

文章编号:1001-9081(2015)07-1945-05

doi:10.11772/j.issn.1001-9081.2015.07.1945

服装类商品属性实体识别

周详^{1*}, 李少波^{1,2}, 杨观赐²

(1. 中国科学院 成都计算机应用研究所, 成都 610041; 2. 现代制造技术教育部重点实验室(贵州大学), 贵阳 550003)

(*通信作者电子邮箱 codant@163.com)

摘要:针对服装类商品标题中的商品属性实体识别问题,提出了一种边界探测规则与条件随机场(CRF)相结合的混合方法。首先,使用统计方法挖掘隐蔽的实体提示字信息;然后,以字为粒度对三种统计成词指标及其内涵进行了阐释;接着,基于统计成词指标和提示字信息设计了实体边界探测规则;最后,基于经验风险最小化给出了规则中阈值的确定方法。在与字标注的CRF模型的对比实验中,总体准确率、召回率、F1值分别提升了1.61%、2.54%和2.08%,验证了对于实体边界探测规则的有效性。所提方法可用于电子商务信息检索(IR)、电子商务信息抽取(IE)、查询意图识别等任务。

关键词:命名实体识别;服装类商品;条件随机场;电子商务

中图分类号: TP391.1 文献标志码:A

Entity recognition of clothing commodity attributes

ZHOU Xiang^{1*}, LI Shaobo^{1,2}, YANG Guanci²

(1. Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu Sichuan 610041, China;

2. Key Laboratory of Advanced Manufacturing Technology, Ministry of Education (Guizhou University), Guiyang Guizhou 550003, China)

Abstract: For the entity recognition of commodity attributes in clothing commodity title, a hybrid method combining Conditional Random Field (CRF) with entity boundary detecting rules was proposed. Firstly, the hidden entity hint character messages were obtained through a statistical method; secondly, statistical word indicators and their implications were interpreted with a granularity of character; thirdly, entity boundary detecting rules was proposed based on the entity hint characters and statistical word indicators; finally, a method for identifying threshold values in rules was proposed based on empirical risk minimization. In the comparison experiments with character-based CRF models, the overall precision, recall and F1 score were increased by 1.61%, 2.54% and 2.08% respectively, which validated the efficiency of the entity boundary detecting rule. The proposed method can be used in e-commerce Information Retrieval (IR), e-commerce Information Extraction (IE) and query intention identification, etc.

Key words: Named Entity Recognition (NER); clothing commodity; Conditional Random Field (CRF); e-commerce

0 引言

命名实体识别(Named Entity Recognition, NER)是自然语言处理领域的一项基础任务,被广泛应用于分词、词性标注、信息抽取、信息检索、自动问答系统等任务中。开放领域的命名实体识别研究主要针对日期、时间、数量、人名、地名、和机构名六类命名实体。其中中文机构名的识别较为困难,其主要的识别方法可以分为3类:基于规则的方法、基于统计的方法以及规则与统计相结合的方法。基于统计的方法中条件随机场(Conditional Random Field, CRF)^[1]综合了最大熵模型和隐马尔可夫模型的优点^[2],在分词、序列标注领域得到广泛的应用。文献[3–5]采用规则与统计模型相结合的方法研究命名实体识别,取得了较好的结果。

随着我国电子商务的蓬勃发展,商务信息处理研究得以推进。当前商务信息处理中的命名实体识别研究^[6–9]主要围绕品牌名(brand, BRA)、系列名(series, SER)、型号名(type,

TYP)、产品名(product, PRO)展开。但商品的种类繁多,属性不尽相同,这些研究并不适用于所有商品。

服装类商品在电子商务中占有巨大的市场份额,对服装类商品命名实体识别的研究可为提供电商分词、信息抽取、用户意图识别等提供支撑,具有较强的应用价值。本文针对服装类商品,将品牌名、风格名(style, STY)、材质名(material, MAT)、局部特征名(local feature, LOC)和商品名(commodity, COM)五类服装类商品属性实体(Clothing Commodity Attributes Entity, CCAE)作为识别的对象,结合基于字标注的条件随机场和实体边界探测算法,取得了较好的识别结果。

1 任务界定与分析

服装类商品属性实体目前尚没有统一的定义。在面对服装类商品时,用户较少关心系列名、型号名,更多地关心风格、材质以及服装局部的特征。由此引出本文要研究的属性实体:

收稿日期:2015-02-04;修回日期:2015-04-04。

基金项目:国家科技支撑计划项目(2012BAF12B14);国家自然科学基金资助项目(51475097)。

作者简介:周详(1989–),男,河南宜阳人,硕士研究生,主要研究方向:自然语言处理;李少波(1973–),男,湖南岳阳人,教授,博士生导师,博士,主要研究方向:计算智能、智能系统;杨观赐(1983–),男,湖南嘉禾人,副教授,博士,主要研究方向:计算智能、智能系统。

品牌名 指服装类商品的品牌,不包含系列名。如“雅*·自由自在”中“雅*”品牌名,“自由自在”则不是。

风格名 指表示服装整体风格的词,包括区域风格名如“韩版”“日式”“英伦风”“波西米亚”等,场景风格名如“休闲”“运动”“商务”“学院风”“机车风”等,以及主观感受风格名如“文艺”“甜美”“清新”“性感”等。

材质名 指商品主体部分所用的材质如“棉”“麻”“桑蚕丝”“羽绒”“涤纶”“莫代尔”“竹炭纤维”等,以及材质本身的属性如“亮面”“亚光”“高弹”“免烫”等。

局部特征名 指服装一个部分的特征,包括领型特征、袖口特征、衣襟特征、裤脚或裙摆特征等。

商品名 指可单独用于指代商品的物品名名称,如“t恤”“polo 衫”“针织衫”“棉衣”“羽绒服”“披肩”“袜子”等。

本文选用电子商务中的商品标题语料,从中识别上述实体。商品标题语料与产品名识别中常使用的新闻语料、评测语料相比实体密度更大,但存在用词不规范、随意组合等缺点。服装类商品实体本身的特殊性,以及商品标题语料的特点都给识别带来了难度:

1) 中文品牌的识别可类比中文人名识别,不同的是人名中的姓氏用字范围确定、可穷举,具有明显的提示作用;而品牌的用字范围无法确定,其用字倾向也更隐蔽。

2) 服装的材质可能是混合材质,因此材质名中存在一部分随意组合的现象,如“莱卡棉”“棉麻”“丝绒”等,占标注语料中材质名总数的 15.79%。

3) 局部特征名中存在一部分与风格名、商品名、材质名交叠的情况,如“可脱卸帽”“时尚领”“貉子毛领”“衬衫领”等,占标注语料中局部特征名总数的 10.74%。

4) 商品名实体存在与风格名、材质名、局部裁剪特征名的交叠的情况,如“休闲裤”“运动服”“羊毛衫”“呢大衣”“蝙蝠衫”等,占标注语料中商品名总数的 20.96%。

2 识别方法

2.1 整体流程

以字为标注粒度的条件随机场模型能够较好地捕捉和利用实体的用字倾向,从而识别出新实体中的一部分字,在命名实体识别中得到广泛应用;其对于实体边界的识别仅依赖于对训练集内部实体首字或尾字的考虑,无法利用更广泛的无标注语料中的统计成词指标。本文在字标注的条件随机场模型的基础上,引入基于统计成词指标的实体边界探测规则,帮助修正字标注条件随机场模型对于实体边界识别的不足。该方法实现流程如图 1 所示。

整个流程主要分为两部分:第一部分是基于条件随机场对服装类商品实体作初步识别。包括原子特征的选择以及特征模板的构建;训练模型,并利用模型对服装实体初步识别。第二部分是基于规则的实体边界修正。包括提示字集的构建、统计指标哈希表的构建;边界规则的确定;对条件随机场标注结果进行边界修正。

2.2 基于条件随机场服装实体识别

2.2.1 条件随机场简介

条件随机场^[1]是一种概率无向图模型。图中节点分为输入节点和输出节点,每个输出节点代表一个目标状态,通过最大化输入到输出节点上的条件概率来估计当前输入的目标状

态,每个条件概率通过势函数的归一化乘积来计算。线性链条件随机场^[10]是条件随机场的一种,被广泛应用于分割和序列标注等任务。记 $x = \{x_1, x_2, \dots, x_n\}$ 为输入序列, $y = \{y_1, y_2, \dots, y_n\}$ 为输出状态序列,则线性链条件随机场可被形式化表述为:

$$P(y|x) = \frac{1}{Z(x)} \exp\left(\sum_{i,k} w_k f_k(y_{i-1}, y_i, x, i)\right) \quad (1)$$

其中:

$$Z(x) = \sum_y \exp\left(\sum_{i,k} w_k f_k(y_{i-1}, y_i, x, i)\right) \quad (2)$$

其中: $f_k(y_{i-1}, y_i, x, i)$ 是在位置 i 的特征函数,分为转移特征与状态特征两类; $Z(x)$ 是归一化参数; w_k 是特征函数 f_k 对应的权值,也是待学习的参数。

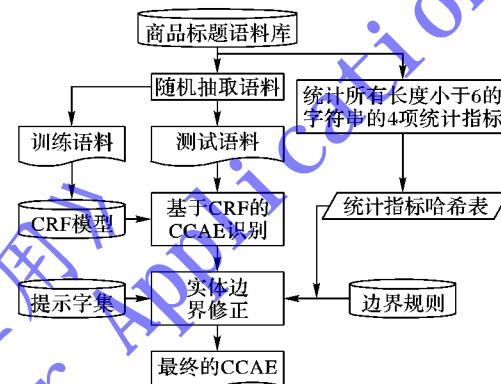


图 1 服装类商品实体识别总体流程

2.2.2 原子特征与特征模板

1) 原子特征。

① 单个字字面。

② 单个字在标题中离散化相对位置 (Discrete Relative Position, DRP)。统计 2 万条标注语料中组成品牌的字出现的位置,其中有 91.7% 位于标题的 1/4;而组成商品名的字中出现在同一区域的比例仅有 8.81%。受这个现象启发,将字在标题中的相对位置离散化为 4 个值,作为一个原子特征。对于标题 t_j 中的一个字符 c_i ,其相对位置 (Relative Position, RP) 可表示为:

$$RP(c_i, t_j) = pos(c_i)/len(t_j) \quad (3)$$

其中: $pos(c_i)$ 表示 c_i 在标题中的顺序位置(以字符为单位), $len(t_j)$ 表示标题 t_j 的总长度。字符 c_i 在标题 t_j 中的离散化相对位置可表达为:

$$DRP(c_i, t_j) = \begin{cases} 1, & RP(c_i, t_j) \in (0, 0.25] \\ 2, & RP(c_i, t_j) \in (0.25, 0.5] \\ 3, & RP(c_i, t_j) \in (0.5, 0.75] \\ 4, & RP(c_i, t_j) \in (0.75, 1.0] \end{cases} \quad (4)$$

2) 特征模板。

本文以特征模板的方式定义特征函数,每一条模板对应一个特征函数。当前字为基准,前面取 1 个单字,后面取 2 个字,组成 4 个字的窗口;以窗口内的所有单字、单字的离散化相对位置以及所有连续两个字为模板。

2.3 基于统计成词指标的实体边界探测规则

2.3.1 提示字集与统计成词指标

文献 [11~12] 使用了词内聚度、灵活度等概念帮助决定新词的边界,本节将其引入命名实体识别,帮助确定实体边

界。首先,使用统计方法挖掘隐蔽的属性实体提示字集;然后以字为单位重新描述了内聚度和灵活度两个概念,提出词的扩展倾向度量;最后基于此制定了实体边界探测规则。

1) 提示字集。文献[2,13]中提出了利用单字提示字信息帮助识别实体的方法,取得较好的结果。这里考虑用统计的方法挖掘隐蔽的提示字集,通过考察语料中的字与标注类的相关性,将与一类别相关度较高的字作为该类别的提示字。具体地,使用了类内字频率修正的卡方检验法对各类实体中的字进行特征选择:

$$\chi^2(c_i, t_j) = \frac{cf_{ij} * (x_{11}x_{00} - x_{10}x_{01})^2}{(x_{11} + x_{10})(x_{11} + x_{01})(x_{00} + x_{10})(x_{00} + x_{01})} \quad (5)$$

其中: cf_i 是字 c_i 在标注类 t_j 中的字频率, x_{11} 表示标注语料中字 c_i 在标注类 t_j 的共现数, x_{00} 表示都未出现的字数, x_{10} 表示出现 c_i 非 t_j 类的字数, x_{01} 表示 t_j 中非 c_i 的字数。每类按卡方值从大到小抽取得到提示字,共抽取600个不同的字,作为各标注类的提示字集合。表1列举了部分提示字。

表1 实体类别提示字集

实体类	提示字举例	数量
BRA	斯舍迪维伊艾	400
STY	风型韩欧派范	20
MAT	棉绒毛纺呢雪	20
LOC	领肩扣边袖袋	40
COM	衣裤衫裙恤衬	20
O	男女春夏秋冬	100

2) 内聚度。使用基于互信息的方法,来衡量组成一个词的字符串之间的内聚度。设 s 是一个长度为 n 的字符串, $s = c_1c_2\cdots c_kc_{k+1}\cdots c_n$, x, y 是字符串 s 的两个相邻子串。 $x = c_1c_2\cdots c_k$, $y = c_{k+1}c_{k+2}\cdots c_n$,则一个词的内聚度可表示为:

$$cohesion(s) = \min_{1 \leq k \leq n-1} \frac{P(xy)}{p(x)p(y)} \quad (6)$$

要注意,组成 s 两个子串 x, y 之间有固定的先后关系,因此这里使用的不是严格意义上的互信息,而是偏序的互信息。

3) 灵活度。邻接熵(Adjacent Entropy, AE)是新词发现中用于探测词的上下文灵活度的常用特征。本文选取以字为单位的邻接熵,分为左邻字熵(Left Adjacent Character Entropy, LACE)和右邻字熵(Right Adjacent Character Entropy, RACE)。记一个字符串 s 的左邻字集合为LACS(Left Adjacent Character Set),右邻字集合为RACS(Right Adjacent Character Set),则 s 的左邻字熵、右邻字熵分别可表示为:

$$LACE(s) = - \sum_{c \in LACS} P(cs | s) \text{lb}(P(cs | s)) \quad (7)$$

$$RACE(s) = - \sum_{c \in RACS} P(sc | s) \text{lb}(P(sc | s)) \quad (8)$$

其中:式(7)“ cs ”表示字符 c 在字符串 s 的左侧邻接位置出现;式(8)中“ sc ”表示字符 c 在字符串 s 右侧邻接位置出现。

4) 扩展倾向。一个完整的短语倾向于被使用于不同的上下文中;而不完整的短语则倾向于搭配相对固定的邻接字,因此完整的短语邻字熵比不完整的短语大。依据上面的原理,考察一个不完整的短语,其一侧是边界另一侧不是,边界一侧的邻字熵应该比非边界侧的邻字熵大;该短语应该向非边界一侧扩展来探测该侧边界,也就是说应该向邻字熵较小的一侧

扩展。由此,定义语料库中一个字符串 s 的扩展倾向(Extending Tendency, ET)为:

$$ET(s) = \frac{LACE(s) - RACE(s)}{\max(LACE(s), RACE(s))} \quad (9)$$

其绝对值表示扩展倾向的强度;其符号表示该字符串的扩展倾向的方向,正表示从前向后扩展,负表示从后向前扩展。

2.3.2 实体边界探测规则

分别对词频、内聚度和扩展倾向的绝对值设置阈值 α 、 β 、 γ 。基于2.3.1节中对于字串成词指标的分析,作如下假设:
①当字串的频率大于等于 α 时,认为可以使用内聚度和词频对其边界进行判断;
②当字串的内聚度大于等于 β 时,认为该字串是某个词的一部分,否则认为其并非任一词的一部分;
③当字串扩展倾向绝对值大于等于 γ 时,则认为其应该向其扩展倾向符号方向进行扩展。

对于一个已标注字符串或一个未标注的实体提示字 s' ,迭代执行下面的过程:当 s' 在语料库中词频小于 α 时,认为没有足够的词频信息来判断其是否为一个词,返回上一次迭代的边界;当 s' 词频大于 α ,内聚度小于 β 时,返回上一次迭代的边界;当 s' 词频大于 α ,内聚度大于等于 β ,扩展倾向绝对值小于 γ 时,认为其边界完整,不再扩展,返回上一次迭代的边界;当 s' 词频大于 α ,内聚度大于等于 β ,扩展倾向绝对值大于等于 γ 时,判断其沿其扩展倾向方向的邻接字已标注为其他类实体或属于其他类提示字集,返回上一次迭代的边界;否则保存当前字符串边界,沿其扩展倾向方向扩展一字得到获得新边界。当循环次数大于等于4时,返回当前边界。

下面基于经验风险最小化依次确定阈值参数 β 、 γ 和 α 。

首先构建参数学习的训练集。选取已有实体词典中所有长度大于等于2的属性实体作为正例,得样本容量为2506的正例样本集,记为 P 。对已有实体词典中任一长度小于5的属性实体 e_0 ,从标注语料中随机选取一个包含该实体的标题 t_1 ;在 t_1 中,分别从 e_0 左侧增一字得到字串 e_{-1} ,从 e_0 右侧增一字得到字串 e_1 ,如果增字后的字串并非已有词典中的词,则将其作为负例;得到容量为2956的负例样本集,记为 N_1 。对已有实体词典中任一长度大于2的属性实体 e_0' ,从标注语料中随机选取一个包含该实体的标题 t_2 ;在 t_2 中,分别从 e_0' 左侧去一字得到字串 e_{-1}' ,从 e_0' 右侧去一字得到字串 e_1' ,如果去字后的字串并非已有词典中的词,则将其作为负例;得到容量为3058个的负例样本集,记为 N_2 。分别统计集合 P, N_1, N_2 中所有样本在整个语料库中的3种统计成词指标,得到样本库 P', N_1', N_2' 。将 $P' + N_1'$ 作为参数 β 的训练集,将 $P' + N_2'$ 作为参数 γ 的训练集。

对参数 β ,设定风险函数为其训练集中违背假设②的样本数占总样本数的比例;在由经验选定的离散值集合 $B = \{2, 2.5, 3, 3.5, 4, 4.5, 5\}$ 中,调整 β 的值使其风险函数值最小,得到 β 值为3.0。参见图2。

类似地,对参数 γ ,设定风险函数为其训练集中违背假设③的样本数占总样本数的比例,在由经验选定的离散值集合 $R = \{0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4\}$ 中调整 γ 的值使其风险函数值最小,得到 γ 值为0.3。最后,在标注训练语料中对 N_2' 中的所有负例样本使用完整的边界探测规则,设定 α 的风险函数为未被完全修正样本数占总 N_2' 中总样本的比例。依据已经确定的 β 和 γ 值,在由经验选定的离散值集合 $A =$

{20, 25, 30, 35, 40, 45, 50} 内, 调整参数 α 的值, 使其风险函数值最小, 得到 α 值为 30。

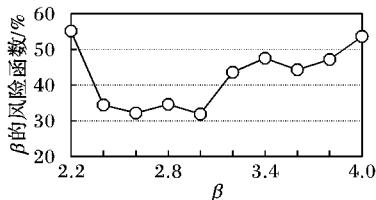


图 2 β 取不同值时其风险函数的变化

3 实验与结果分析

本文使用服装商品标题语料由某网信息技术有限公司提供。对整个服装类商品标题语料库, 去除数字、符号、英文字母, 统计其中所有长度大于 1 小于 6 的连续字符串在语料库中的频度、内聚度、左右邻字熵较大值、扩展倾向, 建立哈希表。

从语料库中随机抽取 22 000 条标题, 人工对其中的 5 类实体进行标注, 实体中的字标注为使用“B-实体类别名称”作为某个实体开始字标注, “I-实体类别名称”作为实体非开始字标注, “O”作为非目标实体标注。其中 20 000 条作为训练集, 其余 2 000 条作为测试集。表 2 给出了训练集和测试集的统计信息。

表 2 标注语料统计信息

实体类别	训练集	测试集
BRA	16 272	1 383
STY	26 517	2 326
MAT	8 824	881
LOC	18 027	2 279
COM	32 833	3 372
总计	102 928	10 241

本文实验使用了开源软件 CRF++^[14], 其中使用了 L-BFGS (Limited-memory Broyden-Fletcher-Goldfarb-Shanno) 算法^[15]训练模型参数。基于统计成词指标的实体边界探测规则使用 Python 实现。设计以下 3 个实验, 对实验结果作为对比分析。

实验 1 在字标注 CRF 框架下, 测试仅使用字面特征的 CRF 的识别性能。

实验 2 在字标注 CRF 框架下, 测试使用字面特征和离散化相对位置特征的 CRF 的识别性能。

实验 3 在实验 2 识别结果的基础上, 使用本文提出的边界探测规则进行实体边界修正, 测试该混合方法的识别性能。

本文使用 3 个评测指标: 准确率、召回率和 F1 测度。对 3 个实验结果的上述 3 个指标进行对比分析, 得到表 3~5。

实验 2 中离散化相对位置的引入对于品牌名和物品名的识别的准确率有小幅度提升, 对于其他类别实体影响较小。这是由于品牌名和物品名的位置分布较其他类实体相对集中, 相对位置的引入对这两类实体有较强的提示作用。

在基于条件随机场识别方法中, 由于标注规模的限制, 训练语料中没有出现的字难于被识别, 因此可能出现某些实体中的字未识别出, 甚至某些实体整个未被识别出的情况。其中品牌名和局部特征名用字较其他类实体更广泛, 更易出现

这种情况, 因此在实验 1 和实验 2 中的识别结果较差。而基于词扩展的实体边界探测规则一定程度上修正了这些错误。结合规则以后, 对局部特征名和品牌名识别效果的提升最为明显, F1 值分别增加了 4.36% 和 2.84%。

表 3 3 个实验的准确率对比 %

实体类别	实验 1	实验 2	实验 3
BRA	86.28	87.24	89.58
STY	94.76	94.89	95.61
MAT	96.75	96.75	96.30
LOC	85.36	84.91	87.67
COM	92.45	93.28	94.09
总体指标	90.99	91.33	92.60

表 4 3 个实验的召回率对比 %

实体类别	实验 1	实验 2	实验 3
BRA	85.03	85.54	88.86
STY	91.70	91.75	92.73
MAT	91.37	91.37	91.49
LOC	77.53	77.53	83.28
COM	90.48	90.93	91.99
总体指标	87.22	87.44	89.76

表 5 3 个实验的 F1 测度对比 %

实体类别	实验 1	实验 2	实验 3
BRA	85.65	86.38	89.22
STY	93.21	93.29	94.15
MAT	93.99	93.99	93.83
LOC	81.26	81.06	85.42
COM	91.46	92.09	93.03
总体指标	89.07	89.34	91.15

另外注意到, 实验 3 中材质名识别准确率有小幅下降。一方面, 一部分表示材质的字“貉”“棉”等, 参与构成某些局部特征名或商品名如“貉毛领”“棉马甲”; 另一方面, 实验 2 识别出的“毛领”“马甲”等词不满足实体边界扩展条件; 因此这些字未被并入目标实体, 而是被错误地判定为材质名, 对准确率造成了不良影响。

总体上, 实验 2 对于识别的准确率和召回率均有小幅提升, 验证了相对位置特征的有效性; 实验 3 相对于实验 1, 准确率提升了 1.61%, 召回率提升了 2.54%, F1 值提升了 2.08%, 验证了本文提出的 CRF 与边界规则结合的方法的有效性。

4 结语

本文针对服装类商品属性实体, 利用大规模语料中的实体边界的统计信息, 构建规则, 修正字标注的条件随机场模型识别结果中的实体边界误差, 取得了较好的结果。但本文仍有一些不足之处:

1) 所用的商品标题语料中有相似的标题成组出现的情况, 在一定程度上降低了这些标题中实体的边界熵, 对实体边界探测造成了干扰。

2) 基于词扩展的实体边界规则可以帮助识别 CRF 训练语料中较少出现, 而在语料库中频繁出现的词, 却无法识别出语料库中较少出现的词。

3)词边界统计指标知识库的获取需要消耗大量计算资源。

4)统计指标阈值的范围需要根据经验人为设定。

本文的识别结果有待进一步提高,下一步可考虑增加实体结构方面规则;也可考虑设计以词为标注单位的条件随机场模型,以便将本文中用于判断实体边界的统计指标直接融入模型,完全由统计模型决定实体边界。

参考文献:

- [1] LAFFERTY J, McCALLUM A, PEREIRA F. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]// ICML 2001: Proceedings of the 2001 International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 2001: 282 – 289.
- [2] FENG Y, SUN L, ZHANG D, et al. Study on the Chinese named entity recognition using small scale tail hints [J]. *Acta Electronica Sinica*, 2008, 36(9): 1833 – 1838. (冯元勇,孙乐,张大鲲,等.基于小规模尾字特征的中文命名实体识别研究[J].电子学报,2008,36(9):1833 – 1838.)
- [3] XIANG X, SHI X, ZENG H. Chinese named entity recognition system using statistics-based and rules-based method [J]. *Journal of Computer Applications*, 2005, 25(10): 2404 – 2406. (向晓雯,史晓东,曾华琳.一个统计与规则相结合的中文命名实体识别系统[J].计算机应用,2005,25(10):2404 – 2406.)
- [4] SHE J, ZHANG X. Musical named entity recognition method [J]. *Journal of Computer Applications*, 2010, 30(11): 2928 – 2931. (余俊,张学清.音乐命名实体识别方法[J].计算机应用,2010,30(11):2928 – 2931.)
- [5] ZHANG J, WANG S, QIAN C. CRF and rules-based recognition of medical institutions name in Chinese [J]. *Computer Applications and Software*, 2014, 31(3): 159 – 162. (张金龙,王石,钱存发.基于CRF和规则的中文医疗机构名称识别[J].计算机应用与软件,2014,31(3):159 – 162.)
- [6] LIU F, ZHAO J, LYU B, et al. Study on product named entity recognition for business information extraction [J]. *Journal of Chinese Information Processing*, 2006, 20(1): 7 – 13. (刘非凡,赵军,吕碧波,等.面向商务信息抽取的产品命名实体识别研究[J].中文信息学报,2006,20(1):7 – 13.)
- [7] ZHANG C, GUO J, XIAN Y, et al. Named entity recognition of the products with English based on conditional random fields [J]. *Computer Engineering and Science*, 2010, 32(6): 115 – 117. (张朝胜,郭剑毅,线岩团,等.基于条件随机场的英文产品命名实体识别[J].计算机工程与科学,2010,32(6):115 – 117.)
- [8] MEI F. Research on product named entity recognition and normalization [D]. Harbin: Harbin Institute of Technology, 2011. (梅丰.产品名实体识别及规范化研究[D].哈尔滨:哈尔滨工业大学,2011.)
- [9] LUO F, XIONG Q, XIAO M. Product named entity recognition based on ontology [J]. *Journal of Wuhan University of Technology: Information and Management Engineering*, 2011, 33(6): 948 – 952. (罗芳,熊前兴,肖敏.基于本体的产品命名实体识别研究[J].武汉理工大学学报:信息与管理工程版,2011,33(6):948 – 952.)
- [10] LI H. Statistical learning methods [M]. Beijing: Tsinghua University Press, 2012: 195 – 197. (李航.统计学习方法[M].北京:清华大学出版社,2012:195 – 197.)
- [11] CHEN F, LIU Y, WEI C, et al. Open domain new word detection using conditional random field method [J]. *Journal of Software*, 2013, 24(5): 1051 – 1060. (陈飞,刘奕群,魏超,等.基于条件随机场方法的开放领域新词发现[J].软件学报,2013,24(5): 1051 – 1060.)
- [12] ZHAO X, ZHANG H. New words identification based on iterative algorithm [J]. *Computer Engineering*, 2014, 40(7): 154 – 158. (赵小宝,张华平.基于迭代算法的新词识别[J].计算机工程,2014,40(7):154 – 158.)
- [13] FENG Y, SUN L, ZHANG D, et al. A rapid algorithm to Chinese named entity recognition based on single character hints [J]. *Journal of Chinese Information Processing*, 2008, 22(1): 104 – 110. (冯元勇,孙乐,李文波,等.基于单字提示特征的中文命名实体识别快速算法[J].中文信息学报,2008,22(1):104 – 110.)
- [14] KUDO T. CRF++: yet another CRF toolkit [EB/OL]. [2014-12-01]. <http://crfpp.googlecode.com/svn/trunk/doc/index.html>.
- [15] LIU D, NOCEDAL J. On the limited memory BFGS method for large scale optimization [EB/OL]. [2014-12-07]. <http://users.iems.northwestern.edu/~nocedal/PDFfiles/limited-memory.pdf>.

(上接第1932页)

- [9] YU Y, WANG Q, KUANG J, et al. An on-line density-based clustering algorithm for spatial data stream [J]. *Acta Automatica Sinica*, 2012, 38(6): 1051 – 1058. (于彦伟,王沁,邝俊,等.一种基于密度的空间数据流在线聚类算法[J].自动化学报,2012,38(6): 1051 – 1058.)
- [10] ZHANG X, FURTLEHNER C, SEBAG M. Data streaming with affinity propagation [C]// Proceedings of the 2008 Machine Learning and Knowledge Discovery in Databases, LNCS 5212. Berlin: Springer, 2008: 628 – 643.
- [11] ZHANG J, CHEN F, LI S, et al. Data stream clustering algorithm based on density and affinity propagation techniques [J]. *Acta Automatica Sinica*, 2014, 40(2): 277 – 288. (张建朋,陈福才,李绍梅,等.基于密度与近邻传播的数据流聚类算法[J].自动化学报,2014,40(2):277 – 288.)
- [12] LIU D, JIANG M. Affinity propagation clustering on oral conversa-

- tion texts [C]// ICSP 2012: Proceedings of 2012 IEEE 11th International Conference on Signal Processing. Piscataway: IEEE, 2012: 2279 – 2282.
- [13] ZHANG X, FURTLEHNER C, GERMAIN-RENAUD C, et al. Data stream clustering with affinity propagation [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2014, 26(7): 1644 – 1656.
- [14] HUANG D, WU T. Density-based clustering algorithm for mixture data sets [J]. *Control and Decision*, 2010, 25(3): 416 – 421. (黄德才,吴天虹.基于密度的混合属性数据流聚类算法[J].控制与决策,2010,25(3):416 – 421.)
- [15] FUJIWARA Y, IRIE G, KITAHARA T, et al. Fast algorithm for affinity propagation [C]// Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence. Piscataway: IEEE, 2011: 2238 – 2243.