情 报 学 报 ISSN 1000 - 0135
第 30 卷 第 2 期 160 - 165, 2011 年 2 月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000 - 0135
Vol. 30 No. 2, 160 - 165 February 2011

doi:10.3772/j.issn.1000-0135.2011.02.007

基于 HTML 树的网页结构相似度研究[1)]

宋明秋 张瑞雪
(大连理工大学系统工程研究所,大连 116023)

摘要 HTML 网页信息是一种半结构化的数据,而且不同网页之间在其结构特征方面都具有一定的相似性。本文就是从信息的结构性角度来研究不同网页信息块之间的相似性,并提出了基于子树最优自由匹配规则的结构相似度度量模型以及利用网页结构相似性提取网页信息的方法。本文中的计算方法都用 python 语言实现。通过实验,本文对不同网页之间的相似度进行了计算和分析,实验数据表明,基于子树最优自由匹配规则的树结构相似度度量模型具有较好的系统性和适用性;通过树结构相似度来确定网页内部元素及两个网页之间的联系,也弥补了传统方法中依赖单调的文本信息比较的不足,使得网页信息提取更加准确,更加迅速。

关键词 HTML 树 结构相似度 自由匹配 信息提取

Study on Web Structural Similarities Based on HTML Tree

Song Mingqiu and Zhang Ruixue
(*Institute of System Engineering*, *Dalian University of Technology*, *Dalian* 116023)

Abstract HTML web information is a kind of semi-structured data, and different web pages always have some kind of similarity in structure. From the perspective of information structure, this paper has studied the similarity between two different blocks of web information, and proposed a new model of calculating structural similarity based on optimally free matching on sub trees and a method of extracting web information by using structural similarity. All of algorithms in this paper are implemented by Python. We have calculated and analyzed the similarity between different web pages through experiment, which shows that our model of calculating structural similarity is of stronger systematicness and applicability. Compared with traditional method which relies on the monotony text information, the new structural-similarity-method makes full use of the relationship between different elements within a page or different pages, which makes web information extracting quicker and more accurate.

Keywords HTML tree, structural similarity, free matching, information extracting

1 引 言

随着 Web 信息资源的爆炸式增长,如何从海量数据中筛选出人们想要的数据就成为了一个富有挑战性的课题。传统的网页信息提取工具大都基于文本信息的匹配,并不能对复杂的结构化网页信息进行准确地比较和取舍。通过挖掘 Web 网页中的结构特性,并通过结构特性来确认所需信息,已成为一种准确迅速提取信息的有效方法[1]。比如通过识别网页正文信息或链接信息的结构特性,可以提取网页中的正文信息块和导航信息块。网页之间的结构相似性研究在搜索引擎领域也得到了广泛的应用[2,3]。将待处理网页与样本网页进行结构比较,根据网页之间的结构相似性程度来消除噪声,提取有较强相关性的网页信息,提高搜索引擎的准确性和

收稿日期: 2009 年 12 月 10 日

作者简介: **宋明秋**,女,1967 年生,大连理工大学副教授,研究方向:数据挖掘、信息安全等。**张瑞雪**,男,1987 年生,大连理工大学硕士研究生,研究方向:数据挖掘。E-mail:zhangruixue08@yahoo.cn。

1) 基金项目: 国家自然科学基金资助项目(70671016)。

效率[4,5]。

参考文献[6]最早提出了最少编辑距离法计算两棵树的相似度,即通过对节点进行删除、插入和移动等简单的操作,从一棵树转化到另一棵树,用最少的操作步数来衡量两棵树的相似度。该方法在以后的研究中得到了广泛的应用,比如衡量 XML 文档的相似度[7],但该方法允许子树之间跨层匹配,并不适合于层次性较强的 HTML 树结构的比较;相比之下,参考文献[8]提出的基于简单树匹配的结构相似性度量方法更具有层次性和系统性,但其度量过程中使用的匹配规则对树结构的节点数量和顺序要求比较严格,故也不适合一般的网页相似度度量;参考文献[9]、[10]分别通过统计两棵树中节点和链路出现的频率确定树结构相似度,该类方法又过于自由,只统计了节点和链路的相似性,忽略了树结构的系统性,也不适合于度量 HTML 树结构的相似性。总结以上研究方法,考虑到 HTML 网页信息的特点,本文从 HTML 树结构的角度来重新定义网页的相似度,并通过实验将该方法与传统方法进行比较,最后介绍了树结构的相似度在 Web 结构信息压缩中的应用。

本文的实验是建立在构造 HTML 树结构之上的[11],为了提高算法的效率和准确率,还需要在构造 HTML 树之前进行网页清洗,将不必要的注释信息、脚本信息等清洗掉以提高构造 HTML 树效率,在建立树结构之后需要对结构树中不必要的节点(如
等)删掉,以简化计算复杂度和提高度量方法的准确度。

2 基于子树最优自由匹配的结构相似度度量模型

每棵 HTML 树都是由根结点和子树组成的。本文定义树结构相似度的基本思想就是将树的相似度分解为多个子树的相似度,并按照一定的规则将子树的相似度转化成根树的相似度。相似度的传递是一个递归过程,先从根结点向下逐层比较节点信息,然后从叶节点向上逐层传递子树的相似度。

2.1 结构相似度度量模型

定义 2.1:

如图 1 所示,记非空的根树 $T=(N,TC)$,其中 N 是 T 的根结点,TC 是 T 的子树集合,$TC=\{T_1、T_2、T_3、\cdots、T_n\}$,则 T_A、T_B 的相似度为

$$Sim(T_A,T_B)=\begin{cases}0, & (N_A \neq N_B)\\ \dfrac{2}{n_A+n_B}, & (N_A=N_B,n_A=1 \text{ 或 } n_B=1)\\ \dfrac{2}{n_A+n_B}+\dfrac{n_A+n_B-2}{n_A+n_B}*opt(TC_A,TC_B), & (N_A=N_B,n_A>1,n_B>1)\end{cases} \quad \text{(公式 1)}$$

其中,n_A、n_B 分别表示树 T_A、T_B 中节点的个数,$opt(TC_A,TC_B)$ 是对两个子树集合 TC_A、TC_B 按照一定规则进行最大匹配后获得的相似值,在这些规则下,相似度应该满足相似关系,包括自反律:$Sim(T_A,T_A)=1$ 和对称律:$Sim(T_A,T_B)=Sim(T_B,T_A)$。

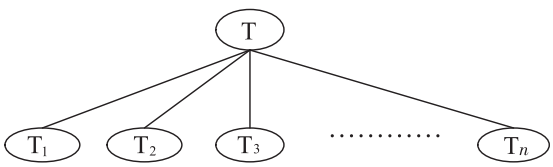

图 1 树结构

2.2 子树最优自由匹配规则

在比较网页的结构时,根据不同的需求可以定义不同的规则来计算 $opt(TC_A,TC_B)$。参考文献[8]在计算两个 HTML 树的结构相似度时,使用的是有序匹配规则,即按照子树的先后顺序进行一对一的匹配,该匹配规则限制了网页结构的比较,不能客观地度量出两个网页的结构相似程度。首先,大部分网页信息是基于半结构化数据的 HTML 语言,表达格式非常自由,对标记元素的顺序没有严格要求;其次,每棵树本身可能含有非常相似的子树,例如电子商务网页中的商品信息列表,它们都具有相同的树结构,但数目不尽相同,对于这类子树集合,如果进行一对一的匹配,不能保证每棵子树都能得到匹配,从而漏掉很多匹配信息。由此,本文提出了自由匹配的规则。

定义 2.2:

自由匹配是对两个子树集合进行匹配时,每个子树都选择对方子树集合中与其最相似的子树作为匹配对象,并将它们的相似度作为计算根树相似度的参考,在匹配过程中,每个子树可以同时作为其他多个子树的匹配对象。

算法公式为:

$$opt(TC_A, TC_B) = \frac{\sum_{i=1}^{m}\{n_{Ai} * \underset{j}{Max}[Sim(T_{Ai}, T_{Bj})]\} + \sum_{j=1}^{n}\{n_{Bj} * \underset{i}{Max}[Sim(T_{Ai}, T_{Bj})]\}}{\sum_{i=1}^{m} n_{Ai} + \sum_{j=1}^{n} n_{Bj}} \quad \text{(公式 2)}$$

2.3 计算步骤

如前所述,两棵树的结构相似度的计算是个分层递归的过程,计算两棵树的相似度首先比较根结点是否相同,如果相同,两棵树就具有可比性,然后分别计算它们子树之间的相似度,并通过子树最优自由匹配规则对两个子树集合进行最优匹配,最后加权求和递归到根树之间的相似度计算。

具体步骤如下:

(1)比较根结点是否相同,如果不同,两个根树的相似度为零,停止计算;否则进入第二步;

(2)统计两个根树的节点数目 n_A 、n_B ,如果 $n_A = 1$ 或 $n_B = 1$,则两棵根树的相似度为 $\frac{2}{n_A + n_B}$,否则分别计算两个子树集合中每个子树之间的相似度,进入第三步;

(3)针对每个子树,从另一个子树集合中选取与之相似度最大的子树作为匹配对象,并将该相似度值作为计算根树相似度的参考值 ref_i ,即 $ref_i = \max_j\{Sim(T_{Ai}, T_{Bj})\}$, $ref_j = \max_i\{Sim(T_{Ai}, T_{Bj})\}$;

(4)以子树的节点数作为权重,计算所有子树总的参考值 $opt(TC_A, TC_B)$,即 $opt(TC_A, TC_B) = \frac{\sum_{i=1}^{m}\{n_{Ai} * ref_i\} + \sum_{j=1}^{n}\{n_{Bj} * ref_j\}}{\sum_{i=1}^{m} n_{Ai} + \sum_{j=1}^{n} n_{Bj}} = \frac{\sum_{i=1}^{m}\{n_{Ai} * ref_i\} + \sum_{j=1}^{n}\{n_{Bj} * ref_j\}}{n_A + n_B - 2}$;

(5)返回两棵根树的相似度,即 $Sim(T_A, T_B) = \frac{2}{n_A + n_B} + \frac{n_A + n_B - 2}{n_A + n_B} * opt(TC_A, TC_B) = \frac{\sum_{i=1}^{m}\{n_{Ai} * ref_i\} + \sum_{j=1}^{n}\{n_{Bj} * ref_j\} + 2}{n_A + n_B}$ 。

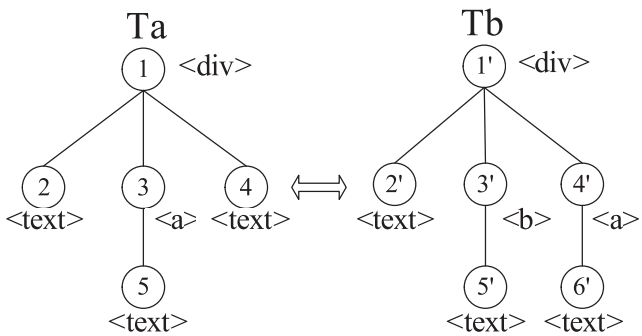

图2 基于子树最优自由匹配的结构相似度计算

以图2中的两个HTML网页片段对应的树结构 T_A 、T_B 为例,两树的根结点相同,而且都有子树。很容易计算出它们子树之间的相似度,并统计出每个子树的节点数,结果如表1所示。然后根据第三步中的公式,计算出每个子树对应的参考值ref。最后按照第5步中的公式计算出两棵根树的结构相似度

$$Sim(T_a, T_b) = \frac{1*1 + 2*1 + 1*1 + 1*1 + 2*0 + 2*1 + 2}{5 + 6} = \frac{9}{11}。$$

表1 子树之间的相似度矩阵

Similarities	2'	3'	4'	ref	Node num
2	1	0	0	1	1
3	0	0	1	1	2
4	1	0	0	1	1
ref	1	0	1		
Node num	1	2	2		

3 实验与分析

基于子树最优有序匹配规则的相似度度量方法与本文提出的基于子树最优自由匹配规则的相似度度量方法都是分层递归地进行计算,都能体现 HTML 树结构的层次性和系统性,本文在实验中主要针对这两类方法分别从计算结果和计算复杂度上做了比较分析。计算过程均由 Python 编程语言实现。

3.1 相似度计算结果比较

选取电子商务网站阿里巴巴的主页 A 及其网站内部两个相近的搜索页面 B、C 做试验。其中:

A 的网址为:http://www.alibaba.com;

B 的网址为:http://www.alibaba.com/trade/search?SearchText=computer;

C 的网址为:http://www.alibaba.com/trade/search?SearchText=telephone。

在计算网页 A、B、C 的树结构相似度之前,首先提取网页的 HTML 树结构,记录他们的节点个数分别为:A,3229;B,3643;C,3587。

然后,计算三个树的结构相似度,并统计计算时间,结果如图 3 所示。

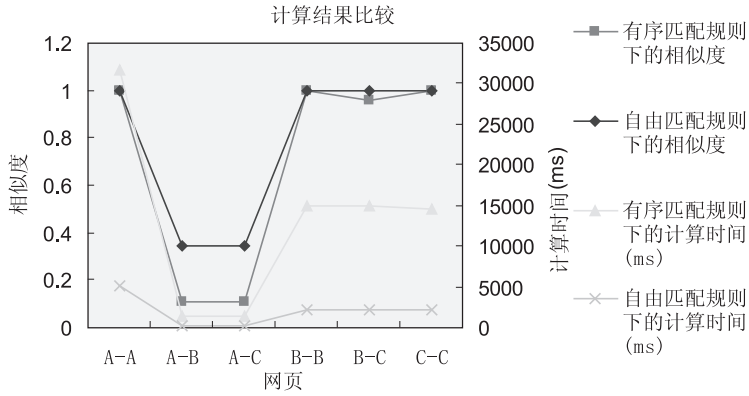

图 3 相似度计算结果比较

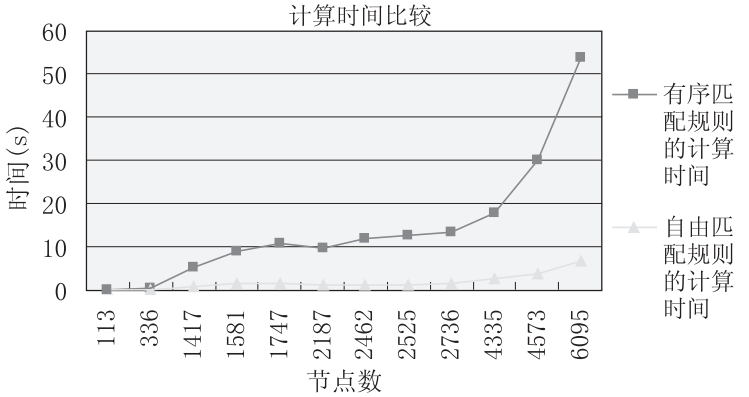

图 4 相似度计算时间比较

从图 3 可以看出:

(1)按照不同的规则进行匹配,特别是在两个网页有较小相似性时,两棵树的相似度会出现很大的差异,并且有序匹配规则下的结构相似度小于等于自由匹配规则下的相似度,这是因为在匹配过程中,有序匹配限制了匹配顺序和匹配的可重复性,导致匹配信息丢失。

(2)使用相同的匹配规则,一般在比较结构相似度大的树时花费的计算时间比较大,特别是在计算树与自身的相似度时,这是因为相似度大就意味着要计算和匹配更多的子树,从而需要更多的计算时间。

3.2 完全相似情况下的计算时间比较

从图 3 可以看出,在计算同样的两个网页的结构相似度时,有序匹配比自由匹配花费更多的时间。为了排除相似度大小对计算时间的影响而更清晰比

较两种算法在计算复杂度上的差异,本文对大小不同的网页进行了自身相似度的计算,并统计出在完全相似情况下不同节点数目的HTML树结构相似度的时间复杂度。实验结果如图4所示。

从图4可以看出,在完全相似情况下,HTML树结构相似度的时间消耗随着节点数目的增多而增加,同时,有序匹配规则下的相似度计算时间消耗比自由规则大,而且这种差距随着节点数目的增多而增加。这是因为虽然两种匹配算法都是将子树相似度矩阵扫描一遍,具有O(N*N)的复杂度,但自由匹配只需找到矩阵每行和每列的最大值,而有序匹配在扫描每个矩阵元素时都要与前面的多个元素进行比较。

4　利用相似度提取商品信息

电子商务网页上的商品信息跟其他的列表信息具有不同的特征:商品信息含有大量的文本信息和链接信息;商品信息的存储树结构比一般的列表信息复杂;商品列表元素数目繁多。所以只通过文本数据进行匹配很难精确地提取商品信息。然而我们可以利用树结构的相似性很容易的提取这些信息。

利用树结构的相似性提取列表信息就是选取具有代表性的商品信息的HTML源码,将其解析成关键树,通过对关键树与网页HTML树结构的每个子树进行依次比较,将相似度满足一定阈值的子树摘取下来,则该子树就是一个商品信息块。图5就是根据阿里巴巴网站的商品列表信息构建的一个关键树。

在提取网页信息的过程中,本文采用深度优先的访问树算法,并假定每个树结构都不与它的子树相似,从而当遇到一个与关键树匹配的树结构时就不再继续比较它的子树。具体算法如图6所示。

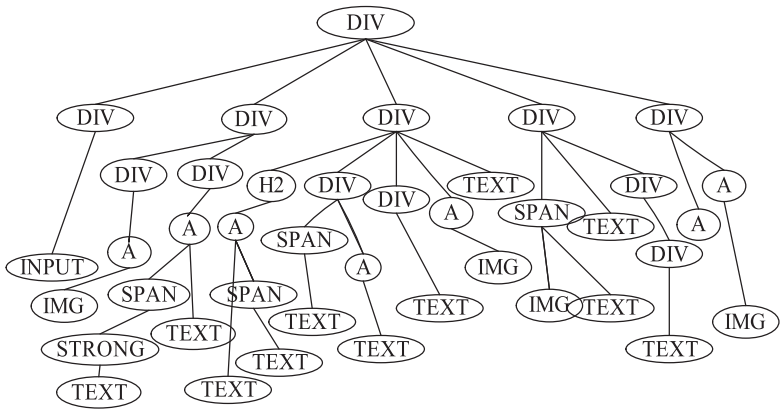

图5　阿里巴巴网站上商品列表信息的关键树

```
Algorithm: Extract(KTrootnode, WTrootnode, Result)//KTrootnode, WTrootnode分
                //别是关键树KeyTree和网页树WebTree的根结点,Result为提取结果
1  if KTrootnode and WHTrootnode have the same tag name then
2      S←similarity(KTrootnode, WTrootnode)        //计算两棵树结构的相似度
3      if S>ST then                                //ST为相似度阈值
4          Result ← Result ∪ {WebTree}             //提取网页信息块
5      end-if
6      return
7  else for each subtreenode in WTrootnode.children    //依次访问孩子结点
8          Extract(KTrootnode, subtreenode, Result)
9      end-for
10 end-if
```

图6　提取商品信息树结构的算法

为验证利用相似度度量提取商品信息的方法的有效性,本文使用 python 编程,针对几个有代表性的电子商务网站做了信息提取实验。本实验设定相似度阈值为 0.8,实验结果如表 2 所示。

表 2　提取商品信息实验结果

实验网页来源	淘宝	阿里巴巴	拍拍网	卓越亚马逊
实验网页中商品数量	42	50	40	30
成功提取的商品数量	40	48	40	27
成功率	95.24%	96.00%	100.00%	90.00%

5　总　结

本文按照子树最优自由匹配的规则定义了两棵树的结构相似度,该方法具有较好的系统性。实验证明,该方法比已有的基于有序匹配的度量方法更简便,更适用于 HTML 网页树结构的相似度度量。从第二个实验结果可以看出,计算两个比较相似的大网页的相似度时会有很大的时间开销,但由于网页树结构本身存在大量相似的子树,我们可以通过压缩冗余的子树来减少网页结构信息的存储,同时,也会降低计算时间。但如何在压缩过程中保持树结构信息的完整性还是一个问题,这也是我们下一步的工作。事实上,网页树结构的相似度度量更多的应用在于信息抽取[12~16],特别是在处理电子商务信息时,我们可以通过比较关键树和商品信息对应的子树,利用树结构的相似性提取相似的信息,实验表明,该方法有较好的准确率。

参　考　文　献

[1] Lee M, Kim Y, Lee K. Logical structure analysis: From HTML to XML[J]. Computer Standards & Interfaces, 2007,29(1):109-124.

[2] Jeh G, Widom J. SimRank: A Measure of Structural-Context Similarity [C]//Proceedings of International Conference on Knowledge Discovery and Data Mining, 2002:538- 543.

[3] Lin Z, King I, Lyu M R. PageSim: A Novel Link-based Similarity Measure for the World Wide Web [C]// Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence,2006:687-693.

[4] Dhyani D, Ng W K, Bhowmick S S. A Survey of Web Metrics[J]. ACM Computing Surveys, 2002, 34(4): 469-503.

[5] Tombros A, Ali Z. Factors Affecting Web Page Similarity [C]//Proceedings of the 27th European Conference on IR Research,2005:487-501.

[6] Tai K C. The Tree-to-Tree Correction Problem[J]. Journal of the ACM,1979,26(3):422- 433.

[7] 潘有能. XML 文档自动聚类研究[J]. 情报学报, 2006,25(2):215-220.

[8] 何昕,谢志鹏. 基于简单树匹配算法的 Web 页面结构相似性度量[J]. 计算机研究与发展,2007,44(z3): 1-6.

[9] Cruz L F, Borisov S, Marks M A, et al. Measuring Structural Similarity Among Web Documents: Preliminary Results [C]//Proceedings of the 7th International Conference on Electronic Publishing,1998: 513-524.

[10] Joshi S, Agrawal N, Krishnapuram R, et al. A Bag of Paths Model for Measuring Structural Similarity in Web Documents[C]//Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining,2003:577-582.

[11] Song M, Zhang R, Gang D. HTML Tree Parsing Algorithm Based on Pre-extracted Data [C]// Proceedings of Eighth International Conference on Mobile Business,2009:249-254.

[12] Gupta S, Kaiser G E, Neistadt D, et al. DOM based Content Extraction of HTML Documents [C]// Proceedings of the 12th international conference on World Wide Web,2003:207-214.

[13] 朱明,王庆伟. 半结构化网页中多记录信息的自动抽取方法[J]. 计算机仿真,2005,22(12):95-142.

[14] 孙承杰,关毅. 基于统计的网页正文信息抽取方法的研究[J]. 中文信息学报,2004,18(5):17-22.

[15] 陈琼,苏文健. 基于网页结构树的 Web 信息抽取方法[J]. 计算机工程,2005,31(20):54-150.

[16] 宋明秋,张瑞雪,吴新涛,等. 网页正文信息抽取新方法[J]. 大连理工大学学报,2009,49(4):594-597.

(责任编辑　马　兰)

基于HTML树的网页结构相似度研究

作者： 宋明秋， 张瑞雪， Song Mingqiu， Zhang Ruixue
作者单位： 大连理工大学系统工程研究所,大连,116023
刊名： 情报学报 ISTIC PKU CSSCI
英文刊名： JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC ANDTECHNICAL INFORMATION
年，卷(期)： 2011,30(2)

参考文献(16条)

1. 宋明秋;张瑞雪;吴新涛 网页正文信息抽取新方法[期刊论文]-大连理工大学学报 2009(04)

2. 陈琼;苏文健 基于网页结构树的Web信息抽取方法[期刊论文]-计算机工程 2005(20)

3. 孙承杰;关毅 基于统计的网页正文信息抽取方法的研究[期刊论文]-中文信息学报 2004(05)

4. 朱明;王庆伟 半结构化网页中多记录信息的自动抽取方法[期刊论文]-计算机仿真 2005(12)

5. Gupta S;Kaiser G E;Neistadt D DOM based Content Extraction of HTML Documents 2003

6. Song M;Zhang R;Gang D HTML Tree Parsing Algorithm Based on Pre-extracted Data 2009

7. Joshi S;Agrawal N;Krishnapuram R A Bag of Paths Model for Measuring Structural Similarity in Web Documents 2003

8. Cruz L F;Borisov S;Marks M A Measuring Structural Similarity Among Web Documents:Preliminary Results 1998

9. 何昕;谢志鹏 基于简单树匹配算法的Web页面结构相似性度量[期刊论文]-计算机研究与发展 2007(z3)

10. 潘有能 XML文档自动聚类研究[期刊论文]-情报学报 2006(02)

11. Tai K C The Tree-to-Tree Correction Problem[外文期刊] 1979(03)

12. Tombros A;Ali Z Factors Affecting Web Page Similarity 2005

13. Dhyani D;Ng W K;Bhowmick S S A Survey of Web Metrics[外文期刊] 2002(04)

14. Lin Z;King I;Lyu M R PageSim:A Novel Link-based Similarity Measure for the World Wide Web 2006

15. Jeh G;Widom J SimRank:A Measure of Structural-Context Similarity 2002

16. Lee M;Kim Y;Lee K Logical structure analysis:From HTML to XML[外文期刊] 2007(01)