●耿 骞 汤艳莉

面向自然语言检索的短语结构索引

摘 要 针对自然语言提问的特点,提出基于短语索引的用户提问的处理方法,给出了短语结构索引的生成方法,设计了提问处理流程。在此方法中,系统接收完整的句子作为提问,采用自然语言处理技术对提问逐步处理,从提问中抽取短语作为检索对象。与关键词相比,短语可以表达更为具体的概念,有助于提高系统的查准率。图 1. 表 1。参考文献 13。

关键词 自然语言检索 短语结构索引 自然语言处理技术 情报检索 索引
分类号 G354

ABSTRACT According to the characteristics of queries in natural language, the authors propose some methods for the processing of user's queries which are based on phrase indexes. Compared with keywords, phrases can represent more concrete concept, and are helpful for the improvement of system efficiency. 1 fig. 1 tab. 13 refs.

KEY WORDS Natural language retrieval. Phrase structural index. Natural language processing technology. Information retrieval. Index.
CLASS NUMBER G354

自然语言检索是信息检索的一种类型。对自然语言检索的研究至今已有 30 多年的历史,但研究进展相对缓慢,其原因就是自然语言检索中的一个最为关键的问题——自然语言处理一直未能得到很好解决。为此,本文提出基于短语索引的用户提问的处理方法,并对支持自然语言检索的索引问题进行探讨。

1 支持自然语言检索的索引及其形式

索引的结构和组织形式服务于检索控制的需求。不同索引的索引能力不同,其所占用的空间和维护时间也不同。对自然语言检索来说,索引形式的确定受信息源的类型和用户提问形式的影响。虽然传统的索引,如基于关键词的全文索引也可以在一定程度上支持自然语言检索,但全文索引丢失了索引词之间的关系,所以依赖全文索引进行检索不是一种最好的方法。从理论上说,能较好完成自然语言检索任务的索引形式应该是语义索引,或功能相当的索引形式。

语义索引是在对文本内容语义特征提取的基础

参考文献

1 Berners-Lee, T., Hendler, J., and Lassila, O. The Semantic Web. Scientific American, 284(5): 34 - 43, May 2001.

2 刘柏嵩.基于知识的语义网:概念、技术及挑战.中国图书馆学报,2003(2).

3 Asuncion Gomez-Perez et al.. Ontology Languages for the Semantic Web. IEEE Intelligent Systems, Jan/Feb 2002: 54 - 60.

4 Michel Klein et al.. Ontology versioning and change detection on the web. In EKAW2002, Spain, October 1 - 4, 2002.

5 Jeff Z. Pan and Ian Horrocks. Metamodeling architecture of web ontology languages. In Proceedings of the Semantic Web Working Symposium, pages 131 - 149, Stanford, July 2001.

6,8 P. F. Patel-Schneider and D. Fensel: Layering the Semantic Web: Problems and Directions, In Proceedings of the 1st International Semantic Web Conference, Lecture Notes in Computer Science (LNCS 2342), Feb 2002.

7 D. Fensel. Language standardization for the Semantic Web: The long way from OIL to OWL. In Proceedings of the Fourth International Conference on Distributed Communities on the Web, Sydney, Australia, April 2002.

刘柏嵩 副研究馆员,博士生。通讯地址:浙江宁波大学 41#信箱。邮编 315211。

(来稿时间:2003-07-28)

上生成的一类索引，其目的是希望在索引中反映出文本标引词之间的内在联系，从而在自动标引过程中过滤文本存在的语言歧义，提高检索的检准率，实现概念级的检索。就目前的研究来看，语义索引主要有两种类型：通过数学方法计算的潜语义索引和通过自然语言理解语义分析得到的索引。

尽管在语义层次建立索引的思想很早就被人们所提出，但由于自然语言理解中语义分析本身所存在的困难，加之人们对在情报检索中如何使用这种索引缺乏深入的研究，所以，到目前为止，尚未见到有使用该种索引的检索系统或相关的检索研究项目。

语义索引虽然能较好地反映文本的语义特征，但目前在计算机语义分析上还有许多需要解决的问题，因而该方法的实施还存在着很多困难。一种可行的方法是采取折中的方案，使用一种基本能反映文本信息索引词之间关系且创建代价小的索引来标识文本。这里我们用一种称为短语索引的标识方法。

2 短语结构索引

2.1 短语结构索引的形式

短语索引是一种建立在短语结构基础上的索引。短语索引在逻辑形式上与句法分析后的得到的短语片断有很大的联系。同时，在一定的情况下，当所要索引的信息文本长度很短时，如仅仅表示一个事实，也可以认为短语索引就是经过句法分析后加工而成的。

在自然语言中，短语是比词更大的意义单位，通常包括两个或两个以上的词语。按照不同的标准，短语分成不同的类型。依据短语的功能，可分为：名词短语、动词短语、形容词短语、介词短语、时间短语等；依据短语内部的句法关系，可分为：主谓短语、动宾短语、主补短语、动补短语和偏正短语等。短语作为完整的语义单元，把它当作信息检索中索引的对象，具有明显的优势：与关键词索引相比，短语保存了更多的原文信息，能准确描述查找的概念，也易于表示复杂概念，从而为提高检准率提供手段；与语义索引相比，短语索引的生成过程简单，维护费用较低。自然语言检索中短语索引研究的重点问题是如何从文本中提取短语以及如何表示短语索引，具体包括短语的表示、权重赋值和匹配三个问题。

2.2 短语提取的方法

从文本中提取短语的方法多种多样，复杂程度不同，但不论采用何种方法，它们都具有以下共同点：选取一个或多个可靠而有效的因素作为提取短语的标准；在抽取算法中使用这些因素，决定哪些词能够组成短语，以何种方式组合。需要考虑与词语相关的因素，包括：词语在文本中的出现频次，词语相似度，多词之间的句法结构，汉语词语中字的个数，语义特征等。采用这些因素的短语提取算法，分为统计方法和句法分析的方法两类。

2.2.1 基于统计的方法

统计方法利用从训练语料中获得的相邻词对出现的频率、词语依存信息、概率和其他因素来抽取可以作为短语的复合词，这些复合词包括：一组单字、固定搭配、N元词串或句法结构。用这种方法得到的短语称为“统计短语”，它们有些不是真正意义上的短语，因而不具有正确的语法结构或语义结构。用统计方法提取短语的基本过程为建立短语词典、对文本和查询分别运用短语构造程序识别出其中的候选短语、为文本和查询赋予短语描述符和为文本和查询赋予单个词语描述符四个步骤。

总的说来，统计方法较为简单，而且在一定范围内，可以较好地解决边界歧义的问题。统计的方法从简单的频率统计到复杂的参数学习不断发展。由于训练语料通常是在语言学理论的指导下进行的，这些使用统计信息的短语抽取方法实际上也运用了语言学知识。

2.2.2 基于句法分析的方法

使用句法分析的方法，首先要知道短语的内部结构以及短语成分的语法功能，以便发现构造短语的规则或模式。短语是句子的构成成分，而句法分析就是分析句子的语法结构，故而从语言学角度出发的抽取短语的方法就是在句法分析的基础上寻找已经定义好的句法短语。

由Dillon和Gray(1983)开发的FASIT系统就是典型的基于规则的短语识别系统。系统实现定义了一组名词短语模板，输入文本中相邻的词和模板去匹配，至少和其中一个模板匹配上的就认为是名词短语。虽然这种方法能识别出正确的短语，但结果中更多的是一些非短语的词串。所以仅使用规则而不考虑上下文信息是不够的，于是研究者开始转向局部句法分析。有时，这种浅层分析可以识别出相当数量的短语作为标引项。Xerox Hull和Grefenstette(1997)使用“浅层分析器”，从文中发现短语，这些短语都满足一定的关系，如主谓关系、动宾关系、动补关系、偏正关系等。实验表明，句法短语的应用

使短语提取的效率提高了15%。

一般来说,用句法分析的方法获得的短语,结构性和完备性较好,而且符合人们的理解方式,但问题在于详尽的句法信息限制了短语作为索引对文本(或用户提问)的表达能力,且支持该方法的语言知识的获得也是十分困难的事情。

2.3 短语索引的表示方法

信息检索系统中,短语的表示方法决定了短语是否能成为高效的标引单元,也就是说,短语如何反映文本内容、如何表示用户提问,才能使查询和文档的匹配过程更加有效。一般来说,短语的表示方式分为三种:直接表示,结构表示和规范化表示。

2.3.1 直接表示法

直接表示法是把抽取的短语直接作为索引项,而不经过任何形式上的转化,内部的结构信息也不保存。这种方法主要用于通过统计得到的短语,由于用统计方法抽取短语时都会预先设定短语的固定长度,则直接表示法就限制了系统处理长度超过固定长度的短语的能力。

2.3.2 结构表示法

结构表示法详细记录了短语结构信息,这些短语通常是对文本或提问进行句法分析之后提取得到的。它以分析树的形式保留了具体的句法信息,并可以通过复杂的表示解决句法上或语义上的模糊性。

汉语中大多数短语结构都可用二叉树的形式表示。其中的标记可以是词性,也可以是短语类型。根据短语的这种特点,北大计算语言所的专家们定义了短语以及短语构成成分的结构,其描述形式如下:

$$[f_1 = v_1$$
$$f_2 = v_2$$
$$\cdots\cdots$$
$$f_n = v_n] \quad n \geqslant 1$$

其中,各特征项 f_i 为原子,特征值 v_i 为原子或复杂特征集。当描述非终结符(即短语)时,各特征项及特征值分别为:

f_1 = 条目,v_1 = 当前短语

f_2 = 句法范畴,v_2 = 当前项所属的句法范畴

f_3 = 左结点,v_3 = 当前项的左结点的复杂特征集

f_4 = 右结点,v_4 = 当前项的右结点的复杂特征集

当描述终结符(即词)时,各特征项及特征值分别为:

f_1 = 条目,v_1 = 当前词条

f_2 = 句法范畴,v_2 = 当前项所属的句法范畴

f_3 = 音节数,v_3 = 当前项的音节数目

f_4 = 外内,v_4 = 当前项为不及物动词时,填"外";当前项为及物动词时,填"内"

f_5 = 义类,v_5 = 当前项所属的语义类代码

结构表示法虽然完整而准确地保存了短语的结构和语义信息,但作为一种索引形式,这种表示方式过于烦琐。在实际应用中,除 Hull(1996)认为语言驱动的浅层句法分析能够提高检索系统识别名词短语的能力,其它试验认为对提高系统的检索效果并未起到太大作用。

2.3.3 规范化表示法

规范化表示法对所抽取的短语作进一步处理,并把它转化为统一的表示形式。它避免了直接表示法和结构表示法存在的问题,起到了折衷的作用。目前国外对规范形式的研究较多,其中使用最多的是"head + modifier"。

首先,从信息检索的角度出发,考虑到名词短语和动词短语在索引和查询中占相当大的比例,在建立索引时仅考虑这两种短语,对名词短语和动词短语可作以下定义:

核心名词短语(Noun Phase)的一般形式为:

$$NP = pre^* \quad head \quad post^*$$

其中 pre 为前修饰词,一般包括形容词、区别词和名词;head 为主要名词;post 为后修饰词或其他补充成分;* 表示零个或多个元素。核心动词短语(Verb Phrase)的一般形式为:

$$VP = subj \quad kernel \quad comp^*$$

subj 是动作的主语,一般为名词或代词;kernel 是主要动词;comp 是其他成分或一般为动作的宾语或介词短语;* 表示零个或多个元素。从短语的定义可以看出,短语中都有一个主要的词作为短语的中心词,如在名词短语中是主要名词(head),动词短语中是主要动词(kernel),而短语中其他的成分都为修饰部分。这样,就可以将短语的规范化形式(Phrase Frame)简化表示为:

$$PF = (h, m)$$

其中,中心词 h 描述的是短语中的中心概念,修饰词 m 对概念作进一步的限制,增强概念的专指性。当从句子中抽取的短语不止一个时,可以表示为$(h_1, m_1); (h_2, m_2); \ldots; (h_n, m_n)$的形式,短语之间用";"分隔开。而且,还允许修饰词为空,此时表示一个单独的中心词。在规范化的短语结构基础上,本文中采用的主索引文件格式如图1所示。

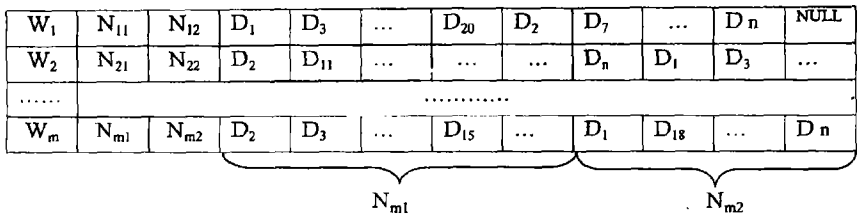

图1 基于短语的主索引文件格式

其中，第一项 W_i 是出现在短语中的词，并只取名词和动词；第二项 N_{i1} 表示 W_i 在短语索引中为中心词的文件数；第三项 N_{i2} 表示 W_i 在短语索引中为修饰词的文件数。从第四项开始的后 N_{i1} 项为短语索引中包含 W_i 的文本号，且 W_i 在短语中为中心词。此后的 N_{i2} 项为短语索引中包含 W_i 的文本号，且 W_i 在短语中为修饰词。除了这个基本索引文件外，还建立辅助短语索引表，其结构如表1。

表1 辅助的短语索引表结构

D_1	$(W_1,W_2);(W_3,W_m);\cdots\cdots$
D_2	$(W_2,W_1);(W_m);\cdots\cdots$
D_3	$(W_m,W_3);(W_1,W_2);\cdots\cdots$
.....	.....
D_n	$(W_2,W_m);(W_3,W_1);\cdots\cdots$

其中，第一项 D_i 为文本号，第二项为文本 D_i 中的各个短语项。进行检索时，用查询短语中的中心词 W_i 或修饰词 W_j 在主索引文件中查找，分别在由 N_{i1} 和 N_{j2} 确定的区域中找到一组文本号，然后由此文本号指引到辅助索引表中，得到 W_i 和 W_j 共现的文本集合，然后判断 W_i 和 W_j 是否出现在同一个短语索引中，如果满足该条件，则匹配成功。

对某些检索情形而言，用户使用自然语言来表达提问更为方便，也更为准确。但这同时也给检索处理带来了困难。由于自然语言检索是针对用户自然语言形式的提问进行处理，所以该类检索就必须利用自然语言处理的方法对提问进行词法、句法乃至语义分析。由于语义分析在技术上所面临的困难，使得在句法层次上处理和理解提问需求成为一种相对可行的方法。为使检索中能有与其相匹配的对象，就相应地需要对源文本进行句法层次的分析和索引，短语结构索引的设计就是为了满足这样一个目的。当然短语结构索引在检索中对用户提问意图的表达和对查全率和查准率提高的程序还有待于在检索实践中进一步检验。

参考文献

1 Tomek Strzalkowski. Natural Language Information Retrieval. Kluwer Academic Publishers, 1999
2 A. T. Arampatzis, T. Tsoris. C. H. A. Koster, Th. P. van der Weide. Phrase－based Information Retrieval. Information Processing & Management, 1998(12)
3 David D. Lewis, Karen Sparck Jones. Natural Language Processing For Information Retrieval
4 Chen Hongbiao. Looking for Better Chinese Indexes: A Corpus-based Approach to Base NP Detection and Indexing. Doctor thesis, 2001
5 Jaana Kristensen. Expanding End-Users Query Statements for Free-Text Searching with a Search-Aid Thesaurus. Information Processing & Management, 1993(6)
6 Li Wenjie. Natural Language processing for Chinese Information Retrieval. Transaction of Tianjin University, 2000(6)
7 张琪玉.自然语言检索中各种因素对检索效率的影响.情报理论与实践,1997(5)
8 邱君瑞.自然语言处理与信息检索系统.情报杂志,2002(3)
9 刘伟权,钟义信.自然语言处理与全文情报检索.情报理论与实践,1997(1)
10 于中华,唐常杰,张天庆.自然语言句法结构的框架树表示方法.小型微型计算机系统,1999(8)
11 李培.句法分析标引方法研究.情报理论与实践,1999(4)
12 刘开瑛,郭炳炎.自然语言处理.北京:科学出版社,1991
13 傅承德.自然语言理解的方法与策略.郑州:河南人民出版社,2000

耿 骞 北京师范大学信息管理系副教授，北京大学在读博士生。通讯地址：北京大学。邮编 100871。
汤艳莉 北京大学在读博士生。通讯地址同上。

(来稿时间:2003-07-31)