也就是说包含"Cancer"的文档是和 Health 类相关的,如果以"Cancer and Breast"作为查询词,那么返回的文档就包含在 Health 类中。由于类别可以进一步分成不同的子类,所以随着抽样的继续,探测查询将按照更多的分类规则被选择,即被具有分类层次结构的分类器所约束。

3 数据集合选择

由于带宽、时效性等资源的限制,不能将查询广播式地传递给每一个数据集进行查询操作,这样会耗费较高的成本,同时也不现实[4,20]。因此,当一个查询被输入,检索代理就需要对数据集进行排序并且决定选择哪些数据集来进行检索,这就是数据集合选择。关

于数据集合选择的方法很多，基本思路是根据选择索引计算出数据集合的 goodness(区分数据集合的优劣)值进行排序，并选择排在前面的数据集合作为查询候选数据集合。因此，数据集合选择问题实质上是如何更好地计算数据库分值的问题[2]。

文献[4]采用词典的方法(lexicon methods)、文档代理方法(documentsurrogate method)、基于分类聚类的选择方法(classification(or clustering)-based collection selection)等对资源选择算法进行了描述，而文献[34]则采用基于资源相关度排序的资源选择技术、基于文献分布状况的资源选择技术、基于检索成本计算的资源选择技术和基于资源内容等级结构的资源选择技术对目前的资源选择技术进行了分析和评价。文献[35]在对信息集选择方法的前期研究成果及分析进行评价的基础上，对当时的语言模型框架方法、查询驱动选择方法等热点进行了评述。最典型的算法涉及到 GlOSS、CORI、CVV、ReDDE 等。由于上述算法公式较多，相关文献都对其进行了较翔实的介绍，笔者在此仅做简单评述。

3.1 GlOSS

早前数据集选择策略是将数据集看做是一个词袋(a big of words)，然后按照和查询词的相似性进行排序[20]。L. Gravano 等[36]提出在协同环境下 GlOSS(glossary of servers server)数据集选择方法。由于该方法当时仅仅在图书馆和信息供应商处使用，且只支持布尔检索，所以也称为 bGlOSS。其基本思想是词频统计信息保存在检索代理 GlOSS 服务器中，当收到查询请求时，计算每个数据库中包含所有查询词文档的数量，GlOSS 按照满足查询请求的文档数的多少来对数据库进行排序，同时也能看到 GlOSS 只需要维护词频信息，和全文本的索引相比较，其索引规模相当小。但对该方法的评估有限，且查询形式不灵活[34]。由于 bGlOSS 的计算过程只涉及到包含查询词项的文档频率和数据库中词项权重的和，而查询词项在每个文档中的频率并没有被体现，因此基于向量空间的检索模型 vGlOSS 被提出。在 vGlOSS 中，查询项和文档的相似度被定义为查询向量和文档向量的内积，数据库与查询项的切合程度，即数据库中所有文档与查询项相似度之和通过 goodness 值来计算。

3.2 CORI

1995 年，J. P. Callan 等[37]提出了一个基于推理网络的概率方法 CORI(a collection retrieval inference)。该方法通过查询项在文档集合中出现的信任概率(belief probabilities)的和来对文档集合进行排序[20]，也可以认为 CORI 方法是一个“大”文档的检索推理网络模型，所谓的“大”文档就是文档数据库。该方法的有效性已在 INQUERY 中得到验证。但对于异构化数据库的检索、文档数据库的描述以及如何将用户需求转化为结构化查询，该方法涉及得比较少。同时研究显示，如果文档集规模呈现出偏态分布，那么 CORI 算法将不能取得好的效果。对此，Si Luo 等[38]提出了在线数据库中估计相关文档分布的方法，这一新方法比 CORI 选择算法具有更高的精度。H. Nottelmann 等[39]首次在大规模测试集上评估了 DTF(the decision-theoretic framework)和 CORI 方法，结果显示在许多情况下 DTF 的效率高于 CORI。DTF 方法不同于以往的资源评价算法，它不仅考虑额外的成本花费，比如时间、费用等，而且还计算了查询数据集的数量和每个数据集返回文档的数量。同时，为了估计检索质量，比如所期望的相关文档的数量，DTF-sample 和 DTF-normal 两种新方法被提出。

3.3 CVV

B. Yuwono 等[25]提出了查询匹配中基于稳定性估计的数据集排序方法，即 CVV(cue-validity variance)。CVV 检索代理仅仅保存文档数据集 DF 的值，其 goodness 值通过文档频数(DF)和线索有效性方差(CVV)两个要素来计算得到。但该方法没有考虑到查询表达式中词项出现的次数，没有给予词项一定的权重。作者在康奈尔大学的 Smart 系统的文本数据集上对其方法进行了测试，实验结果表明，CVV 方法几乎在所有主题文档分布中都表现良好；同时发现 CVV 方法需要检索代理服务器与文档集合服务器之间小量的原数据交换，而且该方法计算简单、存储要求比较低。

3.4 ReDDE

Si Luo 等[40]提出的 ReDDE(the relevant document distribution estimation)数据集选择算法是非协同环境下的一种典型方法，它不再把数据集看成是词袋，或者说看成一个大的单一的文档、词汇的分布。从一定程度说，该方法更能体现文档在数据集中发挥的作用以及各个文档之间的区别。该算法使用相关文档的估计分布来对所有的数据集进行排序，并且选择包含相关文档数最大的那些数据集作为查询数据集。许多相似的算法在 TREC Blog Track[41]中被提出，比如 J. L. Elsas 等[6]和 J. Seo 等[7]受到 ReDDE 算法的启发，对于包含大规模相关主题的博客提出了一种博客选择算法。

ReDDE 算法有很多变形，基本都是通过不同的方法对抽样文档序列中靠前的文档赋予不同的权重以及

估计相关性概率。M. Shokouhi 等[42]提出了一种基于 CSSE(central sample search engine)模型的 CRCS(central-rank-based collection selection)文档集选择方法。CRCS 按照抽样文本在 CSI(抽样文档的索引是通过所有文档集产生的)中的排列顺序赋予不同的权重,抽样文本所属的文档集的得分是按照抽样文本在 CSI 的排序的位置被计算的。通常选择抽样文本序列的前 γ 个结果进行文档集得分的估计,其中关于 γ 的取值研究较少,文献[42]对其取值 50。但该方法在实验中假设每个数据集的检索模型相同、检索效果相当,然而现实情况并非如此。P. Thomas 等[43]提出了 SUSHI 算法,该算法能充分利用抽样文档内容并且不需要训练集,能对抽样文档中未被充分利用的文档进行推理得分,这些得分可以用来优化返回的最终结果。为使文档集选择能在不同的应用中发挥最大化效用(比如高的召回率或者高的准确率),Si Luo 等提出了 UUM(unified utility maximization framework)[44]方法,该方法需要训练信息来估计相关文档的概率。首先使用小规模的训练查询构造一个逻辑转换模型,然后通过抽样文档来估计所有文档的相关性概率,最后依据这些概率,文档按照使不同应用最大化效率来进行排序。这里 UUM 假设所有的文档集(搜索引擎)都使用了有效的检索模型,但实际应用中,如果忽略检索效果将可能降低文档集选择的质量。于是 RUM(returned utility maximization)[45]应运而生,该算法将搜索引擎的检索效果这个因素添加到了联邦检索框架中,在搜索引擎有效性的测试过程,首先发送一个小规模的训练查询到检索文档,然后在搜索引擎返回文档序列后,又在同样的文档集上采用一种有效的集中排序算法排序,最后比较两者结果的一致性。

3.5 分类聚类方法

传统的文档集选择方法是根据文档集的相关性来建立模型的,当然文档选择也可以被看成是分类聚类问题,按照查询将文档集分成不同的类。该方法具有灵活性好、训练集容易获取、机器学习方法易用和检索效果好等优点[41]。文献[46]中提到通过查询的训练集,可将文档集依据 3 类特征进行划分:基于语料库的特征、基于查询类别的特征和点击的特征。E. M. Voorhees 等[47]提出了学习查询返回文档数的两个方法:第一个方法通过发送训练查询并分析返回结果来学习文档集的主题相关性;第二个通过返回文档的数目来训练查询聚类。类似地,S. Cetinta 等[48]提出了 qSIM 资源选择方法,该方法使用过去查询的检索结果来估计获得信息资源的效用,实验表明该方法的效果与 ReDDE 相当。在一系列文献中,P. G. Iperiotis 和 L. Gravano 等[32-33,49]提出了一种叫做有意识的分类(classification-aware)的文档集选择方法。根据抽样文档的词项,作者将每一个文档集映射到层级结构分类树的分支上,每个分支代表一个主题类,主题类的描述是通过文档集合描述的词项统计而产生的。在文档集选择阶段,检索代理比较查询和主题类的描述,然后将查询发送到与其最相关的主题类中。

3.6 其他方法

除了上述常用的数据集合选择方法之外,好多学者也从不同的角度对其进行了研究。D. Hong 等[50]为了使得联邦检索结果多样化,提出在资源选择阶段不仅要考虑结果的相关性,还要考虑结果的多样性,其所提到的两种方法能被广泛地应用到目前存在的资源选择算法中。K. Balog 等[51]为了解决协同分布环境下的 ad-hoc 实体检索问题,提出了 AENN(all that an entity needs is a name.)文档集排序和选择方法。G. Paltoglou 等[52]在非协同分布式信息检索环境下提出了一种新的文档集选择算法,该算法将每个文档集看做整体来建立模型,通过计算抽样文档的相关性得分和在文档集中的位置,选择那些覆盖抽样文档序列最大的文档集来进行检索。该算法的新颖之处是每次动态地估计所要选择文档集的数量。在曾经被聚类算法应用的平台上进行测试,发现该算法和 CORI 效果相当,但比 ReDDE 效果明显。同时在别的平台上测试,也发现该方法比以往的方法在召回率和精确度方面都有所改善。M. Shokouhi 等[53]从文档集描述应该根据文档集(搜索引擎)的变化而被更新的角度出发,研究了过去的文档集描述对当前检索精确度的影响,并提出了 3 个管理更新的策略:CU(constant updates)、QL(updating according to query-logs)、SS(updating according to collection sizes)。J. C. French 等[54]比较了 GlOSS 和 CORI 排序算法的特性,通过和基线方法比较,发现 GlOSS 不能很好地估计相关性排序,而 CORI 在实验环境中显示了对相关性估计的一致好评。文献[55]、[56]等研究了如何处理文档集合选择中文档副本或者重复的问题,提出了文档集之间重合文档的估计方法,以及在最后结果中使文档唯一性最大化的方法。A. L. Powell 等[57]在 6 个测试环境中使用了 3 个测试集来对 CORI、CVV 和 GlOSS 等选择算法进行比较,显示出方法的相似性趋势,但 CORI 方法相比别的方法稳定性更好。

4 结果合并

典型的联邦检索的最后一步就是结果合并。检索代理收到每个被选择的数据集合返回的靠前的文档序列后，需要将这些文档排列成一个单一的序列呈献给用户。但是由于不同的文档集合(搜索引擎)使用不同的检索模型和不同的排序特性，许多数据集返回的文档不能直接进行比较，必须通过结果合并算法计算出不同文档可比较的总体得分，然后才能进行结果的合并。

由于不同数据源在结构、运行平台、检索模型等方面不尽相同，结果合并过程非常复杂，同时也直接影响到呈现给用户的最终效果。文献[58]从选中的数据集合所含文档没有或有少量的重叠、选中的数据集合同构、选中的数据集合异构且所含文档有部分重叠等角度对结果合并算法进行了分类描述，文献[3]、[4]、[18]也对其进行了总结。各种算法中，最典型的莫过于CORI、SSL和SAFE[3,4,58]。

4.1 CORI

在CORI合并算法中[37,59]，文档集c返回的文档的总体得分D_c是基于规范化后的文档得分D'和文档集得分C'线性合并计算得到的。为了使得文档得分在0到1之间，合并计算中的规范化参数建议被设置为0.4和1.4[37]。但是需要说明的是，对于不同类型的查询和文档集，该方法还不能适用[4]。Y. Rasolofo等[60]提出了一种简单的结果合并策略LMS(using result length to calculate merging score)，该方法只需使用文档得分和结果列表长度。由于不需要文档集的统计信息，因此该方法中检索代理不需要存储文档集的相关信息。实验结果显示，在数据集动态更新频繁的网络环境中，该方法更具实用性，同时和CORI方法相比，该方法能产生比较好的或者说至少同等的检索效果。Si Luo等[61]提出了一种基于线性回归和转换模型的结果合并自适应方案，该方法提交查询给集中样本数据集(假设已经建立)和各个不同的数据集，提取各集合中返回的重复文档，以样本数据库返回结果得分作为基准，然后对重复文档做归一化标准处理。在多个类型的搜索引擎上测试，显示该方法的结果优于CORI算法。

4.2 SSL

以前的研究通常假设资源的提供者不仅能相互合作来提供规范化的统计信息，而且搜索客户端能下载到所有被检索到的文档并计算其得分。Si Luo等[62]提出了一种解决结果合并的半学习化方法SSL(semisupervised learning)。通常在结果合并前，能够获得文档集中文档序列的序列号及其得分，而在SSL中，在查询被发送到被选中的数据集的同时，查询也被发送到集中式的抽样数据集。集中式的抽样数据集也会返回包含文档序号和得分的文档序列，这个得分同时也被提供给结果合并算法。通过上述两个文档的得分来训练机器学习算法，使该算法能将特定的文档得分映射为可比较的合并得分。SSL算法提出了两个假设：①从每个被选择的数据库中检索到的一些文档也同时能被集中抽样数据库检索到；②对于小规模的训练样本，即上述文档的两种得分，也能将特定文档的得分映射成可合并的得分。由于重复的文档很普遍，因此机器学习方法基于重复文档的集合而展开。当文档集使用相同的检索模型时，SSL能使用重复文档来训练一个单一模型，该模型能实现文档得分的映射。但如果文档集检索采用不同模型时，SSL就不能训练一个单一的模型来对两种文档得分完成映射，因为得分取值标准不同。SSL算法的两种优势弥补了CORI算法的不足，通过抽样数据集的建模解决了非协同环境下结果合并的许多难题[20]。G. Paltoglou等[63]综合了估计相关性得分算法和回归算法，提出了一种改进的结果合并方法，该方法基于自适应的下载有限数量的文档(被选择)，并通过回归方法来估计其余文档的相关性。实验显示该方法的结果优于CORI和SSL。Lu Jie等[64]对SSL算法进行了改进，提出了分层P2P网络环境下的SESS(score estimation with sample statistics)算法，实验结果表明该算法在分层P2P网络环境下非常有效。

4.3 SAFE

M. Shokouhi等[59]提出了SAFE(sample-agglomerate fitting estimate)方法，该方法是使用抽样数据库中的所有文档的得分，然后通过产生的统计拟合来估计全局得分。与SSL不同，SAFE不依赖于重合文档。其合并过程是当给定一个查询时，抽样文档的结果是原始文档集的子排序，因此对子排序的曲线拟合能估计文档集的全局得分。He Chuan等[65]在SAFE基础上提出了一种权重曲线拟合结果合并方法，该方法通过区分回归模型中精确的排序信息和估计的排序信息，能够准确地估计文档的全局得分。D. Hong等[66]在SSL和SAFE基础上提出了MoRM(mixture of retrieval models)算法，该算法通过集中样本数据库的学习，使用曲线拟合获得文档合并得分，其和SSL、SAFE的不同之处在于该算法在集中抽样数据库上使用了多种检索算法。

4.4 其他方法

除此之外，许多学者也从不同的角度对结果合并

进行了探索。Li Pengfei 等[67]在企业的真实环境下比较了 10 种结果合并算法，并对其结果进行分析，发现轮询调度算法[12]（round-robin）以及 Y. Rasolofo 所提出的方法之一同别的算法比较效果很不理想。Y. Rasolofo 等[68]用来自 15 个新闻网站的的测试集进行了一种合并方法的测试，同需要下载文档和重新计算得分的合并方法相比，该方法是基于当前已获信息（标题、摘要、排序和服务器的有用性）提出的，但两者合并效果相当。Wu Shengli 等[69]从文档重合的角度出发，通过实验评估了包括 SDM[11]（the shadow document method）和 MEM（the multi-evidence method）在内的一系列合并方法在具有重合文档的数据集上的合并结果，发现当文档重合率较高时，SDM 和 MEM 是最好的合并方法，而当重合率较低时，轮询调度算法是最好的合并方法。A. Mourao 等[13]在基于所有搜索引擎排序合并的基础上，提出了一种无监督的后期融合结果合并方法。

5 联邦检索数据集与评价方法

虽然目前存在的许多联邦检索的测试集都使用 TREC 数据集产生，但这还不足以反映真实世界中的搜索引擎，联邦检索测试集应具备如下特征[14]：异构内容、文档集规模和相似性偏态分布、不同检索算法和重合文档。比如 2013 年的 TREC 的联邦网络检索任务中，其测试集不同于许多人工构造的测试集合，其中包含 157 个网络搜索引擎，每个搜索引擎有不同的检索模型，且包含图片、PDF 文档和视频等异构类型的内容[8]。

从联邦检索论文中数据集使用的统计来看，所使用数据集主要分为 TREC 数据集和研究者自建的数据集，如 A. L. Powell 等[57]提出的 SYM－236、UDC－236 和 UBC－100 以及 Si Luo 等[40]提出的 Trec123－100col-bysource、Trec4－kmeans、Trec123－2ldb－60col（“representative”）、Trec123－AP-WSJ－60col（“relevant”）、Trec123－FR-DOE－81col（“nonrelevant”）、Trec123－10col 等。考虑到上述数据集没有涉及到重合文档，M. Shokouhi 等[56]在 TREC GOV 基础上设计了 3 个新的数据集，分别为 Qprobed－280、Qprobed－300 和 Sliding－115。目前建立联邦检索数据集的常用方法是将 TREC 语料集划分进不同的文档集中，好处是许多查询和判断标准被提供，数据集类型丰富，规模大小设置合理，但这种构建的数据集主要从自己的研究角度出发，主题覆盖可能比较狭窄，判断标准单一，不能代表现实中的联邦检索系统环境。

对联邦检索的评价除了传统的相关性、召回率、精确度等之外，在联邦检索的每个阶段都有对应的评价方法。在数据集描述阶段，常用到的方法有词汇相关性度量（CtF，measuring the vocabulary correspondence）、斯皮尔曼等级相关系数（SRCC，Spearman rank correlation coefficient）、KL 距离（KL，Kullback-Leibler divergence）等[21]；数据集选择的评价主要是比较被选择文档返回相关文档的数量，主要方法有 MSE（mean square error）、SRCC 以及 CBR（count-based rangking）、RBR（relevance-based rangking）等[20,37,70]；结果合并的评价主要是比较最终合并结果中相关文档的数量，涉及到计算正确匹配文档的数量、精确度（p@n）以及基线方法等。由于各个阶段之间紧密联系，各阶段不同方法的选择会影响到下个过程的执行，乃至会影响到最终的结果合并，所以从整体过程建立模型来研究结果评价更有说服力，但目前对这方面的研究较少。

6 评述及挑战

在数据集描述中，非协同环境下的基于查询的抽样技术处于本领域的核心位置。自适应抽样的方法（adaptive sampling）能适应动态分布环境的变化，在抽样过程中能避免传统方法抽样的不充分性，并且抽样结束条件减少了人为因素的影响，设置更加合理。收缩技术（shrinkage technique）基于主题相似的数据库所包含词汇更具有相关性的假设，该方法在提高了低频词的覆盖度的同时并不需要额外增加抽样文档的数量。聚焦查询探测方法（focused query probe）应用了一个基于规则的文档分类器来进行抽样，随着探测查询次数的增加，文档数据库的分类将更加精确，因此该方法能更加准确地描述文档数据库的主题。

在数据集选择中，传统的方法是将文档数据库看成一个大词袋，忽略了文档的边界问题，导致检索效率较低。ReDDE 数据集选择算法更能突出文档在数据集中发挥的作用以及各个文档之间的区别，在文档数据库规模相差比较大的情况下，该方法和 CORI 方法相比取得了较好的效果。在 CORI、CVV、GlOSS 3 种经典的数据库排序算法中，CORI 可以在诸多不同的分布式环境中实现稳定、高效的应用，相比之下，GlOSS 与 CVV 方法的性能好坏更多地依赖于集成环境的选取。除了上述通过查询与文档的相关性建立模型的方法外，基于分类聚类的思想也被引入到数据库选择方法中，但其研究目前还是一个比较新的领域，研究成果较少。

在结果合并阶段，以前的技术（比如 CORI）是使用规范的文档和文档数据库两者的得分来计算最终的文

档得分,而目前更多的方法(比如SSL、SAFE)是使用线性回归和曲线拟合来计算合并得分。CORI方法简单易用,但其假设条件是所有数据库的检索模型和评分标准接近,而事实上该假设基本不成立。SSL算法的提出弥补了CORI方法的缺陷,并且对结果合并阶段的研究起到了里程碑的作用。但SSL算法的一个重要问题是用于训练模型的重合文档的数量太少。SAFE算法是对SSL算法的改进,该算法利用了在抽样数据库中的所有文档,用统计的方法来估计它们在原数据库中的得分。因此,SAFE算法并不使用重合文档。从目前的研究来看,如何对不同的检索模型和评分方法进行优化整合将是结果合并阶段研究的一个重要方向。

经过多年的探索和研究,联邦检索研究取得了很大进展,应用领域不断扩大,但也存在许多需要完善和研究的问题:①如何将点击率、锚文本和图谱链接等特性应用到联邦检索中,提高检索效率;②联邦检索的查询扩展;③联邦检索结果的多样化;④如何建立联邦检索不同阶段的关系模型(目前对结果的评价是分别放在不同阶段实施的);⑤大规模数据环境下(如多媒体数据)联邦检索研究;⑥联邦检索中的语义信息描述、用户查询意图的识别;⑦联邦检索数据集的构建等。

参考文献:

[1] Gogoi K,Borthakur J, Sarmah M. Federated search: An information retrieval strategy for scholarly literature [C]//Proceedings the 8th Convention PLANNER-2012 Sikkim University. Gangtok: 2012.

[2] Kopliku A, Pinel-Sauvagnat K, Boughanem M. Aggregated search: A new information retrieval paradigm [J]. ACM Computing Surveys,2014,46(3):41.

[3] 陈家翠. 联邦检索机制及其存在的问题[J]. 图书情报工作,2006,50(6):87-89.

[4] Shokouhi M, Si Luo. Federated search [J]. Foundation and Trends in Information Retrieval, 2011,5(1):1-102.

[5] Avrahami T T, Yau L, Si Luo, et al. The fedLemur project: Federated search in the real world [J]. Journal of the American Society for Information Science and Technology,2006,57(3):347-358.

[6] Elsas J L, Arguello J, Callan J, et al. Retrieval and feedback models for blog feed search [C]//Proceedings of SIGIR. New York: ACM,2008:347-354.

[7] Seo J, Croft W B. Blog site search using resource selection [C]//Proceedings of the 17th ACM Conference on Information and Knowledge Management. New York: ACM,2008:1053-1062.

[8] Demeester T, Trieschnigg D, Nguyen D, et al. Overview of the TREC 2013 federated Web search track [EB/OL]. [2014-08-10]. http://snipdex.org/fedweb.

[9] Callan J, Crestani F, Nottelmann H, et al. Resource selection and data fusion in multimedia distributed digital libraries [C]//Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieva. New York: ACM,2003:363-364.

[10] Si Luo, Callan J, Cetintas S, et al. An effective and efficient results merging strategy for multilingual information retrieval in federated search environments [J]. Information Retrieval,2008,11(1):1-24.

[11] Wu Shengli, Crestani F. Shadow document methods of results merging [C]//Proceedings of the 2004 ACM Symposium on Applied Computing. New York: ACM,2004:1067-1072.

[12] Voorhees E M, Gupta N K, Johnson-Laird B. Learning collection fusion strategies [C]//Proceedings of the 18th Annual International ACM SIGIR Conference. New York: ACM,1995:172-179.

[13] Mourao A, FMartins F, Magalhaes J. NovaSearch at TREC 2013 federated Web search track: Experiments with rank fusion. [2014-08-12]. https://sites.google.com/site/trecfedweb/.

[14] Nguyen D, Demeester T, Trieschnigg D, et al. Federated search in the Wild [C]//Proceedings of the 21th ACM Conference on Information and Knowledge Management. New York: ACM,2012:1874-1878.

[15] 田燕. 中外跨库检索平台的功能分析及展望[J]. 农业图书情报学刊,2009,21(7):5-8.

[16] 李倩. 跨库检索工具分析及在图书馆的应用[J]. 现代情报,2011,31(10):91-94.

[17] 李广建,张智雄. 国外跨库检索系统研究项目及其特点[J]. 情报理论与实践,2004,24(7):444-447.

[18] Crestani F, Markov I. Distributed information retrieval and applications [C]//Proceedings the 35th ECIR Conference. Berlin: Springer-Verlag, 2013:865-868.

[19] Shokouhi M, Baillie M, Azzopardi L. Updating collection representations for federated search [C]//Proceedins of the 21th ACM SIGIR Conference. New York: ACM,2007:23-27.

[20] D'Souza D, Thom J A, Zobel J. Collection selection for managed distributed document databases [J]. Information Processing and Management, 2004,40(3): 527-546.

[21] Callan J, Connell M. Query-based sampling of text databases [J]. ACM Transactions on Information Systems,2001,2(19):97-130.

[22] Gravano L, Chen Chuan, Garcia-Molina H, et al. START'S Stanford proposal for Internet mets-searching [C]//Proceedings of the ACM SIGMOD International Conference on Management of Data. New York: ACM,1997:207-218.

[23] Gravano L, Garc´ia-Molina H, Tomasic A. The effectiveness of GlOSS for the text database discovery problem [C]//Proceedings of the ACM SIGMOD International Conference on Management of Data. New York: ACM,1994:126-137.

[24] Meng Weiyi, Wu Zhonghua, Yu C, et al. A highly scalable and effective method for metasearch [J]. ACM Transactions on Information Systems, 2001,19(3):310-335.

[25] Yuwono B, Lee D. Server raning for distributed text retrieval sys-

tems on the Internet [C] //Proceedings of 5th International Conference on Database Systems for Advanced Applications. Berlin: Springer-Verlag, 1997: 41 - 50.

[26] Gravano L. Querying multiphe document collections across the Internet [D]. Palo Alto: Stanford University, 1997.

[27] Arguello J, Diaze F, Callan J. Sources of evidence for vertical selection [C] //Proceedings of the 32nd International ACM SIGIR Conference. New York: 2009: 315 - 322.

[28] Arguello J, Callan J, Diaz F. Classification-based resource selection [C] //Proceedings of the 18th ACM CIKM. New York: ACM, 2009: 1277 - 1286.

[29] Kim J, Croft B. Ranking using multiple document types in desktop search [C] //Proceedings of the 33rd ACM SIGIR. New York: ACM, 2010: 50 - 57.

[30] Shokouhi M, Scholer F, Zobel J. Sample sizes for query probing in uncooperative distributed information retrieval [J]. Lecture Notes in Computer Science, 2006, 3841: 73 - 75.

[31] Caverlee J, Liu Ling, Bae J. Distributed query sampling: A quality-conscious approach [C] //Proceedings of the 29th Annual International ACM SIGIR. New York: ACM, 2006: 6 - 11.

[32] Ipeirotis P G, Gravano L. When one sample is not enough: Improving text database selection using shrinkage [C] //Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2004: 767 - 778.

[33] Ipeirotis P G, Gravano L. Distributed search over the hidden Web: Hierarchical database sampling and selection [C] //Proceedings of the 28th VLDB Conference. San Fransisco: Morgan Kaufmann Press, 2002: 322 - 333.

[34] 汪语宇,张丽. 集成检索系统中资源选择技术及算法 [J]. 图书情报工作, 2005, 49(10): 29 - 32.

[35] 雷雪. 分布式检索中信息集选择方法研究综述 [J]. 情报科学, 2008, 26(2): 316 - 320.

[36] Gravano L, Garcia - Molinat H, Tomasic A. The effectiveness of GlOSS for the text database discovery problem [C] //Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data. New York: ACM, 1994: 126 - 137.

[37] Callan J P, Lu Z, Croft W. Searching distributed collections with inference networks [C] //Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 1995: 21 - 28.

[38] Si Luo, Lu Jie, Callan J. Distributed information retrieval with skewed database size distributions [C] //Proceedings of the 2003 Annual National Conference on Digital Government Research. Sacramento: Digital Government Society of North America, 2003: 1 - 6.

[39] Nottelmann H, Fuhr N. Evaluating different methods of estimating retrieval quality for resource selection [C] //Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2002: 290 - 297.

[40] Si Luo, Callan J. Relevant document distribution estimation method for resource selection [C] //Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval. New York: ACM, 2003: 298 - 305.

[41] TREC blog track [EB/OL]. [2014 - 08 - 10]. http://trec.nist.gov/data/blog.html.

[42] Shokouhi M. Central-rank-based collection selection [C] //Proceeding of the 29th European Conference on Information Retrieval Research. Berlin: Springer-Verlag, 2007: 160 - 172.

[43] Thomas P, Shokouhi M. SUSHI: Scoring scaled samples for server selection [C] //Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2009: 419 - 426.

[44] Si Luo, Callan J. Unified utility maximization framework for resource selection [C] //Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management. New York: ACM, 2004: 32 - 41.

[45] Si Luo, Callan J. Modeling search engine effectiveness for federated search [C] //Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2005: 15 - 19.

[46] Arguello J, Callan J, Diaz F. Classification - based resource selection [C] //Proceedings of the 18th ACM International Conference on Information and Knowledge Management. New York: ACM, 2009: 1277 - 1286.

[47] Voorhees E M, Gupta N K, Johnson-Laird B. Learning collection fusion strategies [C] //Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 1995: 172 - 179.

[48] Cetintas S, Si Luo, Hao Yuan. Learning from past queries for resource selection [C] //Proceedings of the 18th ACM Conference on Information and Knowledge Management. New York: ACM, 2009: 1867 - 1870.

[49] Ipeirotis P G, Gravano L. Classification-aware hidden-Web text database selection [J]. ACM Transactions on Information Systems, 2008, 26(2): 6.

[50] Hong D, Si Luo. Search result diversification in resource selection for federated search [C] //Proceedings of the 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2013: 613 - 622.

[51] Balog K, Neumayer R, Nørv°ag K. Collection ranking and selection for federated entity search [C] //Proceedings of String Processing and Information Retrieval. Berlin: Springer-Verlag, 2012: 73 - 85.

[52] Paltoglou G, Salampasis M, Satratzemi M. Collection - integral source selection for uncooperative distributed information retrieval environments [C] //Proceedings of the 2008 ACM Workshop on Large-Scale Distributed Systems for Information Retrieval. New York: ACM, 2008: 67 - 74.

[53] Shokouhi M, Baillie M, Azzopardi L. Updating collection repre-

sentations for federated search [C]//Proceedings of the 30th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM，2007：511－518.

[54] French J C，Powell A L，Callan J，et al. Comparing the performance of database selection algorithms [C]//Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM，1999：238－245.

[55] Bender M，Michel S，Triantafillou P. Improving collection selection with overlap awareness in P2P search engines [C]//Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM，2005：15－19.

[56] Shokouhi M，Zobel J. Federated text retrieval from uncooperative overlapped collections [C]//Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM，2007：23－27.

[57] Powell A L，French J C. Comparing the performance of collection selection algorithms [J]. ACM Transactions on Information Systems，2003，21(4)：412－456.

[58] 雷雪，卢涛. 分布式检索中查询结果合并策略研究[J]. 情报理论与实践，2007，30(4)：558－561.

[59] Shokouhi M，Zobel J. Robust result merging using sample-based score estimates [J]. ACM Transactions on Information Systems，2009，27(3)：14.

[60] Rasolofo Y，Abbaci F，Savoy J. Approaches to collection selection and rsults mergingfor distributed information retrieval [C]//Proceedings of the Tenth International Conference on Information and Knowledge Managment. New York: ACM，2001：191－198.

[61] Si Luo，Callan J. Using sampled data and regression to merge search engine results [C]//Proceedings of the 25 th Annual International ACM SIGIR Conference on Research and Development in Information Retrieva. New York: ACM，2002：19－26.

[62] Si Luo，Callan J A. Semisupervised learning method to merge search engine results [J]. ACM Transactions on Information Systems，2003，21(4)：457－491.

[63] Paltoglou G，Salampasis M，Satratzemi M. A results merging algorithm for distributed information retrieval environments that combines regression methodologies with a selective download phase [J]. Information Processing and Management，2008，44(4)：1580－1599.

[64] Lu Jie，Callan J. Merging retrieval results in hierarchical peer-to-peer neworks [C]//Proceedings of the 27th Annual International ACM SIGIR Conference. New York: ACM，2004：25－29.

[65] He Chuan，Hong D，Si Luo. A weighted curve fitting method for result merging in federated search [C]//Proceedings of the 34th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM，2011：24－28.

[66] Hong D，Si Luo. Mixture model with multiple centralized retrieval algorithms for result merging in federated search [C]//Proceedings of the 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieva. New York: ACM，2012：821－830.

[67] Li Pengfei，Thomasn P，Hawking D. Merging algorithms for enterprise search [C]//Proceedings of the 18th Australasian Document Computing Symposium. New York: ACM，2013：42－49.

[68] Rasolofo Y，Hawking D，Savoy J. Result merging strategies for a current news metasearcher [J]. Information Processing and Management 2003，39(4)：581－609.

[69] Wu Shengli，McClean S. Result merging methods in distributed information retrieval with overlapping databases [J]. Information Retrieval，2007，10(3)：297－319.

[70] French J C，Powell A L. Metrics for evaluating database selection techniques [J]. World Wide Web，2000，3(3)：153－163.

作者贡献说明：

杨海锋：构思，资料收集，论文撰写；

陆伟：提出修改意见，定稿。

A Review on Federated Search

Yang Haifeng[1,2]　Lu Wei[1]

[1] School of Information Management，Wuhan University，Wuhan 430072

[2] College of Applied Science，Jiangxi University of Science and Technology，Ganzhou 341000

Abstract: **[Purpose/significance]** The paper summarized research status，and put forward future research direction about federated search. **[Method/process]** Based on a large number of literature research，this paper has summarized and reviewed to federated search. **[Result/conclusion]** Research questions of federated search are mainly related to collection representation，collection selection and result merging. Some algorithms have been proposed in every aspect from different angles. But authoritative data sets and uniform evaluation criteria still relatively scarce. Although theory and technology of federated search were widely used，many new research topics have been raised in big data environment.

Keywords: federated search　collection representation　collection selection　result merging