(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 105117428 A
(43)申请公布日 2015.12.02

(21)申请号 201510471154.3

(22)申请日 2015.08.04

(71)申请人 电子科技大学
地址 611731 四川省成都市高新区(西区)西源大道2006号

(72)发明人 程红蓉 唐明霜 蔡腾远 郭彦伟 张锋

(51)Int.Cl.
*G06F 17/30*(2006.01)
*G06F 17/27*(2006.01)

权利要求书2页 说明书6页 附图2页

(54)发明名称

一种基于词语对齐模型的web评论情感分析方法

(57)摘要

本发明属于自然语言处理情感分析领域,公开了一种基于词语对齐模型的web评论情感分析方法。该方法具体内容包括:从网页上获取评论信息并对这些内容进行预处理;基于改进的机器翻译模型,获取评论中的候选情感词和候选评价对象词;然后利用情感词和评价对象词语之间的情感关系和词语本身的特性指标,从候选词列表中抽取出情感词和评价对象;最后用一种有效的多分类回归模型对评价对象对应的情感词进行情感倾向判定。本发明在多个类别的评论数据集上进行了实验,得到了较好的实验结果。

CN 105117428 A

1.一种基于词语对齐模型的web评论情感分析方法，该方法包括的具体步骤如下：

步骤1、对从互联网上抓取下来的评论数据进行去掉重复的标点符号，去掉网页标签等处理，然后对其进行分词和词性标注。再把标注了的数据按逗号，句号，感叹号切分为短句子。

步骤2、修改基于单词的机器翻译模型，把这种双语的翻译模型应用到单语词语对齐模型中抽取候选的情感词和评价对象词语对。

步骤3、结合情感词和评价对象词语之间的情感关系和词语本身的特性，从候选词对中抽取出精确的情感词和评价对象。

步骤4、用一种有效的多分类回归模型对情感词进行情感极性判定。

2.根据权利要求1所述方法，其特征在于，在步骤2中是将步骤1中的切分为句子的语料库复制生成另一个平行的语料库，这两个相同的语料库作为单语对齐模型的输入语料。

3.根据权利要求1和2所述方法，其特征在于，在步骤3中进一步包括以下步骤：

3.1 根据获得的所有词对计算情感词和评价对象之间的情感关系值。

3.2 计算候选评价对象词的词语特性值 Indicator

由于候选的评价对象词语是具有领域特殊性，本发明用了m个与评论语料领域不相关但规模和原语料库相同的语料，再结合信息熵和文档频率计算候选评价对象的词语指标Indicator。

在候选的评价对象列表中如果一个词有较高的词频且在语料库中分布较均匀，那么该候选词成为评价对象的可能性比较大。在评论语料库C中，把每一篇评论当成是一个独立的类别，语料库C中有n篇评论 $C=\{d_1, d_2, \cdots, d_n\}$，如果某个评价对象词 $t_i$ 在语料库中分布越均匀，则信息熵值越大。把以上描述公式化：

$$IE(t_i) = -\sum_{j=1}^{n} p(d_j, t_i) \log p(d_j, t_i) \qquad (1)$$

以上公式中 $p(d_j, t_i)$ 表示候选词 $t_i$ 在评论 $d_j$ 中出现的概率，具体计算方法如下：

$$p(d_j, t_i) = \frac{tf_{ij}}{\sum_{j=1}^{n} tf_{ij}} \qquad (2)$$

以上公式中 $tf_{ij}$ 表示候选词 $t_i$ 在第 j 篇评论中的词频，如果 $t_i$ 只在一篇评论中出现，那么 $p(d_j, t_i)=1$，则 $\log p(d_j, t_i)=0$，使得 $IE(t_i)=0$。为了后面计算的可行性，$IE(t_i)$ 不能为0，就在公式(2)的分母部分加一个很小的常数项因子 $\varepsilon = 0.0001$. 那么此时的概率公式为：

$$p(d_j, t_i) = \frac{tf_{ij}}{\sum_{j=1}^{n} tf_{ij} + \varepsilon} \qquad (3)$$

根据信息熵会优先选择高频词，但是高频词中可能有普通的常用名词，而遗漏的低频词中也可能存在评价对象。因此，本发明用了m个与评论语料领域不相关的预料库，并结合文档频率进行计算，具体公式如下：

$$Ds(t_i)_j = \begin{cases} \alpha \times \log(1 + df_{in}) & \text{if } df_{out\_j} = 0 \\ \dfrac{\log(1 + df_{in})}{\log(1 + df_{out\_j})} & \text{otherwise} \end{cases} \quad (4)$$

以上公式中 $df_{in}$ 是候选词 $t_i$ 在评论语料库中的文档频率，$df_{out\_j}$ 表示 $t_i$ 在第 j 个与领域不相关的语料库中的文档频率。当 $df_{out\_j} = 0$ 时，候选词 $t_i$ 有很大的概率是具有领域特殊性的评价对象词。因此 α 是一个大于 1 的参数。

通过以上描述，求评价对象的 Indicator 的公式可表示为：

$$I(t_i) = \overline{Ds(t_i)} \times IE(t_i) \quad (5)$$

以上公式中 $\overline{Ds(t_i)} = \frac{\sum_{j=1}^{m} Ds(t_i)_j}{m}$

3.3 计算候选情感词的词语特性值 Indicator

大部分的情感词都不具有领域相关性，如：“好”，“讨厌”，“喜欢”等等。少部分情感词具有领域特殊性，如：餐饮评论中的“可口”。本发明结合文档频率和词语分布比例计算候选情感词的 Indicator。计算公式如下：

$$I(o_i) = \log(1+df_i) \times D_i \quad (6)$$

上式中 $df_i$ 表示候选词 $o_i$ 在评论语料库中的文档频率，$D_i = \frac{\overline{tf_i}}{\sqrt{\frac{\sum_{j=1}^{n}(tf_{ij} - \overline{tf_i})^2}{(n-1)}}}$ 代表了候选词的分布情况。$tf_{ij}$ 是候选词 $o_i$ 在第 j 篇评论中的词频，$\overline{tf_i} = \frac{1}{n} \times tf_{ij}$ 表示 $o_i$ 在所有评论中的平均词频。

3.4 获得精确的情感词和评价对象

为了模型化以上两个因子，本发明构建了一个二分图，然后用一个随机游走算法迭代计算情感词和评价对象的能量值，在候选词列表中能量值高于一定阈值的那些词被选为最终的情感词（或评价对象词）。具体公式如下：

$$E(t) = \lambda \times R \times E(o) + (1-\lambda) \times I_t$$

$$E(o) = \lambda \times R \times E(t) + (1-\lambda) \times I_o \quad (7)$$

其中 E(t) 和 E(o) 分别代表评价对象词和情感词的能量值，R 代表关系矩阵，$R_{ij}$ 表示第 i 个候选的评价对象词和第 j 个候选情感词之间的 Association 权重。$I_t$ 表示候选评价对象的 Indicator 的向量，其中的每个元素值通过公式 (5) 计算得到。$I_o$ 表示候选情感词的 Indicator 的向量，其中每个元素值通过公式 (6) 计算得到。$\lambda \in [0,1]$ 是一个调和参数。

# 一种基于词语对齐模型的web评论情感分析方法

## 技术领域

[0001] 本发明涉及互联网领域、自然语言处理和机器学习领域，具体涉及一种基于词语对齐模型的web评论情感分析实现方法。

## 背景技术

[0002] 随着web2.0的到来和移动互联网的飞速发展，互联网上的信息呈爆炸式的增长，国际电信联盟发布《2014年信息与通信技术》报告称，到2014年年底，全球互联网用户数量将达到约30亿，而且大部分的网络用户也由过去的网络信息获取者变为网络信息的制造者，使得网络内容的数量和网络信息的访问量都在迅猛的增加。情感分析就是对互联网上的信息，如新闻，博文，商品评论，邮件，论坛帖子等内容进行分析和挖掘。

[0003] 伴随着电子商务的规范和发展，在线购物的用户越来越多。线上购物，用户对产品没有一个真实的认知，所以用户会偏向于看商品的评论信息来决定是否购买。对于生产厂家或电商公司而言，要想通过了解某品牌的口碑来判断未来的销售趋势，不再局限于以往的调查问卷或电话回访来获取信息，他们直接通过在线的商品评论就可以获得想要的商品市场反馈信息。除此之外，通过对商品评论进行情感分析进而把商品推荐给用户也是一个很广泛的应用。因此，如何有效的从海量的商品评论中挖掘出深层次的情感信息成为了各行业人们的迫切需求。网络评论的情感分析(Sentiment Analysis/Opinion Mining)也就自然的成为了当前的研究热点。

[0004] 情感分析的主要工作是倾向性信息抽取和倾向性分类。倾向性信息抽取的主要任务是在词语，句子或篇章级别抽取与情感倾向相关的要素，其中对评价对象(opinion target，也称为产品特征product feature)的抽取近年来也有了更细致的工作。Hu和Liu(Hu et al.,2004)认为评价对象往往是评论者经常提及的名词或名词短语，因此用关联规则的方法抽取最小支持率(minimum support)为1%的名词或名词短语作为频繁(frequent)的评价对象。然后抽取包含评价对象的句子中的形容词为情感词(opinion word)，最后结合抽取到的频繁的评价对象和情感词来抽取非频繁(infrequent)的评价对象。Popescu和Etzioni(Popescu et al.,2005)对Hu和Liu的方法进行了改进。首先为每个产品类别定义一系列整体关系标识词(meronymy discriminator)，然后计算整体标识词和名词的点互信息(PMI)值得到该名词为评价对象的可能性。2011年Qiu等人(Qiu et al.,2011)用基于语义关系的双向传播算法(Double Propagation)抽取评价对象。2012年Liu等人(Liu et al.,2012)首次把基于单词的统计机器翻译用到了情感分析中，来联合抽取评价对象和情感词。

[0005] 对于情感倾向性分类包括词语，短句，句子，篇章等不同的粒度，都是文本分类问题。情感倾向性分类，主要有基于监督学习的方法，其中使用得最多的有朴素贝叶斯，支持向量机，最大熵模型，K近邻和条件随机场分类器等，基于半监督学习的方法和无监督学习分方法也有广泛应用于情感分类。半监督和无监督学习的方法虽然在实现上比有监督学习的方法简单，但是情感词之间的语义相似度很难计算，最终的分类结果也没有有监督学习

的分类准确度高，所以现在的情感倾向分类大多数还是使用的监督学习的方法。

## 发明内容

[0006] 基于以上背景技术，本发明提出了一种对产品评论的情感分析方法，目的在于既能对用户的购买提供参考，又能为生产厂商或电商公司提供有效的反馈信息，以便于他们能够对产品进行改进或是向用户推荐产品。

[0007] 本发明所提出方法的主体是抽取情感词和评价对象词语，与传统的单独考虑情感词和评价对象之间的情感关系 (opinion relation) 或是单独依赖词语本身的特性来抽取情感词和评价对象的方法不同，本发明结合了情感关系和词语的特性来联合抽取情感词和评价对象，能够得到更高的抽取准确率。最后用一种有监督的机器学习方法对情感词进行情感分类，进而判定评价对象所对应的情感词的情感倾向。本发明的具体实施步骤如下：

[0008] 1. 数据预处理

[0009] 本发明的数据是通过爬虫程序从网络上抓取的商品评论，这些数据存在书写格式不规范的情况，为了降低对文本分析和情感分类的影响，首先对数据进行了预处理，比如去除空行，去除空格，去除重复的标点以及网页标签等等。再用分词工具对处理后的文本进行分词和词性标注，最后将分词后的文本按标点符号（逗号，句号，感叹号）切分为句子。

[0010] 2. 抽取候选的情感词和评价对象

[0011] 在抽取候选情感词和评价对象之前，本发明基于这样的假设：所有的名词 / 名词短语都是候选的评价对象，所有的形容词 / 动词都是候选的情感词。该假设在之前的情感分析中得到了广泛的应用，并被证明是有效的。有了这种假设，本发明就把抽取情感词和评价对象看成是在文本中抽取（形容词 / 动词，名词 / 名词短语）词对的任务。本发明把基于词语的机器翻译模型 (WTM, Word-Based Translation Model) 改为单语的词语对齐模型来完成抽取词对的任务，具体改进方法是：名词 / 名词短语（形容词 / 动词）对齐到形容词 / 动词（名词 / 名词短语）或对齐到空 (NULL)，让其它词性的词语对齐到它们本身。把第 1 步中处理好的文本复制生成一个平行的语料库，把这两个相同的语料库作为模型的输入数据。对于语料库中的一个含有 n 个单词的句子 $S = \{\omega_1, \omega_2, \cdots, \omega_n\}$，要求得词语对 $A = \{(j, aj) | j \in [1, n]\}$，就要计算如下的概率公式：

[0012] $$P(A|S) \propto \prod_{k=1}^{n} n(\phi_k|\omega_k) \prod_{j=1}^{n} t(\omega_j|\omega_j) d(j|aj, n) \quad (1)$$

[0013] 其中 $t(\omega_j|\omega_{aj})$ 表示第 j 个位置的名词 / 名词短语（形容词 / 动词）和第 aj 位置的形容词 / 动词（名词 / 名词短语）同时出现在句子中的概率信息。如果一个形容词 / 动词和一个名词 / 名词短语在语料库中频繁的出现，那么 $t(\omega_j|\omega_{aj})$ 的值就会较大。$d(j|aj,n)$ 模型化了单词的位置信息，表示位置 aj 的词对准位置 j 的词的概率。$n(\phi_k|\omega_k)$ 模型化了词语对齐的繁衍概率，表示了单词一对多的情况，其中 $\phi_k$ 表示对齐到单词 $\omega_k$ 的词语的个数。通过最大化概率公式 $A^* = \arg\max_A P(A|S)$ 就可以求得词对。

[0014] 3. 获得精确的情感词和评价对象

[0015] 1) 计算情感词和评价对象之间的关系值

[0016] 通过以上方法获得所有的词语对之后，就可以计算名词 / 名词短语和形容词 / 动词之间的对齐概率，公式如下：

[0017] $$P(\omega_o|\omega_t)=\frac{count(\omega_o,\ \omega_t)}{count(\omega_t)} \quad (2)$$

[0018] 其中 $\omega_o$表示形容词 / 动词，$\omega_t$表示名词 / 名词短语，$P(\omega_o|\omega_t)$ 表示名词 / 名词短语到形容词 / 动词的对齐概率，同理可求得形容词 / 动词与名词 / 名词短语的对齐概率。通过对齐概率，就可以公式化情感词和评价对象之间的潜在的情感关系，用 Association 表示。具体公式如下：

[0019] $$Association(\omega_o,\ \omega_t)=(\lambda*P(\omega_t|\omega_o)+(1-\lambda)*P(\omega_o|\omega_t))^1 \quad (3)$$

[0020] 正如之前所描述的那样，本发明是结合潜在的情感关系和词语本身的特性指标（本发明用 Indicator 表示）来抽取情感词和评价对象。由于情感词和评价对象具有不同 Indicator，本发明用了不同的方法来计算它们。

[0021] 2) 计算评价对象的 Indicator

[0022] 本发明是把名词 / 名词短语看作候选的评价对象，这类词具有领域特殊性，同时在语料库中高频且分布均匀的候选词更有可能成为评价对象。基于这两点，本发明用了 m 个与评论语料领域不相关的语料库并结合文档频率和信息熵来计算评价对象的 Indicator。

[0023] 约定在评论语料库 $C=\{d_1,d_2,\cdots,d_n\}$ 中每一篇评论 $d_j$是相互独立的，如果某个评价对象词 $t_i$在语料库中分布越均匀，则信息熵值越大。通过熵值的大小就能很好的反映词语的分布情况，本发明中信息熵的计算公式如下：

[0024] $$IE(t_i)=-\sum_{j=1}^{n}p(d_j,\ t_i)\log p(d_j,\ t_i) \quad (4)$$

[0025] 其中 $p(d_j,t_i)$ 表示候选词 $t_i$在评论 $d_j$中出现的概率，具体计算方法如下

[0026] $$p(d_j,\ t_i)=\frac{tf_{ij}}{\sum_{j=1}^{n}tf_{ij}} \quad (5)$$

[0027] 其中 $tf_{ij}$表示候选词 $t_i$在第 j 篇评论中的词频，如果 $t_i$只在一篇评论中出现，那么 $p(d_j,t_i)=1$，则 $\log p(d_j,t_i)=0$，使得 $IE(t_i)=0$。为了后面计算的可行性，$IE(t_i)$ 不能为 0，本发明就在公式 (5) 的分母部分加一个很小的常数项因子 $\varepsilon=0.0001$。那么此时的概率公式为：

[0028] $$p(d_j,\ t_i)=\frac{tf_{ij}}{\sum_{j=1}^{n}tf_{ij}+\varepsilon} \quad (6)$$

[0025] 基于熵值，高频词会被优先选择，但是高频词中可能有普通的常用名词，如：“人”，“事情”等，而遗漏的低频词中也可能存在评价对象。为了弥补这个缺陷，本发明使用 m 个与评论语料领域不相关但规模相同的语料库，并结合文档频率进行计算，本发明用 $Ds(t_i)_j$表示候选词 $t_i$在第 j 个领域不相关的语料库中的分布信息，具体公式如下：

[0029] $$Ds(t_i)_j=\begin{cases}\alpha\times\log(1+df_{in}) & \text{if } df_{out\_j}=0\\ \frac{\log(1+df_{in})}{\log(1+df_{out\_j})} & \text{otherwise}\end{cases} \quad (7)$$

[0030] 其中 $df_{in}$是候选词 $t_i$在评论语料库中的文档频率，$df_{out_j}$表示 $t_i$在第 j 个与领域不相关的语料库中的文档频率。当 $df_{out\_j}=0$ 时，候选词 $t_i$则有很大的概率是具有领域特殊性的评价对象词。因此为了提高文档频率对分布信息 $Ds(t_i)_j$的影响，设置参数 $\alpha$ 大于 1。

[0031] 最终，求评价对象的 Indicator 的公式可表示为：

[0032] $$I(t_i) = \overline{Ds(t_i)} \times IE(t_i) \quad (8)$$

[0033] 其中 $\overline{Ds(t_i)} = \frac{\sum_{j=1}^{m} Ds(t_i)_j}{m}$

[0034] 3) 计算情感词的 Indicator

[0035] 候选的情感词是形容词 / 动词，大部分都不具有领域相关性，如：“好”，“讨厌”，“喜欢”等等。少部分词语具有领域特殊性，如：餐饮评论中的“可口”，电影评论中的“刺激”。本发明结合文档频率和词语分布比例计算候选情感词的 Indicator。具体公式如下：

[0036] $$I(o_i) = \log(1+df_i) \times D_i \quad (9)$$

[0037] 上式中 $df_i$ 表示候选词 $o_i$ 在评论语料库中的文档频率，$D_i = \frac{\overline{tf_i}}{\sqrt{\frac{\sum_{j=1}^{n}(tf_{ij}-\overline{tf_i})^2}{(n-1)}}}$ 代表了候选词的分布情况。$tf_{ij}$ 是候选词 $o_i$ 在第 j 篇评论中的词频，$\overline{tf_i} = \frac{1}{n} \times tf_{ij}$ 表示 $o_i$ 在所有评论中的平均词频。

[0038] 通过以上 1) ～ 3) 步获得了情感词和评价对象之间的情感关系值 Association 和词语本身的重要指标值 Indicator。把 Association 和 Indicator 结合起来，形成一个筛选候选词的计算指标，称为候选词的能量值 (Energy)。在候选词列表中能量值高于一定阈值的那些词被选为最终的情感词 ( 或评价对象词 )。本发明把该算法模型化为一个偶图 (bipartite graph)，并用重启动的随机游走算法 (Random Walking with Restart) 计算情感词和评价对象的能量值，公式如下：

[0039] $$\begin{aligned} E(t) &= \lambda \times R \times E(o) + (1-\lambda) \times I_t \\ E(o) &= \lambda \times R \times E(t) + (1-\lambda) \times I_o \end{aligned} \quad (10)$$

[0040] 其中 E(t) 和 E(o) 分别代表评价对象词和情感词的能量值，R 是关系矩阵，$R_{ij}$ 表示第 i 个候选的评价对象词和第 j 个候选情感词之间的 Association 权重。$I_t$ 表示候选评价对象的 Indicator 的向量，其中的每个元素值通过公式 (8) 计算得到。$I_o$ 表示候选情感词的 Indicator 的向量，其中每个元素值通过公式 (9) 计算得到。$\lambda \in [0,1]$ 是一个调和参数。

[0041] 4. 情感词情感极性分类

[0042] 根据本发明实施例的方法，最终要获得评价对象对应的情感词的情感倾向。本发明使用一种有效的多分类的回归模型 Softmax 来对情感词进行情感分类，情感词的情感倾向分为三类 (positive,neutral,negative)，在回归模型中分别用 (3,2,1) 来表示类别。对于给定的数据，本发明先用词向量模型将情感词转化为适合 Softmax 的特征向量。Softmax 回归模型的输入数据 $\{(x^{(1)}, y^{(1)}), \cdots, (x^{(n)}, y^{(n)})\}$，其中 $y^{(n)} \in \{1,2,3\}$ 表示类别，输入特征 $x^{(i)} \in R^{n+1}$，表示特征向量 x 的维度为 n+1。通过 5 折交叉验证训练模型参数。最后输出的概率最高的那个类别作为预测类别。在此用一个例子来显示情感分类的结果，如酒店评论：“酒店的房间很温馨，感觉就像回到家一样”，抽取出了情感词和评价对象词对 ( 房间，温馨 )，通过 Softmax 分类，“温馨”对应的情感类别为 3，则对“房间”的评价为“正”，表示为 ( 房间，+)。

## 附图说明

[0043] 图 1 是本发明基于词语对齐模型的 web 评论情感分析方法的整体框架图；

[0044] 图 2 是本发明基于词语对齐模型的 web 评论情感分析方法的处理流程；

[0045] 图 3 是本发明基于词语对齐模型的 web 评论情感分析方法中抽取情感词和评价对象的模型图；

[0046] 图 4 是本发明基于词语对齐模型的 web 评论情感分析方法中针对情感词情感类别判定的流程图。

## 具体实施方式

[0047] 下面参照附图，并结合具体实施例，详细描述本发明的实施例。以下参考附图描述的实施例是示例性的，只是用于解释本发明，而不能理解为对本发明的限制。

[0048] 本发明是基于词语对齐模型的 web 评论情感分析方法，主要是对互联网上的商品评论进行情感分析。如图 1 和图 2 所示，本发明包括以下步骤：

[0049] S1. 从 Web 上获取评论信息。本发明的具体实施例的数据是用爬虫程序从京东网，当当网，携程网和大众点评网站分别抓取的相机评论，书评，酒店评论以及餐饮评论。数据集的具体规模如表 1 所示。

[0050] 表 1 评论数据集

[0051]

| 领域 | 评论篇数 | 评论句子数 |
|---|---|---|
| 相机 | 17052 | 63574 |
| 书 | 9473 | 21630 |
| 酒店 | 2331 | 7365 |
| 餐饮 | 35519 | 346832 |

[0052] S2. 对数据进行预处理

[0053] 以上从 Web 上抓取的数据通常都是不规范的，先去掉网页标签，去掉重复的标点符号等。然后用中科院的分词系统 NLPIR 对文本进行分词，得到语料库 C。然后按逗号，句号，感叹号把每一篇评论切分为句子，得到语料库 C1。

[0054] S3. 从评论信息中获取候选的情感词和评价对象，具体步骤如下：

[0055] 1) 把 S2 步获得文本数据语料 C1 复制生成一个平行的语料库 C2。

[0056] 2) 修改基于单词的机器翻译模型，具体修改策略如下：让名词 / 名词短语（形容词 / 动词）对齐到形容词 / 动词（名词 / 名词短语）或者是 NULL。让其它词性的词对齐到它们本身。

[0057] 3) 把 1) 中的数据集 C1 和 C2 输入到 2) 中修改的模型，最后得到（名词 / 名词短语，形容词 / 动词 /NULL）或者（形容词 / 动词，名词 / 名词短语 /NULL）词对。

[0058] S4. 从候选词对中抽取精确的词对

[0059] 1) 通过 S3 获得了评论语料库中的所有词对，现在就可以计算词对的情感关系 Association，具体计算方法参照公式 (3)。

[0060] 2) 由于候选的评价对象词语是具有领域特殊性，本实施例用了 5 个与评论语料领域不相关且与 C1 规模相同的语料库 D1，D2，D3，D4，D5，再结合信息熵和文档频率计算候选评价对象的词语指标 Indicator，具体参照公式 (8)。

[0061] 3) 大部分情感词是不具有领域特殊性的，如“好”，“喜欢”，“丑陋”等，只有少量的情感词具有领域特殊性，如餐饮评论中的“可口”。基于这一事实，本发明实施例结合词语的文档频率和分布比例获得候选情感词的词语指标 Indicator，具体方法参照公式 (9)。

[0062] 4) 通过 1)、2)、3) 步获得了候选词的情感关系值 Association 以及它们分别的 Indicator 值，为了把这两个因子模型化，本实施例构建了一个偶图 G(V，E，R)，如附图 3，$v_t \in V$，且 $v_o \in V$，$v_t$ 表示候选评价对象词，$v_o$ 表示候选情感词。E 是顶点之间的边集，$v_t$ 和 $v_o$ 之间有情感关系时，则有边。R 表示边上的权重集合，R 中的每个元素由 1) 中计算的 Association 组成。然后用重启动的随机游走算法 (RWR) 参照公式 (10) 迭代计算得到候选词的能量值 (Energy)，候选词列表中能量值大于一定阈值的那些词被选择为最终的情感词和评价对象词。

[0063] 评价标准：本发明实施例用准确率，召回率以及 F1 值作为评价指标，经过人工验证，四个数据集上的准确率，召回率以及 F1 值统计见下表：

[0064] 表 2 实验结果

[0065]

| 数据集 | Precision | Recall | F1 值 |
|---|---|---|---|
| 书评 | 0.64 | 0.92 | 0.75 |
| 相机评论 | 0.60 | 0.76 | 0.67 |
| 酒店评论 | 0.70 | 0.87 | 0.77 |
| 餐饮评论 | 0.63 | 0.85 | 0.72 |

[0066] S5. 情感倾向

[0067] 本发明实施例是用的一种有效的多分类回归模型 Softmax 来对情感词进行情感极性判定。本发明实施例将情感词的情感极性分为正面，中性，负面三个类别，分别用数字 3，2，1 来表示。在训练 Softmax 模型之前，先人工标注情感词的情感类别，在本发明实施例中，先让三个人分别对情感词进行情感标注，这三个人标注的类别出现分歧的情况，则由三个人协商讨论，得出最终的标注结果。然后用一种词向量模型将情感词转化为 n 维的特征向量，本发明实施例用的词向量模型是 word2vector 模型。最后得出各个情感词的预测结果值。具体实验流程见附图 4。

[0068] 以上已经说明性地描述了本发明实施例，以便于本技术领域的研究人员理解本发明，但应该清楚的是，本发明不限于具体实施例的范围。对于本领域的普通技术人员而言，在不脱离本发明的原理和精神的情况下，对这些实施例进行多种修改，变化，替换和变型等，均应包含在本发明所附权利要求所保护的范围之内。

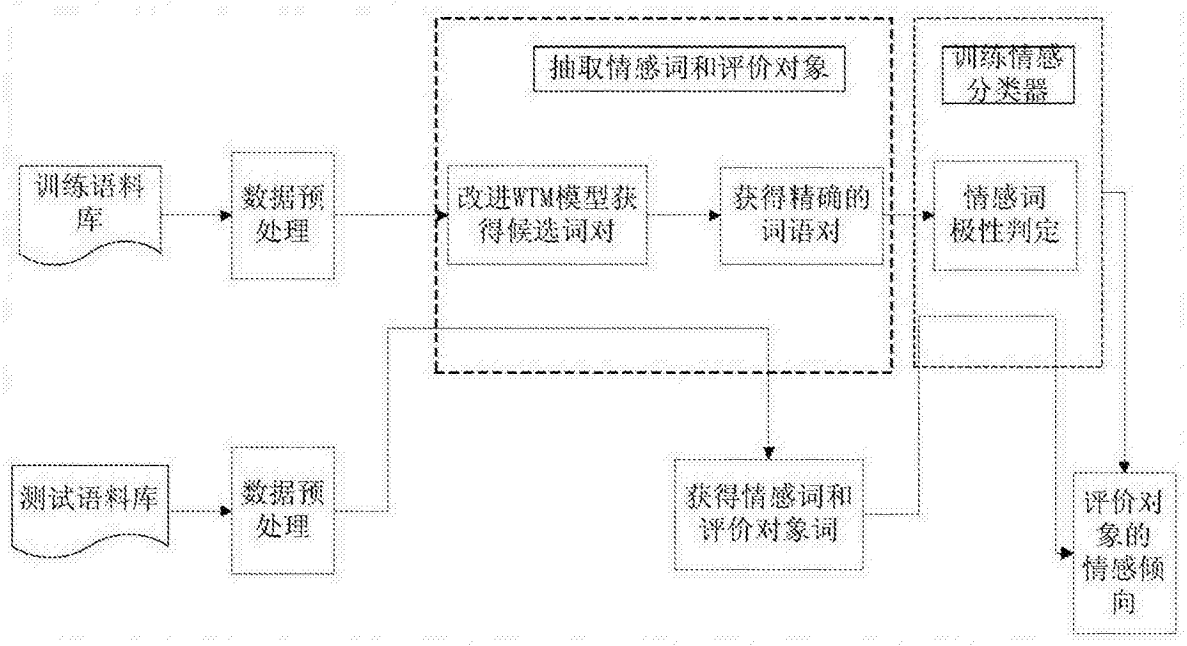
抽取情感词和评价对象
训练情感分类器
训练语料库
数据预处理
改进WTM模型获得候选词对
获得精确的词语对
情感词极性判定
测试语料库
数据预处理
获得情感词和评价对象词
评价对象的情感倾向

图 1

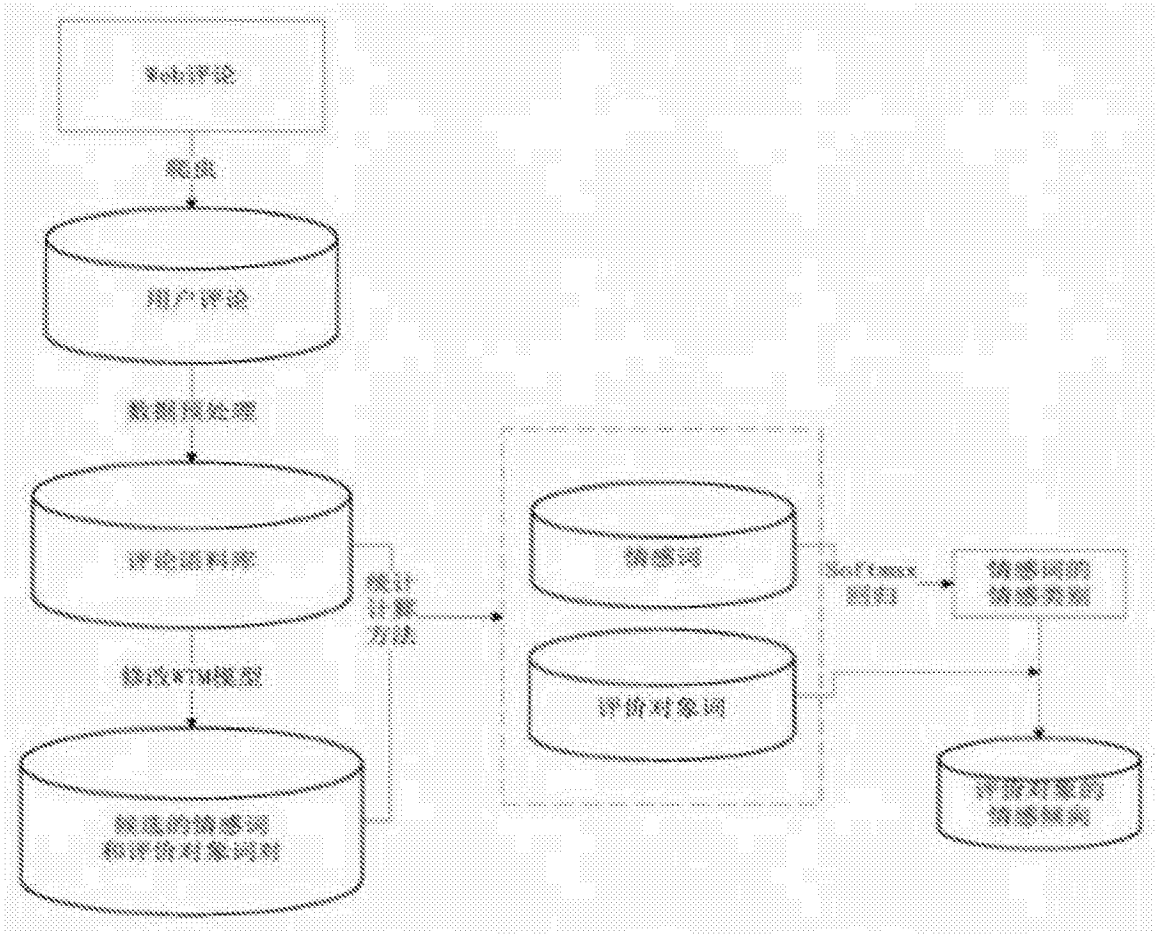

图 2

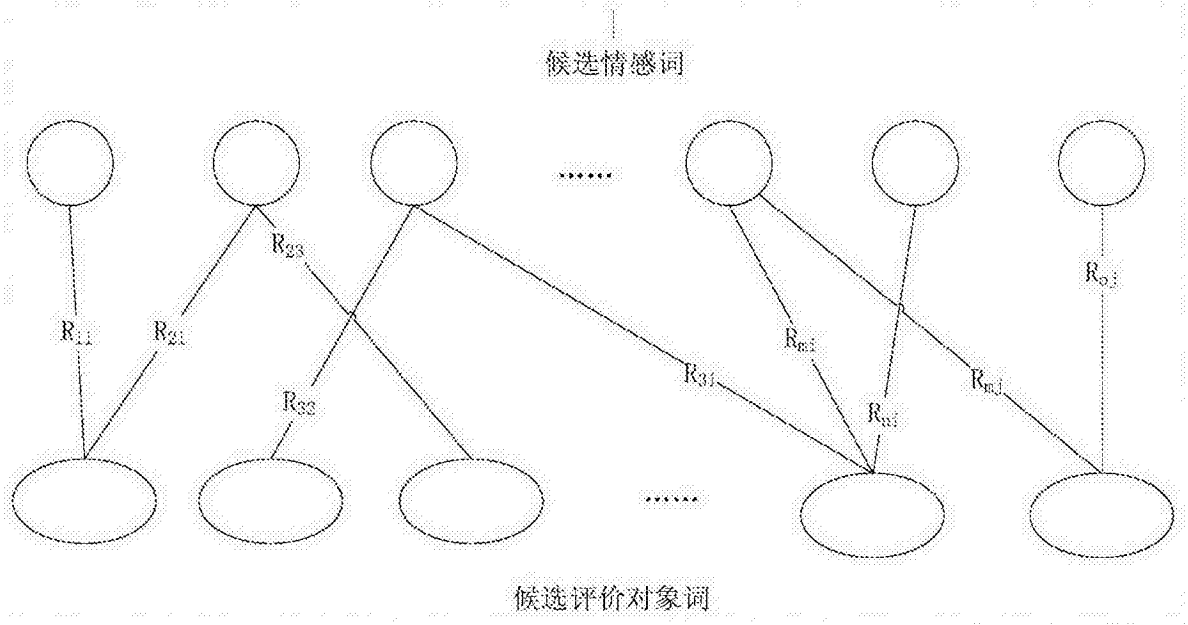
候选情感词
.....
$R_{23}$
$R_{11}$
$R_{21}$
$R_{32}$
$R_{31}$
$R_{m1}$
$R_{m2}$
$R_{m3}$
.....
候选评价对象词

图 3

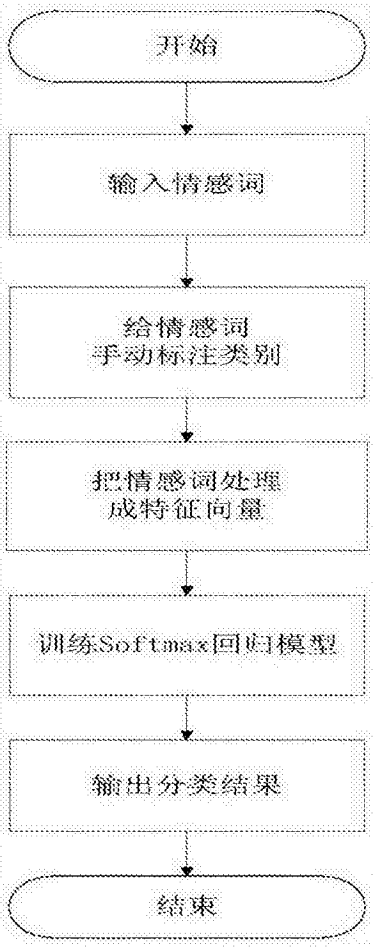
开始
输入情感词
给情感词
手动标注类别
把情感词处理
成特征向量
训练Softmax回归模型
输出分类结果
结束

图 4