·信息资源管理·

基于PLSA模型的Web页面语义标注算法研究

王云英

(湘南学院图书馆 郴州 423000)

摘 要 高效的Web页面语义标注方法是提高Web信息资源利用效率和知识创新的关键。针对当前Web页面语义标注方法存在的问题和Web页面表现出的结构特征和文本特征及其主题分布规律,设计了基于PLSA主题模型的Web页面语义标注算法。该算法分别对Web页面的结构特征和文本特征构建独立的PLSA主题模型,采用自适应不对称学习算法对这些独立的PLSA主题模型进行集成和优化,最终形成新的综合性的PLSA主题模型进行未知Web页面的自动语义标注。实验结果表明,该算法能够显著提高Web页面语义标注的准确率和效率,可以有效地解决大规模Web页面语义标注问题。

关键词 语义标注 PLSA模型 潜在语义主题 标注算法 Web页面

中图分类号 G350 **文献标识码** A **文章编号** 1002-1965(2013)01-0141-04

Research on Web-page Semantic Annotation Algorithm Based on PLSA Model

Wang Yunying

(Library of Xiangnan University, Chenzhou 423000)

Abstract Efficient web-page semantic annotation is the key point to improve the efficient use of web information resource and knowledge innovation. This paper designs a web-page semantic annotation algorithm based on PLSA model according to the structural feature and the text feature existing in web-page to solve the problems of traditional annotation technology. The proposed algorithm constructs PLSA topic model for structural feature and text feature respectively, adopts an adaptive asymmetric learning approach to the integration and optimization of the PLSA model, forms a new comprehensive PLSA model to semantically annotate the unknown web pages automatically. Experimental results demonstrate that this algorithm dramatically improves the accuracy and efficiency of web-page semantic annotation, and can solve the problem of large-scale web-page annotation effectively.

Key words semantic annotation PLSA model latent semantic topics annotation algorithm web pages

0 引 言

Web页面语义标注是实现网络信息资源语义组织和深层利用的基础和核心内容,是构建语义Web关键。自20世纪90年代中期起,引起不同领域的专家学者的广泛关注和积极探索,如Taylor[1]利用句法解析方法和路径分析技术对Web页面信息进行重新描述和标注,在此基础上实现基于概念的生物文本信息抽取。Bertini等[2]针对Web页面的复杂性和多样性,综合运用基于事件和基于对象相融合的信息抽取方法实现对Web页面的语义标注和动态更新。Heath等[3]根据网络问卷调查得出网络信息资源的检索效率和高质量的语义标注呈现正相关,对Web页面进行标注不仅需要关注页面的文本信息,还需要结合页面的结构等信息进行综合标注才能取得理想的结果。Sanchez等[4]利用领域本体和主动学习方法对Web页面内容进行学习,通过对大量训练页面的学习和归纳构建标注模型,然后利用该模型实现基于Web页面内容的自动标注。丁艳辉等[5]提出的基于集成学习和二维关联边条件随机场的Web数据语义自动标注方法,利用训练页面的统计特征和预先抽取的先验知识构造多分类器,根据Dempster合成法则对各分类器的分类结果进行归并,实现训练页面中属性标签和数据元素的识别;在此基础上利用二维边条件随机场模型进行Web数

收稿日期:2012-07-10 修回日期:2012-08-28

作者简介:王云英(1966-),女,馆员,研究方向:图书馆管理和文本标注。

据元素间依赖关系的建模,实现数据元素的自动语义标注。张玉峰等[6]提出的基于数据挖掘的 Web 文本语义分析与标注研究主要利用领域本体和数据挖掘技术实现 Web 页面中文本信息的语义挖掘和自动获取,在此基础上运用聚类方法实现 Web 实例分析与标注、运用关联挖掘和分类方法实现 Web 实例间关系的分析和标注。崔红等[7]提出的基于机器学习的文档语义标注方法,利用有监督的机器学习方法从文本中提取领域规则,实现基于先导词算法的语义标注,解决现有标注系统通用性差的问题。这些研究在一定程度上推动了 Web 页面语义标注的研究和深化,使得 Web 页面所蕴含的信息资源从机器可读向机器可理解迈进,促使机器和人类能够更好地协同和交互。但是,这些研究存在的突出问题在于标注过程中非常重视 Web 页面中的文本信息,而对 Web 页面的结构、布局等视觉信息的重视程度不够,使得现有的标注模型和方法的利用效果不理想,且在标注过程中需要预先手工标注大量训练页面,造成标注结果的可靠性也受到影响。

本文根据目前 Web 页面语义标注存在问题的基础上,通过分析 Web 页面所具备的特征及特征所表现的主题分布规律,设计了基于 PLSA 主题模型的 Web 页面语义标注算法(Web-page Semantic Annotation Algorithm based on PLSA Model, WSAA-PLSAM)。该算法将 Web 页面所具备的每种特征表示为不同的 PLSA 主题模型,通过一种自适应的不对称学习算法进行不同 PLSA 主题模型的集成和融合,形成统一的综合语义标注空间,全面、准确地获取 Web 页面所蕴含的语义知识,提高 Web 页面语义标注的质量和效率。

1 PLSA 模型

PLSA 采用概率模型进行"文本-潜在语义-关键词"三者关系的表达,是 Hoffman[8] 根据利用 LSA 进行文本语义分析建模过程中存在的问题而提出的一种改进模型。其基本原理是:设文档集合 $D=\{d_1,d_2,\cdots,d_i,\cdots,d_N\}$,$D$ 中所有关键词组成的特征集合 $F=\{f_1,f_2,\cdots,f_j,\cdots,f_M\}$,$D$ 中可能包含的潜在语义主题集合 $Z=\{z_1,z_2,\cdots,z_k,\cdots,z_P\}$。对于给定的潜在语义主题 z_k,PLSA 模型假设每个特征关键词 f_j 和其所属的文档 d_i 相互独立,即假设关键词与文档之间、潜在语义主题与文档或关键词之间的概率均服从条件独立。在此假设的基础上,相应的联合分布概率可以表示为:

$$P(d_i,z_k,f_j)=P(d_i)P(z_k|d_i)P(f_i|z_k)$$

利用贝叶斯公式对潜在语义主题 z_k 进行边缘处理可以得到 PLSA 联合概率模型:

$$P(d_i,f_j)=P(d_i)\sum_{K=1}^{K}P(z_k|d_i)P(f_i|z_k)$$

其中,$P(f_i|z_k)$ 表示潜在语义主题在关键词上的分布概率,$P(z_k|d_i)$ 表示文档在潜在语义主题上的分布概率。

PLSA 模型利用 EM(Expectation Maximization, EM)算法进行模型优化和拟合,即利用随机数初始化后,交替执行 E 操作和 M 操作进行迭代计算。E 操作的主要目的是计算任意 (d_i,f_j) 所产生的潜在语义主题 z_k 的先验概率,其计算方法如下:

$$P(z_k|d_i,f_j)=\frac{P(z_k|d_i)P(f_j|z_k)}{\sum_{K=1}^{K}P(z_k|d_i)P(f_j|z_k)}$$

M 操作主要利用 E 操作获取的 $P(z|d,f)$ 进行 $P(z|d)$ 和 $P(f|z)$ 的迭代更新:

$$P(z_k|d_i)=\frac{\sum_{j=1}^{M}n(d_i,f_j)P(z_k|d_i,f_j)}{\sum_{j=1}^{M}n(d_i,f_j)},$$

$$P(f_j|z_k)=\frac{\sum_{i=1}^{N}n(d_i,f_j)P(z_k|d_i,f_j)}{\sum_{m=1}^{M}\sum_{i=1}^{N}n(d_i,f_m)P(z_k|d_i,f_m)}$$

当期望值 $E(L)$ 的边际增长小于预设阀值时停止迭代过程,此时获取的结果认为是最优解,从而得到理想的 $P(f_i|z_k)$ 和 $P(z_k|d_i)$ 的分布情况。$E(L)$ 的计算公式为:

$$E(L)=\sum_{i=1}^{N}\sum_{j=1}^{M}n(d_i,f_j)\log P(d_i,f_j)$$

2 基于 PLSA 模型的 Web 页面语义标注算法

Web 页面所具备的特征可以分为两类[9]:文本特征和结构特征。文本特征主要涉及 Web 页面中存在的大量文本信息,通过其自身的内容表达相应的主题;结构特征主要涉及 Web 页面布局、链接结构和文档结构等信息,对 Web 页面所要表达的内容具备很重要的影响作用。传统的 Web 页面语义标注方法往往忽略 Web 页面所表现出来的结构特征,仅仅重视其文本特征,造成相应的标注方法在使用的过程中效果和质量均不理想。本文综合利用 Web 页面所表现出来的文本特征和结构特征,针对每种特征构建 PLSA 主题模型,利用自适应不对称学习算法进行不同 PLSA 主题模型的集成和优化,形成综合性的统一主题分布。

设 Web 页面的训练集 $D=\{(d_1,c_1),(d_2,c_2),\cdots,(d_N,c_N)\}$,$S_D=\{s_1,s_2,\cdots,s_N\}$ 表示 Web 页面训练集的结构特征集,$C=\{C_1,C_2,\cdots,C_N\}$ 表示 Web 页面训练集的文本特征集,则基于 PLSA 模型的 Web

页面语义标注算法 WSAA-PLSAM 可以描述为:

2.1 **训练阶段** 训练阶段的主要任务是获取文本信息和结构信息之间的相互依存概率的分布,其算法描述如下:

Step1 对于任意 Web 页面 d_i 进行结构解析和文本信息提取,获取表达结构信息的特征向量 $s(d_i)$ 和表达文本信息的特征向量 $c(d_i)$;其中,采用基于规则和树对齐算法[10]获取网页结构特征 $s(d_i)$,对文本信息进行分词处理后,利用 TFIDF 公式进行权值计算,选取前 k 个关键词作为文本信息的特征概念构成特征向量 $c(d_i)$;

Step2 利用获取的 $s(d_i)$ 和 $c(d_i)$ 分别构建 PLSA 主题模型,获取结构信息和文本信息相对应的主题分布 $P_s(s|\alpha)$、$P_s(\alpha|d)$ 和 $P_c(c|\beta)$、$P_c(\beta|d)$;其中,α 和 β 表示主题;

Step3 根据结构特征和文本特征对理解 Web 页面的重要程度确定两个 PLSA 主题模型集成和优化的权重,利用自适应学习算法[11]实现 PLSA 主题的集成,得到新的综合性主题分布 $P(z|d_i)$:

$$P(z_k \mid d_i) = \begin{cases} w_{si}P_s(\alpha_\varphi \mid d_i), \varphi = 1,2,\cdots,m \\ w_{ci}P_c(\beta_{\varphi-m} \mid d_i), \varphi = m+1, m+2, \cdots, m+n \end{cases}$$

其中,w_{si} 和 w_{ci} 分别表示结构特征和文本特征在 Web 页面 d_i 中的权重,m 和 n 分别表示结构特征和文本特征对应的主题数目,$\varphi = m + n$;

Step4 根据集成后的主题分布 $P(z|d_i)$,优化 $P(s|z)$ 和 $P(c|z)$

2.2 **标注阶段** 标注阶段主要利用训练阶段的运行结果,对新的 Web 页面进行自动语义标注,其算法描述如下:

Step5 对于任意新的 Web 页面 d_{new},执行 Step1;

Step6 利用 $s(d_{new})$ 和 $c(d_{new})$,执行训练算法获取 $P(s_{new}|z)$ 和 $P(c_{new}|z)$,从而得到该 Web 页面 d_{new} 的主题分布 $P(z|d_{new})$;

Step7 计算结构特征关键词和文本特征关键词的后验概率:

$$P(s|d_{new}) = \sum_{n=1}^{N} P(s|z_k)P(z_k|d_{new}),\ P(c|d_{new}) = \sum_{n=1}^{N} P(c|z_k)P(z_k|d_{new})$$

Step8 选取后验概率满足预设阀值的若干特征关键词标注 d_{new}。

通过上述操作,就可以通过对少量训练样本的训练和学习,利用 PLSA 主题模型从 Web 页面的结构特征和文本特征两个方面获取相应潜在主题分布,从而有效地实现新 Web 页面的自动语义标注。

3 实验结果分析

为了验证本文设计的基于 PLSA 模型的 Web 页面语义标注算法 WSAA-PLSAM 的有效性,通过收集多个领域的真实数据集对所提出的算法进行实验分析。

3.1 **数据来源** 实验数据主要来自三个不同的领域:手机数据集(Mobile Dataset,MD)、图书数据集(Book Dataset,BD)、论文数据集(Paper Dataset,PD)。其中,MB 数据集主要从在线手机网站上收集的 400 个 Web 页面构成,经过手工标注后,随机抽取 300 个页面作为训练集,剩余的 100 个页面作为测试集。BD 数据集主要从在线图书网站上收集的 400 个 Web 页面构成,经过手工标注后,随机抽取 200 个页面作为训练集,剩余的 200 个页面作为测试集。PD 数据集主要从中国知网、万方等文献数据库中收集的 400 个 Web 页面构成,经过手工标注后,随机抽取 100 个页面作为训练集,剩余的 300 个页面作为测试集。

3.2 **测评指标** 本文采用检验 Web 页面语义标注结果常用的测评指标:查全率(Recall,R)、查准率(Precision,P)、测度 F_1 值,其计算公式为:

$$R = \frac{B}{A}, P = \frac{B}{B+C}, F_1 = \frac{2PR}{P+R}$$

其中,A 表示待标注的 Web 页面数,B 表示正确标注的 Web 页面数,C 表示标注错误的 Web 页面数。

3.3 **实验结果与分析** 本文在上述实验数据的基础上,运用本文提出的 WSAA-PLSAM 算法和只利用结构信息或文本信息构建 PLSA 模型的标注算法进行对比实验分析。实验结果如表 1、表 2、表 3 所示。

表 1 只利用 Web 页面结构信息的标注结果

标注算法	P(%)	R(%)	F_1(%)
MD	67.54	65.64	66.58
BD	54.73	52.96	53.56
PD	45.49	43.89	44.68

表 2 只利用 Web 页面文本信息的标注结果

标注算法	P(%)	R(%)	F_1(%)
MD	82.72	79.98	81.33
BD	71.09	68.65	69.85
PD	58.48	56.75	57.61

表 3 本文提出的综合利用 Web 页面的结构信息和文本信息的标注结果

标注算法	P(%)	R(%)	F_1(%)
MD	86.83	84.79	85.81
BD	82.68	80.55	81.59
PD	79.36	76.98	78.15

通过上述实验结果可以看出,本文提出的基于 PLSA 模型的 Web 页面语义标注算法 WSAA-PLSAM

的整体性能明显优于基于文本信息的标注方法和基于结构信息的标注方法。这是因为 WSAA-PLSAM 算法通过对 Web 页面的结构信息和文本信息分别构建不同 PLSA 主题模型来表示其各自表达的潜在语义主题,然后根据其对整个页面标注的影响程度,利用自适应不对称学习算法进行各个潜在语义主题的集成和优化,形成统一的综合性的潜在语义主题,在此基础上实现 Web 页面的标注过程,从而使得该算法的标注结果明显优于单独使用结构信息或文本信息的标注算法。而仅仅利用 Web 页面所包含的结构信息或文本信息构建 PLSA 主题模型进行 Web 页面语义标注,由于所利用的信息缺乏全面性,使得所构建的潜在语义主题在表达过程中出现偏差,导致标注结果不理想。相比较而言,由于 Web 页面中的文本信息的重要程度高于结构信息,所有单纯依靠文本信息进行标注的结果要优于依靠结构信息的标注结果。

此外,还可以看出,Web 页面语义标注的结果还与训练样本的规模存在一定的正相关,即训练样本越多,构建的 PLSA 主题模型的精度和性能越好,在测试过程中取得的效果也越好,如本文选取的数据集中,MD 数据集训练样本与测试样本的比例为 3:1,BD 数据集训练样本与测试样本的比例为 1:1,PD 数据集训练样本与测试样本的比例为 1:3,在使用过程中,无论采用那种算法进行标注,MD 数据集的标注结果要明显优于 BD 数据集和 PD 数据集。但总的来说,只采用结构信息进行 PLSA 主题模型的构建和标注算法对训练样本的依赖性最强,减少训练样本会导致最终标注结果非常不理想。而融合结构信息和文本信息构建综合性的 PLSA 主题模型的标注算法即使在很少的训练样本的情况下,也能取得较优的标注效果。

4 结束语

随着 Web2.0 和语义 Web 的快速发展和普及,Web 页面集成了众多领域的大量有价值信息,通过有效地 Web 页面语义标注技术和方法对来自不同网站的海量 Web 页面进行标注,是提高网络信息资源利用效率和知识挖掘与创新的关键。本文在分析 Web 页面所具备的结构特征和文本特征的基础之上,设计了基于 PLSA 模型的 Web 页面语义标注算法。该算法利用现有的 PLSA 主题模型,从 Web 页面结构特征和文本特征两个方面构建 PLSA 主题模型,获取结构特征和文本特征所隐含的潜在语义主题;根据结构特征和文本特征对 Web 页面主题分布的影响程度,采用自适应非对称学习算法进行结构特征和文本特征的潜在语义主题集成和优化,形成新的综合性潜在语义主题,在此基础上实现 Web 页面的语义标注。通过在真实数据集上进行的实验表明,综合利用 Web 页面的结构信息和文本信息能够更全面准确地反映 Web 页面所隐含的潜在语义主题,能够显著提高 Web 页面语义标注的质量和效率,可以有效地解决大规模 Web 页面语义标注问题。

参 考 文 献

[1] Tayor A. A Extracting Knowledge from Biological Descriptions [C]. Proceedings of 2nd International Conference on Building and Sharing Very Large-Scale Knowledge Bases,1995:114-119

[2] Bertini M, Cucchiara R, Prati A. An Integrated Framework for Semantic Annotation and Adaptation[J]. Multimedia Tools and Application,2005, 26(3): 345-363

[3] Heath T, Christian B. Semantic Annotation and Retrieval: Web of Data[M]. VerlagBerlin Heidelberg: Springer,2011:201-204

[4] Sanchez D, Isern D, Millan M. Content Annotation for the Semantic web: an Automatic Web-based Approach[J]. Knowledge and Information Systems, 2011, 27(3): 393-418

[5] 丁艳辉,李庆忠,董永权,等. 基于集成学习和二维关联边条件随机场的 Web 数据语义标注方法[J]. 计算机学报,2010,33(2):267-278

[6] 张玉峰,蔡皎洁. 基于数据挖掘的 Web 文本语义分析与标注研究[J]. 情报理论与实践,2010,33(2):85-88

[7] 崔 红,段宇锋,郦 芳. 基于机器学习的生物多样性英文文档语义标注研究[J]. 图书情报知识,2011,32(2):73-77

[8] Hofmann T. Unsupervised Learning by Probabilistic Latent Semantic Analysis [J]. Machine Learning, 2011, 42(1): 177-196

[9] 郑庆华,刘 均,田 锋,等. Web 知识挖掘:理论、方法与应用[M]. 北京:科学出版社,2010:114-116

[10] 朱 凯. 基于结构和视觉特征的网页信息抽取计算的研究与实现[D]. 杭州:浙江大学计算机科学与技术学院,2008

[11] 李志欣,施智平,李志清,等. 融合语义主题的图像自动标注[J]. 软件学报,2011,22(4):801-812

(责编:白燕琼)