基于语义理解的智能搜索引擎研究

陈 林 杨 丹 赵俊芹
(重庆大学软件学院 重庆 400044)

摘 要 本文提出了一种基于自然语言理解的搜索引擎模型。它的核心技术是基于自然语言理解的相关技术,包括从关键词、提问方式、提问重点三个层次对用户查询进行语义分析、特征向量提取及基于该思想建立了面向 Web 网页内容的特征库,提出返回文档排序的算法,基于 Lucene 全文索引工具包建立了搜索引擎,对库中已收入的特征词进行了查询测试,查准率为 86.7%。实验表明,该模型基本实现了对查询短语的理解,对提高搜索引擎的查准率有显著的效果。

关键词 自然语言处理,分词,语义分析,向量空间模型

Research on Intelligent Search Engine Based on Semantic Comprehension

CHEN Lin YANG Dan ZHAO Jun-qin
(School of Software Engineering, Chongqing University, Chongqing 400044, China)

Abstract This article proposes a search engine model which is based on the natural language understanding. It includes a method to analyze users' quest ions in natural language from three layers, that is, keyword, quest ion type and question focus. The analysis consists of semantic analysis, feature extraction and semantic matching. And with this thought the feature base that faces to Web page content is built. In addition, this article proposes an algorithm of returning to the documents arrangement, it investigates implementing retrieval system based on the Lucene toolkit. The feature words, which are collected in the feature base, are tested, and the precision ratio is about 86.7%. The test result indicates that the module can realize the semantic comprehension to query, and it has an evident effect to improve the precision of search engine.

Keywords Natural language process, Word segmentation, Semantic analysis, Vector space model

1 引言

随着互联网的高速发展,网上的信息越来越多,如何在这些海量信息中快速准确地找到所需要的信息也越来越困难。Google、Yahoo、百度、新浪、天网等中英文搜索引擎是人们徜徉信息海洋、获取信息的工具。然而目前这些搜索引擎采用以关键词检索为基础的检索技术[1],用户输入检索关键词向搜索引擎提出请求,而不是以自然语言形式提供的。经过分析,用自然语言来描述对信息的需求比用关键词准确得多,同时用户也容易做到。事实上,用户可能更习惯于用自然语言来描述一个问题而不是用一系列的关键词,例如使用"我想买手机?",那么返回的文档应该包含与手机相关的介绍、参数以及价格等信息,而不是"买 and 手机"。因此,解决这个问题的有效途径是使搜索引擎理解自然语言,即搜索引擎的智能化。本文就查询请求的语义理解进行了初探,以搜索引擎为研究对象,给出基于语义理解的搜索引擎的模型和设计思想。

2 总体流程

总体流程如图 1 所示。用户用自然语言给出查询语句,如"我想买手机",通过语义理解先对句子进行分词,得到一系列词组,然后利用词库提取关键词并且推出特征向量,之后利用信息检索系统检索出文档。

3 自然语言处理

用户以自然语言输入查询语句,系统必须理解用户想搜索的是什么,即:自然语言处理。查询语句理解分为:问句的处理和非问句的处理。

总体上,分为以下主要步骤:

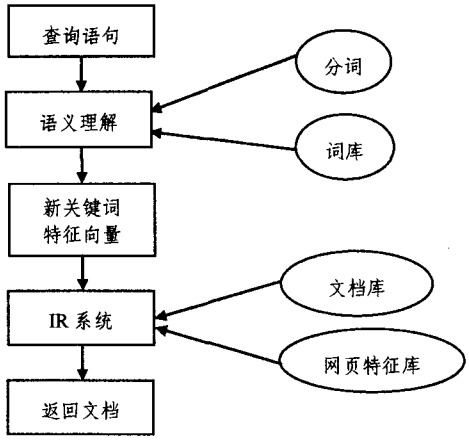

图 1 流程图

Step1(分词处理) 接收用户以自然语言方式表述的提问,使用二元分词对其进行切分,一系列词组(详见 3.1 节);

Step2(关键词与特征词提取) 根据问题分析的需要,提取关键词、问句特征词以及提问侧重点,得到包含语义信息的关键词串和特征词串(详见 3.2 节);

Step3(语义分析及查询扩展):对的关键词进行中英文转化、同义和近义词转化等处理,依据问句特征词进行提问方式识别,再对提问侧重点进行分析,最终形成包含语义信息的 n 元特征向量(又称 n 元组)(详见 3.3 节)。

3.1 分词

众所周知,英文是以词为单位的,词和词之间是靠空格隔开,而中文是以字为单位,句子中所有的字连起来才能描述一个意思。例如,英文句子 I am a student,用中文则为:"我是

陈 林 硕士,主要研究领域为自然语言处理、企业信息化;杨 丹 教授,博士生导师,主要研究领域为调度理论及其应用、企业信息化、数字图像处理;赵俊芹 硕士,主要研究领域为企业信息化。

一个学生”。计算机可以很简单通过空格知道 student 是一个单词，但是不能很容易明白“学”、“生”两个字合起来才表示一个词。把中文的汉字序列切分成有意义的词，就是中文分词，有些人也称为切词。我是一个学生，分词的结果是：我 是 一个 学生。

目前常用的分词方法有最大匹配法（Maximum Match 即 MM 法）、最小匹配法、逐词遍历匹配法等[2]。词是传统检索系统的常用的索引单元。因此在索引汉语句子之前，必须有一个预处理的过程将句子切分成能词，也可以是 N 元组（N-Gram）[3]。一般采用基于词典的分词方法，最大匹配算法通常被用来解决切分歧义。未登录词是基于词典方法的一个主要问题。特别是对于一些在检索中起重要作用的专有名词，如果未在词典中出现，将不会被索引。采用 N 元组作为索引单元即 N 元分词，不需要任何语言学知识。这种方法将文本切分成包含两个、三个或是更多相邻汉字的单元。在检索中常用的为二元组和三元组，二元组（bi-gram）将文本中所有相邻汉字均作为索引项。例如有一个字符串 $C_1C_2C_3C_4$，则由它生成的索引项为 C_1C_2，C_2C_3，C_3C_4。由于 75% 的汉语常用词是由两个汉字组成的[4]，因此使用二元分词语是一种有效的方法。这种方法的主要优点是简单和易于实现；另一方面，由于基于词频统计匹配切分对于未登录词的切分相对较弱，因此对于某些地名、人名等专有名词的检索效果远差于一元和二元切分。这是造成基于词频统计匹配切分的召回率低于机械切分的主要原因。例如，对于词典中没有的查询词，比如“大亚湾”，基于词典分词的方法将把它切分成 3 个单字。如果使用二元分词，它将被切分为“大亚”和“亚湾”。如果这两个索引单元同时在文档中出现，文档中包含“大亚湾”这个查询词的概率很大，这个概率远高于文档中出现 3 个单字时查询词出现的概率。

根据微软中国研究院的在第五届和第六届文本检索会议（TREC 5&6）中的实验结果[5]，使用二元分词，平均检索精度是 0.4244。使用一个较大规模的词典（220k 条目）进行最大匹配分词，平均精度是 0.3907。将最大匹配切分得到的长词汇，切成单字后一并索引，可将精度提高到 0.4290。如果加入未登录词识别模块，词典分词的检索精度可达到 0.4302。从以上数据可以看出，两类方法的性能基本相当。分词的方法略好，但是需要更多的资源。使用二元分词速度快，实现简单，效率高。对于查询中句子的切分问题，我们所采取的策略为，先将句子分词，然后切分成二元组可以得到更好的检索性能。

它的工作原理用形式化语言表述如下：

字串 $C=\{C_1\ C_2\ C_3\cdots\ C_n\}$，经过二元分词处理后产生分词串为 $S=\{S_1\ S_2\ S_3\cdots\ S_n\}$，其中 $S_i=C_iC_{i+1}$。

3.2 关键词与特征词提取

有关研究表明[6]，中文信息存在一定的结构。我们通过分析用户表述搜索目标的自然语言短语，认为这些短语通常可分为两部分，即关键词和特征词。特征词可以用若干个特征词条进行描述[7]，为了让搜索引擎能够理解表述搜索目标的自然语言短语，需要建立大规模的网页特征库，即知识库。搜索引擎通过知识库获得表述搜索目标的自然语言短语所要求的网页中应包含词汇，然后根据这些词汇对网页进行检索。

我们从关键词、提问方式、提问重点三个角度对查询进行分析。这里的关键词指包含重要信息的词。提问方式是指问句的类型，如人名类、地点类等等，稍后将详细讨论。提问重点指问题的侧重点，是体现相似问题之间差异的一个重要因素，比如提问“http 协议的定义是什么?”和“http 协议的作用是什么?”，前者的重点是强调“定义”或“概念”，后者则强调“作用”或“用途”。系统通过以下 3 个步骤进行分析提取：

Step1（关键词提取） 这是最重要的步骤。本系统把查询中的名词、数词和具有动宾结构的动词作为关键词，我们为关键词设定了级别，将经过分词直接得到的名词和数词作为主关键词，分词过程中出现的具有动宾结构的动词作为次关键词。

主关键词（P_KW_i）提取，可以描述为：

Extrac－ primary － keyword ($W_1, W_2, \cdots, W_n$) :→ ($P\_KW_1, P\_KW_2, \cdots, P\_KW_m$)

次关键词（S_KW_i）提取，可以描述为：

Extrac－ secondary－keyword ($W_1, W_2, \cdots, W_n$)→ ($S\_KW_1, S\_KW_2, \cdots, S\_KW_n$)

比如：“我想买手机和电脑”一句进行关键词提取后，得到结果如下.

Keywords：手机（P_KW_1），电脑（P_KW_2），买（S_KW_1）。

如果用户的查询语句不是问句而是陈述句，则忽略第二步和第三步。

Step2（问句特征词（SW_i）提取） 如果用户的查询语句是问句。比如：“http 协议的定义是什么?”则根据系统定义的问句类型，抽取能代表某类问题的特征词。

Extrac－ semanicword ($W_1, W_2, \cdots, W_n$) :→ ($SW_1, SW_2, \cdots, SW_m$)

比如：“http 协议的定义是什么?”它的特征词是：是/什么（SW_1）

Step3 结合问句类型，按规则（同义词扩展后根据提问方式）提取问句的侧重点（EW_i）。

Extrac－emphasisword ($W_1, W_2, \cdots, W_n$):→ ($EW_1, EW_2, \cdots, EW_m$)

得到“http 协议的定义是什么?”问句侧重点为：概念（EW_1）

3.3 语义分析及查询扩展

提取出特征词后，还需要对关键词、提问方式和提问重点进行语义分析，用于形成新的检索条件。

关键词 其语义处理的目的是找出相应的同义词，并将英文词汇转换成对应的中文词汇。为此，本文设计了表 1 所示的词表，定义了与某个词相关的同义词，以及词的使用频度、词性等信息。其中词的使用频度可以在使用过程中动态调整。词表设计时不区分中英文，从而解决了汉语与英语的不同语种混和处理问题。另外由于同义词可能有多个，为便于检索并节省存储空间，我们将同义词存储为词的编号，词号之间用特殊符号隔开，如表 1 所示。

表 1 词表

词号	词	同义词号	使用频率	词性
159	电脑	159;550;669	500	n
550	计算机	159;550;669	300	n
669	computer	159;550;669	550	n
…	…			

同义词（我们讲中文转化也叫做同义词）转化形式化为：Get－ synonymy － word (KW_i) : → KW_{i1}, KW_{i2}, …,

KW_{ik})

提问方式　识别提问方式对理解提问意图有一定帮助，且对于辨别问句的细微差异起一定作用，为此，我们总结提炼出了如表2所示提问方式。

表2　提问方式表

编号	名称	疑问词
A	时间	什么时候
B	人名	谁
C	事物	是什么
D	地点	哪里
E	原因	为什么
F	方式	怎么样

提问方式的识别主要根据表2中提到的疑问词进行，可以描述如下：Get→wayofasking（SW_1，SW_2，…，SW_l）：→（Way）

比如："http协议的定义是什么？"根据表2我们得到提问方式属于C类。再根据特征词"定义"以及同义词转换得到查询的侧重点是"概念"。

陈述句　对于陈述句我们同样需要理解查询意图。为此，总结提炼出了如表3所示动词分类。

表3　动词分类表

词类名	词例
供求类	供求、供应、供货、供给、卖、售、需求、求购、买…
修理类	修理、维护、修补、补救，纠正…
…	…

比如："我想买手机。"根据表3我们知道"买"属供求类。再根据特征词"手机"以及同义词转换得到查询的侧重点是"供求类"，即关于手机的供求信息。

我们将问句和陈述句的侧重点归纳为特征向量 $SE=(E_1, E_2, \cdots, E_n)$，$E_i$ 为查询语句的特征值。

3.4　面向Web网页内容的特征库

文本的语义是基于概念之上的，而词是概念的基本构成单位[1]。无论是一篇文章还是一个句子，都表达了某些特定的信息，那么，它们应该使用特定的词来表达这些信息。例如一个介绍手机的网页中，一定包含参数名词（如资料、优点、报价）。这里的"报价"预示着网页中将包含某些特定的词，从某种意义上说是对网页的分类，它不一定出现在网页中，但反映了网页的特征，表达了网页内容的语义。这些参数名词构成了"供求"这个词的特征，是包含在网页中的表达网页内容语义的一组词汇。所谓面向Web网页内容的特征库是存储了特征词条的数据库。我们对大量的Web网页的内容进行了分析，概括出网页的特征词并从中抽取出特征向量，建立了面向Web网页内容的特征库。

4　检索及排序

4.1　基于Lucene工具建立搜索引擎

Lucene是一个基于Java的全文信息检索工具包，它不是一个完整的搜索应用程序，而是一个用Java写的全文索引引擎工具包。Lucene目前是Apache Jakarta家族中的一个开源项目，也是目前最为流行的基于Java开源全文检索工具包。

Lucene的检索方法是基于向量空间模型（Vector Space Model，VSM）。查询和文档都被表示成为向量。文档和查询之间的相似度通过向量夹角的余弦值表示，公式如下。

$$\mathrm{Sim}(q,d)=\cos(q,d)=\frac{\sum_{t\in q\wedge d}(w_{q,t}\cdot w_{d,t})}{\sqrt{\sum_{t\in q}(w_{q,t})^2\cdot\sum_{t\in d}(w_{d,t})^2}}$$

其中 $w_{q,t}=\log(N/f_t)+1$；$w_{d_t}=\log(f_{d,t})+1$；$f_{d,t}$ 是项 t 在文档 d 中出现的频率。N 是文档总数，f 是包含项 t 的文档数目。

4.2　返回文档排序

由于Lucene的基础排序算法只关注了词频和逆词频，这样虽然很好地利用了关键词在文档中的作用，但是没有体现网页的重要性，另外PageRank没有真正解决相关性即搜索词和页面的相关程度。为此，我总结提炼出新的排序算法。

公式如下：

$$Score\_d=k1\times Oldscore+K2\times PR+K3\times Simscore$$

其中 $Oldscore$ 是由基础排序算法计算出的记录文档的得分，PR 是通过PageRank算法得分，$Simscore$ 是查询特征词与文档特征信息相似程度，$K1$，$K2$，$K3$ 为权重系数。

$Oldscore$ 公式如下：

$$Oldscore=sum\_t(tf\_q\times idf\_t / norm\_q\times tf\_d\times idf\_t / norm\_dt\_t)$$

$Oldscore$：$Document(d)$ 的得分；sum_t：$Term(t)$ 的总和；tf_q：查询中 t 的频度的平方根；tf_q：d 中 t 的频度的平方根；idf_t：$\log(numDocs/docFreq\_t+1)+1.0$；$numDocs$：索引中 $Document$ 的数量；$docFreq\_t$：包含 t 的 $Document$ 的数量；$norm\_q$：$sqrt(sum\_t((tf\_q\times idf\_t)^2))$；$norm\_d\_t$：在与 t 相同域的 d 中tokens数量的平方根；

PR公式如下：

$$PR(A)=(1-d)+d(PR(1)/C(1)+\cdots+PR(n)/C(n))$$

d 为阻尼因子，一般设为0.85；$PR(i)$ 表示一个指向 A 页的网站其本身的PageRank得分；$C(i)$ 表示该页面所拥有的导出链接数量。

$Simscore$ 网页与特征词的相似度，同样也是通过向量夹角的余弦值表示。

结束语　本文阐述了在搜索引擎中实现理解查询语句的方法，以及改进返回文档的排序算法可以使网页与查询语句的相关度更精确，更好地实现对查询语句的理解。

基于语义理解的元搜索引擎能够理解自然语言短语描述的搜索目标，不需用户对目标网页进行分析，使用户从繁琐的检索规则中解脱出来；利用知识库增加了检索信息，使返回的网页更符合用户的要求，减轻了用户从大量网页中寻找目标的繁重劳动。

参考文献

[1] 杜阿宁，方滨兴，胡铭曾，等. 中文交互式网络搜索引擎及其自学习能力[J]. 计算机工程与应用，2003，39(10)：148-150

[2] 刘颖. 计算语言学[M]. 北京：清华大学出版社，2002

[3] Schubert F，Li Hui. Chinese Word Segmentation and Its Effect on Information Retrieval[D]. Information Processing & Management，2002

[4] Wu Z M，Tseng G. Chinese Text Segmentation for Text Retrieval：Achievements and Problems[J]. Journal of the American Society for Information Science，1993，44(9)：532-542

[5] Gao Jianfeng. An Empirical Study of CLIR at MSRCN，2001[C]. Shanghai：International Workshop ILT&CIP-2001 on Innovative Language and Chinese information processing Technology，2001

[6] 尤昉，李涓子，王作英. 基于《知网》的中文信息结构抽取研究[J]. 计算机工程与应用，2002，38(18)：56-58

[7] 周强，冯松岩. 构建知网关系的网状表示[J]. 中文信息学报，2000，14(6)：21-27