(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 112905771 A
(43)申请公布日 2021.06.04

(21)申请号 202110184849.9

(22)申请日 2021.02.10

(71)申请人 北京邮电大学
地址 100876 北京市海淀区西土城路10号
申请人 国网山东省电力公司电力科学研究院

(72)发明人 芦效峰 王文婷 李睿凡

(51)Int.Cl.
G06F 16/332(2019.01)
G06F 16/33(2019.01)
G06F 40/284(2020.01)

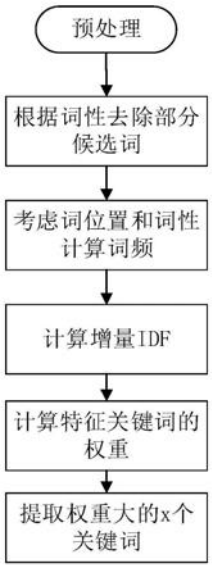
权利要求书1页 说明书3页 附图1页

(54)发明名称

基于词性和位置的特征关键词提取方法

(57)摘要

本发明公开了一种基于词性和位置的特征关键词提取方法，包括以下步骤：第一步，文本预处理，得到候选关键词；第二步，去除特定词性的候选关键词，考虑词性和词位置计算加权词频；第三步，计算文本中候选关键词的增量逆文档频率；第四步，计算文本中候选关键词的权重；第五步，按照权重从大到小对文本候选关键词进行排序，并选择权重最大的x个词作为文本的关键词。基于词在文本中出现的位置、词性的因素来优先选择关键词，提高了关键词提取的正确率。



CN 112905771 A

1.基于词性和位置的特征关键词提取方法,其特征在于提取特征关键词有下列处理步骤:

步骤一,对文本进行预处理,包括分词,去除文本中的停用词和和标点符号;如果文本是英语,进行英文大小写转换,词形还原;

步骤二,去除特定词性的词,去除不适合作为文本关键词的词性的词;考虑词位置和词性计算词频,计算每个候选词的加权词频的方法为

$$tf(t,d)=\begin{cases}(1+\alpha+\beta)*\frac{n_{t,d}}{\sum_k n_{k,d}} & if\ \ t\in T\ \ and\ \ t\ \ is\ \ a\ \ noun.\\ (1+\alpha)*\frac{n_{t,d}}{\sum_k n_{k,d}} & if\ \ t\in T\ \ and\ \ t\ \ isn't\ \ a\ \ noun.\\ (1+\beta)*\frac{n_{t,d}}{\sum_k n_{k,d}} & if\ \ t\in C\ \ and\ \ t\ \ is\ \ a\ \ noun.\\ \frac{n_{t,d}}{\sum_k n_{k,d}} & if\ \ t\in C\end{cases};$$

其中$n_{t,d}$表示词t在文档d中的次数,$\sum_k n_{k,d}$表示文档d中所有词次数的总和,T表示由标题部分的词组成的集合,C表示由正文部分的词组成的集合,α表示词出现在标题中的权重增加值,0<α<1,β是词性的权重增加值,0<β<1,α和β的值由人工设定;

步骤三,计算文本中关键候选词的增量逆文档频率,统计当前时间数据集中文档总数N_c,及当前时间段数据集中包含词t的文档数量n(t,c),计算随文本集动态更改的逆向文件频率的方法为

$$idf(t,c)=log(\frac{N_c}{n(t,c)+1});$$

N_c表示当前时间段数据集中文档总数,n(t,c)表示当前时间段数据集中包含词t的文档数量,由于每个时间段数据集中数量是变化的,因此N_c、n(t,c)是随着时间动态改变的;

步骤四,计算文本中关键词候选词的权重,计算文本中关键候选词权重的方法为

$$weight(t,d)=\frac{tf(t,d)*idf(t,D)}{\sqrt{\sum_{k\in d}(tf(k,d)*idf(k,D))^2}};$$

其中k是文档d中的所有参与计算的候选词,D是当前时段数据集;

步骤五,将步骤四计算的所有关键词候选词按照权重从大到小排序,选择权重最大的x个词作为文本的特征关键词。

2.根据权利要求1所述的基于词性和位置的特征关键词提取方法,其特征在于:根据步骤二的处理方式去除关键词候选词中一些不适合作为文本关键词的词性的词,例如副词,数字,量词,连词,定冠语,介词,比较级形容词,情态助词,人称代词,前限定词。

3.根据权利要求1所述的基于词性和位置的特征关键词提取方法,其特征在于:根据步骤二的处理方式计算出每个候选关键词的加权词频。

4.根据权利要求1所述的基于词性和位置的特征关键词提取方法,其特征在于:根据步骤四的处理方式计算文本中关键词候选词的权重。

基于词性和位置的特征关键词提取方法

技术领域

[0001] 本发明涉及一种文本特征关键词提取方法,是基于词性和位置的特征关键词提取方法。

背景技术

[0002] 在文本挖掘领域,TF-IDF是一种特征提取或者特征降维的方法。TF-IDF的主要思想是某个词在一篇文档中出现的概率越大,即词频TF越高,而在其他文档中很少出现,则说明这个词具有很高的辨识度,对关键词特征的重要性也越高。但是TF-IDF方法经常挑选中词频很高但是实际意义却很小的词作为关键词,例如“经常”、“重要”等经常在文章中出现的词,但是没有体现文章的主题思想。

[0003] 另一个问题是,TF-IDF方法中文章中每个词的逆文档频率IDF是一个常数值。但是在实时网络话题检测中,由于采集的文章或数据来自于网络,因此采集的数据集是动态变化,这样固定的IDF集合不能很好表示动态变化数据集中词的逆文档频率。此外,词是否是停用词、词在文章中出现的位置、词性对词的特征权重以及是否是关键词都有影响,但是传统TF方法计算词频不考虑上述因素对词权值的影响。已有的研究也没有将词在文章中的位置和词性两因素一起考虑,这导致很多不能反映文章主题信息的词被误认为关键词。

发明内容

[0004] 由于TF-IDF算法提取关键词特征有上述不足之处,因此我们提出了基于词性和位置的特征关键词提取方法,以下详细介绍此方法的细节:

[0005] 步骤一,对文本进行预处理,包括分词,去除文本中的停用词和和标点符号;如果文本是英语,进行英文大小写转换,词形还原;

[0006] 步骤二,去除特定词性的词,去除不适合作为文本关键词的词性的词;考虑词位置和词性计算词频,计算每个候选词的加权词频的方法为

[0007] $$tf(t,d)=\begin{cases}(1+\alpha+\beta)*\frac{n_{t,d}}{\sum_k n_{k,d}} & if\ t\in T\ and\ t\ is\ a\ noun.\\ (1+\alpha)*\frac{n_{t,d}}{\sum_k n_{k,d}} & if\ t\in T\ and\ t\ isn't\ a\ noun.\\ (1+\beta)*\frac{n_{t,d}}{\sum_k n_{k,d}} & if\ t\in C\ and\ t\ is\ a\ noun.\\ \frac{n_{t,d}}{\sum_k n_{k,d}} & if\ t\in C\end{cases};$$

[0008] 其中$n_{t,d}$表示词t在文档d中的次数,$\sum_k n_{k,d}$表示文档d中所有词次数的总和,T表示由标题部分的词组成的集合,C表示由正文部分的词组成的集合,α表示词出现在标题中的权重增加值,0<α<1,β是词性的权重增加值,0<β<1,α和β的值由人工设定;

[0009] 步骤三,计算文本中关键候选词的增量逆文档频率,统计当前时间数据集中文档总数N_c,及当前时间段数据集中包含词t的文档数量n(t,c),计算随文本集动态更改的逆向文件频率的方法为

[0010] $$idf(t,c)=log(\frac{N_c}{n(t,c)+1});$$

[0011] N_c表示当前时间段数据集中文档总数，n(t,c)表示当前时间段数据集中包含词t的文档数量，由于每个时间段数据集中数量是变化的，因此N_c、n(t,c)是随着时间动态改变的；

[0012] 步骤四，计算文本中关键词候选词的权重，计算文本中关键候选词权重的方法为

[0013] $$weight(t,d)=\frac{tf(t,d)*idf(t,D)}{\sqrt{\sum_{k\in d}(tf(k,d)*idf(k,D))^2}};$$

[0014] 其中k是文档d中的所有参与计算的候选词，D是当前时段数据集；

[0015] 步骤五，将步骤四计算的所有关键词候选词按照权重从大到小排序，选择权重最大的x个词作为文本的特征关键词。

[0016] 跟现有发明相比，本发明的上述技术方案的有益效果如下：

[0017] 1、该方法适用于动态变化的数据集，很好表示动态变化数据集中词的逆文档频率；

[0018] 2、去掉停用词，并关注词在文本中出现的位置、词性的因素来判断一个词是不是文本的关键词，可避免很多与文本主题无关的词误认为是关键词的情况，提高关键词提取的正确率。

附图说明

[0019] 图1是基于词性和位置的特征关键词提取方法的工作流程图

具体实施方式

[0020] 为使本发明要解决的技术问题、技术方案和优点更加清楚，下面将结合附图及具体实施例进行详细描述：

[0021] 如图1所示，为该方法的流程图，输入数据为实时文本集，输出为文本的关键词，即经过位置和词性加权计算的排序后关键词，具体实施过程中的步骤如下：

[0022] 步骤一，对文本进行预处理；

[0023] 步骤1-1：对文本进行分词，并去除文本中的停用词和和标点符号；

[0024] 步骤1-2：如果文本是英语，转到1-3步骤；如果是中文则转到步骤二；

[0025] 步骤1-3：进行英文大小写转换，进行词形还原；

[0026] 步骤二，去除特定词性的词，考虑词位置和词性计算词频；

[0027] 步骤2-1：去除不适合作为文本关键词的词性的词，关键词候选词中副词，数字，量词，连词，定冠语，介词，比较级形容词，情态助词，人称代词，前限定词不适用作为文本关键词，去除这些词性的词；

[0028] 步骤2-2：考虑词位置和词性计算加权词频，计算每个候选词的加权词频的方法为

[0029] $$tf(t,d)=\begin{cases}(1+\alpha+\beta)*\frac{n_{t,d}}{\sum_k n_{k,d}} & if\ t\in T\ and\ t\ is\ a\ noun.\\ (1+\alpha)*\frac{n_{t,d}}{\sum_k n_{k,d}} & if\ t\in T\ and\ t\ isn't\ a\ noun.\\ (1+\beta)*\frac{n_{t,d}}{\sum_k n_{k,d}} & if\ t\in C\ and\ t\ is\ a\ noun.\\ \frac{n_{t,d}}{\sum_k n_{k,d}} & if\ t\in C\end{cases};$$

[0030] 其中$n_{t,d}$表示词t在文档d中的次数，$\sum_k n_{k,d}$表示文档d中所有词次数的总和，T表示由标题部分的词组成的集合，C表示由正文部分的词组成的集合，α表示词出现在标题中的权重增加值，0<α<1；β是词性的权重增加值，0<β<1，α和β的值可以人工设定，一般设β<α，根据经验可以设置α＝0.2，β＝0.1；

[0031] 步骤三，计算文本中关键候选词的增量逆文档频率；

[0032] 步骤3-1：统计当前时间数据集中文档总数N_c，及当前时间段数据集中包含词t的文档数量n(t,c)；

[0033] 步骤3-2：计算随文本集动态更改的逆向文件频率的方法为

[0034] $$idf(t,c)=log(\frac{N_c}{n(t,c)+1});$$

[0035] N_c表示当前时间段数据集中文档总数，n(t,c)表示当前时间段数据集中包含词t的文档数量，由于每个时间段数据集中数量是变化的，因此N_c、n(t,c)是随着时间动态改变的；

[0036] 步骤四，计算文本中关键词候选词的权重；

[0037] 步骤4-1：计算文本中每个关键候选词的权重，计算方法为

[0038] $$weight(t,d)=\frac{tf(t,d)*idf(t,D)}{\sqrt{\sum_{k\in d}(tf(k,d)*idf(k,D))^2}};$$

[0039] 其中k是文档d中的所有参与计算的候选词，D是当前时段数据集；

[0040] 步骤五，选择权重最大的x个词作为文本的特征关键词；

[0041] 步骤5-1：将所有关键词候选词按照步骤4-1计算的权重从大到小排序；

[0042] 步骤5-2：选择权重最大的x个词作为文本的特征关键词，x根据需要人工设定。

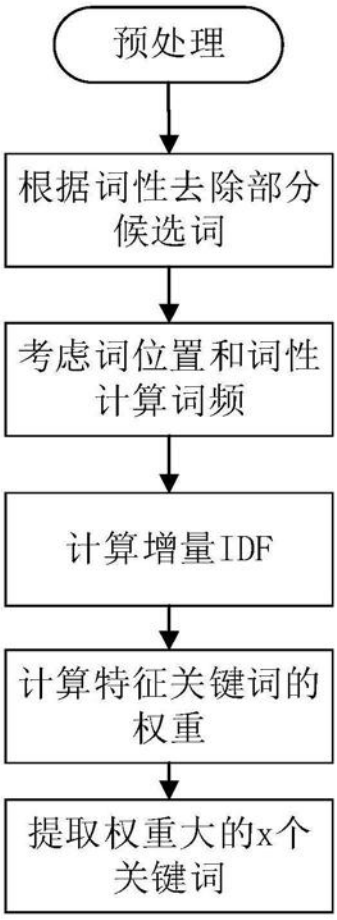

预处理
根据词性去除部分候选词
考虑词位置和词性计算词频
计算增量IDF
计算特征关键词的权重
提取权重大的x个关键词

图1