网络食品安全的歧义性消解算法

刘金硕[1] 邓莹莹[2] 邓 娟[2]

(武汉大学计算机学院 武汉 430072)[1] (武汉大学国际软件学院 武汉 430072)[2]

摘 要 以网络食品安全信息为研究对象,旨在提出一个能够解决食品安全领域专有名词指代不明的歧义消解算法。文中采用的歧义消解算法是在改进的TF-IDF特征选择算法的基础上,结合了隐含马尔可夫模型(HMM)和SVM分类器,从而实现专有名词的歧义消解。提出了一个在TF-IDF的基础上增加两个加权因子的特征提取算法LN-TF-IDF。实验表明,以202831条文本实验所得的准确率和召回率的调和平均值F1值为评价标准,设计的基于改进TF-IDF的食品安全领域歧义消解算法的效果比基于传统TF-IDF的歧义消解算法平均提升了7.31%,且在不同时间抓取的实验数据集下,本算法的效果也相对稳定。

关键词 食品安全,歧义消解,隐含马尔可夫模型,TF-IDF,支持向量机

中图法分类号 TP391 **文献标识码** A

Disambiguation Algorithm Design and Implementation of Food Safety Issues in Network

LIU Jin-shuo[1] DENG Ying-ying[2] DENG Juan[2]

(Computer School, Wuhan University, Wuhan 430072, China)[1]

(International School of Software, Wuhan University, Wuhan 430072, China)[2]

Abstract The article aimed to put forward a disambiguation algorithm which can correctly classify the unknown terms, based on the food safety information in network. The disambiguation algorithms used in this paper combines the hidden Markov model(HMM) and SVM classifier to achieve terminology disambiguation, based on the improved TF-IDF feature selection algorithm. This paper proposed a new feature extraction algorithm LN-TF-IDF with two additional weighting factors on traditional TF-IDF. Experiments show that, the improved TF-IDF disambiguation algorithm designed in the field of food safety enhances the effect of disambiguation by average 7.31% on the 202831 texts. It was compared with the traditional TF-IDF text feature selection algorithm, with the F-measure as evaluation criteria. At the same time, the effect of the algorithm is relatively stable on different experimental data sets obtained from different time.

Keywords Food safety, Disambiguation, HMM, TF-IDF, SVM

1 引言

歧义消解就是指在一个命名实体指称项可对应多个命名实体概念时,确定该指称项实际指向的命名实体概念的过程[1]。当前,网络舆情信息中关于食品安全问题的信息在逐步增多。食品安全领域中有许多专业名词,比如化学成分名、食品名等。随着科技的发展,食品的种类及其含有的化学成分越来越多,在给这些食品或者化学成分命名时,不可避免会出现重名的情况,加上这些物质可能存在别名,就使得产生歧义的可能增加。消除这些专有名词的指代歧义,对网络食品安全问题的快速监测有着重要的意义。

国内外关于不同类型的歧义消解的研究成果已有很多。复旦大学的龚凌晖[1]以条件随机场为基本框架,设计并实现了一种中文命名实体识别的系统,并在此基础上基于潜在语义分析实现了对命名实体的歧义消解;北京大学的何径舟[2]提出了使用特征选择和最大熵模型结合的词义消歧方法;Pedersen T[3]发表在北美计算语言学协会的文章,采用决策树的方法完成词义消歧;Johannes Hoffart, Mohamed Amir Yosef等发表在自然语言处理的经验方法学报上的文章[4],采用结合实体的先验概率、上下文之间的相似性和候选人的实体以及所有提到的候选实体之间的连贯性,来构建有向图,从而消除歧义;哈尔滨工业大学的戴祥鹰[5]利用一种自底向上的凝聚层次聚类方法,在迭代的过程中,采用全连接的方式计算簇间相似度实现文本聚类,最终消除同名情况的人名歧义;北京大学的韩伟[6]提出了一个基于关键词-属性的模型,结合相似度计算级聚类的人名消歧方法。其他可用于消歧的方法如PageRank[7]、隐含马尔科夫模型[8]、支持向量机[9]等也有相应的成果。

虽然现有的歧义消解研究成果已经有很多,但食品安全领域的歧义消解在国内的研究相对较少,已有的研究成果中有可借鉴的方法,但总体针对性不够强。所以本文提出了根据网络食品安全领域中的专有名词可能存在的歧义来设计歧

本文受国家自然科学基金项目(61303214)资助。

刘金硕 副教授,主要研究方向为文本挖掘、网络内容安全等,E-mail:liujinshuo@whu.edu.cn;邓莹莹 女,硕士生,主要研究方向为文本挖掘;邓 娟 博士,副教授,主要研究方向为模式识别、图像处理和高性能计算,E-mail:dengjuan@whu.edu.cn(通信作者)。

义消解算法。

本文设计的算法的整体思路为：采用 HMM 模型和改进的 TF-IDF 算法分别得到对应文本中食品领域的专有名词集及特征词集和特征向量集，从而为消歧过程提供原始输入，并利用 SVM 分类器实现专有名词的歧义消解。其中，本文的研究重点在于 TF-IDF 文本特征选择算法的改进。

本文第 1 节介绍背景；第 2 节介绍本文使用到的相关理论和技术；第 3 节详细介绍本文的消歧算法；第 4 节介绍算法的实验和结果分析；最后总结全文。

2 相关理论与技术

2.1 隐含马尔可夫模型

隐含马尔可夫模型是马尔科夫链的一个扩展，是一个双重随机过程。在这个过程中具体的状态序列不可知，只知其状态转移概率，即模型的状态转换过程是不可观察的，而可观察事件的随机过程是隐蔽的状态转换过程的随机函数[10]。

2.2 TF-IDF 文本特征选择算法

TF-IDF(Term Frequency-Inverse Document Frequency，词频-逆文本频率)是一种用于信息搜索和信息挖掘的常用加权技术[11]。TF-IDF 算法通过计算 TF * IDF 来衡量文本集中词条的权重。TF-IDF 是基于评估函数的特征选择算法中比较常见的一种文本特征选择及权重评估的方法。

2.3 支持向量机

支持向量机(Support Vector Machine，SVM)是一种机器学习方法，是在统计学习理论的基础上对线性分类器提出的一种非常有潜力的分类技术[9]。它在解决小样本、非线性及高维模式识别中表现出许多特有的优势，并能够推广应用到函数拟合等其他机器学习问题中[12]。

3 基于 TF-IDF 的消歧算法

本节主要介绍本文的食品安全领域歧义消解算法的详细设计。本算法是在改进的 TF-IDF 特征选择算法的基础上，采用训练后的 HMM 识别和提取专有名词，改进的 TF-IDF 算法进行文本的特征选择，然后利用 SVM 分类器将产生歧义的专有名词从其所有可能的含义列表中分类到正确的含义中，从而完成食品安全领域专有名词的歧义消解。

3.1 算法总体设计

算法流程如图 1 所示。

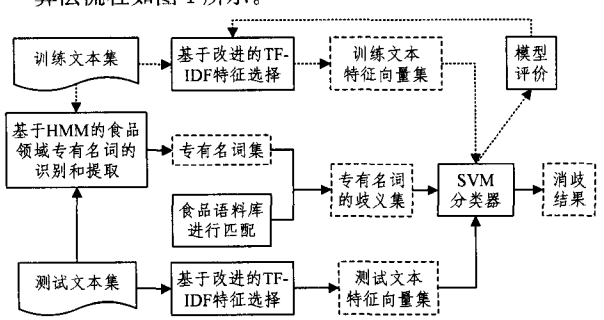

图 1 歧义消解算法流程

其中，虚线框表示经过相应过程的处理后产生的结果，即比如对数据集进行基于 HMM 的化学名词和食品名称的识别与提取后，会产生该数据集的专有名词集。

本文的算法主要包括以下步骤。

1. 数据集处理：包括消歧流程图中“基于 HMM 的专有名词的识别与提取”和“基于 LN-TF-IDF 的文本特征提取”这两个模块。基于 HMM 的专有名词的识别与提取采用鲍姆-韦尔奇算法来进行 HMM 模型训练，并完成提取过程，得到数据集的“专有名词集”，并通过与语料库进行匹配，可得到相应专有名词的歧义集；基于 LN-TF-IDF 的文本特征提取采用本文所改进的 TF-IDF 算法完成，得到数据集的文本特征特征向量集。

2. SVM 分类器训练：不断调整文本特征词选取的阈值，不断迭代文本特征提取过程，直到找到最优的文本特征词集为止，从而训练得到一个当前 SVM 分类算法情况下的最优 SVM 分类器。

3. SVM 分类器消歧：采用 HMM 模型和改进的 TF-IDF 算法分别得到对应文本中食品领域的专有名词集和特征词集，为消歧过程提供原始输入，并利用 SVM 分类器实现专有名词的歧义消解。

其中，本文的研究重点在于对文本特征选择算法的改进，以为食品安全领域的文本消歧提供基础。

3.2 基于 HMM 的名词识别与提取

本文采用著名的鲍姆-韦尔奇算法来实现无监督的模型训练，实验结果表明，该训练方法有很好的效果。其训练过程的主要步骤是：

1. 定义模型参数状态转移概率 $P(O_t \mid S_t)$ 和输出概率 $Q(S_t \mid S_{t-1})$：

$$P(O_t \mid S_t)=\frac{P(O_t,S_t)}{P(S_t)} \tag{1}$$

$$Q(S_t \mid S_{t-1})=\frac{P(S_{t-1},S_t)}{P(S_{t-1})} \tag{2}$$

其中，O_t 表示当前观测到的序列，即根据分词结果得到的文本中所有的有意义的名词数据集(为保证可观测序列的准确性，会对分词的结果进行人工标注和筛选)；S_t 表示隐含的状态序列，即文本中名词对应的词的归属集(比如：化学名词、食品名、机构名等)。

2. 根据大量的文本找到一组初始的模型参数，即对应的转移概率和生成概率。该模型参数能够产生输出序列 O，并把该初始模型记为 M_0，根据该模型计算产生输出序列(即具体观测信号)的概率值 $P(O \mid M_0)$，同时记录该模型产生输出序列的所有可能路径以及对应的概率，相当于给式(1)和式(2)提供了标注的训练数据。

3. 根据式(1)和式(2)计算出新的模型参数，这也是更准确的模型的输入，把该模型称为 M_1，不断重复上述过程，直至找到最优的模型 M_b，将该模型作为最终的专有名词识别的工具。

整个 HMM 模型的训练通过不断迭代优化，得到局部的最优解，初始模型参数的选择会影响模型的最终识别效果。利用训练得到的最终 HMM 模型从分词的数据集中识别并提取食品领域的专有名词，然后将专有名词与食品语料库进行匹配，则可得到该专有名词的含义列表及歧义集。

3.3 基于 TF-IDF 的特征选择算法的改进

借鉴 TF-IDF 的算法思想，通过上下文分析及关键特征词的邻域组合词分析文本歧义域，对 TF-IDF 算法增加评估因子进行优化，得到评估函数 LN-TF-IDF(其中 L 表示特征词词长，N 表示是否为邻域词，TF 表示词频，IDF 表示词的逆话题频率)，从而确定食品安全领域专有名词的消歧算法实现。

以下为消歧评估算法的具体内容：

$$TF(t_k,d)=\frac{tf(t_k,d)}{\sum_{t_k\in d} tf(t_k,d)} \tag{3}$$

$$IDF_{t_k}=\frac{\log N}{\log n_{t_k}} \tag{4}$$

其中，$tf(t_k,d)$为特征词 t_k 在文本 d 中的频数，N 表示话题集，n_{t_k} 表示话题集中含有 t_k 的话题数量。

在文本的特征提取过程中，一般来说，词长越长，其专指性越强，而概括性越弱，所以长词应该具有相对更高的权重，所以设置加权因子 L：

$$L(t_k,d)=\frac{1}{l+1} \tag{5}$$

其中，l 表示特征词的词长。

在特征提取的过程中，一般来讲，如果一个特征词与较多的特征词是邻域词，那么这个特征词与文本主题的相关性应该更大，它所具有的权重也就应该相对较高。经过大量抽样数据统计，可以得出文本中一个特征词条与其他特征词条是否是邻域词的权重计算，$Y_k=1.0$ 表示是邻域词的权重，而 $Y_k=0.1$ 则表示不是邻域词；设 T_k 为特征词 t_k 在相应位置出现的次数，则它的邻域加权因子为：

$$N(t_k,d)=\frac{\sum(Y_k*T_k)}{\sum T_k} \tag{6}$$

整合式(3)－式(6)可得到最后得出的特征词的权重值：

$$W_k=\frac{tf(t_k,d)}{\sum_{t_k\in d} tf(t_k,d)}*\frac{\log N}{\log n_{t_k}}*\frac{l}{l+1}*\frac{\sum(Y_k*T_k)}{\sum T_k} \tag{7}$$

通过该算法对给定文本的特征词做处理，得到特征词的权重值，当该值大于设定的阈值时，把它加入到特征集，作为分类器的输入。在本文的处理过程中，通过不同的阈值设定，不断迭代训练过程，直到找到最优的文本特征词集为止。将采用此算法得到的特征词集作为训练 SVM 分类器的输入，进而得到效果更优的 SVM 分类器。

3.4 利用 SVM 分类器的歧义消解过程

本文采用构造 SVM 分类器来实现最后的消歧步骤，充分利用 SVM 分类的优点，根据由 HMM 算法得到的歧义词含义列表，对给定的专有名词在当前文档中划分类别，从而实现消歧过程。本文将在训练文本集上通过 HMM 模型和改进的 LN-TF-IDF 算法分别得到的对应文本的食品领域的专有名词集和特征词集作为训练 SVM 分类器的输入(具体流程参考图 1 中与 SVM 分类器相关的流程图示)。

SVM 的训练过程包含了模型评价模块，模型评价准则以召回率和准确率的调和平均数 F1 为评价因子，然后将结果反馈给文本特征词选择过程，进而调整阈值，不断迭代该过程，从而得到一个当前 SVM 分类算法情况下的最优 SVM 分类器。

分类器训练好之后，采用同样的方法对目标文本集进行处理，得到的数据集作为 SVM 分类器的输入，分类器对输入的数据进行分类操作，得到目标歧义词的确切含义，从而实现消歧的过程。

4 实验与结果分析

在 Windows 7 64 位的操作系统中，4GB 内存，双核处理器的硬件条件下，在 MyEclipse10 中，用 java 实现 LN-TF-IDF 特征选择算法，这是本文歧义消解算法的创新点，也是重点。HMM 和 SVM 的研究已经很成熟，本文并未做改进。本文就采用已经实现的鲍姆-韦尔奇算法来进行 HMM 模型训练，完成专有名词的提取。同样地，本文使用的是研究成熟的 SVM 分类器，只是改进了产生训练 SVM 分类器输入的方法，从而改进 SVM 分类器的效果。

4.1 数据集

本实验所采用的数据集是本课题组从一些门户网站和微博抓取的信息。抓取数据的主要门户网站包括食品安全_99 健康网、中国食品安全论坛官方网站、人民网食品频道、食品药品安全舆情网、舆情在线_新华网、食品安全腾讯微博等。

在本次实验中，采用的数据集包括 202831 条文本，是在不同的时间段抓取的。实验前，将实验数据集平均分为 4 组，用来测试在不同的数据集下，本文所设计的歧义消解算法的稳定性。

实验时，将每一组实验数据集里的文本分为训练文本集及实验文本集，分配比例为 2∶1，训练文本集用来训练 HMM 模型并完成训练文本集的专有名词的识别与提取，以及通过改变特征提取时的阈值，训练得到一个当前 SVM 分类算法情况下的最优 SVM 分类器。实验文本集用来验证本文所设计的消歧算法的有效性。

4.2 实验结果与分析

本实验的目的是验证本文所设计的食品安全领域特征选择算法的有效性和稳定性。

本文采用的特征选择算法是经过优化的，通过对比算法优化前后对食品安全领域专有名词歧义消解的效果来验证算法的有效性。所以实验时需要在训练文本集上分别采用优化前和优化后的 TF-IDF 特征选择算法对文本进行特征选择。在用实验文本集进行消歧实验的时候，对实验文本的特征提取也要分别用优化前和优化后的 TF-IDF 算法。同时，为了测试算法的稳定性，实验中采用了在不同时间段抓取的 4 个数据集进行实验。

为了得到最优的分类器，首先对不同阈值下分类器的效果进行实验，实验时所选取的阈值主要包括：0.1、0.05、0.01、0.005 等。

取不同阈值时的实验结果如表 1 所列(以准确率、召回率和 F1 值为评测标准)。

表 1 取不同阈值时的实验结果

	阈值	准确率 P/%	召回率 R/%	F1/%
1	0.1	88.65	43.65	58.50
2	0.05	83.75	51.66	63.90
3	0.01	77.64	77.56	77.60
4	0.005	64.99	74.44	69.39

从实验结果可以很容易地看出，当阈值为 0.01 时，分类器效果最佳，所以阈值取 0.01。

不同数据集上的实验结果表 2 所列(以准确率、召回率和 F1 值为评测标准)。

表 2 不同数据集上的实验结果

		特征选择算法	准确率 P/%	召回率 R/%	F1/%
一	1	传统的 TF-IDF 算法	67.22	73.21	70.09
	2	改进的 LN-TF-IDF 算法	77.26	80.14	78.67
二	3	传统的 TF-IDF 算法	61.75	68.72	64.95
	4	改进的 LN-TF-IDF 算法	70.02	74.85	72.35
三	5	传统的 TF-IDF 算法	65.63	71.17	68.29
	6	改进的 LN-TF-IDF 算法	75.36	76.12	75.74
四	7	传统的 TF-IDF 算法	74.14	72.76	73.44
	8	改进的 LN-TF-IDF 算法	78.26	80.28	79.25

(下转第 26 页)

Biology and Bioinformatics (TCBB), 2013, 10(5): 1162-1175

[26] Saleh S, Olson B, Shehu A. A population-based evolutionary search approach to the multiple minima problem in de novo protein structure prediction [J]. BMC Structural Biology, 2013, 13 (1): 1-28

[27] Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures [J]. Proceedings of the National Academy of Sciences of the Unitized States of America, 2000, 97(19): 10383-10388

[28] Kortemme T, Morozov A V, Baker D. An orientation- dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes [J]. Journal of molecular biology, 2003, 326(4): 1239-1259

[29] 张贵军,郝小虎,周晓根,等. 动态步长蛋白质构象空间搜索方法[J]. 吉林大学学报(工学报)

[30] Huang E S, Samudrala R, Park B H. Scoring functions for ab initio protein structure prediction [M]// Protein Structure Prediction: Methods and Protocols. 2000

[31] Keaver-Fay A, Tyka M, Lewis S M, et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules [J]. Methods in Enzymology, 2011, 487: 545-574

[32] Gront D, Kulp D W, Vernon R M, et al. Generalized fragment picking in Rosetta: design, protocols and applications [J]. PLoS One, 2011, 6(8): e23294

[33] 郝小虎,张贵军,周晓根,等. 一种基于片段组装的蛋白质构象空间优化算法[J]. 计算机科学, 2015, 42(3): 237-240

[34] Storn R, Price K. Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces [J]. Journal of Global Optimization, 1997, 11(4): 341-359

[35] Storn R. Differential evolution design of an IIR-filter [C]// Proceedings of IEEE International Conference on Evolutionary Computation, 1996. Nagoya, 1996: 268-173

[36] Berg B A, Neuhaus T. Multicanonical ensemble: a new approach to simulate first-order phase transitions [J]. Physical Review Letters, 1992, 68(1): 9-12

[37] Bradley P, Misura K M, Baker D. Toward high-resolution de novo structure prediction for small proteins [J]. Science, 2005, 309 (5742): 1868-1871

[38] Zhang Y, Kolinski A, Skolnick J. TOUCHSTONE II: a new approach to ab initio protein structure prediction [J]. Biophysics Journal, 2003, 85(2): 1145-1164

[39] Wu S T, Skolnick J, Zhang Y. Ab initio modeling of small proteins by iterative TASSER simulations [J]. BMC Biology, 2007, 5(1): 1-10

(上接第 9 页)

其中,不同实验组使用不同的数据集,同一组使用的是相同的数据集,即 1、2 使用相同数据集,一、二使用不同数据集。为了更直观地表示结果,将上述实验数据改为柱状图表示,如图 2 所示。

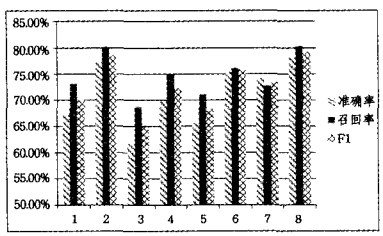

图 2 实验结果柱状图

从图 2 可以看出,在不同的数据集上,本文所设计的算法的歧义消解效果都有较为明显的提升,且效果稳定。具体来讲,基于本文改进的 LN-TF-IDF 算法的歧义消解方法的 F1 比基于传统的 TF-IDF 算法的歧义消解方法的 F1 平均提升了 7.31%。

实验结果说明,本文所设计的算法在解决食品安全领域专有名词的歧义问题中有较好的稳定性和有效性。

结束语 总体来看,本文设计的歧义消解算法的歧义消解的关键步骤在于利用 SVM 分类器消歧,主要工作在于 TF-IDF 算法的优化。该改进的算法增加考虑了特征词的上下文关系及特征词的长度在特征词权重计算中的影响,这样计算得到的特征词权重弥补了传统的 TF-IDF 算法的不足,在一定程度上提高了特征词选择的准确性,进而提高了歧义消解的有效性。这说明特征选择在歧义消解中的重要性。

参 考 文 献

[1] 龚凌晖. 中文命名实体识别与歧义消解研究[D]. 上海:复旦大学, 2011

[2] 何径舟,王厚峰. 基于特征选择和最大熵模型的汉语词义消歧[J]. 软件学报, 2010(6): 1287-1295

[3] Pedersen T. A Decision Tree of Bigrams is an Accurate Predictor of Word Sense [C]// Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics(NAACL-01). Pittsburgh, PA, 2001

[4] Hoffart J, Yosef M A, Bordino H, et al. Robust Disambiguation of Named Entities in Text[C]// Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, Scotland, UK, 2011: 782-792

[5] 戴祥鹰. 文本聚类在话题检测与人名消歧中的应用研究[D]. 哈尔滨:哈尔滨工业大学, 2010

[6] 韩伟. 人名消歧研究与实现[D]. 北京:北京大学, 2014

[7] 李永亮,黄曙光,鲍蕾. 一种基于 PageRank 算法和知网的词义消歧方法[J]. 计算机应用与软件, 2011, 28(4): 213-215

[8] 徐钟. 隐含马尔科夫模型在中文实体分类中的应用及研究[D]. 南昌:南昌大学, 2012

[9] Mena B H, van K M. A Hybrid Approach for Robust Multilingual Toponym Extraction and Disambiguation [C]// International Conference on Language Processing and Intelligent Information Systems. Warsaw, Poland, 2013

[10] 廖浩,李志蜀,王秋野,等. 基于词语关联的文本特征词提取方法[J]. 计算机应用, 2007, 27(12): 3009-3012

[11] 平源. 基于支持向量机的聚类及文本分类研究[D]. 北京:北京邮电大学, 2012

[12] 范昕炜. 支持向量机算法的研究及其应用[D]. 杭州:浙江大学, 2003