信息技术领域术语字频、词频及术语长度统计*

李芸　王强军

（北京语言文化大学，北京 100083）

E-mail: liyun@blcu.edu.cn; wangqj@blcu.edu.cn

摘要：本文对两本信息技术术语词典中的术语的用字、用词、术语长度等进行了统计，并做了比较分析。另外，还对术语系统的经济指数等做了一些初步的探讨。

关键词：信息技术；术语；术语用词；术语用字；术语长度

引言

术语是语言词汇的一部分，研究术语系统中的用字、用词情况以及术语的语法结构、语义结构将有助于术语的标准化工作，还会对术语的自动提取提供有用的计量参照。本文对两本信息技术领域术语词典中的术语作了细致的统计和比较，总结出一些规律。这两本词典是《英汉网络技术词汇》和《计算机科学技术百科全书》，均由清华大学出版社出版。本次统计使用的术语量分别是：《英汉网络技术词汇》（以下简称“网络技术”）56,609 条，《计算机科学技术百科全书》（以下简称“百科全书”）1,291 条，总字数分别为 282,068 字、6,754 字。所有术语均由人工分词，人工校对，错误率在千分之一以下。

1 术语的长度

本文统计了术语的两种长度，一个是术语包含的字数，叫做术语含字长度，另一个是术语包含的词数，叫做术语含词长度。

*本项目受到教育部人文社会科学研究规划基金资助(01JA740008)。

1.1 术语含字长度

1.1.1 网络技术术语

表1 网络技术术语含字长度分布表

术语含字长度	术语条数	百分比（%）	术语例子
2	4,354	7.6914	蠕虫
3	5,446	9.6204	熟模式
4	**15,306**	**27.0381**	梳状编码
5	**9,070**	**16.0222**	X 系列建议
6	**11,557**	**20.4155**	直接存取装置
7	5,909	10.4383	可选分时软件包
8	4,691	8.7636	撕断纸带式转换法
9+	6	0.0106	偏置量数据区的区距
合计	56,609	100.00	

术语平均含字长度：

总字数/术语总条数=282068/56609＝4.9827 （字/条）

1.1.2 计算机百科全书术语

表2 百科全书术语含字长度分布表

术语含字长度	术语条数	百分比（%）	术语例子
1	8	0.6197	区
2	67	5.1898	信念
3	107	8.2881	感知器
4	**338**	**26.1813**	联想记忆
5	**245**	**18.9775**	波尔兹曼机
6	**231**	**17.8932**	小脑网络模型
7	139	10.7668	自组织映射模型
8	69	5.3447	句法模式识别方法
9	42	3.2533	基于合一的语法理论
10	17	1.3168	Hopfield 神经网络模型
11	14	1.0844	自然语言处理的句法分析
12	7	0.5422	离散事件系统仿真输出分析
13	3	0.2324	离散事件系统仿真建模方法学
15	2	0.1549	人工神经网络在模式识别中的应用
16	1	0.0775	通用键盘汉字编码输入方法评测规则
20	1	0.0775	基于ISO/IEC10646和Unicode的汉字编码字符集
合计	56,609	100.00	

术语平均含字长度：

总字数/术语总条数=6754/1291=5.2316（字/条）

1.1.3 比较

百科全书的术语含字长度稍大于网络技术的术语含字长度。可以认为，术语含字长度在5字左右，与邢红兵统计（参见文献[4]）基本相同。

1.2 术语含词长度

1.2.1 网络技术术语

表3 网络技术术语含词长度分布表

术语含词长度	术语条数	百分比（%）	术语例子
1	4,412	7.7938	蠕虫
2	**20,146**	**35.5880**	模拟 视频
3	**19,778**	**34.9379**	直接 存取 装置
4	**10,860**	**19.1842**	可选 分时 软件 包
5	1,398	2.4696	可 编程 多路 转接 器
6	15	0.0265	包 装配 器 和 拆卸 器
合计	56,609	100.00	

术语平均含词长度：

总词数/术语总条数= 154559/56609 = 2.7303 （词/条）

1.2.2 计算机百科全书术语

表4 百科全书术语含词长度分布表

术语含词长度	术语条数	百分比（%）	术语例子
1	74	5.7320	映射
2	**458**	**35.4764**	计算 理论
3	**440**	**34.0821**	小脑 网络 模型
4	216	16.7312	数字 计算 误差 分析
5	66	5.1123	光纤 分布 式 数据 接口
6	23	1.7816	自然 语言 处理 的 词法 分析
7	10	0.7746	汉字 编码 字符 集 标准 体系 结构
8	2	0.1549	通用 键盘 汉字 编码 输入 方法 评测 规则
9+	2	0.1550	人工 神经 网络 在 模式 识别 中 的 应用
合计	1,291	100.00	

术语平均含词长度：

总词数/术语总条数= 3747/1291 = 2.9024 （词/条）

2 词频统计

对构成术语的词在术语系统中的词频、词的长度进行统计，详见下表。

2.1 网络技术术语

构成术语的不同的词有9,858个，其中有19个西文单词或缩略词。出现频度最高的前20个词见下表。

表5 网络技术术语用词的高频词表（前20个）

序号	1	2	3	4	5	6	7	8	9	10
词	器	网络	系统	通信	数据	机	程序	控制	法	信号

频度	4,487	3,184	2,451	2,253	1,858	1,666	1,619	1,582	1,420	1,309
序号	11	12	13	14	15	16	17	18	19	20
词	交换	传输	网	信息	协议	服务	处理	终端	式	线路
频度	1,248	1,151	1,142	1,102	1,095	1,080	1,056	963	805	781

表 6　网络技术术语用词的词长分布表

词长	不同词的个数	出现次数（词次）	例词
1	**710**	**28,217**	器
2	**8,787**	**125,369**	方法
3	235	814	无线电
4	98	127	自顶向下
5	25	29	印度尼西亚
6	3	3	斯堪的纳维亚
合计	9,858	154,559	

从上表可以看出，构成术语的词的多是 1 字词和 2 字词，而且这些词分布也很广。

2.2 计算机百科全书术语

构成术语的不同的词有 989 个，其中有 94 个西文单词或缩略词。出现频度最高的前 20 个词见下表。

表 7　百科全书术语用词的高频词表（前 20 个）

序号	1	2	3	4	5	6	7	8	9	10
词	机	器	计算	系统	语言	程序	存储	处理	数据	图象
频度	134	132	117	89	88	55	51	50	44	43
序号	11	12	13	14	15	16	17	18	19	20
词	方法	汉字	设计	控制	式	输入	库	性	软件	技术
频度	41	41	39	35	33	30	28	28	26	25

表 8　百科全书术语用词的词长分布表

词长	词数（个词）	词频（词次）	例词
0	10	13	Z
1	**143**	**728**	主
2	**772**	**2,929**	最优
3	44	55	中日韩
4	20	22	人助机译
合计	989	3,747	

从上表可以看出，构成术语的词的多是 1 字词和 2 字词，而且这些词分布很广，与表 6 有相似之处。

3 字频统计

3.1 网络技术术语

构成术语的不同的字有 1,967 个汉字，还有 16 个符号和西文单词或缩略词。出现频度最高的前 20 个字见下表。

表9　网络技术术语用字的高频字表（前20个）

序号	1	2	3	4	5	6	7	8	9	10
字	信	网	器	路	通	数	线	络	电	接
频度	6,062	5,344	4,615	4,172	3,597	3,470	3,249	3,209	3,202	3,085
序号	11	12	13	14	15	16	17	18	19	20
字	程	制	系	统	机	传	分	用	号	理
频度	2,397	2,784	2,698	2,541	2,421	2,272	2,206	2,164	2,151	2,136

3.2 计算机百科全书术语

构成术语的不同的字有691个汉字，另外有86个符号和西文单词或缩略词。出现频度最高的前20个字见下表。

表10　百科全书术语用字的高频字表（前20个）

序号	1	2	3	4	5	6	7	8	9	10
字	机	计	算	器	语	数	理	法	系	程
频度	170	161	154	150	127	121	114	109	101	100
序号	11	12	13	14	15	16	17	18	19	20
字	统	言	字	图	设	式	型	序	制	存
频度	92	89	78	75	64	63	63	63	62	61

4 术语系统的其他指数

4.1 术语系统的经济指数

定义1. 在一个术语系统中，术语总条数除以构成术语的不同词的总数所得的商，称为该术语系统的经济指数，记作E，单位是“条/词”。

其计算公式可表示为：E = 术语总条数/不同词的总数　　(1)

网络技术术语系统的经济指数是：

$E = 56609/9858 = 5.7424$

百科全书术语系统的经济指数是：

$E = 1291/989 = 1.3054$

在大多数术语系统中，$E>1$；如果$E \leq 1$，则说明术语系统设计的经济效应不高（冯志伟，1997）。比较上述两个术语系统的经济指数，我们发现，网络技术术语系统的经济指数较高。也就是说，每个词平均可构成5.7424条术语。可见，网络技术术语系统有较高的经济效应。当然，E受到系统中术语数量的强烈影响，一个是56,609条，一个是1,291条。再深入分析其差别原因，还有可能是因为网络技术词典中的术语条目是按英文术语的不同翻译视为不同的词条而造成的，百科全书中的术语不会有这种情况。

4.2 词的术语构成频率

定义2. 在一个术语系统中，构成术语的词的总数除以构成术语的不同词的总数所得的商，称为词的术语构成频率，记作F，单位为“次”。

其计算公式可表示为：F = 总词数/不同词的总数 (2)

网络技术术语系统的词的术语构成频率：

F = 154559/9858 = 15.6785

网络技术术语系统的词的术语构成频率：

F = 3747/989 = 3.7887

F 的值不小于 1，即 F≥1；对于同一个术语系统来说，词的术语构成频率 F 不能小于术语系统的经济指数 E，即 E≤F，因为术语的总条数小于等于术语用词的总词频。F 表示构成术语的每个词平均出现多少次。因此，这个值可以代表这些词构成术语的平均频率。术语系统中的高频词越多，该系统中词的术语构成频率也就越高。比较上面的两个 F 值，可以知道网络技术术语系统中的词的术语构成频率较高。

4.3 术语形成的经济率

根据定义 1 和定义 2，（2）除以（1），得到

F/E =总词数/不同词的总数*不同词的总数/术语总条数 =总词数/术语总条数

又根据术语平均含词长度公式，得到

F = E * L

其中 L=总词数/术语总条数

即系统的经济指数 E 与术语的平均含词长度 L 的乘积，恰恰等于词的术语构成频率 F。其实，这就是术语形成的经济率——FEL 公式（详见冯志伟，1997，P127-128）。

变换 FEL 公式，可以得到：

E = F/L

即术语系统的经济指数 E 与词的术语构成频率 F 成正比，而与术语的平均含词长度 L 成反比。如果想提高术语系统的经济指数，可以从这两个方面来考虑。一般来说，术语的平均含词长度不能改变过大，所以增加词的术语构成频率是比较好的方法。

5 结束语

对术语系统中的字频、词频和术语长度进行统计分析是术语研究的基础部分，从中可以发现术语构成的一些基本规律。限于篇幅，本文给出了一些统计列表，分析不算太多。我们认为，更进一步的研究还应该在术语的语言学方面找出规律，包括术语的语法结构、语义结构以及术语的生长模式等等。

参考文献：

[1] 冯志伟，现代术语学引论，北京：语文出版社，1997。

[2] 章鸿猷主编，英汉网络技术词汇，北京：清华大学出版社，2000。

[3] 张效祥主编，计算机科学技术百科全书，北京：清华大学出版社，1998。

[4] 邢红兵，计算机领域汉英术语的特征及其在语料分布规律，术语标准化与信息技术，，2000 年第 4 期。

作者简介：李芸（1970—），女，博士生，主要研究领域为动态语言知识更新，术语自动提取和分类；王强军(1973—)，男，博士生，主要研究领域为术语自动提取，动态语言知识更新。

Character Frequency, Word Frequency and Length of Terms in the Field of Information Technology*

LI Yun, WANG Qiangjun

(Beijing Language and Culture University, Beijing 100080, China)

E-mail: liyun@blcu.edu.cn; wangqj@blcu.edu.cn

Abstract: This paper lists the character frequency, word frequency and length of terms in the field of Information Technology in detail, and makes comparative analysis between terms from two dictionaries. Then, it presents additional index and principle in the terminology system.

Key words: Information Technology; terms; characters of terms; words of terms; length of terms;

* Supported by the Ministry of Education Social Science Foundation of China under grant No. 01JA740008