

专业信息检索系统中索引项提取策略的研究

熊桂喜, 陆壮飞

(北京航空航天大学 计算机学院, 北京 100083)

摘要:索引项(Index Unit)的提取是中文全文检索领域的关键技术之一。将通用搜索引擎的索引项提取策略应用于某一专业领域的检索系统中,会出现因标引词典无法覆盖该领域的专业词汇而造成的查准率偏低和因辞典不断加入专业词汇而造成检索效率降低的矛盾。介绍了一种面向专业领域的索引项提取策略,通过在提取过程中区分索引项和专业索引项并分别计算其权值,提升专业索引项与目标文档的相似度。在北京公安交通管理领域的网页数据集进行实验,证明该索引策略在查询专业领域信息时可提供较高的查准率并显著提高检索效率。

关键词:专业检索;索引权重;倒排文件

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2008)09-0019-03

Research on Index Unit Selection Strategy in Domain - Oriented Information Retrieval

XIONG Gui-xi, LU Zhuang-fei

(School of Computer Science, Beihang University, Beijing 100083, China)

Abstract: Index unit selection is one of the key technologies of Chinese full-text retrieval. General purpose index unit selection method is not suitable for domain-oriented information retrieval because there will be a conflict between the low accuracy and recall rate caused by a base-line dictionary which can not cover domain-oriented words and the decrease of performance caused by the growing dictionary size. Introduces a domain-oriented index unit selection strategy. It is able to increase the similarity between domain-specific index units and documents by selecting domain-specific and general index units and calculating their weight separately. In this paper, use documents from Beijing Public Traffic Management website as experimental data set, show that the domain-oriented index unit selection strategy can provide better accuracy and significantly improved retrieval efficiency.

Key words: domain-oriented information retrieval; index weight; inverted file

0 引言

搜索引擎系统普遍采用基于倒排文件的全文索引技术。在倒排索引文件的构建过程中,索引项(Index Unit)的提取是最基本也是最重要的步骤,其效果将直接影响索引的效率和索引文件标引的准确率及系统的检索效率。索引项的提取主要由分词、过滤、降噪等步骤组成。在英文文档中,由于英语词汇可以由空格自然的划分开,因此可以直接依据空格提取出单词索引项;对于用中文或其他一些没有自然分隔符的语言写成的文档,需要使用专门的分词/索引项提取策略。

对应于不同的分词方法,中文信息检索领域通常使用如下三种基本的索引项:

①单字索引项:将中文文本中每个字都作为单独的索引项。

② n 元语法(n -gram)索引项:将文本中所有相邻的 n 个字符合并作为索引项。

③词索引项:从文本中提取具有语义的词作为索引项。

文献[1]的研究表明,相对于单字和 n 元语法,使用词作为索引项可以大幅提高短语查询的准确性,同时便于系统利用语言学知识。实际应用中的通用检索系统通常使用以上两种甚至三种索引项相结合的方式提取索引项,以提高检索查全率。文献[2]的研究表明使用词索引+单字索引,或词索引+2元语法组合的索引效果较为理想。

目前,中文自动分词的成熟技术都是基于分词词典的机械型分词方法。通用检索系统所使用的中文分词词典主要包含:基本词;常见人名;常见地名;专有名词;日期和数字;噪声词等词典。专业检索系统在包含

收稿日期:2007-12-26

基金项目:“十五”国家科技攻关计划(2005BA414B04)

作者简介:熊桂喜(1964-),男,硕士,副教授,研究方向为企业应用系统集成(EAI)、智能交通系统(ITS)。

这些类别的词汇基础上,还应包括:专业词汇(包括专业名称、人名、地名、产品名等等);专业短语。如将通用检索系统的索引项提取策略应用于某一专业领域检索系统,则会出现因为无法充分提取专业索引项而导致的查准率、检索效率偏低的问题。

为了解决上述问题,采用了一种分别提取专业索引项和通用索引项,分别为其计算权值的策略,并以北京市公安交通管理检索系统为例进行了验证实验。

1 专业索引项提取策略

文中的索引项提取策略主要着眼于解决以下两个问题:

第一,由于词库规模的限制,通用检索系统无法完全提取某一专业领域的索引项。以北京公安交通管理领域的专业短语为例,若使用通用索引项提取策略处理文本“高粱桥斜街”,提取的结果为“高粱/桥/斜街”三个索引项。在检索过程中,虽然最终返回的结果会根据匹配度排序算法将正确的文档,即同时出现这三个索引项的文档列在最前端,但也会返回大量无关的文档,如文档库中所有包含“桥”的文档。系统的查准率低,而且耗时大。

第二,通用检索系统使用的索引项提取策略不适合专业索引项的提取,主要体现在两方面:首先,虽然系统通过扩大词库规模可以提取出更多的索引项,但无法在识别出新加入的专业短语的基础上进一步提取出短语中包含的基本词,使索引项的提取精度降低,进而降低检索操作的召回率。另一方面,词库的内容混杂且规模不利于控制,为检索系统的管理工作带来困难。

1.1 词典组织

第一级词典用于提取专业索引项。首先需要建立专业词典。建立专业词库主要有三种方法:

- 1) 对专业语料资源进行人工分词;
- 2) 对专业语料资源进行词频统计学习;
- 3) 整理专业数据库中的专有名词(如地名、产品名)。

文中实验部分使用北京航空航天大学计算机学院针对北京市公安交通领域整理的 3 万词词库,其中包括北京市各街道名称 7437 条,桥梁名称 3035 条,路口名称 5505 条,车站名称 5145 条,以及其他公安交通专业术语约 9000 条。该词库基本覆盖了北京地区公安交通管理业务所涉及的专业词汇及短语。

第二级词典即为通用的基本分词词典。文中实验中使用北京航空航天大学计算机学院整理的 20 万字通用分词词典。

1.2 索引项提取原理

虽然基本词+单字和基本词+二元语法等索引项提取策略在通用检索领域都有较优秀的表现^[2],尤其体现在查全率方面。但若将其应用在专业领域,则暴露出准确率和检索效率偏低,不适合专业检索要求的现象。文中采用的方法由两级提取操作构成,每一级都使用词作为索引项。第一级为专业索引项提取,在提取过程中使用基于专业词典的正向最大匹配分词算法处理待索引文档。第一级索引项提取完成后,对结果集进行过滤,只从中取出与专业词典匹配的部分作为待索引文档的专业索引项。第二级为基本词索引,使用基于基本词库的正反双向最大匹配算法再一次处理待索引文档,并进行歧义识别、噪声剔除等操作。第二级提取出的索引项,包括匹配成功的词和未被识别的单字、符号等等,作为待索引文档的基本索引项。经过两级提取,专业索引项和基本索引项都作为待索引文档的有效索引项存入倒排文件,如图 1 所示。

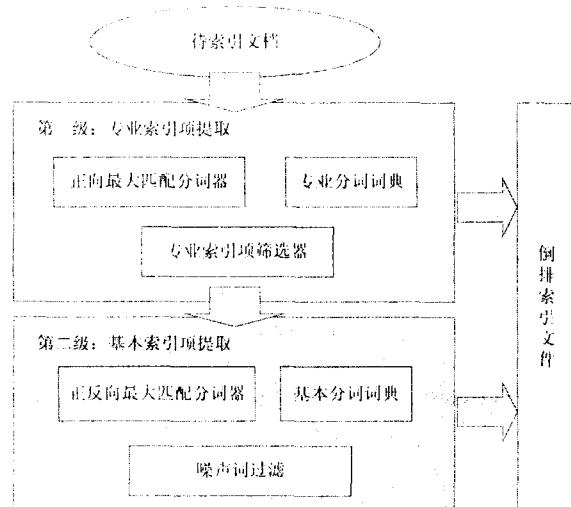


图 1 专业索引器逻辑结构

例如:处理文本“高粱桥斜街(交大东路至大慧寺路路口)禁止机动车通行”时,先通过第一级分词提取出“高粱桥斜街/交大东路/大慧寺路路口”三个专业索引项,再通过第二级分词将文本划分为“高粱/桥/斜街/交大/东路/至/大慧寺/路/路口/禁止/机动车/通行”等基本索引项,则最终提取出的索引项即为“高粱桥斜街/交大东路/大慧寺路路口/高粱/桥/斜街/交大/东路/大慧寺/路/路口/禁止/机动车/通行”。如图 2 所示。

1.3 索引项权重的计算

文中采用基于向量空间模型的加权算法对提取出的索引项进行加权。向量空间模型于 20 世纪 60 年代末由 G. Salton 等人提出,是目前普遍采用的信息检索模型。向量空间模型将检索文档和检索词(关键词)表

示为向量的形式,根据向量表示的颗粒度大小不同,计算向量每一维的元素(字、词或短语)的索引项权值,然后依向量空间的相似度计算结果来排列检索结果^[3]。

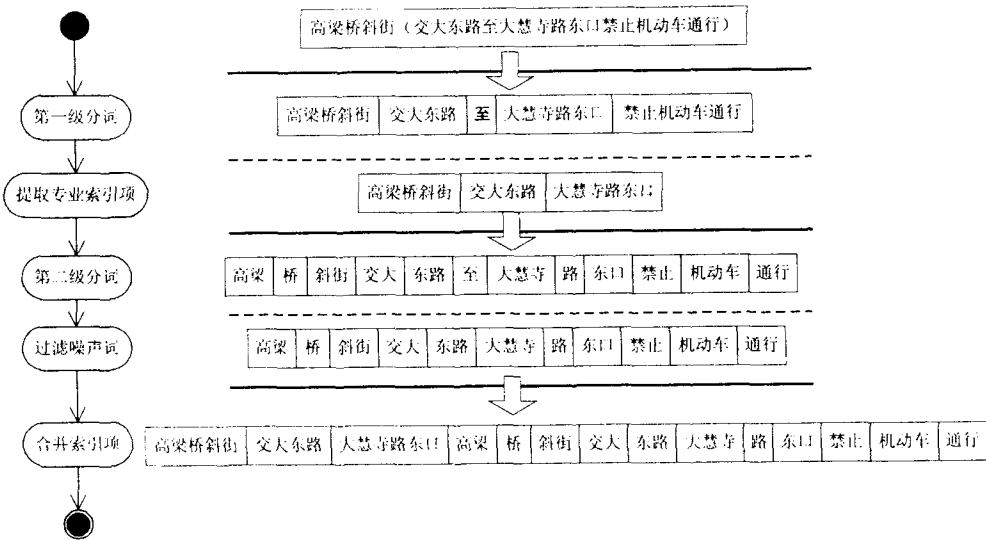


图 2 索引项提取流程及示例

对于基本索引项,采用一种改进的 TFIDF 方法计算索引项的权值。将索引项表示为 $F = \{f_1, f_2, \dots, f_n\}$, 每个文档 d_i 表示为所有索引项权值组成的文档向量 $d_i = (w_{1i}, w_{2i}, \dots, w_{ni})$, 分别计算索引项相对于文档的 TF 值与 IDF 值,并以 TF 值与 IDF 值的乘积作为索引项相对于文档的权重。

TF(Term Frequency) 是索引项频率,记为 $TF(f_i, d_j)$, 基本思想是利用索引项在文档中出现的频次为索引项加权,索引项在文档中出现的次数越多,该文档与检索就越相关,二者的相似度越大。

IDF(inverse document frequency) 是逆文档频率,设 m 是文档集中文档的数目, $DF(f_i)$ 是文档集中含有 f_i 的文档数目, IDF 的定义可由公式(1)和(2)表示:

$$IDF(f_i) = \log \frac{m}{DF(f_i)} \quad (1)$$

或:

$$IDF(f_i) = \log(1 + \frac{m}{DF(f_i)}) \quad (2)$$

TF 从局部上反映了单个文档与检索项之间绝对的相关性,而 IDF 则从整个文档集的全局出发,从全局上反映了每个文档与检索项之间相对的相关性,侧重考虑了文档之间的差异性。文中采用的算法的权重综合考虑了 TF 和 IDF,定义 w_{ij} (第 i 个索引项相对于第 j 篇文档的权重) 计算方法如公式(3) 所示:

$$w_{ij} = \alpha * \frac{TF(f_i, d_j)}{\max_k TF(f_k, d_j)} * \log(1 + \frac{m}{DF(f_i)}), \alpha \geq 1 \quad (3)$$

公式中的 α 为不小于 1 的常数,在计算专业索引

项权值时适当增大其值(文中实验部分取 α 值为 1.5), 提高该文档与专业索引项的相似度。在实际应用中, 对于由多个基本索引项组成的专业短语索引项还可以

使用 Fagan 提出的短语索引项加权算法对其进行进一步加权^[4]。

2 测试实验及测试结果

文中测试实验选用的目标文档集为北京航空航天大学计算机学院整理的北京公安交通管理网页库,该库包含的网页主要来自北京市公安局网站(www.bjjtgl.gov.cn),截止至 2007 年 9 月 1 日。测试方案参考了

TREC^[5]及中文 WEB 信息检索评测^[6],旨在针对专业索引策略的查准率和检索效率进行初步评测。测试集由 10 条查询组成,包括了地名、专业名词、机构名和法规名等。

测试集分为 TD(Topic Distillation, 主题提取)和 NP(Named Page search, 指定页面查询)两类任务。TD 的主要目的是测试系统对于一个特定主题发现一组关键资源的能力,通常根据在前 n 个结果中有几个正确的答案来进行判断,记作 $P(n)$ 。文中选取 n 值为 10, 即 $P(10)$ 作为本测试的评价标准。NP 的主要目的是评测系统对指定页面的查询能力,评价的方法为取正确结果在结果列表中排位的倒数作为测量值。

表 1 为测试集,表 2 为对文中采用的专业索引项提取策略与基本词+单字索引策略的检索效果对比数据。

表 1 查询测试集

编号	任务类别	查询语句	查询描述
1	TD	西直门桥	地名查询
2	TD	高粱桥斜街	地名查询
3	TD	四道口东口	公交站名查询
4	TD	营运机动车	专业名词查询
5	TD	天网行动	专业名词查询
6	TD	交强险	专业名词简称查询
7	TD	朝阳支队劲松队	查询组织机构信息
8	NP	72 号令	查询机动车登记规定(72 号令)的信息
9	NP	东城交通支队事故科	查询组织机构信息
10	NP	2007 年 9 月驾驶证停止	查询 2007 年 9 月机动车驾驶证停止使用的公告

(下转第 25 页)

作为类复合层次上嵌套属性的索引,可以通过遍历简单的索引来实现复杂的检索和查询。

设 H 表示实例模式类的聚集 $C_1 C_2 \cdots C_n$, H 上的一条检索路径定义为 $P = C_i A_1 A_2 \cdots A_n (n > 1)$, 其中 C_i 表示 H 中的一个类, A_i 表示 C_i 中的一个属性, 则 C_i 是类 C_{i-1} 的一个属性 A_{i-1} 的值域 ($1 < i \leq n$)。

由于库中的实例模式类具有多种属性和成分, 嵌套索引同时也是对多种成分检索的有效接口。通过各个实例模式类的公共属性, 可以对多个不同的实例模式类进行联合检索。由此可见, 嵌套索引是体现面向对象特征检索的索引形式, 直接关系到面向对象的分析和检索效率。

3 结束语

随着软件设计模式的日益增多, 在各个学术和技术领域, 模式都在有意无意地应用着, 把这些人类知识的精华进行有效复用的意义是非常大的, 有效地复用模式需要首先有效地组织和管理模式, 文中对软件设计模式的特性进行分析, 提出一种面向对象的软件实例模式库来存储设计模式, 并给出相应的库组织结构、

管理和索引方法。

采用面向对象的方法对软件设计模式进行划分、组织和封装, 克服了传统数据库系统普遍存在的数据类型表达能力弱, 复杂查询功能差等缺点, 提高了模式知识的共享性和重用性, 增强了对实例模式的检索和管理能力。力图在此基础上研制出一个高效且真正实用的设计模式的复用和支持系统。

参考文献:

- [1] Gamma E, Helm R, Vlissides J J. Design Patterns: Elements of Reusable Object - Oriented Software[M]: [s. l.]: Addison - Wesley, 1995.
- [2] 王晓庆, 曾文英. 设计模式中的面向对象原则及其子模式[J]. 计算机工程, 2003, 29(9): 192 - 193.
- [3] 孟祥文, 邵维忠. 设计模式特化和模式库组织[J]. 计算机工程, 2002, 28(5): 36 - 37.
- [4] Su S, Lam H, Eddula S, et al. Osam: An object - oriented knowledge base management system for supporting advanced application[J]. ACM SIGMOD, 1993, 22(2): 12 - 40.
- [5] 雷光复. 面向对象的新一代数据库系统[M]. 北京: 国防工业出版社, 2000.

(上接第 21 页)

由表 2 中两种索引项提取方式的横向对比数据可以看出, 两种索引项提取方式在执行 TD ($N = 10$) 和 NP 任务时均可获得令人满意的效果。但专业索引项提取策略在检索效率上大大优于通用索引项提取策略, 全部结果中的无关文档数量也明显减少。

表 2 测试结果对比数据

	专业词 + 基本词索引项				基本词 + 单字索引项			
	P(10)	R	结果数量	时间(ms)	P(10)	R	结果数量	时间(ms)
1	10/10	/	32	6	8/10	/	418	28
2	4/4	/	4	≤ 5	4/10	/	372	21
3	9/9	/	9	≤ 5	9/10	/	3797	118
4	8/10	/	25	≤ 5	8/10	/	5151	162
5	5/5	/	5	≤ 5	5/10	/	52	8
6	6/6	/	6	≤ 5	6/10	/	41	11
7	3/10	/	1687	38	3/10	/	5802	181
8	/	1	2	≤ 5	/	1	30	8
9	/	1	17	≤ 5	/	1	6594	139
10	/	1	157	15	/	1	226	32

3 结束语

文中提出的索引项提取策略可在不降低基本索引项提取精度的前提下有效地提取专业索引项。并经过

实验证明, 此方法提取的索引项可以较明显地提高专业检索系统的检索效率, 降低硬件开销。

参考文献:

- [1] Nie Jian - Yun, Gao Jiangfeng, Zhang Jian, et al. On the use of words and n - grams for Chinese information retrieval[C] // In: Proceedings of the fifth international workshop on Information retrieval with Asian languages. Hong Kong, China: [s. n.], 2000: 141 - 148.
- [2] He Hongzhao, Gao Jianfeng. Finding the better indexing units for Chinese information retrieval[C] // In: Proceeding of the first SIGHAN workshop on Chinese language processing - Volume 18. [s. l.]: [s. n.], 2002: 1 - 7.
- [3] Salton G. A Vector Space Model for Automatic Indexing[J]. Communications of the ACM, 1975, 18: 613 - 620.
- [4] Fagan J L. Experiments in Automatic Phrase Indexing For Document Retrieval: A Comparison of Syntactic and Non - Syntactic Methods[D]. Upson Hall Ithaca, NY, USA: Computer Science/Arts & Eng. Dept., Cornell University, 1987.
- [5] TREC[EB/OL]. 2004. <http://nlp.mit.edu/books/chapters/0262220733chap1.pdf>.
- [6] SEWM[EB/OL]. 2007 - 03. <http://www.cwrf.org/Evaluation/CWT.html>.