DOI:10.3969/j.issn.1673-4785.201405066
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20150302.1106.006.html

基于概念簇的多主题提取算法

马甲林[1,2],张永军[1,2],王志坚[1]
(1.河海大学 计算机与信息学院,江苏 南京 211100; 2. 淮阴工学院 计算机工程学院,江苏 淮安 223003)

摘 要:现实世界存在着大量的多主题文本,多主题在信息检索、图书情报等领域有着广泛的应用。传统主题提取算法大多是针对文本整体提取一个主题,且存在缺乏语义信息、向量高维和稀疏等缺陷。以《知网》为知识库,构建概念向量表示文本,根据概念的语义及上下文背景对同义词进行归并、对多义词进行排歧,并利用概念间语义关系实现语义相似度计算;在此基础上提出基于概念簇的多主题提取算法 MEABCC,该算法通过对概念进行聚类,得到多个主题簇;在使用 K-means 算法进行概念聚类时,通过"预设种子"方法对其进行改进,以弥补传统 K-means 算法对初始中心的敏感性所引起的时空开销不稳定、结果波动较大的缺陷。实验结果表明,该算法具有较好的准确率、召回率和 F_1 值。

关键词:语义;稀疏;上下文背景;知识库;概念簇;多主题提取; K-means;MEABCC
中图分类号:TP18 **文献标志码**:A **文章编号**:1673-4785(2015)02-0261-06

中文引用格式:马甲林,张永军,王志坚. 基于概念簇的多主题提取算法[J]. 智能系统学报, 2015, 10(2): 261-266.
英文引用格式:MA Jialin, ZHANG Yongjun, WANG Zhijian. Multi-topic extraction algorithm based on concept clusters[J]. CAAI Transactions on Intelligent Systems, 2015, 10(2): 261-266.

Multi-topic extraction algorithm based on concept clusters

MA Jialin[1,2], ZHANG Yongjun[1,2], WANG Zhijian[1]
(1. College of Computer and Information, Hohai University, Nanjing 211100, China; 2. School of Computer Engineering, Huaiyin Institute of Technology, Huaian 223003, China)

Abstract: There are a large number of multi-topic documents existing in the real world, and the extraction of multi-topic is widely used in the fields of information retrieval, library science and intelligence. In the traditional theme extraction algorithm, in most cases a theme is extracted for the whole text, which lacks of semantic information and has high-dimensional vector and sparse defects. Setting concept vectors to represent text based on the repository of cnki.net, merging synonyms and discriminating polysemy according to the semantic of concepts and context, thereby achieving the computation of semantic similarity in light of the semantic relation among concepts. The multi-topic extraction algorithm based on the concept of clusters (MEABCC) is proposed. The MEABCC acquires multiple topics by clustering concepts. The conceptual clustering made by K-means algorithm is improved through the method of presetting "default seed", which makes up the undulating time and space overlay and the unstable results. This happen to be caused by sensitivity to initial centers of traditional K-means algorithm. The experiments showed that MEABCC has good accuracy, recall and F_1 values.

Keywords: semantic; sparsity; context; knowledge base; concept clusters; multi-topic extraction; K-means; MEABCC

收稿日期:2014-06-01. **网络出版日期**:2014-03-02.
基金项目:国家自然科学青年科学基金资助项目(11201168).
通信作者:马甲林. E-mail:majialin@126.com.

现实世界存在着大量的多主题文本,据统计36.85%文章包含多个主题,Sekine 和 Nobata 主持的

一项研究表明,日本新闻文章中的 44.62%在谈论多个话题。从文本中提取反映不同观点的多个子主题,在信息检索、图书情报和信息安全等领域有着非常广泛的应用[1-2]。大多数传统主题提取方法是针对一篇文章从整体考虑提取一个主题,未能区分出文内混杂的多个子主题,文献[3]认为子主题体现在主观句子的语义中,提出 CRF 模型从主观句子的极性角度提取子主题,该方法以形容词、副词词性判断句子语义的贬褒极性,未涉及其他语义信息;文献[4]使用滑动窗口的方法可以从网络评论文本提取局部子主题,适用于网络评论文本;另外,常用的 LDA(latent Dirichlet allocation)模型提出于 2003 年,该模型虽然目前使用广泛,但 LDA 是一个完全基于统计的方法,在向量空间模型(VSM)下存在向量高维和稀疏、忽略词汇语义及上下文背景等问题,同时提取过程受到同义词和多义词的干扰,因而在质量和效率上表现欠佳[3-5]。

本研究利用《知网》知识库,采用概念向量模型(CVM)取代传统 VSM 模型表示文本,同时在 CVM 模型下同义词将被自动归并,再根据上下文语义相关性对多义词进行排歧处理;其次通过计算概念的语义相似度取代传统相似度计算,在此基础上提出基于概念簇的多主题提取算法(MEABCC),该算法采用无监督学习的方法,通过改进经典 K-means 算法对文本概念进行聚类后得到多个子主题簇,其中,使用"预设种子"方法改进来 K-means 算法,以弥补传统 K-means 算法 K 个初始中心选择的随机性所引起的时空开销不稳定、结果波动较大的缺陷。

1 概念向量模型

文本处理的首要问题是文本表示,本研究以中科院计算机语言信息工程研究中心董振东主持创立的《知网》为知识库,建立基于概念的向量模型来表示文本。

1.1 同义词和多义词处理

《知网》是一个以汉语和英语词汇所代表的概念为描述对象,以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。在《知网》中,词汇语义描述被定义为概念 。每一个词可以表达为几个概念,概念是由一种知识表示语言(DEF)来描述,这种用来描述概念的"词汇"又叫义原,相比词汇的规模,义原的数量很少。《知网》定义了 1 500 多个义原,分为 3 类:基本义原、语法义原和关系义原,DEF 中基本义原反映了概念的主要语义,例如:词汇"爱好者",在《知网》中用 DEF 的基本义原为:DEF = {Human|人, * Fondof|喜欢,# WhileAway|休闲},所表达的意思是:"爱好者"是个人,这个人喜欢某个东西,本词语是和休闲相关[7],它们之间存在语义相关性。在《知网》中,如果某个词只有一个意思,那么这个词对应唯一的概念,而多义词往往对应多个概念,为了找到某个多义词在文中的具体含义,作如下定义:

定义 1 对于任意中文词汇 c_0,在《知网》中描述其对应概念的 DEF 的基本义原集为 $\{c_1, c_2, \cdots, c_m\}$,$(m >= 1)$ 则称 c_0 与 $\{c_1, c_2, \cdots, c_m\}$ 属于同一个语义类。

语义类不仅与概念对应,而且与描述概念的 DEF 对应,语义类揭示了词语之间的语义联系,描述某个 DEF 的基本义原在语义上是相关的,某个语义类和文章语境相符时,文中很可能出现该语义类包含的词汇,利用这一语言现象可以消除词汇歧义。如图 1:多义词"水分",在语义类包含{"植物"、"土壤"、"阳光"、"生长"}中"水分"的含义是指"物体内含有的水",而在语义类包含{"经济"、"数据"、"增长"、"报告"}中"水分"的含义是指"夹杂不真实成分"。

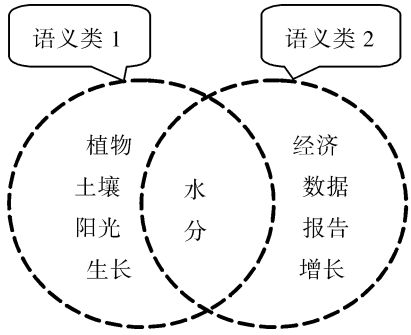

图 1 "水分"语义类示意图

Fig.1 The semantic class schematic diagram of 'moisture'

由于汉语的复杂性,同一篇文章中一词多义和同义词的情况非常多,单纯的机械词频统计方法无法处理涉及词汇语义的问题,这是影响文本主题提取质量的一个重要因素。为了解决多义词排歧和同义词归并问题,本研究利用《知网》,同义词在概念映射阶段被归并到同一概念上;多义词对应多个概念,根据语义类成员词和上下文背景的语义相关性来为多义词选择适合该文语境的语义类。定位多义词在文中最佳语义类的思路是:如果某个语义类所属成员词汇在本篇文中出现权值之和越大,说明该语义类比其他语义类更符合文章主题,则该语义类是该多义词的在此文中最合适的语义类。词汇 w_i 在文章中所含的信息量 $H(w_i)$ 计算公式为[7]

$$H(w_i) = -\mathrm{TF}(w_i, \mathrm{ST}) \times \log[p(w_i)] \quad (1)$$

式中:ST 表示待处理文本,TF(w_i,ST)表示词汇 w_i 在文中出现的频率,$P(w_i)$为词 w_i 的概率分布。

定义 2　多义词 c,它的第 i 个语义类 L_i 权值为[7]

$$C_{\text{Weight}}(L_i) = \sum_{j=1}^{n} H(w_j) \times \log_2 n \qquad (2)$$

式中:n 为某个语义类 L_i 成员词在文中出现的个数。语义类权值越大,该语义类成员词对文章主题的贡献越大。

定义 3　多义词 c,在《知网》中对应多个语义类,选择符合该文背景的最佳语义类公式为

$$\text{Best}_{cL_i} = \text{Max}(C_{\text{Weight}}(L_i)) \qquad (3)$$

1.2　概念向量构建算法

传统基于特征词的向量空间模型(VSM),认为向量是正交的,即词汇之间互不相关。显然,这和现实情况不符,众所周知,文献中各个词汇之间存在着复杂的语义联系[5]。利用《知网》知识库,构建概念向量模型来表示文本,可以建立起词汇之间语义联系,为后续进一步的语义计算提供了可能。CVM 构建过程首先对文本进行分词和预处理后得到文本的特征集,然后对特征集中的每个特征进行概念映射;特征词到概念的映射过程中大量的同义词被归并到相同的概念中,实现了强度较大的降维;其次利用《知网》概念描述语义的特点,根据语义类和上下文背景的相关性,实现多义词排歧,其构建算法如下。

算法 1　概念向量构建算法

输入:文本 T;

输出:文本 T 的概念向量 $\boldsymbol{T}$。

$$\boldsymbol{T} = [(G_1,(C_1,\cdots,C_i))\quad (G_2,(C_2,\cdots,C_j))\quad \cdots\quad (G_q,(C_q,\cdots,C_k))]。$$

步骤如下:

1)使用中科院 ICTCLAS 分词系统对 T 进行分词后得 $\boldsymbol{T}=[C_1\quad C_2\quad \cdots\quad C_n]$;

2)利用信息增益(IG)初步提取 T 特征后得到 $\boldsymbol{T}=[C_1\quad C_2\quad \cdots\quad C_m]$,其中 $m\leqslant n$;

3)依次查询《知网》知识库,对特征词进行概念映射;

①查询《知网》,若 T 的特征词 C_m 对应唯一的概念,则 C_m 为单义词或同义词,直接获取 C_m 的概念,转至 4);

②:若 C_m 对应多个概念,则 C_m 为多义词,所以 C_m 对应多个语义类表示为 $\{L_1,L_2,\cdots,L_p\}$ $(p\geqslant 1)$,采用如下步骤为 C_m 进行多义词排歧:

```
For i=1 to p
  {
利用式(1)计算语义类 L_i 所有成员词汇的信息量;
   利用式(2)计算 L_i 权值;
  }
Next i;
```

利用式(3)为 C_m 选择最佳语义类,最终将 T 的所有特征映射为概念,$\boldsymbol{T_G}=[(C_1,G_1)\quad (C_2,G_2)\quad \cdots\quad (C_m,G_m)]$;

4)对 $\boldsymbol{T_G}$ 按照概念进行整理合并得到:

$$\boldsymbol{T_G} = [(G_1,(C_1,\cdots,C_i))\quad (G_2,(C_2,\cdots,C_j))\quad \cdots\quad (G_q,(C_q,\cdots,C_k))]$$

式中:G_q 为 $\boldsymbol{T_G}$ 集合中无重复的概念,$q,i,j,k\leqslant m$;//现实同义概念的归并;

5)输出文本 T 对应概念向量 $\boldsymbol{T}$

$$\boldsymbol{T} = [(G_1,(C_1,\cdots,C_i))\quad (G_2,(C_2,\cdots,C_j))\quad \cdots\quad (G_q,(C_q,\cdots,C_k))]。$$

2　多主题提取算法

对于单主题提取,机械统计的主题提取方法通过词频统计按照权值大小抽取主题句,能够得到质量达到简单应用级别的主题句[6]。然而,现实中存在着大量的多主题文献,单纯的统计方法无法抽取多主题。本研究提出的 MEABCC 多主题提取方法是以 1.2 节提出的概念向量来表示文本,利用《知网》中义原的树形层次体系结构计算义原相似度,进而计算概念的相似度,然后通过改进 K-means 算法对组成文本的概念进行聚类,形成多个子主题概念簇。

2.1　概念相似度计算

相似度是衡量 2 个词汇语义关系的一个重要指标,涉及到词语的词法、句法、语义甚至语用等多方面的信息。其中,对词语相似度影响最大的是词的语义。在《知网》中,词汇被描述为概念,词汇的相似度计算就转化为对概念的相似度计算。词语距离与词语相似度之间有着密切的关系。2 个词语的距离越大,其相似度越低;反之,2 个词语的距离越小,其相似度越大[8]。

《知网》通过多个义原来描述概念,义原之间存在着各种复杂的关系,如:上下位关系、同义关系、对义关系等。其中,最重要的是上下位关系,所有的义原根据上下位关系构成了一个树状的义原层次体系,所以可以通过计算义原距离得到概念的距离进而获得概念的相似度[9]。假设 2 个义原在义原树

层次体系中的路径距离为 d,d 的计算过程如下:

设义原集中的任意一个义原为 w_i,L_i 为义原 w_i 在概念树中的深度,a 为距离初始阈值,b 为满足不等式 $\max(L)<a/b$ 的一个正实数,则 w_i 与其父节点的距离定义为[9]

$$d(w_i,\text{parent}(w_i)) = a - L_i b \tag{4}$$

任意 2 个义原 w_i、w_j 之间的距离定义为[9]

$$d(w_i,w_j) = \omega_k.[a - \max(L_i,L_j)b] \tag{5}$$

式中:ω_k 表示第 k 种关系对应的权重,通常取 $\omega_k \geqslant 1$。可以验证,上述定义符合对距离函数的数学要求,式(4)、(5)反映出义原在义原层次树中的位置越深,二者之间的距离越小,即越相似。

定义 4 任意 2 个义原 (w_i,w_j) 之间的语义相似度为[9]

$$\text{Sim}(w_i,w_j) = \frac{\theta}{d(w_i,w_j) + \theta} \tag{6}$$

式中:d 是 w_i 和 w_j 在义原层次体系中的路径长度,是一个正整数。θ 是一个可调节的参数。

定义 5 设概念 U 和 V 分别由义原组 $(p_{u1}, p_{u2},\cdots, p_{un})$ 和 $(p_{v1}, p_{v2},\cdots, p_{vm})$ 描述,则 U、V 相似度为

$$\text{Sim}(U,V) = \frac{(U,V)}{\sqrt{(U,U)\cdot(V,V)}} \tag{7}$$

式中:$(U,V) = \sum_{i}^{n}\sum_{j}^{m}\text{Sim}(p_{u_i},p_{v_j})$。

定义 6 概念 U 由义原组 $(p_1, p_2,\cdots, p_n)$ 表示,概念集 C 由概念集合 $\{C_1,C_2,\cdots,C_m\}$ 组成,概念 U 和概念集 C 的相似度定义为 U 和 C 中所有概念相似度的最大值[7]:

$$\text{Sim}(U,C) = \text{Max}\{\text{Sim}(U,C_i) \mid C_i \in C\} \tag{8}$$

2.2 MEABCC 算法

当前主题提取的方法主要有 2 类:基于机械统计的方法和基于语法语义分析的方法。统计的方法能够有效利用文章表层信息抓住文章关键词汇,收集文章原句输出主题,优点是通用性好,适用于非受限区域,然而,其几乎完全忽略词汇语义信息,难以得到质量较高的主题,且不易提取多主题。基于语法语义分析的主题提取方法被认为比传统的基于机械统计的方法更符合语言规律,提取的主题质量较高,但其要求极高的人工智能技术和完备的专家系统,以及领域受限等问题导致应用困难[10]。

本研究提出了基于概念簇的多主题提取算法(MEABCC),其思路是:利用《知网》知识库丰富的语义信息,将文本表示成为概念向量模型,改进 K-means 算法对概念进行语义聚类,形成多个子主题概念簇,进而得到文章对应的多个子主题关键词集。

聚类算法有很多种,最典型有效的划分法之一是 K-means,K-means 算法是从样本中随机取出 K 个样本作为初始聚类中心,再通过迭代,计算每个类的中心,每个样本被归入到最近的中心,重新计算类中心,直到类中心不再改变。使用 K-means 算法进行聚类,首先要选取 k 个点作为初始聚类中心,然后进行反复的迭代,由于初始中心选择具有随机性,会导致结果和耗时随不同的初始输入而波动,从而引起算法不可预测的复杂度[11]。为了解决这一问题,借鉴传统基于统计的主题提取思想,文章的主题很大程度上反映在词共现上,做进一步的延伸,文章中的同义词往往围绕某一个主题,而同义词在概念向量模型中表现为同一个概念,因而在多主题提取中,本研究提出根据概念向量中每个概念包含文章词的个数大小进行排序,选取包含文章词个数最多的前 K 个概念作为 K-means 聚类的初始中心的"预设种子",这种方法可以克服 K-means 算法的对初始中心的敏感性。

定义 7 概念集 C 由概念集合 $\{C_1,C_2,\cdots,C_n\}$ 组成,则 C 中心点为[7]

$$\text{certer}C = \frac{\sum_{i=1}^{n} C_i}{n} \tag{9}$$

式(9)由计算文本集合中心点的方法所得。

基于概念簇的多主题提取算法具体步骤如下:

算法 2 基于概念簇的多主题提取算法

基本流程如下:

输入:文本 T,聚类的个数参数 k,主题个数 k_1,其中 $k_1<k<n$;

输出:T 的 k_1 个子主题句集合 $\{(st_{11},st_{12},\cdots,st_{1u}),(st_{21},st_{22},\cdots,st_{2v}),\cdots(st_{k11},st_{k12},\cdots,st_{k1w})\}$。

步骤如下:

1)调用算法 1 得到文本 T 的语义概念向量 $\boldsymbol{T}$

$$\boldsymbol{T} = [(G_1,(c_1,\cdots,c_i)) \quad (G_2,(c_2\cdots,c_j)) \quad \cdots \quad (G_q,(c_q,\cdots,c_k))]$$

2)从 $\boldsymbol{T}$ 中选择 K 个包含词汇数目最多的概念作为聚类初始类中心 $\boldsymbol{T_Z} = [G_{m1} \quad G_{m2} \quad \cdots \quad G_{mk}]$,其中 $mk \leqslant q$;

3)根据相似度计算式(8)计算每个概念与 K 个类中心的相似度,将对应概念分配到相似度最大的类中;

4)利用式(9)重新计算各类的中心点;

5)重复 3)和 4)直到类的中心点不再改变,得

到 K 个类别的概念集:{ Φ_1 },{ Φ_2 },…,{ Φ_k } };

6)选择包含概念个数最多的前 k_1 个概念集合,得到组成 k_1 个子主题的概念集合:{ Φ_1 },{ Φ_2 },…,{ Φ_{k_1} } },进而得到 k_1 子主题对应文章中 k_1 个关键词汇集合:{ $(c_{11},c_{12},\cdots,c_{1t})$,$(c_{21},c_{22},\cdots,c_{2j})$,… $(c_{k_11},c_{k_12},\cdots,c_{k_1t})$ }。

3　实验及结果分析

目前还没有已标注主题的中文文本标准语料库,复旦大学自然语言处理实验室的公开的标准语料库共包含 20 个类别,19 637 篇文档,但均未标注主题,考虑到工作量因素,本研究从该语料库 5 个类别中选择篇幅较长、多主题特征较为明显的 500 篇文档,经从事汉语言工作的专业人员进行多主题词标注后作为实验样本。实验结果评判采用通用的准确率(P)、召回率(R)和综合指标 F_1。

$$P=\frac{\text{反映主题的主题句数量}}{\text{抽取出的主题句总数}}$$

$$R=\frac{\text{抽取出的主题句的数量}}{\text{文本中实际主题句总数}}$$

$$F_1=\frac{2PR}{P+R}$$

3.1　参数估计

为了得到算法 2 中初始聚类簇参数 k 的最恰当的值,根据测试样本的实际篇幅长短、文章结构等情况,经汉语专业人士分析,每篇样本抽取子主题个数 k_1 的值取 3,并人工为每篇样本标注了 3 个子主题作为标准值,在 $k_1=3$ 的情况下实验分析 k 取值,图 2 反映出 k 在不同取值下准确率、召回率和 F_1 的变化情况。

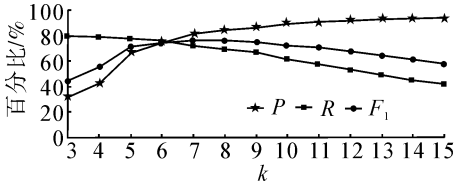

图 2　不同 k 值下 P、R 和 F_1 变化

Fig.2　The accuracy and recall rate and F_1 under different k

由图 2 可以看出,每篇样本抽取 3 个子主题的情况下,MEABCC 算法随着 k 值的增大提取主题的准确率不断提高,而召回率在降低,这是由于 k 值增大导致聚类簇细化,所以准确率逐渐上升;通常情况下算法召回率是确定的,但在本实验中,随着 k 值的增大类别不断细化,在选取前 3 个($k_1=3$)最大子主题的时,引起了召回率下降;为了找到最合适的 k 值,分析图 2 的 F_1 指标,从综合指标 F_1 的趋势上看,F_1 的最高点出现在 $k=7$ 时,所以算法 2 在本实验样本对象下最适合的取值是 $k=7$,需要说明的是 k 的取值是和要处理的文章的有关。

3.2　算法测试

为了测试通过“预设种子”的方法改进 K-means 算法提取多主题的质量,实验样本仍然为预备的 500 篇文档,采用 3.1 节参数实验中获得的结果,取 $k=7$,子主题个数 k_1 为 3,首先采用传统 K-means 算法,随即产生 k 个初始中心的方法实验 5 次,和 MEABCC 提取主题结果统计如表 1 所示。

表 1　K-means 和 MEABCC 多主题提取结果统计

Table 1　K-means and MEABCC more topic extraction result statistics

算法\指标	次数	准确率/%	召回率/%	F_1/%	耗时
K-means	第 1 次	61.3	56.8	59.0	3′51″
	第 2 次	76.8	65.1	70.5	6′73′
	第 3 次	49.4	52.3	50.8	5′21″
	第 4 次	78.9	57.7	66.7	8′01″
	第 5 次	50.1	68.0	57.7	4′21″
MEABCC	1 次	81.7	68.9	74.8	3′39″

从表 1 数据可以看出,传统 K-means 在 5 次随即产生初始中心的情况下,结果的准确率、召回率以及综合指标 F_1 值都非常不稳定,算法耗时变化较大,这是由于传统的 K-means 算法对初始聚类中心较敏感,导致结果和耗时随不同的初始输入波动较大。为消除这种缺陷,本研究结合主题提取特点,每个主题往往包含多个具有相同概念的词,概念成员词构成了一个围绕该概念的语义中心,因而可根据概念在文中出现成员词的数量大小,预设出可能性最大的 K 个初始中心,从而改进 K-means,不但提取的主题质量较高,算法的执行效率也有较大的提高。

4　结束语

向量空间模型下的传统主题提取方法忽略词语间的语义联系,缺乏语义信息,提取的主题质量不高,不适合提取多主题。本研究利用《知网》,构建概念向量模型来表示文本,对同义词进行归并,对多义词进行语义排歧;实现了概念的语义相似度计算;采用无监督学习的方法,提出基于概念簇的多主题提取算法(MEABCC),该算法通过合理“预设初值”,改进经典 K-means 后对概念进行聚类,得到多个子主题簇。实验测试结果反映出 MEABCC 算法效果和效率均较优。

参考文献：

[1] TANG Jie, YAO Limin, CHEN Dewei. Multi-topic based query-oriented summarization[C]//Proceedings of the SIAM International Conference on Data Mining. Sparks, USA, 2009: 1141-1152.

[2] LAMIREL J C. Multi-view data analysis and concept extraction methods for text[J]. Knowledge Organization, 2013, 40(5): 305-319.

[3] NA Fan, LI Huixian, and WANG Chao. Research on sentiment analyzing in multi-topics texts[J]. Advances in Computer Science, Intelligent System and Environment, 2013, 105: 581-586.

[4] FU Xianghua, LIU Guo, GUO Yanyan, et al. Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon[J]. Knowledge-Based Systems, 2013, 37: 186-195.

[5] ZENG Jianping, DUAN Jiangjiao, WANG Wei, et al. Semantic multi-grain mixture topic model for text analysis[J]. Expert Systems with Applications, 2011, 38: 3574-3579.

[6] 刘金岭.基于降维的短信文本语义分类及主题提取[J]. 计算机工程与应用, 2010, 46(23):159-161.
LIU Jinling. Dimensionality reduction of short message text classification and thematic extraction of semantic[J]. Computer Engineering and Applications, 2010, 46(23): 159-161.

[7] 白秋,金春霞,周海岩.概念向量文本聚类算法[J]. 计算机工程与应用, 2011, 47(35): 155-157.
BAI Qiuchan, JIN Chunxia, ZHOU Haiyan. Text clustering algorithm based on concept vector[J]. Computer Engineering and Applications, 2011, 47(35): 155-157.

[8] 江敏,肖诗斌. 一种改进的基于《知网》的词语语义相似度计算[J]. 中文信息学报, 2008, 22(5): 84-89.
JIANG Min, XIAO Shibin. An improved word similarity computing method based on HowNet[J]. Journal of Chinese Information Processing, 2008, 22(5): 84-89.

[9] 刘金岭.基于语义的高质量中文短信文本聚类算法[J]. 计算机工程, 2009, 35(10): 201-205.
LIU Jinling. High quality algorithm for chinese short messages text clustering based on semantic[J]. Computer Engineering, 2009, 35(10): 201-205.

[10] LLORET E. Manuel palomar text summarisation in progress: a literature review[J]. Artificial Intelligence Review, 2012, 37: 1-41.

[11] XU Junling, XU Baowen, et al. Stable initialization scheme for K-means clustering[J]. Wuhan University Journal of Natural Sciences, 2009, 14: 24-28.

作者简介：

马甲林，男，1981 年生，博士研究生，主要研究方向为自然语言处理。曾获第 12 届全国多媒体课件大赛三等奖、江苏省高等学校优秀多媒体教学课件二等奖、淮安市科技进步奖三等奖、发明专利 1 项、参编教材 1 部，发表学术论文 7 篇。

张永军，男，1978 年生，讲师，博士研究生，主要研究方向为中文信息处理、文本数据挖掘、发表学术论文 8 篇，参编教程 1 部。

王志坚，男，1958 年生，教授，博导，主研方向为基于网络的计算机应用技术、软件复用、基于网络的软件系统集成技术，主持国家“863”项目、江苏省基金项目等多项，出版专著多部。