*Computer Science and Application* 计算机科学与应用, 2013, 3, 64-68
http://dx.doi.org/10.12677/csa.2013.31012 Published Online February 2013 (http://www.hanspub.org/journal/csa.html)

# Improved TF-IDF Keyword Extraction Algorithm*

**Xiaolin Wang, Lin Yang, Dong Wang, Lihua Zhen**

School of Computer, Anhui University of Technology, Ma'anshan
Email: wxl@ahut.edu.cn, yl_5682@126.com

Received: Nov. 22nd, 2012; revised: Dec. 16th, 2012; accepted: Dec. 25th, 2012

**Abstract:** According to the TF-IDF extract algorithm, this paper proposes a new extraction algorithm based on the words frequency statistics. Combining with sections mark technology, this algorithm assigns corresponding position weight to the words located in different position and calculates the words similarities with the same parts of speech which have a high counter in the result of the word segmentation, then merge the words with a higher similarity, finally we get the keyword sorted by the weight via the TF-IWF algorithm. This method optimized the traditional Chinese keyword extract algorithm, which take little notice of the higher similarity words, and lead to low-accuracy. The results show the new approach has better algorithm performance compared with the previous TF-IDF algorithm and the keywords set extracted can generally express the content of the article.

**Keywords:** Hownet; Word Similarity; TF-IDF; Keyword Extraction

# 改进的 TF-IDF 关键词提取方法*

**王小林，杨　林，王　东，镇丽华**

安徽工业大学计算机学院，马鞍山
Email: wxl@ahut.edu.cn, yl_5682@126.com

收稿日期：2012 年 11 月 22 日；修回日期：2012 年 12 月 16 日；录用日期：2012 年 12 月 25 日

**摘　要：**在 TF-IDF 算法基础上，提出新的基于词频统计的关键词提取方法。利用段落标注技术，对处于不同位置的词语给予不同的位置权重，对分词结果中词频较高的同词性词语进行词语相似度计算，合并相似度较高的词语，通过词语逆频率 TF-IWF 算法，按权值排序得到关键词。这种改进算法解决了传统中文关键词提取方法中对相似度高的词的不重视而导致关键词提取精度不高的问题。实验结果表明，改进的算法结果在准确率和召回率上较原有的 TF-IDF 算法上都得到较好的提升，使得提取的关键词集合能较好体现文本内容。

**关键词：**知网；词语相似度；TF-IDF；关键词提取

## 1. 引言

关键词是能够表达文档中心内容的词语，常用于计算机系统标引论文内容特征、信息检索、系统汇集以供读者检阅。由于它的出现和发展，使得计算机检索技术得到发展。关键词提取是文本挖掘领域的一个分支，是文本检索、文档比较、摘要生成、文档分类和聚类等文本挖掘研究的基础性工作。

*基金项目：国家自然科学基金资助项目(编号：61003311)；安徽省高校省级自然科学基金资助项目(编号：KJ2011A040)。

目前，关键词自动提取技术可分为三大类。1) 基于统计的方法，如 TF，TF-IDF[1,2]算法，简单快捷，TF 提取文本高频词作为候选关键词，TF-IDF 采用文本逆频率 IDF 对 TF 值加权取权值大的作为关键词，Turney 对此方法作了实验证明。但 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 算法的精度并不是很高，尤其是当文本集已经分类的情况下；2) 基于词共现图的方法，如 Keyword[3,4]

算法。该类算法，是建立在词频统计的基础上，将词语及其语义关系映射到词共现图，N 个顶点的词共现图只能包含 N-1 条边。利用该图计算每个顶点的 Key 值，Key 值的大小代表顶点的重要性，选取若干个重要顶点，即为该文档的关键词字。耿焕同[3]等人旨在找出频率不高但对中心内容贡献度大的词语，但算法需要设置过多的参数，如顶点数，边数等，因此常造成边界上关键词的取舍问题，影响算法的稳定性和精度；3) 基于词语网络的方法，如 SWN[5,6]模型算法，BC[7]算法。词语网络技术是词共现图技术的一个发展，以单词为顶点，以共现关系为边，计算网络节点的 Key 值来衡量词语在文本里的重要程度，董洛兵[5]、马力[6]等人的小世界网络模型和张敏[7]等人的社会网络模型均基于词语网络的方法对文本进行处理，但该类算法不仅需要耗费较长时间去构造词语网络，而且过程中需要设置过多的参数，不仅产生了较大的时间空间复杂度，而且对于大型文本严重影响算法的稳定性。而且考虑到影响网络中节点重要性的指标有多个，而文献[7]等只考虑了其中一种网络节点中心度，不能对其结果有较强的说服力。

综合上述量化结果，本文提出：1) 采用词语相似度研究成果，统计文本相似度高的词语，并进行合并，引入文本段落标注技术，进一步优化文本分词后的统计结果；2) 提出 TF-IDF 改进算法 TF-IWF (Term Frequency-Inverse Word Frequency)，将文本逆频率更换成词语逆频率，使其权值更能表达每个词语在语料库中的重要程度。用本文方法在 VC++6.0 环境下完成文本的关键词自动提取实验，其结果与经典统计方法实验结果进行对比，准确率和召回率显著提高。

## 2. 改进的文本预处理技术

考虑到在文本中存在很多像“计算机”和“电脑”类似的词语，而在一般的文本词频统计当中，都是分开来对待，这样在基于词频统计的分词算法中将严重影响所求结果的精确度。张颖颖[8]、索红光[9]等人分别采用建立同义词链和基于词汇链的技术来解决此类问题。本文利用《知网》计算文本分词后的词语相似度，取大于一定阈值的词语进行合并处理，以提高同义词或近义词的权重。

词语相似度是一个主观性很强烈的概念。从信息论的角度出发，Dekang Lin[10]认为任何两个事物的相似度取决于他们的共性和个性。在中文分词领域，主要是指两个词语在不同的上下文中，可以相互替换使用而不改变文本的句法语义结构的程度[11]。本文结合对《知网》的深入认识，采用文献[12]的词语相似度计算方法。

针对传统算法提取的词语不能表征关键词的文本分布情况，本文在词语预处理过程中加入词语段落标注技术，并结合一定的数据结构完成实验。在算法执行过程中，三元组$<w_i, fre_i, v_i>$表示所处理文本的结果集，其中 $w_i$ 是词语，$fre_i$ 是词语 $w_i$ 加权后出现次数，$v_i$ 是词语在文本中的位置权重。四元组$<w_i, w_j, sim_{ij}, fre_i + fre_j>$表示对三元组中部分词语计算相似度后的集合，其中 $sim_{ij}$ 表示词语 $w_i$、$w_j$ 的相似度，$fre_i + fre_j$ 表示两个词语的词频之和。

考虑到文本词语两两相互计算相似度，将产生较大的计算量，在计算过程中消耗过多的时间，从计算效率的角度讲，这是不可取的。在计算 $sim_{ij}$ 时，考虑不同词性对词语相似度的影响度很低[12]，词频过低的词语对计算结果影响也很低两方面因素，将四元组中具有相同词性，且词频大于约定阈值 $fre$(本文取值为 2)的词语 $w_i$、$w_j$ 进行相似度计算。这种设计大大减少了计算次数，在算法执行效率上大大提高。

具体算法步骤如下：

1) 对语料库文本进行分词；

2) 给分词结果的文本分别进行位置加权处理，按如下规则：

文本第一行是标题，赋予权值 5*；段首第一个词等于“摘要”，则赋予权值 3*；段首第一个词等于“关键词”或“关键词”，则赋予权值 5*；段首第一个词等于“结论”，赋予权值 3*；其它，每段首赋予权值 1*；

3) 去停用词(虚拟词，语气助词，副词，符号，一个字的词……)并统计各文本词频，得到$<w_i, fre_i, v_i>$三元组；

4) 进行词语相似度计算方法，计算三元组里词频 $fre_i > 2$(词频小于 3 的词忽略)的所有词语相似度 $sim_{ij}$。当 $sim_{ij} \geq 0.85$ 则认为两个词语相似度极高，在文本上下文中可以替换，将返回四元组$<w_i, w_j, sim_{ij}, fre_i + fre_j>$，并在原三元组里删除词语 $w_j$；

5) 在三元组$<w_i, fre_i, v_i>$中，查找四元组$<w_i, w_j,$

$sim_{ij}$, $fre_i + fre_j$>中的词语。当它们第一个元素代表的词语相同时，将三元组的 $fre_i$ 替换为四元组中的 $fre_i + fre_j$。重新组成三元组<$w_i$, $fre_{isumsim}$, $v_i$>，其中 $fre_{isumsim} = fre_i + fre_j$；最后得出三元组集合即为改进文本预处理算法后的结果。

## 3. 改进的关键词提取算法

传统的 TF-IDF (Term Frequency-Inverse Document Frequency)是一种用于资讯检索与文本挖掘的常用加权技术。它是一种基于训练集的统计方法，用以评估一个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。字词的重要性随着他在文本中出现的次数成正向增加，但同时会随着他在语料库中出现的频率成反向下降。TF-IDF 加权的各种形式经常被搜索引擎应用，作为文件与用户查询之间相关程度的度量或评级。计算公式如公式(1)所示。

$$TF-IDF_{i,j} \to TF_{i,j} \times IDF_i = \frac{n_{i,j}}{\sum_k n_{k,j}} \times \log \frac{|D|}{\left|\{j : t_i \in d_j\}\right|} \quad (1)$$

其中，*TF* 部分分子 $n_{i,j}$ 表示词语 $t_i$ 在文本 $j$ 中的频数，分母表示文本中所有词语的频数和；IDF 部分 $|D|$ 表示语料库 $d$ 的文档数，$\left|\{j : t_i \in d_j\}\right|$ 表示本语料库 $d$ 中包含文档 $j$ 中词语 $t_i$ 的文档数。

在本质上 IDF 是一种试图抑制噪音的加权，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。这对于大部分文本信息，并不是完全正确的。IDF 的简单结构并不能使提取的关键词，十分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。尤其是在同类语料库中，这一方法有很大弊端，往往一些同类文本的关键词被掩盖。例如：语料库 D 中教育类文章偏多，而文本 $j$ 是一篇属于教育类的文章，那么教育类相关的词语的 IDF 值将会偏小，使提取文本关键词的召回率更低。

本文在此基础上，提出词语逆频率方式计算加权算法 TF-IWF (Term Frequency-Inverse Word Frequency)，如公式(2)所示：

$$TF-IWF_{i,j} \to TF_{i,j} \times IWF_i = \frac{n_{i,j}}{\sum_k n_{k,j}} \times \log \frac{\sum_{i=1}^{m} nt_i}{nt_i} \quad (2)$$

其中 TF 部分分子 $n_{i,j}$ 表示词语 $t_i$ 在文本 $j$ 中出现的次数，分母表示文本 $j$ 中所有词语频词和，IWF 部分分子表示语料库中所有词语频数之和，$nt_i$ 表示词语 $t_i$ 在语料库中出现的总频数。IWF 部分的含义是对语料库词语总数与待查文本中该词出现在语料库中的次数比求对数。这种加权方法降低了语料库中同类型文本对词语权重的影响，更加精确地表达了这个词语在待查文档中的重要程度。在传统方法 TF-IDF 所求的权值一般很小接近 0，精确度也不是很高，公式(2)的计算结果刚好能解决最后权值过小的问题，本实验中将精确度保留 6 位有效数字，使得计算结果更加精确。结合文本预处理技术，本文文本关键词提取算法分为两大部分：

1) 生成语料库词语统计结果集

①在复旦大学的中文语料库 20 个类别中挑选经济，政治，体育，科技等 10 类文档共 1000 篇文本构成文本语料库；

②采用第二部分文本预处理算法，忽略词语相似度计算步骤，对 1000 篇文章分别处理，并将结果存放在二元组<$w_i$, $fre_i$>集合中，其中 $w_i$ 表示词语，$fre_i$ 表示该词语出现的频次；

③合并所有统计结果二元组，相同词语频数相加，得到二元组<$w_i$, $fre_{isum}$>，并将其写入 Train_Set.DAT，其中 $fre_{isum}$ 表示词语 $w_i$ 在语料库中出现的总频次。

假如每次对测试文本进行关键词提取时，都要对该语料库进行训练，这将耗费大量时间。所以在本文实验中，考虑到该方法在实际应用中的可行性，本文算法只需要对语料库进行一次训练，并将语料库训练结果保存在 Train_Set.DAT 中。在每次对测试文本进行关键词提取时，只需要读取 Train_Set.DAT 文件，调用其存储的二元组<$w_i$, $fre_{isum}$>即可。

2) 提取文本关键词

①采用文本预处理技术处理测试文本，得到三元组<$w_i$ , $fre_{isumsim}$, $v_i$>；

②采用 TF-IWF 加权算法对三元组<$w_i$, $fre_{isumsim}$, $v_i$>进行处理，提取文本总词语个数的 5%，作为测试文本的关键词集 *set*。

结合第一部分和第二部分，本实验具体实现过程如图 1 所示。

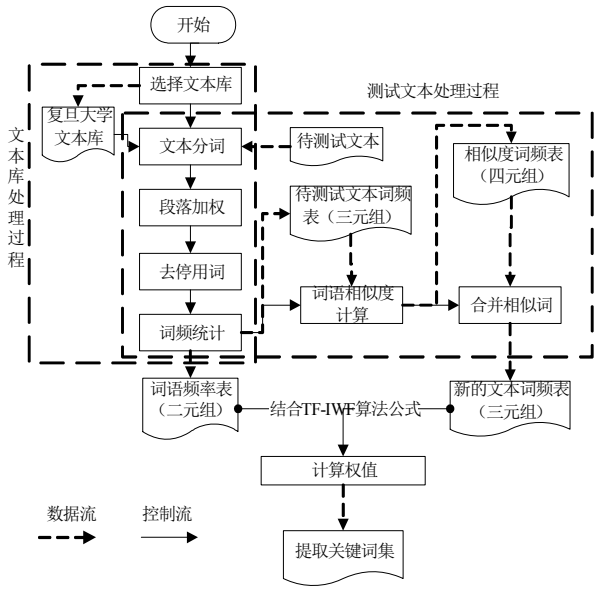


**Figure 1. Algorithm flow chart of keyword extraction**
**图 1. 关键词提取算法流程图**

## 4. 实验结果

由于一般文本都没有手动标引出主题关键词，只能用过标题了解文本大意，而提取出的关键词也只是用来便捷了解文本大意，根据其提取结果对文本理解具有一定的主观性。为验证本实验的正确性，本文从复旦大学语料库中随机选取已标引出关键词的经济，政治，体育，科技等 10 类不相关文本各 10 篇，作为测试文本集，使用 TF，TF-IDF 和本文算法进行对比试验。由于一般关键词提取算法的评估都是通过特定的评估函数与人工提取的关键词进行比较。在此，本文也利用精确率、召回率来评测函数性能，其定义如下：

$$\text{准确率} = \frac{\text{提取正确的关键词数}}{\text{提取的关键词数}} \tag{3}$$

$$\text{召回率} = \frac{\text{提取正确的关键词数}}{\text{本中的关键词数}} \tag{4}$$

为满足不同类型大小文本提取关键词个数的客观性，公式(3)中提取的关键词数设置为该文档中总词语个数的 5%，实验结果如表 1 所示。

表 1 中的所有数据都是对 10 篇文档进行计算后的均值。表 2 中的数据是针对表 1 数据计算的三种算法的平均精确率和平均召回率。由上述数据可知，本文算法提取的关键词精确率、召回率明显优于传统算法，进一步验证了算法的可取性。另外，在其中一篇

**Table 1. Results comparison of keyword extraction between the algorithm proposed in this paper and the traditional one (TF, TF-IDF)**
**表 1. 本文算法和传统算法(TF, TF-IDF)关键词提取结果对比**

| 方法指标 / 实验数据 | 精确率 | | | 召回率 | | |
|---|---|---|---|---|---|---|
| | TF | TF-IDF | 本文方法 | TF | TF-IDF | 本文方法 |
| Economy (10) | 0.390 | 0.376 | 0.413 | 0.610 | 0.665 | 0.703 |
| Politics (10) | 0.332 | 0.328 | 0.347 | 0.643 | 0.641 | 0.670 |
| Sports (10) | 0.304 | 0.312 | 0.331 | 0.612 | 0.632 | 0.655 |
| Law (10) | 0.330 | 0.321 | 0.371 | 0.569 | 0.590 | 0.618 |
| Education (10) | 0.346 | 0.348 | 0.351 | 0.611 | 0.649 | 0.691 |
| History (10) | 0.402 | 0.410 | 0.433 | 0.671 | 0.680 | 0.711 |
| Transport (10) | 0.309 | 0.311 | 0.379 | 0.588 | 0.576 | 0.619 |
| Environment (10) | 0.360 | 0.356 | 0.379 | 0.581 | 0.602 | 0.692 |
| Energy (10) | 0.368 | 0.381 | 0.399 | 0.610 | 0.656 | 0.728 |
| Space (10) | 0.345 | 0.361 | 0.481 | 0.673 | 0.645 | 0.771 |

**Table 2. The average accurate rate and the recall rate among the three approaches**
**表 2. 各方法平均精确率和平均召回率**

| | TF | TF-IDF | 本文方法 |
|---|---|---|---|
| 平均精确率 | 0.349 | 0.350 | 0.388 |
| 平均召回率 | 0.617 | 0.634 | 0.686 |

经济类文章《论中国经济的繁荣与萧条》[13]中，本文算法不仅提取出较高词频：经济(123)，还能提取出词频很低的主题词繁荣(12)，萧条(9)以及能表达文本内容的科教兴国(4)，变革(16)等。而在 TF 中只能提取出词语频率较高的经济(123)，投资(65)，政策(52)，在 TF-IDF 算法中还能提取出变革(16)，保障(2)，但是这些低频词并不能准确表达文章大意。

## 5. 结论

本文利用《知网》计算词语相似度，合并相似度较高的词语，并且词语频次叠加，反映出词语频率较高词语在文本中的重要性。新的 TF-IWF 算法中，将 IDF 转换为 IWF，将词频比作为文本候选关键词去噪音的权值，有效的抑制了与测试文本同类语料库对所提取关键词权重的影响，修正了 TF-IDF 算法的偏差。在算法实现过程中，根据文章结构和词语位置，给予

不同位置的词语不同的位置权重。经过实验表明，本文算法相较其他算法，效果更优，得出的关键词能基本反映文本内容。

## 参考文献 (References)


[1] P. D. Turney. Learning algorithms for keyphrase extraction. Information Retrieval, 2000, 2(4): 303-336.

[2] I. H. Witten, G. W. Paynteer, E. Frank, et al. KEA: Practical automatic keyphrase extraction. The 4th ACM Conference on Digital Libraries, Berkeley: ACM Press, 1999: 254-256.

[3] 耿焕同, 蔡庆生, 于锟等. 一种基于词共现图的文档主题词自动抽取的方法[J]. 南京大学学报(自然科学), 2006, 42(2): 156-162.

[4] Y. Matsuo, Y. Ohsawa and M. Ishizuka. KeyWorld: Extraction keywords in a document as a small world. Discovery Science, 2001, 2226: 271-281.

[5] 董洛兵. 基于 SWN 理论的文本复合关键字提取算法的研究[D]. 西安电子科技大学, 2006.

[6] 马力, 焦李成, 白琳. 基于小世界模型的复合关键词提取方法研究[J]. 中文信息学报, 2009, 5: 121-128.

[7] 张敏, 耿焕同, 王煦法. 一种利用 BC 方法的关键词自动提取算法研究[J]. 小型微型计算机系统, 2007, 28(1): 189-192.

[8] 张颖颖, 谢强, 丁秋林. 基于同义词链的中文关键词提取算法[J]. 计算机工程, 2010, 36(19): 93-95.

[9] 索红光, 刘玉树. 一种基于词汇链的关键词抽取方法[J]. 中文信息学报, 2006, 20(6): 25-30.

[10] D. K. Lin. An information-theoretic definition of similarity. Proceedings of the 15th International Conference on Machine Learning (ICML'98), San Francisco: Morgan Kaufmann Publishers Inc., 1998: 296-230.

[11] 刘群, 李素建. 基于《知网》的词汇语义相似度计算[A]. 第三届汉语词汇语义学研讨会论文集[C]. 台北, 2002: 59-76.

[12] 王小林, 王义. 改进的基于知网的词语相似度算法[J]. 计算机应用, 2011, 11(31): 3075-3090.

[13] 陈乐一. 论中国经济的繁荣与萧条[URL], 2005. http://www.kesum.com/zjzx/mjzl/hunan/cly/200512/8970.html