

基于 HNC 语义分析的中文标题分类方法

陈磊

中国科学院声学所 声场声信息国家重点实验室 II 部 北京 100080

摘要: 文章的主题与其标题句具有强相关关系, 标题句各语义角色可对文献分类提供参考信息, HNC 理论是语义分析的有力工具。由此出发, 本文提出基于 HNC 语义分析的分类算法。即对标题句先进行句类分析, 求得各语义角色, 由角色的概念类别得到文献的分类。

关键词: 文献分类 语义分析 HNC 理论 自然语言理解

A Chinese Document Title's Classification Method Based on HNC Semantic Analysis

Chen lei

State Key Lab Section II, Acoustics Institute of Chinese Academy of Sciences

P. O. Box 2712, Beijing 100080, P.R.C

Abstract: The title of a document is closely related to its topic, semantic cases of the title can give some information to document classification, HNC theory can provide a powerful means to do semantic parsing. Hence, a Chinese Document Classification Method Based on HNC is on light here. Firstly, it does Sentence-Category parsing on the title, then we can get all cases of the title, at last the document classification can be extracted from Concept-Category of cases.

Keywords: DOCUMENT CLASSIFICATION, SEMANTIC ANALYSIS, HNC, NLP

一 引言

文献分类就是对文献集进行类别划分。该技术随着 Internet 的飞速发展, 显得愈发重要。自从 1991 年诞生以来, WWW 已经发展成为拥有约 1 亿用户和近千万个站点、600G 信息容量的巨大分布式信息空间, 并且这个数字仍以每 4 至 6 个月翻一番的速度增加。文献分类一般通过统计方法或知识工程方法来实现, 由于统计方法相对简单的机制, 目前为大多数实用文献分类系统所使用。[5]近年来, 将汉语分析技术与统计相结合的一些新方法相继提出, 如[3]。

本文提出一种利用语义分析技术的汉语文献标题分类方法。它首先利用 HNC 句类分析系统对文献标题进行语义分析, 得出标题的主语义块分布(包括 A、B、C 和 E 块), 进而核查各语义块的概念类别, 得到该标题文献的类别。

标题同文献主题具有密切关系。特别是在新闻类文献中, 作者在标题中会极大地浓缩

出文献的人物、事件、时间、空间等等。据统计，500 篇人民日报和 200 篇港、台、新（新加坡）的新闻稿测试样本中，标题与文章主题的基本符合率达到了 91%。[3]可见，依据文献标题对其进行分类分析是可行的，同时还将节约计算时间和数据存储空间。

二 文献标题分类模型

目前主要针对新闻稿进行分类，我们将分类规定为：对象+类别，如：{北约，南联盟} + 军事，中国 + 经济等等。

分类模型定义为一四元组：

$$SC = \{T, Co, Cc, Sca\}$$

T：系统所处理的文献标题集

Co：对象的概念类别集

Cc：类别的概念类别集

Sca: $T \rightarrow Co \times Cc$ 标题分类函数，其中的 Co, Cc 为标题的“对象内容”分解，通过 HNC 句类分析得到。

其中，Co 多为表示国家、地域、组织、机构、个人等的概念，如：pj01(国家)、wj01(地区)、pe(组织)、p(人)等。

Cc 为提供语境信息的概念类别，重要有八类。

语境编号	HNC 概念类别	语境类别
1	A	专业活动:政治、经济、军事、法律、科技、教育、卫生
2	Y6 y:6, 9, c	劳作与服务
3	y7 y:6, 9, c	交往与娱乐
4	y8 y:6, 9, c	记忆与想象
5	D	社会规约
6	6m m:0—5	人类本能及生理活动
7	7	心理活动及精神状态
8	8	思维活动

目前 Cc 重要针对 1 号语境进行类别划分。a 概念节点表请参见[1](89—91 页)。标题句句类分析文献分类算法是关于语义的计算方法，是对标题更深入的分析。例如：[3]中提到的“地域+类别”形式的分类模型，对于以下语料：

南联盟石化工厂遭到北约轰炸。

依据其标题分析算法 g_T ，结果应为{ 欧洲 + 经济 }，其中，V1={南联盟}，V0={石化、工厂}，可得到：{石化、工厂}→经济；{南联盟}→欧洲。由此得到该文献的分类为{欧洲 + 经济}，但该文献的政治和军事性应更为突出。

三 标题句句类分布

HNC 理论定义了 7 种基本句类, 分别命名为: 作用句, 效应句, 转移句, 关系句, 过程句, 状态句, 判断句。相应的符号为: X, Y, T, R, P, S, D。([1] 论文 2)

每种基本句类都有自己的“个性”, 例如: 状态句是所有基本句类中唯一可以没有 E 语义块的句类。在进行句类分析时, 应该利用这些宝贵信息。为了设计强健高效的标题句分类算法, 我们对标题句的基本句类分布作了统计, 基本上掌握了实际文本中的标题句句类分布的特点, 同时也对基于句类的分析算法进行了原理验证。

标题句来源: 《人民日报》电子版的国际新闻稿标题。对其进行人工句类分析, 标注句类代码, 获得各句类分布为:

	基本句类	复合句类	混合句类
分布 (%)	76.6	18.7	4.7

基本句类中子类分布:

	X	Y	T	R	S	P	D
分布 (%)	42.6	8.5	21.8	7.5	3.6	7.5	8.5

四 句类分析分类算法

4.1 Sc 算法:

1. 利用 HNC 句类分析系统, 对文献标题进行句类分析 [2], 确定特征语义块 E 和广义对象语义块 JK 及其句类 J。
2. 核查特征语义块的语境信息 InfoOfContext (E)。
3. 核查广义对象语义块 JK 的语境信息 InfoOfContext (JK); 核查广义对象语义块 JK 的对象信息 InfoOfObject (JK);
4. 由 J、InfoOfContext (E)、InfoOfContext (JK) 生成文献类别 (Category)
5. 由 J、InfoOfContext (E)、InfoOfObject (JK) 生成文献对象 (Object)

4.2 示例:

1 由特征语义块 E 决定文献类别

特征语义块 E 与文献类别存在着密切的关联信息, 这是因为“一个句子的基本语义信息就蕴含在 E 块中” “所谓一个句子的基本信息就是指它所表达的关于作用效应链的某一或某些环节的信息。” [1] E 块决定句类信息, 决定了广义对象语义块的值, E 块同时也表

达了明确的语境信息。例如：“轰炸、袭击、空袭”等等概念表明了文献类别与军事有关，而“倒闭、盈利、兼并”等涉及经济类文献。从 E 块入手，不但可预期广义对象语义块，确定对象的取值，进而引发语境信息，而且可直接映射语境信息。由语境信息进行文献分类。

例如：北约 将 继续 轰炸 南联盟。

A QE EQ E B (式 4-1)

上述主题句经过 HNC 句类分析系统后，将得到如式 4-1 的物理表示式。其中，A 为作用者；QE 为 E 语义块的“上装”，对 E 块起修饰和限制的作用；EQ+E 为 E 块的构成，EQ 表示 E 块前修饰，E 表示特征语义块主体；B 为对象语义块。该句中 E 块“轰炸”的 HNC 语义标注为：

HNC (概念层次网络符号)：(va42, v352) (式 4-2)

上述知识项中，HNC 符号表示“轰炸”兼具有 a42 行和 353 行的 v(动词)概念，其中，a42 可展开为 a-4-2，表示 a 行(专业活动)下的 4 分行(军事)下的 2 分行概念(战争)。352 可展开为 3-5-2，表示基元概念中的 3 行(效应)下的 5 分行(立与破)下的 2 行(破)的概念。经过此标注，“轰炸”该概念的语义属性为：效应-立与破-破，语境信息表示为：专业活动-军事-战争，InfoOfContext (E) = a4 (军事)。由此，文献的类别可确定为：军事-战争。该句为 X (作用句)，作用者 A 与作用对象 B 为“轰炸”行为的双方，一个主动，一个被动，文献的对象为 {北约，南联盟}。该标题所对应的文献分类为 {北约，南联盟} + 军事。

2 由对象 JK 确定文献标题类别

在 E 不提供语境知识的情况下，可利用广义对象的语境知识确定文献类别。特定的对象必然与相应的语境有关，例如：国会，政府，警察等均强关联于政治，而贸易、顺差、逆差等强相关于经济。特别是在 Y (效应句)中，Y 多为效应类基元概念，缺乏语境信息，但效应对象 YB 和效应内容 YC 带有充分的信息。

例如：中国 冶金 工业 形势|好转。 (式 4-3)

YB Y

“好转”无明确的语境知识，InfoOfContext (E) = NULL，YB 中“冶金”HNC 符号为：gva21，“工业”为：ga21，所以 YB 带有明确的 a2 (经济)语境知识，InfoOfContext (JK) = a2，结合 Y 句类特点，可确定文献类别为经济，对象可从 YB 中发现 {p、pe、pj2} 等概念取得。最后的结果为 中国 + 经济。

3 广义对象语义块与特征语义块共同决定

在 InfoOfContext (E)、InfoOfContext (JK) 均非空的情况下，二者协调可更可靠的决定类别。

例如：法国 汽车工业 | 上半年 | 出现 盈利。(式 4-4)

YB fK YQ Y

盈利：(vva2, vv3a1)，InfoOfContext (E) = a2，“工业”：ga21，InfoOfContext (JK) = a2，InfoOfObject (JK) = 法国，所以，文献类别为：法国 + 经济

4 对象+类别模型的功能提高

● 解决对象多重分类问题

依据对象而不是简单的地域属性，可以更准确地表示事件双方的关系，同时也可支持检索方的智能信息查询。例如：

美国 准备 对南联盟 发动地面进攻。(式 4-6)

A QE B XQ X

由 E 的主体“进攻”可确定其语境为 a4 (军事), 文献主题类别为军事。根据作用句的句类信息, 确定“进攻”动作的发出者为“美国”, 动作的对象为“南联盟”。文献主题对象为: {美国, 南联盟}。

当用户检索“美南关系”, “美国海外军事”等时, 上述文献可快速地确定入选。

● 解决类别交叉分类问题

真实文献的类别不一定是单一的, 比如军事和政治、文教和科技等就经常出现于同一文献内, 因此在文献标题中也可能会出现类别交叉的现象。本文通过在标题内挖掘多重语境信息的方法解决该问题。

例如: 日本中小學生 出现 道德真空。(式 4-7)

YB Y YC

该句为效应句, 其语境信息由 YB、YC 提供。“学生”: pa72, “道德”: gd01, 分属于 1 号和 5 号语境。所以其文献类别为 {教育 (a7) + 社会规范 (d01)}。

五 结论

根据语义对文献进行分类可充分地利用文献标题内的信息, 从而可更加快速准确地得到结果。利用 HNC 理论和其句类分析系统, 我们提出了以上的原型系统, 在理论上展示了所解决问题的广度与深度。

当然, 该模型还有不少待提高之处, 如尚未充分考虑文献标题句的语言学特点; 尚未将 E 块的语义引入分类模型; 尚未考虑对象之间的复杂关系。在下一段的工作中将继续开展相关研究, 并设计基于语义的文献全文分析分类方法。

参考文献

- [1] 黄曾阳 《HNC (概念层次网络) 理论》清华大学出版社, 1998
- [2] 晋耀红 基于 HNC 理论的句类分析系统的设计与实现 中国科学院声学所硕士学位论文, 1998
- [3] 战学刚 姚天顺, 基于汉语分析的中文分类方法 《1998 中文信息处理国际会议论文集》清华大学出版社, 1998
- [4] 麻志毅, 姚天顺, 基于情境的文本理解, 计算机科学, 1998. 3
- [5] 吴立德等, 大规模中文文本处理, 上海: 复旦大学出版社, 1997