

Web 多媒体数据源质量模型

邓 松^{1,2}, 王映龙^{1,2}, 何火娇^{1,2}, 余 鹰¹

(1 江西农业大学 软件学院, 江西 南昌 330045;

2 江西省高等学校农业信息技术重点实验室, 江西 南昌 330045)

摘 要: 文中分析了多媒体数据源质量评估的多个关键因素, 提出了一种针对于多媒体数据源的质量估计模型。模型从合时性、资源数量、使用数量、信息丰富度等四个维度对数据源质量进行了质量评估, 并对源质量计算中的规范化方法进行了改进。实验证明, 利用该模型对数据源质量进行估计具有很好的准确性和可操作性。

关键词: 数据源; 质量维度; 遗传算法; 质量模型

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2011)11-0157-03

Quality Estimation Model of Multimedia Data Source of Web

DENG Song^{1,2}, WANG Ying-long^{1,2}, HE Huo-jiao^{1,2}, YU Ying¹

(1 Department of Software, Jiangxi Agriculture University, Nanchang 330045, China;

2 Key Laboratory of Agricultural Information Technology of Jiangxi College, Nanchang 330045, China)

Abstract: We analyse the key factors of quality assessment of web multimedia data sources and propose a new model to estimate the quality of web multimedia data sources. The model estimate the quality of data source from four dimensions: Timeliness, Resources Quantity, Use Quantity, Content Richness and improve the standardized method. The experiment result shows that the quality model is of good veracity and operability to evaluate the quality of multimedia web data sources.

Key words: data source; quality dimensions; genetic algorithm; quality model

1 引言

随着互联网的发展, 如何在 Web 中找到用户所需要的高质量的查询结果引起人们的关注。依据查询关键字获取与查询最相关的数据源研究, 目前已经有了较为高效的选择算法。而在数据源选择过程中考虑源本身质量的研究, 目前还是处于初步阶段。不同领域的数据源有着不同的质量维度, 难以用一个统一的维度指标衡量所有的数据源。因此如何根据数据源类型的特点选取最核心、最适合的质量维度是一个关键问题。文中针对多媒体的数据源(例如软件资源类数据源、音乐、影视类数据源)的共同特点, 选择了一组相应的质量维度, 从合时性、信息丰

富度等多个质量指标对多媒体数据源的质量进行了评估。

2 相关研究

在过去近十年间, 出现了较多基于查询相关的数据源选择高效算法, 比较经典的算法有 gGLOSS/vGLOSS^[1]、CORI^[2]等, 然而对数据源质量的评估研究还没有受到足够多的重视。目前, 基于数据质量的数据源选择方案主要是首先选择合适的数据源质量评价指标, 然后建立相关的质量评价模型, 最后依据评价打分, 进行数据源的选取。现有的考虑数据源自身质量的源选择方法主要有以下两种: ①考虑在集成环境下的基于质量的数据源选取方法; ②独立的

基于质量的源选取方法.

Web 数据集成环境下的数据源选取,法国学者 Abounaga 做了较为深入的研究. 2007 年, Abounaga 研究了在 Web 数据集成环境下利用数据源质量评价模型进行数据源的选择问题^[3]. 国内,苏州大学鲜学丰等人在 2010 年提出了一种在集成环境下以迭代方式进行 Web 数据源选取和集成的方法^[4]. 该数据源选择方案的核心在于质量增益函数,为了做到机器自动评价,文中的质量模型选取了一致性、完整性、冗余性等三个客观的质量维度. 独立的数据源评价方案,国内学者研究相对较多. 2008 年苏州大学鲜学丰等人分析了影响 Deep Web 数据源质量评估的若干因素,提出了一种 Deep Web 数据源质量评估模型^[5]. 而后,苏州大学胡鹏昱等人又提出了一种扩展的 Deep Web 质量估计模型^[6]. 在 2009 年苏州大学胡鹏昱等人综合考虑 Deep Web 接口查询能力、接口页面质量、服务质量三大方面因素,也提出一种深网数据源质量估计模型^[7].

不同类型的数据源,具有不同的核心质量维度. 以上的基于质量的源评价模型都未考虑到多媒体资源型数据源的特点,目前还没有研究者分析过影响评估多媒体类型数据源质量的因素和评估该类型源质量的方法. 文中分析了多媒体数据源质量评价的多个因素,提出了一种新的数据源质量评估模型,并采用遗传算法进行了实验验证.

3 Web 多媒体数据源质量的维度分析

多媒体数据源包含复杂格式的压缩下载文件,视频,音乐等 Web 信息资源,因此影响评估多媒体数据源的主要维度与普通的文本型数据源有较大的不同. 考虑到机器自动评价的必要性,文中主要选择一些能够客观度量的源质量维度用于数据源评价. 以软件资源型多媒体数据源为例,对多媒体 Web 数据源核心页面的特点进行分析,如图 1 所示.

(1) 资源数量 Size(D): 用户作为数据源的选择者,在进行多媒体数据源选择的时候,通常会挑选那些拥有丰富数据资源的 Web 数据源. 多媒体数据源都建有分类或者英文字母查询索引,利用网络爬虫可以较容易的获取某分类索引中资源的数量. 因此,Size(D)可以用下式表示:

$$\text{Size}(D) = \frac{N}{n} \cdot \sum_{i=1}^n \text{count}(C_i) \quad (1)$$

其中, N 表示多媒体数据源中可下载文件、音频或视频种类数量, n 为多媒体数据源中实验抽样类型的



图 1 多媒体 Web 数据源核心页面
数量. $\text{count}(C_i)$ 代表某类型下资源的数量.

(2) 信息丰富度 Richness (D): 资源相关的信息包括了软件大小、软件性质、更新时间、开发商、软件类别、应用平台等,多媒体资源相关信息越丰富,用户越能做出准确的选择. 信息丰富度可用下式计算: $\text{Richness}(D) = \text{Fields}(\text{corepage})$, 也就是核心页面(下载或者播放页面)中与资源相关字段数量.

(3) 合时性 Timeliness(D): 用户选择一个多媒体资源的时候,倾向于选择那些常更新的数据源进行下载或者播放音视频文件. 合时性可以用下式表示:

$$\text{Timeliness}(D) = \frac{N_{\text{fresh}}}{12} \cdot \frac{365 - 3.65 \times \text{day}}{365} \quad (\text{day} \leq 100) \quad (2)$$

其中, N_{fresh} 表示在最后一次更新前的一年中资源有过更新的月份的数量, day 表示最后一次更新到当前的间隔天数,当该值大于等于 100 时, Timeliness 直接置为 0.01.

(4) 使用数量 Use(D): 多媒体数据源中的下载文件,音频或视频质量难以直接度量,但是可以通过下载次数或播放次数间接度量. 使用数量可用下式计算:

$$\text{Use}(D) = \frac{U_i}{n} \quad (3)$$

其中, U_i 为某资源的用户下载或者播放次数, n 为抽样资源总数量.

4 多媒体 Web 数据源质量模型

为了对多媒体数据源质量进行评估,文中提出了一种多媒体数据源质量评估模型 $Q\text{-mode} = \{D, W, R, S\}$. 其中 $D = \{d_1, d_2, \dots, d_m\}$ 代表一组数据源集合, $W = \{w_1, w_2, \dots, w_m\}$ 代表一组数据源维度集合, $R = \{r_1, r_2, \dots, r_m\}$ 代表一组数据源权重

集合, $S = \{s_1, s_2, \dots, s_m\}$ 为数据源得分集合. 利用维度乘以相应的权重计算数据源得分. 有了维度计算公式与数据之后, 就可以对数据源进行评分. 但还要解决以下二个问题: ① 维度标准的范围不一致; ② 各个维度的重要性不一样.

对于第一个问题, 文中采用质量维度标准化处理的策略. 标准化也就是要让不同维度的取值范围分布在同一区间, 如 $(0.0, 1.0)$ 区间, 使它们具有可比性. 可以采用最小-最大规范化方法.

$$v' = \frac{v - \min_A}{\max_A - \min_A} \quad (4)$$

其中, $\max_A$ 代表属性 A 的最大值, $\min_A$ 代表属性 A 中的最小值. 研究表明, 有的数据源资源数量可达几十万, 而有的数据源资源数仅为几千, 有的平均下载次数高达数万, 有的数据源平均下载次数仅仅 0.99 次, 在这种情况下, 直接采用该规范化方法, 某维度上相对较小的一些数据值将没有区分度. 因此, 本文针对多媒体数据源的特点提出了改进的最小-最大规范化方法, 该方法分为两步: ① 将数值悬殊的维度上的数据值按照数据长度 + 小数点 + 原数值的方式进行编码, 例如 1234566 编号后的结果为 7.1234566; ② 将编码后的数据值按照普通的最小-最大规范化处理.

对于权重问题, 文中采用改进的遗传算法寻找维度的最优权重分布. 改进遗传算法使用实数编码, 因此便于数值优化. 选择适应度函数的时候, 按照一组维度可能的权重分布, 对样本集中已经过专家人工评定的数据源进行评估, 用 Kendall's τ 距离来衡量评估效果.

5 实验结果分析

为了验证提出的多媒体数据源质量模型的有效性, 手工选取了 52 个软件资源下载类的多媒体数据源进行了实验, 其中软件资源类数据源具有与音频、视频数据源相同的核心维度. 首先对 25 个数据源进行专家人工排序, 然后采用遗传算法根据训练数据获取最优的质量模型维度的权重分布, 剩余 27 个数据源作为测试集用于测试效果. 该实验分别采用了普通的和改进的最小最大规范化方法处理原始数据, 并比较了在这两种情况下的源质量模型选择的效果.

用 P 表示协调对数量, Q 表示不协调对个数, 则 P_1 和 P_2 之间的 Kendall's τ 距离为: $(P - Q) / (P + Q)$. 采用 Kendall's τ 距离来衡量两序列间的相

似度, 用来作为遗传算法的适应度函数. 图 2 为在两种不同规范化情况下, 质量模型的执行性能.

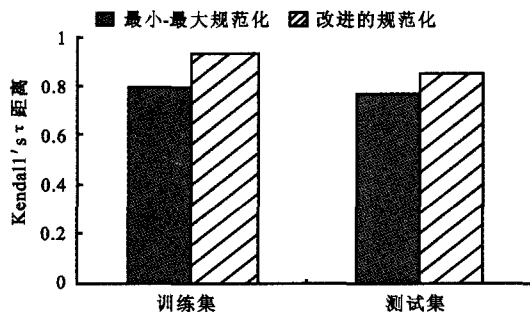


图2 两种方法下的源选择执行性能

从图2可以看出, 文中数据源质量评价模型具有较高的精度. 在两种方法下, 测试集的 Kendall's τ 距离值均大于 0.757, 改进后的值达到了 0.848.

原始数据经过最小最大规范化后再利用遗传算法获取的权重分布为 $(0.1460, 0.1349, 0.6605, 0.0586)$, 而改进后的分布权重为 $(0.0927, 0.3452, 0.4531, 0.1090)$. 改进后的训练集合 Kendall's τ 距离值达到 0.909, 因而可以较为真实的反映权重分布. 权重分布向量中间两列分别代表资源数量维度和使用维度. 由于有的多媒体数据源数量维度中的值或使用维度中的数据值相差很悬殊, 用传统的最大最小规范化来进行数据的预处理, 会使得遗传算法寻优获取的权重分布不能达到最优. 图3显示了两种方法下选择不同数量数据源的精确性, 从图中可以看出, 改进后的规范化方法可以取得更好的效果.

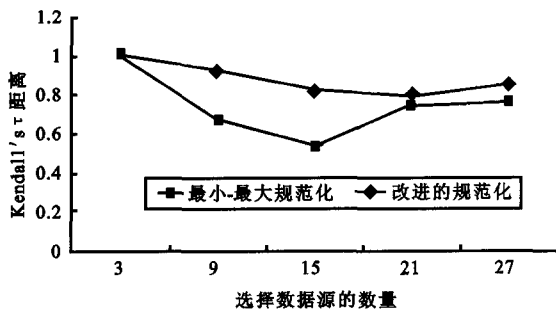


图3 选择不同数量数据源时算法的精确性

6 结束语

文中分析了影响评估多媒体数据源质量的若干维度, 提出了一种新的针对于多媒体 Web 数据源的质量评价模型. 针对于源质量维度上的数据值悬殊导致遗传算法获取不到准确的权重分布的问题, 提出了一种改进的规范化算法. 通过对真实的数据源

(下转第 164 页)

参考文献:

- [1] 杨士元,胡梅,王红. 模拟电路软故障诊断研究[J]. 微电子学与计算机, 2008, 25(1): 1-8.
- [2] 田裕鹏. 红外检测与诊断技术[M]. 北京: 化学工业出版社, 2006.
- [3] 彭敏放,沈美娥,贺建飏,等. 容差电路软故障检测与定位[J]. 电工技术学报, 2009, 24(3): 222-228.
- [4] 廖薇,许春冬,刘锦高. 基于神经网络的模拟电路故障诊断研究[J]. 微电子学与计算机, 2010, 27(5): 125-128.
- [5] 谢涛,何怡刚,姚建刚,等. 基于小波包和自组织网络的模拟电路故障诊断[J]. 微电子学, 2009, 39(2): 190-193.
- [6] 朱大奇,刘永安. 故障诊断的信息融合方法[J]. 控制与决策, 2007, 22(12): 1321-1328.
- [7] Erik Berglund, Joaquin Sitte. The parameterless self-

organizing map algorithm [J]. IEEE Transactions on neural networks, 2006, 17(2): 305-316.

- [8] 袁海英,陈光祜. 故障诊断中基于神经网络的特征提取方法研究[J]. 仪器仪表学报, 2007, 28(1): 90-94.

作者简介:

吴正苗 女, (1979-), 硕士研究生. 研究方向为集成电路测试与诊断、数字图像处理.

彭敏放 女, (1964-), 博士, 教授, 博士生导师. 研究方向为电路测试与诊断、电气设备状态监测、智能信息处理、电磁场理论及其应用.

谭虎 男, (1986-), 博士研究生. 研究方向为集成电路测试与诊断、复杂网络理论及应用.

沈美娥 女, (1966-), 硕士, 副教授. 研究方向为集成电路测试与诊断、计算机仿真与建模.

(上接第 159 页)

进行验证, 说明了该模型可以有效评估多媒体 Web 数据源的质量.

参考文献:

- [1] Gravano L, Chang C, Garcia-Molina H, et al. STAR-TS: stanford proposal for internet meta-searching[C] // Proceedings of the 20th ACM-SIGMOD International Conference on Management of Data. Tucson, Arizona: ACM, 1997: 207-218.
- [2] Souza D, Zobel J, Thom J. Is CORI Effective for Collection Selection an Exploration of parameters, queries, and data [C] // Proceedings of Australian Document Computing Symposium. Melbourne, Australia: IEEE, 2004: 41-46.
- [3] Abounaga A, ElGebaly k, Wong D. BE: user guided source selection and schema mediation for internet scale data integration[C] // Proceedings of the IEEE Interna-

tional Conference on Data Engineering. Istanbul, Turkey: IEEE, 2007: 1513-1514.

- [4] Xian XF, Zhao PP, Yang YF. Efficient selection and integration of hidden web database[J]. Journal of computers, 2010, 5(4): 500-507
- [5] 鲜学丰, 方巍, 赵朋朋, 等. 一种 Deep Web 数据源质量评估模型[J]. 微电子学与计算机, 2008, 25(10): 47-50.
- [6] 胡鹏昱, 苗忠义, 崔志明, 等. 扩展的 Deep Web 质量估计模型研究[J]. 微电子学与计算机, 2008, 25(9): 24-27.
- [7] 胡鹏昱, 赵朋朋, 方巍, 等. 深网数据源质量估计模型. 计算机工程, 2009, 35(9): 204-207.

作者简介:

邓松 男, (1982-), 博士研究生, 讲师. 研究方向为 Web 数据挖掘、知识工程.

王映龙 男, (1974-), 博士, 教授. 研究方向为数据挖掘.