tf-idf

维基百科，自由的百科全书

tf-idf（英语：**t**erm **f**requency–**i**nverse **d**ocument **f**requency）是一种用于信息检索与文本挖掘的常用加权技术。tf-idf是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。tf-idf加权的各种形式常被搜索引擎应用，作为文件与用户查询之间相关程度的度量或评级。除了tf-idf以外，互联网上的搜索引擎还会使用基于链接分析的评级方法，以确定文件在搜索结果中出现的顺序。

目录

原理

在一份给定的文件里，**词频**（term frequency，tf）指的是某一个给定的词语在该文件中出现的频率。这个数字是对**词数**（term count）的归一化，以防止它偏向长的文件。（同一个词语在长文件里可能会比短文件有更高的词数，而不管该词语重要与否。）对于在某一特定文件里的词语t_i来说，它的重要性可表示为：

$$\mathrm{tf_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

以上式子中$n_{i,j}$是该词在文件d_j中的出现次数，而分母则是在文件d_j中所有字词的出现次数之和。

逆向文件频率（inverse document frequency，idf）是一个词语普遍重要性的度量。某一特定词语的idf，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取以10为底的对数得到：

$$\mathrm{idf_i} = \lg \frac{|D|}{|\{j : t_i \in d_j\}|}$$

其中

- $|D|$：语料库中的文件总数
- $|\{j : t_i \in d_j\}|$：包含词语t_i的文件数目（即$n_{i,j} \neq 0$的文件数目）如果词语不在数据中，就导致分母为零，因此一般情况下使用$1 + |\{j : t_i \in d_j\}|$

然后

$$\mathrm{tfidf}_{i,j} = \mathrm{tf}_{i,j} \times \mathrm{idf}_i$$

某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的tf-idf。因此，tf-idf倾向于过滤掉常见的词语，保留重要的词语。

例子

有很多不同的数学公式可以用来计算tf-idf。这边的例子以上述的数学公式来计算。词频（tf）是一词语出现的次数除以该文件的总词语数。假如一篇文件的总词语数是100个，而词语“母牛”出现了3次，那么“母牛”一词在该文件中的词频就是3/100=0.03。而计算文件频率（IDF）的方法是以文件集的文件总数，除以出现“母牛”一词的文件数。所以，如果“母牛”一词在1,000份文件出现过，而文件总数是10,000,000份的话，其逆向文件频率就是lg（10,000,000 / 1,000）=4。最后的tf-idf的分数为0.03 * 4=0.12。

在向量空间模型里的应用

tf-idf权重计算方法经常会和余弦相似性（cosine similarity）一同使用于向量空间模型中，用以判断两份文件之间的相似性。

tf-idf的理论依据及不足

tf-idf算法是创建在这样一个假设之上的：对区别文档最有意义的词语应该是那些在文档中出现频率高，而在整个文档集合的其他文档中出现频率少的词语，所以如果特征空间坐标系取tf词频作为测度，就可以体现同类文本的特点。另外考虑到单词区别不同类别的能力，tf-idf法认为一个单词出现的文本频数越小，它区别不同类别文本的能力就越大。因此引入了逆文本频度idf的概念，以tf和idf的乘积作为特征空间坐标系的取值测度，并用它完成对权值tf的调整，调整权值的目的在于突出重要单词，抑制次要单词。但是在本质上idf是一种试图抑制噪声的加权，并且单纯地认为文本频率小的单词就越重要，文本频率大的单词就越无用，显然这并不是完全正确的。idf的简单结构并不能有效地反映单词的重要程度和特征词的分布情况，使其无法很好地完成对权值调整的功能，所以tf-idf法的精度并不是很高。

此外，在tf-idf算法中并没有体现出单词的位置信息，对于Web文档而言，权重的计算方法应该体现出HTML的结构特征。特征词在不同的标记符中对文章内容的反映程度不同，其权重的计算方法也应不同。因此应该对于处于网页不同位置的特征词分别赋予不同的系数，然后乘以特征词的词频，以提高文本表示的效果。

参考资料

- Salton, G. and McGill, M. J. 1983 *Introduction to modern information retrieval*. McGraw-Hill, ISBN 0-07-054484-0.
- Salton, G., Fox, E. A. and Wu, H. 1983 Extended Boolean information retrieval. *Commun. ACM* 26, 1022–1036.
- Salton, G. and Buckley, C. 1988 Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24 (5): 513–523.

外部链接

- Term Weighting Approaches in Automatic Text Retrieval (http://portal.acm.org/citation.cfm?id=866292)
- Robust Hyperlinking (http://elib.cs.berkeley.edu/cgi-bin/pl_dochome?query_src=&format=html&collection=Wilensky_papers&id=3&show_doc=yes)：An application of tf–idf for stable document addressability.
