信息技术领域术语提取的初步研究*

王强军　李芸　张普

（北京语言文化大学，北京 100083）

E-mail: wangqj@blcu.edu.cn, liyun@blcu.edu.cn, zhangpu@blcu.edu.cn

摘要： 本文对信息技术领域术语自动提取方法进行了实验，提出了领域相减的术语提取方法，即根据流通度理论，利用术语在不同领域中的不同流通度值进行术语提取。评价了领域相减法在术语自动提取中的作用。

关键词: 术语提取；流通度；领域相减；语言信息处理

引言

术语集中体现和负载了一个学科领域的核心知识，术语的变化在一定程度上反映了一个学科领域的发展变化。术语提取对于信息检索、信息提取、数据挖掘等语言信息处理研究以及了解、把握一个学科领域的发展现状、未来趋向等具有重要的理论和现实意义。本文通过对信息技术领域术语自动提取的尝试，探讨对于术语的认识和面向特定领域的术语提取方法。

1 基本概念和本文的讨论范围

下面给出本文用到的几个名词的解释，和对它们的认识：

术语（Term）：在一个学科领域中使用，表示该学科领域内概念或关系的词语。术语可以是词，也可以是短语。术语可以只在一个学科领域中存在，也可并存于多个学科领域中。

一般词语（Common Words）：一个学科领域中除了术语之外的词语都叫做一般词语。所有学科领域中一般词语的并集构成了一般词语的全集。一般词语的全集加上所有学科领域的术语构成语言交际的词语的全集。

学科领域（Field）：人类知识的一门分科或一个专业范围。本文采用的学科分类体系以人类知识体系为框架，以便于进行术语提取和其他语言信息处理为原则。详见李芸、王强军（2001）。

信息技术（Information Technology）：应用信息科学的原理和方法研究信息的产生、获取、变换、传输、存储、处理和利用的工程技术，又称信息工程。信息技术是在计算机、通信和控制技术的基础上发展起来的。

流通度（Circulation）：一个语言单位流行通用的程度。它揭示了一个语言单位在社会生活中发展演变的过程。详见张普（1999a）

* 本项目受到教育部人文社会科学研究规划基金资助(01JA740008)。

术语提取的实质是确定术语的前界和后界。按照术语的前后界有无明显标记，术语可分为三类：有前后界标记的，有前界或后界标记的，无前后界标记的。第一类如科学论文中的关键词，往往都有明显的标志，说明它们是该篇文章的重点词汇，也是信息检索的重要途径之一。第二类中又有很多种情况，例如跟在某些词（称为、叫做等）后面的极有可能是术语；又如文本中的双语词语，一般是一种语言的术语后跟一个带括号的注释，这个括号即可看作是前面术语的后界标记。第三种情况是没有任何标记的术语。它们混杂在文本中，数量大，分布广，是术语提取的重点所在，也是难点所在。

本文着重讨论第三类术语，无前后界标记术语的提取。它包括两方面的工作：

1. 确定术语的前后边界，保证它是一个合法的语言单位；
2. 把它跟一般词语去分开，保证它是一个术语而非一般词语。

2 无前后界标记术语的提取方法

2.1 术语的一般特征

张普（2001）论述了术语和一般词语的关系，指出：

（1） 术语一般只在一个或几个特定的领域流通，只有该特定领域的人使用，而一般词语是各个领域都流通，是所有使用该语言的人通用的。

（2） 术语不仅只在本领域流通，一般说术语也都是本领域的高流通度的词语。

（3） 术语不仅在本领域是高流通度的，离开了特定领域，其流通度一般趋近于零。例如：半数致死量、氯代三环芳烃类化合物、多氯代二苯。

（4） 一般词语集合在每个领域中都是共用的，所以基本上是个常数；术语是各个专门领域独用的词语；各个领域互不相同。

（5） 每个学科领域的词语集合由一般词语集合加上这个领域的术语组成。

2.2 技术路线

基于以上认识，可以有如下无前后界标记术语的提取方法：

（1） 确定待提取术语的领域，称作待处理领域；

（2） 选定一个在术语使用上与待处理领域区别较大的领域，称作对照领域；

（3） 计算各领域词语的流通度；

（4） 对两个领域内词语流通度相减，确定阈值，去除一般词语，得到处理领域的候选术语表；

（5） 重复步骤（1）—（3），直到去除的一般词语数量极少时停止；

（6） 利用其它手段进一步缩小候选术语表的范围。

2.3 全切分

在上述技术路线中有一个重要的问题，就是汉语分词的歧义切分问题。按照一般的做法，要确定什么是术语，先要确定什么是词；要确定什么是词，就不能回避汉语的分词问题。如果那样，这里提到的方法就要打一个大的折扣，因为汉语分词是语言信息处理的一个老大难问题，这个老大难问题的“最后一公里”集中在未登录词语的识别和歧义切分上，而有意义的术语提取就是要提出新术语——属于特定领域的未登录词。所以在步骤（1）和步骤（2）之间实际上有一个不受未登录词和歧义切分问题影响的语料切分过程，我们称作全切分，即把任意长度（不超过句长）的一个语言片段都看作一个候选术语来进行处理。

3 实验和结果

3.1 语料的选取和领域的确定

实验选取信息技术领域语料 10 万字进行手工术语提取，作为标准答案。选取理论语言学作为对照领域，选取信息技术领域和理论语言学语料各 10 万字进行实验。

从 10 万字的语料中共提取出术语 6635 词次（Token），合 1854 词形（Type），平均出现频率 3.58。术语（词次）总字数 28568 字，平均术语长度 4.31 字。术语长度集中在 2 字到 5 字，所占比例均在 10% 以上，这跟邢红兵（2000）所列术语长度集中于 4 字和 6 字的结论不同。原因有二：（1）邢文所处理的是英汉双语术语，其汉语部分跟本文所讨论的单语术语有所不同；（2）人工提取术语存在一定的不确定因素，跟各人对术语的理解有关。详见下表：

表 1 术语占总语料的比例

	术语	一般词语	总语料
字数	28568	71086	99654
百分比（%）	28.67	71.33	100.00
词次（Token）	6635	—	—
词形（Type）	1854	—	—

表 2 术语长度比例

长度（字符数）	1	2	3	4	5	6	7	8	9
条数	28	329	255	539	225	169	112	62	26
百分比%	1.51	17.75	13.75	29.07	12.14	9.12	6.04	3.34	1.40
长度（字符数）	10	11	12	13	14	15	16	>16	总数
条数	41	14	14	8	7	8	5	26	1854
百分比%	2.21	0.76	0.76	0.43	0.38	0.43	0.27	1.40	100.00

3.2 领域相减方法

对信息技术领域和对照领域的语料进行全切分，按照频率排序，减去相同的部分，得到候选术语 3375 条，与手工作出的答案相比，其中包含术语 1005 条。由此得出准确率 P＝1005 / 3375 = 30%，召回率 R = 1005 / 1854 = 54.2%。如果扩大两个领域的比较范围，则随着召回率的升高，准确率相应下降。图 1 显示了二者的关系。

需要指出的是，以上领域相减是在对语料进行简单操作后进行的，所以准确率和召回率相对较低。但是由此可以得到一个比较小的候选术语表。在此基础上再进行进一步的筛选，就不会费很大力气。

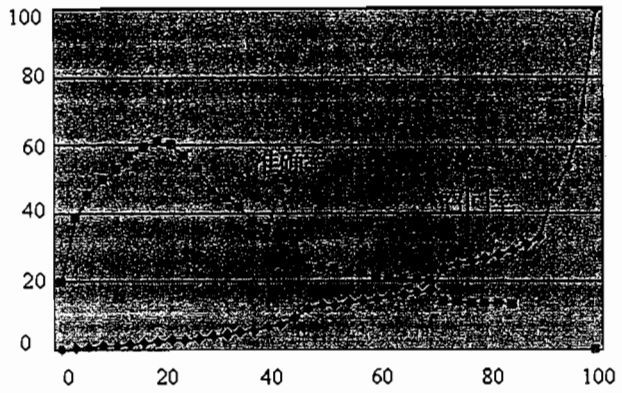

图 1 召回率和准确率

图 1 中的递增曲线（蓝色）是召回率曲线，另一条曲线是准确率曲线（红色）。

3.3 关于对照领域

这次实验只选了一个对照领域，理论语言学。语料量 10 万字。实验结果表明，尽管感觉上计算机和语言学中间有一个计算语言学的交叉学科因而显得比较接近，但是对领域相减提取术语的影响并不太大。用对照语料和 3.1 的标准答案相比重合的有 18 条，占总条数 1854 的 1%。这十八条是：表、层、对象、方法、过程、句法、类、事件、谓词、系统、显示、写、信息、语句、语言、语义、属性。它们在两个领域中的出现频率并不相等，数据表明除了“谓词”和“语句”之外，其它 16 条在信息技术领域语料中的频率远高于对照领域，从而不会在相减时去掉这些术语。

4 结束语

本文采用领域相减的术语提取方法对信息技术领域进行术语提取实验。通过对实验数据进行分析，认为领域相减的方法在术语提取中有一定的作用，可以在不受分词问题影响的全切分方法下，较快缩小候选术语表的规模，为后续筛选和提取提供方便。

参考文献:

[1] GB 10112－88，确立术语的一般原则与方法

[2] 张普（2001），流通度在ＩＴ术语识别中的应用分析——关于术语、术语学、术语数据库的研究，辉煌二十年——中国中文信息学会二十周年学术会议论文集，2001 年 11 月。

[3] 张普（1999），关于大规模真实文本语料库的几点理论思考，语言文字应用 1999 年第 1 期。

[4] 张普（1999a），关于语感与流通度的思考，语言教学与研究，１９９９年第 2 期。

[5] 张普（1999b），关于网络时代语言规划的思考，语文研究，１９９９年第 3 期。

[6] 隋岩、张普（1999），１９９７中文报纸媒体流通度分析，１９９９年计算语言学年会论文集。

[7] 邢红兵（2000），基于第三代语料库的信息领域术语动态更新，语言文字应用，2000 年第 2 期。

[8] 邢红兵（2000），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。

[9] 李芸、王强军（2001），信息技术领域术语自动提取研究，辉煌二十年——中国中文信息学会二十周年学术会议论文集，2001 年 11 月。

作者简介： 王强军（1973—），男，博士生，主要研究领域为术语提取；李芸(1970—)，女，博士生，主要研究领域为信息提取。张普，男，博士生导师，教授，主要研究领域为动态语言知识更新。

Automatic Term Extraction in the Field of Information Technology*

WAGN Qiangjun[1] LI Yun[1] ZHANG Pu[1]

[1](*Beijing Language and Culture University, Beijing 100080, China);*

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

Abstract: In this paper, we presented a method for automatic term extraction named *lingyu xiangjian*, which means to shorten the candidate term list in multi fields corpora. The result was analyzed and evaluation was given.

Key words: Term extraction; circulation; LingYu XiangJian; language information processing

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