XML 文档结构相似测度研究*

闫利国[1]，贺 飞[1,2]

(1. 清华大学 软件学院；2. 清华大学 计算机科学与技术系，北京 100084)

摘 要：为了满足基于 Web 的 XML 数据信息的近似搜索、信息分类以及数据交换的需求，提出一种新的有效地鉴定 XML 文档间结构相似度的标准。该标准包含了 XML 文档的结构信息和节点嵌套的语义信息，可以有效地给出 XML 文档间的结构相似测度。通过实验证明该标准具有高度的准确性和有效性。

关键词：可扩展标记语言；结构相似测度；编辑距离

中图法分类号：TP391 文献标识码：A 文章编号：1001-3695(2006)03-0044-03

Research on Evaluating Structural Similarity between XML Documents

YAN Li-guo[1], HE Fei[1,2]

(1. *School of Software, Tsinghua University*; 2. *Dept. of Computer Science & Technology, Tsinghua University, Beijing* 100084, *China*)

Abstract: For sake of increasing requirement about approximate search, data cluster and data exchange from XML documents in the Web, a new effective metric for evaluating structural similarity between XML documents is brought forward. It's accurateness and effectiveness are testified by the experiments.

Key words: XML; Structural Similarity; Edit Distance

1 引言

随着 Web 的广泛使用，XML(eXtensible Markup Language，可扩展标记语言)以其巨大的通用性和灵活性逐渐成为了企业所关注的焦点，也为基于 Web 的信息交换带来了新的希望。它不仅可以标记非结构化的数据，也可以标记高度结构化的数据，如数据库中数据。但是 XML 数据是半结构化的，在搜索处理这些半结构化的数据信息时，已有的用于处理较为结构化数据的 XML 查询语言(XQuery Language，XQL)就不再适用，特别是在用户需要查找与某一信息相关(但不完全一致)的数据时。因此在搜索异构的 XML 数据时，还应该研究基于 XML 文档的近似搜索技术。近似搜索技术的基础是要能够准确地度量所查询信息与文档、文档与文档间的相关性与相似性。传统的信息检索技术也可以应用于 XML 文档，但是这种基于向量空间模型的检索技术并不能反映 XML 文档中节点嵌套结构的语义信息[8]。本文中给出一种有效地确定 XML 文档(实例)间相似测度的标准，不仅考虑了 XML 文档的结构信息，而且考虑了 XML 文档中节点的语义信息。

尽管 XML 的文档结构可以由一个描述该文档的 DTD(Document Type Definition，文档类型定义)或 XML Schema 决定，但是 Web 上的 XML 数据往往很难找到与之相应的 DTD 或 XML Schema，并且从一个 XML 文档提取描述其结构的 DTD 或 XML Schema 并不容易。因此本文描述的方法是基于 XML 文档实例的。

收稿日期：2005-03-19；修返日期：2005-05-15

基金项目：国家“863”计划资助项目(2003AA4Z3210，2003AA413031)

本文主要介绍近年来国内外工作者在 XML 文档结构相似测度方面所作的研究及其发展状况，介绍了有效评价 XML 文档结构相似测度的标准，以及 XML 文档结构信息提取与建模、结构相似测度标准定义等。

2 相关研究

最近几年，许多学者对 XML 文档的相似测度问题进行了广泛研究。其中比较传统的一种方法是把 XML 文档之间的相似性用树之间的编辑距离(Edit Distance)来度量[4]。编辑距离是指由一棵树转变为另一棵树的最少操作步骤。这些操作包括节点改名、插入和删除等。Nierman 和 Jagadish[5] 对该方法进行了优化，又增加了子树的插入和删除操作，这使得计算两个文档间的相似度变得更加灵活。但是它的计算代价却很高，至少是 $O(n^2)$。

另外，在某些情况下，采用编辑距离无法区分树与树的差别，从而无法合理地对 XML 文档进行分类。如图 1 所示，Doc_1 和 Doc_2 之间的编辑距离与 Doc_2 和 Doc_3 之间的编辑距离相等，因为将源树转变为目标树仅仅只要两步改名操作。但是图 1 中从语义的角度讲 Doc_2 和 Doc_3 应该属于同一类，即它们的相似度应该高于 Doc_1 和 Doc_3(或 Doc_2)。

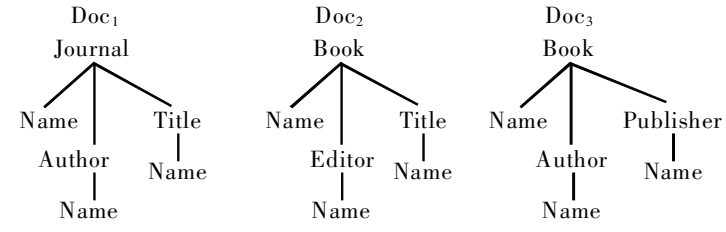


图 1 文档间的 Tree Distances

Sergio Flesca 在文献[2]中把 XML 文档进行编码(该编码

不同于本文所述的编码), 然后将 XML 文档表述为一个傅里叶展式。该方法的缺点在于对快速傅里叶变换的依赖性很强, 这就影响了其对 XML 文档处理的速度, 其代价最少也为 $O(n\lg n)$。另外更重要的是, 该文对 XML 文档建模时所采用的模型并没有合理地将 XML 文档结构中的相关信息表述出来, 因此它在计算不同 XML 文档间的相似测度值时其效果并不是很理想, 见本文第 4 节。

在文献[3]中, Wang Lian 等人定义了判断相似结构之间距离大小的标准。这种方法需要对每个 XML 文档进行分析处理, 并根据文档的结构信息构建一个有向图(S-G 图)。测量 XML 文档间相似度大小的标准就是建立在这个有向图的基础之上。标准的实质是用两个有向图中所包含的公共边数与两个文档中边数较大的一个比值大小来确定文档间距离大小。但是许多结构不同的 XML 文档却可能由几乎完全相同的元素(或属性)构成, 因此这个标准存在一定的局限性, 因为它仅仅考虑了公共边的数目, 而没有考虑每条边在 XML 文档中所起的语义作用。如图 1 所示, Doc_1, Doc_2 和 Doc_3 之间彼此仅仅有一条公共边, 如果按照文献[3]中所定义的标准, 这三个文档之间的距离相等, 应属于同一个分类; 但是从语义的角度讲它们却应归属两个完全不同的类别(Journal 和 Book)。

本文介绍一种不同的计算 XML 文档间相似测度的标准。该标准不同于上述提到的各项标准, 它考虑了每一个节点的语义特征, 可以根据 XML 文档的不同结构有效地计算其结构相似测度值。

3　测定 XML 文档的结构相似度

XML 文档格式作为当前网络信息传输的主要方式, 其通常含有大量的信息实体, 并且这些信息实体只归属于有限的几类信息结构模式。因此一般的 XML 文档会含有大量相似结构的冗余信息。为了有效测定 XML 文档间的结构相似测度, 我们所关心的只是与 XML 文档结构有密切联系的信息, 而不是文档中表示事物具体信息的数据本身。测定 XML 文档间的结构相似测度需要筛选并去除 XML 文档中与其结构不相关的冗余信息, 提取 XML 文档的有效结构并为之建模。根据模型并依据判定 XML 文档间结构相似度的标准即可得到 XML 文档间的结构相似测度值。整个测定过程的重点包含: 提取 XML 文档的有效结构、结构建模和相似测度标准定义。下面将对各个部分分别加以详细描述。

3.1　提取 XML 文档的有效结构

XML 文档实例是由许多具有具体数据值的元素和属性(本文认为某个元素的属性节点与该元素的子元素节点是等价的, 即元素的属性将视为其子元素节点来处理)构成, 往往同一个元素名可能拥有多个数据值。但是我们感兴趣的是 XML 文档的结构, 而不是具体的数据值。与其实质的文档结构(图 2(b))相比, 这些重复信息(图 2(a))是多余的。

但是图 2(b)所表示的 XML 文档结构并不是最有效的, 它同样包含有冗余的节点信息(Name 节点)。当前已有许多学者对 XML 文档的特征进行了深入研究, 并提出许多提取 XML 文档结构的方法。本文将基于文献[7]中提出的状态最小化算法对 XML 文档进行处理, 即对图 2(b)所示的 XML 文档结构进行优化, 使其最小化。图 2(c)即为所得的 XML 文档的最小有效结构。

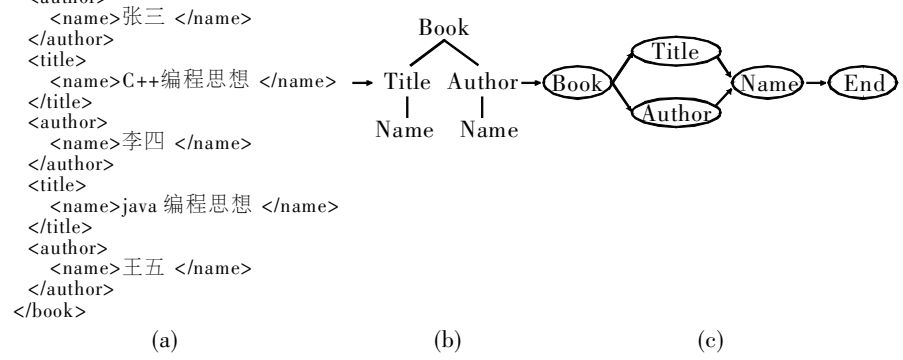


图 2　提取 XML 文档的有效结构

3.2　结构建模

经过第 3.1 节的处理, 可获得表征 XML 文档最小有效结构的有向图。为了能合理地表述 XML 文档结构的信息, 下一步需要在该有向图的基础上为 XML 文档建模, 以表达出文档中每个节点所具有的语义特征。

建模标准是为每一个节点赋予权值, 以表征该节点的语义影响范围。节点的权值满足如下定义。

定义 1　End 节点的权值为 0; End 节点的邻节点其权重为 1; End 节点的非邻节点, 其权值为该节点到 End 节点所有可达路径上节点的权值之和。

这里定义了 End 节点的邻节点其权重为 1, 因为每个叶子节点的语义影响范围只有它自己本身。而 End 节点的非邻节点, 其语义影响范围是该节点到 End 节点所有可达路径上节点总和。其值越大它在该 XML 文档中的语义影响也就越大。图 3 即为图 1 中所示文档相应的最小有效结构图(赋权有向图)。其中 Doc_1 的结构图中节点 N(Name)为 End 节点的邻节点, 权值为 1; A(Author)节点和 T(Title)节点到 End 节点分别只途经 N 节点, 所以其权值为 1; J(Journal)节点到 End 节点有三条路径: J→A→N→End, J→T→N→End, J→N→End; 因此 J 节点的权值为 5。Doc_2 和 Doc_3 的结构图与 Doc_1 类似。

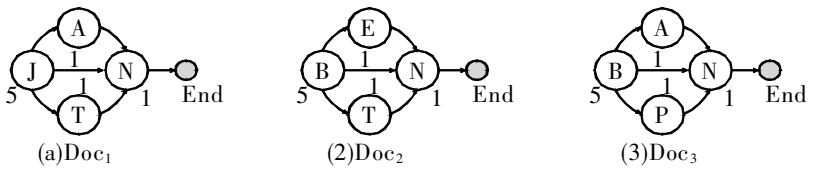
下面根据这个模型, 给出确定 XML 文档间相似测度的标准。设 D_1 和 D_2 分别为 XML 文档 $Document_1$ 和 $Document_2$ 的最小有效结构的加权有向图, 则 $Document_1$ 和 $Document_2$ 之间的相似测度值满足如下定义。

定义 2　给定文档 D_1, D_2, 它们之间的相对相似测度值定义为

$$Sim(D_1, D_2) = \frac{D_1 \text{ 和 } D_2 \text{ 所有共公节点的权值之和}}{D_1 \text{ 和 } D_2 \text{ 所有节点的权值之和}}$$

由于表征 XML 文档最小有效结构的有向图中每个节点都是唯一的, 如果两个文档所含有的公共节点数越多, 且语义影响的权重值之和越大, 那么这两个文档从语义的角度就越相似。这就避免了仅仅考虑公共边数目多少来决定文档相似测

度值的局限性[2]。



(a)Doc_1 (2)Doc_2 (3)Doc_3

图 3 Doc_1, Doc_2 和 Doc_3 的最小有效结构图

因此, 由定义 2, 据图 3 有

$$Sim(Doc_1, Doc_2) = \frac{2}{8} = \frac{1}{4} = Sim(Doc_1, Doc_3),$$

而 $Sim(Doc_2, Doc_3) = \frac{6}{8} = \frac{3}{4}$。满足 Doc_2 和 Doc_3 从语义角度讲相似度高于 Doc_1 和 Doc_3(Doc_2) 的特点。

4 实验结果

通过一些实验数据与文献[2] 中实验数据的比较, 证明本文所采用的标准在测定 XML 文档间相似测度值方面具有极高的准确性和有效性。

一般情况下, 一个 XML 的文档结构是由 DTD 或 XML Schema 来决定的。因此, 如果一些 XML 文档实例满足某个 DTD 或者 XML Schema, 那么它们的结构应该是比较相近的。下面我们给出一些实验结果的统计数据。用于实验的 XML 文档实例来源于五个不同的 DTD, 这些 DTD 在文献[2] 上曾经使用过(图 4)。统计数据主要由两部分组成: 一部分是满足同一个 DTD 的任意两个 XML 文档实例之间相似测度值的平均值; 另一部分是满足不同 DTD 的任意两个 XML 文档实例之间相似测度值的平均值。

按照这五个 DTD, 利用 XML 文档生成工具[6] 分别为它们生成 200 个 XML 文档实例。这是一个 SUN 公司提供的用 Java 书写的工具, 它可以生成满足各种 DTD 的 XML 文档实例。同时它还支持 RELAX Namespace, RELAX Core, TREX 和 W3C XML Schema Part 1 中的一个子集。根据这1 000个 XML 文档实例我们给出同一 DTD 下任意两个 XML 文档实例之间相似测度的统计平均值, 以及不同 DTD 间任意两个 XML 文档实例间相似测度的统计平均值, 如表 1 所示。

表 1 实验结果统计平均值

$Sim(S_i, S_j)$	S_1	S_2	S_3	S_4	S_5
S_1	0.994 742	0.517 226	0.206 502	0.176 347	0.153 223
S_2	0.517 226	0.993 407	0.164 563	0.157 247	0.149 283
S_3	0.206 502	0.164 563	0.960 795	0.196 202	0.159 570
S_4	0.176 347	0.157 247	0.196 202	0.983 1887	0.95 3246
S_5	0.153 223	0.149 283	0.159 570	0.953 246	0.997 280

其中 S_i ($1 \le i \le 5$) 表示满足第 i 个 DTD 的 XML 文档实例的集合。$Sim(S_i, S_j)$ ($1 \le i, j \le 5$) 表示分别属于 S_i 和 S_j 两个集合中的任意两个 XML 文档实例相似测度的平均值。

表 2 的实验结果是文献[2] 中给出的, 这里引用的目的是与本文的实验结果进行比较。通过比较 $Sim(S_i, S_j)$ ($1 \le i, j \le 5$) 的值, 可以看出, 本文所给出的测定标准对属于不同 DTD 的 XML 文档实例的区分效果明显好于表 2 给出的结果。

表 2 文献[2] 中的实验结果

$Sim(C_i, C_j)$	C_1	C_2	C_3	C_4	C_5
C_1	0.965 5	0.641 8	0.815 3	0.482 2	0.393 5
C_2	0.641 8	0.964 8	0.748 5	0.658 6	0.503 7
C_3	0.815 3	0.748 5	0.961 9	0.540 2	0.431 3
C_4	0.482 2	0.658 6	0.540 2	0.978 2	0.681 7
C_5	0.393 5	0.503 7	0.431 3	0.681 7	0.945 2

5 结论

本文中定义了一个用于计算 XML 文档间结构相似度的标准。该标准不仅考虑了 XML 文档单纯的结构信息, 而且考虑了每一个节点的语义影响。实验结果也证明了本文所述标准的有效性。该标准可以用于 XML 文档的近似搜索、XML 文档聚类、XML 文档结构抽取, 以及基于 XML 文档的数据交换平台。

参考文献:

[1] N Garofalakis, A Gionis, R Rastogi, *et al*. XTRACT: A System for Extracting Document Type Descriptors from XML Documents[C]. Dallas, Texas: Proceedings of ACM SIGMOD Conference on Management of Data, 2000. 165-176.

[2] S Flesca, G Manco, E Masciari, *et al*. Detecting Structural Similarities between XML Documents[C]. Proceedings of the 5th International Workshop on the Web and Databases, WebDB, 2002.

[3] Wang Lian, David Wai-Lok Cheung, Nikos Mamoulis, *et al*. An Efficient and Scalable Algorithm for Clustering XML Documents by Structure[J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(1): 82-96.

[4] K Zhang, R Stgatman, D Shasha. Simple Fast Algorithm for the Editing Distance Between Trees and Related Problems[J]. SIAM Journal on Computing, 1989, 18(6): 1245-1262.

[5] A Nierman, H V Jagadish. Evaluating Structural Similarity in XML Documents[C]. Madison, Wisconsin, USA: Proceedings of the 5th International Workshop on the Web and Databases, WebDB, 2002.

[6] http://wwws.sun.com/software/xml[EB/OL]. 2004-09-20.

[7] Alfred V Aho, Ravi Sethi, Jeffrey D Ullman. Compilers: Principles, Techniques, and Tools[M]. Publisher: Addison-Wesley, Hardcover, 1986. 796.

[8] 郑仕辉, 周傲英, 张龙. XML 文档的相似测度和结构索引研究[J]. 计算机学报, 2003, (9): 1116-1122.

作者简介:

闫利国 (1979-), 男, 山西太原人, 硕士, 主要研究领域为软件工程; 贺飞 (1980-), 男, 博士, 主要研究领域为算法、形式化方法、企业信息系统。