目 录

张庆国

第五章　自动标引

5.1　自动标引和人工标引

情报检索的最终目的在于检出满足用户需要的文献资料，但是在检索系统中，文献却是以其主题的标识化形式——标引词来表示的。检索过程也是通过情报提问中的检索词与检索系统中的标引词的匹配而实现的。标引词（检索词）是联接文献与提问的中介，是联系情报与其用户的中介。由此可见标引工作在情报检索工作中的重要地位。

对于标引工作，存在着标引质量和标引效率两方面的要求。

标引质量，所选中的标引词能否全面、准确地反映文献的主题，直接决定着该篇文献能否正确地被用户通过该篇文献的标引词集合而了解原文献，直接决定着检索系统能否正确地将检索提问与其匹配，从而直接决定该篇文献能否为需要它的用户检出。

标引效率，由于标引工作在质量上的极高要求，由于标引工作在自身上的难度，标引工作的效率是比较低的，然而它又是检索系统中不可缺少的重要部分，是文献进入检索系统的必经环节，因此其效率问题是必须解决的。

对于现代化的、自动化的情报检索系统来说，手工方式的标引工作显然是不适宜的。长期以来，人们一直致力于自动标引的研究工作，也取得了一些成就，但就目前情况看来，自动的标引工作方式仍未占据标引工作的统治地位，在有些检索系统中，同时存在着自动和手工两种标引手段，在有些检索系统中，手工标引是主要的标引方式。

手工标引与目前水平上的自动标引（以下简称自动标引）相比较，各自特点如下：

1. 手工标引效率较低，自动标引效率较高。

标引是一个智力活动过程。对于人工标引来说，它需要标引者具有相当的普通知识、专业知识、文献知识和标引知识，认真领会文献内容，经过归纳、分析、综合等思维活动，选择出最恰当的若干个标引词。这种工作难度使得即使是较高水平的标引人员（甚至是标引专家）也需要耗费一定的精力和时间才能完成对一篇文献的成功标引。自动标引则不同。当把标引工作需要的方法、规律、要求以及待标引的文献输入计算机后，计算机会自动、快速地按照人们规定的方法、规律和要求对文献进行处理。

2. 手工标引质量较高，自动标引质量较低。

如前所述，标引是一个智力活动过程，因此，由具有较高水平的标引员标引出的文献，其可信度是比较高的，能在一定程度上满足人们对标引工作的要求。自动标引则不同。目前的自动标引技术中并没有什么“智能”因素，它是通过文献及其内容成份（句子、短语、词组、词片、字等）的外部形式实现的。这些方法固然有一些理论根据和在实际工作中成功的经验，但它毕竟是沿着人工标引不同的道路发展的。目前水平上的标引质量还不太令人满意。而且，我们还应注意到，标引的目的在于协调标引者和检索者，目前对提问的“标引”工作（选择检索词的工作）还是人工进行的，与自动的文献标引方式相对照，二者在形式上的结果是否完全一致，恐还值得研究。

3. 手工标引的一致性较差，自动标引的一致性较好。

所谓标引的一致性，是指在不同的时间和环境下，对同一篇文

献标引的若干个结果的一致程度。手工标引除受标引工作的一般规律制约外，还受许多偶然因素的影响，这使得手工标引工作的"再现性"较差。同一篇文献由不同的人同时标引，或由一个人在不同的时间和场合标引，结果往往会不一致；自动标引则不同，自动标引是机械的，在输入关于标引工作的一般性方法、规律、要求之后，除非有意识地输入新的指令性信息，可在任何时候再现出同样的标引结果。

4 手工标引多数是赋词标引，自动标引基本上是抽词标引

赋词标引是指标引用词是从词表中抽出赋给该篇文献的，换言之，赋词是规范词。抽词标引是指标引用词是从文献中抽取出来的，只要它被标引者认为恰当表达了文献主题，则不考虑词表对词汇的限制，换言之，抽词是自由词。自动标引工作在抽词标引领域里进展较快。

文献资料自动标引的研究工作目前还远未完成，但它无疑是标引工作必然的发展方向，也是自动化情报检索中的一个重要领域。我们在这里选择介绍一些国内外的自动标引技术方案。

5.2 西文自动标引方案简介

5.2.1 词频统计原理

西文的自动标引方法基本上都是基于词频统计的，为此，我们先介绍词频统计的思想。

人们在长期的文献工作和标引工作中发现，在一篇文献中，某个词在该篇文献中的重要程度往往与该词在该篇文献中的出现频率有关，一个词越重要，它在文献中出现的相对频率就越高，因此，

有可能通过统计词汇在文献中的出现频率，判断它们对于该篇文献的重要程度，从而在其中选取某些频率适当的词作为该篇文献的主题标识。

进一步地，人们还发现了词与词频之间的一些在量上的关系规律。将一篇文献中的各个不同词按其出现频率降序排序，那么各词的频率与该词在这个序列中的序号的乘积近似地等于一个常数。用公式表示为

$$F_i \cdot r_i = C \qquad (5-1)$$

其中 F_i 是某词在一篇文献中的出现频率，r_i 是该词在上述序列中的序号，C 是一常数。这个统计规律被称为 Z_{ipf} 定律。

例如，有人对出现的1000000个词汇进行了统计，其中出现频率最高的十个词的频率及其序号情况如下表所示

序号(r)	词	词频(F)	r·(F/1000000)
1	the	69971	0.070
2	of	36411	0.073
3	and	28852	0.086
4	to	26149	0.104
5	a	23237	0.116
6	in	21341	0.128
7	that	10595	0.074
8	is	10099	0.081
9	was	9816	0.088
10	he	9543	0.095

表5—1 Zipf 定律的一个实例

并不是频率越高的词越有主题标识意义和检索意义。因为这些频率最高的词往往是一些功能词（如上表中的词）和专指度不高的其他词（如某一学科中的通用和常用词汇）。这些词显然不适合甚至不可能用作标引词。与此相反的一个情况是，频率太低的词也不适合作标引词。因为它们在文献中只是偶而出现或很少出现的，我们可以认为它们不代表文献的基本主题和主要内容。

用来作为标引词的那些词汇往往是中频词。

我们可以设想一个指标，标识能力。这个指标表示词作为标引词的合适程度。从高频词出发，频率最高的词作为标引词的能力是很弱的。随着频率的降低，相应的词逐渐能够表示该篇文献的主题和内容。在中频词的某一点上，标识能力达到最大值。但以后则随着频率的继续下降而下降。图示如下。

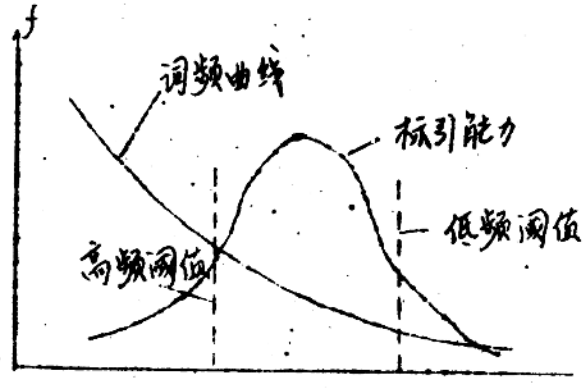

图 5.1　Zipf 定律和标引能力

选择对于一篇文献适合的标引词是如此。选择对于一个检索系统适合的标引词集合同样如此。设该检索系统共有 N 篇文献，选定标引词集合的步骤如下。

1. 对于每一篇文献，计算其中每一个不同词的出现频率，即

主题词k在文献i中的频率F_{ik}。

2. 对于每一个标引词，计算其在全部N篇文献中的出现频率TF_K。

$$TF_K = \sum_{i=1}^{N} F_{ik}$$

3. 按照TF_K的降序，将所有主题词排序。指定一个适当的高频阈值。弃去全部在这个高频阈值之上的词。这些被弃去的词由于在文献集合中出现过频，其出现与否并不过分影响检索性能。

4. 指定一个适当的低频阈值，同样弃去全部在这个低频阈值之下的词。

5. 剩下的中频词被用来标引这N篇文献。

下面给出几种标引方法。每一种标引方法的核心是一个赋权函数。按照这个赋权函数为一篇文献中的各"准标引词"赋权之后，其中权值较高的若干个词便作为该篇文献的标引词。

5.2.2 逆文献频率法

这种方法的理论基础是：假定词的重要程度与该词在该篇文献中的出现频率成正比，而与用这个词标引过的文献篇数成反比（因为，在越少的文献中出现这个词，则这个词对这些文献的标识能力就越强，或者说，区别这些文献与其他文献的能力就越强）。

先看后一个指标，记这个文献篇数为D_K，即在整个检索系统的N篇文献中，有D_K篇文献中出现了主题词k。我们用"逆文献频率因子"来表示D_K对标引词k的权值的影响因子。对逆文献频率因子的一个可能的计算方法是

$$\log_2 \frac{N}{D_K} + 1 = \log_2 N - \log_2 D_K + 1 \qquad (5-2)$$

其中N是整个系统中的文献篇数。

例如，在一个有1000篇文献的集合中，Alpha出现在100篇文献中，Beta出现在500篇文献中，Gamma出现在900篇文献中。于是与这三个词相应的三个逆文献频率因子的值分别为4.322，2.000和1.132。我们看到，有某词出现的文献篇数越少，则该词的逆文献频率因子的值越高，从而对于这个词来说，可以更容易地把这些文献检索出来。

令W_{ik}为文献i中词k的权，我们可以规定以下公式

$$W_{ik} = F_{ik} \cdot (\log_2 N - \log_2 D_K + 1) \qquad (5-3)$$

5.2.3 信号——噪音法

利用信息论的方法也可以给标引词赋权。我们知道，在信息论中，一个信号（或一个词）所含内容的信息量可以通过该信号（或该词）在一次通信（或一个文本）中出现概率的反比函数来表示。该词的出现概率越高，则该词的信息量越小。一个词的信息量定义为

$$I = -\log_2 p \qquad (5-4)$$

其中p是该词的出现频率。

例如，词Alpha在10000个词中只出现了一次，则其出现概率为0.0001，而其信息量为

$$I = -\log_2(0.0001)$$
$$= -(-13.278)$$

仿照香农的信息量公式，我们可以定义一个标识词对一个有N篇文献的文献集合的“噪音”，

$$Noise_K = \sum_{i=1}^{N} \frac{F_{iK}}{TF_K} log_2 \frac{TF_K}{F_{iK}} \qquad (5-6)$$

这个值与该词在文献集合中的“聚集度”成反比变化。即对于绝对平均的分布，该词在每篇文献中出现的频率相同。这时的“噪音”值为最大。例如，设词K在每篇文献中都出现且仅出现一次（$F_{iK}=1$，$i=1, 2, \cdots N$），则

$$Noise_K = \sum_{i=1}^{N} \frac{1}{N} log_2 \frac{N}{1}$$

$$= log_2 N$$

再看另一个极端的情况，即绝对聚集的情况。词K只出现在一篇文献i中（这时显然有$F_{iK}=TF_K$）。则噪音值

$$Noise_K = \frac{TF_K}{TF_K} log_2 \frac{TF_K}{TF_K}$$

$$= 1 log_2 1$$

$$= 0$$

显然，在噪音和标引词的专指度之间存在着联系。因为宽泛的、不专指的词总是趋向于在文献集合中较为分散，因而也有较高的噪音。所以，噪音的反比函数可用来作为标引词权值的因子。一种称为标引词K的信号的函数可定义如下，

$$Signal_K = log_2 TF_K - Noise_K \qquad (5-7)$$

对于噪音的最大值（$F_{iK}=1$，$i=1, 2, \cdots, N$），

"信号"值为0。因为那时有

$$log_2 TF_K = Noise_K = log_2 N$$

而另一方面，当标引词k仅出现在一篇文献中时，噪音值$Noise_K = 0$，而信号值为最大

$$Signal_K = log_2 TF_K - Noise_K$$

$$= log_2 TF_K$$

$$= log_2 F_{iK}$$

其中F_{iK}是标引词k在这唯一的一篇文献中的出现频率。

在原理上，可以将一篇文献中的各标引词按其信号值的降序排列。这种函数有利于将少量文献（信号值高的标引词出现的文献）从其他文献中区别出来。于是，在考虑F_{iK}的值后，我们可以定义一个复合函数作为标引词k在文献i中的权值

$$W_{iK} = F_{iK} \cdot Signal_K \qquad (5-8)$$

5.2.4 词辨别值法

标引工作有着双重目的，一个是标识文献使之在检索时能够实际被检索出来，另一个是区别各篇文献，即要求赋予文献的标引词具有辨别文献的功能。这后一个任务是非常重要的，因为它直接决定检索的查准率。上两节我们介绍的算法中已经考虑了这个因素，本节我们再介绍一种主要考虑这个因素的标引方法——词辨别值法。

考虑一个文献集合。令D_i和D_j为不同的两篇文献，每篇文献各自用一个标引词集合标引，我们可以用一个指标"匹配度"（$Similar(D_i, D_j)$）来度量文献D_i和D_j的相似程度。在

典型的匹配度函数中。往往把两篇没有什么共同点的文献的匹配度规定为0。把两篇完全一致的文献的匹配度规定为1，而大多数文献对的匹配度值则在0与1之间。

能满足上述要求的匹配度计算公式有很多。例如余弦公式。设文献——语词矩阵中D_i和D_j的两行分别是$(a_{i1}, a_{i2}, \cdots, a_{iM})$和$(a_{j1}, a_{j2}, \cdots, a_{jM})$。则令

$$Simolar(D_i, D_j) = \frac{\sum_{k=1}^{M} (a_{ik} \cdot a_{jk})}{\sqrt{\sum_{k=1}^{M} a_{ik}^2 \cdot \sum_{k=1}^{M} a_{jk}^2}}$$

设除$i = j$之外的所有文献对D_i和D_j的匹配度都已计算出。我们可以得到一个平均匹配度$\overline{Similar}$，其值为

$$\overline{Similar} = C \cdot \sum_{i=1}^{N} \sum_{j=1}^{N} Similar(D_i, D_j) \qquad (5-9)$$

其中C是一常数。其值可以任意指定（如指定为$\frac{1}{N(N-1)}$）。这个表达式给出了文献在文献空间中“聚集”的程度。当所有的文献都相同时，所有的$Similar(D_i, D_j) = 1$。这时的平均匹配度达到最大值

$$\overline{Similar} = C \cdot \sum_{i=1}^{N} \sum_{j=1}^{N} 1$$

$$= C \cdot N(N-1)$$

空间密度可以以更方便的方法计算。这种方法是。构造一个人为的“平均文献”$\bar{D}$，作为一个中心，在这个中心中，标引词被假定具有平均的频率。即，标引词k的平均频率被定义为

$$\bar{F}_K=\frac{1}{N}\sum_{i=1}^{N}F_{iK} \qquad (5-10)$$

于是平均匹配度可以通过以下方式计算：

$$\bar{S}imilar=C\cdot\sum_{i=1}^{N}Similar(\bar{D},D_i) \qquad (5-11)$$

考虑一个文献集合，其平均匹配度为$\bar{S}imilar$，现在把这个集合中的所有标引词k移除，并记移除掉标引词k后这个文献集合的平均匹配度为$(\bar{S}imilar)_K$。首先设k是一个宽泛的高频词，并且有均匀的分布，则把它除去后，会降低平均的文献匹配度，即

$$(\bar{S}imilar)_K<\bar{S}imilar$$

其次，设k是一个专指度很高的低频词，在原文献集合中，它只在少量文献中有较高的权值，而在其他文献中却权值很小或为零，那么把它移除后，会增加文献集合的平均匹配度，即

$$(\bar{S}imilar)_K>\bar{S}imilar$$

于是，对每个标引词k，可定义词辨别值DV_K

$$DV_K=(\bar{S}imilar)_K-\bar{S}imilar \qquad (5-12)$$

在计算了每个标引词k的DV_K值之后，可将这些词按DV_K值的降序排列，排序后我们将会发现，序号越小的词越专指，序号越