一种基于混合判定模型的复合概念抽取方法

欧阳柳波[1,2],邹北骥[1],刘丽杰[2]
(1. 中南大学信息科学与工程学院,湖南长沙 410083;2. 湖南大学软件学院,湖南长沙 410082)

摘 要: 从大规模领域语料库中抽取领域概念,现有方法不能有效识别复合概念.本文提出一种基于混合判定模型的复合概念抽取方法,首先对文本进行分词处理,为每个词条添加词条标签,并对词条集进行噪音词消除和同义词合并处理,然后统计词条的加权词频,根据词条标签值计算位置亲和度和位置匹配度,判定和筛选可组合成复合概念的原子词条,最后通过设置不同复合深度值,实现多重复合概念抽取.采用不同规模语料库进行抽取实验,实验结果表明本文方法具有更高的召回率和准确率.

关键词: 语料库;领域概念;复合概念;加权词频;词条标签;位置亲和度;复合深度
中图分类号: TP18 **文献标识码:** A **文章编号:** 0372-2112 (2013)03-0488-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2013.03.012

A Method of Compound Concept Extraction Based on Hybrid Judgment Model

OUYANG Liu-bo[1,2], ZOU Bei-ji[1], LIU Li-jie[2]
(1. *School of Information Science & Engineering, Central South University, Changsha, Hunan* 410083, *China*;
2. *School of Software, Hunan University, Changsha, Hunan* 410082, *China*)

Abstract: The existing methods could not identify compound concept effectively from large-scale domain corpus. This paper proposes a method of compound concept extraction based on a hybrid model. Firstly, we make segmentation processing for corpus texts and add entry label for each term. We secondly remove noise words and merge synonyms for the entry set. Then we count the weighted term frequency, the location affinity degree, the location matching degree, and make a stepwise estimation to identify composite concept with atomic terms. Ultimately we realize the extraction of multiple-compound concept via giving different compound depth. On the foundation of the extraction method, we carried out the experiments with different corpora for compound concept extraction. The results indicated the method has higher recall and precision.

Key words: corpus; domain concept; compound concept; weighted term frequency; entry label; location affinity; compound depth

1 引言

本体[1]是智能信息检索,产品信息建模,自然语言理解,语义网构建[2]和Web服务处理[3,4]等领域应用的基石,本体技术已成为当前人工智能的研究热点之一.领域本体[5]是指描述某领域知识的专门本体,它形式化地描述了领域概念之间的关系,可很好地表达领域活动及其所具有的特性和规律,是人机之间、机器和机器之间互相理解的语义基础.由于领域知识的复杂性、大规模级和多变性,利用计算机自动构建和评价领域本体有着重要意义.领域概念是领域本体的基础组成部分,如何从领域文本中准确地提取领域概念是领域本体自动构建和评价的基础工作.

本文以领域语料库,即领域文本集为研究对象,领域文本是领域知识的主要来源.领域概念是领域知识的基本单元,客观上表达一个独立、特定的语义,通过对领域文本进行分词处理和词频统计获得.由于中文句子中词间无间隙的特点,对中文文本进行有效的词切分是一个难题,相关的中文分词系统,如中科院的ICTCLAS[6],哈工大的LTP等,虽然能有效实现词切分,但其基础词库中未收录的领域新词,分词系统不能识别,导致一个领域复合概念被切分为多个词,不能满足应用需求.随着社会的发展,各领域中大量新概念层出不穷,因此对新概念识别的研究十分必要.本文将分词系统不能正确识别的领域概念统称为复合概念.

收稿日期:2012-03-21;修回日期:2012-07-25

目前新概念的识别方法主要有两种[7]:一种是基于规则的方法,基于规则的方法对新概念识别的准确率较高,但是规则定义和提取非常复杂、繁琐和多变.另一种是基于统计的方法,当前对新概念识别方法的研究大多以该方法为基础;如 Peng 等人提出一种利用条件随机域[8]的中文词法切分方法,通过计算置信度来识别新概念;Xu Sun 等人提出一种利用新词构成的规则特性和词串统计的方法[9,10],检测新概念,但对汉语成语等的识别度不高,对连续出现的长串新词无法正确切分识别;陈建超提出词序列频率有向网的中文组合词提取算法[11],根据文本信息,建立词序列有向矩阵,根据有向矩阵的统计结果,获取新概念,但抽取出的垃圾串较多;崔世起等人提出一种基于大规模语料库的新词抽取方法[7],通过垃圾串过滤机制,识别新概念,对垃圾串的识别度有提高,但基于模式串的机制对一些非模式串的新概念无法识别.

本文提出一种基于混合判定模型的复合概念抽取方法,该方法以统计方法为基础,不需要设计新概念提取规则,只需要处理分词后的重要词条的有关信息;采用加权词频统计方法衡量词条的领域相关性,同时增强噪音词处理功能,一定程度上提高了对垃圾词串的识别率;为了获取词条信息,本文定义词条标签,为每个重要词条添加语义标签、句子标签、位置标签等,用以计算词条间的位置亲和度与位置匹配度,并据此判定复合概念,取代一般的文本字符串匹配扫描方法.

2　混合判定模型

2.1　混合判定模型

领域概念的自动抽取,首先利用分词系统对领域文本进行分词处理,获得词条集,然后根据加权词频统计公式,计算各个词条的加权词频,其加权词频大于阈值的词条即可认定为领域概念.但现有分词系统对分词库中未包含的新概念不能正确识别,而是将其错误的切分成多个能识别的最大单元的原子词条,本文将这些经过加权词频筛选出来的词条统称为原子领域概念.如领域概念"领域本体构建",ICTCLAS 分词系统将其切分成"领域"、"本体"、"构建"三个原子领域概念,而我们需要的是领域概念"领域本体构建".

为了能将这些领域概念从语料库中正确自动地抽取出来,本文提出一种基于混合判定模型的复合概念抽取方法,该方法的具体思想如图 1 所示.

步骤 1　利用分词系统,对全部语料文本进行分词处理,同时对切分好的词条添加词条标签,形成原子词条库;根据噪音词表和同义词表,去除噪音词条并合并同义词条,形成新的原子词条库.

步骤 2　计算每个词条的加权词频,选取出加权词频不小于阈值的原子词条,形成原子领域概念集;仍以"领域本体构建"为例,"领域"、"本体"、"构建"三个词条的加权词频均大于取定阈值,即可认定三个词条均为原子领域概念.

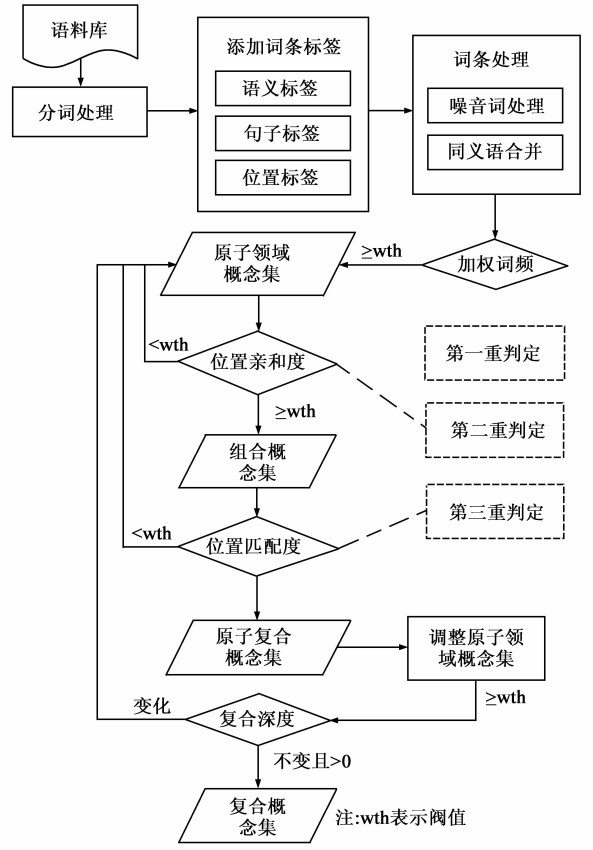

图1　复合概念混合判定模型

步骤 3　分析原子领域概念集里是否有原子领域概念可组合成组合概念.常规的方法是对原子领域概念任意组合,然后与领域文本进行逐串匹配,查看是否有这样的组合词,这种方法工作量大且复杂.本文提出一种对词条添加词条标签的方法,根据原子领域概念的句子标签,计算它们在同一句内的概率即位置亲和度,进而减少词条组合运算量.如原子领域概念"领域"、"本体"的句子标签相同的个数不小于阈值,即可认为原子领域概念"领域"和"本体"可构成组合概念"领域本体"或"本体领域";否则计算另两个原子领域概念的位置亲和度.

步骤 4　检测组合概念是否在领域文本中真实存在.本文提出位置匹配度思想,取代常规的逐串匹配方法,根据原子领域概念的位置标签,检查它们的位置是否按固定次序相邻,如果是,则可认定该组合概念在领域文本中真实存在,由其构成的组合概念可能是复合概念.如"领域"、"本体"两原子领域概念的位置标签按"领域本体"这一固定次序相邻的匹配量不小于阈值,

即可认定组合概念“领域本体”为原子复合概念,其复合深度加 1;否则计算另两个原子领域概念的位置匹配度.

步骤 5 检测原子复合概念是否可继续复合.调整原子领域概念集,将构成原子复合概念的原子领域概念从原子领域概念集中删除,并将该原子复合概念加入原子领域概念集.查看原子领域概念集中的原子领域概念的复合深度在本次循环结束后是否有变化,其中复合深度没有变化且大于 0 的原子领域概念即为本次循环得到的复合概念集,并将复合概念集从原子领域概念集中删除.

利用递归思想,继续步骤 3 ~ 5,直到原子领域概念集不能再组合出新的原子复合概念为止.如复合概念“领域本体”,与原子领域概念“构建”,重复步骤 3、4 得到新的原子复合概念“领域本体构建”,下一次循环结束后,其复合深度不再增加,则 “领域本体构建”即为我们需要的正确的领域复合概念.

从上述混合判定模型的分析可知,混合判定模型就是一种对复合概念识别过程中的多重判定,它依据加权词频、位置亲和度和位置匹配度三个层面分别降低复合概念识别的计算量.

2.2 模型的相关概念

以下相关定义中的词条可以是原子词条、原子领域概念或复合概念.

定义 1 句子标签.句子标签是对文档中的每个简单句子进行标注,进而获得每个句子的位置标识值.一个句子 t 的句子标签 Label_ $T(t)$ 用一个四元组来标识,表示如下:

$$\text{Label_}T(t)=(d,f,p,s) \tag{1}$$

其中:d 为文档编号,f 为段落类型编号,p 为段落编号,s 为句子编号.

定义 2 位置标签.位置标签是指某词条在文档中的具体位置标识,由句子标签和该词条在所属句子中出现的次序决定.一个词条 c 的位置标签 Label_ $P(c)$ 用一个六元组来标识,表示如下:

$$\text{Label_}P(c)=(\text{Label_}T(t),o_f,o_l) \tag{2}$$

其中:o_f 为构成词条 c 的第一个原子词条在句子 t 中出现的次序号,o_l 为构成词条 c 的最后一个原子词条在句子 t 中出现的次序号,那么对于原子词条 c,则有 $o_f=o_l$.

定义 3 语义标签.语义标签表示词条的词性和词义,一个词条 c 的语义标签 Label_ $S(c)$ 用一个二元组来标识,表示如下:

$$\text{Label_}S(c)=(q,m) \tag{3}$$

其中:q 为词条的词性,m 为词条的词义.相同名称的词条可能有多个语义标签,分别代表具有相同词条名称的不同概念.如“打”字,在“打毛衣”中是“编织”的意思,在“打麦子”中是“收割”意思,因此“打”具有多个语义标签.

定义 4 复合深度.复合深度是指词条 c 由原子词条进行组合的次数,用 h 表示,原子词条的复合深度为 0.如“软件设计”由“软件”和“设计”两个原子词条进行 1 次组合而成,则其复合深度为 1;“软件设计工具”由“软件”、“设计”和“工具”三个原子词条构成,进行 2 次组合构成,其复合深度为 2;依此类推.

定义 5 词条标签.一个完整的词条标签由该词条的位置标签、语义标签和复合深度组合而成,即一个词条 c 的词条标签 Label(c)用一个八元组来标识,表示如下:

$$\text{Label}(c)=(d,f,p,s,o_f,o_l,q,m,h) \tag{4}$$

定义 6 位置亲和度.位置亲和度是指两个词条在同一文档中且处于同一句子中的次数,即两个词条句子标签相同的个数.位置亲和度反映两个词条的潜在可组合程度,两个词条的位置亲和度越大,构成复合概念的可能性也越大.

两个词条 c_i、c_j,T 是包含词条 c_i、c_j 的句子 t 构成的集合,词条 c_i 和 c_j 的位置亲和度 $A(c_i,c_j)$,表示如下:

$$h(c_i,c_j)=\begin{cases}1, & \text{if } (c_i.\text{Label_}T(t)=c_j.\text{Label_}T(t))\\ 0, & \text{if } (c_i.\text{Label_}T(t)\neq c_j.\text{Label_}T(t))\end{cases} \tag{5}$$

$$A(c_i,c_j)=\sum_{t\in T}h(c_i,c_j) \tag{6}$$

定义 7 位置匹配度.位置匹配度是指位置亲和度大于某一阈值的两个词条,其位置标签按固定次序相邻的次数.位置匹配度大于某一阈值,则认为该两个词条可构成一个复合概念.

对于位置亲和度大于阈值的两个词条 c_i、c_j,T 是包含词条 c_i、c_j 的句子 t 构成的集合,词条组合 c_{ij} 的位置匹配度 $M(c_i,c_j)$,表示如下:

$$g(c_i,c_j)=\begin{cases}1, & \text{if } (c_i.\text{Label_}P(c_i).o_l\\ & =c_j.\text{Label_}P(c_j).o_f-1)\\ 0, & \text{if } (c_i.\text{Label_}P(c_i).o_l\\ & \neq c_j.\text{Label_}P(c_j).o_f-1)\end{cases} \tag{7}$$

$$M(c_i,c_j)=\sum_{t\in T}g(c_i,c_j) \tag{8}$$

定义 8 加权词频.本文采用 Robertson 提出的领域词频加权的权重计算公式 BM25F[12].

$$w_j(\bar{d},C)=\frac{(k_1'+1)\text{tf}_j}{k_1'((1-b)+b\dfrac{\text{dl}'}{\text{avdl}'})}\log\frac{N-\text{df}_j+0.5}{\text{df}_j+0.5} \tag{9}$$

其中,tf_j 指加权后的第 j 个概念在文档 $\bar{d}$ 中的词

频,dl′是加权后的文档长度,avdl′是加权后的平均文档长度,k_1' 是加权后的自由参数.df_j 是指文档频率,b 是自由参数,N 为文献集合 C 中的文档数.文献集合 C 是某特定领域的文档集合,词频系数 w_f\{题名,关键词,摘要,正文\},具体设定根据领域具体情况而定.

计算概念集合中的每个备选概念 j 在语料集合 C 中的每篇文档 $\bar{d}$ 的权重 $w_j(\bar{d}, C)$以及该概念 j 在该文档中的词频 tf_j.

$$F_j = \sum_{\bar{d} \in C} w_j(\bar{d}, C) \times tf_j \tag{10}$$

3　复合概念抽取

3.1　系统功能模块

根据混合判定模型,可将系统划分成三大核心功能模块,具体模块功能如图 2 所示.

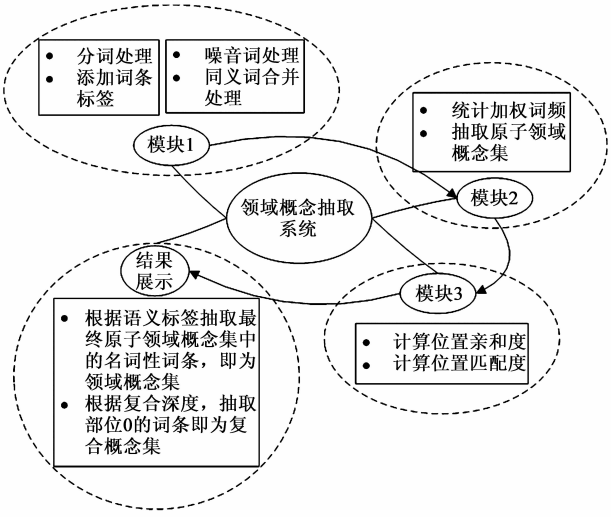

图2　系统功能模块结构图

如图 2 所示,模块 1 的主要操作有:①分词处理;②添加词条标签;③去除噪音词;④同义词合并.其中分词处理采用中科院的 ICTCLAS 分词系统实现;对语料中如"的,得,啊,呀,等"等噪音词去除,通过自行构建符合领域特征的噪音词库的方式进行处理;同义词合并通过词条标签中的语义标签进行处理.模块 1 的核心功能实现是对词条添加词条标签,详细操作过程参见算法 1.

模块 2 的主要作用是根据词条的加权词频,筛选原子领域概念集,进一步缩小原子词条库的规模量,减少计算量,具体操作过程参见算法 2.

模块 3 的主要功能是将原子领域概念集中隐含的复合概念正确地组合起来,提高领域概念抽取的精确度.复合概念的具体抽取实现算法参见算法 3 与算法 4.

抽取结果展示方式有两种:一种是根据词性展示抽取出来的全部领域概念,包含复合概念和非复合概念,因为领域概念均为名词,所以通过词性判断,只需将语义标签中标注是名词的领域概念集中的所有概念抽出即可;另一种是根据复合深度展示抽取结果,可将复合概念和非复合概念分别展示.复合概念有一个最明显的特征就是其复合深度大于 0,根据这一特征,抽取领域概念集里复合深度大于 0 的领域概念即为复合概念.

3.2　复合概念抽取算法

根据混合判定模型,复合概念抽取相关算法如下.

算法 1　词条标签算法.

在 ICTCLAS 分词系统的基础上,引入词条标签标注机制,实现在对领域文本进行分词处理的同时,对分词后的词条添加词条标签.词条标签包含语义标签、句子标签、位置标签和复合深度 4 个组成部分,分别对应不同的功能应用.其中语义标签主要应用于同义词条合并和抽取结果展示的处理;句子标签与位置标签用于加权词频、词条位置亲和度和位置匹配度的计算;复合深度用于复合概念规模控制和抽取结果展示.

算法过程:

输入:语料库中的领域文本.

①扫描领域文本,根据定义 1,对扫描到的简单句 t 添加句子标签 $Label_T(t)$;

②扫描简单句 t,根据定义 2,对扫描到的词条 c_i 添加位置标签 $Label_P(c_i)$;

③继续扫描词条 c_{i+1},如果扫描到"。"、"?"、"!"、";"、":"、" "标点符号,返回①;否则根据定义 2,对扫描到的词条 c_i 添加位置标签 $Label_P(c_{i+1})$;

④重复步骤① ~ ③步,直到扫描完整个语料库,输出带有词条标签的原子词条集$\{c_1, c_2, \cdots, c_n\}$.

根据定义 5 可知,由于词条标签的添加,使得同一词条的词条标签也不相同,于是构成了该词条的词条标签集,其具体表示为:$\{Label(c_i)_1, Label(c_i)_2, \cdots, Label(c_i)_n\}$.

算法 2　加权词频算法.

Robertson 提出的加权词频计算公式 BM25F,是典型的概率检索模型,它保证了词频不会成为衡量领域概念的唯一标准,该算法综合考虑了词频、文档频率、关键字来源字段的重要程度、文献长度等多种权重因素来确定一个词条是领域概念.经过系统模块 1 的处理后,根据原子词条库里的词条携带的词条标签信息和语料文本信息,利用 BM25F 公式,进行加权词频统计,根据语料库规模等级,取定阈值,筛选出原子领域概念集.

算法过程:

输入:原子词条集,加权词频阈值 wth.

①根据定义 8,利用词条的词条标签和语料文本信息,计算各个词条的加权词频 F_{c_i};

②如果 $F_{c_i} \geqslant$ wth,保留 c_i,否则删除 c_i;

③输出按加权词频 F_{c_i} 由大到小排序的原子领域概念集 $\{c_1, c_2, \cdots, c_n\}$.

算法 3 位置亲和度算法.

由定义 6 可知,位置亲和度算法利用两原子领域概念的句子标签,反映两个原子领域概念的潜在可组合程度,其作用是删去不可组合的原子领域概念,进而减少位置匹配度的运算量.

算法过程:

输入:原子领域概念集 $\{c_1, c_2, \cdots, c_n\}$,亲和度阈值 ath.

①获取原子领域概念 c_i、c_j,其词条标签集分别为 $\{\text{Label}(c_i)_1, \text{Label}(c_i)_2, \cdots, \text{Label}(c_i)_m\}$,$\{\text{Label}(c_j)_1, \text{Label}(c_j)_2, \cdots, \text{Label}(c_j)_n\}$;

②根据定义 6、定义 1 和两原子概念的词条标签集,获取两原子领域概念的句子标签相同的词条标签集的词条标签集,计算 $A(c_i, c_j)$;

③如果 $A(c_i, c_j) \geqslant$ ath,输出 $A(c_i, c_j)$,和句子标签相同的该两原子领域概念构成的词条标签集;否则返回步骤①.

算法 4 位置匹配度算法.

通过位置亲和度算法得到可能构成复合概念的两原子领域概念,但还要验证该组合概念是否在语料文本中真实存在,大多新概念识别算法[5~9]大都采用词串匹配统计的方法进行验证,本文则只利用可构成组合概念的两个原子领域概念的位置标签,查看两原子领域概念是否按固定次序相邻,计算其位置匹配度,即可判断组合概念是否是复合概念.

算法过程:

输入:句子标签相同的两原子领域概念构成的词条标签集,$\{\text{Label}(c_i)_1, \text{Label}(c_i)_2, \cdots, \text{Label}(c_i)_a\}$,$\{\text{Label}(c_j)_1, \text{Label}(c_j)_2, \cdots, \text{Label}(c_j)_a\}$,加权词频阈值 wth,匹配度阈值 mth.

①根据定义 2 和词条标签集,获取两原子领域概念的位置标签集;

②利用定义 7,计算 $M(c_i, c_j)$,获取位置匹配成功的词条标签集 $\{\text{Label}(c_i)_1, \text{Label}(c_i)_2, \cdots, \text{Label}(c_i)_m\}$,$\{\text{Label}(c_j)_1, \text{Label}(c_j)_2, \cdots, \text{Label}(c_j)_m\}$;

③如果 $M(c_i, c_j) \geqslant$ mth,合并原子领域概念 c_i、c_j 成复合概念 c_ic_j,将匹配的词条标签集给予复合概念 c_ic_j,根据定义 2,修改 c_ic_j 的位置标签,令 $c_ic_j.\text{Label_}P(c_ic_j).o_f = c_i.\text{Label_}P(c_i).o_f$,$c_ic_j.\text{Label_}P(c_ic_j).o_l = c_j.\text{Label_}P(c_j).o_l$,根据定义 4,修改 c_ic_j 的复合深度,令 $c_ic_j.\text{Label}(c_{ij}).h = c_i.\text{Label}(c_i).h + c_j.\text{Label}(c_j).h + 1$,形成新的词条标签集 $\{\text{Label}(c_{ij})_1, \text{Label}(c_{ij})_2, \cdots, \text{Label}(c_{ij})_m\}$,将 c_ic_j 加入到原子领域概念集;

④将词条标签集 $\{\text{Label}(c_i)_1, \text{Label}(c_i)_2, \cdots, \text{Label}(c_i)_m\}$ 从原子领域概念 c_i 的词条标签集中删除,将词条标签集 $\{\text{Label}(c_j)_1, \text{Label}(c_j)_2, \cdots, \text{Label}(c_j)_m\}$ 从原子领域概念 c_j 的词标签集中删除,重新计算原子领域概念 c_i 的加权词频值 F_{c_i},如果 $F_{c_i} \geqslant$ wth,保留原子领域概念 c_i,否则删除,c_j 同理;

⑤输出原子领域概念集中复合深度没有变化且大于 0 的原子领域概念即为复合概念;

⑥返回到位置亲和度算法,递归计算,直到不再有原子复合概念为止,循环结束.

4 实验

为了验证本文方法对复合概念提取结果的有效性和准确性,本文与近期两个具有代表性的算法进行对比实验,一个是基于词序列频率有向网的中文组合词提取算法[11],记为 $E1$;另一个是基于大规模语料库的新词检测算法[7],记为 $E2$;本文所提算法记为 $E0$.

本文采用准确率和召回率作为算法衡量标准,准确率表示抽取结果中正确识别的新概念所占的比率,召回率表示识别出的新概念是否全面.这两个指标分别从广度与深度方向综合对实验结果进行评估.它们的计算方法如下:

准确率计算公式如公式(11)所示:

$$P = \frac{C}{N} \times 100\% \tag{11}$$

其中,C 表示正确抽取的新概念数,N 表示抽取出的新概念数.

召回率计算公式如公式(12)所示:

$$R = \frac{C}{A} \times 100\% \tag{12}$$

其中,C 表示正确抽取的新概念数,A 表示语料库中所有的正确的新概念数.

4.1 实验设计

本实验从知网随机选取与软件工程领域相关的文档 200 篇,从 20 篇文档起,依次增加 20 篇文档,直到 200 篇,构建 10 个不同规模的测试语料库语,将三种算法在这 10 个语料库上进行实验.

为了实现相同环境下实验结果的有效对比分析,每种语料库规模下,三种算法的阈值取定相同,但本文提出的算法对阈值取定有三种情况,为了实现对等分析,本文将加权词频、位置度和度与位置匹配度的阈值取定与其它两种算法相同.

人工抽取新概念的代价大且效率低,但其抽取结果的准确率最高,其抽取结果可作为黄金对比标准.本文采用的语料库规模量不大,为了实现与黄金标准的对比分析,本实验对语料库进行人工抽取,在黄金标准下对三种算法的抽取结果进行对比分析.

4.2 实验结果

根据实验设计,三种算法在不同语料规模下,其新概念抽取的总数 N 和抽取结果的正确数 C,如表 1 所示:

表 1　不同语料规模下三种算法提取的新概念结果

语料规模	人工抽取	*E*0		*E*1		*E*2	
	A	*N*	*C*	*N*	*C*	*N*	*C*
20	154	151	128	201	130	96	77
40	276	278	236	372	237	166	136
60	384	396	337	534	334	224	186
80	476	482	425	679	418	272	228
100	582	584	528	851	517	325	275
120	679	681	621	1054	612	371	316
140	753	750	694	1201	684	405	347
160	814	809	758	1327	735	429	370
180	886	876	829	1469	804	461	398
200	924	917	876	1530	843	471	408

由表 1 的结果可知,*E*0 在三种规模语料库下检测到的新概念正确数是最多的,而 *E*2 在三种规模语料库下检测到的新概念正确数与总数都最少,*E*1 检测到的新概念总数最多,但其新概念检测的正确数比 *E*0 略少.

根据公式(11)与公式(12)计算其准确率 P 和召回率 R,三个算法抽取的准确率和召回率对比情况如图 3 和图 4 所示.

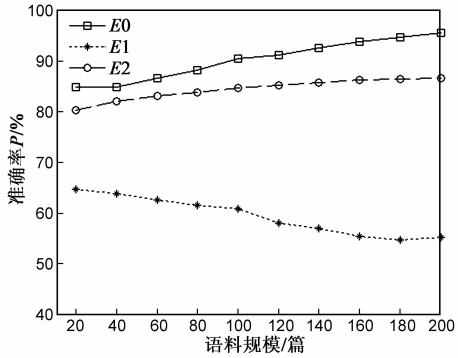

图3　不同语料规模下三种算法的准确率变化曲线

图 3、4 的实验结果,直观地展示了三种算法对新概念检测的有效性.*E*0 在每组实验中的 P 与 R 都最高,且随着语料库规模的不断升级,其 P 与 R 均有上升之势.*E*1 在每组实验中的 P 都最低,并且随着语料库规模的不断升级,其 P 不断下降,但其 R 结果比较稳定,且较高.*E*2 在每组实验中的 P 较高,但 R 最低,随着语料规模的升级,其 P 逐步提高,但 R 逐渐减小.

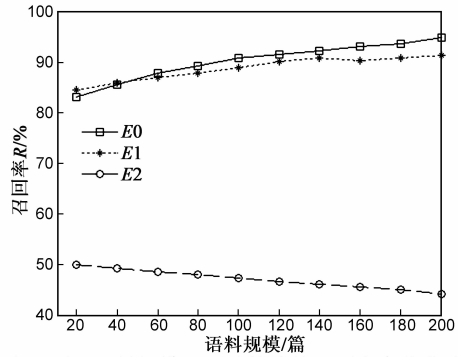

图4　不同语料规模下三种算法的召回率变化曲线

这是因为影响 *E*1 的 P 与 R 的主要因素有以下三点:①偏重于串频的统计,*E*0 与 *E*1 相比,增加了加权词频,一定程度上降低了串频的影响.②噪音词缀的影响,*E*1 没有有效的噪音词缀识别机制,使得很多满足其抽取算法但是错误的新概念被提取出来,如“用户的体验”.③抽取算法存在缺陷,比如“有向无环图”,其抽取算法抽取出“有向”、“有向无”、“有向无环图”、“向无”、“向无环图”、“无环图”等新概念,但正确的有“有向无环图”,其余五个都是错误的新概念.随着语料规模的不断升级,因素①的影响有所下降,但因素②和③的影响不断增加,而语料库文档质量下降(较口语化)对因素②和③的影响最为明显,所以其 P 有下降之势,但因为抽取出的新概念较多,所以其 R 比较稳定.

影响 *E*2 的 P 与 R 的主要因素有以下三点:①采用词性过滤机制,虽然词性过滤能有效减少数据处理量,但同时不能检测那些特殊的新概念.②基于模式的新概念检测算法,由于汉语的多样性与多变性,模式不可能完全满足新概念需求,且制定模式需要较多的工作量与理论研究.③模式串过滤采用词频统计,虽然其采用垃圾尾词典方法,比 *E*1 降低了噪音词缀的影响,也大大降低了语料库文档质量对其的影响,但其模式识别依旧采用词频统计,且只采用部分词频较高的词尾,在一定程度上限制了新概念的识别.随着语料规模的不断升级,其抽取的新概念的增加率有所提高,但远不及语料库中新概念的增加量,所以其 P 有上升之势,但是相应的 R 逐渐减小.

本文提出的 *E*0 算法的 P 与 R 主要受以下两个因素影响.①基于串频的统计,虽然采用加权词频技术,降低了其影响,但一些词频较低的新概念依旧不能被检测.②噪音词处理机制,该机制能有效地降低垃圾词缀的影响,但同时使得一些特殊的包含噪音词的新概念被拆分,不能正确抽取.但相比 *E*1 与 *E*2,*E*0 算法的 P 与 R 受到的影响明显较小,所以其 P 与 R 都最高,且随着语料库规模的不断升级,因素①的影响下降,因素

②的影响相对稳定,所以其 P 与 R 均保持上升之势.

列举系统抽取的部分复合概念,如表 2 所示:

表 2 部分复合概念结果展示

计算机软件	软件工程	数据挖掘	数据表示
数据预处理	机器学习	异常检测	人工智能
模式识别	可行性分析	需求分析	设计文档
软件版本	测试用例	测试结果	软件缺陷
数据挖掘技术	软件开发过程	软件开发阶段	代码检测
有向无环图	数据流	计算机病毒	关键字
可重用代码	可重用	文本聚类方法	标识符
数据库设计	软件行为建模	设计模式	用户体验

4.3 算法时间复杂度比较

$E1$ 算法的时间复杂度为 $O(n^2)$,$E2$ 算法的时间复杂度为 $O(n)$.本文提出的 $E0$ 算法受句子规模 n 与句子中所含的原子领域概念个数 n_c 的影响,其时间复杂度为 $O(n_c^2 n)$.

虽然 $E0$ 算法的时间复杂度比 $E2$ 算法略高,但是随着语料规模 n 的不断升级,句子中原子领域概念个数 n_c 经过噪音词等处理后变少且逐渐稳定,相对于 n 而言可视为常量,可以认为 $n_c n \approx n$,故其时间复杂度约为 $O(n_c n)$.更进一步得到,$E0$ 算法的时间复杂度趋近于 $O(n)$,则 $E0$ 与 $E2$ 算法的时间复杂度相差不大;$E1$ 算法的时间复杂度明显高于其他两种算法,且随着语料规模 n 的不断升级,差异越来越明显.

5 结论与展望

本文提出的复合概念抽取方法,使用加权词频、噪音词处理、同义词合并、添加词条标签等技术,减少重复串搜索的工作量,增加了垃圾词缀识别机制,对复合概念的抽取有较好的效果.对于大规模语料库领域概念抽取,其自动抽取算法,减少了人工工作量,且实验结果证明系统具有较高的准确率与召回率.当语料库出现更新后,无需对之前的海量级语料进行重复处理,只需利用之前的领域概念抽取结果,建立标准领域概念集,通过累加统计的方法,对新添加的语料文本进行领域概念抽取处理即可得到准确有效的结果,大大减化了工作量.本系统对语料库处理与基于语料库的本体自动构建、本体自动进化、本体领域评估等领域具有较高的应用价值.

本方法虽然在一定程度上能够准确抽取复合概念,但对于那些加权词频较高的垃圾词的识别存在问题,如表 4 中的"可重用",只是一个修饰词,并不归属于领域概念,但本系统不能将其正确识别;此外,对于一些低频的新概念,本系统的检测性能不高.在保证较高的领域概念识别的准确率与召回率的基础上,如何根据抽取出的领域概念提取领域关系是下一步研究的重点.

参考文献

[1] 李善平,尹奇▮,胡玉杰,郭鸣,付相君.本体论研究综述[J].计算机研究与发展,2004,41(7):1041 – 1052.
LI Shan-Ping, YIN Qi-wei, HU Yu-jie, GUO Ming, FU Xiang-jun. Overview of researches on ontology[J]. Journal of Computer Research and Development, 2004, 41(7): 1041 – 1052. (in Chinese)

[2] 叶育鑫,欧阳丹彤.混合语义约简和选择估值优化 SPARQL[J].电子学报,2010,38(5):1205 – 1210.
YE Yu-xin, OUYANG Dang-tong. Optimize SPARQL by combining semantic reduction and selectivity estimation[J]. Acta Electronica Sinica, 2010, 38(5): 1205 – 1210. (in Chinese)

[3] 邱田,李鹏飞,林品.一个基于概念语义近似度的 Web 服务匹配算法[J].电子学报,2009,37(2):429 – 432.
QIU Tian, LI Peng-fei, LIN Pin. A Web service matching algorithm based on semantic similarity of concepts[J]. Acta Electronica Sinica, 2009, 37(2): 429 – 432. (in Chinese)

[4] 李曼,王大治,杜小勇,王珊.基于领域本体的 Web 服务动态组合[J].计算机学报,2005,28(4):644 – 650.
LI Man, WANG Da-zhi, DU Xiao-yong, WANG Shan. Dynamic composition of Web services based on domain ontology[J]. Chinese Journal of Computers, 2005, 28(4): 644 – 650. (in Chinese)

[5] 陈刚,陆汝钤,金芝.基于领域知识重用的虚拟领域本体构造[J].软件学报,2003,14(3):350 – 355.
CHEN Gang, LU Ru-Qian, JIN Zhi. Constructing virtual domain ontologies based on domain knowledge reuse[J]. Journal of Software, 2003, 14(3): 350 – 355. (in Chinese)

[6] Huaping Zhang, Hongkui Yu, Deyi Xiong, Qun Liu. HHMM-based chinese lexical analyzer ICTCLAS [A]. Proceedings of the Second SIGHAN Workshop on Chinese Language Processing [C]. Morristown, NJ USA: Association for Computational Linguistics, 2003, 184 – 187.

[7] 崔世起,刘群,孟遥,于浩,西野文人.基于大规模语料库的新词检测[J].计算机研究与发展,2006,43(5):927 – 932.
Cui Shi-qi, Liu Qun, Meng Yao, Yu Hao, Nishino Fumihito. New word detection based on large-scale corpus[J]. Journal of Computer Research and Development, 2006, 43(5): 927 – 932. (in Chinese)

[8] Fuchun Peng, Fangfang Feng, Andrew McCallum. Chinese segmentation and new word detection using conditional random fields [A]. Proceedings of the 20th International Conference on Computational Linguistics [C]. Morristown, NJ USA: Association for Computational Linguistics, 2004. 562 – 568.

[9] Xu Sun, Yaozhong Zhang, Takuya Matsuzaki, Yoshimasa Tsuruoka, Jun'ichi Tsujii. A discriminative latent variable chinese segmenter with hybrid word/character information [A]. Pro-

ceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics [C]. Morristown, NJ USA: Association for Computational Linguistics, 2009. 56 – 64.

[10] Ruiqiang Zhang, Keiji Yasuda, Eiichiro Sumita. Chinese word segmentation and statistical machine translation [J]. ACM Transactions on Speech and Language Processing (TSLP), 2008, 5(2): 1 – 19.

[11] 陈建超,郑启伦,李庆阳,严桂夺.基于词序列频率有向网的中文组合词提取算法[J].计算机应用研究,2009,26(10):3746 – 3749.
CHEN Jian-chao, ZHENG Qi-lun, LI Qing-yang, YAN Gui-duo. Chinese combined-word detection based on directed net of word-sequence frequency [J]. Application Research of Computers, 2009, 26(10): 3746-3749. (in Chinese)

[12] Stephen Robertson, Hugo Zaragoza, Michael Taylor. Simple BM25 extension to multiple weighted fields [A]. Proceedings of the thirteenth ACM international conference on information and knowledge management (CIKM) [C]. New York, USA: ACM Press, 2004. 42 – 49.

作者简介

欧阳柳波　男,1972 年 6 月生,湖南衡阳人,博士研究生,副教授,主要研究方向为智能信息处理和软件工程.
E-mail: oylb@hnu.edu.cn

邹北骥　男,1961 年 12 月生,江西南昌人,博士,教授,博士生导师;主要研究领域:计算机视觉、图形图像处理和软件工程.
E-mail: bjzou@vip.163.com

刘丽杰　女,1988 年 10 月生,河南漯河人,硕士研究生,主要研究方向为领域本体评价和软件应用评估.
E-mail: 631993405@qq.com