基于 Web 语料的概念获取系统的研究与实现*)

余 蕾[1,2] 曹存根[1]

(中国科学院计算技术研究所 北京 100080)[1] (中国科学院研究生院 北京 100080)[2]

摘 要 互联网网页中存在大量的专业知识。如何从这些资源中获取知识已经成为 10 多年来的一个重要的研究课题。概念和概念间的关系是知识的基本组成部分,因此如何获取并验证概念,成为从文本到知识的过程中的重要步骤。本文提出并实现了一种自动从 Web 语料中获取概念的方法,该方法利用了规则、统计、上下文信息等多种方法和信息。实验结果表明,该方法达到了较好的效果。

关键词 中文信息处理,知识获取,概念获取,概念验证

Concept Extraction and Verification from Web Corpus

YU Lei[1,2] CAO Cun-Gen[1]

(Institute of Computing Science, Chinese Academy of Sciences, Beijing 100080)[1]

(Graduate School, Chinese Academy of Sciences, Beijing 100080)[2]

Abstract There is a large amount of knowledge on the Web pages. How to intelligently acquire knowledge from the massive information on Web pages has become a very important task. Concepts as well as inter-conceptual relations and inter-attribute relations of concepts are the main parts of knowledge. Therefore how toacquire and verify concepts is an important step in the knowledge acquisition. This paper proposes a hybrid approach to automatically extract concepts from large Web corpus. The hybrid approach makes use of rules, statistic, and context information to identify and verify concepts. The experiment shows very good performance of this method for extracting concepts.

Keywords Chinese information processing, Knowledge acquisition, Concept acquisition, Concept verification

1 引言

近年来,随着计算机应用和网络技术的不断发展,知识获取的来源、数量和形式也随之发生了根本的变化,一些传统的知识处理和知识组织方式面对新的应用遇到了困难,如传统知识库的构建方法已经很难适用于大型知识库的构建。越来越多的研究者开始构建基于本体的知识库。本体是一种能在语义和知识层次上描述系统的概念模型,其目的在于以一种通用的方式来获取领域中的知识,从而实现知识在不同的应用程序和组织之间的共享和重用。本体越来越广泛地应用到很多领域,如信息检索、机器翻译、知识管理、电子商务、智能教学和信息集成等。但是,本体的建立、维护和修改是一项耗时、耗力、又易错的工作。

本体由一个概念(词汇)术语集和这些概念间的关系构成,其中概念的获取是自动构建本体的重要部分。本文根据中文的特点,设计了一个适合中文的、通用的、与领域无关的自动抽取中文概念和概念的实例的算法。实验证明,该方法是可行和有效的。

1.1 概念和概念词

知识中的概念(包括常识、专业知识、新闻流通领域等方面)是用词语表示的,概念获取的本质是从词汇到概念的映射过程。概念获取的实质就是能代表概念的词汇的获取,尤其是新词新语和未释义词;概念是没有歧义的,它能够唯一地、准确地指向现实世界中的实体或对象。概念词不同于概念,它是一个能够指称概念的词。一个概念可以由多个词表示,例如同义词,而一个词也可能表示多个概念,例如多义词。概念词和概念不是等价的,只有在概念词本身没有歧义的情况下,两者才是等价的。因此概念词首先是一个词,其次它才代表某个概念。在语言处理和知识工程领域,我们可以从形式上区分概念和概念词:概念词是一个独立的词,而概念则是由一组语义相同或近似的词构成的集合。

概念获取从本质上来说是能承载概念的词汇的获取,而概念词和术语、新词语等既有区别又有联系。术语是在特定的专业领域中使用的,是一种具有很强的领域特征的词语,因而术语抽取(term extraction)的处理对象是大量的领域文本,而概念获取并不限定某个具体的领域,所以概念获取的处理对象是开放的文本语料;新词语识别(unknown word identification)的目标是那些没有收录在词典中的新词,包括专有名称、复合词、派生词和数字型的复合词(numeric-type compounds),对词典包含的已知词并不十分关注。而概念词既包括已知词,也包括一部诸如命名实体之类的未知词,但不处理诸如时间、货币、数量等数字型的复合词。

1.2 相关工作以及概念获取的困难

与概念获取比较相似的工作是术语识别,术语是在特定专业领域中一般概念的词语指称。支持术语识别的知识可分为内在知识和外在知识,内在知识一般指与术语及其构成相

*)自然科学基金(#60273019、60573064、60573063 和 60496326)和国家重点基础研究发展计划(2003CB317008 和 G1999032701)资助。

余 蕾 硕士研究生,主要研究方向为知识获取;曹存根 博士生导师,主要研究方向为人工智能。

关的词形、句法、语义或者统计信息，而外在知识一般指术语的上下文信息或者外部资源，如词典(dictionary)、同义词库(thesauri)、本体或者语料库等。术语提取的方法大致可以分为以下几类：语言学方法[1~3]、统计方法[4,5]、混合方法[6,7]以及上下文模式匹配[8,9]的方法。文[10]提出了一种混合式的考古学领域概念获取方法，文[11]提出了一种基于植物学本体获取植物名的算法，文[12]介绍了从 Web 语料中获取具有上下位关系的概念对的方法。

从汉语的特点来看，汉语缺乏形态变化，没有性、数和格的变化标志，汉语的分词效果也在一定程度上影响概念词自动识别的性能。从需要解决问题的特点分析，中文概念词获取的困难在于：

(1)对于一些在语料库中出现频度低的概念词，很难识别。

(2)由于要获取的概念词不是某个专业领域的，也不是属于某个类型的新词语(例如地名)，因此概念词边界很难确定。

(3)某些词或短语本身具有多种含义，要在一定的上下文中才能判断它所代表的含义。

2 系统的总体框架

基于概念词本身的特点，我们提出了下面一个总体框架(如图 1 所示)来获取和验证概念词。其中，获取和验证概念词的过程需要利用两种重要的信息：上下文模式和概念词构成规则。如果单纯利用人工来得到这两种信息，需要耗费很大的时间和精力，所以我们在研究利用这两种信息来进行概念词抽取验证的同时，也注意到用统计和机器学习的方式来获取这两种信息，以降低人工的干预。另一方面，我们目前用于抽取概念词的上下文模式还不是很丰富，用模式学习的方法来获取一些上下文模式，可以丰富概念词的获取模式库。

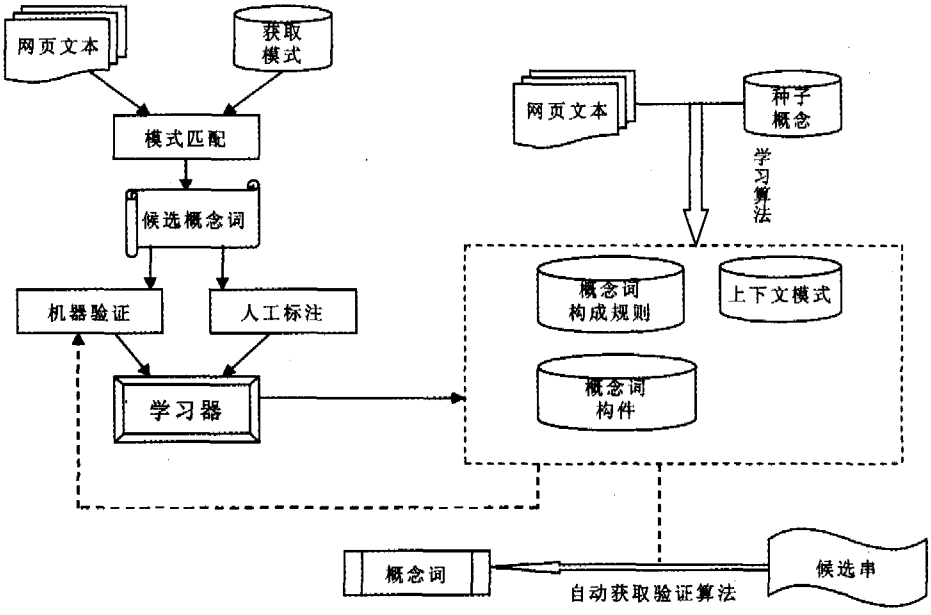

图 1 概念词自动获取和验证的总体框架

根据各个部分的功能，这个框架可以分成以下几个部分。

基础资源：为概念词抽取提供大规模真实文本的语料库、训练语料库以及概念词获取模式库，这些基础资源是获取概念词和概念词识别知识的主要来源。

学习器：具有学习概念词抽取和验证知识的功能。该学习模块以概念词库和训练语料库作为学习概念词抽取和验证知识的来源，借助中文信息处理领域的切分、词性标注等工具对语料库和概念词库进行语言分析，并利用统计等学习算法学习识别和验证概念词的知识。学习到的识别和验证概念词的知识包括：

- 概念词在真实语料中的上下文模式
- 概念词构成规则
- 概念词构件(构成概念词的词语或短语)

概念词识别和验证知识资源：包括概念词构成规则、概念词构件和概念词的上下文模式。这种资源通过学习模块通过统计的方法和自然语言处理方法从基础资源中学习到，并通过少量人工整理，放在资源库中，供抽取验证模块使用。

概念词抽取验证模块：利用概念词抽取知识资源，在语料库中通过模式匹配，二次概念词定位，概念构词规则抽取，以及开放验证等手段对概念进行抽取和验证。

3 概念词抽取和验证的基本流程

我们利用两种模式从文本中获取候选概念词，首先是在大语料库中进行模式匹配，利用上下文模式从文本中获取一个候选串，然后把候选串中可能含有的多个候选概念按照某种方式提取出来，并对这个候选概念词进行多层次的概念验证。

为了提取和验证词串中的概念，我们要综合利用概念词的 3 个特征：

(1)上下文模式特征

利用上下文模式，在第一次句型匹配得到的候选串的基础上，抽取里面含有的多个候选概念词，或者剥离概念词两边的附着成分。此外，候选串的上下文特征也可以用于概念词验证。

(2)词形-句法模式特征

利用概念的词形-句法模式特征(概念词构成规则)，可以提取出概念词并给出一定意义上的概念词可信度。

(3)概念词构件统计特征

概念词内部存在着一些概念词构件，在大语料中它表现了一种比较好的统计特征，我们利用统计的方法获取了这些类似新词语的成分，然后利用词典里已有的词和这些概念词

构件来进行概念词抽取。

句型模板(获取模式)对于概念数不确定的例句不太适用,同时也无法在句型中表现嵌套匹配。正则表达式由于考虑回溯,影响匹配速度,无法满足海量语料的处理。解决这个问题的方案是正则表达式和句型模板混合使用。一次句型匹配面向原始语料,使用句型模板在语料中匹配;概念词界定过程用分割词把匹配的字符串分隔成句块;然后在句块中用正则表达式表示的概念词构成规则对其中的概念词进行抽取。最后利用混合的方法对概念词进行选择评价和验证。对于那些比较有把握的概念词,我们把获取概念词构件的算法应用在这些词串上,获取更多的概念词构件来提高概念词抽取的查准率和查全率(如图 2 所示)。

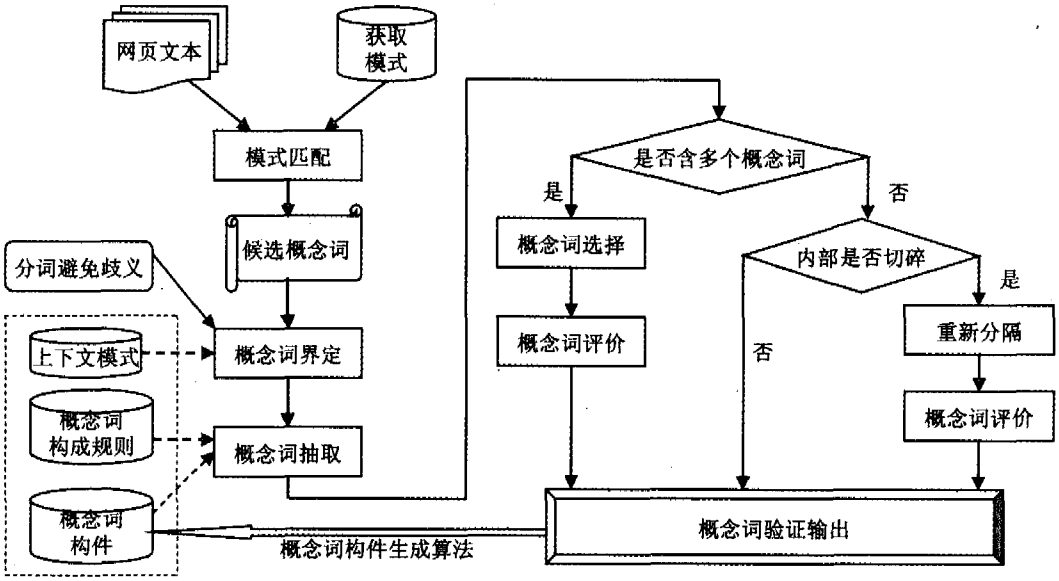

图 2　概念词抽取和验证的基本流程

3.1　模式匹配

从获取模式库中选取模式,在语料库中以句子为单位进行匹配,找到符合模式的句子,进行模式标记,然后从模式匹配句子中抽取出可能蕴含概念的部分,作为候选串,用于下一步的概念抽取验证。候选串有可能利用了概念左边界或右边界信息也可能同时利用了这两种信息。由于最终程序利用的匹配句型(上下文模式)只有一个,匹配得到的词串往往还不准确,它往往会出现以下三种情况:

(1)多个概念词用一些连接成分隔开

例如候选串"其主要原料是牡蛎",其中有两个概念词:"主要原料"、"牡蛎";

(2)真正的概念词两边有一些附着成分

例如候选串"其实是苏联军事情报局",其中有一个概念词"苏联军事情报局";

(3)这个词串根本不是概念词

例如候选串"有时她会哈哈笑",其中没有概念词。

3.2　概念词界定

概念词界定的主要任务是用分割词或分割符把候选串分隔成句块。我们对候选串分词后的词语查找分割词词典,目的是防止某些单字分割词是词语的一部分。例如,"中"、"的"属于分割词词典,如果经过分词,"中国近代史博物馆"不会被错误地分隔为"国近代史博物馆",但"中国近代史博物馆中的文物"就会被分隔为"中国近代史博物馆"和"文物"。

算法　分隔句块的算法

输入:候选串

输出:若干句块

Step 1. 对候选串进行分词,并作简单的标注;

Step 2. 在分词的基础上,对每一个词查找分割词词典;

Step 3. 把这个候选串按照分割词词典分隔为若干句块;

3.2.1　分词模块

分词是实现概念词提取的基础。现在我们给出切分规则。对下列特殊字符或字符串进行切分:

·句末点号和句中点号。句末点号包括句号、问号、叹号。句中点号包括逗号、顿号、分号、冒号。

·汉字字符串:由若干个连续的汉字字符构成的字符串。

·非汉字字符串:由若干个连续的非汉字字符构成的字符串。

需要指出的是,对下列两种情形,我们不做切分处理:

(1)对满足标点符号启发规则的字符串不做切分。例如,对于规则

Left(〈S〉,〈"〉)(Right(〈S〉,〈"〉)→NoSeg(S)

表示对于字符串 S,若 S 位于引号之间,则 S 不做切分,其中 Left(〈S〉,〈"〉)表示 S 左邻引号",Right(〈S〉,〈"〉)表示 S 右邻引号",NoSeg(S)表示 S 不做切分。

(2)若字符串 S 位于小括号、或中括号、或大括号、或单引号、或双引号之间,则 S 不做切分。

我们采用最大向前匹配法进行分词。词语 W 的词性标注为 W 在词典中的词性集合,可能为一个或多个。需要说明的是,算法仅利用 W 的可能词性集合来辅助提取候选概念词,并不对 W 在句子中的词性进行辨别。

3.2.2　句块分隔模块

分割词是指标点符号(包括全角和半角)、虚词和一些用上下文学习算法获取的一些常见概念上下文(一般是一些多字虚词和构词能力比较差的动词)。算法的过程可以用下面的有限状态自动机来表示(如图 3 所示)。自动机的状态含义和输入字符含义分别如表 1 和表 2 所示。

在找句块的左边界词时,判断依据一个分割词与非分割词左右相连,则认为是一个句块的左边界词。如果多个分割词连接出现,则需要找到最后一个与分非分割词相连的词;在找句块的右边界词时,判断依据一个分割词与非分割词右左

相连，则认为是一个句块的右边界词。如果多个分割词连接出现，则需要找到第一个与分非分割词相连的词。对候选串按分割词分隔后，候选串被分为一个或多个句块。

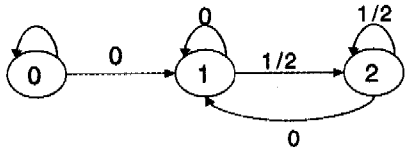

图 3　句块分隔自动机

表 1　状态含义

状态	含义
0	初始状态
1	表示发现了句块的左边界
2	表示发现了句块的右边界

表 2　输入字符含义

输入字符	含义
0	表示该字符不是分割词
1	表示该字符是分割词
2	表示该字符是字符终结符

3.3　概念词抽取

在上一步，我们把候选串用一些分割词和分割符号分隔成了多个句块，这样每个句块中含有的概念个数一般都不会超过一个。所以接下来的任务就是从句块中把概念抽取出来并对概念做出评价。具体的方法是，利用学习器获取的概念词构成规则（主要是词性规则）在句块中抽取匹配的概念词，并对概念词做出评价。由于规则很多，规则用正则表达式表示，能够降低规则的数目。所有的规则放在一个文本文件中，用数字给出它的准确率和优先级，用户可以根据应用的不同领域动态增删规则。

定义（概念置信度）　广义上讲“置信”指的是正确的概率，“置信度”是评价这个概率的一种量度，概念置信度表示评价概念词能够指代概念的可靠程度。

3.3.1　正则表达式

表 3　正则表达式中的字符说明

字符	说明	举例
+	一次或多次匹配前面的字符	“an+”与“an”和“ann”匹配，但与“a”不匹配
*	零次或多次匹配前面的字符或子表达式	an* 匹配“a”和“ann”
|	左右两边表达式之间“或”关系，匹配左边或者右边	“n|ann”与“n”或“ann”匹配 “(n|a)nn”与“nnn”或“ann”匹配
()	标记一个子表达式的开始和结束位置	(a|b)n 与“an”或“bn”匹配

正则表达式提供了功能强大、灵活而又高效的方法来处理文本。正则表达式的模式匹配法可以快速地分析大量的文本以找到特定的字符模式；提取、编辑、替换或删除字符串。正则表达式是从左向右去匹配目标字符串的一组模式。大多数字符在模式中表示它们自身并匹配目标中相应的字符。举个最简单的例子，模式“ann*”匹配了目标字符串“vbanv”中与其匹配的一部分“an”。

由于正则表达式表达能力强，我们实现了一个正则表达式匹配的程序，除了我们规则中需要字符集外，这个正则表达式程序支持以下几种扩展的字符，表 3 给出了这几种字符的说明和具体的例子。

使用的规则是学习器得到的概念词构成规则，每个规则第一项是词性构成规则的正则表达式，第二项是它的精确率，第三项是规则的优先级；表 4 给出所使用规则的部分示例。

表 4　用于抽取概念词的规则（部分）

规则	精确率	优先级
n+	0.98	90
(h|a|b)nn*	0.99	100
vv*knn*	0.99	99
nn*vnn*	0.85	100

本文用到的词性符号的说明：n 名词、b 区别词、v 动词、a 形容词、h 前接成分、k 后接成分、g 语素、q 量词 、f 方位词、d 副词、m 数词。

3.3.2　概念词的抽取和选择

顺次匹配规则数组，找到与规则匹配的序列提取对应的概念词，并对提取的概念词按照冲突消解原则进行选择。如果发现句块中提取的概念词在分词的过程中被切碎，重新对句块进行定界和抽取。

算法　概念词的抽取和选择算法

输入：概念词构词规则集、句块

输出：抽取的概念词

Step 1. 找到句块的第一个词性序列 L（GetFindFirstPOSLis）；

Step 2. 从规则集中匹配并提取对应的概念词（MatchAmongRules(L)）；

Step 3. 如果成功，转步骤 5，否则，下一步；

Step 4. 取得句块的下一个词性序列 L（GetNextPOSLis），若 L 不为空，到步骤 2；若 L 为空，算法结束；

Step 5. 候选概念词处理（ConceptSelection）。

对于同一个规则，只找出一个最长的匹配。下面给出了一个示例（如表 5 所示）。

表 5　概念词抽取示例

句块	北京/天安门/也许
句块词性序列	nnd
匹配的规则	nn*
匹配最长的符合规则的词性序列	nn
提取的概念词	北京/天安门

规则匹配冲突消解原则：

Rule1：存在相交关系，合并这两个序列和对应概念串，并认为这个合并的词性序列可能是一条新的规则；

Rule2：存在包含关系，保留最长的匹配概念串；

Rule3：不存在相交关系和包含关系，对应的多个概念串都保留。

3.4　概念词验证

有时通过一个概念的内部构词特征并不能确定某些候选概念是否是概念，或者本身一个短语就存在歧义，这时就需要结合概念的上下文信息和词间的关联度来对候选概念进行验证。如果某些语言单位如果经常出现在概念词的上下文环境中，那么这些语言单位可以作为概念词的左右边界，为上下文中概念词的鉴定提供依据。借助于大规模语料库和概念词

库，可以获得围绕概念词各方面的统计特征，从而为概念词的判定与鉴别提供依据。概括说来，概念的上下文特征有以下两种。

·边界特征：有界定符(《》、“”、‘’)，界定词(“的”，“是”)，动词驱动词等出现；

·频率特征：作为候选概念的次数，即出现在多个符合模式的句子中。

定义(概念支持度) 衡量开放语料库对确认候选概念是概念的证据的量，以(概念出现在限定上下文的频度/语料大小)来衡量。

算法 基于模式的概念验证的算法

输入：待验证概念集，验证模式集合

输出：已验证概念及验证过程中匹配的模式

Step 1. 把语料库按预先定义的大小分块，这样的目的是如果在一定大小的语料中就达到了可以验证候选概念 C 是概念就没有必要验证程序跑完整的上 T 的语料。如果没有达到，继续增量验证。

Step 2. 从待验证概念集中读一个概念 C。

Step 3. 把语料库中取出一块用于验证。

Step 4. 获取概念 C 的上下文模式信息并存入概念 C 的上下文模式数组中，获取出现的网页信息和词频。

Step 5. 计算概念 C 上下文模式数组中在上下文模式库出现的个数 count，并根据词频计算指标概念支持度 support(C)，如果 support(C)＞阈值 T，则把概念 C 的上下文模式数组中没有在上下文模式库中出现的模式补充到候选上下文模式中；返回 step2。

Step 6. 第 4 步未能验证转 step3 继续验证。

4 实验结果分析

基于上面提到的概念获取框架，我们在 VC^{++} 环境下实现了一个概念获取系统(Concept Acquisition and Verification System)。

4.1 概念获取系统评价方法

我们选用查准率(Precision)、查全率(Recall)和未登录概念词识别比率(Recognition-rate)作为实验结果的评价方法。

首先定义如下参数。

·N_{ks}：由概念获取验证系统提取和知识工程师都提取，并经知识工程师确认的概念词个数；

·N_k：由知识工程师提取和确认的概念词个数；

·N_s：由概念获取系统提取的概念词个数；

·N_c：由概念系统和分词系统都提取的经过确认的概念词个数；

·N_w：由分词系统识别的经过确认的概念词个数。

查准率、查全率和未登录概念识别比率分别定义如下：

$$\text{Precision}=\frac{N_{ks}}{N_s},\ \text{Recall}=\frac{N_{ks}}{N_k},\ \text{Recognition-rate}=\frac{N_w-N_c}{N_{ks}-N_c}$$

未登录概念识别比率是反映概念获取系统和分词系统在识别未登录概念词的性能。这个比值越小，说明概念获取系统识别未登录概念词的个数越多，性能越好。

4.2 实验评估与分析

我们以一个 400M 的 Web 语料为基础资源，进行概念获取和验证。提取约 410296 个概念词。对语料分组随机抽取进行了评估：平均查准率为 78.6%，查全率为 83.0%。

我们将北京大学计算语言研究所的分词系统[13]和本文提出的概念获取系统的实验结果进行了比较。以随机抽取的五十个候选串为例，分词系统提取了 5 个正确概念词，概念获取系统提取了 44 个正确概念词，10 个错误概念词，其中概念获取系统比分词系统多提取了 40 个未登录概念词。概念获取系统的正确率为 81.5%，未登录概念词识别比率为 2.5%。

4.2.1 试验结果示例

句块分隔结果：

阿富汗塔利班最高领导人奥马尔已经躲藏起来

句块—阿富汗塔利班最高领导人奥马尔躲藏起来

海龙大厦，东临中关村大街，北四环的新开盘的鼎好电子商城则可看做一个集大成者。

句块—海龙大厦 东 中关村大街 北四环 新开盘 鼎好电子商城 看做一个集大成者

下面是 2 组概念词抽取验证结果，其中第一行是候选串，其下是从中提取出的概念词，输出格式是概念词、概念词的构成模式、句块的词性序列、匹配的模式、置信度：

阿富汗塔利班最高领导人奥马尔已经躲藏起来

阿富汗塔利班最高领导人奥马尔 nnnnangnn nnnnangnn nn*ann*gg*nn* 0.97

海龙大厦，东临中关村大街，北四环的新开盘的鼎好电子商城则可看做一个集大成者。

海龙大厦 nn nn nn* 0.99

中关村大街 nn nn nn* 0.99

北四环 ggn ggn gg*nn* 0.97

新开盘 nv nv nn*vv* 0.85

鼎好电子商城 nann nann nn*ann* 0.95

集大成者 vnk vmqvnk vnn*k 0.90

4.2.2 错误分析

(1)切分歧义造成的错误

阿根廷人为“屠夫

阿根廷 n nb nn* 0.99

屠夫 n n nn* 0.99

在分词过程中由于切分歧义这个串被切分为“阿根廷/人为/“/屠夫”，因此“阿根廷人”未能正确提取。

(2)专业术语结构过分复杂造成的错误

动力分散内燃液传摆式列车

燃液传摆式列车 gnnnnn nafgnnnnn gg*nn* 0.97

这个串切分并标注后的结构为“动力/n 分散/a 内/f 燃/g 液/n 传/n 摆/n 式/n 列车/n ”，在构成模式中没有可以完整匹配的模式，所以只提取了其中的一部分。这种错误会随着概念词构件的增加而逐渐减少。

总结以及将来的工作 本文提出并实现了一个概念抽取和验证的统一框架。利用学习算法得到的概念词构成规则、上下文模式以及概念词构件，综合了规则、句型、正则模式、统计等方法，该框架均衡了效率和各项指标，达到了较好的效果。

虽然本文的研究基本达到预期目标，在以后的研究中，还需要在以下几个方面作进一步的研究：

(1)对于专名，它们有一些独特的特征可循，例如中英文

(下转第 195 页)

的任意形状，并要求获得整个数据集的全局信息，这在内存有限情况下对动态数据集进行聚类是难以适用的。如何在有限内存情况下对动态数据集进行增量式聚类是我们下一步研究工作的方向。

参 考 文 献

1 Ester M, Kriegel H -P, Sander J, et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: Proc. 2nd Int Conf on Knowledge Discovery and Data Mining, Portland, OR, 1996. 226～231

2 Han Jiawei, Kamber M, Fan Ming, et al. 数据挖掘：概念与技术. 北京：机械工业出版社，2001

3 Ankerst M, Breunig M, Kriegel H-P, et al. OPTICS: Ordering Points To Identify the Clustering Structure. In: Proc. ACM SIGMOD'99, Int Conf. on Management of Data, Philadelphia, PA, 1999

4 Breunig M M, Kriegel H-P, Ng R T, et al. LOF: identifying density-based local outliers. In: Proc. ACM SIGMOD 2000 Int Conf on Management of Data, Dalles, TX, 2000

5 Tang Jian, Chen Zhixiang, Ada Wai-chee Fu, et al. A Robust Outlier Detection Scheme for Large Data Sets. In: http://www.cs.panam.edu/～chen/papers.html

6 Zhou Yong-Feng, Liu Qing-Bao, Deng Su, et al. An Incremental Outlier Factor Based Clustering Algorithm. In: the First International Conference on Machine Learning and Cybernetics, Nov2002, CHINA

7 Jin Wen, Tung A K H, Han Jiawei. Mining Top-n Local Outliers in Large Databases. In: Proc. ACM KDD 2001, San Francisco, California USA

（上接第 156 页）

21 Sion R. Query Execution Assurance for Outsourced Databases. In: Proceedings of the 31st VLDB Conference, Trondheim, Norway, 2005

22 Khanna S, Zane F. Watermarking maps: hiding information in structured data. SODA, 2000. 596～605

23 Thruaisingham B. Recursion Theoretic Properties of the Inference Problem in Database Security [R]. In: MTP291. MITRE Corp, Bedford, Mass, 1990. 21～33

24 Itoh T. On Lower Bounds for the Communication Complexity of Private Information Retrieval. IEICE Transactions, 2001, E84- A (1)

25 Su T, Ozsoyoglu G. Controlling FD and MVD Inferences in Multilevel Relational Database System [J]. IEEE Transactions on Knowledge and Data Engineering, 1991, 3(4): 474～485

26 Ng W, Lau Ho-Lam. Effective Approaches for Watermarking XML Data. In: Proc. of DASFAA, LNCS, 2005, 3453: 68～80

27 WMDB System Architecture. http://www.cs.stonybrook.edu/～sion/projects/wmdb, 2004

28 王正飞，王曼，汪卫，等. 数据库中加密字符数据的存储与查询 [J]. 计算机研究与发展，2004, 41(Suppl. 10): 66～71

29 王晓峰，王尚平. 秘密同态技术在数据库安全中的应用 [J]. 计算机工程与应用，2003, 14: 194～196

30 Li Yingjiu, Guo Huiping, Jajodia S. Tamper Detection and Localization for Categorical Data Using Fragile Watermarks. In: Proceeding of the DRM'04, October 25, 2004, Washington, DC, USA

31 Li Yingjiu, Swarup V. Fingerprinting Relational Databases - Schemes and Specialties. IEEE Transaction on Dependable and Secure Computing, 2005, 2(1): 34～45

32 张敏，徐震，冯登国. 数据库安全 [M]. 北京：科学出版社，2005. 163～169

33 朱虹，史凌云，张勇. 多级安全数据库系统推理问题研究 [J]. 计算机工程与应用，2004, 13: 179～181

（上接第 165 页）

人名的组字相对来说比较稳定，地名和机构名也有比较好的后缀特征可以利用。如果把概念词抽取中把专名识别独立地作一个模块，利用其中独有的一些规律，对专名进行推测和验证，系统的准确率会更高。

（2）概念词内部语义关系的获取：分析复合概念词与构成它的基本概念词之间的语义关系，从而为领域 Ontology 的辅助构建、概念词辅助定义、概念词的辅助翻译提供支持。

（3）概念词关系的获取：概念之间存在着上下位关系、部分关系、同指关系、同义关系、反义关系等关系；利用概念词内部的语义信息和语料中的分布特点，可以进一步获取概念词间的关系。

参 考 文 献

1 Bourigault D. Surface Gramatical Analysis for the Extraction of Terminological Noun Phrases. In: Proceedings of COLING 92. 977～981

2 Frantzi T K. Incorporating Context Information for the Extraction of Terms. In: Preceedings of ACLEACL' 97

3 Wu shih-hung, Hsu wen -Lian. A semi-automatic domain ontology acquisition tool from Chinese Corpus [C]. In: Proc. of the 19th International Conference on Computational Linguistics (COLING)2002, Taipei, Taiwan, 2002. 1313～1317

4 Dunning T. Accurate Methods for the Statistical of Surprise and Coincidence. Association for Computational Linguistics, 1993, 19 (1): 61～76

5 Pantel P, Lin Dekang. A Statistical Corpus-based Term Extractor. In: Canadian Conference on AI, 2001. 36～46

6 Enguehard C, Pantera L. Automatic Natural Acquistion of a Terminology. Journal of Quantitative Linguistics, 1994, 2(1): 27～32

7 Luo S F, Sun M S. Two-Character Chinese Word Extraction Based on Hybrid of Internal and Contextual Measures. In: Proceeding of ACL2003, Sapporo, Japan, 2003

8 Riloff E. Automatically constructing a dictionary for information extraction tasks. In: Proceedings of the Eleventh National Conference on Artificial Intelligence, 1993. 811～816

9 郑家恒，杜永萍，宋礼鹏. 农业病虫害词汇获取方法初探. 见：孙茂松，陈群秀. 语言计算与基于内容的文本处理. 北京：清华大学出版社，2003. 61～66

10 张春霞. 领域文本知识获取方法研究及其在考古领域中的应用：[博士论文]. 北京：中科院计算所，2005

11 罗贝，吴洁，曹存根，等. 从文本中获取植物知识方法的研究. 计算机科学，2005, 32(10): 6～13

12 刘磊，曹存根. 一种基于"是一个"模式的下位概念获取方法. 计算机科学(已录用). 2006

13 http://icl.pku.edu.cn/icl%5Fres/segtag98/