●陆 伟

元素级 XML 检索模型构建的关键问题与解决方案研究*

摘 要 与传统信息检索不同的是 XML 要实现元素级的检索,其核心是元素级检索模型的构建。而 XML 文档内上下文元素的相关性、元素之间信息的重复性以及元素大小的不一性等则是构建模型时面临的核心问题。解决办法是:构建基于 BM25 元素级 XML 检索模型,构建基于上下文的元素级 XML 检索模型 BM25E,过滤重复元素,进行可检索元素的选择和太小元素的处理。表 1。图 1。参考文献 19。

关键词 信息检索 XML 元素检索 检索模型 模型构建

分类号 G354

ABSTRACT Different from traditional information retrieval, XML is used to realize element-level retrieval, with the core of constructing an element-level retrieval model. In this paper, the author analyzes key problems in the construction of the model, and proposes some solutions, such as constructing a BM25-based element-level XML retrieval model, constructing a context-based element-based XML retrieval model BM25E, filtering duplicated elements, and selecting searchable elements and processing too small elements. 1 tab. 1 fig. 19 refs.

KEY WORDS Information retrieval. XML element retrieval. Retrieval model. Model Construction.

CLASS NUMBER G354

1 概述

传统信息检索研究往往关注于非结构化信息(自由文本),而很少关注文档结构所蕴涵的语义信息。XML 作为半结构化信息的标记语言,不仅需要考虑如何从文档中找到相关信息,而且也需要考虑相关信息的结构和粒度问题。也就是说,XML 检索与传统信息检索的不同之处在于它不仅要求实现文档级的检索,而且需要实现元素级的检索(所谓文档级检索是指检索结果返回的记录是一个个 XML 文档,而元素级检索是指检索结果要求返回的记录是 XML 文档中的一个个元素。实际上 XML 文档级检索要求返回的是根元素,因而也可以看成是元素级检索的一个特例,这里,将它分为文档级检索和元素级检索)。而即使是文档级的检索,XML 检索也与普通文本信息检索有所不同。XML 文档通常包含一些子域(元素),如 IEEE 提供的 INEX 2005 XML 数据集就包括文章标题、摘要、正文、章节标题、参考文献及附录等,研究证明探讨文档的内部结构对提高检索性能有一定帮助。与文档级 XML 检索不同的是,XML 元素级检索在检索的目标与用户需求、元素与内容的存储与索引、检索算法与模型构建、检索结果的呈现及相关反馈等多方面都有很多不同之处,尽管已经出现了一些 XML 检索系统,如 STORED,XISS\R,Tamino 和 HYREX 等,然而迄今为止,尚未出现公认的 XML 检索模型,关于 XML 上下文(包括文档与元素、元素与元素之间)的相关性、重复元素的过滤、可检索元素的选择和太小元素的处理等问题仍然需要做大量的实验和用户实证研究。本文就将在对上述问题进行分析的基础上,探讨元素级 XML 信息检索模型的构建。

2 XML 元素检索模型构建的关键问题分析

2.1 XML 元素上下文的相关性

XML 文档的内部结构信息对提高文档级检索的性能有一定帮助。而对于元素级 XML 检索来说,对于文档内某一特定元素,其他元素都可以视为它的上下文元素,由于文档内信息的关联性,我们有理由相信,这些元素信息并不是孤立的,上下文元素应该在一定程度上影响该特定元素信息的相关性,进而影响元素检索的性能。然而具体是否有影响,影响程度如何,还需要在实验中予以验证。

* 本文系国家社会科学基金项目"基于 XML 的多媒体信息检索模型及实现研究"(编号 06CTQ006)的研究成果之一。

目前,国际上已有部分学者直接或间接地做了部分工作,其中探讨最多的是元素父子之间的继承关系,如 Abolhassani[1]、Geva[2]以及 Ogilive[3]等用不同方法归并子元素的权重以获取父元素的权重值;也有学者试图通过探讨文档检索与元素检索的关系来间接对此问题进行探讨,如 Sigurbjornsson[4]和 Mass[5]等用如下公式探讨文档权重对元素权重的影响:

$$Sn = DocPivot \times Sa + (1 - DocPivot) \times Sc \quad (1)$$

其中 DocPivot 为 0 ~ 1 之间的调适变量,Sa 为文档权重,Sc 为元素权重,Sn 为归并后的元素权重。Mass 采用该方法在 2004 和 2005 年度的 INEX 活动中取得了较大成功。然而上述两种方法都是采用线性权重归并的方式,正如 Robertson 所指出的那样,在词频、文档长度、域值合并等角度存在着缺陷[6]。本文将从词频归并的角度予以研究。

2.2 重复元素的问题

XML 文档中元素以类 B + 树的层级结构呈现,这意味着位于文档中某个位置的词可能同时属于多个具有继承关系的元素。用该词检索时,这些具有继承关系的元素都将同时出现在结果集中,它们被称为重复元素。显然,返回所有的重复元素并不是理想选择,如何过滤这些重复元素并确定应该返回哪些元素到结果集中就成为构建检索模型时要考虑的另一个重要问题。

为研究此问题,INEX 2005 对此进行了规定并设立了两种检索策略,即 Thorough retrieval Strategy 和 Focused retrieval strategy[7]。前者不考虑重复元素的状况,检索的所有元素都将在结果集中按元素权重得分由大到小排列;后者从元素的穷尽性(Exhaustivity)和专指性(Specificity)两个角度对重复元素的过滤做了规定,即给定查询语句,希望能够尽可能返回穷尽性和专指性最高的元素,而且结果集中不允许有重复元素出现。Kazai[8]给出了计算元素穷尽性和专指性的模型 nXCG,该方法成为 INEX 2005 的官方评价方法,学者们试图利用它探索重复元素过滤的最佳方法和算法模型,然而到目前为止,关于重复元素过滤的研究还处于起步阶段。

2.3 可检索元素和太小元素问题

与传统全文信息检索不同的是,XML 文档中的元素大小参差不齐,从整个 XML 文档到文档中的一个单词都可能是一个独立的相关元素,如 IEEE XML 数据集中的作者信息等元素。一方面,针对不同的文档集合,不同的检索需求,要求检索的元素类型是不同的,我们把纳入检索目标内的元素称为可检索元素,XML 元素级检索模型设计之前一个要重点考虑的问题即是可检索元素的选择问题。另一方面,长度参差不齐的元素对检索模型的设计也带来了极大挑战,在目前的检索模型中,尤其是那些考虑了文档长度信息的检索模型,对于同样的词频,长度较短的元素往往排在结果集的前列。而实际上,有些元素长度太小并没有包含完整的信息,以至于它们虽然相关但也不适合作为结果返回,这些元素被称为太小(too small)元素。如何解决太小元素的问题并尽可能在结果集中过滤掉也是需要重点考虑的问题之一。

3 模型构建的解决方案

与传统文本检索相比,XML 元素级检索存在很多不同之处。然而两者又有很大的紧密联系,从目前来看,大部分 XML 检索模型都是在传统信息检索模型的基础上发展起来的。为论述方便,我们首先在一定的假设基础上应用 BM25 概率检索模型实现 XML 元素级检索,然后在此基础上,提出针对各问题的解决方案。

3.1 基于 BM25 的元素级 XML 检索模型

BM25 模型是由 Robertson 等[9]提出的,它是典型的概率检索模型,包含多个变种,用以实现不同的信息检索目的,如 BM25b,BM250,BM251 等。本文首先直接应用该模型实现 XML 元素检索,这主要是基于两个假设:(1)XML 文档集中的每个元素都可以被看做是一个文档;(2)无论是相同文档内的元素还是不同文档内的元素,它们彼此之间都是独立的。

基于此假设可以给出如下定义:给定文档集 C,它共包含 n 个元素,即 $e = 1, \cdots, nE$。对于给定元素 e,查询词 j 的词频为 $tf_{e,j}$,el_e 为元素长度,$avel$ 为平均元素长度,则应用 BM25 模型可以得到如下公式:

$$w_j(e,d,C) = \frac{(k_1+1)tf_{e,j}}{k_1((1-b)+b\dfrac{el_e}{avel})+tf_{e,j}} \log \frac{N - df_j + 0.5}{df_j + 0.5} \quad (2)$$

从公式 2 可以看到,我们并没有用元素词频 ef_j 代替文档词频 df_j,关于这一点还需要实验研究,公式 2 中的另一个问题是 $avel$,我们分别使用了文档长度、所有元素平均长度、可检索元素平均长度、评估集合元素平均长度等作为它的值,研究发现,它们在检索性能方面并没有太大差异[10]。

3.2 基于上下文的元素级 XML 检索模型 BM25E

很明显,BM25 模型是建立在所有元素彼此之间互不相关的假设之上,但实际上,相同文档内的元素彼此之间并不是孤立的,它们之间存在着或多或少的语义关系,这种关系使得上下文元素在一定程度上将影响文档内元素的相关性,因此在构建检索模型时,有必要将这些上下文元素信息纳入到模型考虑的范围。线性归并域加权词频法是由 Robertson 等提出

的[11]，笔者等利用该方法实现了XML文档级的检索[12]，并进而以该方法为基础，提出了元素级的域加权检索模型BM25E[13]，该公式为：

$$wf_j(e,d,C)=\frac{(k'_1+1)tf'_{e,j}}{k'_1((1-b)+b\frac{el'_e}{avel'})+tf'_{e,j}}\log\frac{N-df_j+0.5}{df_j+0.5} \quad (3)$$

其中$tf'_{e,j}$是加权后第j个词在元素e中的词频，el'_e是加权后元素的长度，$avel'$是加权后的平均元素长度，k'_1是加权后的自由参数。

表1 BM25E在INEX 2005 CO. THOROUGH上的部分评价结果

结果集	INEX官方评价指标							
	Quantization: strict, Overlap = off							
	nxCG(10)	rank	nxCG(25)	rank	nxCG(50)	rank	MAep	rank
Run1	0.0538	10	0.1174	2	0.1539	2	0.0256	7
Run2	0.0500	13	0.1189	1	0.1931	1	0.0241	12
Run3	0.0231	34	0.0355	36	0.0962	16	0.0172	27

BM25E模型的基本思想是，XML文档集中的每个元素都可以被看做是一个文档，与基于BM25的元素级XML检索模型的不同之处在于给定特定元素，该模型允许其从所在文档内的特定上下文元素中继承相关信息。应用该模型，我们参加了INEX 2005年度XML检索活动，所提交的结果集在CO. THOROUGH任务的一些评价指标中排在所有参加者的前列，如表1显示了当忽略元素重复问题时3个结果集返回高相关元素在共55个结果集中的得分排名情况[14]。但由于时间等因素的限制，我们在实际实现上，只选取了3个特定域即文章标题、章节标题和摘要作为可继承的上下文元素，在这方面还需要做进一步的工作。

3.3 重复元素的过滤

重复元素是指文档内有继承关系的元素。对重复元素，只要针对每一文档，分别过滤掉之内的重复元素，使保留下来的元素彼此之间无继承关系即可。至于如何过滤重复元素，目前并无定论，我们在具体实现上，采用了两种方法：

(1)权重直接过滤法。该方法是最简单易行的一种方法，其目标是尽量保留权重高的元素。首先对整个查询候选结果集进行排序，然后在该排序结果集中按照权重由高到低的顺序选取元素到最终结果集中，每选取一个元素时都要和最终结果集中同一文档内的元素进行比较，如果该元素与其中任一个元素有继承关系，则忽略该元素，直至整个过滤结束。该方法的缺点是没有考虑元素节点各权重的分配情况，我们将在进一步的研究中考虑元素节点权重分配因素。

(2)最佳元素法。上一种方法总是保证文档内权重最高的元素会保留在最终结果集中，但权重最高，并不代表该元素就是最佳元素。例如，给定元素A,B,C,D,E,F，其层次关系及权重如图1所示。根据最优过滤法，E,D,C将依次出现在结果集中，但有时父元素却有更高的相关性，此时希望元素B而不是E和D出现在结果集合中。这种情况下，我们的做法是设定一个阀值，当一个元素有两个或指定个子元素的权重大于该阀值时，则不管其子元素中是否有元素的权重大于该元素的权重，都将该元素优先置于结果集中。如在图1中，当设定如果有两个及以上的子元素权重大于1时，则返回父元素，根据该规则，元素B,C将依次出现在结果集中；如果设定阀值为0.9，则将只有元素A出现在结果集中。从INEX 2006官方实验评价结果来看，我们利用该方法提交的结果总体上排在中下游[15]。关于该方法是否可以进一步改进优化，仍有待进一步探索。

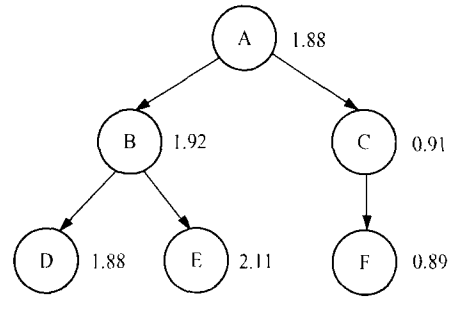

图1 XML文档的元素B+树表示及各元素权重得分实例

3.4 可检索元素的选择及太小元素的处理

关于可检索元素的选择问题，已有详细论述[16]。在具体实现上，可以有两种方法：其一是索引时控制，即在对XML元素进行索引时，根据设置的可检索元素来建立元素路径索引，而忽略对其他元素的索引；另一种方法是在检索时控制，即对文档集内所有元素建立索引，在检索时，根据设置获取可检索元素结果集。这两种方法各有优缺点，前者使得检索时速度较快，但降

低了灵活性,每当可检索元素发生变化时,都需要重新建立索引;后者的优点是灵活性较大,但检索时运行速度相对较慢。至于选择何种方法,可以根据实际情况确定,如 Mass 等在实验中只对 article,abs,sec,ssl,p,ipl,bm,bib 等重要元素建立了索引[17],而笔者等则是在检索时对可检索元素进行控制。

对太小元素的处理,到目前为止,主要有以下方法:其一,通过设置可检索元素限制小元素标签,如 Mass 只对少数重要元素建立索引,一定程度上解决了太小元素的处理问题,实验结果也证明这种方法有较高的查准率,该方法的缺陷在于包含在其他类型元素内近 30% 的相关元素被漏检,设置的可检索元素中也有一些元素太小不能作为相关元素。其二,在检索时,设定元素大小的阀值,阀值小于设定值时将被过滤掉。元素的大小既可以是元素的字节,也可以是元素中包含的词的数量,如 Geva 在实验中将该值规定为 25 个单词[18]。该方法在一定程度上弥补了第一种方法中的缺陷,但由于未对可检索元素进行控制,检索的查准率有所降低。Geva 在实验中进一步将两者结合起来使用,但同样存在着漏检问题。其三,在检索模型中,加入元素长度参数,使检索排序偏向于大元素[19],即在同样的词频下,元素的权重与元素大小正相关。该方法使得太小元素出现在结果集中的机会大大减少,却产生了另一个缺陷,并不是所有大元素都比小元素更相关,该方法也与传统的 tf-idf 模型相矛盾。

我们在实验中采用第一种方法和第二种方法相结合的方式,同时对 BM25 模型中的参数 b(该参数反映了文档相关性与文档长度的关系)进行调适,来解决太小元素的问题,在一定程度上也借鉴了第三种方法。利用这两种方法,我们参加 INEX 2005 和 INE 2006 的年度活动,它们亦被大多数 INEX 参加者所采用。

4 结束语

本文分析阐述了 XML 元素级检索需要解决的关键问题,提出了对应的解决方案并在实验中有具体实现。然而,XML 检索目前还是一个相对较新的课题,仍需要做进一步深入研究。尽管我们提出了面向上下文的元素级检索模型 BM25E 并在参加 INEX 2005 时应用该模型取得了不错的成绩,但是我们只对数据集中的少数域做了加权实验,当考虑更多的元素域时,模型的复杂度将大大增加,参数调适也将更加困难,是否需要引入基因算法等对参数进行调适还需要认真考虑。而关于重复元素的问题,由于这一问题目标的多元性,迄今提出的各种方法仍无法确切衡量其优劣,需要对重复元素的过滤目标做出更多研究。关于元素的选择问题,实验证明文中提出的两种方法基本可行,而关于太小元素的处理问题,笔者针对 INEX IEEE 数据集提出了一个集成的解决方案,但其普适效果如何,尚需要更多的数据集予以验证,如何在解决太小元素问题的同时降低漏检率,仍需要更多研究。

参考文献

1 M. Abolhassani, N. Fuhr, S. Malik. HyREX at INEX 2003. Proceedings of the Second Workshop of the Initiative for The Evaluation of XML Retrieval(INEX). 2003,27-32

2,18 S. Geva. GPX-Gardens Point XML Information Retrieval at INEX 2004. INEX 2004, LNCS. 2005 240-253

3,19 P. Ogilvie, J. Callan. Hierarchical Language Models for XML Component Retrieval. INEX 2004. LNCS. 2005, 224-237

4 B. Sigurbjornsson, J. Kamps, M. Rijke, An element based approach to XML Retrieval, Proceedings of the Second Workshop of the Initiative for The Evaluation of XML Retrieval (INEX). 2003,15-17

5,17 Y. Mass, M. Mandelbrod, Component Ranking and Automatic Query Refinement for XML Retrieval, INEX 2004, LNCS. 2005,73-84

6,12 陆伟. Stephen Robertson. 基于域加权词频法的 XML 文档级检索模型及其实现. 中国图书馆学报,2006(6)

7 M. Lalmas. INEX 2005 Retrieval Task and Result Submission Specification. Preproceedings of INEX 2005

8 G. Kazai, M. Lalmas. INEX 2005 Evaluation Metrics. INEX 2005, Lecture Notes in Computer Science. 2006,16-29

9 S. Robertson, S. Walker. Some simple effective approximations to the 2-Poisson model forprobabilistic weighted retrieval. Proceedings of the 17th annual international ACM SIGIR-conference on Research and development in information retrieval. 1994,345-354

10 W. Lu, S. Robertson, A. Macfar lane. Investigating Average Element Length for XML Retrieval by Using BM25. (待发)

11 S. Robertson, H. Zaragoza, M. Taylor. Simple BM25 Extension to Multiple Weighted Fields. CIKM'04. 2004,42-49

13,14 W. Lu, S. Robertson, A. Macfarlane. Field-Weighted XML Retrieval Based on BM25. Proceedings of INEX 2005. LNCS. 2006,126-137

15 INEX 2006. http://inex. is. informatik. uni-duisburg. de/2006/,2006,1-20.

16 S. Robertson, W. Lu, A. Macfarlane. XML-structured documents: retrievable units and inheritance. FQAS 2006. Springer LNCS. 2006,121-132

陆 伟 博士,副教授。通讯地址:武汉大学信息资源研究中心。邮编 430072。

(来稿时间:2007-01-26)