基于概念的中文博客情感极性聚类分析①

申　莹，徐东平，庞　俊

(武汉理工大学 计算机科学与技术学院，武汉 430063)

摘　要：通过基于概念的聚类方法，对博客作者的情感极性进行分析。在知网情感词汇库的基础上，将概念引入向量空间模型。首先，提取博客文本情感词，利用基于情感词概念的向量空间模型完成对博客文本的表示。然后，使用 k-means 算法对博客文本进行聚类，完成对博客情感极性的分析。在向量空间模型中使用概念作为特征项，提高了对博客作者情感极性分析的精度。实验证明基于概念的向量空间模型比传统基于词语的向量空间模型在博客文本情感聚类上具有更好的性能。

关键字：概念；向量空间模型；知网；情感极性；聚类分析

Clustering Analysis of Sentiment Polarity for Chinese Blogs Based on Concept

SHEN Ying, XU Dong-Ping, PANG Jun

(Department of Computer Science and Technology, Wuhan University of Technology, Wuhan 430063, China)

Abstract: A clustering method based on concept was provided to analyse the sentiment polarity for Chinese Bloggers. The concept is introduced into Vector Space Model (VSM) on the basis of HowNet. Firstly, sentiment words are extracted from blog texts which would be expressed by VSM with the concept of sentiment words. Secondly, blog texts are clustered with k-means algorithm to finish the analysis of sentiment polarity for Chinese Blogs. The precision of sentiment polarity analysis of Chinese Blogs is improved with concept as feature in VSM. The experiment proves the concept based VSM to be of better performance than traditional term based VSM in clustering analysis of Chinese Blogs on sentiment polarity.

Key words: concept; vector space model (VSM); HowNet; sentiment polarity; clustering analysis

1　引言

随着 Web2.0 的迅速发展，博客（Blog）逐渐成为 Internet 上一种重要的舆论媒体。在 Blog 中，网民可以任意发表对时事的看法，其包含了大量的情感极性信息。所谓情感极性是指人对客观事物的好、恶，褒、贬，支持、反对等态度。肯定的情感极性表达了一种愿望、需求，或赞美、肯定的感情；否定的情感极性则表达了一种不愉快或不受欢迎的感情[1]。通过对 Blog 作者情感极性的分析，可以掌握广大网民对时事的态度，是掌握网络舆情的一个重要途径。本文通过将 Blog 文本情感词映射到概念层，用聚类方法分析 Blog 作者的情感极性。基于概念的博客情感极性聚类分析模型如图 1 所示。

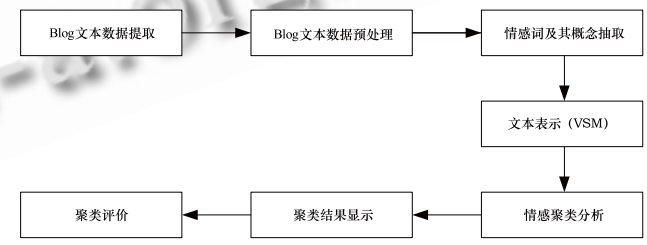

图 1　基于概念的中文博客情感极性聚类分析模型

由图 1 可知，本文首先提取 Blog 文本数据，并对数据进行处理，然后基于知网情感词汇库从博客文本中提取情感词及概念，在文本表示时将情感词概念引入到向量空间模型。建立基于概念的向量空间模型后，进行情感极性的聚类分析，显示聚类图 1 基于概念的

① 收稿时间:2010-11-15;收到修改稿时间:2010-12-24

中文博客情感极性聚类分析模型结果，最后评价聚类结果。传统的聚类方法采用向量空间模型表示文本，以单个的词语作为特征项，忽略了每个特征项之间的语义联系，词频维数过高，聚类算法计算复杂度较高。在本文中，用这种基于概念的聚类方法来分析 Blog 作者情感极性，降低了聚类时间复杂度，也增加了情感极性分析的精度。本文将首先介绍基于概念的向量空间模型，然后再讲解中文博客情感极性聚类分析。

2 基于概念的向量空间模型

建立基于概念的向量空间模型，其中情感词的概念来源于知网提取，故本部分先介绍知网。

2.1 知网简介

知网英文名称为 HowNet，是一个以汉语和英语的词语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[2]。HowNet 中有两个主要的概念：“概念”和“义原”。“概念”是对词汇语义的一种描述，每个词可以表达为几个概念（即一词多义）。“概念”是一种知识表示语言来描述的，这种知识表示语言所用的词汇叫做“义原”。 “义原”是用于描述一个“概念”最小意义的单位，通过有限的义原，组合成一个无限的概念集合。HowNet 中描述一个词语的格式如下：

W_X= 词语 ；

G_X= 词语词性；

E_X= 词语例子；

DEF= 概念定义。

通过处理，选取 HowNet 中每个词语信息中“W_X= 词语 ；G_X= 词语词性；DEF= 概念定义”这三部分信息，完成情感词概念的提取。

2.2 基于概念的向量空间模型的建立

2.2.1 Blog 文本数据获取及预处理

图 1 中，首先根据确定的话题（关键词），通过 Google Blog Search 搜索所需的博客文本。然后，通过爬虫爬取一定数量的博客文本。博客搜索结果的标题和摘要是和话题密切相关的，而且比较集中的包含博客作者的情感和观点，本文选取这两部分作为研究对象。

Blog 文本数据的预处理主要是通过采用中国科学研究院开发的 ICTCLAS 分词系统[3]来完成的。ICTCLA 分词系统是一款使用最广泛的分词系统之一，分词性能优越，准确率超过 98%。通过分句、分词和去除停用词步骤，完成对数据的预处理[4]。

2.2.2 情感词概念获取

本文主要研究的是博客文本的情感极性，经过预处理后，需要基于 HowNet 词典抽取博客文本中的情感词。HowNet 中包含了 4566 个正向的中文情感词和 4370 个负向的中文情感词。先建立一个情感词典的情感词表，将这些情感词放入词表中。通过查表的方式判断通过处理后的分词文本中的词是否是情感词，若是则提取情感词，然后通过 HowNet 词典提取情感词的 DEF 项；否则返回分词文本继续判断下一词语。该获取情感词概念流程如图 2 所示。情感词的一词多义现象不是很多，本文选取情感词的第一义项作为情感词的概念。

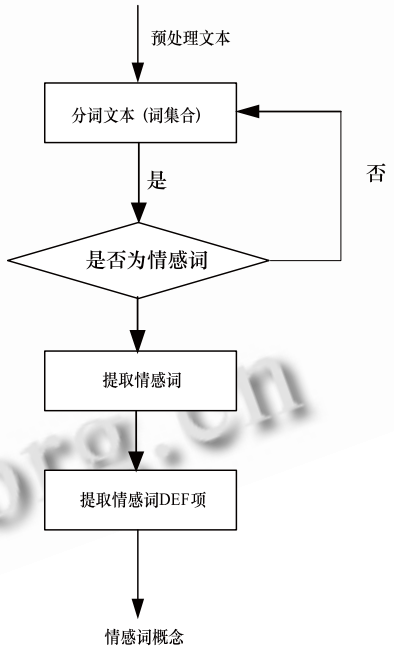

图 2 获取情感词概念流程图

2.2.3 构造基于概念的向量空间模型

有了前面的基础工作，可以将提取的情感词概念作为特征项来表示每一个博客文本。向量空间模型是使用最为广泛的文本表示模型，本文选取向量空间模型（Vector Space Model ，VSM）来表示博客文本。

向量空间模型（VSM）是 Salton G 等人在 20 世纪 60 年代提出的，其思想是把文本表示成由特征项构成的向量空间中的一个点，通过计算向量之间的距离判断文本之间的相似程度[5]。在传统基于词语的向量

空间模型中，因为没有考虑词之间存在的概念相似情况，所以影响了数据聚类的准确性[6]。本文引入词语的概念到向量空间模型，构造基于概念的向量空间模型。以概念作为特征项，而不是用单个词语，这样有效的降低了文本向量模型的维数，提高了特征提取的精度和效率。

传统的用词语为特征项的 VSM 为 $V_i(W_1, W_2, \cdots, W_n)$，为 w_i 第 i 个词语特征项的权重。由概念作为特征项的 VSM 表示为：$V_j(C_1, C_2, \cdots, C_n)$（其中 $k \le n$），C_j 为第 j 个概念特征项的权值。计算情感词概念特征项权值采用经典的 TF-IDF[7]公式计算，把词语的频率换成概念的频率，如公式(1)所示。

$$C(t,d)=\frac{tf(t,d)\times\log(N/n_t+0.01)}{\sqrt{\sum_{t\in d}[tf(t,d)\times\log(N/n_t+0.01)]^2}} \quad (1)$$

其中，$C(t,d)$ 表示概念特征项 t 在博客文本 d 中的权重。$tf(t,d)$ 表示概念特征项 t 在文本 d 中的概念频率，N 为博客文本的总数，n_t 为文本集中出现概念特征项 t 的文本数，分母为归一化因子。

3 Blog情感极性聚类分析

构建了基于概念的向量空间模型，所有博客文本转化为向量的形式。也就是博客文本表示成空间中的一个个的数据点，对博客文本的情感极性的聚类也就是对这些数据点进行聚类。将具有情感极性的博客文本通过聚类，分成不同的簇，完成对博客情感极性的分类。

3.1 k-means 算法情感极性聚类分析

k-means 的简洁和效率使得其成为所有聚类算法中最为广泛使用，适用于大规模数据集。将博客文本表示成基于概念的向量，采用经典 k-means 算法对博客文本进行聚类分析。k-means 算法是一种基于划分方法的聚类算法。其主要的思想是：先随机挑选出 k 个数据对象作为初始的簇中心，然后计算其他数据对象到初始簇中心点的距离，将这些数据点分配到距离最近的初始簇中心所在的簇中。全部数据点被分配后，每个聚类的簇中心根据聚类中现有的数据点重新计算簇中心，分配其他数据对象。如此循环往复，直至满足算法某个终止条件。算法结束条件有三个，只需要满足其中一个即终止：

1) 每个簇的成员（数据点）没有发生变化；

2) 每个簇中心发生变化；

3) 误差平方和（Sum of Squared Error, SSE）局部最小。

误差平方和（SSE）计算公式（2）如下：

$$SSE=\sum_{j=1}^{k}\sum_{X\in Cj}dist(X,m_j)^2 \quad (2)$$

其中 k 表示需要的聚类数目，C_j 指第 j 簇，mj 表示 C_j 的簇中心。一般，用某个簇的所有数据点的平均向量来计算该簇的簇中心。X 表示输入的数据集中的一个数据点。$dist(X,mj)$ 表示数据点 X 与簇中心 mj 的距离[8]。

本文为方便实验结果评价，令 k=3。就是将博客文本的情感极性分为三类：正向、负向和中立。

3.2 文本相似性

文本相似性的度量是文本聚类的主要依据。博客文本表示成为向量后，文本的相似性的度量转换成向量的相似性度量。最常用的文本相似性度量的方法有夹角余弦表示法和欧几里得距离表示法[9]。夹角余弦相似函数公式(3)如下：

$$Sim(X_1,X_2)=\frac{X_1\bullet X_2}{\|X_1\|\times\|X_2\|} \quad (3)$$

X_i，X_j 表示两个文本向量，$\|X_i\|$（i = 1，2）表示向量 X_i 的模，也就是文本 X_i 中含有概念特征项的数量。上述公式的“•”表示两个文本特征向量的内积。

欧几里德距离函数计算公式公式（4）如下：

$$dist(X_i,X_j)=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\cdots+(x_{ir}-x_{jr})^2} \quad (4)$$

X_i，X_j 分别表示两个文本向量（$x_{i1}, x_{i2}, \cdots, x_{ir}$）和（$x_{j1}, x_{j2}, \cdots, x_{jr}$）。

本实验采取夹角余弦相似函数计算方法来表示文本的相似性。

4 实验结果与分析

本实验使用的是一台个人计算机，该机器的配置为：Windows XP 操作系统，Intel（R）Celeron（R）CPU 2.26GHz，512MB 内存。并采用 Microsoft Visual C++6.0 集成开发环境作为编程工具。

为了评价聚类结果，人工标注所有的博客搜索结果。首先将搜索结果标注为两类：相关和不相关。然后，相关的搜索结果根据情感极性的不同标注为三类：正向，中立和负向。表达赞成，表扬，支持，肯定等情感的结果标注为正向；表达反对，批评，否定等情

感的结果标注为负向；没有表达情感的结果标注为中立。

4.1 实验步骤

本实验的具体步骤按照图 1 所示，主要概括如下：

1）Blog 文本数据采集

本实验主要采集的数据是使用 Google 博客搜索引擎查询话题“建国大业”，使用爬虫爬取获取一定数量的博客搜索结果（包括标题和摘要部分）。

2）Blog 文本数据预处理

使用分词系统 ICTCLA 对博客搜索结果进行分句、分词以及去除停用词的预处理。

3）情感词及概念抽取

从 HowNet 情感词典中提取经过预处理的博客搜索结果中的情感词，并提取情感词的 DEF 项，作为其概念。

4）文本表示

使用 VSM 表示模型，将情感词概念作为特征项表示博客搜索结果。同时，将情感词作为特征项表示博客搜索结果，作为实验结果的对比组。

5）情感极性聚类分析

使用 k-means 算法对基于词语的特征项表示的 VSM 模型和基于概念的特征项表示 VSM 模型进行博客搜索结果的聚类分析。

6）聚类结果的显示和评价

聚类分析的结果分布用饼图表示，并用 Ground Truth 评价方法来评价聚类结果。Ground Truth 方法选用常用的精度（Precision）、熵（Entropy）和边缘索引（Rand Index）三个度量参数来评价聚类。Precision 的值越大表示聚类的准确度越高，聚类性能也越好；Entropy 值越小表示聚类越稳定，性能越好；Rand Index 值越大表明聚类结果与已标记分类好的结果项越相似，聚类的效果越好[10]。

4.2 实验结果分析

本实验选取 300 条搜索结果，其中 100 条正向，100 条中立和 100 条负向。分别使用基于词语的向量空间模型和基于概念的向量空间模型对这 300 条搜索项进行聚类分析。因采用 k-means 算法初始簇中心的随机性，每次聚类的结果不一样，所以分别实验 10 次，取平均值来评价聚类结果。

其中使用基于概念的向量空间模型进行聚类分析的聚类结果分布如图 3 所示。图 3 是聚类结果的显示，其中簇 1 占整个搜索项的 67%，簇 2 占整个搜索项的 24%，簇 3 占整个搜索项的 9%。

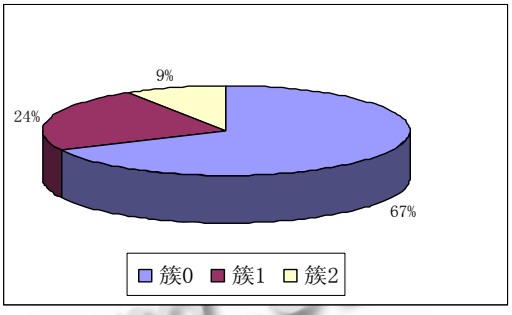

图 3 基于概念的向量空间模型聚类结果显示

如表 1 所示，在 300 条搜索结果项中，提取的情感词有 635 个，提取了 623 个情感词概念。采用情感词概念为特征项的项数少于情感词作为特征项，向量空间模型的维度减少。

表 1 基于 HowNet 词典特征项项数

情感词	情感词概念
635	623

表 2 情感极性聚类的两种模型性能比较

VSM	Precision	Entropy	Rand Index	Run Time（s）
based on Term	0.3710	1.5728	0.4964	25.448
based on Concept	0.3869	1.5583	0.5232	23.535

我们分别采用使用词语和概念作为特征项的向量空间模型分别进行聚类分析。从表 2 中可以看出，Ground Truth 评价模型中的三个参数，其中基于概念的向量空间模型中 Precision 和 Rand Index 值均高于基于词语的向量空间模型，Entropy 的值低于基于词语的向量空间模型。由表 1 可知，以概念为特征项少于以词语作为特征项，故表 2 中显示基于概念的向量空间模型聚类时间也减少。表 2 中的四个参数说明博客情感极性聚类，基于概念的向量空间模型性能较好与基于词语的向量空间模型。

5 总结与展望

本文通过建立一种基于概念的向量空间模型，使用 k-means 的聚类算法对博客情感极性进行分析，最后使用 Ground Truth 评价模型分别对以概念和词语作为特征项的向量空间模型的聚类结果进行评估。实验

(下转第 121 页)

2 Luisa MR, Elena V, Juan PC, María ÁP, et al. A proposal of user interface for a distributed asynchronous remote evaluation system: An evolution of the QUESTOURnament tool. Proc. 9th IEEE Int'l Conf. on Advanced Learning Technologies. Riga, 2009.75–77.

3 康海燕,樊孝忠,汤世平.基于 J2EE 的在线测评系统的研究与设计.计算机工程,2004,13:169–171.

4 何静雯.ACM/ICPC 评测系统综述.计算技术与自动化,2005,4:405–409.

5 王辉,胡新华,张广泉.集群式程序设计竞赛评测系统设计与开发.计算机应用与软件,2009,9:119–122.

6 李哲.在线程序竞赛评判系统的设计与实现[硕士学位论文].大连:大连理工大学,2008.

7 周伟明.多核计算与程序设计.武汉:华中科技大学出版社,2009.

8 董昊.Linux 在多核处理器上的负载均衡原理.淘宝核心系统团队博客. [2010-11-11]. http://rdc.taobao.com/blog/cs/?p=379.

9 Wikipedia. Processor Affinity. [2010-11-11]. http://en. Wikipedia.org/wiki/Processor_affinity.

10 Foong A, Fung J, Newell D. An in-depth analysis of the impact of processor affinity on network performance. IEEE Transactions on Networks, 2004,1:244–250.

11 Hollinger D.Time Measurement. [2010-11-11]. http://www.cs.rpi.edu/~hollingd/comporg/notes/timing/timing.pdf.

12 Ahmed SA, Muhammad AR, Shusmita AS, et al. Secured programming contest system with online and real-time judgment capability. Proc. of the 8th Int'l Conf. on Computer and Information Technology, 2005.

13 张胜,洪明.基于 Pocket PC 的 IDE 设计与实现.计算机系统应用,2008,17(11):14–19.

14 Reinders J. Intel Threading Building Blocks: Outfitting C++ for Multi-core Processor Parallelism. O'Reily, 2007.

15 Grune D. The software and text similarity tester SIM. [2010-11-11]. http://www.few.vu.nl/~dick/sim.html.

(上接第 75 页)

证明，基于概念的向量空间模型比基于词语的向量空间模型的聚类分析性能更好。本文中还有很多值得进一步研究的地方。对于情感词概念的选取，本文直接选用了第一义项，没有具体的概念排歧工作。在使用 k-means 算法进行聚类分析时，每次随机选取的簇中心不一样，使得聚类的结果不稳定。今后需要在情感词概念提取以及算法优化方面加强研究，使聚类结果性能更好。

参考文献

1 杨勇涛.Web 舆情观点挖掘关键技术研究.成都:电子科技大学,2009.

2 董振东,董强.HowNet's HomePage.Http://www.keenage.com, 2010.

3 夏天,樊孝忠,刘林.利用 JNI 实现 ICTCLAS 系统的 Java 调用.计算机应用,2004,24(12):177.

4 胡静,蒋外文,朱华.Web 文本挖掘中数据预处理技术研究.现代计算机(专业版),2009,(3):48-50.

5 陈龙,范瑞霞,高琪.基于概念的文本表示模型.计算机工程与应用,2008,44(20):162-164.

6 刘金岭.基于语义的高质量中文短信文本聚类算法.计算机工程,2009,35(10):201-202.

7 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.计算机应用,2007,27(12):3010.

8 Liu B. Web Data Mining. Chicago: Springer Press, 2006. 120-121.

9 游春晖.基于语义情感倾向的文本相似度计算.西安:电子科技大学,2008.

10 Pang J, Xu DP, Feng S, et al. A Novel Clustering Approach Based on Graph Similarity for Chinese Blogs on Authors' Sentiment. The 7th International Conference on Fuzzy Systems and Knowledge Discovery. Yantai: Yantai University Press, 2010. 2344-2348.