大数据背景下文档关键词抽取方法的探讨

◆孟晓燕　赵卫红

（青岛黄海学院　山东青岛　266427）

摘要：关键词抽取是借用计算机从文档中选择出能够反映主题内容的词，提供一个简短的内容摘要，便于用户获取文档信息。在当今大数据时代，在文本处理的许多领域，关键词抽取都是一项重要技术。选取关键词的目的是运用关键词最大限度反映出文档内容，研究从文档集中选取关键词的方法。本文通过引入能够与文档相关程度的指标 D_{ij}（分数），建立出 D_{ij} 的数学模型，给出文档关键词抽取方法。

关键词：关键词抽取；TextRank 算法；多文档 ；聚类

0 引言

在大数据背景下，关键词提取在文本处理的许多领域，都成为一项重要技术。大数据是全体数据，追求精确度和因果关系都变得意义不大，寻找事物之间的相关关系变得更加重要。在面对大量文档时，人们想通过阅读关键词来了解大致意思，所以如何较好提取关键词尤为重要。周锦章等[1]将文档集进行词向量表征，通过构建 TextRank 的转移概率矩阵，提出一种基于词向量与 TextRank 的关键词抽取方法。罗燕等[2] 运用词频统计规律改进传统 TE-IDF 算法，改善了关键词的提取效果。门家乐[3]提出了如何用 TextRank 做关键词提取。目前关键词提取的主流方法有基于隐含主题模型的 LDA[4]、基于 TF-IDF [5]词频统计的关键词抽取，基于词图模型 TextRank[6]的关键词抽取。

1 关键词概念

一般来说，文档的主题要通过一些特定的，能够体现主题的词语来刻画，这样的词叫作关键词。对于文档，首先是要确定一个文档的关键词。我们可能猜测文档中最频繁出现的词语应该是最重要最有资格充当关键词。但是，这个直觉实际情况恰恰相反。出现最频繁的大部分词语都是那些类似于“the”或者“and”等常见词。这些词语通常用于辅助表达，但本身不携带任何含义。实际上，英语中几百个常见词，往往在文档分类之前就被去掉。

事实上，描述主题的词语往往都是罕见。从信息论角度看，用罕见的词语当作关键词比起相对常见的词做关键词，更能引起人们的注意，能获得更大的信息量。但是，并非所有罕见的词语在做关键词时同等重要。一方面，某些在整个文档集合中极少出现的词“notwithstanding”（尽管），“ albeit”（虽然）并不能提供多少有用的信息，当然做检索词语是不合适的。另一方面，比如，“chukker” (马球戏的一局)的词虽然和上述词语一样罕见，但是该词语却能提示我们文档明显和马球运动有关。上述两类罕见的词语区别在于它们是否在部分文档中反复出现有关。也就是说，类似“albeit”的词语第一次出现并不会增加它多次出现的可能性。但是，如果一篇文章中出现“chukker”，那么随后可能会提到“first chukker”（第一回），“second chukker”（第二回）发生什么，以此类推。也就是说，如果这类词在文档中出现，那么他它们很可能反复出现。罕见词“chukker”具有两个特点：一是罕见，二是连续性。

我们一旦确立了罕见词语做关键词，那么不能做关键词的罕见词看作是“噪音”。下面，我们将给出尽可能避免噪音的一种获取最大信息量的检索词语选择方法。

2 关键词选择方法

为了特定搜索目的，按照以下步骤完成互联网上调查。

（1）文档集

选定 m 个检索词，在 Google 依着这 m 个检索词查询，获得相应的 m 类文档：$N=N_1+N_2+\cdots+N_m$，假设这些子文档集总和为 N，建立由 N 个子文档构成的文档集。

（2）词项（词组）集

为了对 N 个文档赋予关键词，对所有文档逐一地进行分词。分词是按照一定的规范重新组合成词项的过程。中文分词是文本挖掘基础。对于输入一段中文，成功的中文分词，可以达到电脑可以自动识别语句含义的效果。对所有 N 的个文档进行分词后，我们获取了“词项（词组）”集合。在这个词项（词组）集合中的每一个词项（词组）可能成为某一文档的关键词。当然，并不是在词项（词组）集合中的词都能称为关键词。一个词项（词组）能不能成为关键词，就要看这个词项（词组）能不能代表文档的信息。

（3）词（词组）出现的概率

给词项集中的词项（词组），在 N 个文档集中,第 i 个词项（词组）出现 n_i 个文档中，词项（词组）i 在整个文档集中出现的频率为

$$p_i=\frac{n_i}{N} \qquad (2)$$

（4）词项（词组）的信息量

(2）式中 P_i 表示词项 i 在 N 个文档集里出现的概率。而概率越小，信息量

$$I_i=-\log_2 P_i \qquad (3)$$

越大。也就是词项越罕见，该词项信息量越大。

哪个文档里出现词项（词组）i 出现的次数越多，越满足连续性，这个词项（词组）作为该文档关键词的可能性越大。

给文档集中的文档编号，j 表示文档集中的文档编号，用 F_{ij} 表示第 j 个文档中出现第 i 个词项（词组）出现的次数,f_{ij} 表示第 i 个词项（词组）在第 j 个文档里出现的相对次数。

$$f_{ij}=\frac{F_{ij}}{\text{在文件j中出现最多词项的频数}} \qquad (4)$$

定义　称

$$D_{ij}=f_{ij}\times I_i=-\frac{f_{ij}}{\text{在文件j中出现最多词项的频数}}\times\log_2 P_i \qquad (5)$$

为词项（词组）i 在文档 j 中的得分。

【例】假定文档集中有 $N=2^{20}=1048576$ 篇文档，并词项 1 在其中 $2^{10}=1024$ 个文档中出现，假定文档 5 中，词项 1 出现 20 次（假定这也是在这个文档中词语出现最多的次数）

$D_{15}=-f_{ij}\times\log_2 P_1=1\times10=10$

词项 1 在文档 5 中得分为 10。

（5）赋予文档关键词

对文档集（N 个文档）中的指定的文档 j，计算所有词项在该文档中的得分，得分最多的词项作为文档 j 的关键词。

基于关键词的得分，按照分数由大到小，给关键词排序，确定文档的关键词。

结束语：本文通过引入能够与文档相关程度的指标 Dij（分数），建立出 Dij 的数学模型，给出文档关键词抽取方法。词项 Dij 与词项出现的概率及词项所含信息量有关，本文给出的关键词抽取方法理论简单易懂，只是运用了简单的概率、$-\log_2 p_i$ 与信息量 Ii 呈负相关关系等数学知识，并且该方法操作简单，可行性强。本文只是给出理论方案，没有给出计算机运行程序，在推广方面仍存在不足，这点是我继续研究的方向。

参考文献：

[1]周锦章，崔晓辉.基于词向量与 TextRank 的关键词提取方法.计算机应用研究[J/OL]，2019，36（5）.[2018-03-09]

[2]罗燕，赵书良，李晓超等.基于词频统计的文本关键词提取方法[J] 计算机应用.2016，36（3）：718-725.

[3]门家乐.基于 TextRank 的关键词提取算法. 探索与观察.

作者简介：

第一作者简介：孟晓燕(1981－)， 汉，女，山东菏泽人，本科，副教授，主要研究方向高等数学、应用数学。

第二作者简介：赵卫红（1978.12－），女，籍贯：山东青岛，学历：本科，单位：青岛黄海学院，职称：副教授，职务：教师，研究方向：高等教育，英语教学与研究。