Journal of Computer Applications ISSN 1001-9081 2014-08-10
计算机应用,2014,34(8):2311-2316,2331 CODEN JYIIDU http://www.joca.cn

文章编号:1001-9081(2014)08-2311-06 doi:10.11772/j.issn.1001-9081.2014.08.2311

基于模拟退火的在线 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)

(* 通信作者电子邮箱 jyhan@njupt.edu.cn)

摘 要:针对基于训练模型或用户交互的 Web 数据质量评估方法不能在线响应,也不能获取内容事实内涵的问题,提出一种基于模拟退火(SA)的在线 Web 文档内容数据质量评估(QASA)方法。首先,通过在 Web 上搜集主题相关文档,构建目标文档的相关空间,进一步采用开放式信息抽取技术抽取文档内容的事实;然后,采用 SA 技术在线构建两个最重要的数据质量维度即准确性和完整性的参照;最后,通过比对目标文档和维度参照的事实来量化数据质量维度。实验结果表明,QASA 方法可以及时返回近似最优解,并保持与离线算法等同或高于 10% 的精度。该方法不仅能满足实时响应的要求,而且具有高的评估精度,可应用于在线识别高质量的 Web 文档。

关键词:数据质量;Web 文档;模拟退火;维度;事实

中图分类号:TP311.13;TP18 **文献标志码**:A

Data quality assessment of Web article content based on simulated annealing

HAN Jingyu*, CHEN Kejia

(*College of Computer Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing Jiangsu* 210003, *China*)

Abstract: Because the existing Web quality assessment approaches rely on trained models, and users' interactions not only cannot meet the requirements of online response, but also can not capture the semantics of Web content, a data Quality Assessment based on Simulated Annealing (QASA) method was proposed. Firstly, the relevant space of the target article was constructed by collecting topic-relevant articles on the Web. Then, the scheme of open information extraction was employed to extract Web articles' facts. Secondly, Simulated Annealing (SA) was employed to construct the dimension baselines of two most important quality dimensions, namely accuracy and completeness. Finally, the data quality dimensions were quantified by comparing the facts of target article with those of the dimension baselines. The experimental results show that QASA can find the near-optimal solutions within the time window while achieving comparable or even 10 percent higher accuracy with regard to the related works. The QASA method can precisely grasp data quality in real-time, which caters for the online identification of high-quality Web articles.

Key words: data quality; Web article; Simulated Annealing (SA); dimension; fact

0 引言

人们经常在线获取 Web 文档,其数据质量直接影响获取的数据的价值。数据质量公认为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]。然而 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 文档是松散的自然语句序列,缺少模式约束来保证质量;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。目前 Web 数据质量评估方面已经有许多工作[4-11]。但这些工作多采用离线训练模型或采用用户交互的方式来衡量数据质量,不适合在线处理的场合,也不能从事实内涵角度揭示 Web 文档内容是否和现实一致。

为了实现准实时的 Web 文档数据质量鉴别,提出基于模拟退火的数据质量在线评估(Quality Assessment based on Simulated Annealing, QASA)方法。QASA 采用模拟退火(Simulated Annealing, SA)在线构建数据质量维度参照,通过比对目标文档和参照的事实来量化数据质量维度。该方法由以下两个连续步骤构成:1)相关文档识别和事实提取。根据目标文档内容,在 Web 上搜集主题相关文档,并抽取其中的事实,从而构建目标文档的主题相关空间。2)在线维度参照构建和维度量化。在相关空间中,采用模拟退火在线构建两个最重要的数据质量维度即准确性和完整性的参照,然后将目标文档事实和维度参照进行比较,量化数据质量维度。

该方法的创新之处在于:1)利用模拟退火在线构建维度参照,可在规定的时间内返回处理结果,满足在线处理要求;2)通过信息抽取,将文档转化为事实集合,从事实内涵角度量化 Web 文档内容是否和现实世界一致。在实际数据上的实验表明 QASA 是 Web 文档数据质量在线评估的有效方法。

1 相关工作

1.1 Web 文档数据质量评估

目前 Web 文档内容质量评估主要采用基于模型或基于用户交互的方法。基于模型的方法提取内容中的特征,训练模型来进行质量评估:文献[4]综合文本、评论和网络三个方面特征,采用支持向量机模型对 Web 文档的数据质量进行评

收稿日期:2014-02-14;修回日期:2014-03-22。 基金项目:国家自然科学基金资助项目(61003040,61100135)。

作者简介:韩京宇(1976-),男,吉林白山人,副教授,博士,CCF 会员,主要研究方向:数据管理、知识库; 陈可佳(1980-),女,江苏淮安人,副教授,博士,主要研究方向:机器学习、信息系统。

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

估;文献[5]根据文本长度,采用机器学习推断Wikipedia文档数据质量好坏;文献[6]利用文档长度、词性、Web特征和可读性等,用最大熵理论训练模型,评估Wikipedia文档质量。基于用户交互的方法根据用户和文档交互特征来评估数据质量:文献[7-9]根据文档的修改历史和作者信誉度来估计Web内容质量的好坏;文献[10]利用用户信誉度和文本内容的依赖关系来计算文档质量;文献[11]根据用户的交互模式来判定Wikipedia文档的数据质量好坏。但这些方法侧重于离线处理,同时也不能抓取文档的事实内涵。

1.2 模拟退火技术

模拟退火算法在1982年由Kirkpatrick等首次提出[12]。它依据固体物质退火过程与问题求解过程的相似性来设计搜索策略,是一种启发式随机搜索算法。模拟退火算法不只是接受最优解,也会以一定的概率接受近似最优解。目前模拟退火技术在组合优化[13]、实时处理[14]方面得到广泛应用。

1.3 信息抽取技术

近年信息抽取技术获得长足进步[15-19],这个领域专注于从文档中抽取实体或事实,主要分成基于模式的、基于规则的和基于统计学习的三类方法[15-17]。这些方法针对特定模式或需求进行信息抽取,要预先知道种子模式。近年,针对开放域的信息抽取技术收到广泛关注[18-19],该类方法不依赖特定的特征,直接从文本中抽取事实,具有通用性。但针对Web文档内容抽取数据质量参照则鲜有工作。

2 相关文档识别和事实抽取

目标文档的相关空间通过以下两个步骤识别。

1)搜集备选相关文档。根据目标文档题目和关键字在Web上搜索相关文档,选取PageRank值高于设定阈值的文档,以获得普遍认可的备选相关文档。

2)过滤不相关文档。目标文档的相关空间由所有和目标文档描述相同主题的文档构成。给定目标文档和备选相关文档,如果两篇文档描述相同的主题,其词法应具有较高的相似度。本文提出在n-gram空间中[20]识别主题密切相关文档,如算法1所示。

算法1 determineRelevantSpace。

Input:target article P, possibly relevant articles $\{P_1,P_2,\cdots,P_M\}$, lexical similarity threshold $0<\gamma<1$。

Output:relevant space Ω。

$\Omega\leftarrow\varnothing$;

remove stop words and extract stem for each word in P and $P_i(1\leqslant i\leqslant M)$;

calculate word n-gram vectors $\boldsymbol{vec}^p$ and $\boldsymbol{vec}^i(1\leqslant i\leqslant M)$;

foreach P_i do

  if $\cos(\boldsymbol{vec}^p,\boldsymbol{vec}^i)>\gamma$ do

    $\Omega\leftarrow\Omega\cup P_i$;

  end

end

return Ω;

算法1首先去除每篇候选文档中的停止词(stop word)并提取词干;然后,根据备选文档中包含的单词构建n-gram向量,向量中的每个分量代表对应单词在文档中出现的频率;最后,根据备选文档和目标文档向量的余弦积决定是否过滤掉备选文档。与目标文档的余弦积超过规定阈值的所有文档构成目标文档的相关空间。

为了获取相关文档的事实内涵,采用信息抽取技术[18]获取文档的事实纲要。

定义1 事实纲要。一个文档P的事实纲要$corp(P)$是文档内容包含的事实集合。每个事实是一个3元组$f=(h,v,t)$,其中:h是首元素,t是尾元素,v是谓词。

3 质量维度参照在线构建和维度量化

为了基于事实内涵计算文档相似性,事实相似度定义如下。

定义2 事实相似度。给定两个事实$f_1=(h_1,v_1,t_1)$,$f_2=(h_2,v_2,t_2)$,事实相似度$sim_{\text{fact}}(f_1,f_2)=0.25\times es(h_1,h_2)+0.5\times ps(v_1,v_2)+0.25\times es(t_1,t_2)$。其中:$es$是元素相似度,$ps$是谓词相似度。

定义3 元素相似度。给定两个元素h_1和h_2(或t_1和t_2),元素相似度es计算如下:

1)如果h_1和h_2在字面上相同,$es(h_1,h_2)=1$;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,则$es(h_1,h_2)=\frac{1}{n}\sum_{i=1}^{n}sem(w_1^i,w_2^i)$,其中$sem(w_1^i,w_2^i)$是通过查找本地同义词典[21]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。

类似地计算谓词相似度ps,不再赘述。

为了量化目标文档的准确性和完整性,要在目标文档的相关空间构建准确性和完整性参照。从理论上说,准确性参照包含目标文档各个事实的最准确表达,完整性参照包含目标文档主题的最完整表达。但在有限时间内,穷尽相关空间的所有可能组合计算准确性参照和完整性参照是不现实的。为此,提出采用模拟退火,在规定时间内返回近似的最优解,从而满足在线处理要求。

模拟退火是一种启发式的概率算法,其模拟金属退火过程,寻找整个系统内能最低的状态。它从一个足够高的温度$T=T_{\max}$开始融化系统,逐步降低温度并计算每个状态的内能,直至收敛。在每个温度平台,最多迭代L次,称为一个周期(epoch)。如果L次迭代完成,或系统提前到达到平衡,则降到下一个更低温度。这里平衡是指一个非递减的内能值序列。在每次迭代中,相对于当前状态S,邻近状态S'以概率$P(e,e',T)=\min\left(1,\exp\left(\frac{e-e'}{T}\right)\right)$被接受,其中:$e$和$e'$分别代表$S$和$S'$状态的内能,$T$代表温度。

3.1 准确性参照构建和准确性量化

给定目标文档,其准确性参照包含目标文档每个事实的最准确表达。为了识别最准确表达,事实表述定义如下。

定义4 事实表述。给定目标事实f和事实表达fs,如果满足$sim_{\text{fact}}(f,fs)=1$,则称$fs$是$f$的一个事实表达。

不失一般性,相关空间Ω中各个相关文档彼此独立。对于同一个事实,不同文档可能会有字面上不同的事实表述。如果一个事实被相互独立的多个文档所描述,则该事实是准确事实。

定义5 准确事实。给定支持度阈值$\beta(0<\beta\leqslant 1)$,假设

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

事实 f 有 n 个字面上不同的事实表达 $\{fs_1, fs_2, \cdots, fs_n\}$,其支持度分别是 $sup(fs_1), sup(fs_2), \cdots, sup(fs_n)$。如果满足 $\left(\sum_{i=1}^{n} \sup(fs_i)\right)/|\Omega| \geqslant \beta$,则称 f 是一个准确事实。这里支持度 $sup(fs_i)$ 是相关空间中包含 fs_i 的相关文档数目。

给定目标文档事实纲要 $corp(P) = \{f_1, f_2, \cdots, f_n\}$,构建准确性参照即寻找被支持的准确事实最多的相关文档子集合。形式化如下,给定目标文档纲要 $corp(P) = \{f_1, f_2, \cdots, f_n\}$ 和相关空间 $\Omega = \{P_1, P_2, \cdots, P_M\}$,寻找相关空间的一个子集合 $R^{acc} \subseteq \Omega$,满足 $\sum_{i=1}^{n} AFV(f_i)$ 最大化。这里

$$AFV(f_i) = \begin{cases} 1, & f_i \text{ 在 } R^{acc} \text{ 中是一个准确事实} \\ 0, & \text{其他} \end{cases} \quad (1)$$

每个子集合 R^{acc} 对应一个退火状态 $S = (R^{acc}, \boldsymbol{AFV}[\,], AFV^{sup}[\,])$,$\boldsymbol{AFV}[\,]$ 是准确事实向量,其每一个元素取值 0 或者 1,$AFV^{sup}[i]$ $(0 \leqslant i < n)$ 是支持事实 f_i 的事实表达集合。为了在既定时间内返回解,采用算法 2 构造目标文档的准确性参照。

算法 2　constructAccuracyBaseline。

Input: target article facts $corp(P) = \{f_1, f_2, \cdots, f_n\}$, relevant space $\Omega = \{P_1, P_2, \cdots, P_M\}$, agreement fact threshold β, epoch length L, upper bound temperature T_{max}, sequilibrium threshold ε^{acc}。

Output: optimal accuracy baseline state。

```
//initialization
HH ← ∅, VH ← ∅, TH ← ∅;
construct Hash tables HH, VH and TH based on facts in Ω;
S ← initAccState(corp(P), Ω, HH, VH, TH, β);
e ← E^acc(S), S_best ← S, e_best ← e, T ← T_max;
//iteration
while {T > 0} do
  epochLen ← 0;
  foreach P_i ∈ Ω do
    S' ← arg min{E^acc(S.R^acc + P_i), E^acc(S.R^acc − P_i),
         E^acc(S.R^acc ± P_i)};
    e' ← E^acc(S');
    S ← S' with the probability min(1, exp((e − e')/T));
    epochLen ++;
    if e < e_best then
      S_best ← S, e_best ← e;
    end
    if time is out then
      return S_best;
    end
    if S is an equilibrium point or epochLen > L then
      T ← T − 1;
    end
  end
end
return S_best;
```

整个算法分成两个阶段,即初始化和迭代。初始化时,首先在内存中创建 3 张哈希表,即 Head Hash(HH)、Verb Hash(VH)和 Tail Hash(TH)。3 张表分别根据相关空间中事实的首元素、谓词和尾元素中包含的单词,将事实映射到对应桶中,以支持事实搜索。桶中每个键 w 对应一个条目 (id_{art}, f),f 是文档 id_{art} 中包含 w 的事实。在创建了哈希表后,采用算法 3 初始化搜索空间,如下所示。

算法 3　initAccState。

Input: target article facts $corp(P)$, relevant space Ω, three Hash indices HH, VH and TH, agreement fact threshold β。

Output: initial state S。

```
randomly choose a subset of Ω, denoted as R;
S.R^acc ← R;
foreach f_i ∈ corp(P) do
  LL ← ∅;
  foreach word w ∈ f_i.h do
    SF ← facts in bucket w of HH;
    LL ← LL ∪ SF;
  end
  foreach word w ∈ f_i.v do
    SF ← facts in bucket w of VH;
    LL ← LL ∪ SF;
  end
  foreach word w ∈ f_i.t do
    SF ← facts in bucket w of TH;
    LL ← LL ∪ SF;
  end
  Filter out the facts in LL that are not fact statements of f_i by
    consulting local WordNet;
  if |LL| / |Ω| ≥ β then
    S.AFV[i] ← 1, S.AFV^sup[i] ← LL;
  end
end
return S;
```

在迭代阶段,随着温度的逐步降低寻找新的解状态。在每一个温度,最多迭代 L 次,每次对各个状态轮流进行相关文档的增、删或替换操作,以寻求能量最低的新状态。状态 S 的能量定义为:

$$E^{acc}(S) = n - \sum_{i=1}^{n} S.AFV[i] \quad (2)$$

能量 $E^{acc}(S)$ 越小,子集合 R^{acc} 中被支持的准确事实越多。在迭代中,将当前状态能量值 e 和最近的 K 个能量值进行比较,以决定是否到达平衡。假设 $L' = \{e_1, e_2, \cdots, e_K\}$ 是最近的 K 个能量值,对于给定的阈值 ε^{acc},如果对于任意的 $e_j \in L'$,$|e - e_j| / |e| < \varepsilon^{acc}$,系统达到一个不能跳出的状态,则降低温度。

算法 2 的运行时间主要由迭代决定。一个文档包含 $|corp(P)|$ 个事实,假设一个文档包含的事实个数的上界是 W,通过哈希,对每个事实的搜索在常数时间内完成。因此,算法 2 的时间复杂度是 $O(T_{max}L|\Omega|W)$。由于 W 是一个常数,算法 2 的时间复杂度主要由温度上界 T_{max},周期上界 L 和相关空间大小 $|\Omega|$ 决定。

假设最终状态对应的准确性参照中有 m 个准确事实,则准确性为:

$$accu(P) = m/n \quad (3)$$

3.2　完整性参照构建和完整性量化

完整性指目标文档在多大程度上覆盖主题相关事实。给定一个目标文档 P,其完整性参照是其主题文档应包含的相关事实。完整性是目标文档包含的事实数目和参照中包含的事实数目的比值。为了构建目标文档 P 的完整性参照,需要识别 P 中事实的同现事实。

定义 6　同现事实。给定目标文档的一个事实 f 和相关空间 Ω,同现事实 $CF(f)$ 是一个事实集合 $\{cf_j \mid 1 \leqslant j \leqslant m\}$。每个

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

事实 cf_j 满足如下两个条件：

1）邻近性。cf_j 和 f 在相关文档中出现在同一小节区域，以确保 cf_i 和 f 的相关性。

2）同现性。cf_j 和 f 同时出现在 $\delta^* \mid \Omega \mid$（$0 < \delta \leqslant 1$）个以上的相关文档，以确保两者高概率地同现。

定义7 完整性参照。给定一个包含 n 个事实的目标文档 $corp(P) = \{f_1, f_2, \cdots, f_n\}$，其完整性参照是 $corp(P) \cup \left(\bigcup_{i=1}^{n} CF(f_i)\right)$。

给定目标文档事实纲要 $corp(P) = \{f_1, f_2, \cdots, f_n\}$ 和相关空间 $\Omega = \{P_1, P_2, \cdots, P_M\}$，构建完整性参照归结为寻找 Ω 的子集合 R^{com}，从而满足最大化 $\bigcup_{i=1}^{n} CF(f_i)$。该问题采用如算法4的模拟退火过程来解决。

算法4 constructCompletenessBaseline。

Input: article facts $corp(P) = \{f_1, f_2, \cdots, f_n\}$, relevant space Ω, upper bound temperature T_{max}, epoch threshold L, co-occurrence threshold δ, equilibrium threshold ε^{com}。

Output: optimal completeness baseline state。

```
//initialization
HH^CF ← ∅, VH^CF ← ∅, TH^CF ← ∅;
construct Hash tables HH^CF, VH^CF and TH^CF based on facts in Ω;
S ← initCompleteState(corp(P), Ω, HH^CF, VH^CF, TH^CF, δ);
e ← E^comp(S)                          //calculate completeness energy
S_best ← S, e_best ← e, T ← T_max;
while T > 0 do
  epochlen ← 0;
  foreach P_i ∈ Ω do
    S' ← arg min{S.R^com + P_i, S.R^com − P_i, S.R^com ± P_i};
    e' ← E^comp(S');
    S ← S' with probability min(1, exp((e − e')/T));
    epochlen ++;
  end
  if {e < e_best} then
    S_best ← S;
    e_best ← e;
  end
  if time is out then
    return S_best;
  end
  if S is an equilibrium point or epochLen > L then
    T ← T − 1;
  end
end
return S_best;
```

每个状态 $S = (R^{com}, CFS[])$，其中：R^{com} 是相关文档子集合，$CFS[i]$（$0 \leqslant i < n$）是 f_i 的同现事实集合。整个算法由初始化和迭代两个阶段组成。在初始化阶段，根据同现事实中包含的单词构建3张内存哈希表 HH^{CF}、VH^{CF} 和 TH^{CF} 来索引同现事实。然后采用算法5进行初始化，如下所示。

算法5 initCompleteState。

Input: article facts $corp(P)$, relevant space Ω, Hash indices HH^{CF}, VH^{CF}, TH^{CF}, and co-occurrence threshold δ。

Output: completeness baseline state S。

```
S.R^com ← a subset of Ω is randomly chosen;
foreach f_i ∈ corp(P) do
  CF_i ← ∅; LL ← ∅;
  foreach word w ∈ f_i.h do
    LL ← LL ∪ f_i's equivalent facts in bucket w of HH^CF;
  end
  foreach word w ∈ f_i.v do
    LL ← LL ∪ f_i's equivalent facts in bucket w of VH^CF;
  end
  foreach word w ∈ f_i.t do
    LL ← LL ∪ f_i's equivalent facts in bucket w of TH^CF;
  end
  foreach fs ∈ LL do
    CF_i ← CF_i ∪ adjacent facts of fs;
  end
  foreach cf ∈ CF_i do
    if occurrence times of cf is above δ* |Ω| then
      S.CFS[i] ← S.CFS[i] ∪ cf;
    end
  end
end
return S;
```

在算法5中，对每一个事实 $f_i \in corp(P)$，首先根据哈希索引找到邻近事实，然后通过过滤选取共现事实。一个完整性状态 S 对应的能量为：

$$E^{comp}(S) = n \cdot n - \sum_{i=1}^{n} \mid S \cdot CFS[i] \mid \tag{4}$$

共现事实越多，状态 S 的能量越低。

在迭代阶段，随着温度降低，逐步寻找新的解状态。在每一个温度，最多执行 L 次迭代。每次迭代对相关文档依次进行增、删或替换操作，以寻求能量更低的状态 S'。通过比较当前状态 S 和新的状态 S' 的能量，以概率 $\min\left(1, \exp\left(\frac{e - e'}{T}\right)\right)$ 决定是否迁移到新状态 S'。

算法4的运行时间亦主要由迭代决定。假设一个文档包含的事实个数的上界是 W，一个小节区域包含的事实个数的上界是 k。通过哈希，目标文档的共现事实在 Wk 时间内识别。因此，算法4的时间复杂度是 $O(T_{max} L \mid \Omega \mid Wk)$。这里，$W$ 是一个常数，k 是一个很小的常数。

迭代结束时，$CFS[i]$ 的每个元素对应同现事实集合。此时，目标文档 P 的完整性为：

$$comp(P) = \frac{\mid corp(P) \mid}{\left| corp(P) \cup \left(\bigcup_{i=1}^{n} CF(f_i)\right) \right|} \tag{5}$$

4 实验

实验在2.4 GHz双核CPU、2 GB内存的台式机上实现。一个数据集是从Wikipedia上下载的1 200篇关于计算机科学（http://en.wikipedia.org/wiki/computer_science）和科学家（http://en.wikipedia.org/wiki/scientist）的文档，不妨记为Wiki。该数据集按照Wikipedia社区标准（en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial Team/Assessment）划分成6个质量类，分别是FA（Featured Article）、GA（Good Article）、B（B-Class）、C（C-Class）、ST（Start-Class）和SU（Stub-Class）。易知文档的准确性和完整性符合FA > GA > B > C > ST > SU（这里 > 表示优于）。另外一个数据集是从Web上收集的380篇文档，涵盖了科学、文化、艺术、历史、体育和战争等主题，不妨记为hybrid。对第二个数据集从准确性和完整

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

性两个维度进行人工分类。文档划分成 4 个质量类:Ⅳ、Ⅲ、Ⅱ和Ⅰ(Ⅳ > Ⅲ > Ⅱ > Ⅰ)。每个文档由 5 个用户分别独立投票,根据票数决定其数据质量类别。

对每一个目标文档,按照下述步骤搜集 20 个相关文档构成相关空间:

1)根据目标文档的题目和关键字利用 Google 搜集 PageRank 值高于 0.7 的 60 个主题相关文档。

2)根据文本内容中单词的 3-gram 相似性,识别最相似的 20 个相关文档构成目标文档的相关空间。对每个相关文档在线调用开放式信息抽取工具 ReVerb(http://reverb.cs.washington.edu/)获取其事实纲要。

准确性和完整性参照构建的平衡阈值通过取样获得,分别设置为 $\varepsilon^{acc}=0.1$,$\varepsilon^{com}=0.13$。数据集的准确支持阈值和同现阈值通过试错法(trial and error)获得。对 Wiki 数据集,准确支持阈值 $\beta=0.25$,同现阈值 $\delta=0.37$。对于 hybrid 数据集,$\beta=0.31$,$\delta=0.34$。除非特别说明,设置退火周期长度 $L=1050$,温度上界 $T_{max}=400$。取 10 次运行的平均值作为实验结果,以避免随机偏差。实验从在线评估精度、影响性能的参数和相关工作比较三个方面进行。

4.1　在线评估精度

QASA 方法给目标文档的数据质量维度一个位于 0 和 1 之间的评估值。为了识别维度评估值和数据质量类的对应关系,作如下处理:对 Wiki 数据集,设共有 N 篇文档,N_1 篇 FA,N_2 篇 GA,…,以此类推;设按 QASA 的维度分值将文档降序排序为 $P_1,\cdots,P_i,\cdots,P_N$;则前 N_1 篇文档属于 FA 类,其后的 N_2 篇文档属于 GA 类,……,以此类推。类似可以识别 hybrid 数据集上文档的质量类。图 1 和图 2 分别给出两个数据集上各个质量类的评估精度(准确性权重 0.6,完整性权重 0.4)。这里,质量类 i 的评估精度定义如下:

$$prec(i)=n_i/N_i \tag{6}$$

其中:N_i 代表质量类 i 的实际文档数目,n_i 代表 QASA 方法正确识别出的文档数目。

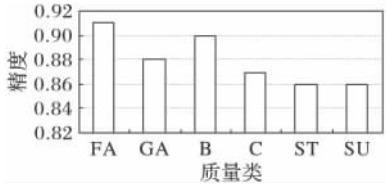

图 1　Wiki 数据集上评估精度

为了验证构造的维度参照的精度,设定温度 $T_{max}=10000$ 以几乎实现穷尽搜索,以此结果作为标准的维度参照。按式(7)计算参照错误率:

$$err_i=|m^i-M|/M \tag{7}$$

其中:m^i 是 QASA 算法在下降 i 个温度后,识别出的准确事实(或同现事实)数目;M 是标准参照中包含的事实数目。

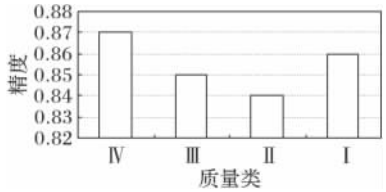

图 2　hybrid 数据集上评估精度

图 3 显示了在构建准确性参照和完整性参照时,随着温度的降低错误率的变化情况。可以看到错误率在最初急剧下降,然后逐渐趋于平缓。对于准确性参照,从图 3 可以观测到,下降 16℃后错误率下降到 42%,温度下降 256℃后,错误率下降到 9%;对于完整性参照,下降 16℃后,错误率下降到 50%,下降 256℃后,错误率下降到 14%。这表明 QASA 方法在绝大多数情况下能够及时返回可接受的解。

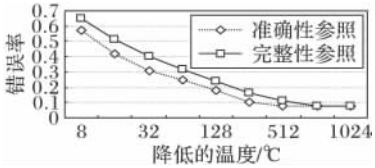

图 3　维度参照错误率随温度变化

4.2　影响性能的参数

本节实验周期长度 L 和相关空间大小 $|\Omega|$ 对评估精度和运行时间的影响。

4.2.1　周期长度对评估精度和运行时间的影响

设定 $|\Omega|=20$,$T_{max}=400$,将周期长度从 5 上升到 2805。图 4 显示了识别出来的准确事实和同现事实比例随周期长度的变化情况。可见随着周期长度变大,越来越大比例的准确事实和同现事实被识别。但当周期长度足够大时,长度的增大对识别出的事实数目几乎没有影响。这是因为当周期长度足够大时,可能状态被已被搜索的概率越来越大。

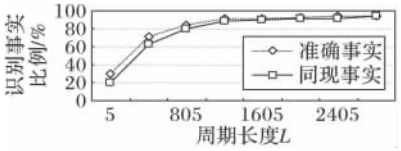

图 4　识别出的事实比例随周期长度变化

设定 $|\Omega|=20$,$T_{max}=400$,将周期长度从 5 上升到 2805。图 5 显示了准确性参照和完整性参照的构建时间随周期长度的变化。可以观测到运行时间随着周期长度的增长而线性变化。

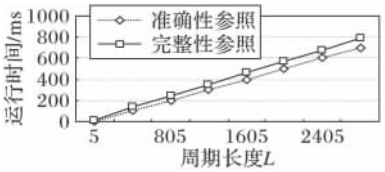

图 5　运行时间随周期长度变化

4.2.2　相关空间大小对评估精度和运行时间的影响

设定周期长度 $L=1050$,$T_{max}=400$,将相关空间大小 $|\Omega|$ 从 5 提高到 30。图 6 显示了识别出的准确事实和同现事实的比例随相关空间大小的变化情况。可见随着相关空间变大,准确性参照和完整性参照的精度会提高。但相关空间大到一定程度,这种效果迅速消失。这是因为相关空间大小足够大时,新增加的文档对准确事实和同现事实的支持度的影响变得非常微弱。

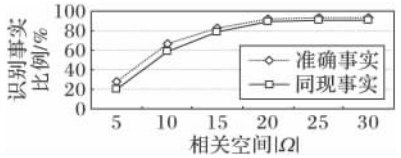

图 6　识别出的事实比例随相关空间大小变化

为了验证运行时间的变化,设定 $L=1050$,$T_{max}=400$,将相关空间大小从 5 增加到 30。图 7 显示了构建准确性参照和完整性参照的运行时间随相关空间大小的变化情况。可见运行时间随相关空间而线性增长。

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

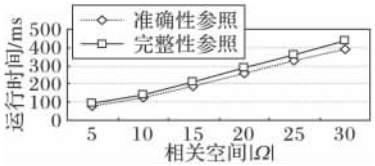

图7 运行时间随相关空间大小变化

4.3 相关工作比较

目前尚缺少在线的Web文档内容数据质量评估方法，将本文方法和文献[4]中采用训练支持向量机的数据质量评估(Quality Assessment based on Support Vector Machine，QASVM)方法作了比较。QASVM方法首先离线训练质量评估的支持向量机，然后根据训练的模型在线评估Wikipedia文档质量。实验在Wiki数据集上进行，图8和9显示了QASA方法和QASVM方法分别在FA和ST两个质量类上的精度比较。

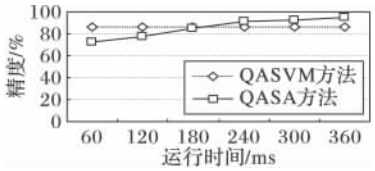

图8 FA质量类上精度比较

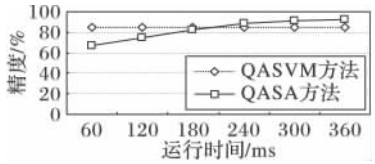

图9 ST质量类上精度比较

可见，在FA质量类上，QASA在运行时间小于180 ms的情况下，精度稍逊色于QASVM方法，但随着迭代时间增加，QASA方法的精度很快优于QASVM方法；在ST质量类上，QASA在运行时间小于200 ms的情况下，精度稍逊色于QASVM方法，但此后，QASA方法的精度要高于QASVM方法。在其他质量类上表现出类似的趋势，篇幅有限，不再赘述。这是因为QASA方法以内容的事实为基础，采用退火算法以概率逼近最优解，当时间非常短时结果不够理想，但绝大部分情况会给出一个高精度的结果。由于QASA方法不需要离线训练模型，因此更适合在线处理场合。

5 结语

为了实现Web文档及时、准确的信息获取，Web文档内容的数据质量在线评估是亟待解决的问题。本文提出在抽取文档事实的基础上，用模拟退火实现准实时的在线评估，保证在任何时候返回一个可接受的近似最优解。本文方法根据事实语义量化数据质量，不依赖任何特征，是一种通用的在线数据质量评估方法。实验表明QASA是实现Web文档数据质量在线评估的有效途径。

参考文献：

[1] AEBI D, PERROCHON L. Towards improving data quality [C]// Proceedings of the 1993 International Conference on Information Systems and Management of Data. Washington, DC: IEEE Computer Society, 1993: 273 - 281.

[2] BATINI C, CAPPIELLO C, FRANCALANCI C, *et al.* Methodologies for data quality assessment and improvement [J]. ACM Computing Surveys, 2009, 41(3): 8 - 75.

[3] BOUZEGHOUB M, PERALTA V. A framework for analysis of data freshness [C]// Proceedings of the 2004 International Information Quality Conference on Information System. Washington, DC: IEEE Computer Society, 2004: 59 - 67.

[4] DALIP D H, GONCALVES M A, CRISTO M, *et al.* Automatic assessment of document quality in Web collaborative digital libraries [J]. Journal of Data and Information Quality, 2011, 2(3): article 14.

[5] BLUMENSTOCK J E. Size matters: word count as a measure of quality on Wikipedia [C]// Proceedings of the 17th International Conference on World Wide Web. New York: ACM Press, 2008: 1095 - 1096.

[6] RASSBACH L, PINCOCK T, MINGUS B. Exploring the feasibility of automatically rating online article quality [EB/OL]. [2013-08-10]. http://upload.wikimedia.org/wikipedia/wikimania2007/d/d3/RassbachPincockMingus07.pdf.

[7] ZENG H, ALHOSSAINI M A, DING L. Computing trust from revision history [C]// PST'06: Proceedings of the 2006 International Conference on Privacy, Security and Trust: Bridge the Gap Between PST Technologies and Business Services. New York: ACM Press, 2006: 33 - 40.

[8] ZENG H, ALHOSSAINI M A, FIKES R, *et al.* Mining revision history to assess trustworthiness of article fragments [C]// Proceedings of the 2006 International Conference on Collaborative Computing: Networking, Applications and Worksharing. New York: ACM Press, 2006: 1 - 10.

[9] HU M, LIM E P, SUN A. Measuring article quality in Wikipedia: models and evaluation [C]// Proceedings of the 16th ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2007: 243 - 252.

[10] YU S, MASATOSHI Y. Assessing quality scores of Wikipedia article using mutual evaluation of editors and texts [C]// Proceedings of the 22nd ACM Conference on Information and Knowledge Management. New York: ACM Press, 2013: 1727 - 1732.

[11] LIU J, 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.

[12] JOHNSON D S, ARAGON C R, McGEOCH L A, *et al.* Optimization by simulated annealing: an experimental evaluation [J]. Operations Research, 1991, 39(3): 78 - 406.

[13] WANG X, XU X, WANG Z. A profit optimization oriented service selection method for dynamic service composition [J]. Chinese Journal of Computers, 2010, 33(11): 2104 - 2115. (王显志，徐晓飞，王忠杰. 面向组合服务收益优化的动态服务选择方法[J]. 计算机学报，2010，33(11)：2104 - 2115.)

[14] TAN C M. Simulated annealing [M]. Vienna: InTech Publisher, 2008: 77 - 88.

[15] DALVI N, KUMAR R, SOLIMAN M. Automatic wrappers for large scale Web extraction [C]// Proceedings of the 37th International Conference on Very Large Databases. New York: VLDB Endowment, 2011: 219 - 230.

[16] XIAO S, HE Y. Approach of Chinese event IE based on verb argument structure [J]. Computer Science, 2012, 39(5): 161 - 164. (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.)

(下转第2331页)

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

MAE 变化曲线 CF(1)、CF(2)、CF(3),在相同的最近邻取值下,当权重不同时,得出的 *MAE* 变化曲线不同;但本文算法的 *MAE* 值明显小于传统的协同过滤算法,且权重近似平均时,其 *MAE* 值达到最小值。在冷启动数据集下,本文方法可有效地提高推荐的准确率。

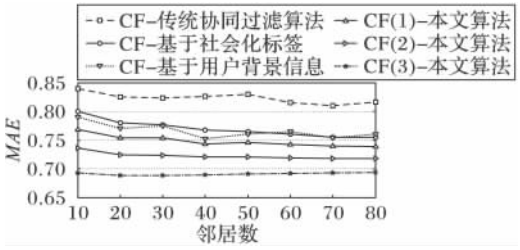

图 2　推荐算法的 *MAE* 比较

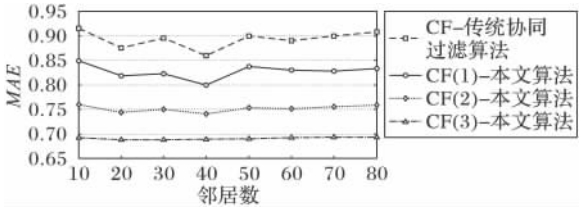

图 3　冷启动数据集下算法的 *MAE* 比较

5　结语

随着协同过滤算法在电子商务中的广泛应用,如何提高其推荐精度已成为非常重要的研究问题。本文提出的结合用户背景信息和社会化标签的协同过滤算法,能够有效地提高算法的推荐精度和数据稀疏及冷启动问题。由于电子商务平台与社交网络平台是独立运行的,本文验证未能使用综合电子商务和社交网络的标准数据集,而是先计算两个数据集用户评分之间的相似性得出一个用户相似类,再将两个数据集整合成本文需要的数据集来进行验证,存在一定的不足。未来工作是集成电商和社交网络平台数据,对本文的工作给予进一步的验证。

参考文献:

[1] HUANG L, LI D. A review of information recommendation in social media [J]. CAAI Transactions on Intelligent Systems, 2012, 7(1): 1 – 8. (黄立威,李德毅. 社交媒体中的信息推荐[J]. 智能系统学报,2012,7(1):1 – 8.)

[2] RICCI F, ROKACH I, SHAPIRA B, *et al*. Recommender systems handbook [M]. Berlin: Springer, 2011: 145 – 186.

[3] YU H, LI Z. A collaborative filtering recommendation algorithm based on forgetting curve [J]. Journal of Nanjing University: Natural Sciences, 2010, 46(5): 520 – 527. (于洪,李转运. 基于遗忘曲线的协同过滤推荐算法[J]. 南京大学学报:自然科学版,2010,46(5):520 – 527.)

[4] FENG Y, LI J, XU H, *et al*. Collaborative recommendation method improvement based on social network analysis [J]. Journal of Computer Applications, 2013, 33(3): 841 – 844. (冯勇,李军平,徐红艳,等. 基于社会网络分析的协同过滤推荐方法改进[J]. 计算机应用,2013,33(3):841 – 844.)

[5] WANG H, MU Y, YU X. Resource recommendation in social tagging system based on collaborative matrix factorization [J]. Application Research of Computers, 2013, 30(6): 1739 – 1741. (王海雷,牟雁超,俞学宁. 基于协同矩阵分解的社会化标签系统的资源推荐[J]. 计算机应用研究,2013,30(6):1739 – 1741.)

[6] SYMEONIDIS P, TIAKAS E, MANOLOPOULOS Y. Product recommendation and rating prediction based on multi-modal social networks [C]// Proceedings of the 5th ACM Conference on Recommender Systems. New York: ACM Press, 2011: 61 – 68.

[7] JIA D, ZENG C, PENG Z, *et al*. A user preference based automatic potential group generation method for social media sharing and recommendation [J]. Chinese Journal of Computers, 2012, 35(11): 2381 – 2391. (贾大文,曾承,彭智勇,等. 一种基于用户偏好自动分类的社会媒体共享和推荐方法[J]. 计算机学报,2012,35(11):2381 – 2391.)

[8] ZHAO Y, DONG J, DONG J. Tag recommendation for blogs based on social tagging [J]. Computer Engineering and Design, 2012, 33(12): 4609 – 4613. (赵亚楠,董晶,董佳亮. 基于社会化标注的博客标签推荐方法[J]. 计算机工程与设计,2012,33(12):4609 – 4613.)

[9] JIA D, ZHANG F. A collaborative filtering recommendation algorithm based on double neighbor choosing strategy [J]. Journal of Computer Research and Development, 2013, 50(5): 1076 – 1084. (贾冬艳,张付志. 基于双重邻居选取策略的协同过滤推荐算法[J]. 计算机研究与发展,2013,50(5):1076 – 1084.)

[10] ZHUANG J, WANG M, YE M. Agriculture information recommendation system based on content filtering [J]. Computer Engineering, 2012, 38(11): 38 – 41. (庄景明,王明文,叶茂盛. 基于内容过滤的农业信息推荐系统[J]. 计算机工程,2012,38(11):38 – 41.)

[11] XIA N, SU Y, QIN H, *et al*. Method for personalized user profiling in social tagging systems [J]. Journal of Computer Applications, 2011, 31(6): 1667 – 1670. (夏宁霞,苏一丹,覃华,等. 社会化标签系统中个性化的用户建模方法[J]. 计算机应用,2011,31(6):1667 – 1670.)

[12] ADOMAVICIUS G, TUZHILIN A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(6): 734 – 749.

(上接第 2316 页)

[17] YANG S, LIN H, HAN Y. Automatic data extraction from template-generation Web pages [J]. Journal of Software, 2008, 19(2): 209 – 223. (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.)

[18] ETZIONI O, FADER A, CHRISTENSEN J, *et al*. Open information extraction: the second generation [C]// Proceedings of the 22nd International Joint Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2011: 3 – 10.

[19] SIMÕES G, GALHARDAS H, GRAVANO L. When speed has a price: fast information extraction using approximate algorithms [C]// Proceedings of the 39th International Conference on Very Large Databases. New York: VLDB Endowment, 2013: 1462 – 1473.

[20] MAYS E, DAMERAU F J, MERCER R L. Context based spelling correction [J]. Information Processing and Management, 1991, 27(5): 517 – 522.

[21] Princeton University. WordNet: a lexical database for English [EB/OL]. [2013-09-10]. http://wordnet.princeton.edu/.

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net