基于主题划分的分布式检索混合结果合并技术研究*

何莉　林鸿飞

大连理工大学计算机科学与工程系 大连 116024

E-mail: lihedlut@yahoo.cn

摘　要：结果合并是分布式信息检索的最后一个环节，将直接影响检索结果排名，因而准确、有效的结果合并策略对提高检索性能非常重要。本文提出了一种基于主题划分的分布式检索混合结果合并方法。该方法通过文本聚类把文档划分为主题集合，对于给定的用户查询，基于主题进行检索，并利用检索结果中文档的集合得分、排名及RSV值等信息做逻辑回归拟合，完成结果合并过程。实验表明，与传统的结果合并算法相比，本文方法的检索效果得到显著提高，且克服了传统方法过度依赖中间结果得分的缺点。

关键词：分布式检索，混合结果合并，逻辑回归

A Hybrid Results Merging Algorithm for a Topical-Based Distributed Information Retrieval

Li He, Hongfei Lin

Department of Computer Science and Engineering, Dalian University of Technology, Dalian 116024

E-mail: lihedlut@yahoo.cn

Abstract: Results merging, which is the last step of distributed information retrieval, has a direct impact on the final ranking of search results, thus how to merge search results precisely so as to improve retrieval performance has become an important issue. In this paper, we introduce a hybrid result merging strategy for topic-based distributed information retrieval. Firstly, the whole collection was clustered into several sub-collections, each of which represent an implicit topic of the entire collection, and then the individual results from each cluster are merged by fitting a logical regression model taking the collection selection score, ranking information and RSV information into consideration. The experimental result shows that our approach consistently improves retrieval performance over traditional merging methods, which usually rely on document relevance scores returned from remote collections.

Keywords: Distributed Information Retrieval, Hybrid Result Merging, Logistic Regression

1　前言

当今社会，爆炸性增长的网络信息不但给用户提供了丰富的知识来源，同时也给检索系统带来了巨大的挑战。搜索引擎索引页面通常能够达到几十亿个到上百亿个，如果每次都查询全部页面，而用户只浏览其中的前十到二十个结果，那么就造成了巨大的资源浪费。如果能够只对这几十到上百亿页面中的一部分进行搜索就能得到和检索全部数据相似甚至更好的结果，那无疑对于搜索引擎的建设具有重要意义。因此，大规模数据查询过程中，必须引入分布式信息检索技术以解决上述两个方面的问题。

分布式信息检索的优势在于各检索系统之间分工协作，能够处理海量数据并快速响应用户查询，具有很高的检索效率。目前其主要研究的方向包括：文档集合的表示、选择和结果合并[1]等。其中，结果合并过程是将多个检索服务器返回的查询结果进行合并，以形成单一的结果列表返回给用户，它是整个分布式信息检索的最后一个也是至关重要的一个环节，合并策略的选取直接影响检索结果的质量。因此，本文着重研究分布式信息检索过程中的结果合并策略。

*本文承国家自然科学基金资助项目（60373095，60673039）和国家“八六三”计划项目（2006AA01Z151））的资助。

对于用户提交的查询,检索系统必须依据被查询的信息集合中，文档的重叠程度等因素对结果合并策略加以调整。早期的合并策略是根据检索所得的相关文档原始得分来实现结果合并。但是，这些文档的相关得分通常由于其所处集合之间的差异而不能直接比较。因此，许多研究围绕着归一化不同信息集合中文档原始得分展开。如:CORI[2]、LMS[3]和 CombMNZ[4]。但是，以 CORI 为代表的上述算法在归一化文档得分的过程中，必须知道各文档的最高和最低得分，即需要花费巨大的代价在各集合之间进行通信。

基于逻辑回归模型的结果合并策略被证实比上述归一化文档原始得分的算法更加有效[5]。该模型凭借各个信息集合返回的中间结果列表信息[6]实现结果合并，而不需要得到每篇文档的相关得分。但是，由于没有利用文档的全局性信息，在该模型的训练中容易产生较大的偏差。

基于全局信息的下载策略[7]就合并结果而言，无疑是所有结果合并算法中最好的。它将所有信息集合检索所得文档全部下载到本地，建立全局索引，以此计算文档的集合无关性全局得分，并根据该得分形成最终的结果合并列表。但是，下载算法需要占用大量的传输带宽和空间，容易造成用户等待时间过长，在现实环境中极不适用。

因此，本文提出一种基于主题划分的分布式检索混合结果合并策略。该策略既能利用基于逻辑回归模型，有效归一化文档得分的特性，又能加入文档的全局性信息以避免模型训练的偏差。文中通过为各个信息集合建立混合结果合并模型，消除了集合之间的差异，从而获得归一化文档相关性得分，以此完成结果合并过程。

本文余下章节的安排如下：第二部分详细描述了基于主题划分的分布式检索混合结果合并策略。第三部分介绍了实验过程，并通过对比不同的结果合并算法，验证了混合结果合并策略的优异性。第四部分对工作进行总结并提出了继续研究的方向。

2 基于主题划分的分布式检索混合结果合并策略

基于主题划分的分布式检索混合结果合并策略将所有文档依据其不同的主题划分为多个文档集合。用户检索时，选择与查询最相关的主题文档集合进行检索，并依据检索所得文档的排名、所属集合得分和文档 RSV 值为每个检索的信息集合建立逻辑回归模型，得到与集合无关的文档检索得分，并最终根据此得分合并检索结果文档，完成整个分布式检索的结果合并过程。

2.1 主题的构建与文档集合的选择

上述基于主题的分布式检索过程中，文档集合的划分是首先需要处理的问题。用户检索时，希望相关的查询结果能够尽可能的集中。而简单的按照文档大小、信息来源划分文档集合的方法显然不能满足需求。因此，本文采用 K-Means 聚类算法按照主题对文档集合进行划分，使得同一主题的相关文档聚集在同一文档集合内。这样，和用户相关的查询结果就能聚集在某几个集合中，因而用户无需对所有文档进行查询，而只需要选择这些和用户相关的集合进行检索就可以得到满意的结果，极大地提高了检索效率。

选择与用户查询相关的文档集合这一过程中，本文采用了 CORI[2]算法，得到文档集合的查询相关性得分，以此选择最符合用户需求的集合进行检索，并从中得到用以建立混合结果合并模型的文档 RSV 信息，这一过程将在 2.2 节详细描述。

2.2 文档 RSV 信息

混合结果合并策略在合并结果文档过程中，没有利用和集合相关的文档得分进行结果排序，而是充分考虑文档的全局信息，采用全局性文档 RSV 值建立结果合并模型。这样设计避免了不同集合内部特性对文档得分造成的偏差。文档的 RSV[8-9]（retrieval status value）值是通过在选定的信息集合中，全部或者部分下载有限数目检索所得文档，建立全局样本索引获得的。

其中，检索主题集合所得文档 d 与用户查询 q 的可能相关性，与该文档的 RSV 值存在以下映射关系：

$$f:\Re \to [0,1], f(RSV(d,q)) \approx \mathrm{P}(rel \mid q,d) \tag{1}$$

P(rel|q,d) 根据文[10]全概率定理又可以做如下推导：

$$\begin{aligned}&\mathrm{P}(rel \mid q,d) = \mathrm{P}(rel \mid q \leftarrow d) * \mathrm{P}(q \leftarrow d) + \mathrm{P}(rel \mid \neg(q \leftarrow d)) * \mathrm{P}(\neg(q \leftarrow d))\\ &= \mathrm{P}(rel \mid \neg(q \leftarrow d)) + [\mathrm{P}(rel \mid q \leftarrow d) - \mathrm{P}(rel \mid \neg(q \leftarrow d))] * \mathrm{P}(q \leftarrow d)\\ &= f(\mathrm{P}(q \leftarrow d)) = f(\sum_{t \in q} P(q \leftarrow t) * P(t \leftarrow d))\end{aligned} \tag{2}$$

因此，在 INQUERY 系统中，文档 d 的 RSV 值可以最终表示如下：

$$RSV(d,q) = \sum_{t \in q} (0.4 + 0.6 * \frac{tf(t,d)}{tf(t,d) + 0.5 + 1.5 * \frac{dl(d)}{avgdl}} * \frac{\log(\frac{|DL| + 0.5}{df(t)})}{\log(|DL| + 1.0)}) * \frac{tf(t,q)}{ql(q)} \tag{3}$$

其中，tf(t,d)表示文档 d 中关键词 t 的数目，dl(d)是文档 d 的长度，avgdl 是平均文档长度，|DL|是信息集合大小，df(t)代表含有关键词 t 的文档数目，tf(t,q)代表用户查询 q 中关键词 t 的数目，ql(q)是用户查询 q 的长度。

上述下载文档以得到全局性文档 RSV 值，为各个信息集合建立混合结果合并模型的过程，充分利用了下载策略基于全局信息进行结果合并的优势，同时只下载有限数目的文档避免了下载算法容易造成带宽负荷过重等缺点，提高了结果合并效率。而这一过程中，确定下载文档的数目和选择下载文档是其中最重要，也是必须被考虑的内容。

2.3 文档下载过程

如 2.2 节所述，文档下载的目的是获得全局性的 RSV 值以建立结果合并模型。因此，下载过少数目的文档显然无法提供全面的信息，而过多的文档虽然能够加大全局性信息，但同时也制约了训练速度，加重带宽负荷。众所周知，和用户查询相关的文档通常列在检索结果列表的前面。因此，本文在选择文档下载时，采用从各个信息集合检索结果列表的第一位开始下载前 N 个结果文档。其中，这前 N 个文档既可以从前往后逐个下载，也可以按一定的比率下载（如：可以按奇数顺序下载前 1,3,5…N 个文档）。这里本文选择从第一个文档开始逐个下载结果文档以建立全局样本索引计算 RSV 值，得到文档的全局信息。

2.4 混合结果合并模型的构建

利用上述文档 RSV 值和文档排名、所属集合得分信息，就可以为每个主题信息集合建立结果合并模型，从而得到文档的集合无关性得分。这一过程中，很重要的一点是如何才能提供一个准确的从文档所属集合得分、RSV 值和排名信息到文档无关性得分的映射关系。文[10]指出在文档排名信息和文档无关性得分之间存在着非线性的映射关系，以非线性的逻辑回归模型能很好的反应二者之间映射。因此，本文同样为基于主题的各个信息集合建立逻辑回归结果合并模型，其中文档 D_i 的集合无关性得分与文档排名信息、集合得分信息和文档 RSV 值的映射关系如下：

$$P(D_i \ is \ \mathrm{Rel} \mid rank, Cscore, RSV) = \frac{e^{a+b*\ln(rank)+c*Cscore+d*RSV(d,q)}}{1+e^{a+b*\ln(rank)+c*Cscore+d*RSV(d,q)}} \tag{4}$$

rank 代表从主题信息集合中，检索所得文档 D_i 的排名，其中采用 rank 的对数值进行计算可以加大文档之间排名的差异性从而更好的训练模型，Cscore 代表该信息集合的得分，RSV 代表文档 D_i 的 RSV 值。通过调整参数 a、b、c 和 d，可以对上述回归模型进行曲线拟合，得到不同的回归曲线，从而针对不同的用户查询，为每个信息集合建立不同的结果合并模型，获得相应的

集合无关性文档得分，同时始终保证这个得分在 0 和 1 之间。

2.5 拟合回归曲线

为了得到最好的文档集合结果合并模型，本文对公式（4）代表的曲线进行拟合。通过对公式（4）稍加变形，可以得到线性表达式（6）如下所示：

$$y = \frac{e^{a+b*\ln(rank)+c*Cscore+d*RSV}}{1+e^{a+b*\ln(rank)+c*Cscore+d*RSV}} \tag{5}$$

$$\log it(y) = a + b*\ln(rank) + c*Cscore + d*RSV \tag{6}$$

公式（6）中自变量 y 和变量 ln(rank)、Cscore 和 RSV 只存在线性的关系，因此可以对其建立线性回归模型，将公式（6）转化为矩阵的表示形式。

$$Y = X*B + e \tag{7}$$

$$Y=\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}, X=\begin{bmatrix} 1 & \ln(rank_1) & Cscore_1 & RSV_1 \\ 1 & \ln(rank_2) & Cscore_2 & RSV_2 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 1 & \ln(rank_n) & Cscore_n & RSV_n \end{bmatrix}, B=\begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix}, e=\begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ e_n \end{bmatrix} \tag{8}$$

对于上述线性回归模型进行曲线拟合的目的在于评估参数 a、b、c 和 d，最大程度的降低观察值 y 和模型所得值之间的差异 e。因此，文中采用最小二乘法来估计并得到回归模型参数。

$$B = (X'X)^{-1}X'Y \tag{9}$$

其中，用以估计回归模型的样本值对($\ln(rank_i)$, $Cscore_i$, RSV_i,y_i)中，$rank_i$ 是信息集合 C 检索所得第 i 个文档的结果排名，$Cscore_i$ 代表集合 C 的得分，RSV_i 是第 i 个文档的 RSV 值，y_i 代表文档 i 的用户查询相关性判断，为全局性信息，0 代表该文档与用户查询完全不相关，1 代表完全相关。当决定系数 R^2 接近于 1 时，回归模型就趋近于样本数据，即回归曲线得到了最优拟合。

$$R^2 = 1 - \frac{SS_E}{SS_T} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{10}$$

其中，SS_E 和 SS_T 分别代表残缺平方和与回归平方和。上述参数估计的过程中，需要逐个下载的文档，不断计算决定系数 R^2 和更新回归模型直到 R^2 接近于 1。因此，本文在训练过程中设置一个阈值，当决定系数超过该阈值时即停止下载文档和更新模型。这样通过下载少量数目的文档即可以得到比较好的混合结果合并模型，提高了检索效率，文中决定系数的阈值设为 0.9。

3 实验

3.1 实验数据

本文选择 TREC2004 Ad Hoc 检索任务的语料完成实验，将全部 5.16G，3,481,008 个文档基于内容，采用 K-Means 聚类算法划分为 10 个主题集合，（最大集合为 923M，最小集合为 76.5M），并使用编号 1-50 的 TOPICS 中 Title 信息作为用户查询语句，其平均查询长度为 8 个单词。

3.2 实验结果

实验中，选择和查询相关的集合这一阶段，本文采用了经典的 CORI 算法，以此选择与用户查询最相关的主题集合，并得到这些集合的用户查询相关性得分。这一过程中，用户可以在这多个和查询相关的主题集合上，建立不同的检索模型得到检索结果文档，也可以所有信息集合都

采用相同的检索模型进行检索。如果以相同的检索模型进行检索，根据 Raw-Merging 算法的思想，各个集合的结果之间本身就存在一定的可比性。这里，本文采用后一种方法，以 Indri Toolkit[11] 为各个信息集合建立检索模型，得到各集合检索结果文档。

混合结果合并模型的训练过程中，根据下载各个信息集合中检索返回的文档，可以建立全局样本索引，该样本索引可以近似的体现全局索引的信息。这样，文档的 RSV 值、排名和所在集合得分就可以基于该样本索引，通过 INQUERY 检索算法得到，并以此训练各个信息集合的混合结果合并模型，最终获得各文档的全局相关性得分。

为了验证混合结果合并模型中各个参数的有效性和影响最终结果排名的重要因素，本文将文档的 RSV 信息从上述混合结果合并模型中移除，只余文档排名和集合得分信息参与模型训练。此时，剩余的都是集合相关的参数信息，因此本文称该模型为 Collection-Dependent Result Merging，简称 CD Merging。进一步，我们移除集合得分信息，只剩文档排名信息作为模型训练参数，称其为 Rank Merging。最后，将集中式数据检索结果（Centralized retrieval）、Raw-Merging、CORI、Rank Merging、CD Merging 和混合结果合并策略（Hybrid Result Merging）在相关主题集合上做实验对比，检索结果如表 1 所示。

表 1 各合并算法检索结果比较

结果合并算法	P@5	P@10	P@15	P@20
Raw-Merging	0.4000	0.4180	0.4200	0.3850
CORI	0.4040	0.4140	0.4187	0.3870
Rank Merging	0.3840	0.4020	0.4040	0.4050
CD Merging	0.4200	0.3960	0.3973	0.3840
Hybrid Result Merging	0.5880	0.5660	0.4733	0.4030
Centralized retrieval	0.5240	0.5100	0.4920	0.4720

3.3 结果分析

表 1 表明，CORI 算法表现的比较稳定，而 Rank Merging 的效果不佳，多数指标低于 Raw-Merging 的合并结果。这说明只利用和集合本身特性相关的文档排名信息训练结果合并模型是远远不够的。同时，由于 Raw-Merging 的各项结果指标和 CORI、Rank Merging 十分相近似甚至略高，也验证了以同一检索模型在各个集合中检索得到的文档之间存在一定可比性这一假设。

CD Merging 算法在 P@5 指标上超出除混合结果合并和集中式数据检索外其他算法，这一事实说明了信息集合得分对于结果合并有很大影响，特别是对于调整前 5-10 个检索结果排名有重大影响。即加入信息集合得分这一因素后，如果各个信息集合之间的文档排名相同，来自与用户查询相关性高的信息集合中的文档，将比来自相关性低的信息集合中的文档在最终结果列表中排名靠前，这一结果也符合人们的常规认识。因此，本文在训练混合结果合并模型中加入信息集合得分是合理的，能够帮助调整模型，使之更好的拟合文档相关性得分分布。

表 1 同样表明，混合结果合并模型得到的结果已经很贴近于集中式检索结果。这一现象充分说明了两点：第一，和用户查询相关的文档通过 K-Means 主题聚类后的确实分布在少数信息集合中，因此只检索相关主题信息集合的分布式检索结果才能够接近或者超过集中式检索结果。这样，用户确实只需要检索与查询相关的主题集合就能够得到近似于检索所有文档的查询结果，大大节约了检索时间，符合基于主题的分布式检索的设计初衷；第二，通过全部或者部分下载检索所得文档，得到的全局样本索引能够很好的反映全部数据分布下的全局索引。利用全局样本索引计算得到文档 RSV 值，结合信息集合得分和文档排名信息可以有效调整混合结果合并模型，使得回归曲线能够极好的拟合真实数据分布，对于训练混合结果合并模型影响重大。

综上所述，基于主题的分布式检索混合结果合并模型结合了基于全局信息下载算法的优点，同时通过下载有限数目的文档信息训练结果合并模型，达到了比较好的实验结果并且能够快速、高效的响应用户的查询，满足了分布式检索的各项需求。

4 总结和展望

本文主要描述了基于主题划分的分布式检索混合结果合并策略。该策略通过在各个信息集合中，下载有限数目检索所得文档训练混合结果合并模型，将信息集合得分、文档排名和 RSV 值映射成信息集合无关的文档相关性得分。文中详细说明了结果合并过程所涉及的算法和模型，并最终以实验结果阐述了混合结果合并策略的有效性。

下一步，我们将在此基础上考虑集合选择过程。集合选择是分布式检索中相当重要的一个环节，但是本文对此考虑不多，仅采用经典的 CORI 算法而没有做任何优化，为了更加贴近真实的分布式检索环境，下一步研究工作将针对改进集合选择过程展开。

参 考 文 献

[1] 张刚,周昭涛,王斌.基于主题的分布式信息检索技术研究.计算机工程, 2006, 32(12): 80-82.

[2] J Callan, L Z Lu, W Croft. Searching distributed collections with inference networks. In the Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, 1995: 21-28.

[3] Y Rasolofo, F Abbaci, J Savoy. Approaches to collection selection and results merging for distributed information retrieval. In the Proceedings of the Tenth International Conference on Information and Knowledge Management, 2001: 191-198.

[4] S Beitzel, E Jensen, A Chowdhury, D Grossman, O Frieder, N Goharian. Fusion of effective retrieval strategies in the same information retrieval system. Journal of the American Society for Information Science and Technology, 2004, 55(10): 859-868.

[5] G Paltoglou, M Salampasis, M Satratzemi. Results merging algorithm using multiple regression models. In Proceedings of the 29th European Conference on Information Retrieval (ECIR), 2007: 173-184.

[6] L Si, J Callan. Using sampled data and regression to merge search engine results. In the Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2002: 19-26.

[7] G Paltoglou, M Salampasis, M Satratzemi. Hybrid Result Merging. In the Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, 2007: 321-330.

[8] H Nottelmann, N Fuhr. From Uncertain Inference to Probability of Relevance for Advanced IR Applications. ECIR 2003: European conference on IR research No25, 2003, 2633: 235-250.

[9] A Imafouo, X Tannier. Retrieval Status Values in Information Retrieval Evaluation. Lecture Notes in Computer Science, 2005, 3772: 224-227.

[10] A Calve, J Savoy. Database merging strategy based on logistic regression. Information Processing and Management, 2000, 36(3): 341-359.

[11] http://www.lemurproject.org/indri/