文章编号: 1000-8608(2007)03-0439-05

一种基于同义词词典的模糊查询扩展方法

马晖男*, 吴江宁, 潘东华

(大连理工大学 系统工程研究所, 辽宁 大连 116024)

摘要: 在信息检索系统中,查询扩展是一种非常有效的改进检索性能的方法. 为此,提出一种基于同义词词典的模糊查询扩展方法. 该方法中的同义词词典是基于著名的语义词典WordNet中的同义词集合建立的,同义词之间的贴近度[0, 1]使用Tanimoto系数获得. 利用该词典,能够进行较好的查询扩展. 将该方法与向量空间模型结合应用于文本信息检索系统中,所构造的检索模型相当于一种简单的语义模型,并且可以根据阈值来控制查询扩展的程度. 所得试验结果表明,使用该查询扩展方法的信息检索系统较常规信息检索系统的检索性能有一定改善.

关键词: 模糊查询扩展; 同义词词典; 信息检索
中图分类号: TP391 **文献标识码:** A

0 引 言

随着网络信息时代的到来,信息日新月异,并呈指数增长,人们已经生活在海量的信息数据世界中. 为使用户快速、高效地找出与其需求相关的信息,高性能信息检索系统显得越来越重要.

在传统的基于关键词的信息检索系统中,用户给定的查询语句与文本常常被解析成独立的实词,主要为名词、动词、形容词、副词. 利用这类系统进行检索时,常会有两种现象出现: (1) 检索时,系统将名词、动词、形容词、副词单独看待,尽管在改进的模型中加入了一些逻辑运算符,但始终无法清晰地表示这些词间的语义关系,尤其是形容词和副词脱离上下文后,很难识别其正确的语义,而由于这些语义不确定的实词参与检索过程,系统常常产生许多与用户查询不相关的信息; (2) 检索时,用户表达查询需求可以使用异形同义的词语(同义词),因此倘若仅仅使用给定查询词进行检索,忽略其同义扩展词,常会导致系统漏检很多与查询语义相关的信息. 反之,若在检索时使用同义词替换,则更多的用户感兴趣的信息能够被检索出来.

查询扩展(query expansion, QE)方法[1~3]可以避免检索系统受限于用户指定的查询项,使得其能够在更广泛的意义或概念上检索出与用户需求相关的信息,解决因用词不同而造成的检索效率不高的问题. 查询扩展主要有3种方法: 以查询语句为基础的查询扩展,以语料库为基础的查询扩展,以及以语言分析特征为基础的查询扩展.

本文采用以查询语句为基础的扩展策略,提出一个基于同义词词典的模糊查询扩展方法,构建符合使用者信息需求的检索机制,以增进文本检索系统的检索效益.

1 同义词词典的结构

1.1 同义词词典的建立

通常,查询扩展是以同义词词典为基础进行的,为此研究中首先建立了模糊同义词词典,该词典能够针对给定的词及其词义,提供它的一组同义词以及相应的对给定词的贴近度.

目前,广泛使用的权威的英文同义词词典为WordNet[4]. WordNet是由普林斯顿大学的心理学家、语言学家和计算机工程师联合设计的一个基于认知语言学的英语词典,它最具特色之处是根据词义而不是词形来组织词汇信息. 从这方面讲,WordNet更像是一部语义词典而不是普通词典. WordNet按照单词的意义组成一个"单词网络",如果查找某一单词,利用该网络能迅速获得

收稿日期: 2005-12-10; 修回日期: 2007-03-10.
基金项目: 日本佳思腾株式会社资助项目.
作者简介: 马晖男*(1977-),女,博士.

这个单词所有词性的所有词义,这些叫做“sense”,根据不同的“sense”可获得其对应的同义词[5].

利用 WordNet虽然可以得到每个词义的同义词,但却缺少同义词之间的贴近度. 研究表明,词性相同的同义词尽管词位意义(sense)相同,但在具体应用时其词位指示的客观对象(denotation)、应用中词语的具体所指(reference)和应用中词语所指客观对象(referent)均可不同. 为了衡量同义词间不同的贴近度,本文提出建立一个模糊同义词词典. 该词典中,每一对同义词之间的贴近度用 0~1的某个小数度量. 根据计算得到的贴近度,可以对扩展词的相近程度进行确定. 设定扩展阈值,则可对原查询向量进行模糊扩展. 使用该模糊同义词词典,能够将合成短语进行不同阈值下的扩展.

贴近度的计算主要以合成短语的共现率作为主要考量因素. 此外,还考虑了扩展短语对文本的隶属程度这一因素.

假设两个短语经常在给定文本集合中共同出现,可认为它们表示相同的概念. 因此,查询语句中最初的短语 t_k 与扩展短语 t_e 之间的贴近度可以由短语之间的相关系数 r_{ke} 来确定,该系数以 WordNet词典为依据,这里采用了 Tanimoto系数[6]:

$$r_{ke} = \frac{n_{ke}}{n_k + n_e - n_{ke}} \tag{1}$$

式中: n_k 代表词典中包含短语 t_k 的同义词集合数,n_e 代表词典中包含短语 t_e 的同义词集合数,n_{ke} 代表词典中既包含短语 t_k 又包含短语 t_e 的同义词集合数.

对于与短语 t_k 相对应的所有扩展短语的贴近度可构成一个相关系数矩阵,用 $\boldsymbol{R}_k$ 表示,其形式如下:

$$\boldsymbol{R}_k = (r_{ke}) \quad (k, e = 1, 2, \cdots, N_k) \tag{2}$$

式中: r_{ke} 表示短语 t_k 与短语 t_e 之间的贴近度,$r_{ke} \in [0, 1]$,N_k 为扩展短语的个数. 对于 $k \neq e$,则 $r_{ke} = r_{ek}$;对于 $k = e$,则 $r_{kk} = r_{ee} = 1$.

实际应用中,考虑到词的搭配原则以及扩展短语的统计意义,并不是所有相关的扩展短语在给定文本中均能替换原短语,基于这一点,本文定义了一个反映扩展短语对文本重要程度的隶属函数,其值计算公式如下:

$$\_{ed}_i = 1 - \prod_{\substack{m=1 \\ m \neq e}}^{N_k} (1 - r_{me}),\ r_{me} \in \boldsymbol{R}_k \tag{3}$$

式中 $\_{ed}_i (e = 1, \cdots, N_k)$ 表示短语 t_e 对文本 $\boldsymbol{d}_i$ 的隶属程度,$\_{ed}_i \in [0, 1]$,它是作用在所有扩展短语上的一个负代数积的补数.

据此,可以利用 $\_{ed}_i$ 来衡量扩展短语对文本的重要程度,并通过预设定的隶属度阈值对所有的扩展短语进行删减处理.

1.2 同义词词典的结构表示

为节省篇幅,本文只给出形容词词性的模糊同义词词典的结构表示,其他词性的同义词词典结构表示与其类似. 本文建立的模糊同义词词典主要由两级表构成: 第一级表为形容词词典,保存所有形容词、每个形容词根据其词意所对应的 ID编号、该 ID编号所对应的 sense,如表 1所示;第二级表为形容词同义词词典,保存该形容词的 ID编号、该 ID编号所对应的同义词的 ID编号、第一个 ID编号所对应的形容词的词形、同义词之间的贴近度,如表 2所示. 研究中,根据 WordNet提供的同义词集合计算贴近度并建立相关系数矩阵,存储到两级表中.

表 1 第一级形容词词典

Tab. 1 The first class of adjective thesaurus

ID1	Adj	Description
⋮	⋮	⋮
18	Fast	Sense1: refer to something, such as activity or movement, marked by great speed
19	Fast	Sense2: Firmly fixed or fastened
108	Quick	Sense1
256	Tense	Sense1
⋮	⋮	⋮

表 2 第二级形容词同义词词典

Tab. 2 The second class of adjective synonymy thesaurus

ID2	ID1	Syn	Degree
⋮	⋮	⋮	⋮
18	18	Fast	1.0
18	108	Quick	0.9
19	19	Fast	1.0
19	256	Tense	0.8
108	108	Quick	1.0
108	18	Fast	0.9
256	256	Tense	1.0
256	19	Fast	0.8
⋮	⋮	⋮	⋮

表 3和表 4分别给出了形容词词典中第一、二级表各字段名称以及所对应的字段说明、数据类型、数据长度.

表 3　第一级形容词词典结构

Tab. 3　The first class of adjective thesaurus structure

字段名称	字段说明	数据类型	数据长度
ID1	某一个词意(sense)对应的编号	Int	4
Adj	形容词	Char	25
Description	该形容词的某一个词意描述	Vchar	50

表 4　第二级形容词同义词词典结构

Tab. 4　The second class of adjective synonymy thesaurus structure

字段名称	字段说明	数据类型	数据长度
ID2	与第一级表相对应的形容词编号	Int	4
ID1	该形容词的某一个同义词在第一级表中对应的编号	Int	4
Syn	与 ID2对应的词意的形容词词形	Char	25
Degree	ID2与 ID1的贴近度	Float	8

2　基于同义词词典的模糊查询扩展

基于同义词词典的模糊查询扩展方法与传统扩展方法的差别在于其不仅能够完成传统的查询扩展(阈值设为 0),而且还能根据检索结果精度的要求,调整阈值,进行不同程度的查询扩展,以求更优的检索结果.

2.1　模糊查询扩展方法

查询扩展方法早已被应用于文本信息检索系统中[3],其动机是解决由于用户不能够准确表达查询中的关键词的用词,而导致相关文本检索困难的问题. 但对于合成短语的扩展方法还没有文献涉及,本文正是以此为出发点,以查询语句为基础进行扩展,提出了一个新的模糊查询扩展方法,以此改善信息检索系统的检索效率.

具体做法是首先将查询语句中的修饰词(形容词、副词)与被其修饰的中心词(名词、动词)组成整体关键词(本文称其为合成短语)作为查询项,以此避免修饰词单独检索带来的无用信息;然后,利用同义词词典对修饰词和中心词分别进行模糊扩展并重组,形成新的统计意义下的候选合成短语;最后,根据扩展的查询向量,采用改进的向量空间模型(modified vector space model, MVSM),在文本数据库中进行检索.

进行文本信息检索时,用户输入的查询语句经解析后,可表示为向量形式:

$$\boldsymbol{q} = (w_1 \quad w_2 \quad \cdots \quad w_k \quad \cdots \quad w_l)^{\mathrm{T}} \tag{4}$$

其中 w_k 表示查询向量中查询索引项所包含的短语 t_k 的权重值,若查询语句中中心词无修饰词修饰,t_k 为独立实词短语;若有修饰词修饰,t_k 为合成短语.

进行模糊查询扩展时,原查询向量需增加扩展查询索引项的信息. 假设给定的查询向量 $\boldsymbol{q}$ 中包含 l 个合成短语,若所有合成短语根据同义词词典均可被扩展,则可得到扩展的候选合成短语. 如果将这些候选合成短语直接加到最初的查询向量中,新的扩展的查询向量则可表示为

$$\begin{aligned}\boldsymbol{q} = (&w_1 \quad w_{1,1} \quad w_{1,2} \quad \cdots \quad w_{1,N_1} \quad w_2 \quad \cdots \\ &w_k \quad w_{k,1} \quad \cdots \quad w_{k,e} \quad \cdots \quad w_{k,N_k} \quad \cdots \quad w_l \\ &w_{l,1} \quad \cdots \quad w_{l,N_l})^{\mathrm{T}}\end{aligned} \tag{5}$$

$w_{k,e}$ 是与 t_k 相关的扩展的候选合成短语 t_e 的权重,其中 $e = 1, \cdots, N_k$, N_k 为 t_k 根据同义词词典扩展出的短语总数量.

考虑到扩展的合成短语在一定程度上与原来的合成短语相关,以及它对文本的重要程度这一因素,本文用由式(3)中得到的 $\_ed_i$ 来减弱扩展合成短语对文本的依附程度,其中 $\_ed_i \in [0, 1]$. 修正后的扩展的候选合成短语权重计算公式如下:

$$\begin{aligned}&w_{k,e} = tf_{ie} \times idf_e \times \_ed_i; \\ &e = 1, \cdots, N_k\end{aligned} \tag{6}$$

式中: tf_{ie}(词频)是指文本 d_i 中合成短语 t_e 出现的次数,idf_e(逆文本频率)是指文本集合中文本数与包含当前合成短语 t_e 的文本数之比的对数,其计算公式如下:

$$idf_e = \log(|D| / DF(t_e)) \tag{7}$$

其中 $|D|$ 为文本集合中的文本数量,$DF(t_e)$ 为文本集合中含有 t_e 的文本数量.

将式(6)标准化,最终得到

$$w'_{k,e} = \frac{tf_{ie} \times idf_e \times \_ed_i}{\sum_{e=1}^{N_k} (tf_{ie} \times idf_e \times \_ed_i)^2} \tag{8}$$

将标准化的权重值进行排序处理,根据用户提出的检索条件,设定扩展阈值进行筛选,确定最终保留在查询向量中的扩展短语作为最后的查询索引项.

2.2　模糊查询扩展方法在文本信息检索中应用

研究中,文本信息采用向量空间模型[7,8]来表示,该模型由 Salton等于 1975年首次提出,其在数学表达上具有简洁、易于操作的特性,使得它在文本信息检索系统中得到广泛应用.

使用向量空间模型,可以将文本、查询语句在多维空间中进行表示. 当进行检索时,将文本集中每个文本进行特征表示,根据关键词出现的频率确定关键词权重,并表示成向量的形式,然后计

算文本与查询语句的相关程度,其值可用文本向量与查询向量之间的夹角余弦值来确定. 相关度为 0~ 1间的值,越趋近于 1,表示该文本与查询需求越匹配.

但向量空间模型自身存在一个缺点,即当查询语句中的关键词表示为一个向量时,这些关键词之间是正交的,不存在任何依附关系,其后果是文本中的上下文关联得不到反映. 将本文提出的模糊查询扩展方法与传统的向量空间模型相结合,将特征提取由传统的提取关键词转为提取合成短语,由于有了修饰词修饰和限制中心词的这种语义环境的体现,文本中部分语义关系在一定程度上能够得以反映.

应用本文提出的模糊查询扩展方法的信息检索系统的主要构架如图 1所示: 图中虚线部分为模糊查询扩展的过程. 具体描述是根据用户给出的查询语句,通常它会包含带有修饰词修饰中心词这样的合成短语,通过词性标注,并使用同义词词典,可以分别得到修饰词以及中心词的同义词,将这些扩展的修饰词、中心词重新进行统计意义下的语义组合,便可得到新的合成短语. 利用所得短语构造向量空间模型,并计算文本向量与查询向量之间的相关度,根据相关度值大小进行文本排序,最后,得到与查询语句语义相关的文本.

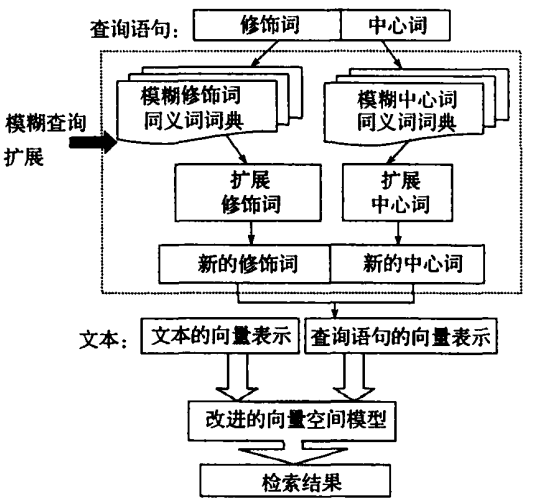

图 1 应用模糊查询扩展方法的信息检索系统主要构架

Fig. 1 Architecture of information retrieval system based on fuzzy query expansion

3 应用模糊查询扩展方法得到的试验结果

对于本文中提出的模糊查询扩展方法,本文利用国际标准语料库进行了试验测试. 试验中,采用 TREC语料库[9]中的 La-times文本集,其中包含 3 319篇文本. 针对系统需要,筛选了 42条查询语句,这些查询语句具有以下特点: (1)至少包含一个被修饰词修饰的中心词(本试验选用的是形容词修饰名词的合成短语)作为查询项; (2)根据所建立的同义词词典,每条查询语句中的合成短语都能够被扩展.

例如查询语句为 infectious diseases: How many infectious diseases are there in the world?

根据同义词词典,进行模糊同义查询扩展,在检索结果中,不仅包含"infectious diseases"合成短语的文本,包含其他合成短语如{ infectious, contagious, communicable, epidemic} × {diseases, illness, sickness}的文本也在检索结果之列.

试验中,合成短语试验,即 Modification search,阈值设为 1. 0;所进行的模糊查询扩展试验,修饰词、中心词的查询扩展阈值都设为 0. 6.

本试验主要通过精确率和召回率[10]这两个指标来评价所提出的基于模糊查询扩展的信息检索系统的检索效率,由于篇幅所限,对于每条查询语句的精确率和召回率不作描述,这里只给出它们的平均值,见图 2.

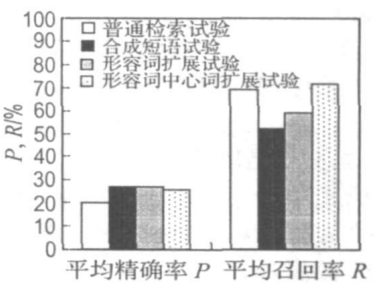

图 2 基于同义词词典的模糊查询扩展方法进行文本信息检索试验结果柱形图

Fig. 2 Bar chart of applying fuzzy query expansion method based on synonymy thesaurus to information retrieval system

从图 2中可以看出,使用本文提出的基于同义词词典的模糊查询扩展方法进行文本信息检索时,获得的精确率和召回率与传统检索系统相比均有所提高. 即将该模糊查询扩展方法应用于文本信息检索中,除了在检索结果中获得了更多的相关文本,更重要是,可以去除大量信息垃圾,为用户节约宝贵的阅读时间. 试验得到的另一结果是,扩展后的短语数量在 8~ 10个有较好的检索

结果.

4 结　语

本文提出了基于同义词词典进行模糊查询扩展的方法,并将模糊查询扩展技术与传统的向量空间模型相结合应用于文本信息检索系统中. 使用标准英文 TREC语料库中的 La-times测试文本集合进行的试验表明,与基于传统的 VSM 模型的检索系统相比,文中改进的检索系统其精确率和召回率均有所提高.

利用提出的模糊查询扩展方法,首先,使用合成短语作为查询项,使得检索过程能在一个比较窄的、精确的检索范围内进行;其次,对这些合成短语进行了同义扩展,使得文本中虽然没有出现与原查询项词形匹配的词,但与该查询项语义相关的文本能够被检索出来. 这种方法是提高文本信息检索效率的有效途径.

参考文献:

[1] LEE H M, LIN S K, HUANG C W. Interactive query expansion based on fuzzy association thesaurus for Web information retrieval [C]// **Proceedings of the 10th IEEE International Conference on Fuzzy Systems**. Australia: [s n], 2001: 724-727

[2] LIM J, SEUNG H, HWANG J, *et al*. Query expansion for intelligent information retrieval on internet [C] // **Proceedings of Parallel and Distributed Systems International Conference**. Washington: IEEE Computer Society, 1997: 656-662

[3] 贺宏朝, 何丕廉, 陈　霞. 利用人工和自动生成的资源进行中文信息检索查询扩展 [J]. 计算机工程与应用, 2002, **21**: 18-20

[4] Cognitive Science Laboratory, Princeton University. WordNet [EB/OL]. [2003-10-10] http: // www.cogsci.princeton.edu/~ wn/

[5] 姚天顺, 朱靖波, 张　俐, 等. 自然语言理解—— 一种让机器懂得人类语言的研究: 第 2版 [M]. 北京: 清华大学出版社, 2002 122-148

[6] MANDALA R, TOKUNAGA T, TANAKA H. Query expansion using heterogeneous thesauri [J]. **Inf Process and Manage**, 2000, **36**: 361-378

[7] SALTON G, WONG A, YANG C S. A vector space model for automatic indexing [J]. **Commun of the ACM**, 1975, **18**(11): 613-620

[8] JING Li-ping, HUANG Hou-kuan, SHI Hong-bo. Improved feature selection approach TFIDF in text mining [C] // **Proceedings of 1st Information Conference on Machine Learning and Cybernetics**. Beijing: [s n], 2002 944-946

[9] NIST. Text Retrieval Conference [EB/OL]. [2003-04-08] http // trec.nist.gov/

[10] 钱学森图书馆医学分馆. 信息检索基础知识: 检索效率及评价 [EB/OL]. [2007-01-10] http: // 202.117.24.24/html/xjtu/kejian/yxkj/pages/bjjc/chapter1/7.htm

An approach to fuzzy query expansion based on synonymy thesaurus

MA Hui-nan*, WU Jiang-ning, PAN Dong-hua

(Res. Inst. of Syst. Eng., Dalian Univ. of Technol., Dalian 116024, China)

Abstract Query expansion (QE) has been proved to be one of effective methods for improving the performance of the information retrieval (IR) system. Therefore, a new fuzzy QE method based on synonymy thesaurus is proposed, and the synonymy thesaurus is built based on the famous lexical database WordNet. In the synonymy thesaurus, the similarity between the synonyms is [0, 1], which is obtained by Tanimoto coefficient. By using this synonymy thesaurus, query expansion can be done well. Then the fuzzy QE method is introduced into the document information retrieval system together with the modified vector space model. The experimental results show that the developed information retrieval system has got more effective performance than before by using the fuzzy query expansion method. One feature of the proposed information retrieval model is that it can be treated as one of simple semantic models. Another feature is that the expansion degree is controllable based on different thresholds.

Key words fuzzy query expansion; synonymy thesaurus; information retrieval