# 中文叙词表本体概念定义注释的自动构建研究*

田金凤[1] 曾新红[1,2] 黄华军[2] 林伟明[2]
[1](深圳大学计算机与软件学院 深圳 518060)
[2](深圳大学图书馆 深圳 518060)

【摘要】设计面向综合性中文叙词表本体的叙词概念定义抽取方法,获得良好的实验效果并已投入实际应用。其中,基于"高频词与句子向量"和"TF * IDF 向量"两种定义抽取算法提出的二维相对量的融合算法,能够更有效地抽取出前两种方法的良好结果,有效信息提高比一般可达到60%。

【关键词】中文叙词表本体 OTCSS 定义抽取 向量空间模型 高频词与句子向量 TF * IDF 向量 二维相对量
【分类号】TP18 TP301.6

# Research on Automatic Construction of Definition Notes for Concepts in OntoThesaurus

Tian Jinfeng[1] Zeng Xinhong[1,2] Huang Huajun[2] Lin Weiming[2]
[1](College of Computer and Software, Shenzhen University, Shenzhen 518060, China)
[2](Shenzhen University Library, Shenzhen 518060, China)

【Abstract】The paper proposes some methods of definition extraction for concepts in the comprehensive OntoThesaurus. They achieve good experiment effects and are applied to the actual OTCSS. Among them, an integrated algorithm named "two - dimensional relative quantity" based on "high - frequency words vector" and "TF * IDF vector" is presented. This algorithm can much effectively extract good results from that of the first two methods, and the effective information improving ratio can reach 60% generally.

【Keywords】OntoThesaurus OTCSS Definition extraction VSM High - frequency words vector TF * IDF vector Two - dimensional relative quantity

## 1 引 言

概念的定义是人们理解该概念最直接、最清晰的信息。W3C 的推荐标准 SKOS(简单知识组织系统)为概念的定义专门设置了一个注释属性 skos:definition 进行描述,以提供一个概念的本意的完整解释[1]。

中文叙词表本体(OntoThesaurus)融合了叙词表与本体[2],是一种同时具备二者特征的知识组织系统[3]。综合性中文叙词表本体包含了各个领域的词汇,其中不乏一些专业术语,一般用户在使用和参与其建设时可能会遇

收稿日期:2011-09-22
收修改稿日期:2011-11-04
*本文系广东省哲学社会科学"十一五"规划项目"中文知识组织系统的形式化语义描述标准体系研究"(编号GD10CTS02)和广东省自然科学基金团队项目"新型计算模式及其软件开发方法研究"(编号:10351806001000000)的研究成果之一。

到障碍。如果单纯依靠修订专家手工添加叙词的定义注释将是一项非常耗时耗力的任务。因此，自动构建叙词概念的定义注释成为中文叙词表本体共建共享系统(OntoThesaurus Co - construction and Sharing System, OTCSS)[4,5]需要具备的功能。这一功能的实现可以极大地减轻修订专家的负担，也可以进一步给广大的网络用户带来方便。

针对上述问题，本文力求找到一种解决 OTCSS 中叙词定义注释自动构建的方法。以《中国分类主题词表》为例，设计了面向综合性中文叙词表本体中叙词概念的定义注释语句抽取方法，获得了良好的实验效果，且已应用于实际的 OTCSS 系统。

## 2 研究背景及本文的主要工作

### 2.1 定义抽取研究现状

定义抽取是一个较新的研究课题，是信息抽取[6]的一个新的分支，其目的是使人们在网络搜索时能迅速定位到词语的定义信息，提高信息抽取的有效性和实用性。国内外在这方面的研究起步都较晚，但是也取得了一定的研究成果[7-14]，主要集中于自然科学领域，对于人文领域的术语定义抽取还鲜有涉及。主要研究成果包括：

(1)定义规则匹配模式

贾爱平[7]提出了关于术语定义的识别研究课题，通过对大量科技文献真实语料的阅读，总结出了 8 种定义语言模式和 5 种定义排除语言模式，并提出了定义抽取的评价方法。这些定义规则的总结为后来的定义自动抽取奠定了形式化的基础。很多人利用这些规则对定义进行自动化抽取，并得到了不错的结果。

(2)智能匹配

张榕等[8]提出了结合句子术语定义隶属度和高频词与句子向量的智能匹配算法。该算法首先通过计算词语的术语定义隶属度来计算句子的术语定义隶属度，根据所得的结果对句子降序排列，取前 5 个；在所有候选定义语句中选取出现频率最高的 15 个词作为一个向量，即高频词向量。然后将候选定义语句作为另一个向量，计算两个向量之间的余弦相似度。同样根据计算结果对句子降序排列，取前 5 个；最后将这两者计算结果结合起来，取综合值最高的一个句子。这种方法适合领域性强的术语，但对于综合性的词表作用有限。

(3)SP

其核心算法是质心算法[9]。主要思想是通过计算术语在句子中的位置得到统一的模式，模式中包含有槽，为每一种位置的槽中可能出现的词语加权。利用训练集得到模式和权值，再将训练得到的模式和权值应用到网络中去。这种方法是建立在英文语法的基础上，中文定义语句中术语出现位置灵活，且已有贾爱平[7]总结了符合中文语法习惯的定义规则模式，所以该算法对本文帮助有限。

(4)最大熵分类器

将目标术语(Target Term)提交给搜索引擎；返回以目标术语为中心的 250 字长度的片段，称作窗口(Window)；将窗口分类成可接受的语句和不可接受的语句两种类型；系统返回定义语句[10]。同 SP 方法一样，都是建立在英文语法的基础上。

(5)SVM 分类器

该算法需要两类特征量：术语特征量，指术语词频、术语长度、术语内部连接度及术语邻接上下文的信息熵等；术语定义中的词汇特征，它是利用句子的术语定义隶属度来描述的[11]。这种词汇特征与张榕等[8]提出的方法类似。SVM 分类器根据这两个特征量进行计算，筛选得到最终结果。这种方法的主要目的是利用已有的定义模式发现伴随着定义语句出现的新术语，与本文出发点不同。

### 2.2 主要工作

本文以网络为语料来源，利用搜索引擎从网络中获取所需的定义注释信息，寻找合适的定义匹配模式；根据中文叙词表本体的相关理论，设计了面向以中国分类主题词表为代表的综合性叙词表中的叙词抽取定义注释语句的方法。

主要创新点在于：

(1)在借鉴前人研究成果的基础之上，做了以下改进：在抽取实际语料的过程中，针对历史叙词和地理叙词，分别提出了一种新的抽取模式：String + [“,”] + “[“]史称” + theme + [”] 和 theme + [“,”] + [“是|就”] + (“位于”|“坐落”) + String；在前人基础上扩展了高频词向量的计算范围，使其计算可以灵活使用。

(2)将文本检索领域的 TF * IDF 算法引入定义抽取领域，以弥补仅使用高频词造成的抽取结果具有片

面性。

(3)提出了基于高频词与句子向量和TF * IDF向量两种定义抽取算法的二维相对量的融合算法,更有效地抽取良好的定义语句。

## 3 面向综合性中文叙词表本体的叙词概念定义抽取方法

### 3.1 叙词定义规则匹配

通过对网络数据和规则模板的研究与分析,构建了符合本系统要求的定义模板,即定义句子匹配规则。两个句子的结束符,包括句号、问号和感叹号之间的字符串,如果能有一个字符串同以下某一个模板匹配,则这个字符串就是要抽取的候选定义句子。

定义注释规则的匹配模式用类似正则表达式的方式来表达。将句子分为几个可选的匹配项,每一项之间用加号(+)连接,其中方括号表示里面的内容可以出现也可以不出现,圆括号表示里面的内容必须出现一项,圆括号中用竖线分开的内容表示可选项,即只出现其中一项即可。String表示出现在句子中的词语串,不限定其内容。双引号("")表示里面的内容必须与原文中的内容一致。theme表示需要抽取其定义注释信息的叙词。

本文所使用的定义注释抽取规则如表1所示[7](其中加下划线的为叙词):

表1 本文所用的定义规则匹配模式及例句

| | 模式 | 例句 |
|---|---|---|
| 模式1 | theme + [",","]("即"\|"是"\|"指"\|"就是"\|"是指"\|"指的是") + String | 阿旃陀是古代梵语阿蘧提耶的音译,意为无想。 |
| 模式2 | String + ("称"\|"即"\|"叫"\|"称为"\|"叫做"\|"就是"\|"称之为"\|"定义为") + theme | 奉行不结盟,并确实没有结盟的国家就是不结盟国家。 |
| 模式3 | theme + ["主要"]("包含"\|"包括") + String | 并串行转换器包含触发器和复用器,并且用于N个并行数据比特并且以N倍时钟速度将它们串行移出到发射机。 |
| 模式4 | theme + ("由"\|"是由") + String | 保险法由合同法和业法两部分组成。 |
| 模式5 | theme + ["也"\|"又"\|"简"]("称"\|"叫"\|"称为"\|"称之为") + String | 春节又称元日、元旦、无正、元辰、元朔、岁旦、岁首、岁朝、新正、首祚、三元或年、过年,为夏历新年的第一天。 |
| 模式6 | theme + ":" + String | 八国联军:1900年英美德法俄日意奥八个帝国主义国家的侵华联军。 |
| 模式7 | "所谓" + theme + [",","]("即"\|"是"\|"指"\|"就是"\|"是指"\|"指的是") + String | 所谓城市中心论,是对欧洲资本主义国家的无产阶级以城市为中心的革命道路理论的简称。 |
| 模式8 | String + [",","] + "["]史称" + theme + ["] | 公元960年后周大将赵匡胤借口有外敌入侵,调集兵力出大梁(今河南开封),至陈桥驿(今开封东北)授意将士给他穿上黄袍拥立他为帝,成为宋朝的开国皇帝,史称陈桥兵变。 |
| 模式9 | theme + [",","] + ["是\|就"] + ("位于"\|"坐落") + String | 阿斯旺水坝位于埃及开罗以南900公里的尼罗河畔。 |

其中,模式8和模式9是在实际语料的抽取过程中总结得出的,而模式1至模式7均改编于贾爱平[7]总结的定义规则模式。

除了使用定义注释抽取规则以外,还使用了以下定义注释排除规则:

(1)String(?|!)

以问号结尾的句子和以感叹号结尾的句子往往都不能清晰地解释一个词语。

(2)|String| <15 或 >500

这个规则对句子的长度做了一个限制。

(3)|letter| >100

如果一句话中的英文字母连续超过100个,说明这句话的杂乱信息较多。

### 3.2 向量空间模型

(1)理论基础

向量空间模型[15-17](Vector - Space Model, VSM)因为将文档的内容形式化为多维向量空间中的一个点,将文档以向量的形式定义到实数域中,使模式识别中的各种成熟算法得以采用,提高了自然语言文档的可计算性与可操作性,所以成为最常用的文本表示模型。

①文档:通常指文章中一定长度的片断,如句子、段落或整篇文章。本文将句子看作文档,因此在下文中句子和文档不加区分。

②特征项:通常使用VSM中语言单元,如字、词、词组等,则文档就被看作是由其包含的特征项所组成的集合。本

文将词看作文档的特征项。

③权重:特征项的权值。本文为此设计了两种表征权重的方法。

④相似度:有任意两个文档 $D_1(w_{11},w_{12},\cdots w_{1k}\cdots,w_{1n})$ 和 $D_2(w_{21},w_{22},\cdots w_{2k}\cdots,w_{2n})$,它们的相似度记作 $Sim(D_1,D_2)$。向量空间模型的相似度也有多种计算方法,如内积、余弦值等[15,16]。最常用的是余弦系数,这也是本文采用的方法,公式如下[15]:

$$Sim(D_1,D_2)=\cos\theta=\frac{\sum_{k=1}^{n}w_{1k}\times w_{2k}}{\sqrt{(\sum_{k=1}^{n}w_{1k}^2)(\sum_{k=1}^{n}w_{2k}^2)}} \quad (1)$$

(2)高频词与句子向量

记作 hifre * sent 权值法,利用高频词向量与句子向量的余弦值来衡量定义语句。其主要思想是以一句话为单位,把每一句话看作一个文档 D,对其进行分词、停用词过滤等处理,将文档 D 中的每一个词作为一个特征项 $t_k$,统计所有经过处理后的文档中的词语出现的次数 $N_{tk}$,即把所有词语都记录在一个 Hifre_word 数组中,然后统计它们在文档集 C 中出现的频率 $f_{tk}$,用词频 $f_{tk}$ 作为特征项的权值,并按照降序排列,即高频词向量 $hifre(f_{tk1},f_{tk2},\cdots f_{tki}\cdots,f_{tkn})$。进而统计所有词语在每一个文档中的出现次数 $N'_{tk}$,将词语出现在单个文档中的频率 $f'_{tk}$ 作为特征项的权值,计算出每个文档的权值,作为句子向量 $sent(f'_{tk1},f'_{tk2},\cdots f'i\cdots,f'_{tkn})$,其中 $1\leqslant i\leqslant n$。最后计算出高频词向量与句子向量之间的相似度,并按照降序排列。两者相似性越大,权值越高,公式如下[8]:

$$Sim(hifre,sent)=\cos\theta=\frac{\sum_{i=1}^{n}f_{tki}\times f'_{tki}}{\sqrt{(\sum_{i=1}^{n}f_{tki}^2)(\sum_{i=1}^{n}f'^2_{tki})}} \quad (2)$$

本文中并没有限制高频词的个数,根据句子的实际情况灵活调整向量,不拘泥于一个限定值,计算更方便、灵活,同时得出的结果也更具有参考价值。

(3) TF * IDF 向量

记作 TF * IDF 权值法,利用特征词的频率及逆文档频率来衡量定义语句。

①TF(Term Frequency)[15]:词频,表示特征词 $t_k$ 在某文档 D 中的出现频率,表征一个词语与某个文档的相关性。一个特征词的 TF 向量记作 $tf(f_{tk1},f_{tk2},\cdots f_{tki}\cdots,f_{tkn})$。

②IDF(Inverse Document Frequency)[15]:逆向文档频率,表示特征词 $t_k$ 在整个文档集 C 中的出现频率,表征了一个词语普遍重要性的度量。IDF 值越大,说明特征词 $t_k$ 对文档 D 的代表性越强,反之,则越弱。一个特征词的 IDF 向量记作 $idf(d_{t1},d_{t2},\cdots d_{ti}\cdots,d_{tn})$。

③计算出 TF 向量与 IDF 向量之间的相似度,并按照降序排列。两者相似性越大,权值越高,公式如下:

$$Sim(tf,idf)=\frac{\sum_{i=1}^{n}f_{tki}\times d_{ti}}{\sqrt{(\sum_{i=1}^{n}f_{tki}^2)(\sum_{i=1}^{n}d_{ti}^2)}} \quad (3)$$

### 3.3 二维相对量融合算法

hirfre * sent 量算法和 TF * IDF 算法从两种不同角度计算得到良好的定义语句,并且每一个候选定义语句权值的取值范围也有所区别。笔者提出了二维相对量的融合算法,从这两种算法的计算结果中得到更好的定义语句。

二维相对量采用了数学中向量基的概念,把 hifre * sent 看作一个平面的向量 hs,将权值最小的向量值作为该平面的基 $e_1$,则句子 i 在此方向上的坐标值为该句子向量的权值与基 $e_1$ 的差值 $h_i$,这样利用 hifre * sent 向量的权值与基 $e_1$ 的差值组成了新的向量 hs $(h_1,h_2,\cdots h_i\cdots,h_n)$;同样地,把 TF * IDF 看作是另一个平面的向量,将权值最小的向量权值作为该平面的基 $e_2$,则句子 i 在此方向上的坐标值为该句子向量的权值与基 $e_2$ 的差值 $d_i$,这样利用 TF * IDF 向量的权值与基 $e_2$ 的差值又组成一个新的向量 $td(d_1,d_2,\cdots d_3\cdots,d_n)$。由于在平面中向量的平移不影响向量的值,可以将这两个平面的基平移到一起,利用它们组成一个坐标系。由于是利用向量值最小的句子作为基,从理论上来讲,在这个坐标系里,距离原点最远的点就是此叙词最好的定义语句。根据平面上的点的距离公式 $\sqrt{x^2+y^2}$,得到二维相对量的计算公式如下:

$$two\_d=\sqrt{\alpha\times hs^2+\beta\times td^2} \quad (4)$$

其中,α 和 β 是调整系数。通过式(4)可以抽取两种算法中较好的定义注释语句。根据实验分析,hifre * sent 向量值的取值范围在 0.8 到 0 之间,而 TF * IDF 向量值的取值范围在 0.5 到 0 之间,α 和 β 取值范围为(0.3,0.7)得到的效果最好。

## 4 中文叙词表本体概念定义注释的自动构建方案及系统实现

### 4.1 自动构建方案

根据 OTCSS[4] 的特点,提出了符合 OTCSS 需要的

定义注释信息自动构建方案，如图 1 所示：

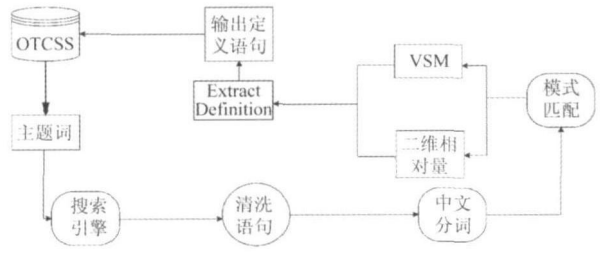


图 1　中文叙词表本体概念定义注释自动构建方案

该方案主要分为两个部分：

（1）数据准备与清理：从 OTCSS 中接收叙词，将接收到的叙词交给成熟的商业搜索引擎去搜索并下载与此叙词有关的语句，剥离下载的语句信息中的网页标记等。

（2）数据筛选与抽取：将上一步骤得到的数据进行分词等处理，利用合适的规则对语句进行筛选，并利用向量空间模型算法和二维相对量融合算法抽取良好的定义语句。

### 4.2　系统实现

本系统有两大功能模块：数据下载部分主要实现叙词的抽取，通过商用搜索引擎来搜索与叙词有关的定义信息；数据处理部分主要实现通过算法来筛选在网络上搜索到的定义语句。系统流程如图 2 所示：

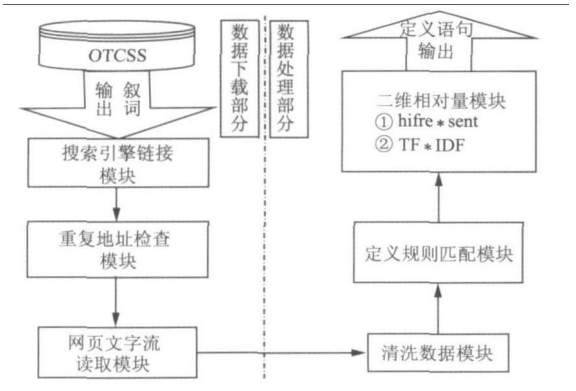


图 2　系统流程图

### 4.3　在 OTCSS 中的应用

将上述方案嵌入到 CCT1_OTCSS 知识提取模块中，系统通过后台运行，将筛选后的结果按照权值从高到低进行排序并存储于数据库中。

OTCSS 系统针对修订专家的各种需求，提供易操作的网络界面，帮助他们选择和修改叙词的定义信息以完善本体。实例如图 3 和图 4 所示：

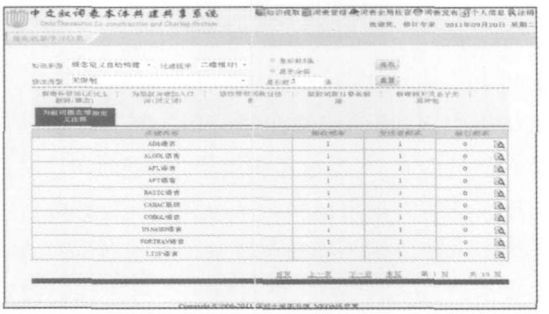

图 3　提取界面

图 4　选择和修改界面

## 5　实验结果及分析

本研究从“中国分类主题词表本体共建共享系统”（CCT1_OTCSS[5]）提供的 OWL 文件中共选取了 650 个叙词作为实验词汇。该语料中的叙词涉及领域宽广，重点选取了法学、地理、宗教三个人文社科领域的叙词和计算机、工业技术、手工业三个自然科学领域的叙词。其中地理领域 120 个，法学领域 100 个，宗教领域 100 个，计算机领域 110 个，工业技术领域 110 个，手工业领域 110 个。

由于是借助商业搜索引擎返回相关网页，所以查全率不是笔者关心的问题，此处只评价查准率。本系统是为完善 OTCSS 的功能而实现，需要给修订专家推荐抽取出的叙词的前三个（或更多）定义注释语句，因此对查准率做了细微的区分：

（1）精确查准率（Accurate Precision Rate，APR）：对所有叙词而言，定义注释语句排在第一位所占的比率。

（2）粗略查准率（Rough Precision Rate，RPR）：对所有叙词而言，定义注释语句排在前三位所占的比率。

两种评价方法的计算公式规定如下：

$$APR = \frac{b}{a} \times 100\% \quad (5)$$

$$RPR = \frac{b+c}{a} \times 100\% \quad (6)$$

其中，a 表示实验语料中所有叙词的数目；b 表示实验结果中，叙词的定义注释语句排在其所有抽出语句第一位的数目；c 表示实验结果中，叙词的定义注释语句排在其所有抽出语句第二位或第三位的数目。

为了考察使用二维相对量融合算法对提高查准率有多少帮助，还引入了一个指标——有效信息提高比（Effective Information Improving Rate，EIR），公式如下：

$$EIR = \frac{New - Old}{Old} \times 100\% \quad (7)$$

其中，New 表示新算法得出的实验结果数目，Old 表示原算法得出的实验结果数目。有效信息提高比考察新算法相对于原算法的有效程度，其比值越大，表明新算法的有效性越高，反之，则表示新算法的有效性越低。

本文中原算法有两种：hifre * sent 向量与 TF * IDF 向量，新算法为二维相对量融合算法。需要分别计算新算法相对于两种原算法的有效信息提高比。二维相对量融合算法相对于 hifre * sent 算法有效信息提高比，记作 H_EIR；二维相对量融合算法相对于 TF * IDF 算法有效信息提高比，记作 T_EIR。公式如下：

$$H\_EIR = \frac{N_d - N_h}{N_h} \times 100\% \quad (8)$$

$$T\_EIR = \frac{N_d - N_t}{N_t} \times 100\% \quad (9)$$

其中，$N_d$ 表示二维相对量融合算法得到的实验结果中，定义注释语句排第一位的数目；$N_h$ 表示 hifre * sent 算法得到的实验结果中，定义注释语句排第一位的数目；$N_t$ 表示 TF * IDF 算法得到的实验结果中，定义注释语句排第一位的数目。

需要强调的是，这里只针对定义语句排名第一的情况计算有效信息提高比，其原因有两点：在统计定义语句排名时比较容易，且最能体现算法的有效性；三种排名顺序全部考虑进去，分析时比较复杂，无法一概而论。

二维相对量融合算法的实验结果，包括定义注释语句的数目分布，以及每个领域的精确查准率与粗略查准率，如表 2 所示：

表 2　叙词的定义注释语句分布及查准率

| 项目 | 人文社科 | | | | 自然科学 | | | | 大综合 |
|---|---|---|---|---|---|---|---|---|---|
| | 地理 | 法学 | 宗教 | 综合 | 计算机 | 工业技术 | 手工业 | 综合 | |
| a | 120 | 100 | 100 | 320 | 110 | 110 | 110 | 330 | 650 |
| b | 46 | 35 | 39 | 120 | 41 | 39 | 30 | 110 | 230 |
| c | 68 | 55 | 42 | 165 | 41 | 30 | 36 | 107 | 272 |
| APR | 38.3% | 35.0% | 39.0% | 37.5% | 37.3% | 35.5% | 27.3% | 33.3% | 35.4% |
| RPR | 95.0% | 90.0% | 81.0% | 89.1% | 74.5% | 62.7% | 60.0% | 65.8% | 77.2% |

为计算有效信息提高比，分别统计得到 hifre * sent 向量、TF * IDF 向量以及二维相对量三种算法所得到的排名第一的定义注释语句的数目。为对比起见，还统计了三种算法得到的排名第一的定义注释语句中，其重合的语句数目及分布情况，并计算出每一种算法在各个领域的有效信息提高比，如图 5 和表 3 所示。部分实验结果如表 4 所示。

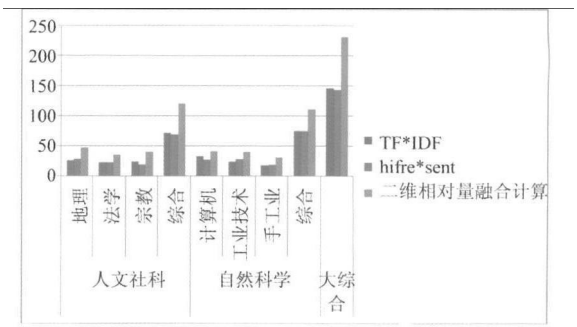


图 5　排名第一的定义注释语句的数目对比

## 6　结　语

本文针对 OTCSS 系统存在的大部分叙词概念缺少定义注释的不足，提出了可以缓解上述问题的概念定义注释自动构建方案，获得了良好的实验效果：对于实验语料，提取前三个句子，人文社科领域的叙词定义查准率可达到 80% 以上，自然科学领域的叙词定义查准率也能达到 60% 以上；且随着提取数量的增加，查准率也会随之提高。

表3 叙词的定义注释语句抽取情况及有效信息提高比

| 项目 | 人文社科 | | | | 自然科学 | | | | 大综合 |
|---|---|---|---|---|---|---|---|---|---|
| | 地理 | 法学 | 宗教 | 综合 | 计算机 | 工业技术 | 手工业 | 综合 | |
| TF * IDF | 26 | 22 | 23 | 71 | 33 | 23 | 18 | 74 | 145 |
| hifre * sent | 28 | 22 | 19 | 69 | 27 | 28 | 19 | 74 | 143 |
| 二维相对量融合算法 | 46 | 35 | 39 | 120 | 41 | 39 | 30 | 110 | 230 |
| 重合语句 | 18 | 13 | 10 | 41 | 19 | 13 | 11 | 43 | 84 |
| T_EIR | 76.92% | 59.09% | 69.57% | .69.01% | 24.24% | 69.57% | 66.67% | 48.65% | 58.62% |
| H_EIR | 64.29% | 59.09% | 105.26% | 73.91% | 51.85% | 39.29% | 57.89% | 48.65% | 60.84% |

表4 部分实验结果

| 叙词领域 | 定义注释语句 |
|---|---|
| 地理 | 阿非利加人(Afrikaner)是在18世纪初和19世纪初到达南非内陆地区第一批移民之前来到开普(Cape)地区的欧洲人(荷兰、法国和德国)后裔。<br>阿斯旺水坝位于埃及开罗以南900公里的尼罗河畔。 |
| 法学 | 法律制裁是指由特定国家机关对违法者依其法律责任而实施的强制性惩罚措施。<br>所谓法学流派,是指对法学领域中某一重大理论或问题持相同或相近的观点而形成的群体。 |
| 宗教 | 八正道是佛弟子修行的八项内容:"正见解、正思想、正语言、正行为、正职业、正精进、正意念、正禅定。"<br>由于加尔文改革了天主教的传统教义,故又称"Reformed Churches",汉译为归正宗,该宗实行长老制,由信徒推选长老与牧师共同管理教会,所以亦称长老宗。 |
| 计算机 | 并串行转换器包含触发器和复用器,并且用于N个并行数据比特并且以N倍时钟速度将它们串行移出到发射机。<br>程序性中断:在程序执行的过程中,发现了程序性质的错误或出现了某些特定状态而产生的中断。 |
| 工业技术 | 真空技术主要包括真空获得、真空测量、真空检漏和真空应用四个方面。<br>我们称最大正应力和最小正应力作用平面为主平面。 |
| 手工业 | 弹性针布针间容易充塞纤维,降低梳理效能,需要定期用抄针工具把充塞的纤维抄去,称为抄针。<br>一般分为吹、打、弹、拉四大类,拨弦乐器也叫弹拨乐器,它是由拨弦振动发音的乐器。 |

将其整合到OTCSS系统后,实践证明,本文提出的解决方案具有实用价值,系统可在后台批量地从网络中搜索和筛选出概念的定义注释信息,修订专家可以灵活使用文中提到的三种算法对候选定义语句进行提取,为需要释义的叙词概念增加定义注释。欢迎登录深圳大学图书馆NKOS研究室网站(http://nkos.lib.szu.edu.cn)进行实际操作[5]。

## 参考文献:


[1] W3C. SKOS Simple Knowledge Organization System Reference: W3C Recommendation [EB/OL]. [2010-02-23]. http://www.w3.org/TR/skos-reference/.

[2] 宋炜,张铭. 语义网简明教程[M]. 北京:高等教育出版社,2004:22.

[3] 曾新红. 中文叙词表本体——叙词表与本体的融合[J]. *现代图书情报技术*,2009(1):34-43.

[4] 曾新红,明仲,蒋颖,等. 中文叙词表本体共建共享系统研究[J]. *情报学报*,2008,27(3):386-394.

[5] 深圳大学图书馆NKOS研究室. 中国分类主题词表本体共建共享系统CCT1_OTCSS CCT1_OTCSS[DB/OL]. [2011-09-23]. http://nkos.lib.szu.edu.cn:8080/ThesaurusProjectForCCTWL/login.jsp.

[6] Riloff E, Jones R. Learning Dictionaries for Information Extraction by Multi-Level Boots trapping[C]. In: *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI-99)*, Florida. AAAI Press / The MIT Press, 1999.

[7] 贾爱平. 科技文献中术语定义的语言模式研究[D]. 北京:北京语言大学,2002.

[8] 张榕,宋柔. 术语定义提取研究[J]. *术语标准化与信息技术*,2006(1):29-32.

[9] Cui H, Kan M Y, Chua T S. Unsupervised Learning of Soft Patterns for Generating Definitions from Online News[C]. In: *Proceedings of the 13th World Wide Web Conference*, NewYork. 2004: 90-99.

[10] Lampouras G, Androutsopoulos I. Finding Short Definitions of Terms on the Web Pages[C]. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. 2009: 1270-1279.

[11] 荀恩东,李晟. 采用术语定义模式和多特征的新术语及定义识别方法[J]. *计算机研究与发展*,2009,46(1):62-69.

[12] 许勇,荀恩东,贾爱平,等. 基于互联网的术语定义获取系统[J]. *中文信息学报*,2004,18(4):37-43.

[13] Joho H, Sanderson M. Retrieving Descriptive Phrases from Large Amounts of Free Text[C]. In: *Proceedings of the 9th International Conference on Information and Knowledge Management*. New York: ACM Press, 2000: 180 - 186.

[14] Klavans J L, Muresan S. Evaluation of DEFINDER: A System to Mine Definitions from Consumer - Oriented Medical Text[C]. In: *Proceedings of the 1st ACM/IEEE Joint Conference on Digital Libraries*. Virginia. ACM Press, 2001: 201 - 202.

[15] 宗成庆. 统计自然语言处理[M]. 北京: 清华大学出版社, 2008.

[16] 程显毅, 朱倩, 王进. 中文信息抽取原理及应用[M]. 北京: 科学出版社, 2010.

[17] 黄萱菁, 夏迎炬, 吴立德. 基于向量空间模型的文本过滤系统[J]. *软件学报*, 2003, 14(3): 435 - 442.

(作者 E - mail: zengxh@szu.edu.cn)

# 《现代图书情报技术》特邀专栏组稿

《现代图书情报技术》是中国科学院主管、中国科学院国家科学图书馆主办的计算机信息管理技术方面的学术性刊物。刊物拥有清晰的定位,即以跟踪技术的研究、应用、交流为主体,服务于广大信息技术人员。

本刊从2004年起开设不定期栏目——《特邀专栏》,每一期专栏集中发表关于某个特定方面的技术研发与应用的研究型文章,汇集科研成果、聚焦研究前沿。

## 1 《特邀专栏》操作办法及流程

(1)本栏目特邀国内外知名专家、学者、教授担任专栏主编,专栏的设立一般由期刊的策划编辑和特邀专栏主编沟通,根据国内外图书情报技术学科的发展需要提出选题。

(2)选题一旦确定后,由特邀专栏主编承担稿件的组织,审核并撰写前言。一期特邀专栏一般为4 - 6篇文章为宜。稿件组织过程中,策划编辑将与特邀专栏主编进行定期的沟通,及时掌握稿件的撰写情况,并对稿件的撰写提出适当的建议和意见。

(3)稿件经特邀专栏主编审核通过,提交给编辑部。后期由策划编辑负责与作者的联系沟通及安排出版等事宜。

(4)专栏的选题一旦确定后,将确定基本时间表。一般的操作周期为3 - 5个月。以正式确定特邀专栏题目为起始点,在1个月内确定约请论文的作者和题目,3个月内确定初稿,5个月内确定采用稿。

## 2 《特邀专栏》稿件内容要求

(1)深入反映本专栏选题方向的前沿研究成果或重大应用成果,侧重理论研究、技术分析、系统论证或设计等,注意理论与实践相结合。

(2)特邀专栏稿件应该主要是原始性和原创性研究论文,也可以有一篇综述性论文,但综述性论文必须可靠地覆盖该方向的原始核心文献。

(3)文章按照严谨的学术文章体例写作,即明确扼要地界定研究问题,简要说明研究方法,系统精炼地描述国际国内发展状况,进而详细地描述作者自身研究工作的技术线路及研究结果。

(4)特邀专栏的一系列文章应注意覆盖专栏选题所涉及的各个研究方向和多个研究单位,充分覆盖可能存在的多种观点和技术线路。

(5)充分承认前人/别人的工作,充分引证所参考引用的文献(尤其是本研究工作中的原始核心文献和国内最先出现的研究文献),严格遵守著录规范。

## 3 《特邀专栏》稿件格式要求

(1)论文版式请参照本刊网站"下载专区"中"论文模板"。

(2)多个作者时,请注明通信作者,并注明各个作者的单位。

(3)每篇稿件以6 - 8千字为宜(按篇幅字数计算,包括图、表)。