主题词表及其语义环境的计算机构建与应用研究[1)]

田 梅 王兰成 侯 双

(南京政治学院上海分院信息管理系,上海 200433)

摘要 电子叙词表建构及其计算机管理是提高信息检索质量的重要保证,本文介绍主题词表及其辅表的语义关系,探讨主题词表、范畴索引和词族索引表的计算机构建与设计,文章分别给出了实现的数据结构与算法。

关键词 计算机应用 主题词表 语义环境 算法

Applied Research of Semantic Electronic Thematic Words

Tian Mei, Wang Lancheng and Hou Shuang

(*Shanghai Political College Information Center, Shanghai* 200433)

Abstract This paper describes a method of the applied research based on semantic electronic thematic words. The method organizes the thesaurus with supplement words. The thematic words as well as category and hierarchy indexing have been realized and designed by computer. The technical issues concerned have been discussed and applied within the computer processing system.

Keywords computer application, thematic words, semantic environment, algorithm.

1 引 言

主题词表是一种主题检索系统所用的检索词的词汇表,设有参照系统和各种索引,以显示词间语义关系和提供各种查词途径。长期以来,该表作为一种将文献、标引人员和信息用户的自然语言转换成规范化语言的术语控制工具,在文献标引和信息检索等方面有着广泛的应用。

近几年来,随着计算机网络技术迅速发展,广大用户对计算机网上资料信息检索的质量提出更高要求,而其关键因素是信息的自动处理和控制能力[1]。电子叙词表的建构及其计算机管理已成为提高领域信息检索质量的重要保证,对主题词表的语义环境及其检索的实现研究也成为当务之急。本文将介绍主题词表及其辅表的语义关系,研究主题词表、范畴索引和词族索引表的计算机构建与设计,文中给出了相关的实现算法。

2 主题词表的语义关系及其应用需求

叙词是经过规范化处理的、能显示叙词之间语义关系的、有标引和检索意义的词或词组。所有有序化叙词之和就构成了叙词表。叙词表中的语义关系包括等同关系、属分关系、相关关系。词的等同关系是指一组词或词组在概念上完全相同或意义接近,属分关系指概念内涵相同、外延范围大小不同的词之间的关系,族首词则是一种特殊的属分关系,其

收稿日期:2002 年 11 月 28 日

作者简介:田梅,女,1978 年生,硕士研究生,主要研究方向为计算机信息管理。王兰成,男,1962 年生,在职博士生,现为解放军南京政治学院信息资源管理教研室主任,教授,硕士研究生导师,研究领域为数据库与信息处理。侯双,男,1976 年生,硕士研究生,主要研究方向为数字图书馆。

1) 本研究得到南京政治学院"十五"科研项目基金的部分资助

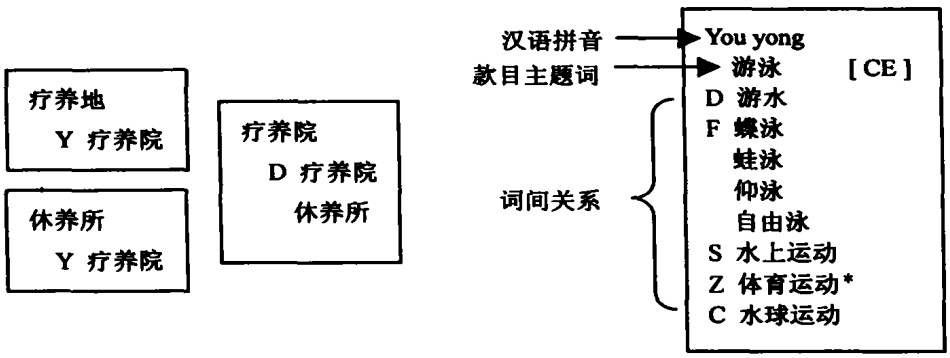

图 1　主题词词款中各参照项构成的语义网络示例

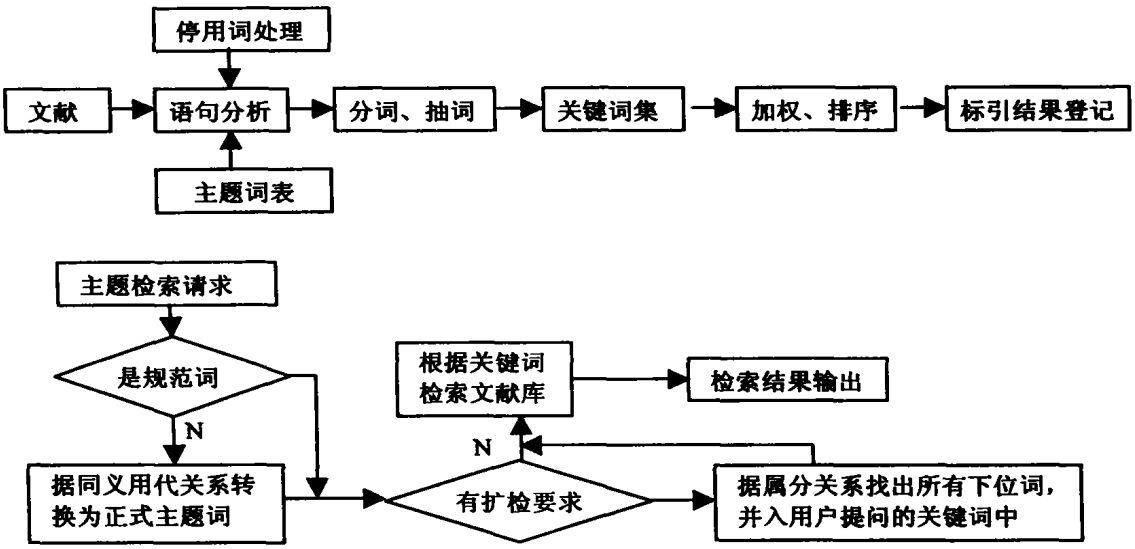

图 2　文献自动标引和检索的流程

外延范围最大，相关关系是指概念内涵之间语义联系的词间关系[2]。

语义关系在叙词表中的表现，以图 1《中国档案主题词表》的款目为例，词款目中各参照项构成了一个词的语义网络。其中 D(代)与款目词是等同关系，这些词都是非叙词，由款目词代替；F(分)是指款目词的下位词，S(属)指款目词的上位词，C(参)指款目词的相关词，Y(用)表示的词都是正式叙词，Z(族)表示的词称为族首词，也就是最上位词。

一条词的款目中，带有“Y”项的为非正式叙词，相反带有“D”项为正式叙词。叙词与“Y”项或“D”项之间是等同关系，Y、D 之间存在着转化关系，如“疗养地”、“休养所”、“疗养院”条目可以表现这种转化关系。“游泳”词条中，“游泳”是正式叙词，其与 D 项“游水”是等同关系，“游水”是非正式主题词，“游泳”与 F 项、S 项与“游泳”、Z 项与“游泳”是属分关系，“游泳”与 C 项是相关关系。

3　主题词表及语义环境的设计与实现

3.1　自动标引与检索对主题词表设计的需求

文献自动标引和检索的流程如图 2 所示。

自动标引通过停用词表的处理，来剔除无用词，通过主题词表实现标引词的词表控制和概念转换。主题标引停用词表应包含中文标点符号、普通汉语的停用字、词以及档案领域的停用字、词。主题词表，应提供多入口，以及丰富的语义关系。并且提供等同、属分、相关、范畴语义关系的控制。此外标引结果若含有主题词表中未收录的词应该在临时表中进行登记，当该词的引用超过一定取值时将其在主题词扩展表中进行登记[3]。

文献检索系统利用主题词表的用代关系规范用

户提问的主题词、用属分关系实现扩检，因此主题词表在设计时应该对 Y、D、F 项数据访问提供方便入口。

3.2 主题词语义环境的数据结构

根据上面的需求分析，经过概念分析、逻辑设计组织数据库结构，确定叙词表、语义关系表和索引的设计。首先由计算机对纸质词表进行扫描和加工，形成一个初表。然后根据数据量对数据库的各表结构进行设计。其中，叙词表含有叙词项、正式叙词标记、连接语义关系表的指针。语义关系表，包括 S、F、Y、D、C 各项的内容和指针。索引表包括范畴索引和词族索引，编制程序由前面两个表生成。

(1)停用词表(stoplist)

停用词表应包含中文标点符号、普通汉语的停用字、词以及档案领域的停用字、词。该词表主要用于控制自动标引的切分词操作，是独立于《中档表》、实现计算机词表控制特有的数据表，需要根据实际情况不断补充。该表数据库结构如表 1 所示：

表 1

字段名	id	unuse
数据类型	N	C
含义	位置标识	停用词

(2)叙词表(main)

所有《中档表》正式叙词和非正式叙词，都收录在该词表中(见表 2)。

如表 2 所示该表提供主题词、拼音及主题词出现频率的统计，提供语义关系表入口及范畴索引表入口，此外对于非正式叙词提供了与其同义的正式叙词在该表的指针。tf 字段提供的频率统计有助于标引和检索权值计算和选词。对于一般要求的检索和标引，本表即可实现控制功能。

(3)叙词关系表(relation)

当检索中要求扩检，以及标引要求较高时要用到用、代、参等语义关系，因此还要设计叙词关系表，通过叙词表中提供入口来实现访问。该表提供了主题词、范畴索引入口、词族索引入口、上位词、参项(见表 3)。其中 hnextclass、hclass 字段是记录词族索引的辅助字段，将在词族索引表中给出说明。

3.3 词表算法的设计与实现

停用词表一般由标引人员采用录入的方式实现。Main 和 relation 表则通过算法由《中档表》同时产生。dictionary.recnum、main.ID、relation.rid 分别为各表位置标识，均与记录顺序保持一致，因此分别在三个表中相同位置的记录对应 recnum、ID、rid 值相等。

依据这个原则，为算法设计带来方便。算法参见相关论文。

表 2

字段	ID	tw	pinyin	tf	rid	y	cn
类型	N	C	C	F	N	N	C
含义	主题词在该表中的地址标识	主题词	拼音	该词出现频率	该词在语义关系表中的地址	为 0 表示该词为正式叙词，为大于 0 的数，表示正式叙词在该表中位置	该叙词范畴号

表 3

字段	rid	tw	cid	locs	loc	locf	hid	hnextclass	hclass
类型	N	C	N	N	C	C	N	C	N
含义	主题词该表中的地址标识	主题词	该词范畴索引中的位置	记录该词上位类在本表中的地址	该词参考项的位置	记录该词下位类多个词在本表中的地址	该词在词族索引中的位置	记录该词下位词的下位词在该表中的地址	该词在词族索引表中的级别

4　范畴索引与词族索引表的实现算法

4.1　范畴索引表的设计与实现

主表中的全部主题词按类目分类排列，提供按类查词的途径，这就形成了范畴索引[4]。《中档表》category 存放范畴表的所有范畴号及其所代表的类别，以及范畴号对应的词在叙词表的位置(见表 4)。

Category 的生成算法：

1　Begin

2　建立 category 的表结构，打开 dictionary 和 relation 表，均定位在第一条记录的位置

3　do while dictionary 表未结束

4　在 category 中从第一条记录开始扫描到表尾

5　category 中有 category.cn = = dictionary.index

6　dictiongary.recnum 写入 category.id，用“#”分隔

7　category 中没有 category.cn = = dictionary.index

8　category 中增加一条记录，将 dictionary.index 赋给 category.cn，将 dictiongary.recnum 赋给 category.id

9　赋 category.id 值为其记录号(由 recn()获得)

10　将 category.cid 赋给 relation.cid，转到(4)

11　(category 扫描结束)dictionary 和 relation 指针均向下移动一位，转到(3)

12　(循环结束)

13　关闭各表

14　End

表 4

字段名	cid	cn	mean	id
数据类型	N	C	C	N
含义	范畴号在该表中位置标识	范畴号	表示范畴号对应的类别	在 main 中的位置，以“#”为间隔

表 5

字段名	HID	hh	Ridclass2	next
数据类型	N	C	N	N
含义	词族族首词在该表中位置标误	词族族首词	二级词在叙词关系表中的位置	具有下级词的 relation 表中的记录指针

category 建立好了以后，按 cn 进行升序排列，这样就形成了按范畴号字母顺序的范畴表。

4.2　词族索引表的设计与实现

所谓词族是把属性相同的主题词按其概念等级阶梯式的排列而成的概念体系。词族索引是把主表中具有属种关系、包含关系和整体部分关系的正式主题词，按规定属分级别展开全显示的一种词族系统。Hearchy 表在构造中采用指针访问、relation 表的方法，再在 relation 表中通过 hnextclass 字段记录下一级访问指针，通过 hclass 字段记录该记录主题词的级别。实际上形成了一颗先序遍历的树。这样设计灵活，不仅实现词族索引的层次结构，而且访问入口多，在 relation 表中可以直接获取族首词的指针信息(见表 5)。

Hierarchy 的生成算法：

1　Begin

2　建立 hierarchy 的表结构

3　打开 relation 表，定位到第一条记录

4　do while 表未结束

5　tw 字段以“*”结尾 在表中增加一字段，写入 hierarchy.hh 字段中

6　relation.locf 写入 hierarchy.locf 中

7　根据 hierarhy.locf 的指针在 relation 表中定位
　　定位的记录 locf 字段不为空(仍有下位类)共有 n 条

8　定位的记录 locf 字段全为空转到(18)

9　将该 n 条记录指针记在 hierarchy.next 中

10　I = 2

11　do while 未处理完这 n 条记录

12　处理 hierarchy.next 中记录的下一条记录，定位到 relation 表中，将 I 赋给 relation.hclass

13　对 relation.locf 字段所指本表中位置进行访问

14　访问到的记录 locf 字段为空转到(13)

15　若访问到的记录 locf 字段不为空，赋 I + 1 给 hclass，并将该记录登记到访问前的记录的 hnextclass 字段中，复建立指针和向下遍历的操作，直到所访问的记录 locf 字段为空，转到(13)

16　(n 条记录处理完毕)relation 表向下移动一条记录，转到(4)

17　(relation 表结束)关闭所有表

18　End

综上，我们建立了叙词表、叙词关系表、停用词表以及范畴索引、词族索引，这将极大的方便计算机以词表控制的自动处理技术。当然根据具体应用不同需求，词表结构还要做一些细微调整，本文所讲的构造和生成词表的方法仍然适用。

参　考　文　献

1　Lassila O，Swick R.R. Resource Description Framework (RDF)

Model and Syntax Specification. World Wide Web Consortium Recommendation. 1999. http://www.w3.org/TR/REC-rdf-syntax/

2 张进等.计算机信息检索软件设计原理.武汉大学出版社,1994

3 王兰成等.基于中国档案主题词表的自动标引控制研究.情报学报,2002,21(2):177~180

4 W.C.Mann and S.A.Thompson. Rhetorical Structure Theory: A Framework for the Analysis Texts. USC/Information Science Institute Research Report RR-87-190(1987)

(责任编辑 芮国章)