**XML快速检测算法研究[[1]](#endnote-1)**

X X

（X X X X X X X X X X X X X X X X X ，X X XXXXXX）

**摘 要**: 互联网X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X算法。

**关键词** XML，查询，索引

**中图分类号 TP 393 文献标识码 A**

**Research on Rapid Detection Algorithm for XML Documents**

XXXXX XXX

(X X X X., X XXXXXX, X)

**Abstract** Cyber X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X model．

**Keywords** XML，Query，Indexing

# **引言**

XML[5]作为一种自定义的数据格式，具有许多传统数据形式所不具备的特点。X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X。变化检测技术通常能为一个查询系统提供两种手段，即信息的累积更新能力和条件触发能力。

目前，在变化检测算法方面的研究主要有两类。X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X。下面首先给出了相关的定义，然后重点介绍了相关的算法，最后给出了算法分析以及部分实验结果。

# **概念定义**

**定义1** 假定X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X假定lab(r)=r。

**定义2** 定义X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X路径。

**定义3**定义X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X 无定义。

**定义4**定义X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X ，则VID(v')>VID(v)。

**定义5**定义X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X ，则MIID(v')>MIID(v)。

**定义6** 定义X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X且val(k1)≠val(k2)}。

**定义7**给定X X X X X X X X X X X X X X X X X X X X X X X X X X X X X条件：

1)如果X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X 节点的值。

2)否则，X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X ，KLab(v)。

**定义8**给定X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X ．Type(v))

1)如果X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X ．val(v))

2)否则，X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X ，则KHash(v)=hash(KHash(v1')⋯KHash(v N'))

**定义9**定义X X X X X X X X X X X X X X X X X X X X X X X X X ，PHash(e1)=PHash(e2)}

**定理1** 给定X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X ，有：

e1、e2X X X X X X X X X X X KHash(e1)=KHash(e2)

证明：X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X 相同。

充分性。X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X 。由此推出T1'、T2'叶节点的KHash值不存在一一对应相同的关系，同理推出KHash(e1)≠KHash(e2)。

# **F-Diff算法描述**

整个检测过程包括四个步骤。X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X 。限于篇幅，下面重点讨论过滤算法和映射算法。

## 3.1． 过滤算法

**输入** X X X X X X X X X X X X X X X X X

**输出** X X X X X X X X X X X X X X

**初始化**X X X X X X X X X X X X X X X X X X X X X X

**第一步** X X X X X X X X X X X X X X X X X

X X X X X X X X X X X X X

X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X

**第二步**

取X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X，KHash(em))

取X X X X X X X X X X X X X X X X X X 序列A2=(KHash(e1') ，⋯，KHash(en'))

取X X X X X X X X X X X X ，

X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X

X X X X X X X X X X X X X X X X X

X X X X X X X X X X X X X X X

X X X X X X X X X X X X X X X

X X X X X X X X X X X X X X XX X X X X X X X X X X X X X X

X X X X X X X X X X X X X X X X X X X X X X X X X X X X X

**第三步** X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X

## 3.2． 映射算法

**输入** X X X X X

**输出** X X X X X X X

**初始化** X X X X X X X X X X X X X X X X X X X X

**第一步** X X X X X X X X X X X X X X，

X X X X X X X X X X X X X X

X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X

**第二步**

X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X

X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X

X X X X X X X X X X X X X X X，

X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X

X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X

X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X

X X X X X X X X X X X X X X X

X X X X X X X X X X X X X X

X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X

X X X X X X X X X X X X X X X X X X X X X X X X X X X X X

**第三步**X X X X X X X X X X X X X X X X X X X X

X X X X X X X X

**算法分析**

在整个检测过程中，文档解析与哈希计算的算法复杂度为O(|T1|×log(|T1|)+|T2|×log(|T2|))，基本上与X-Diff相同，只是在节点哈希计算上由于增加了特征路径信息、子集排序以及路径哈希的计算，使得实际时间开销略有增加。X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X 。由此可以看出，本算法对文档变化检测过程中最复杂的匹配过滤和映射建立等环节进行了较大程度的优化，从算法复杂度和实际时间开销两方面都取得了明显的改进。

# **实验结果**

测试环境包括：Intel Core 2 Duo E8400 CPU，2G内存，Windows XP平台，IBM XML4C 3.0.1。X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X 。下一步的工作是扩大数据的背景领域，研究实际应用数据中特征标记域的特点以及相关的判定算法。



图1 1%变化率X X X X X X

参考文献

1. Berk E．HtmlDiff：A Differencing Tool for HTML Documents．

Student Project，Princeton University．

1. Chawathe S,et al．Change Detection in Hierarchically Structured Information．In：Proc．of the ACM SIGMOD Intl．Conf．on Management of Data．MontreaI, June 1996
2. Curbera F P，Epstein D A．Fast Difference and Update of XML Documents．XTech’99，San Jose，March 1999
3. Zhang K．A New Editing based Distance between Unordered Labeled Trees．Combinatorial Pattern Matching．1993．1：254~265
4. World Wide Consortium．Extensible markup language(xml)1.0．http://www.w3.org/TR/REC-xml，2000
5. Douglis F．et al．The AT＆T Internet Difference Engine：Tracking and Viewing Changes on the Web World Wide Web，1998.1(1):27~44
6. Maruyama H, Tamura K, Uramoto R. Digest values for DOM

(DOMHash)proposal．IBM Tokyo Research Laboratory．

http://www.trl.ibm.CO.jp/projects/xml/domhash.htm．1 998

1. Wang Y．et al．X-Diff：A Fast Change Detection Algorithm for XML Documents．Submitted for publication
2. Zhang K，Shasha D．Simple Fast Algorithms for the Editing Distance between Trees and Related Problems．SIAM Journal of Computing．1989，18(6)：1245~1262

1. X X（XXXX - ），X ，X X ，X X X X X X X X X X X X X X X X X X X X X ，X X X X X X X X X X X X X X X ，XXXXXXXXXXXXXXXXXXXXXX [↑](#endnote-ref-1)