基于事实抽取的 Web 文档内容数据质量评估

韩京宇 陈可佳
(南京邮电大学计算机学院 南京 210003)

摘 要 Web 文档内容数据质量评估决定获取数据的有用性。基于词法或用户交互进行质量评估的方法缺乏通用性,也不能获取内容的事实内涵。因此提出基于事实的质量评估方法(Fact-based Quality Assessment,FQA)。首先在 Web 上构建目标文档上下文,并抽取 Web 文档内容的事实;然后分别采用投票和图迭代策略,构建准确性和完整性维度的参照;最后,比对目标文档和维度参照的事实,量化准确性和完整性。该方法不依赖特定特征,基于事实内涵量化数据质量维度,可取得高的评估精度。实验结果证明了 FQA 方法的优越性。

关键词 数据质量,Web 文档,准确性,完整性,质量维度,事实
中图法分类号 TP311.13 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.11.047

Ranking Data Quality of Web Article Content by Extracting Facts

HAN Jing-yu CHEN Ke-jia
(College of Computer Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract Data quality assessment of Web article content helps identify useful data. Exiting approaches not only heavily rely on lexicon features or user interactions to obtain quality indicators, but also can not capture the content' semantics. A fact-based quality assessment (FQA) approach was proposed in this article. Given one target article, the approach starts with the identification of alternative context by collecting relevant articles and extracting facts from every article. Then, the accuracy baseline is constructed by voting, and the completeness baseline is constructed by iterations over fact graphs. Finally, data quality dimensions, including accuracy and completeness are calculated by comparing the facts of the target article with the established dimension baselines. Based on the facts of target article content, rather than particular features, FQA approach can quantify data quality dimensions with high precisions. The superior performance of FQA was verified in the experiments.

Keywords Data quality, Web article, Accuracy, Completeness, Quality dimensions, Fact

1 引言

Web 文档是 Web 数据的主要载体,其数据质量决定着获取的 Web 数据的价值。Web 文档内容的数据质量评估极具挑战性,原因在于:(1)相比结构化数据,Web 文档内容是松散的语句序列,缺少模式约束保证质量。(2)Web 上缺少有效的规范和审核机制,数据拷贝方便,低质网页容易泛滥。(3)尤其是在 Web 2.0 时代以后,每个用户都可以自由创建、修改网页,非专业水准的编辑更容易产生数据质量问题。

数据质量被公认分解成若干数据质量维度来衡量,主要包括准确性、完整性、新鲜性、一致性等[1-3]。近年,Web 数据质量评估受到了广泛关注[4-10],主要集中在对网站质量的评估和对 Web 文档内容质量的评估。其中,对 Web 文档内容质量的评估主要分成基于特征、基于用户反馈和基于编辑历史 3 类方法。但这些方法从 Web 文档词法或用户交互提取质量指标,缺少通用性,并且也不能从事实内涵角度揭示 Web 文档内容是否和现实一致。

本文提出一种基于事实抽取的数据质量量化方法:该方法将文档内容看作事实的集合,利用 Web 上主题相关数据构建两个最重要质量维度(准确性和完整性)的参照。然后,通过比对目标文档和参照来量化准确性和完整性。该方法不仅不依赖特征和用户交互,而且能取得好的评估精度。本文的主要贡献在于:(1)提出根据文档内容事实来评估数据质量,克服了基于特征或交互类方法不能从事实内涵角度量化数据质量的弊端。(2)提出利用 LDA(Latent Dirichlet Allocation)主题模型[11]来识别与目标文档主题密切相关的上下文文档。(3)提出在上下文中,分别通过投票和图迭代构建准确性和完整性参照;然后比对目标文档和参照量化质量维度。理论分析和实验表明,该方法是 Web 文档内容数据质量评估的有效方法。

2 相关工作

2.1 Web 数据质量评估

Web 网站数据质量评估集中于评估框架和具体评估方

到稿日期:2014-01-10 返修日期:2014-03-10 本文受国家自然科学基金项目(61003040,61100135),中央高校基本科研业务费专项资金项目(LGZD201324)资助。

韩京宇(1976-),男,博士,副教授,CCF 会员,主要研究方向为数据管理、知识系统,E-mail:jyhan@njupt.edu.cn;陈可佳(1980-),女,博士,副教授,CCF 会员,主要研究方向为机器学习、信息系统等。

法的研究。文献[4]提出针对 Web 网站的数据质量定义、发布和清洗框架。文献[9]提出一个根据社会网络评估数据质量的框架。文献[10]通过检测词法错误来判断各个网站的数据质量。文献[12,13]提出利用数据源拷贝关系和条件依赖等识别数据质量。

网页内容质量评估主要分成基于文本特征、基于用户和基于编辑历史 3 类方法。基于文本特征的方法提取内容中的文本特征来进行质量的评估:文献[5]综合文本、评论和网络 3 个方面特征对 Web 文档质量进行评估;文献[8]用文本长度推断 Wikipedia 文档数据质量的好坏;文献[14]利用文档长度、词性、Web 特征和可读性等,结合最大熵理论训练模型,评估 Wikipedia 文档质量。基于用户的方法根据用户和文档交互特征来评估数据质量:文献[6]根据作者信誉度来估计 Web 内容质量的好坏;文献[15]根据用户的交互模式来判定 Wikipeda 文档质量好坏。基于编辑历史的方法根据网页的编辑历史来推断数据质量的好坏。文献[7]根据网页编辑历史来识别可信的文档。文献[16]提出从编辑历史挖掘频繁子序列,提取不同的质量类别特征,进一步通过聚类和分类来推断数据质量的好坏。这些工作的局限性在于根据特征或用户反馈来提取质量线索,不具有通用性,也不能抓取内容的事实内涵。

2.2 信息抽取技术

近年,信息抽取技术获得长足进步[17-21]。这个领域专注于从文档中抽取实体或事实。其主要分为基于模式的、基于规则的和基于统计学习的 3 类方法[17-19]。这些方法针对特定模式或需求进行信息抽取,需要预先知道种子模式。近年,针对开放域的信息抽取技术受到广泛关注[20,21]。但针对 Web 文档内容抽取数据质量参照则鲜有工作。

3 FQA 方法步骤和 LDA 主题模型

基于事实的 Web 文档数据质量评估分 3 个步骤:

1)上下文生成。对每一个目标文档 P,根据其题目和关键字搜索网页;然后,根据单词的 n-gram[22] 词法相似度和 LDA 模型的主题相似度过滤不相关网页,构建上下文。

2)维度参照提取。对上下文中每一 Web 文档内容抽取其事实;然后,在上下文的事实集合中投票确定准确性参照;采取图迭代算法,确定完整性参照。

3)维度量化。通过比较目标文档事实和维度参照的事实,计算准确性和完整性。

本文利用 LDA 主题模型识别相同主题的上下文文档。该模型如图 1 所示,其假设每个文档根据如下过程产生:

1)根据参数为 α 的 Dirichlet 分布,取样生成文档主题向量 $\theta=(\theta_1,\cdots,\theta_i,\cdots,\theta_K)$,以此代表各个主题在文档中出现的概率。

2)文档中每个单词 s 根据如下过程产生。

a)根据参数为 θ 的多项式分布,选取一个隐含主题 z_n;

b)根据概率 $p(w|z_n,\beta)$,选取单词 s,其中 β 是一个矩阵,代表每个主题中各个单词出现的频率。

这里 α 和 β 是料库参数,可由变分贝叶斯(Variational Bayesian)或吉布斯抽样(Gibbs Sampling)获得[11]。

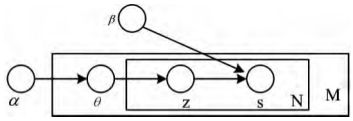

图 1 LDA 模型图示

4 构建目标文档上下文

对待评估的目标文档 P,根据题目和关键字在 Web 上搜索 PageRank 高于设定阈值的文档,以获得普遍认可的相关文档。在前 N 个相关文档中确定 P 的上下文,上下文包含的文档和目标文档论述相同的主题内容。不失一般性,如果两个文档论述相同的主题内容,其词法相似度亦会较高。因此,综合词法相似度和主题相似度识别目标文档的上下文。

定义 1 给定目标文档 P、相似度阈值 $\upsilon(0<\upsilon\leqslant 1)$ 和词法相似度权重 $\rho(0<\rho<1)$,P 的上下文是 $\Omega(P)=\{P_1,\cdots,P_i,\cdots,P_n\}$。每个文档 P_i 满足 $\rho\times sim_{lex}(P,P_i)+(1-\rho)\times sim_{topi}(P,P_i)\geqslant\upsilon$。其中 $sim_{lex}(P,P_i)$ 是 n-gram 空间中的词法相似度[22],$sim_{topi}(P,P_i)$ 是主题相似度。

为了计算主题相似度,提出在相关文档构成的料库上计算 LDA 模型的 α 和 β 后,采用算法 1 计算任意一篇文档的主题分布。算法 1 中 φ_{ik} 表示第 i 个词由主题 k 生成的概率,满足 $\sum_{k=1}^{K}\varphi_{ik}=1$。由于 φ_{ik} 和每个主题权重 Y_k 相互影响,算法 1 中采用迭代来计算文档 P 的主题分布 $(Y_1,\cdots,Y_k,\cdots,Y_K)$。

算法 1 articleTopics

输入:Dirichlet distribution parameter $\alpha=(\alpha_1,\cdots,\alpha_k,\cdots,\alpha_K)$, $\beta=(\beta_1,\cdots,\beta_k,\cdots,\beta_K)$ where β_k is a vector, article P with length N_P

输出:Topic distribution of P, denoted as Γ_P

```
foreach topic k ∈ 1,…,K do
initialize Yk = NP/K;
end
repeat
  foreach word si(1 ⩽ i ⩽ NP) do
    foreach topic k ∈ 1,…,K do
      φik ← βksi * exp(Ψ(Yk));
    end
    normalize φi;
  end
  foreach topic k ∈ 1,…,K do
    Yk ← αk + Σ_{i=1}^{NP} φik
  end
until convergence;
normalize (Y1,…,YK);
return ΓP ← (Y1,…,YK);
```

根据经验,多模型会取得更好的学习效果[23],设定 $K=12,24,48,96$。目标文档和相关文档的主题相似度根据算法 2 计算。该算法中,选取主题向量的最大余弦相似度作为目标文档和相关文档的相似度。

算法 2 topicSimilarity

输入:Target document P, relevant document P_r, a series of LDA models$\{(\alpha_k,\beta_k,K)|K=12,24,48,96\}$

输出:Topic similarity $sim_{topi}(P,P_r)$

```
ret ← ϕ;
foreach K ∈ {12,24,48,96} do
  ts ← articleTopics(αk,βk,P);
  tr ← articleTopics(αk,βk,Pr);
  ret ← ret ⋃ cosine(ts,tr);
end
return the largest value in ret;
```

5 参照提取和维度量化

为了构建准确性和完整性参照，首先利用开放式信息抽取(open information extraction)工具 ReVerb(http://reverb.cs.washington.edu/)和词性标注工具[24]抽取每个文档的事实纲要。

定义 2 一个文档 P 的事实纲要 $corp(P)$ 是该文档包含的事实集合。每个事实是一个三元组 $f=(h,v,t)$，h 是首元素，t 是尾元素，v 是谓词。

抽取上下文所有文档的事实纲要，同时构建 3 张哈希表 HH、TH、VH 和 1 张同义词表 T。其中 HH、TH、VH 分别以首元素、尾元素和谓词中包含的单词词干为键，支持对事实的搜索。T 是一个本地化的 WordNet (http://wordnet.princeton.edu/)同义词典。

构建参照时，定义事实 f 的支持度 $sup(f)=\frac{|equ(f)|}{|\Omega|}$，以过滤支持度比较低的事实描述。其中，$|\Omega|$ 表示上下文中包含的文档个数，$|equ(f)|$ 表示含 f 的同义事实的文档个数。为了识别同义事实，事实相似度定义如下。

定义 3 给定两事实 $f_1=(h_1,v_1,t_1)$，$f_2=(h_2,v_2,t_2)$，事实相似度 $sim_{fact}(f_1,f_2)=0.25*es(h_1,h_2)+0.5*ps(v_1,v_2)+0.25*es(t_1,t_2)$。其中，$es$ 是元素相似度，ps 是谓词相似度。

定义 4 给定两个首元素(或尾元素)h_1 和 h_2，其元素相似度 es 计算如下：

1)对元素预处理并抽取词干后，如果 h_1 和 h_2 在字面上相同，则 $es(h_1,h_2)=1$；否则，转向 2)。

2)假定 $h_1=\langle w_1^1,\cdots,w_1^i,\cdots,w_1^n\rangle$，$h_2=\langle w_2^1,\cdots,w_2^i,\cdots,w_2^n\rangle$ 具有相同的词性序列 $es(h_1,h_2)=\frac{1}{n}\sum_{i=1}^{n}sem(w_1^i,w_2^i)$，其中 $sem(w_1^i,w_2^i)$ 是通过查找同义词典确定的两个单词 w_1^i 和 w_2^i 的语义相似度；否则，转向 3)。

3)如果 h_1 是一个单词，而 h_2 是一个名词词组、形容词词组或副词词组，计算 h_1 和 h_2 的核心成分来确定 h_1 和 h_2 的相似度；否则，转向 4)。

4)如果 h_1 和 h_2 是词组，通过词组类型和其包含的核心成分来确定两者的相似度。

类似地，可以计算谓词相似度 ps，不再赘述。

5.1 目标文档的准确性计算

5.1.1 构建准确性参照

定义 5 给定目标文档 P，其准确性参照是一个事实集合。该集合中每个事实是目标文档相应事实的最精确表达。

由于上下文中各个文档相互独立，提出在所有文档的事实纲要上，采用投票确定准确性参照中的事实表达。这通过识别备选事实和事实投票两个步骤实现。在识别备选事实时，以目标事实中的单词为键，在哈希表中查找与目标事实相似的事实表达，如算法 3 所示。

算法 3 extractCandiFacts

输入：Thesaurus T，fact hash tables HH，VH，TH，target fact fs

输出：Top Q candidate facts

```
ret ← ϕ;
(h,v,t) ← fact components of fs;
set1 ← query(T,HH,VH,fs,h,v);
set2 ← query(T,HH,TH,fs,h,t);
set3 ← query(T,VH,TH,fs,v,t);
ret ← set1 ∪ set2 ∪ set3;
sort facts in ret in descending order based on similarity w. r. t. fs;
return top Q ones in ret;
```

其中 query 例程根据事实分量在哈希表中查找相似事实，如算法 4 所述。

算法 4 query

输入：Thesaurus T，hash tables H_1，H_2，target fact fs，two components of fs，denoted as c_1 and c_2

输出：Similar facts w. r. t. fs

```
F ← ϕ; S1 ← retrieve c1′ equivalents from T;
S2 ← retrieve c2′ equivalents from T;
foreach s1 ∈ S1 do
  F1 ← retrieve facts from H1 or H2 based on s1;
  foreach s2 ∈ S2 do
    F2 ← retrieve facts from H1 or H2 based on s2;
    F ← F ∪ (F1 ∩ F2);
  end
end
return F;
```

准确性参照按照算法 5 计算获得，准确性参照的事实在备选事实集合中根据可信度投票获得。

定义 6 给定一个事实 f，其可信度 $conf(f)=\frac{1}{3}conf_h(f)+\frac{1}{3}conf_v(f)+\frac{1}{3}conf_t(f)$，其中

$$conf_h(f)=\frac{|\Lambda^{hvt}|}{|\Lambda^{vt}|},conf_v(f)=\frac{|\Lambda^{hvt}|}{|\Lambda^{ht}|},$$

$$conf_t(f)=\frac{|\Lambda^{hvt}|}{|\Lambda^{hv}|}$$

这里 Λ^{hvt} 是与 f 等价的事实集合。Λ^{vt}、Λ^{ht}、Λ^{hv} 分别是与(v,t)、(h,t)和(h,v)分量等价的事实集合。

算法 5 constructAccuracyBaseline

输入：target article profile corp(P)，thesaurus T，fact hash tables HH，VH，TH，confidence threshold μ

输出：Accuracy baseline of P

```
Σ ← ϕ; // Initialize the accuracy baseline
foreach fs ∈ corp(P) do
  X ← extractCandiFacts(T,HH,VH,TH,fs);
  if μ > conf(f)(∀ f ∈ X\ fs)
  then
    ftop ← null;
  else
    conflist ← sorts all facts in X based on their confidence;
    ftop ← top fact of conflist;
    end
    Σ ← Σ ∪ ftop;
end
return Σ;
```

假定目标文档最多有 M 个事实，算法 5 的运行时间由目标文档事实个数和算法 3(extractCandiFacts)的磁盘 I/O 次数决定。查询算法 query 的时间复杂度取决于在词典表 T 和哈希表上的磁盘 I/O。由于事实分量的单词个数上界是一个常数，假设每个单词最多有 L 个同义词，可知算法 5 的时间复杂度是 $O(ML^2)$。由于 $M\gg L$，因此构造准确性参照的时间主要取决于目标文档包含的事实个数 M。

5.1.2 准确性度量

准确性是指目标文档的内容表达在多大程度上与事实相符。

定义 7 一个包含 n 个事实的文档 P，其准确性定义为 $acc(P)=\frac{1}{n}\sum_{j=1}^{n} sim_{fact}(f_j, f_j^b)$。其中 f_j^b 是 f_j 在准确性参照中的对应事实。如果对应事实为 $null$，则 $sim_{fact}(f_j, null)=0$。

5.2 目标文档的完整性计算

5.2.1 构建完整性参照

为构建完整性参照，将上下文事实集合抽象成无向图：无向图每个节点代表一个事实，每条边代表两个事实的相似度。参照图具有如下性质：(1)覆盖上下文所有事实；(2)不同事实以不同的相似度关联。完整性参照图通过如下两个步骤构建。

步骤 1 参照图初始化

每个唯一事实 f_i 对应图中一个节点，每个节点赋予初始完整性值 $s(f_i, 0)=1$。如果两个事实的相似度大于设定的阈值 θ，构建以事实相似度为权重的无向边。

步骤 2 参照图求精

如果节点是一个孤立点，完整性值不变化。如果节点和邻近事实节点连通，该节点值迭代如下：

$$s(f_i,t)=s(f_i,t-1)-\frac{1}{2^t}\sum_{j\in con(f_i)}\frac{s(f_j,t-1)}{|con(f_i)|+\sum_{f_k\in con(f_j)}\frac{1}{sim_{fact}(f_k,f_j)}}$$

其中，t 是迭代计数器，$con(f_i)$ 是与节点 f_i 直接相连的节点集合。给定收敛阈值 η，如果所有节点中两次迭代变化的最大值小于 η，则迭代停止。完整性参照迭代求解的原理在于每个节点覆盖一定的信息量，相邻节点的重叠信息量通过迭代消除，直至收敛到一个稳定的值。

定理 1 完整性参照图中各个节点的完整性值收敛于 0 和 1 之间。

证明：节点 f_i 的完整性值在迭代过程中构成一个序列 $s(f_i,0), s(f_i,1), \cdots, s(f_i,t), \cdots$，现证明其有界并且单调。

有界性：用归纳法证明 $\forall t\geqslant 1, \frac{1}{2^t}\leqslant s(f_i,t)\leqslant 1$ 成立。

(1) 当 $t=1$ 时，$s(f_i,1)=1-\frac{1}{2}\sum_{j\in con(f_i)}\frac{1}{|con(f_i)|+\sum_{f_k\in con(f_j)}\frac{1}{sim_{fact}(f_k,f_j)}}\geqslant\frac{1}{2}$ 成立。

(2)假设在 $t=n$ 时，$\frac{1}{2^t}\leqslant s(f_i,t)\leqslant 1$ 成立。当 $t=n+1$ 时，

$$\begin{aligned}
&s(f_i,n+1)\\
&=s(f_i,n)-\frac{1}{2^{n+1}}\sum_{j\in con(f_i)}\frac{s(f_j,n)}{|con(f_i)|+\sum_{f_k\in con(f_j)}\frac{1}{sim(f_k,f_j)}}\\
&\geqslant s(f_i,n)-\frac{1}{2^{n+1}}\sum_{j\in con(f_i)}\frac{1}{|con(f_i)|+\sum_{f_k\in con(f_j)}\frac{1}{sim(f_k,f_j)}}\\
&\geqslant\frac{1}{2^n}-\frac{1}{2^{n+1}}\sum_{j\in con(f_i)}\frac{1}{|con(f_i)|+\sum_{f_k\in con(f_j)}\frac{1}{sim(f_k,f_j)}}\\
&\geqslant\frac{1}{2^{n+1}}
\end{aligned}$$

单调性：在相邻两次迭代过程中 $s(f_i,t)-s(f_i,t+1)=\frac{1}{2^{t+1}}\sum_{j\in con(f_i)}\frac{s(f_j,t)}{|con(f_i)|+\sum_{f_k\in con(f_j)}\frac{1}{sim_{fact}(f_k,f_j)}}\geqslant 0$。由此可知，迭代中各个节点的完整性值单调递减。

各节点完整性值由于在迭代中是一个有界单调递减数列，因此必收敛于一个介于 0 和 1 的值。证毕。

5.2.2 完整性度量

完整性表示目标文档在多大程度上覆盖主题相关事实。为了计算完整性，按照与构建完整性参照相同的方法，构建目标文档的事实图。然后，完整性按照定义 8 计算。

定义 8 目标文档 P 的完整性是 $comp(P)=\frac{\sum_{f_i\in P}s(f_i)}{\sum_{f_j\in B^P}s(f_j)}$，其中 $s(f_i)$ 和 $s(f_j)$ 分别是目标文档和参照的节点的完整性值。

6 实验分析

从 Wikipedia 上下载了 1200 篇英文 Web 文档，实验在双核 i3 CPU@2.4GHz 和 2048M 内存的笔记本上运行。下载的文档已经被 Wikipedia 社区划分为 Featured Article (FA)，Good Article (GA)，B-Class (B)，C-Class (C)，Start-Class (ST)，Stub-Class (SU) 6 个质量类(http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment)中的一类。其中上下文相似度阈值 υ、词法权重 ρ 以及收敛阈值 η 通过采样确定。实验从 FQA 方法的性能、影响参数和相关工作比较 3 个方面展开。

6.1 精度和运行时间

6.1.1 评估精度

从参照精度和维度评估精度两个方面分析 FQA 方法性能。一个维度参照 B 的精度是 $prec(B)=\frac{|B^{FQA}\cap B^{man}|}{|B^{man}|}$，其中 B^{FQA} 是通过 FQA 方法识别的参照，B^{man} 是人工参与识别的参照。图 2 显示了各个数据质量类的准确性参照的精度，图 3 显示的是各个数据质量类的完整性参照的精度。可见，准确性参照的精度在 92%到 95%之间，完整性参照的精度在 86%到 91%之间。

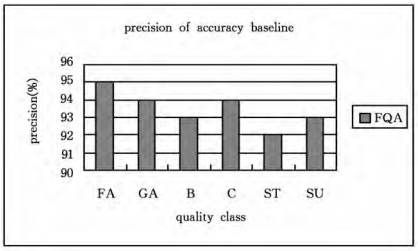

图 2 准确性参照的精度

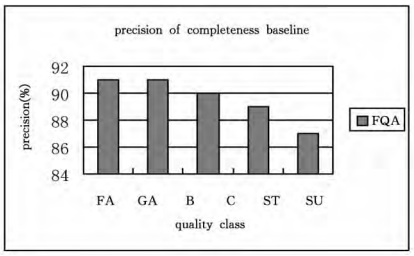

图 3 完整性参照的精度

针对每篇目标文档，FQA 方法计算一个维度(或维度组

合)的评估值。为了计算评估精度,采用如下方法将 FQA 的评估值映射到已经标注的数据质量类,以实现评估精度指标的量化。设数据集有 N 篇文档,$N_1,\cdots,N_6$ 代表每个数据质量类的文档数目,易知 $N_1+\cdots+N_6=N$ 成立。根据 FQA 给出的评估值,N 个文档降序排列为 $P_1,\cdots,P_N$,易知前 N_1 个文档属于 FA 类,随后 N_2 个文档应属于 GA 类,依此类推。每个质量类 i 的评估精度 $prec(i)=\frac{|X\cap Y|}{|Y|}$,其中 Y 是在质量类 i 中的文档,X 是采用 FAQ 方法判定属于 i 类的文档。

图 4 显示了分别基于准确性(acc.)、完整性(comp.)及准确性和完整组合(acc.+comp.)进行评估的精度。基于准确性评估的精度在 90%到 93%之间,基于完整性进行评估的精度在 84%到 90%之间。可见,基于准确性进行评估的精度高于基于完整性进行评估的精度。基于准确性和完整性组合(准确性和完整性权重分别是 0.64 和 0.36)进行评估取得最好的效果,其评估精度在 94%到 97%之间。

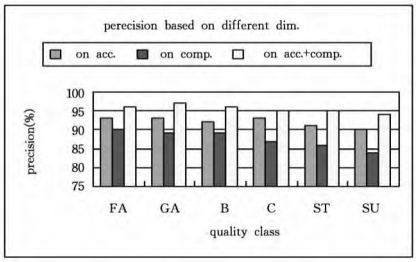

图 4 基于不同维度的数据质量评估精度

6.1.2 运行时间因素分析

图 5 显示出构建准确性参照的时间,运行时间随目标文档事实数目的变化。可见,运行时间随目标文档事实数目线性增长。

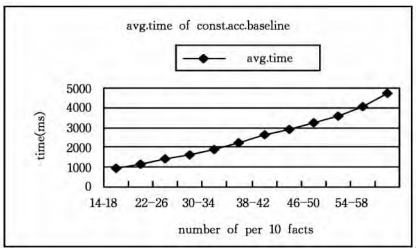

图 5 构建准确性参照的时间

完整性参照构建时间由迭代次数决定,而迭代次数由收敛阈值 η 决定。图 6 显示出迭代次数随错误率的变化情况,错误率是所有节点在两次连续迭代间的平均值差。可见,在经过最初的若干次迭代后,出错率迅速趋于稳定。为此,根据图中事实的数目,把图分成 7 个组。对于每个组,收敛阈值 η 设定在曲率变化最快的地方,如表 1 所列。

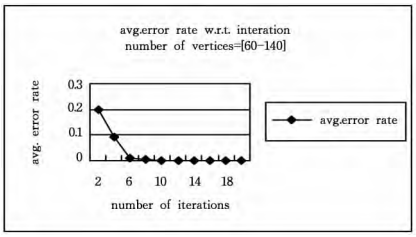

图 6 平均错误率随迭代次数的变化

表 1 不同数目事实的收敛阈值 η 设定

# of facts(×10)	6—14	15—22	23—30	31—38	39—46	47—54	55—62
η	0.009	0.013	0.020	0.021	0.027	0.033	0.035

6.2 上下文阈值对评估精度的影响

上下文阈值 υ 决定了上下文中文档的数目,进一步决定了备选事实的数目。图 7 显示了在 FA 质量类上下文的平均文档数目随阈值的变化,易知,阈值 υ 越小,上下文所包含的文档数目越多。图 8 显示了在 FA 质量类上质量评估精度随阈值的变化。可见,随着阈值减小,评估精度逐渐变大。这是因为,当阈值很大时,上下文缺少足够的文档来支持相应的事实。但可以看到,如果上下文阈值 υ 过小,上下文会引入过多大噪音,精度也会下降。在其它质量类上的评估精度随上下文阈值的变化显示了相同趋势,就不再赘述。

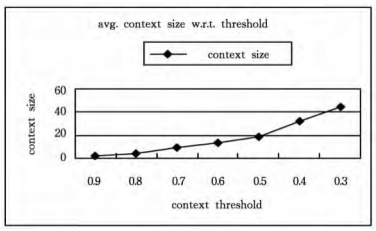

图 7 上下文大小随阈值的变化

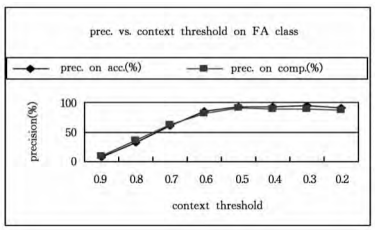

图 8 评估精度随上下文阈值的变化

6.3 相关工作比较

目前对 Web 数据质量进行评估主要基于词法或语法特征提取,比较有代表性的是文献[5]提出的根据特征训练向量机训练模型进行质量评估,不妨将其简记为 SVR。SVR 方法综合文本、评论和网络特征 3 个方面来训练模型。本文对 FQA 方法和 SVR 方法进行了比较。为了保证比较公平,SVR 方法和 FQA 方法各运行 10 次,取 10 次的平均值来对比。SVR 方法采用 2-折交叉验证,即每次实验中,一半文档作为训练集,一半文档作为测试集。FQA 方法利用准确性和完整性两个维度加权来进行质量评估,权重分别取 0.64 和 0.36。从图 9 可以看出,FQA 方法的评估精度比 SVR 方法的平均高出 7 到 12 个百分点。FQA 方法取得更好效果的原因在于:(1)FQA 从事实内涵角度理解内容,更准确地抓住了准确性和完整性的本质。(2)采用 LDA 主题模型可以准确地识别与主题相关的上下文档,能够精确地构造维度参照。

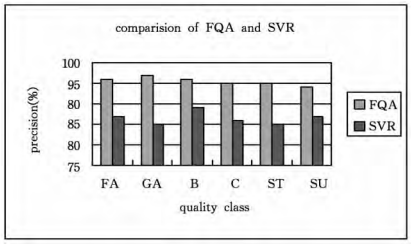

图 9 和相关工作的评估精度比较

(下转第 255 页)

学,2005,19(4):135-140

[11] Bonikowski Z, Bryniarski E, Wybraniec U. Extensions and intentions in the rough set theory[J]. Information Sciences, 1998, 107(1-4):149-167

[12] Zhu W, Wang F Y. Reduction and axiomization of covering generalized rough sets[J]. Information Sciences, 2003, 152:217-230

[13] 魏莱,苗夺谦,徐菲菲,等.基于覆盖的粗糙模糊集模型研究[J].计算机研究与发展,2006,43(10):1719-1723

[14] 徐忠印,廖家奇.基于覆盖的模糊粗糙集模型[J].模糊系统与数学,2006,20(3):141-144

[15] 胡军,王国胤,张清华.一种覆盖粗糙模糊集模型[J].软件学报,2010,21(5):968-977

[16] 胡军,王国胤.覆盖粒度空间的层次模型[J].南京大学学报:自然科学版,2008,44(5):551-558

[17] 胡军,王国胤.粗糙集的不确定性度量准则[J].模式识别与人工智能,2010,23(5):606-615

(上接第 251 页)

结束语 Web 文档数据无处不在,数据质量评估是获得高质量 Web 文档的关键,从而保证数据分析产生有意义的结果。本文提出了一种基于事实提取的 Web 文档内容数据质量评估方法。该方法的优越性在于:(1)不依赖于某类特征,根据从内容中提取的事实来度量准确性和完整性,实现了基于语义内涵的数据质量评估;(2)它是一种对数据质量自动化评估的方法。

理论分析和实验表明了本文方法的有效性。今后将对算法的某些环节进行改进。

参 考 文 献

[1] Aebi D, Perrochon L. Towards improving data quality[C]// Proc. of the international conference on information systems and management of data. New York: ACM, 1993:273-281

[2] 马茜,古峪,张天成,等.一种基于数据质量的多源多模态感知数据获取方法[J].计算机学报,2013,36(10):2010-2131

[3] 郭志懋,周傲英.数据质量和数据清洗研究综述[J].软件学报,2002,13(1):2076-2082

[4] Pernici B, Scannapieco M. Data Quality in Web Information Systems[C]// Proc. of the 21st International Conference on Conceptual Modeling. Berlin Heidelberg: Springer, 2002:397-413

[5] Dalip D H, Cristo M, Calado P. Automatic assessment of document quality in web collaborative digital libraries [J]. ACM Journal of Data and Information Quality, 2011, 2(3):14

[6] Hu Mei-qun, Lim Ee-peng, Sun Ai-xin. Measuring Article Quality in Wikipedia: Models and Evaluation[C]// Proc. of the 16th CIKM. New York: ACM, 2007:243-252

[7] Zeng H, Alhossaini M A, Li D, et al. Computing trust from revision history[C]// Proc. of the 2006 International Conference on Privacy, Security and Trust: Bridge the Gap Between PST Technologies and Business Services. New York: ACM, 2006

[8] Blumenstock J E. Size Matters: Word Count as a Measure of Quality on Wikipedia[C]// Proc. of the 17th International Conference on World Wide Web. New York: ACM, 2008:1095-1096

[9] Knap T, Mlynkova I. Quality Assessment Social Networks: A Novel Approach for Assessing the Quality of Information on the Web[C]// Proc. of QDB of VLDB'10. 2010

[10] Baeza-Yates R, Rello L. On Measuring the Lexical Quality of the Web[C]// Proc. of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality. New York: ACM, 2012:1-6

[11] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003(3):993-1022

[12] Dong Xin, Berti-Equille L, Hu Yi-fan, et al. Global Detection of Complex Copying Relationships Between Sources[C]// Proc. of VLDB Endowment. New York: VLDB Endowment, 2010:1358-1369

[13] Fan Wen-fei. Dependencies Revisited for Improving Data Quality [C]// Proc. of PODS 2008. New York: ACM, 2008:159-170

[14] Rassbach L, Pincock T, Mingus B. Exploring the Feasibility of Automatically Rating Online Article Quality[EB/OL]. [2013-9-10]. http:// upload. wikimedia. org/wikipedia/wikimania2007/d/d3/ RassbachPincockMingus 07. pdf

[15] Liu Jun, Ram S. Who does what: Collaboration patterns in the Wikipedia and their impact on article quality[J]. ACM Transactions on Management Information Systems, 2011, 2(2):1-23

[16] Han Jing-yu, Wang Chuan-dong, Jiang Da-wei. Probabilistic Quality Assessment Based on Article's Revision History[C]// Proc. of the 22nd International Conference on Database and Expert systems Applications (DEXA). Berlin Heidelberg: Springer, 2011:574-588

[17] Dalvi N, Kumar R, Soliman M. Automatic Wrappers for Large Scale Web Extraction[C]// Proc. of the 37th International Conference on Very Large Databases. New York: VLDB Endowment, 2011:219-230

[18] 肖升,何炎祥.基于动词论元结构的中文事件抽取方法[J].计算机科学,2012,39(5):161-164

[19] 杨少华,林海略,韩燕波.针对模板生成网页的一种数据自动抽取方法[J].软件学报,2008,19(2):209-223

[20] Etzioni O, Fader A, Christensen J, et al. Open information extraction: the second generation[C]// Proc. of 22nd International Joint Conference on Artificial Intelligence. California: AAI press, 2011:3-10

[21] Simões G, Galhardas H, Gravano L. When Speed Has a Price: Fast Information Extraction Using Approximate Algorithms[C]// Proc. of the 39th International Conference on Very Large Databases. New York: VLDB Endowment, 2013:1462-1473

[22] Mays E, Damerau F J, Mercer R L. Context based spelling correction [J]. Information processing and management, 1991, 27(5):517-522

[23] Si X, Chang E Y, Gyongyi Z, et al. Confucius and its intelligent disciples: integrating social with search[C]// Proc. of the 36th International Conference on Very Large Databases. New York: VLDB Endowment, 2010:1505-1516

[24] Toutanova K, Klein D, Manning C D, et al. Feature-rich part-of-speech tagging with a cyclic dependency network[C]// Proc. of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2003:173-180