基于检索结果聚类的XML伪反馈技术研究

JIYU JIANSUO JIEGUO JULEI DE XML WEI FANKUI JISHU YANJIU

钟敏娟 著

江西高校出版社
JIANGXI UNIVERSITIES AND COLLEGES PRESS

JIYU JIANSUO JIEGUO JULEI DE XML WEI FANKUI JISHU YANJIU

基于检索结果聚类的XML伪反馈技术研究

钟敏娟 著

江西高校出版社
JIANGXI UNIVERSITIES AND COLLEGES PRESS

图书在版编目(CIP)数据

基于检索结果聚类的XML伪反馈技术研究/钟敏娟著. —南昌:江西高校出版社,2013.12

ISBN 978-7-5493-2250-3

Ⅰ.①基... Ⅱ.①钟... Ⅲ.①智能检索系统-可扩充语言-程序设计-研究 Ⅳ.①G354.4-39

中国版本图书馆CIP数据核字(2013)第304879号

出版发行	江西高校出版社
社　　址	江西省南昌市洪都北大道96号
邮政编码	330046
总编室电话	(0791)88504319
销售电话	(0791)88513417
网　　址	www.juacp.com
印　　刷	天津市天办行通数码印刷有限公司
照　　排	江西太元科技有限公司照排部
经　　销	各地新华书店
开　　本	890mm×1240mm　1/32
印　　张	5.75
字　　数	180千字
版　　次	2013年12月第1版第1次印刷
书　　号	ISBN 978-7-5493-2250-3
定　　价	26.00元

赣版权登字-07-2013-634

CONTENTS 目 录

1. 引言

1.1 研究背景与意义

XML(extensible markup language,可扩展标记语言) 是一种可以用来创建自己标记的标记语言,即所谓的元标记语言(meta - markup language) ,它由万维网协会(World Wide Web Consortium, W3C) 创建,用来克服 HTML 的局限。XML 提出以后,迅速风靡了全世界,在各行各业中得到了大量的应用,并成为网络上信息描述和信息交换的事实标准。

XML 文档的大量涌现,产生了对 XML 数据管理的需求,基于 XML 的信息查询和检索成为一个研究重点[1,2]。国外对 XML 数据的信息检索研究开始于 9 年前,随即成为 DB 和 IR 界的研究热点之一。欧洲 DELOS Network of Excellence for Digital Libraries 与 IEEE Computer Society 于 2002 年共同启动了 Initiative of Evaluation for XML Retrieval(INEX) 创新活动[3],旨在采取统一的记分过程为研究者的 XML 检索方法进行评估,同时也为科研机构比较其成果提供一个论坛。每年一次的 INEX 会议吸引了众多科研机构与学者参与 XML IR 问题的讨论。

XML 数据具有结构化特征,但是与传统的关系数据库中数据不同的是,其结构比较松散。同时,XML 数据中往往存在大量的文本(尤其是在以文本为中心的 XML 文档中) 。XML 的双重特性使得 XML 文档的出现对传统信息检索提出了巨大挑战,主要反映在: ①XML 文档检索以元素而非文档为粒度; ②XML 文档检索要求关键词检索和结构检索相结合; ③XML 文档检索要求有统一的排序机制以适应结构化数据和非结构化文本。

近年来,已有许多针对 XML 文档信息检索技术的研究,其中,如何有效地获得高质量的相关信息从而提高检索结果的性能,成为信息检索领域一个亟待解决的热点问题。目前,搜索引擎是最常用的信息检索工具,用户通过提交有限的几个关键词就能得到查找结果。但是普遍存在检索质量不高的缺点,其中一个很重要的原因就是用户的真实信息需求与提交的查询词之间存在一定程度的偏差。很多时候用户提交的查询表达式非常短小,在缺乏上下文语境下很难完全理解用户的查询意图,这种提交查询的不精确性以及模糊性,使得搜索引擎返回的结果里往往包含了大量无关文档。查询扩展是提高信息检索性能的有效技术手段之一。它通过一定的策略向初始查询中增加一些相关词语形成新的查询,以提供更多有利于判断文档相关性的信息,使用新的查询再次检索文档集,从而使更多的相关文档被检索出来。与普通文档相比,XML 具有自描述性、可扩展性、结构和内容两重特性[4]。因此,在检索过程中,用户的查询意图既可以用查询词来表示,更可以在关键词的基础上加入结构约束来辅助检索,从而提高检索性能。然而,提出有效的结构信息对不了解 XML 文档结构的普通用户来说具有非常大的难度。即使是专业用户也同样存在这个问题,因为 XML 文档具有异构性,专业用户往往也只清楚其中某几类 XML 文档的结构,在对异构文档集进行检索时也较难提出准确的查询表达式。因此,在 XML 信息检索中利用反馈机制来帮助用户形成更准确的查询表达式是提高检索质量的一种有效手段。伪反馈一直以来被认为是一种有效的查询扩展技术。多次 TREC 评测会议表明,伪反馈是一种简单但十分有效的查询扩展技术[5]。但是近年来的研究表明伪反馈方法容易产生“查询主题漂移(query drift)”现象。最主要的原因是伪反馈假设初始检索结果的前 N 篇文档是相关的。而事实上,这个前提假设并不一定总成立,检索结果的前 N 篇文档有可能与查询主题并不相关,从不相关的文档中提取扩展信息显然不合适。

本书正是在此背景下,针对传统伪反馈的“查询主题漂移”问题展开研究,充分利用 XML 的新特性,同时借鉴传统信息检索的方法与技术,希望能够有效地提高 XML 检索质量,对 XML 信息检索领域

的发展有所推动和提高。

1.2 国内外研究现状概述

信息检索里一个最基本的问题就是搜索满足用户信息需求的文档。然而，用户短小的查询往往不能精确地描述其需求，因此，许多学者提出利用查询扩展技术来提高检索性能。在所有的查询扩展方法里，伪反馈不需要用户的参与而受到普遍关注。从目前的研究成果来看，基于伪反馈机制的 XML 信息检索技术的研究还很少，大部分的工作都是针对传统文本的伪反馈。

传统伪反馈方法，其基本思想是假定返回结果的前 N 篇文档是相关的，然后基于这些相关文档进行查询词的扩展。实验结果证实该方法是一种有效的查询扩展技术。然而近年来的研究表明基于传统伪反馈的查询扩展方法容易产生“查询主题漂移(query drift) ”。事实上，返回结果的前 N 篇文档并非都是与查询相关的。文献[6] 指出该前提假设在 N 值较小的时候是合理的，较大的 N 值容易引入大量无关文档，反而会造成查询性能的整体下降，因此，参数 N 的取值在伪反馈的查询扩展中非常敏感。为了较好地解决“查询主题漂移”现象，使得伪反馈方法更加灵活与健壮，目前研究主要围绕以下两方面展开: 一方面如何确定相关文档，形成较高质量的伪相关文档集; 另一方面如何从伪相关文档集合里挑选有效信息进行查询扩展。针对这两大方面的工作，下面将分别展开论述。

(1) 相关文档的确定

造成查询主题漂移的一个很大原因就是查询扩展源的质量不高，这是因为传统伪反馈需要事先固定 N 值，并且假定这 N 篇文档就是相关的。而研究表明[7] 不同的 N 值对最终检索性能有较大的影响。从实际情况考虑，前 N 篇文档并不总是与查询相关，不同的查询主题所获得的初始检索结果质量是不同的。有些查询可能初始检索结果的前 N 篇包含较多的相关文档，质量较高，有些查询可能恰恰相反，初始的结果里存在大量无关文档，因此前 N 篇文档的相关性不能一概而论。这种现实状况不可避免地就会引入大量噪音，对查询扩

展带来干扰。假如我们能够自动地区分哪些是与用户查询相关的文档,哪些是与用户查询无关的文档,就能在源头把噪音信息切除,为有效查询扩展带来前提保障。为此,研究主要涉及两个方向:一是针对 N 值敏感问题展开研究,力图使得基于伪反馈的查询扩展方法更具灵活性和健壮性;另一方向是利用相关的技术和手段,比如对初始检索结果进行调整或者是聚类,从而改进初始检索结果的质量,使得用于反馈的文档都是相关文档。

目前,有很多的研究者针对伪反馈中 N 值敏感问题进行了研究,其解决思路主要围绕以下两种:一种是针对不同的查询条件直接捕获最优反馈文档数目[7~9];另一种思路与之相反,它并不寻求最优 N 值,而是弱化伪相关文档数目的取值对性能的负面影响,来提高伪反馈过程的健壮性,这种思路大部分都是建立在统计语言模型框架里[6,7,10~12]。

文献[6]最早对 N 值敏感问题进行了研究,在语言模型框架下提出了一个新的两阶段混合模型(two - stage mixture model)。新模型优先考虑初始查询条件,利用聚类将所有反馈文档分成两类,即 relevant 和 background,它力图从整个反馈文档集里自动确定真正相关的文档,使得参数设置更加灵活。文献[7]对上述两种思路均进行了探讨,首先提出了利用 CS(clarity score)和 DCG(discount cumulative gain)值作为依据来决定最优反馈文档数目的方法;然后从另一个角度提出了混合模型的构造方法,即不论 N 值大小将所有 N 个查询语言模型捆绑在一起形成一个目标模型,力图平滑不同模型的效果,弱化 N 值的影响。相关实验表明两种方法都能提高整个检索性能。文献[9]对如何获取最优反馈文档数目提出了一个全新的方法,对于给定的一个查询,首先创建语言模型集合来代表在不同参数条件下的扩展形式,然后估计哪种模型能够带来最好的检索性能,利用选择的模型来进行检索。其中模型的选择标准为最大 CS(clarity score)值。

目前,伪反馈方法的健壮性问题仍然没有得到较好地解决,它仍然是个开放和比较活跃的研究领域。

对初始检索结果进行调整[13~16]或者聚类[17~21]是确定相关文档的另一种手段,它们最终目的都是为了获得高质量的反馈文档。早

期的研究中，Mitra[13]等人试图利用自动构造的 fuzzy 布尔过滤器对初始检索结果进行重新排序，以改进初始检索结果的质量；Pedro Amo[14]等人认为，初始检索结果中相关度的排序关系在一定程度上反映了文档的重要性，提出了利用平滑函数来对排在不同位序的文档进行加权，使得排在前面的文档能够在反馈中发挥更大的作用。这些方法均取得了一定的成效。

近几年，学者们从另一角度提出了判断伪相关文档的方法。文献[15]提出了一系列的特征，比如查询词项的熵、文档的初始相关度值、反馈文档与整个反馈文档集的相似度以及扩展词与初始查询词之间的距离等因素，根据这些特征将高质量的相关文档从反馈文档集里过滤出来。文献[16]提出了伪无关文档(Pseudo－Irrelevant Document)的概念。顾名思义，伪无关文档就是与用户查询不相关的文档，令 F_R 代表初始检索结果的前 N 篇文档集，F_I 表示伪无关文档集，X 是超出 N 之外的高分值文档集，Y 是与 F_R 集合中任意文档相似的文档集，则 $F_I = X - (X \cap Y)$。通过在初始检索结果集里去除伪无关文档，从而保证查询扩展能在相关文档中进行。

对初始检索结果聚类也是确定相关文档的一种有效手段。通过聚类对初始检索结果集进行取样与重取样，从而提高伪相关文档质量。

Sakai[17]提出了基于取样的伪相关文档选择标准，对前 N 个返回文档进行筛选。筛选的原则体现了聚类的思想，簇的产生不是利用文档间的相似度，而是共同的词项集合，即文档之间如果包含的词项重复过多，则这些文档应该被聚为一个簇。因此，这样的文档应该被跳过，不应该被选入反馈文档中。文献[18]认为该方法在有些环境下具有局限性，比如在长文档或者短文档两种极端情况下性能不佳，据此提出了改进的聚类方法。该方法认为一个较合理的簇应该考虑双重因素：簇的大小以及簇内部相似度。大簇且低的簇内部相似度表明簇中含有噪音词项的干扰。因此，该文提出了基于局部频率词项的簇内部相似度的聚类过程，并对各簇进行排序，那些分布在位序较前的簇里且同时在该簇中排序靠前的文档被选作伪相关文档。Kyung Soon & Croft[19]提出了基于聚类的重新取样方法来更好地选

择伪相关文档，这个更好的伪相关文档就是领域文档(dominant document) 。文中认为领域文档应该与它最近邻居具有很高的相似度，并且可能属于几个不同的簇; 另一方面，一篇非相关文档理论上应该只形成一个簇，并且没有近邻。基于这种假设，文中采用了 K - nearest neighbors(K - NN) 方法对检索结果的前 N 篇文档进行可重叠聚类。Collins&Callan 在文献[20] 中提出了不确定性是信息检索的特性，比如在查询意图表达，文档语义表达等方面均存在着不确定性。文中采用不确定性分析的取样方法来构造反馈模型，引入了 bootstrap 方法来对初始检索结果的前 N 个文档进行重抽样，并对查询进行不同的变形，来达到提高检索的准确性与稳定性的目的。文献[21]在统计语言模型的框架下提出了一种基于独立分量分析(Independent Component Analysis，ICA) 的统计语义聚类方法，文中假设文档由多个隐含的独立主题混合噪声组成，ICA 可以分离文档中的独立主题形成 ICA 语义空间，并对此进行软聚类。通过用户初始查询与语义聚类的相似度决定使用某个语义聚类下文档估计查询模型。该模型由于滤去了语义聚类下查询不相似文档，因此反馈文档质量得到了提高。

(2) 查询扩展

伪反馈的最终目的是进行查询扩展。在伪相关文档确定之后，如何挑选有效的扩展信息是伪反馈中第二个关键问题。对 XML 文档而言，这种扩展信息不仅包含词项的扩展，而且还有结构约束信息的扩展。

在词项扩展中，查询扩展往往基于各种不同的检索模型。在最早的向量空间模型里，查询扩展主要利用 Rocchio 算法形成新的查询向量，并在 SMART 检索系统中实现了伪反馈机制[22] 。在经典概率模型中，查询扩展主要基于 Roberton/Sparck - Jones 权重[23] 挑选扩展词项。近年来，随着语言模型在信息检索领域中的广泛应用，许多查询扩展技术在语言模型框架下也得到了发展，比如基于混合模型的反馈方法[24] 、基于相关模型的反馈方法[25] 等，它们基本的思想都是利用反馈文档来估计一个更好的查询语言模型。Cao[26] 在混合模型的基础上通过实验验证了传统方法中仅基于词项在文档中的分布情

况的词项扩展标准不充分,事实上,只有大约18%的词对检索质量有帮助,大部分都是不好的或者是中性的词。该文基于SVM的统计分类器提出了选择好的扩展词项的特征:比如词项在伪相关文档和整个文档集中的分布率、词项的共现率、词项的距离信息以及查询词和扩展词项同时出现的文档个数等。文献[27]对相关模型进行了扩展,提出利用词项的位置信息作为线索来推断词项是否与查询主题相关。文中认为,与查询词位置越近的词项,与查询主题越相关。相比以往伪反馈技术中基于文档粒度或者段落粒度进行词项统计,该文以词项为粒度,在统计语言模型框架下结合词项的位置以及距离信息提出了一个全新的位置相关模型PRM(Positional Relevance Model),对反馈文档中相关位置上的词项附以更高的权重。实验结果表明该方法性能优于基于文档粒度的反馈以及基于段落粒度的反馈。

国内利用伪反馈进行查询扩展的方法主要都是在Rocchio框架下进行改进。文献[28]提出了一种基于局部共现的查询扩展方法,利用伪相关文档集中词项与初始查询的共现程度来评估扩展词的质量,并综合考虑了词项在语料集中的全局统计信息,使得选取的扩展词与初始查询所表达的主题具有更好的相关性。文献[29]提出了基于矩阵加权关联规则挖掘的查询扩展方法,在伪相关文档集合里运用MWARM(Matrix - Weighted Association Rules Mining)算法进行矩阵加权关联规则挖掘,构建与原查询项相关的规则库,并据此提出相应的扩展词项权重计算方法。实验结果表明检索性能得到了很好地改善。

以上这些扩展方法都是基于传统文本的伪反馈,并没有针对XML文档来进行。目前,基于XML文档上的伪反馈研究成果极少,大部分的研究工作都是基于相关反馈模型[30~34]。

由Schenkel R和Theobald M共同撰写的文献[30~32]提出了"查询词+标签结构"的查询扩展方法,它是目前所有XML查询扩展文献中最完善的,不仅给出了完整的理论模型还进行了实验验证。文献提出:结构特征包括词项特征(C)、文档特征(D)、祖先特征(A)和祖先后裔特征(AD)。其中,词项特征是指元素节点下面包含的所有词

项;文档特征为文档中元素的后裔里包含的所有 tag - term(元素标记 - 词项)对及其得分;祖先特征为用户标识元素的祖先节点中所包含的所有 tag - term 对及其得分;祖先后裔特征为用户标识元素的祖先节点中所有后裔的 tag - term 对及其得分。该方法基于相关反馈模型,在相关元素的候选集里进行抽取和筛选。从而,扩展后的查询形式为:

//ancestor - tag [A + AD constrains]//* [keyword + C + D constrains]

万常选[33]在相关反馈模型基础上针对 XML 文档提出了一种新的查询扩展方法,不仅包括了查询词的扩展,而且在查询词扩展的基础上还提出了结构扩展方案,最终形成了完整的"内容 + 结构"的查询扩展表达式。在进行查询词扩展时,该文献充分考虑了 XML 文档的结构特性,提出了扩展词的选择标准,除了传统信息检索中的词项频率以及词项的反比文献频率因素以外,还包括与 XML 文档结构特性相关的三个权重影响因素:词项所属元素的语义权重、词项所属元素的层次以及词项与初始查询词间的距离关系。结构扩展方案中,对于在相关文档中共同出现的每一个权重查询词 $term_i$,分别在相关文档中找到公共的语义权重最大的元素 $tagw_i$,$tagw_i - term_i$ 对作为最终结构查询扩展的一个分支结构;文档根元素节点作为结构查询扩展的根节点,由这些分支和根节点共同形成用户的扩展查询表达式。用 INEX 的查询语言 NEXI 表示的扩展查询表达式的形式如下:

//Dr [about(. ,//$tagw_1$,$term_1$) and about(. ,//$tagw_2$,$term_2$) and…]

其中,Dr 为查询文档集中的文档根元素节点,$term_i$ 为扩展查询词,$tagw_i$ 为相关文档中 $term_i$ 所属的公共的语义权重最大的元素。实验数据表明扩展后的查询性能得到了显著提高。

类似的,文献[34]提出了 XML 元素级的反馈算法,在传统伪反馈模型基础上不仅对查询词进行了扩展,而且还针对 XML 的结构特征提出了结构信息的扩展方法。首先进行查询词项的扩展,在 Lemur 系统的打分模型上进行权重的修正,将高权重词项挑选出来并选择与用户查询需求相关的词项作为扩展词项,同时将查询词的权值信息直接加入到系统的评分模型中。结构信息的扩展主要思想是将元素路径视为多个 tag 或者 field 的有序组合,在扩展词项的基础

上对出现在相关元素路径里的每一种 tag 打分，并据此对各个元素的路径进行打分，分值作为元素得分的一部分，路径得分高的元素将在结果中排在前面。

1.3 本书的研究思路与主要研究内容

一个完整的信息检索系统包括数据的组织和存储、用户查询处理和检索输出，以及反馈和查询扩展等三大部分。本书的研究内容属于第三部分，即对搜索引擎返回的初始结果进行伪反馈，从而进行有效的查询扩展。

本书的研究工作主要围绕提高 XML 检索质量，有效解决伪反馈中存在的“查询主题漂移”现象，将研究焦点定位于伪反馈中存在的三个问题的解决方法研究，其具体整体研究思路如图 1 - 1 所示。

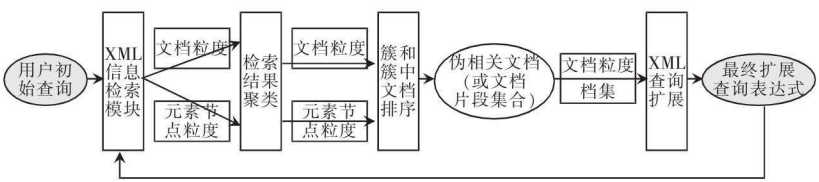

图 1 - 1 研究思路

伪反馈中的首要问题是确定高质量的相关文档，为此，本书采用聚类的技术手段，力图将检索结果中与用户查询意图相关的 XML 文档（或文档片段）聚簇在一起，由于 XML 检索结果的不同返回粒度，因此本书采用不同的方法进行研究，得到不同的聚类结果。然后以此为基准，相继对各个簇和簇中的文档（或文档片段）进行排序，形成高质量的伪反馈文档（或文档片段）集合，最后在得出的伪反馈文档集合里进行相应的查询扩展，得到一个最终的查询扩展式，系统自动根据此查询扩展式进行进一步的检索。

具体研究内容主要包括：

（1）研究了 XML 检索结果聚类问题。XML 检索不仅可以返回整篇文档，而且还能够指向特定片段（如 XML 元素节点）。本书主要面向文本为中心的 XML 数据环境，因此其核心问题就是要针对不同

的返回粒度，结合 XML 文档的内容和结构双重特性提出合理的相似度衡量方法，力图将与查询主题相关的文档(或文档片段) 聚簇在一起。

(2) 研究了基于检索结果聚类的文档(或文档片段) 排序模型。重点是在上述聚类结果的基础上制定相应的排序准则，从而确定高质量的伪相关文档(或文档片段) 集合。具体内容包括候选簇的排序模型以及候选簇中的文档(或文档片段) 排序模型。两者在不同返回粒度的解决方案均有所不同。

(3) XML 查询扩展的研究。利用上述伪相关文档集合中的文档进行查询扩展，不仅研究了词项的扩展，而且还在词项扩展的基础上获得了完整的“内容 + 结构”的查询扩展表达式。

1.4 结构安排

本书的内容共分为 6 章，各章的内容安排描述如下:

第 1 章: 引言，主要介绍本书的研究背景和意义，并从总体上对 XML 伪反馈的研究现状进行回顾，介绍本书的主要研究思路和结构框架。

第 2 章: XML 信息检索与反馈技术，主要介绍本书的研究基础，包括信息检索的模型和性能评价指标、基于反馈的信息检索、XML 的基本概念以及 INEX 评测，为后面章节打下基础。

第 3 章: XML 检索结果聚类，在文本为中心的数据环境下，依据不同的检索返回粒度研究 XML 文档(或文档片段) 间相似度衡量方法。在以文档为返回粒度的聚类中提出了带结构语义的扩展向量空间模型，并以此为基础提出了内容和结构语义相融合的相似性度量方法 CASS; 在以元素节点为返回粒度的聚类中进一步提出了基于词项语义的内容和结构语义相融合的相似性度量方法 LSI - CASS。同时，对聚类中存在的最优划分问题也进行了探索，提出了基于优化初始中心点和评价函数的 k - medoid 聚类算法。

第 4 章: 基于检索结果聚类的文档(或文档片段) 排序模型，主要针对不同检索粒度所获得的不同聚类结果制定相应的排序准则，研

究了候选簇的排序模型以及候选簇中文档(或文档片段) 的排序模型。

第 5 章: XML 查询扩展,主要针对 CO 查询提出了基于伪反馈的关键词扩展和结构扩展。

第 6 章: 结论与展望,对本书的内容进行回顾,并列举了本书的不足之处,以及今后进一步工作的方向。