融合距离度量和高斯混合模型的中文词义归纳模型

张宜浩 刘 智 朱常鹏
(重庆理工大学计算机科学与工程学院 重庆 400054)

摘 要 词义归纳是解决词义知识获取的重要研究课题,利用聚类算法对词义进行归纳分析是目前最广泛采用的方法。通过比较K-Means聚类算法和EM聚类算法在各自词义归纳模型上的优势,提出一种新的融合距离度量和高斯混合模型的聚类算法,以期利用两种聚类算法分别在距离度量和数据分布计算上的优势,挖掘数据的几何特性和正态分布信息在词义聚类分析中的作用,从而提高词义归纳模型的性能。实验结果表明,所提混合聚类算法对于改进词义归纳模型的性能是十分有效的。

关键词 词义归纳,距离度量,高斯混合模型,混合聚类
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.08.045

Chinese Word Sense Induction Model by Integrating Distance Metric and Gaussian Mixture Model

ZHANG Yi-hao LIU Zhi ZHU Chang-peng
(College of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China)

Abstract Word sense induction is an important topic in solving knowledge acquisition of word sense, and the most widely used method to word sense induction is based on cluster analysis algorithm. By comparing K-Means clustering algorithm with EM clustering algorithm on the model of word sense induction, we proposed a new hybrid clustering algorithm by integrating distance metric and Gaussian mixture model, which combine the advantages of distance metric and data distributed computing in the two cluster algorithms respectively to mine the role of geometrical properties and normal distribution information of training data in clustering analysis and then improve the performance of performance of word sense model. Experimental results show that the hybrid clustering algorithm proposed in this paper is very effective to improve the performance of word sense induction model.

Keywords Word sense induction, Distance metric, Gaussian mixture model, Hybrid clustering

1 引言

词义归纳(Word Sense Induction, WSI)又称词义区分,是一种利用机器学习算法从文本的语境中自动判别多义词的词义,并对相似词进行归类的技术。在自然语言理解中,词义知识的获取在知识库构建[1]、词义消歧[2]等诸多领域起着重要的作用,且词义归纳作为词义知识获取的一条重要途径,在多义词的词义内容理解、词义的形式表征等方面具有重要的意义。大量的前期工作也表明,使用词义知识比单纯地使用词形更能够改善信息检索[3]、信息抽取[4]和机器翻译[5]的结果。

目前,针对词义归纳的研究多采用无监督的方法[6],即利用各种聚类技术对多义词所处的语境进行归纳分析。目前具体的研究方法主要包括基于图的方法和基于特征向量的方法。在基于图的方法中,首先将待聚类的元素和特征在图空间中进行定义,然后利用基于图的聚类算法进行词义归纳。Klapaftis等[7]提出了一种无监督的、对歧义词进行层次分组的方法,并利用层次随机图算法对图进行构建以推断其层次结构。唐共波等[8]将多义词的上下文作为特征并构建特征向量,通过计算多义词的词向量与特征向量之间的相似度进行词语消歧。钱涛等[9]提出了一个基于超图的词义归纳模型,该模型使用最大密度超图谱聚类算法发现词义。基于特征向量的方法利用特征选择来构建向量空间,然后使用各种聚类算法进行词义相似度计算和聚类分析,因此,特征选择和聚类算法是基于特征向量的方法的关键。在特征选择方面,TF-IDF、信息增益、卡方检验等是常用的特征选择算法,基本的特征形式包括共现词语、词性、N-gram等。在特征构建方面,研究者前期也做了大量的工作,如:Van等[10]综合利用每个词的分解模型、窗口内的词以及它们的依赖关系进行词义归纳;Lau等[11]利用主题模型自动推导目标词的词义;Huang等[12]构造了一个多粒度的语义空间对歧义词进行表达,同时将词簇和主题的语义空间用于词义归纳。在聚类算法方面,K-Means算法、层次聚类、期望最大化(Expectation Maximi-

到稿日期:2016-11-10 返修日期:2017-02-17 本文受重庆市教委科学技术研究项目(kj1500920, kj1500916),国家自然科学基金项目(61603065)资助。

张宜浩(1982—),男,博士,讲师,CCF会员,主要研究方向为推荐系统、自然语言处理,E-mail:yhzhang@cqut.edu.cn(通信作者);刘 智(1977—),男,博士,副教授,主要研究方向为深度学习、机器视觉;朱常鹏(1981—),男,博士,讲师,主要研究方向为虚拟机。

zation,EM)聚类算法等是最常的聚类算法。研究表明,聚类算法会直接影响词义区分的效果,本文通过分析比较K-Means算法和EM聚类算法各自在词义归纳上的优势,提出一种融合距离度量和高斯混合模型的混合聚类算法来对词义进行聚类分析,旨在在综合利用训练数据的几何特性和正态分布信息以构建词义归纳模型。

2 融合距离度量和高斯混合模型的混合聚类算法

本文提出一种融合距离度量和高斯混合模型的混合聚类算法(Hybrid Clustering Algorithm with Distance Metric and Gaussian Mixture Model,HCDG),旨在在聚类过程中综合考虑样本间的距离度量和高斯分布信息。

2.1 权重矩阵的构建

假定 $\vec{x}$ 和 $\vec{y}$ 是训练数据集合中两个样本的特征向量,则它们之间的距离度量可定义为:

$$dis(\vec{x},\vec{y})=\sqrt{(\vec{x}-\vec{y})S^{-1}(\vec{x}-\vec{y})}$$

其中,S 是这两个特征向量的协方差矩阵。

定义一个具有 N 个顶点的图,图中的每一个顶点代表一个样本,用 $N_k(x_j)$ 表示样本 x_j 的 k 个最近邻样本,则构建样本 x_i 和样本 x_j 之间边的权重 W_{ij} 表示如下:

$$W_{ij}=\begin{cases}1, & \text{if } x_i\in N_k(x_j) \text{ or } x_j\in N_k(x_i)\\ 0, & \text{otherwise}\end{cases}$$

2.2 目标函数的构建

高斯混合模型可以看作不同高斯组件的线性组合,且每个高斯组件都服从高斯分布。假设 $P_i(c)$ 和 $P_j(c)$ 表示两个高斯分布,则这两个分布之间的 Kullback-Leibler 散度可以定义为:

$$D(P_i(c)\parallel P_j(c))=\sum_c P_i(c)\log\frac{P_i(c)}{P_j(c)} \tag{1}$$

为了获得一个对称的公式,可用式(2)来度量两个高斯分布 $P_i(c)$ 和 $P_j(c)$ 间的相似性。

$$\begin{aligned}D_{ij}&=\frac{1}{2}(D(P_i(c)\parallel P_j(c))+D(P_j(c)\parallel P_i(c)))\\&=\frac{1}{2}(\sum_c P_i(c)\log\frac{P_i(c)}{P_j(c)}+\sum_c P_j(c)\log\frac{P_j(c)}{P_i(c)})\end{aligned} \tag{2}$$

定义 $P_i(c)=P(c|x_i)$,再考虑权重矩阵 W_{ij},则可以用式(3)来度量条件概率 $P(c|x)$ 的平滑性。

$$\begin{aligned}R&=\sum_{i,j=1}^{m}D_{ij}W_{ij}\\&=\frac{1}{2}\sum_{i,j=1}^{m}(\sum_c P_i(c)\log\frac{P_i(c)}{P_j(c)}+\sum_c P_j(c)\log\frac{P_j(c)}{P_i(c)})W_{ij}\end{aligned} \tag{3}$$

将式(3)得到的平滑部分和高斯混合模型的似然估计进行线性组合,得到本文提出的混合模型算法的目标函数,如式(4)所示:

$$\begin{aligned}\ell_{new}&=\ell-\lambda R\\&=\sum_{i=1}^{m}\sum_{l=1}^{k}P(c_l|x_i)(\log p(x_i|c_l;\mu,\Sigma)+\log\Phi_l)-\frac{\lambda}{2}\sum_{i,j=1}^{m}\\&\quad(\sum_c P_i(c)\log\frac{P_i(c)}{P_j(c)}+\sum_c P_j(c)\log\frac{P_j(c)}{P_i(c)})W_{ij}\end{aligned} \tag{4}$$

在式(4)中,目标函数由两个部分组成,公式的前一部分是标准的高斯混合模型,后一部分是利用 Kullback-Leibler 散度度量的样本间的相似度,而 λ 则是上述两部分进行线性组合的权重因子。由式(4)中的目标函数可以看出,该算法不仅考虑了数据的正态分布信息,同时也考虑了数据间的几何结构信息,其分别由式(4)中的高斯混合模型和距离相似性矩阵来度量。与标准的 EM 聚类算法相同,由于该式的求导结果不是一个封闭解,因此本文采用期望最大化算法对式(4)中目标函数的最大值进行求解。

2.3 目标函数的求解

本文利用期望最大化(EM)算法对目标函数进行求解时两个步骤是交替进行的:第一步,计算期望(E-Step),根据现有估计值计算最大似然估计;第二步,期望最大化(M-Step),根据 E-Step 求得的最大似然估计重新对各参数进行估值。M-Step 求得的参数估计值被用于下一个 E-Step 计算中,这两个步骤不断交替进行。

(1)计算期望(E-Step)

算法的第一步就是计算隐藏变量 $P(c_i=j|x_i)$ 的后验概率,其表达式中包括了 3 个参数 Φ,μ 和 Σ。利用贝叶斯公式计算其后验概率:

$$P(c_i=j|x_i)=\frac{p(x_i|c_i=j;\mu,\Sigma)p(c_i=j;\Phi)}{\sum_{l=1}^{k}p(x_i|c_i=l;\mu,\Sigma)p(c_i=l;\Phi)} \tag{5}$$

其中,$p(x_i|c_i=j;\mu,\Sigma)$ 的值通过利用高斯密度函数计算得到,而 $p(c_i=j;\Phi)$ 表示数据样本中类别 $c_i=j$ 所占的比例,记为 Φ_j。

(2)期望最大化(M-Step)

在第二步(M-Step)中,需要求解函数表达式的最大似然估计。由于表达式的求导结果不是一个封闭解(closed form),因此需要利用期望最大化算法对其进行优化。M-Step 求解的最终目标就是求解最大似然估计函数中各参数的值。

根据式(4)的目标函数,为了方便计算,可将目标函数 ℓ_{new} 分解为 ℓ_1 和 ℓ_2 两个部分。

假定 $\ell_{new}=\ell_1-\ell_2$,则有式(6)和式(7):

$$\ell_1=\sum_{i=1}^{m}\sum_{l=1}^{k}P(c_l|x_i)(\log p(x_i|c_l;\mu,\Sigma)+\log\Phi_l) \tag{6}$$

$$\begin{aligned}\ell_2&=\frac{\lambda}{2}\sum_{i,j=1}^{m}(D(P_i(c)\parallel P_j(c))+D(P_j(c)\parallel P_i(c)))W_{ij}\\&=\frac{\lambda}{2}\sum_{i,j=1}^{m}(\sum_c P_i(c)\log\frac{P_i(c)}{P_j(c)}+\sum_c P_j(c)\log\frac{P_j(c)}{P_i(c)})W_{ij}\end{aligned} \tag{7}$$

根据式(6)和式(7),在 M-Step 中重新估计得到的 Φ_i 值:

$$\Phi_k=\frac{1}{m}\sum_{i=1}^{m}p(c_k|x_i) \tag{8}$$

下一步的任务是重新估计其他两个参数:均值 μ_k 和协方差 Σ_k。

$$\begin{aligned}&D(P_i(c)\parallel P_j(c))\\&=\sum_c P_i(c)\log\frac{P_i(c)}{P_j(c)}\\&=\sum_{l=1}^{k}P(c_l|x_i)\log\frac{P(c_l|x_i)}{P(c_l|x_j)}\\&=\sum_{l=1}^{k}P(c_l|x_i)\{[\frac{1}{2}(x_j-\mu_l)^{\mathrm{T}}\Sigma_l^{-1}(x_j-\mu_l)-\frac{1}{2}(x_i-\\&\quad\mu_l)^{\mathrm{T}}\Sigma_l^{-1}(x_i-\mu_l)]+O(x_i\parallel x_j)\}\end{aligned}$$

其中，$O(x_i \| x_j)=\log\frac{\sum_{l=1}^{k}N(x_j|\mu_l,\Sigma_l)\Phi_l}{\sum_{l=1}^{k}N(x_i|\mu_l,\Sigma_l)\Phi_l}$。

由于 $O(x_i \| x_j)+O(x_j \| x_i)=0$，因此：

$$\ell_1=\sum_{i=1}^{m}\sum_{l=1}^{k}P(c_l|x_i)(\log\ p(x_i|c_l;\mu,\Sigma)+\log\ \Phi_l)$$

$$\ell_2=\frac{\lambda}{2}\sum_{i,j=1}^{m}(D(P_i(c)\|P_j(c))+D(P_j(c)\|P_i(c)))W_{ij}$$

根据式(4)中的 ℓ_{new}，对 μ_k 求偏导数：

$$\frac{\partial\ell_{new}}{\partial\mu_k}=\frac{\partial\ell_1}{\partial\mu_k}-\frac{\partial\ell_2}{\partial\mu_k}$$

$$=\sum_{i=1}^{m}(x_i-\mu_k)\Sigma_k^{-1}P(c_k|x_i)-\frac{\lambda}{2}\sum_{i,j=1}^{m}\{(x_i-x_j)\Sigma_k^{-1}(P(c_k|x_i)-P(c_k|x_j))\}W_{ij}$$

令 $\frac{\partial\ell_{new}}{\partial\mu_k}=0$，可得：

$$\mu_k=x_i-\frac{\lambda\sum_{i,j=1}^{m}\{(x_i-x_j)(P(c_k|x_i)-P(c_k|x_j))\}W_{ij}}{2\sum_{i=1}^{m}P(c_k|x_i)}\quad(9)$$

根据式(4)中的 ℓ_{new}，对 Σ_k^{-1} 求偏导数，并令 $\frac{\partial\ell_{new}}{\partial\Sigma_k^{-1}}=\frac{\partial\ell_1}{\partial\Sigma_k^{-1}}-\frac{\partial\ell_2}{\partial\Sigma_k^{-1}}=0$，得到对参数 Σ_k 的估计：

$$\Sigma_k=\sum_{i=1}^{m}\varphi_{i,k}+\frac{\lambda\sum_{i,j=1}^{m}\{[\varphi_{j,k}-\varphi_{i,k}](P(c_k|x_i)-P(c_k|x_j))\}W_{ij}}{2\sum_{i=1}^{m}P(c_k|x_i)}\quad(10)$$

根据式(8)－式(10)的计算结果，获得对目标函数中参数 Φ_k，μ_k 和 Σ_k 的估计值，从而利用 EM 聚类算法构建一个新的混合聚类模型。

3　实验与分析

3.1　实验准备和评价

本实验数据是由中科院软件研究所基础软件国家工程研究中心信息检索实验室提供。实验中对词义的归纳采用 F-Score 评估方法，通常认为 F-Score 越高，算法的聚类效果越好。

定义：s_r 表示给定的类别，n_r 表示对应的样本个数；h_i 表示给定的聚类簇，n_i 表示对应的样本个数；而属于类别 s_r 和簇 h_i 的样本数量用 n_r^i 表示，则有：

精确率(Precision)：$P(s_r,h_i)=\frac{n_r^i}{n_r}$

召回率(Recall)：$R(s_r,h_i)=\frac{n_r^i}{n_i}$

类别 s_r 和簇 h_i，F-Score 定义为：

$$F(s_r,h_i)=\frac{2P(s_r,h_i)R(s_r,h_i)}{P(s_r,h_i)+R(s_r,h_i)}$$

对于给定的某一类别 s_r，其 F-Score 定义为 $F(s_r)=\max_{h_i}F(s_r,h_i)$。

聚类算法整体的 F-Score 定义为 $F\text{-}Score=\sum_{r=1}^{c}\frac{n_r}{n}F(s_r)$，其中 c 是聚类的类别数，n 是数据中的样本总数。

3.2　实验数据处理

在词义知识理解中，影响词义的因素大致可分为词法、句法、语义、语用 4 类。句法功能受限于高准确率的句法分析器，目前虽然在句法分析上取得了积极的进展，但对语境类特征的研究尚不成熟；限于当前自然语言处理的水平，对语义的研究仅限于浅层语义分析，还很难实现深层语义分析；至于语用知识，目前则更是难以获取。再结合中文词义归纳数据集的具体情况，本文选用目标词所在句子的一定窗口范围内的高频词作为主要特征词，并结合共现词语、短语结构等信息来构建词义归纳模型的特征向量空间。在特征向量的构建上，特征词的高度稀疏性和特征向量的高维度极大影响着词义的向量表示，为解决此问题，本文利用了哈尔滨工业大学的同义词词林对特征词进行分类抽象表示，首先利用 TF-IDF 算法对目标词窗口范围内的高频词进行筛选，然后利用 TF-IDF 算法计算从同义词词林中提取出的词的分类编码的权重。此方法可大大降低特征向量数据的稀疏度，同时也减少了特征向量的维数，实验也进一步表明了此方法在词义归纳模型构建上的有效性。例如，针对要进行词义归纳的目标词“把握”，首先从该目标词所在的句子中抽取出一定窗口范围内的特征词，如“双手”、“紧紧”等；然后针对每一个特征词在同义词词林中查询其五级编码，如“双手”和“紧紧”分别对应五级编码“Bk08C01”和“Eb09A01”，为了降低特征向量的维度，可选用五级编码的前两级或三级编码对特征词进行抽象表示，然后分别赋予每一级编码不同的特征权重，再利用 TF-IDF 算法计算每一个特征词的权重，从而构建词义归纳模型的特征向量空间。

3.3　实验结果与分析

在本文提出的混合聚类算法中有一个很重要的参数 λ，它表示距离度量在目标函数中的权重，是描述模型平滑性的一个重要参数。当 $\lambda=0$ 时，该混合聚类算法就变成了标准的 EM 聚类。本文通过在词义归纳数据集中的反复实验得出：当 $\lambda=0.2$ 时，算法获得了较好的聚类成绩，词义归纳的 F-Score 结果如表 1 所列。

表 1　中文词义归纳的 F-Score/%

目标词	K-Means	EM	HCDG
暗淡	65.28	62.69	72.50
保安	74.77	67.63	75.11
报销	67.52	70.56	70.21
比重	69.07	69.07	75.50
病毒	65.75	63.77	70.67
材料	50.91	50.04	57.45
参加	65.28	66.67	73.19
程序	65.75	66.67	72.22
冲洗	65.75	62.69	72.83
充电	66.67	63.24	70.00
打断	64.66	67.24	72.31
打开	64.82	52.31	67.70
单纯	63.53	79.74	76.19
导师	66.22	83.58	82.34
东北	64.41	85.55	83.19
东西	52.94	50.63	67.45
肚腩	66.53	63.35	66.71
扼杀	89.67	89.67	90.44
发展	63.92	87.74	87.23
反射	63.32	66.71	64.29
平均	65.84	68.48	73.38

从表 1 中的数据可以看出，本文提出的混合聚类算法对聚类结果的改进是卓有成效的。例如，对于“暗淡”这一目标词，利用传统的聚类算法(K-Means 算法和 EM 聚类算法)进行实验所得的准确率分别为 65.28%和 62.29%，而利用 HC-

DG 算法进行聚类时，其准确率提升到了 72.50%；再如“单纯”一词，利用 K-Means 算法和 EM 聚类算法进行聚类分析时，其准确率分别为 63.53%和 79.74%，这显示了针对“单纯”这一多义词，EM 聚类算法远优于 K-Means 算法；而对于“保安”一词，利用 K-Means 算法和 EM 聚类算法进行聚类分析时，其准确率分别为 74.77%和 67.63%，这显示了针对“保安”这一多义词，K-Means 算法远优于 EM 聚类算法。事实上，上述这种情况是相当普遍的，因此本文提出了 HCDG 算法，在算法中综合考虑了数据的高斯分布信息与几何特征信息。实验结果显示，本文提出的 HCDG 算法能综合利用两种聚类算法的优势，其成绩能达到甚至高于两种聚类算法的最好成绩，且 HCDG 的聚类成绩较为稳定。

根据表 1 中的实验结果，针对绝大多数待消歧词，HCDG 算法都取得了比 K-Means 算法和 EM 聚类算法更好的成绩。但也存在例外，如“单纯”、“导师”、“东北”3 个词，利用 HCDG 算法获得的 F-Score 值要略低于 EM 聚类算法的 F-Score 值，但还是明显高于 K-Means 算法得到的 F-Score 值。进一步分析原因，可能是因为针对这些待消歧词，其样本数据分布比较符合正态分布，因此利用 EM 聚类算法建模获得了较好的实验性能。而对样本的几何特性（如 K-Means 聚类算法）进行聚类分析时，获取的实验性能相对较差（比 EM 聚类算法的性能低了 20%左右）。本文提出的 HCDG 算法是在综合利用两种聚类算法优势的基础上，对两种聚类因素进行加权处理，在一些样本极度偏离几何特征或正态分布信息时，其综合实验性能可能会不可避免地低于使用某一单一聚类算法的最好成绩，但其成绩又明显高于使用某一单一聚类算法的最差成绩。从最终的平均成绩来看，K-Means 聚类算法和 EM 聚类算法分别获取了 65.84%和 68.48%的 F-Score 值，但 HCDG 算法的平均 F-Score 值是 73.38%，远高于前两个单一聚类算法的平均成绩。本文提出的 HCDG 算法通过综合利用两种聚类因素的优势，使其聚类成绩不至于偏向使用某一单一聚类算法的最差成绩，而在绝大多数待消歧词上都取得了非常稳定的成绩，这也是本文提出 HCDG 算法的出发点所在。

为进一步验证本文提出的混合聚类算法在词义归纳模型上的效果，本文将提出的 HCDG 算法和 K-Means、EM 聚类方法进行对比，分别取 3-word window 和 5-word window 内的高频词作为特征候选词进行实验，并对 HCDG 算法中的两个参数（样本的最近邻个数 k 和平衡因子 λ）赋予不同的取值，计算待消歧目标词在词义归纳实验中的平均 F-Score，进行特征选择时分别考虑目标词前后 3 个窗口（3-word window）和 5 个窗口（5-word window）范围内的词语，得到的实验结果如图 1 和图 2 所示。

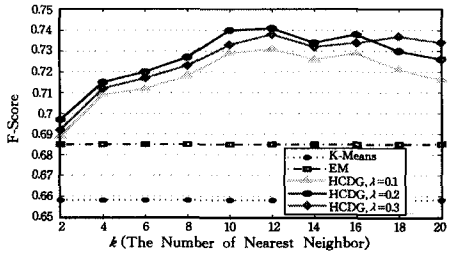

图 1 在 3-word window 上不同聚类算法得到的 F-Score 值

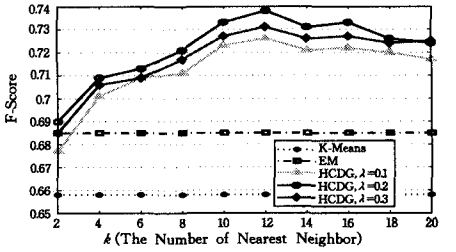

图 2 在 5-word window 上不同聚类算法得到的 F-Score 值

通过图 1 的数据可以看出，在 3-word window 时，本文提出的混合聚类算法取得了较好的结果，且随着最近邻居数目（即 k 值）的增加，实验得到的 F-Score 值是不断增加的，当 $k=12$ 时，模型得到最好的聚类结果；同时也可以看出，平衡因子 λ 的取值对模型聚类性能也有不同程度的影响，当 $\lambda=0.2$ 时，模型得到最好的聚类结果。

图 2 示出了在 5-word window 时 3 种聚类算法在数据集上的 F-Score 值对比，该结果与图 1 的结果比较相似。总体来看，HCDG 算法均取得了较好的结果，这也进一步验证了本文提出的混合聚类算法对于改进词义归纳模型的 F-Score 值的有效性。

结束语 词义归纳指根据给定多义词以及包含该词语的上下文集合，从而自动获取该多义词在语境中的词义和用法。目前，用于词义归纳研究的常用方法是基于特征向量的聚类模型，本文分别利用 K-Means、EM 聚类方法对词义进行聚类分析时发现，两种聚类方法在不同的多义词上分别表现出较优的性能，因此本文提出了一种新的混合聚类算法（HCDG 算法），试图综合利用 K-Means 聚类算法和 EM 聚类算法在距离度量计算及数据分布信息计算方面的优势，充分挖掘训练数据的几何特性和正态分布信息在词义聚类归纳中的作用，从而提高聚类模型的性能。通过在中科院软件研究所提供的中文词义归纳数据集上进行实验，验证了本文提出的混合聚类算法对于改进词义归纳模型 F-Score 值的有效性。

参 考 文 献

[1] CLAUDIO D B, LUIS E A, ROBERTO N. Knowledge base unification via sense embeddings and disambiguation[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language (EMNLP). 2015: 726-36.

[2] LU W P, HUANG H Y. Word Sense Disambiguation Based on Dependency Fitness with Automatic Knowledge Acquisition[J]. Journal of Software, 2013, 24(10): 2300-2311. (in Chinese)
鹿文鹏，黄河燕. 基于依存适配度的知识自动获取词义消歧方法[J]. 软件学报，2013，24(10)：2300-2311.

[3] SCHADD F C, ROOS N. Word-sense disambiguation for ontology mapping: Concept disambiguation using virtual documents and information retrieval techniques[J]. Journal on Data Semantics, 2015, 4(3): 167-186.

[4] YU J, LI C, HONG W, et al. A new approach of rules extraction for word sense disambiguation by features of attributes [J]. Applied Soft Computing, 2015, 27: 411-419.

[5] ETTINGER A, RESNIK P, CARPUAT M. Retrofitting sense-specific word vectors using parallel text[C]// Proceedings of NAACL-HLT. 2016: 1378-1383.

[6] AKKAYA C, WIEBE J, MIHALCEA R. Iterative Constrained Clustering for Subjectivity Word Sense Disambiguation[C]// Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. 2014: 269-278.

[7] KLAPAFTIS I P, MANANDHAR S. Word sense induction & disambiguation using hierarchical random graphs[C]// Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. 2010: 745-755.

[8] TANG G B, YU D, XUN E D. An Unsupervised Word Sense Disambiguation Method Based on Sememe Vector in HowNet [J]. Journal of Chinese Information Processing, 2015, 29(6): 23-29. (in Chinese)
唐共波，于东，荀恩东. 基于知网义原词向量表示的无监督词义消歧方法[J]. 中文信息学报，2015，29(6)：23-29.

[9] QIAN T, JI D H, DAI W H. A Hypergraph Model for Word Sense Induction [J]. Journal of Sichuan University (Engineering Science Edition), 2016, 48(1): 152-157. (in Chinese)
钱涛，姬东鸿，戴文华. 一个基于超图的词义归纳模型[J]. 四川大学学报(工程科学版)，2016，48(1)：152-157.

[10] VAN DE CRUYS T, POIBEAU T, KORHONEN A. Latent vector weighting for word meaning in context[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011: 1012-1022.

[11] LAU J H, COOK P, MCCARTHY D, et al. Word sense induction for novel sense detection[C]// Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. 2012: 591-601.

[12] HUANG Y, SHI X, SU J, et al. Unsupervised word sense induction using rival penalized competitive learning [J]. Engineering Applications of Artificial Intelligence, 2015, 41: 166-174.

(上接第 235 页)

[4] DAI J L. Study on the sparsity problem of collaborative filtering algorithm [D]. Chongqing: Chongqing University, 2013. (in Chinese)
代金龙. 协同过滤算法中数据稀疏性问题研究[D]. 重庆：重庆大学，2013.

[5] DENG A L, ZHU Y Y, SHI B L. A collaborative filtering recommendation algorithm based on item ratingprediction [J]. Journal of Software, 2003, 14(9): 1621-1628. (in Chinese)
邓爱林，朱扬勇，施伯乐. 基于项目评分预测的协同过滤推荐算法[J]. 软件学报，2003，14(9)：1621-1628.

[6] 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.

[7] JI X S, LIU Y B, LUO L M. Similarity measurement based on interest in collaborative filtering [J]. Journal of Computer Applications, 2010, 30(10): 2618-2620. (in Chinese)
嵇晓声，刘宴兵，罗来明. 协同过滤中基于用户兴趣度的相似性度量方法[J]. 计算机应用，2010，30(10)：2618-2620.

[8] CLEGER-TAMAYO S, FERNáNDEZ-LUNA J M, HUETE J F. Top-N news recommendations in digital newspapers [J]. Knowledge-Based Systems, 2012, 27(6): 180-189.

[9] ZHANG X S. Research on collaborative filtering recommendation algorithms for data sparsity [D]. Hefei: University of Science & Technology China, 2011. (in Chinese)
张学胜. 面向数据稀疏的协同过滤推荐算法研究[D]. 合肥：中国科学技术大学，2011.

[10] YU X. Research on recommendation methodsbased on collaborative filtering techniques[D]. Tianjin: Tianjin University, 2009. (in Chinese)
郁雪. 基于协同过滤技术的推荐方法研究[D]. 天津：天津大学，2009.

[11] FAN B, CHENG J J. Collaborative filtering recommendation algorithm based on user's multi-similarity[J]. Computer Science, 2012, 39(1): 23-26. (in Chinese)
范波，程久军. 用户间多相似度协同过滤推荐算法[J]. 计算机科学，2012，39(1)：23-26.

[12] MILLER B N, ALBERT I, LAM S K, et al. MovieLens unplugged: experiences with an occasionally connected recommender system[C]// Proceedings of the 8th International Conference on Intelligent User Interfaces. ACM, 2003: 263-266.

[13] LUO X, OUYANG Y X, XIONG Z, et al. The effect of similarity support in K-Nearest-Neighborhood based collaborative filtering [J]. Chinese Journal of Computers, 2010, 33(8): 1437-1445. (in Chinese)
罗辛，欧阳元新，熊璋，等. 通过相似度支持度优化基于 K 近邻的协同过滤算法[J]. 计算机学报，2010，33(8)：1437-1445.

[14] STECK H. Evaluation of recommendations: rating-prediction and ranking[C]// Proceedings of the 7th ACM Conference on Recommender Systems. ACM, 2013: 213-220.

[15] SHI Y, LARSON M, HANJALIC A. Unifying rating-oriented and ranking-oriented collaborative filtering for improved recommendation[J]. Information Sciences, 2013, 229(6): 29-39.

[16] Apache. Mahout[EB/OL]. [2016-06-03]. http://mahout.apache.org.

[17] MovieLens datasets [EB/OL]. [2016-06-16]. http://grouplens.org/datasets/movielens.

[18] KOREN Y, BELL R M, VOLINSKY C. Matrix factorization techniques for recommender systems[J]. IEEE Computer, 2009, 42(8): 30-37.