

综合结构和内容的 XML 文档相似度计算方法

赵宁宁, 梁意文

(武汉大学 计算机学院, 湖北 武汉 430072)

摘 要: 提出了一种综合考虑 XML 文档内容和结构信息的文档相似度计算方法. 通过使用不同的方法分别计算文档内容信息相似度和结构信息相似度, 然后赋予二者不同的权重将二者综合起来, 得到文档的综合相似度. 在真实数据集上的实验结果表明, 综合结构和内容信息的方法能够提高计算 XML 文档相似度的准确性.

关键词: 内容相似度; 结构相似度; XML 相似度; 向量空间模型; 路径频率

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2016)04-0069-04

Combining Structure and Content Similaritiesmeasure for XML Document

ZHAO Ning-ning, LIANG Yi-wen

(Computer School, Wuhan University, Wuhan 430072, China)

Abstract: This paper proposed a document similarity calculation method considering the XML document content and structure information in this paper. Different methods was used to calculate the document content similarity and structural information, and different emphasis was laied on them. Then the comprehensive similarity of the document can be attained. Experimental results on real data sets show that the method integrated structure and content information can improve the accuracy of calculation of XML documents similarity.

Key words: content similarity; structure similarity; XML similarity; VSM; path frequency

1 引言

随着网络技术的发展, 网络上的信息呈爆炸式增长. 可扩展标记语言 (eXtensive Markup Language, XML) 作为一种半结构化语言, 被广泛地应用.

XML 文档可以看作由内容和结构两部分组成. 现有的 XML 文档相似度计算方法主要是基于文档结构的或者是基于文档内容的, 针对半结构的 XML 文档, 综合考虑文档结构和内容信息的相似度计算方法并不多. 只考虑文档结构或者只考虑文档内容的相似度计算方法, 在很多应用场合得不到理想的结果. 完整的 XML 相似度计算应该既考虑文档的结构信息也考虑文档的内容信息.

计算 XML 文档结构相似度时, 通常使用树形

结构、基于路径的方法和图形结构表示文档结构信息, 文档的表示方式不同, 相似度的计算方法也不同^[1]. 文献[2]使用基于树的表示方法, 利用树的“编辑距离”的方法进行相似度计算, 这种方法的缺点是复杂度较高, 文献[3]使用基于图的方式表示 XML 文档, 通过计算文档间共同出现的边点集合得到结构相似度.

关于文档内容信息挖掘经典的方法是文献[4]提出的向量空间模型 (Vector Space Model, VSM) 和文献[5]提出的特征项权重计算方法. 主成分分析法、潜在语义分析法也被广泛应用于文档内容相似度的计算中用于特征降维.

综合考虑文档内容和结构信息的相似度计算方法有, 文献[6]提出的结构链接向量模型, 将文档分为若干结构单元, 然后为结构单元分别构造特征向

量. 文献[7]使用位图索引技术将文档表示为三维矩阵, 将路径和文本信息融合在特征项里. 这两种方法虽然综合考虑了文档结构和内容信息, 但是均是使用高维向量表示文档信息, 复杂度较高.

本文采用综合比较法, 分别从文档内容和文档结构两个侧面得到相似度值, 再赋予二者不同的权重将二者综合起来, 作为整个 XML 文档的相似度. 既综合考虑了内容和结构信息, 又避免使用高维向量表示文档, 有效降低了计算复杂度. 通过实验证明, 该方法可以有效的计算 XML 文档的相似度.

2 文档表示模型

2.1 文档内容表示模型

文本从本质上说是人类自然语言的片段, 而对于计算机来说则是由众多字符型构成的非结构化信息. 要计算文本间的相似度, 需通过数学建模将文本表示成计算机能够处理的形式. 向量空间模型是目前最为成熟和应用最为广泛的文本表示模型之一.

向量空间模型的基本概念如下所示.

(1) 文档(Document): 一般是指一个完整的文本或者是文本的部分片段. 在这里, 文档即是文本, 包含 XML 文档的全部文本信息.

(2) 特征项(Term): 是指在文档内容中抽取的一些能够代表全文内容的基本单位. 这样文档 D 就可以用这些特征项的集合来表示, 即 $D(t_1, t_2, t_3 \cdots t_k \cdots t_n)$, 其中 t_k 表示第 k 个特征项.

(3) 特征项权重(Term Weight): 对于文档的各个特征项, 为其赋予一定的数值, 表示特征项在文档中的重要程度, 这些数值即是特征项对应的权重. 这样文档 D 就表示为 $D(t_1, w_1; t_2, w_2; t_3, w_3 \cdots t_k, w_k \cdots t_n, w_n)$, 可简化表示为 $D(w_1, w_2, w_3 \cdots w_k \cdots w_n)$, 其中 w_k 表示第 k 个特征项的权重.

(4) 向量空间模型(VSM): 结合特征项和特征项的权重, 文档 D 就表示为 $D(t_1, w_1; t_2, w_2; t_3, w_3 \cdots t_k, w_k \cdots t_n, w_n)$, 这时可以把 $t_1, t_2, t_3 \cdots t_k \cdots t_n$ 看成一个 n 维坐标系中的坐标轴, $w_1, w_2, w_3 \cdots w_k \cdots w_n$ 是对应的坐标值, 那么由 $t_1, t_2, t_3 \cdots t_k \cdots t_n$ 分解得到的正交矢量组就构成了一个文档内容向量空间, 文档则被映射成为其中的一个点, 因此文档 $D(w_1, w_2, w_3 \cdots w_k \cdots w_n)$ 可以看成是 n 维空间中的一个向量, $D(w_1, w_2, w_3 \cdots w_k \cdots w_n)$ 就称为文档 D 的向量表示或者向量空间模型.

(5) 内容相似度(Content Similarity): 用来表示两个文档内容之间的相似程度, 本文采用余弦系数

法来计算文档内容的相似度.

向量空间模型用特征项及其权重来表示文档内容, 把文档内容之间的相似度比较转化为向量之间的运算, 很大程度上降低了问题的复杂性. 由以上概念可以看出, 特征项的提取和权重的计算是运用向量空间模型进行文本相似度计算的关键.

2.2 文档结构表示模型

XML 文档的结构与文档内容的组织方式有关. 很多信息比如节点名称、节点属性、节点间的关系等都可以用来计算文档的结构相似度. 为了简化计算, 仅考虑最重要的节点名称信息. 使用文档结构树表示文档的结构信息, 而文档结构树由不同的节点路径组成. 给定文档集 $D\{d_1, d_2, \cdots, d_n\}$, 可以从中提取路径集合 $P\{p_1, p_2, \cdots, p_f\}$. 几个相关概念介绍如下所示.

(1) 路径(Path): 路径 p_i 由从根节点到叶子节点的节点名称组成. 其中叶子节点是指包含文本信息的节点.

(2) 结构模型(Structure Modeling): 将路径 $p_1, p_2, \cdots, p_f$ 看成一个 f 维坐标系中的坐标轴, 文档 d_i 的结构信息可以表示为路径向量空间里的一个向量 $\{p_{i,1}, p_{i,2}, \cdots, p_{i,f}\}$, 对应的坐标值为路径在文档 d_i 中的出现次数.

(3) 结构相似度(Structure Similarity): 用来表示两个文档结构之间的相似程度, 在这里用欧氏距离来度量文档结构的相似度.

3 文档相似度的计算

3.1 文档内容相似度

文档内容相似度的计算采用传统的向量空间模型, 首先进行预处理, 从 XML 文档的叶子节点提取文档的内容信息, 过滤掉停用词后提取特征项, 然后计算特征项的权重进而得到文档内容的向量, 最后通过计算向量间的夹角得到文档内容相似度的值. 具体步骤如下所示.

3.1.1 特征项的提取

根据向量空间模型, 要计算文本的相似度, 首先要对文本内容进行特征提取. 文本的特征项应该具备两个特征: 一是彻底性, 指它能够覆盖文本主题的程度; 二是专门性, 指它能反映文本的独特具体内容. 由于词语作为特征项在文档中出现的频率较高, 呈现一定的统计规律, 所以向量空间模型主要采用词语作为文档的特征项.

选取词语作为文档的特征项, 需要将指定的文

本序列切分成有独立意义的单个词语. 无论是英文文本还是中文文本, 都包含很多虚词以及在很多文章里都出现的高频词, 对文档相似度计算来说没有意义. 因此, 在提取特征项时, 需要根据通用的中英文停用词表(stop word list)去除停用词.

3. 1. 2 特征项权重的计算

特征项的权重直接影响着文本相似度的计算结果, 权重越大表示特征项在文档中越重要. 本文采用文献[5]提出的 TF-IDF (Term Frequency Inverse Document Frequency) 方法计算特征项的权重. TF-IDF 方法是目前使用最广泛的权重计算方法, 它是一种统计方法, 主要包含以下两个部分.

(1)词频(Term Frequency):某个特征项在文本中出现的频率,反映该特征项在文本中的重要程度.

(2)反文档频率(Inverse Document Frequency):反映某个特征项在整个文本集中的重要程度,特征项在所有文本中出现的频率越高,表示该特征项的区分能力越差.

TF-IDF 方法的权重计算公式如下:

$$\omega_k = \text{tf}_k \times \text{idf}_k$$

(1)

用 N 表示文本集中所有文本数, n 表示包含第 k 的特征项的文本数, 并对上式进行归一化处理, 消除文本长度不同对权重的影响, 最后得出权重的计算公式为

$$\omega_k = \frac{\text{tf}_k \times \log\left(\frac{N}{n}\right)}{\sqrt{\sum_{k=1}^n \text{tf}_k^2} \times \left[\log\left(\frac{N}{n}\right)\right]^2}$$

(2)

3. 1. 3 文档内容相似度的计算

本文采用余弦系数法计算文档内容相似度, 余弦系数法实际上是利用向量长度对向量内积法进行规范化后的结果. 假设有两文档 A 和 B , 二者的文本向量分别表示为文本 $\mathbf{A}(a_1, a_2, \dots, a_k \dots a_n)$ 和文本 $\mathbf{B}(b_1, b_2, \dots, b_k \dots b_n)$, a_k 和 b_k 分别表示两个文本中对应特征项的权重. 根据余弦系数法, 两个文本之间的文本相似度计算公式为

$$\text{contSim}(A, B) = \frac{\sum_{k=1}^n a_k \times b_k}{\sqrt{\sum_{k=1}^n a_k^2} \times \sqrt{\sum_{k=1}^n b_k^2}}$$

(3)

3. 2 文档结构相似度

文档结构相似度的计算采用基于路径频率的向量化方法, 步骤如下:

3. 2. 1 构建路径频率向量

要构建路径频率向量, 需先根据文档集 D 中的文档结构, 构建对应的文档结构树, 然后提取路径集

合 P . 然后计算每一条路径在文档 d_k 中出现的频率, 举例说明如下:

现有一文档集 D , 包含三个 XML 文档 $\{d_1, d_2, d_3\}$, 文档结构树如图 1 所示. 其中, R 表示文档根节点的节点名称, E_i 表示内部节点或叶子节点的节点名称, 尖括号里的 t_i 表示文档的内容. 各文档的结构信息用矩阵表示如表 1 所示.

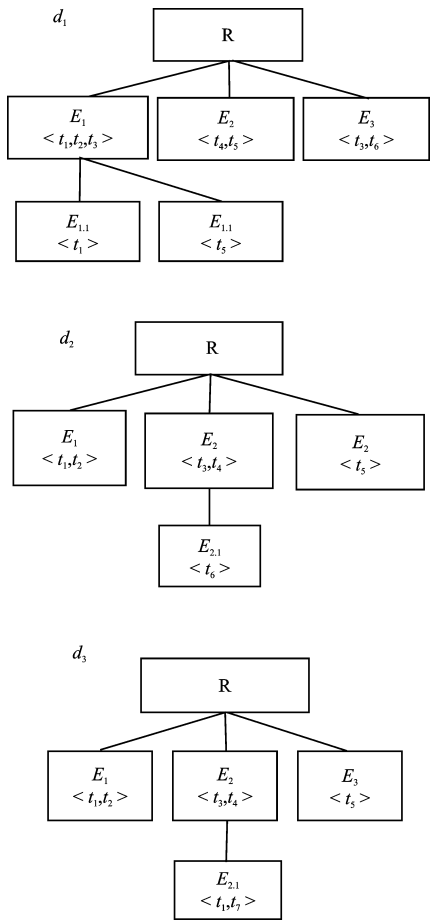


图 1 文档集 D 中的文档结构树

表 1 文档集 D 的文档结构矩阵

路径/文档	d_1	d_2	d_3
R/ E_1	1	1	1
R/ E_1 / $E_{1.1}$	2	0	0
R/ E_2	1	2	1
R/ E_2 / $E_{2.1}$	0	1	1
R/ E_3	1	0	1

3. 2. 2 文档结构相似度的计算

根据文档结构矩阵, 可以得出各个文档的结构向量, 本文采用欧氏距离公式计算文档的结构相似度. 假设有两文档 A 和 B , 二者的结构向量分别表示为: 文档 $\mathbf{A}(p_{a_1}, p_{a_2}, \dots, p_{a_f})$ 和文档 $\mathbf{B}(p_{b_1}, p_{b_2}, \dots,$

p_{b_j}), p_{a_i} 和 p_{b_i} 分别表示两个文档中对应路径出现的次数. 根据欧氏距离公式, 两文档的距离相似度计算公式为

$$\text{structSim}(A, B) = \sqrt{\sum_{i=1}^f (p_{a_i} - p_{b_i})^2} \quad (4)$$

3.2.3 文档相似度计算

本文采用综合比较法, 分别从文档内容和文档结构两个侧面计算文档相似度, 再赋予二者不同的权重, 最终得到文档的综合相似度值. 计算公式如下:

$$\begin{aligned} \text{docSim}(A, B) = & \text{contSim}(A, B) \times \varphi + \\ & \text{structSim}(A, B) \times (1 - \varphi) \end{aligned} \quad (5)$$

式中, φ 取值范围为 0 到 1, 表示文档内容信息在比较文档相似度时占的比重. φ 值越大, 表示在比较相似度时, 文档内容信息越重要, 结构信息越不重要. φ 的值可以根据需求动态调整. 对于不同的数据集, 不同的应用, φ 的值也不尽相同. 对于按照文档结构分类属于同一类型的文档集, 文档结构类似, 因此文档结构对分类结果影响较小, 因此在计算综合相似度时应该赋予结构相似度较小的权重; 对于按照文档结构分类不属于同一类型的文档集, 文档结构和内容都有差异, 在计算综合相似度时, 二者都不应该被忽略.

4 实验结果及分析

4.1 实验设计

为了检验本文提出的综合法的有效性, 与传统的基于内容和基于结构的相似度计算方法进行比较实验. 本文实验采用的数据集为生活中的真实数据集^[8], 由 101 个来自 20 个网站的 XML 文档组成. 分为 automobile, movie, reference, software 4 类, 其中 20 个属于 automobile 类, 20 个属于 movie 类, 41 个属于 reference, 20 个属于 software.

实验的评价标准采用信息检索系统中使用的准确率 and 召回率思想, 计算二者的加权平均得到 F_1 分数, 通过评价文档分类的正确性来反应文档相似度计算方法的有效性. 与信息检索系统一样, 将所有的文档分为 4 类, A: 计算所得相似度大于零, 属于同类的文档; B: 计算所得相似度大于零, 属于不同类文档; C: 计算所得相似度为零, 属于同类的文档; D: 计算所得相似度为零, 属于不同类的文档. 因此定义准确率为 $\text{precision} = A / (A + B)$, 定义召回率为 $\text{recall} = A / (A + C)$, F_1 分数的计算公式如下:

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (6)$$

在进行实验时, 分别从 automobile, movie, reference, software 4 类中随机选取一个文档作为比较的基准, 与所有样本进行相似度计算. 为了得到此数据集最优的内容权重值 φ , 将数据集中剩下的文档分为训练集和测试集两部分进行实验. 首先对于训练集, 从 0 到 1 逐步调整内容权重值 φ 的大小, 计算 F_1 的平均值, 实验结果如图 2 所示.

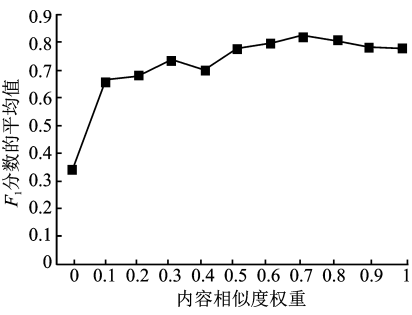


图 2 训练集实验结果

由图 2 所示的结果可知, 对于本文使用的数据集, 内容权重值 φ 取 0.7 比较合适. 然后对测试集进行实验, φ 的值固定取 0.7, 分类结果如表 2 所示.

表 2 实验结果

方法	F_1				平均值
	automobile	movie	reference	software	
本文方法	100.00	100.00	97.20	95.27	98.12
VSM	100.00	100.00	95.28	93.14	97.11
path	85.13	87.26	89.64	91.47	88.38

4.2 实验结果与分析

表 2 中, VSM 那行是使用传统的向量空间模型和 TFIDF 权重计算得到的 F_1 结果, 仅仅考虑文档的内容信息, 忽略了文档的结构信息. path 那行是使用基于路径频率的文档结构相似度计算方法得到的 F_1 结果, 仅仅考虑文档的结构信息. 实验结果表明, 与仅考虑文档内容或结构的方法相比, 本文采用的综合方法准确率更高.

5 结束语

本文提出了一种综合考虑文档内容信息和结构信息的 XML 文档相似度计算方法. 分别从内容和结构两个方面计算文档相似度, 然后赋予二者不同的比重将二者结合起来. 实验结果表明, 综合考虑文档结构和内容信息, 可以提高相似度计算的准确度. 本文综合考虑了文档的内容和结构信息, 但是忽略了文档的语义信息, 这将是下一步需要改进的重点.

(下转第 76 页)

以正常启动,而传统方案就会因数据信息不全而无法正确启动。

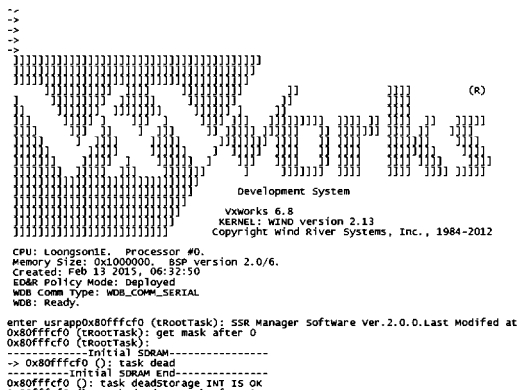


图7 方案测试图

6 结束语

本文针对 NAND FLASH 工作原理、坏块信息存储方式的特点进行分析,提出了一种高可靠性的坏块表存储方案设计。将坏块信息进行独立、冗余存储,提高坏块表的可靠性,进而提高整个 NAND FLASH 存储器的可靠性。经实验可知,该算法在坏块信息区出现使用坏块,信息存储失败时,能够自主调整坏块信息的存储位置,重新存储坏块信息,并不影响下次启动。该算法也会在系统故障掉电时,检测出坏块信息区的冗余备份是否都完成更新,并检测

出未更新的备份,屏蔽其内容,通过已更新的信息表重新启动。

该算法能够有效避免坏块信息的存储,更新出现的风险,提高 NAND FLASH 存储系统的可靠性。

参考文献:

- [1] 高怡祯. 基于闪存的星载大容量存储器的研制[D]. 北京:中国科学院,2004.
- [2] 彭军. 基于 NAND Flash 的多路并行存储系统的研究与实现[D]. 长沙:湖南大学,2013.
- [3] 贾源泉. 基于 NAND FLASH 的多路并行存储系统的研究与实现[D]. 长沙:国防科学技术大学,2011.
- [4] 盛江坤,邱孟通,姚志斌,等. NAND 型 Flash 存储器总剂量效应实验研究[J]. 原子能科学技术,2014, 48 (8): 1502-1507.
- [5] 宁飞. 大容量存储中 NAND Flash 坏块的管理方法[J]. 电子测试,2010(12):64-68.
- [6] 王永年. n 中取 k 表决系统的可靠性评估[J]. 电子质量,2009,30(10):37-38.

作者简介:

李 珊 女,(1989-),博士,助理研究员.研究方向为星载存储技术. E-mail:orchidmg@126.com.

(上接第 72 页)

参考文献:

- [1] Algergawy A, Mesiti M, Nayak R, et al. XML data clustering: An overview[J]. ACM Computing Surveys (CSUR), 2011, 43(4): 25.
- [2] Brzezinski D, Piernik M. Adaptive XML stream classification using partial tree-edit distance[M]//Foundations of Intelligent Systems. Berlin:Springer International Publishing, 2014: 10-19.
- [3] Hossain M S, Angryk R A. Gdclust: A graph-based document clustering technique [C] // Data Mining Workshops, 2007. Omaha: IEEE, 2007: 417-422.
- [4] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.
- [5] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval[J]. Information processing & management, 1988, 24(5): 513-523.

- [6] Yang J, Cheung W K, Chen X. Learning element similarity matrix for semi-structured document analysis [J]. Knowledge and Information Systems, 2009, 19 (1): 53-78.
- [7] Yoon J P, Raghavan V, Chakiram V. BitCube: A three-dimensional bitmap indexing for XML documents [C]//Scientific and Statistical Database Management, 2001. Fairfax: IEEE, 2001: 158-167.
- [8] Kurt A, Tozal E. Classification of xslt-generated web documents with support vector machines [M] // Knowledge Discovery from XML Documents. Berlin , Heidelberg:Springer, 2006: 33-42.

作者简介:

赵宁宁 女,(1992-),硕士研究生.研究方向为信息检索、人工免疫学. E-mail:349177535@qq.com.

梁意文 男,(1962-),教授.研究方向为人工免疫学.