(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 103617157 A
(43)申请公布日 2014.03.05

(21)申请号 201310661778.2

(22)申请日 2013.12.10

(71)申请人 东北师范大学
地址 130024 吉林省长春市人民大街 5268 号

(72)发明人 孙铁利 杨凤芹 周旭 孙红光 吴迪

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

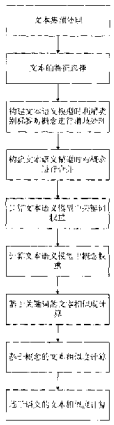
权利要求书3页 说明书8页 附图4页

(54)发明名称

基于语义的文本相似度计算方法

(57)摘要

本发明提供了一种基于语义的文本相似度计算方法，涉及面向文本的智能信息处理技术领域。其目的在于解决常规的文本向量空间模型及余弦相似度无法进行语义相关判断的问题。基于语义的文本相似度计算包括以下步骤：对文本集进行预处理，提取出初始特征词，将其表示成由关键词和概念两部分组成的向量模型；然后分别计算关键词部分的语义相似度和概念部分的语义相似度，通过对两部分进行求和最终得到文本的语义相似度。



CN 103617157 A

1. 一种基于语义的文本相似度计算方法，其特征在于采用基于关键词和基于概念的混合语义相似度计算方法计算文本间的语义相似度，具体包括如下步骤：

文本预处理：对文本集进行预处理，去停用词；

特征选择，应用卡方统计方法选择文本集的特征：针对每个类别，分别计算各个关键词和类别的卡方值，根据卡方值的大小对关键词进行降序排列，设定一个阈值 γ，过滤掉低于确定 γ 的全部关键词，从而得到每个文本的初始表示向量；

构建文本语义表示模型：文本的语义表示模型由关键词和概念两部分构成，即 $D = \{t_1, t_2, \ldots, t_m, c_1, c_2, \ldots, c_n\}$，其中 D 表示文本的特征集合，$t_1$ 表示关键词特征，c_i 表示概念特征；对于文本的初始表示向量中不在知识库中的词，保留原形作为文本语义表示模型中关键词特征；对于出现在知识库中的词，利用概念转化规则将其转化成知识库中的概念，作为文本语义表示模型中的概念特征；概念转化规则包括按照一定顺序进行以下两个操作：首先结合文本所在类别的类标签对概念进行词义消歧处理，确定概念的确切词义，然后利用概念层次转换对概念进行合并处理，以充分挖掘概念间的语义关联，接着，计算文本语义模型中关键词权重，最后，结合词语自身的特征权重以及词和类别之间相似度的信息计算概念的权重值；

基于关键词的文本相似度计算，该部分主要包括两方面处理：一是计算每对关键词间的相似度，二是在关键词相似度基础上计算文本相似度；在计算关键词间的相似度时依赖于以下假设：如果一个词语和其他词语之间存在某些联系，那么它们通常会共同出现在一些文档中；以此为基础，基于关键词的相似度计算方法首先根据语料库构建一个关键词相似度矩阵，然后通过对关键词对间的相似度加权求和取平均的方式得到文本相似度；

基于概念的文本相似度计算，该步骤主要包括两部分，一是计算每对概念间的相似度，二是在概念相似度的基础上计算文本相似度；在计算概念的相似度时，根据 Lin 提出的经典的概念相似度计算公式计算概念间的相似度，构建一个概念相似度矩阵，然后通过对概念对间的相似度加权求和取平均的方式得到文本相似度；

基于语义的文本相似度计算，最后对基于关键词的文本相似度计算结果和基于概念的文本相似度计算结果进行求和从而确定文本间的语义相似度。

2. 如权利要求 1 所述的一种基于语义的文本相似度计算方法，其中利用概念层次转换对概念进行合并处理，以充分挖掘概念间的语义关联包括：

根据知识库中概念之间的继承关系，依次找到每个概念的第 r 层上位概念，用第 r 层上位概念来表示当前概念；对于概念 c_1 和 c_2，如果 c_1 是 c_2 的子概念，c_2 是 c_1 的父概念，那么它们之间的关系可表示为 $c_1 < c_2$；进一步地，如果没有任何概念 c_3 处于 c_1 和 c_2 之间，那么 c_1 就是 c_2 的直接下位概念，c_2 是 c_1 的直接上位概念，它们之间的关系可表示为 $c_1 <_d c_2$；一个概念 c_i 的第 r 层上位概念的定义如下：

$$H(c_i, r) = \{c \mid c_i <_d c_1 <_d \ldots <_d c_r = c\} \quad (1)$$

其中，$H(c_i, r)$ 表示的是 c_i 的第 r 层上位概念，r 是概念在知识库中的层次数。

3. 如权利要求 1 所述的一种基于语义的文本相似度计算方法，其中计算文本语义模型中关键词权重包括：关键词 t 在文档 d 中的权重计算采用 tfidf 计算方法，计算公式如下所示：

$$w(t,d)=tf(t,d)\times\log(\frac{|D|}{n}) \quad (2)$$

其中，tf(t，d) 是词频，它表示词 t 在文档 d 中出现的频率，|D| 为文档总数，n 表示包含词 t 的文档数。

4. 如权利要求 1 所述的一种基于语义的文本相似度计算方法，其中计算文本语义模型中概念权重包括：概念权重计算公式为

$w(c,d_k) = tf(c,d_k)\times idf(c)\times rel(c,l_i|d_k)$ (3)

其中，$rel(c,l_i|d_k)$ 表示概念 c 和其所在文本 d_k 所属类别的类标签 l_i 之间的相似度，$w(c,d_k)$ 是概念 c 在文本 d_k 中的权重，idf(c) 是概念 c 的反文档频率，$tf(c,d_k)$ 是词频，它表示概念 c 在文档 d_k 中出现的频率，$idf(c) = \log(|D|/n)$，|D| 为文档总数，n 表示包含概念 c 的文档数；

当概念在知识库中的层次 $r > 1$ 时，其权重根据以下公式迭代计算：

$$w(c_r,d_k)=\sum_{c_{r-1}} w(c_{r-1},d_k) \quad (4)$$

其中，$c_{r-1} <_d c_r$。

5. 如权利要求 1 所述的一种基于语义的文本相似度计算方法，其中计算每对关键词间的相似度包括：设 $T = \{t_1, t_2, \ldots, t_m\}$，表示未出现在知识库中的关键词构成的集合，基于关键词的相似度计算方法根据语料库构建一个基于统计的相似度矩阵 $A = (a_{ij})_{m\times m}$，该矩阵的每一个元素 a_{ij} 是每一对属于 T 中关键词 t_i 和 t_j 之间的相似度值，其计算公式如下所示：

$$a_{ij}=sim(t_i,t_j)=\frac{\vec{t_i}\cdot\vec{t_j}}{|\vec{t_i}|\cdot|\vec{t_j}|}=\frac{\sum_{\forall d_k} w_{ki}\cdot w_{kj}}{\sqrt{\sum_{\forall d_k}(w_{ki})^2}\cdot\sqrt{\sum_{\forall d_k}(w_{kj})^2}} \quad (5)$$

其中，w_{ki} 表示关键词 t_i 在文本 d_k 中的权重，w_{kj} 表示关键词 t_j 在文本 d_k 中的权重。

6. 如权利要求 1 所述的一种基于语义的文本相似度计算方法，其中在关键词相似度的基础上计算文本相似度包括：假设两个文本 d_1 和 d_2 的表示模型中分别包括 l 和 k 个不在知识库中出现的关键词，则基于关键词的方法定义两个文本间的相似度如以下所示：

$$sim_{vs}(d_1,d_2)=\frac{(\sum_{i=1}^{l}\sum_{j=1}^{k} w_{1i}\times w_{2j}\times a_{ij})}{lk} \quad (6)$$

其中，$sim_{vs}(d_1, d_2)$ 表示两个文本 d_1 和 d_2 的相似度。

7. 如权利要求 1 所述的一种基于语义的文本相似度计算方法，其中计算每对概念间的相似度包括：在计算概念间的相似度时，根据 Lin 提出的经典的相似度计算公式计算概念间的相似度，其计算公式如下所示：

$$sim_{lin}(s_1,s_2)=\frac{2\log(p(LCA(s_1,s_2)))}{\log(p(s_1))+\log(p(s_2))} \quad (7)$$

其中，$LCA(s_1, s_2)$ 是指词义 s_1 和 s_2 的最低共同祖先，s_1 和 s_2 分别是概念 c_1 和 c_2 经过词义消歧之后对应的语义，该相似度的取值范围在 0 和 1 之间；p(s) 为当前词 s 在知识库中出现的概率，即当前词的下位概念（包括其本身）与知识库中所有概念个数的比值。

8. 如权利要求 1 所述的一种基于语义的文本相似度计算方法，其中在概念相似度的基础上计算文本相似度包括：设 $C = \{c_1, c_2, \ldots, c_n\}$ 是文本表示模型中的概念集合，构建概

念相似度矩阵 $P = (p_{ij})_{n\times n}$，该矩阵的每一个元素 p_{ij} 是概念 c_i 和 c_j 之间的相似度，计算 p_{ij} 的公式如下：

$$p_{ij} = sim(c_i, c_j) = sim_{lin}(s_i, s_j) \quad (8)$$

假设两个文本 d_1 和 d_2 的表示中分别包括 m 和 n 个概念，则基于概念的相似度计算方法将 d_1 和 d_2 之间的相似度定义为如下的形式：

$$sim_{wn}(d_1, d_2) = \frac{(\sum_{i=1}^{m}\sum_{j=1}^{n} w(c_i, d_1)\times w(c_j, d_2)\times sim(c_{1i}, c_{2j}))}{mn} \quad (9)$$

其中，如果 c_i 或 c_j 是知识库中最底层概念则按照公式 (3) 计算 $w(c_i, d_1)$ 或 $w(c_j, d_2)$，否则按照公式 (4) 计算 $w(c_i, d_1)$ 或 $w(c_j, d_2)$。

9. 如权利要求 1 所述的一种基于语义的文本相似度计算方法，其中对基于关键词的文本相似度计算结果和基于概念的文本相似度计算结果进行求和从而确定文本间的语义相似度包括：计算公式如下

$$sim(d_1, d_2) = sim_{vs}(d_1, d_2) + sim_{wn}(d_1, d_2) \quad (10)$$

其中，$sim(d_1, d_2)$ 表示文本间的语义相似度。

基于语义的文本相似度计算方法

• 技术领域

[0001] 本发明涉及面向文本的智能信息处理技术领域，尤其涉及基于关键词的文本语义相似度计算方法和基于概念的文本语义相似度计算方法。

• 背景技术

[0002] 随着互联网的飞速发展，信息技术的发展也突飞猛进，各类信息资源的数量以惊人的速度增长，如何通过精确地计算文本间的相似度快速而又准确地检索出信息是当前亟待解决的问题。

[0003] 文本相似度的计算方法在计算机技术的各个领域获得应用，例如在文本检索领域 (Text Retrieval)，文本相似度可以改善搜索引擎的召回率 (Recall) 和准确度 (Precision)；在文本挖掘领域 (Text Mining)，文本相似度作为一个测量方法用来发现文本数据库中潜在的知识；在基于网页的图像检索 (Image Retrieval) 领域，可以利用图像周围的描述性短文本来提高准确率。此外，文本相似度计算方法也可以应用到其他一些研究领域，包括文本概括 (Text Summarization)，文本分类 (Text Categorization) 和机器翻译 (Machine Translation) 等领域。

[0004] 常规的文本相似度计算的大致步骤为：首先，将待进行相似度计算的文本进行预处理，然后利用特征选择算法对特征进行抽取，构建一个传统的空间向量模型，再利用余弦相似度计算公式进行文本的相似度计算。

[0005] 对文本表示模型而言，现在普遍使用的还是 Salton 和 McGill 提出的向量空间表示模型，它是一个由词和文档构成的矩阵，词和词之间是独立的个体，将文本转化为向量空间的点。每个样本可以看成是多维的点，如一个数据集 P 有 s 个样本点，则 $P = \{p_1, p_2, \ldots, p_s\}$，在一个 n 维的空间中，每一个样本点 p_i 可用一个 n 维的属性向量表示 $< p_{i1}, p_{i2}, \ldots, p_{in} >$，其中 $1 \leqslant i \leqslant s$；其中 P_{im} 表示的是第 m 个属性（特征项）在第 i 个样本中的权重。

[0006] 向量空间模型的最大优点是实现简单。它把文本这种非结构化形式进行了数值化的表示，把文本看成多维空间中的一个点，对文本的计算可以通过向量计算得出，降低了复杂度。常用的文本处理方法在结构化文本时通常采用词袋表示模型，该模型有以下不足：(1) 未考虑两个词的语义关联，两个语义相近的词却被看成了两个独立的特征。(2) 在不同上下文中的同一个词的语义不能被很好地鉴别出来。

[0007] 对文本相似度计算而言，常用的文本间相似度计算方法是余弦相似度方法，该方法将文本看作空间中的一个点并将其表示为向量形式，利用向量之间的夹角大小来定量地计算文本间相似度，该方法没有考虑文本间具有相同语义的特征词，不能充分体现文本之间的语义相似性。

[0008] 为解决常规的文本相似度计算的上述问题，本发明提供了一种基于语义的文本相似度计算方法。

• 发明内容

[0009] 本发明提供一种基于语义的文本相似度计算方法，其目的在于解决常规的文本向量空间模型及余弦相似度无法进行语义相关判断的问题，能够提高文本相似度计算的精度，以满足各种智能文本信息处理的需求。

[0010] 本发明的上述目的是这样实现的，详细说明如下：

[0011] 一种基于语义的文本相似度计算方法，其特征在于采用基于关键词和基于概念的混合语义相似度计算方法计算文本间的语义相似度，具体包括如下步骤：

[0012] 文本预处理，对文本集进行预处理，去停用词；

[0013] 特征选择，应用卡方统计方法选择文本集的特征：针对每个类别，分别计算各个关键词和类别的卡方值，根据卡方值的大小对关键词进行降序排列，设定一个阈值 γ，过滤掉低于确定 γ 的全部关键词，从而得到每个文本的初始表示向量；

[0014] 构建文本语义表示模型：文本的语义表示模型由关键词和概念两部分构成，即 $D = \{t_1, t_2, \ldots, t_m, c_1, c_2, \ldots, c_n\}$，其中 D 表示文本的特征集合，$t_i$ 表示关键词特征，c_i 表示概念特征；对于文本的初始表示向量中不在知识库中的词，保留原形作为文本语义表示模型中关键词特征；对于出现在知识库中的词，利用概念转化规则将其转化成知识库中的概念，作为文本语义表示模型中的概念特征；概念转化规则包括按照一定顺序进行以下两个操作：首先结合文本所在类别的类标签对概念进行词义消歧处理，确定概念的确切词义，然后利用概念层次转换对概念进行合并处理，以充分挖掘概念间的语义关联，接着，计算文本语义模型中关键词权重，最后，结合词语自身的特征权重以及词和类别之间相似度的信息计算概念的权重值；

[0015] 基于关键词的文本相似度计算，该部分主要包括两方面处理：一是计算每对关键词间的相似度，二是在关键词相似度基础上计算文本相似度；在计算关键词间的相似度时依赖于以下假设：如果一个词语和其他词语之间存在某些联系，那么它们通常会共同出现在一些文档中；以此为基础，基于关键词的相似度计算方法首先根据语料库构建一个关键词相似度矩阵，然后通过对关键词对间的相似度加权求和取平均的方式得到文本相似度；

[0016] 基于概念的文本相似度计算，该步骤主要包括两部分，一是计算每对概念间的相似度，二是在概念相似度的基础上计算文本相似度；在计算概念的相似度时，根据 Lin 提出的经典的概念相似度计算公式计算概念间的相似度，构建一个概念相似度矩阵，然后通过对概念对间的相似度加权求和取平均的方式得到文本相似度；

[0017] 基于语义的文本相似度计算，最后对基于关键词的文本相似度计算结果和基于概念的文本相似度计算结果进行求和从而确定文本间的语义相似度。

[0018] 其中利用概念层次转换对概念进行合并处理，以充分挖掘概念间的语义关联包括：

[0019] 根据知识库中概念之间的继承关系，依次找到每个概念的第 r 层上位概念，用第 r 层上位概念来表示当前概念；对于概念 c_1 和 c_2，如果 c_1 是 c_2 的子概念，c_2 是 c_1 的父概念，那么它们之间的关系可表示为 $c_1 < c_2$；进一步地，如果没有任何概念 c_3 处于 c_1 和 c_2 之间，那么 c_1 就是 c_2 的直接下位概念，c_2 是 c_1 的直接上位概念，它们之间的关系可表示为 $c_1 <_d c_2$；一个概念 c_i 的第 r 层上位概念的定义如下：

[0020] $$H(c_i, r) = \{c \mid c_i <_d c_1 <_d \ldots <_d c_r = c\} \quad (1)$$

[0021] 其中，$H(c_i, r)$ 表示的是 c_i 的第 r 层上位概念，r 是概念在知识库中的层次数。

[0022] 其中计算文本语义模型中关键词权重包括：关键词 t 在文档 d 中的权重计算采用 tfidf 计算方法，计算公式如式 (2) 所示：

[0023] $$w(t,d)=tf(t,d)\times\log(\frac{|D|}{n}) \quad (2)$$

[0024] 其中，tf(t,d) 是词频，它表示词 t 在文档 d 中出现的频率；|D| 为文档总数，n 表示包含词 t 的文档数。

[0025] 其中计算文本语义模型中概念权重包括：概念权重计算公式为

[0026] $w(c,d_k) = tf(c,d_k)\times idf(c)\times rel(c,l_i|d_k)$ (3)

[0027] 其中，$rel(c,l_i|d_k)$ 表示概念 c 和其所在文本 d_k 所属类别的类标签 l_i 之间的相似度，$w(c,d_k)$ 是概念 c 在文本 d_k 中的权重，idf(c) 是概念 c 的反文档频率，$tf(c,d_k)$ 是词频，它表示概念 c 在文档 d_k 里出现的频率，$idf(c) = \log(|D|/n)$，|D| 为文档总数，n 表示包含概念 c 的文档数。

[0028] 当概念在知识库中的层次 $r > 1$ 时，其权重根据以下公式迭代计算：

[0029] $$w(c_r,d_k)=\sum_{c_{r-1}} w(c_{r-1},d_k) \quad (4)$$

[0030] 其中，$c_{r-1} <_d c_r$。

[0031] 其中计算每对关键词间的相似度包括：设 $T = \{t_1, t_2, \ldots, t_m\}$ 表示未出现在知识库中的关键词构成的集合，基于关键词的相似度计算方法根据语料库构建一个关键词的相似度矩阵 $A = (a_{ij})_{m\times m}$，该矩阵的每一个元素 a_{ij} 是每一对属于 T 中关键词 t_i 和 t_j 之间的相似度值，其计算公式如下所示：

[0032] $$a_{ij}=sim(t_i,t_j)=\frac{\vec{t_i}\cdot\vec{t_j}}{|\vec{t_i}|\cdot|\vec{t_j}|}=\frac{\sum_{\forall d_k} w_{ki}\cdot w_{kj}}{\sqrt{\sum_{\forall d_k}(w_{ki})^2}\cdot\sqrt{\sum_{\forall d_k}(w_{kj})^2}} \quad (5)$$

[0033] 其中，w_{ki} 表示关键词 t_i 在文本 d_k 中的权重，w_{kj} 表示关键词 t_j 在文本 d_k 中的权重。

[0034] 其中在关键词相似度的基础上计算文本相似度包括：假设两个文本 d_1 和 d_2 的表示模型中分别包括 l 和 k 个不在知识库中出现的关键词，则基于关键词的方法定义两个文本间的相似度如公式 (6) 所示：

[0035] $$sim_{vs}(d_1,d_2)=\frac{(\sum_{i=1}^{l}\sum_{j=1}^{k} w_{1i}\times w_{2j}\times a_{ij})}{lk} \quad (6)$$

[0036] 其中，$sim_{vs}(d_1,d_2)$ 表示两个文本 d_1 和 d_2 的相似度。

[0037] 其中计算每对概念间的相似度包括：在计算概念间的相似度时，根据 Lin 提出的经典的相似度计算公式计算概念间的相似度，其计算公式如下所示：

[0038] $$sim_{lin}(s_1,s_2)=\frac{2\log(p(LCA(s_1,s_2)))}{\log(p(s_1))+\log(p(s_2))} \quad (7)$$

[0039] 其中，$LCA(s_1, s_2)$ 是指词义 s_1 和 s_2 的最低共同祖先，s_1 和 s_2 分别是概念 c_1 和 c_2 经过词义消歧之后对应的语义，该相似度的取值范围在 0 和 1 之间；p(s) 为当前词 s 在知识库中出现的概率，即当前词的下位概念（包括其本身）与知识库中所有概念个数的比值。

[0040] 其中在概念相似度的基础上计算文本相似度包括：设 $C = \{c_1, c_2, \ldots, c_n\}$ 是文本

表示模型中的概念集合，构建概念相似度矩阵 $P=(p_{ij})_{n\times n}$，该矩阵的每一个元素 p_{ij} 是概念 c_i 和 c_j 之间的相似度，计算 p_{ij} 的公式如下：

[0041] $$p_{ij}=sim(c_i,c_j)=sim_{lin}(s_i,s_j) \quad (8)$$

[0042] 假设两个文本 d_1 和 d_2 的表示中分别包括 m 和 n 个概念，则基于概念的相似度计算方法将 d_1 和 d_2 之间的相似度定义如下的形式：

[0043] $$sim_{wn}(d_1,d_2)=\frac{(\sum_{i=1}^{m}\sum_{j=1}^{n}w(c_i,d_1)\times w(c_j,d_2)\times sim(c_{1i},c_{2j}))}{mn} \quad (9)$$

[0044] 其中，如果 c_i 或 c_j 是知识库中最底层概念则按照公式 (3) 计算 $w(c_i,d_1)$ 或 $w(c_j,d_2)$，否则按照公式 (4) 计算 $w(c_i,d_1)$ 或 $w(c_j,d_2)$。

[0045] 对基于关键词的文本相似度计算结果和基于概念的文本相似度计算结果进行求和从而确定文本间的语义相似度包括：计算公式如下

[0046] $$sim(d_1,d_2)=sim_{vs}(d_1,d_2)+sim_{wn}(d_1,d_2) \quad (10)$$

[0047] 其中，$sim(d_1,d_2)$ 表示文本间的语义相似度。

[0048] 本方法所提供的技术方案的积极效果是：和常用的基于向量空间模型的文本表示方法不同，本发明将文本表示成关键词 + 概念的形式。在提取概念特征时，利用类别信息对概念进行词义消歧处理，并利用层次关系对概念进行转化，以达到充分挖掘文本中概念间语义关系的目的。在计算文本间的相似度时，分别计算关键词对的相似度和概念对的相似度，从而克服了向量空间模型的维数高、稀疏问题给文本相似度计算带来的影响。

• 附图说明

[0049] 本发明将通过示例，参考下述附图以更进一步的阐述：

[0050] 图 1 是本发明实现基于语义文本相似度计算的系统流程图。

[0051] 图 2 是概念映射层数 r 不同时 Reuters 数据集的分类结果比较。

[0052] 图 3 是概念映射层数 r 不同时 20Newsgroup 数据集的分类结果比较。

[0053] 图 4 是几种算法性能在 Reuters 数据集上的分类结果综合比较。

[0054] 图 5 是几种算法性能在 20Newsgroups 数据集上的分类结果综合比较。

• 具体实施方式

[0055] 为了使本技术领域的人员更好地理解本发明方案，下面结合附图对本发明做进一步的详细说明。

[0056] 如附图 1，包括以下几个步骤：

[0057] 文本集预处理。对文本集进行预处理，去停用词，将无益于分类处理的代词、介词、连词等高频词过滤掉。然后用基于规则依赖的提取词根方法对单词做词形变换，这有助于集中文本的特征，减少储存所需的空间。

[0058] 文本的特征选择。对文本集词语进行特征选择，目的是去除一些对确定恰当的文本表示意义不大的词语。应用卡方统计方法来选择文本集的特征。针对每个类别，分别计算特征和类别的卡方值，根据卡方值的大小对关键词进行降序排列，设定一个阈值 γ，过滤掉低于确定 γ 的全部关键词，从而得到每个文本的初始表示向量。

[0059] 构建文本语义表示模型：利用类别标签对概念进行消歧处理。构建文本语义表示模型时，首先将文本的特征表示成关键词特征和概念特征两部分，即 $D = \{t_1, t_2, \ldots, t_m, c_1, c_2, \ldots, c_n\}$，其中 D 表示文本的特征集合，$t_i$ 表示关键词特征，c_i 表示概念特征。构建文本的关键词特征时，将知识库中不存在的词保留原形作为语义模型中关键词特征部分。对于出现在知识库中的词，利用概念转化规则将其转化成知识库中的概念，作为文本语义表示模型中的概念特征；概念转化规则包括按照一定顺序进行以下两个操作：首先结合文本所在类别的类标签对概念进行词义消歧处理，确定概念的确切词义，然后利用知识库中的概念层次转换对概念进行合并处理，以充分挖掘概念间的语义关联，接着，计算文本语义模型中关键词权重，最后，结合词语自身的特征权重以及词和类别之间相似度的信息计算概念的权重值。以下将具体分析上述操作过程：

[0060] 1、结合文本所在类别的类标签对概念进行词义消歧处理，确定概念的确切词义：

[0061] 一个词在不同的上下文中有不同的含义，但文本中的每个词和文本的类别之间有着密切的关系。本发明通过计算词的每个语义和类标签之间的相似度来确定该词在本类别文本中的确切语义，相似度最大的语义就是该词的当前语义。具体方法是：对词 t 和它的一系列语义 $s_t = \{s_{1t}, s_{2t}, \ldots, s_{kt}\}$，其中，k 是 t 的语义个数，计算和当前类别标签词义 s_l 相似度最大的语义，计算公式如公式 (1) 所示。

[0062] $$s(t) = \arg\max_{1 \le i \le k} \mathrm{sim}(s_{it}, s_l) \quad (1)$$

[0063] 其中，l 是类别标签名，s_l 是类别标签名的语义，最后确定 s(t) 就是词 t 在当前类别 l 中的确切语义。

[0064] 2、利用概念层次转换对概念进行合并处理，以充分挖掘概念间的语义关联：

[0065] 本发明通过知识库中的语义关系信息来丰富文本的表示，利用概念上下位关系将一些具有相同语义的概念进行合并，有效地解决同义词问题，克服单纯从字面上考虑词义而失去词间关联性的问题。根据知识库中概念之间的继承关系，依次找到每个概念的第 r 层上位概念，用第 r 层上位概念来表示当前概念。这种转化不仅可以大大降低向量的维数，而且能更准确地表达文本的语义。对于概念 c_1 和 c_2，如果 c_1 是 c_2 的子概念，c_2 是 c_1 的父概念，那么它们之间的关系可表示为 $c_1 < c_2$；进一步地，如果没有任何概念 c_3 处于 c_1 和 c_2 之间，那么 c_1 就是 c_2 的直接下位概念，c_2 是 c_1 的直接上位概念，它们之间的关系可表示为 $c_1 <_d c_2$。一个概念 c_i 的第 r 层上位概念的定义如公式 (2) 所示。

[0066] $$H(c_i, r) = \{c \mid c_i <_d c_1 <_d \ldots <_d c_r = c\} \quad (2)$$

[0067] 其中，$H(c_i, r)$ 表示的是 c_i 的第 r 层上位概念，r 是概念在知识库中的层次数。

[0068] 3、计算文本语义模型中关键词权重。关键词 t 在文档 d 中的权重计算采用 tfidf 计算方法，计算公式如式 (3) 所示。

[0069] $$w(t,d) = tf(t,d) \times \log(\frac{|D|}{n}) \quad (3)$$

[0070] 其中，tf(t,d) 称作词频 (Term Frequency)，它表示词 t 在文档 d 里出现的频率，|D| 为文档总数，n 表示包含词 t 的文档数。

[0071] 4、计算文本语义模型中概念权重。本方法结合了词语自身的特征权重以及词和类别之间的相似度信息。本发明认为作为类标记的词语具有更大的通用性，词在文本中的权重应该与该词和当前文本所属类别之间的相似度有关，如果该词和文本所属类别越相似，

则表明该词和该类关联度越高。据此提出的概念权重计算公式如式 (4)。

[0072] $w(c,d_k) = tf(c,d_k) \times idf(c) \times rel(c,l_i|d_k)$ (4)

[0073] 其中，$rel(c,l_i|d_k)$ 表示概念 c 和其所在文本 d_k 所属类别的类标签 l_i 之间的相似度，$w(c,d_k)$ 是概念 c 在文本 d_k 中的权重，idf(c) 是概念 c 的反文档频率，$tf(c,d_k)$ 是词频，它表示概念 c 在文档 d_k 里出现的频率。$idf(c) = \log(|D|/n)$，|D| 为文档总数，n 表示包含概念 c 的文档数。

[0074] 当概念在知识库中的层次 $r > 1$ 时，其权重根据公式 (5) 迭代计算。

[0075] $$w(c_r,d_k) = \sum_{c_{r-1}} w(c_{r-1},d_k) \quad (5)$$

[0076] 其中，$c_{r-1} <_d c_r$。

[0077] 根据词和类别的相似度以及词的权重，调整语义向量模型中概念的权重，在一定程度上量化地表示了文本中包含的抽象语义信息。

[0078] 基于关键词的文本相似度计算。该部分主要包括两方面，一是计算每对关键词间的相似度，二是在关键词相似度的基础上计算文本相似度。在计算关键词间的相似度时依赖于以下假设：如果一些词语之间存在某些语义联系，那么它们通常会共同出现在一些文本中。以此为基础，基于关键词的相似度计算方法首先根据语料库构建一个基于统计的关键词相似度矩阵 A，然后通过对关键词对间的相似度加权求和取平均值的方式得到文本相似度。具体如下：

[0079] 1、基于语料库的方法求解关键词间的相似度：设 $T = \{t_1, t_2, \ldots, t_m\}$ 表示未出现在知识库中的关键词构成的集合，基于关键词的相似度计算方法构建一个基于统计的相似度矩阵 $A = (a_{ij})_{m \times m}$，该矩阵的每一个元素 a_{ij} 是每一对属于 T 中关键词 t_i 和 t_j 之间的相似度值，其计算公式如式 (6) 所示。

[0080] $$a_{ij} = sim(t_i,t_j) = \frac{\vec{t_i} \cdot \vec{t_j}}{|\vec{t_i}| \cdot |\vec{t_j}|} = \frac{\sum_{\forall d_k} w_{ki} \cdot w_{kj}}{\sqrt{\sum_{\forall d_k} (w_{ki})^2} \cdot \sqrt{\sum_{\forall d_k} (w_{kj})^2}} \quad (6)$$

[0081] 其中，w_{ki} 表示关键词 t_i 在文本 d_k 中的权重，w_{kj} 表示关键词 t_j 在文本 d_k 中的权重。

[0082] 2、计算文本相似度：假设两个文本 d_1 和 d_2 的表示模型中分别包括 l 和 k 个不在知识库中出现的关键词，则基于关键词的方法定义两个文本间的相似度如公式 (7) 所示。

[0083] $$sim_{vs}(d_1,d_2) = \frac{(\sum_{i=1}^{l} \sum_{j=1}^{k} w_{1i} \times w_{2j} \times a_{ij})}{lk} \quad (7)$$

[0084] 其中，$sim_{vs}(d_1, d_2)$ 表示两个文本 d_1 和 d_2 的相似度。

[0085] 基于关键词的文本相似度计算方法有效地避免了文本的向量空间模型表示所导致的文本向量高维且稀疏、严重影响文本相似度计算问题。

[0086] 基于概念的文本相似度计算。该部分主要包括两方面，一是计算每对概念间的相似度，二是在概念相似度的基础上计算文本相似度。在计算概念间的相似度时，根据 Lin 提出的经典的相似度计算公式计算概念间的相似度，其计算公式如公式 (8) 所示。

[0087] $$sim_{lin}(s_1,s_2) = \frac{2\log(p(LCA(s_1,s_2)))}{\log(p(s_1)) + \log(p(s_2))} \quad (8)$$

[0088] 其中，LCA(s_1, s_2) 是指词义 s_1 和 s_2 的最低共同祖先，s_1 和 s_2 分别是概念 c_1 和 c_2 经过词义消歧之后对应的语义，该相似度的取值范围在 0 和 1 之间。p(s) 为当前词 s 在知识库中出现的概率，即当前词的下位概念（包括其本身）与知识库中所有概念个数的比值。

[0089] 在概念相似度计算的基础上，通过对概念对间的相似度加权求和取平均的方式得到文本的相似度。设 $C = \{c_1, c_2, \ldots, c_n\}$ 是文本表示模型中的概念集合，构建概念相似度矩阵 $P = (p_{ij})_{n\times n}$ 该矩阵的每一个元素 p_{ij} 是概念 c_i 和 c_j 之间的相似度，可表示为公式 (9) 的形式。

[0090] $$p_{ij} = sim(c_i, c_j) = sim_{lin}(s_i, s_j) \quad (9)$$

[0091] 假设两个文本 d_1 和 d_2 的表示中分别包括 m 和 n 个概念，那么基于概念的相似度计算方法将 d_1 和 d_2 之间的相似度定义为公式 (10) 的形式。

[0092] $$sim_{wn}(d_1, d_2) = \frac{(\sum_{i=1}^{m}\sum_{j=1}^{n} w(c_i, d_1)\times w(c_j, d_2)\times sim(c_{1i}, c_{2j}))}{mn} \quad (10)$$

[0093] 其中，如果 c_i 或 c_j 是知识库中最底层概念则按照公式 (4) 计算 $w(c_i, d_1)$ 或 $w(c_j, d_2)$，否则按照公式 (5) 计算 $w(c_i, d_1)$ 或 $w(c_j, d_2)$。

[0094] 基于语义的文本相似度计算。该单元根据基于关键词的文本相似度计算结果和基于概念的文本相似度计算结果，计算最终的文本语义相似度，其计算公式如式 (11) 所示。

[0095] $$sim(d_1, d_2) = sim_{vs}(d_1, d_2) + sim_{wn}(d_1, d_2) \quad (11)$$

[0096] 这种混合的相似度计算方法充分利用了文本表示中关键词的语义信息和概念的语义信息。因此，本方法能够获取更精确的文本相似度。

[0097] 为了探究本发明中基于语义的文本相似度计算方法的性能，发明者将本发明应用到文本分类问题中，对比的实验方法如下：

[0098] 基准方法：采用关键词表示模型，利用余弦方法计算文本间的相似度；

[0099] 方法 1：采用关键词表示模型，利用基于关键词的文本相似度计算方法计算文本间的相似度；

[0100] 方法 2：采用概念 + 关键词的语义表示模型，利用余弦相似度方法计算文本间的相似度；

[0101] 方法 3：采用概念 + 关键词的语义表示模型，利用本发明中的基于语义的文本相似度计算方法计算文本间的相似度。

[0102] 本实验采用 F 值的宏平均和微平均作为分类结果的评价指标，使用的实验数据来自两个标准的英文数据集，它们是 Reuters21578 和 20Newsgroup。在 Reuters 这个数据集中，本实验选用了来自 5 个类别的 1756 篇文章，其中 1272 篇作为训练集，484 篇作为测试集，这 5 个类别是 Grain、Trade、Interest、Crude 和 Earn。在数据集 20Newsgroup 中，选择了来自 9 个类别的 5066 篇文章，这 9 个类别是 talk.politics.guns、talk.politics.misc、rec.sport.baseball、Sci.space、Alt.atheism、sci.crypt、Sci.med、rec.sport.hockey 和 rec.motorcycles，其中 4160 篇用作训练文档，906 篇用作测试文档。

[0103] 图 2 和图 3 是合并概念层数 r 对文本分类结果的影响。实验结果表明，在概念映射阶段，利用上下位关系对概念进行合并时，并不是合并的层数 r 越高分类效果越好。在本发明的实验中，当层数为 1 或 2 时可得到最优值。

[0104] 图 4 是基准方法和其他 3 种方法在 Reuters 数据集的 5 个类上的实验比较结果，图 5 是基准方法和其他 3 种方法在 20Newsgroup 数据集的 9 个类上的实验比较结果。实验结果表明，本发明提高了文本相似度计算的精度，具有较优的性能。

[0105] 显然，本领域的技术人员可以对本发明进行各种改动和变型而不脱离本发明的精神和范围。这样，倘若本发明的这些修改和变型属于本发明权利要求及其等同技术的范围之内，则本发明也意图包含这些改动和变型在内。

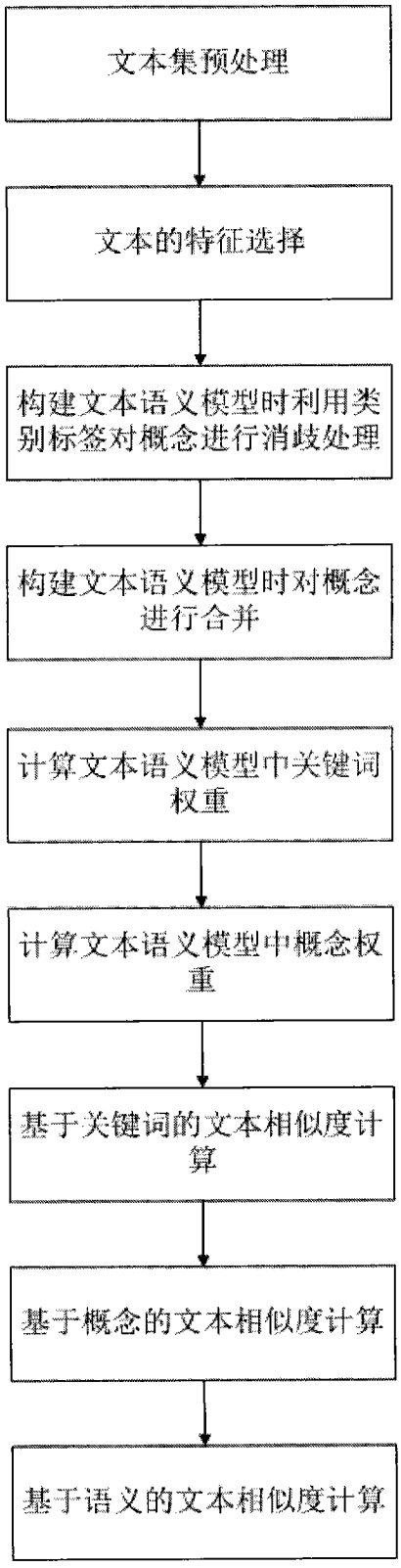

文本集预处理
文本的特征选择
构建文本语义模型时利用类别标签对概念进行消歧处理
构建文本语义模型时对概念进行合并
计算文本语义模型中关键词权重
计算文本语义模型中概念权重
基于关键词的文本相似度计算
基于概念的文本相似度计算
基于语义的文本相似度计算

图 1

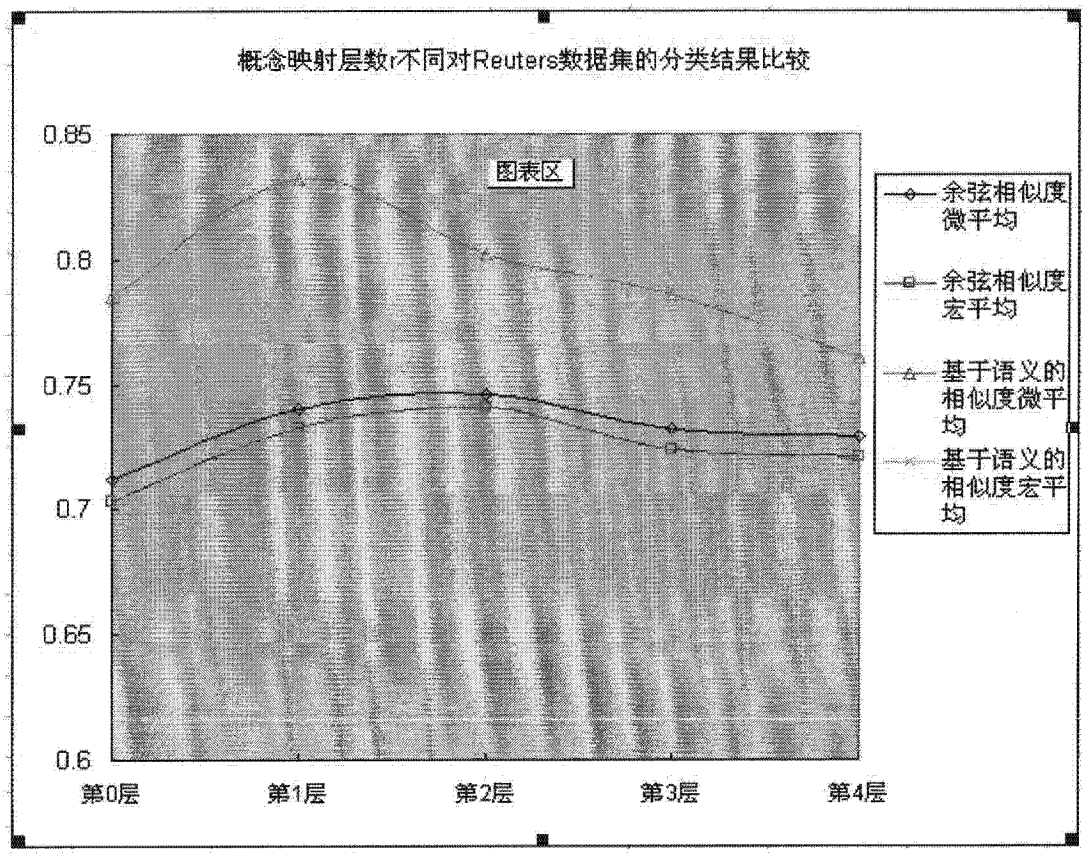

概念映射层数r不同对Reuters数据集的分类结果比较
图表区
0.85
0.8
0.75
0.7
0.65
0.6
第0层
第1层
第2层
第3层
第4层
余弦相似度微平均
余弦相似度宏平均
基于语义的相似度微平均
基于语义的相似度宏平均

图 2

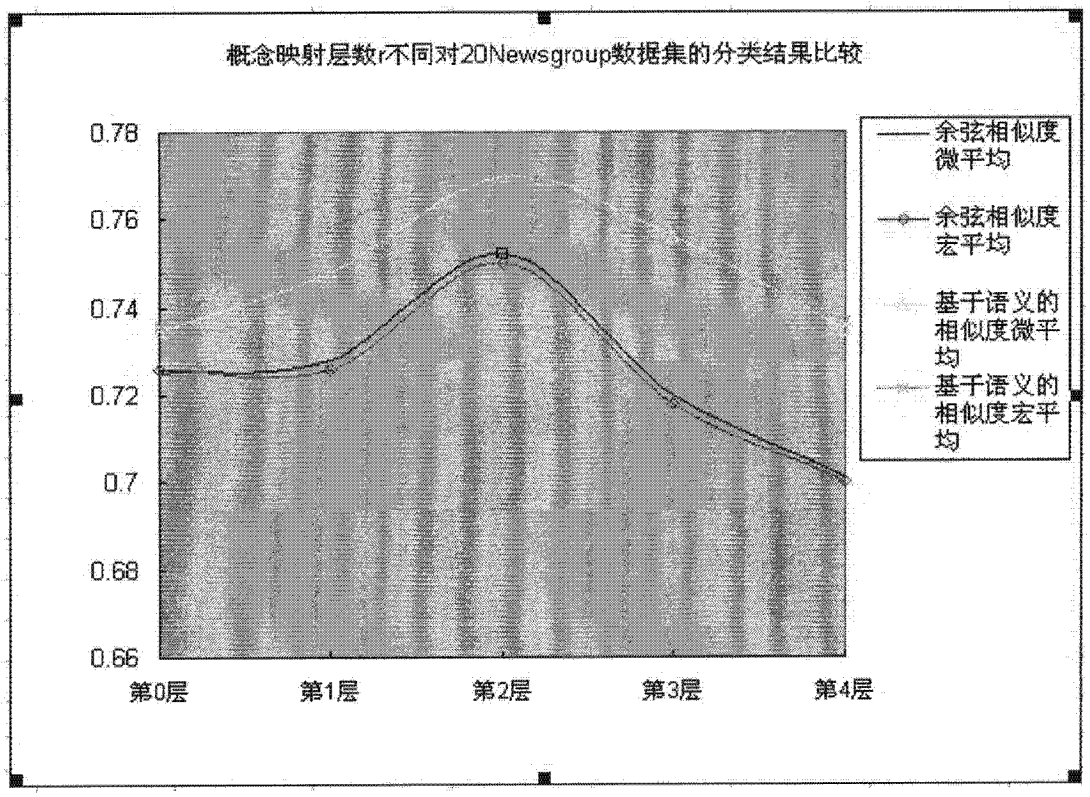

概念映射层数r不同对20Newsgroup数据集的分类结果比较
0.78
0.76
0.74
0.72
0.7
0.68
0.66
第0层
第1层
第2层
第3层
第4层
余弦相似度微平均
余弦相似度宏平均
基于语义的相似度微平均
基于语义的相似度宏平均

图 3

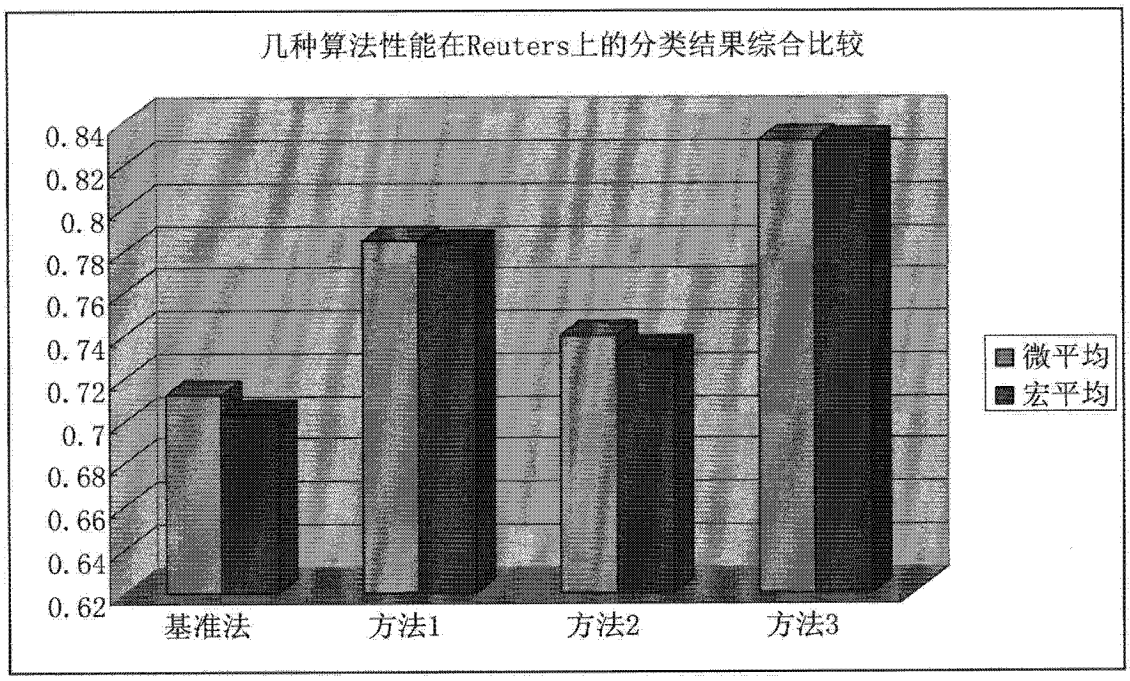

几种算法性能在Reuters上的分类结果综合比较
0.84
0.82
0.8
0.78
0.76
0.74
0.72
0.7
0.68
0.66
0.64
0.62
基准法
方法1
方法2
方法3
微平均
宏平均

图 4

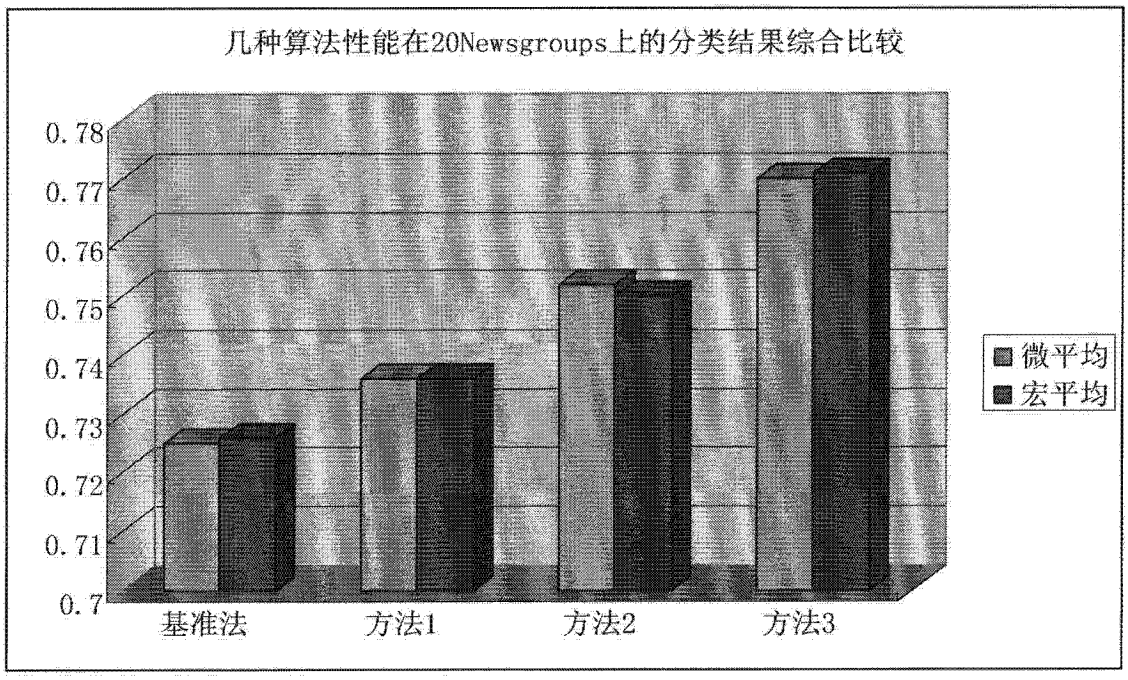

几种算法性能在20Newsgroups上的分类结果综合比较
0.78
0.77
0.76
0.75
0.74
0.73
0.72
0.71
0.7
基准法
方法1
方法2
方法3
微平均
宏平均

图 5