中文叙词表本体概念间等同关系自动构建研究

方　圆[1]，曾新红[2,1]
（1.深圳大学计算机与软件学院　深圳　518060）
（2.深圳大学图书馆　深圳　518060）

摘　要　中文叙词表本体（OntoThesaurus）是一种新型的、同时具备叙词表和本体特征的知识组织系统。其配套系统“中文叙词表本体共建共享系统”（OTCSS）采用的共建方法是依靠一种机器辅助的人工构建方法（类似于WIKI）。这种方式的好处是修订意见准确率高、修订内容完整等，但是也存在修订信息时效性不稳定的缺点，其成功还需要依赖网络用户的大量参与。本文提出了中文叙词表本体中等同关系的自动构建方案。其中，基于web搜索引擎的模式匹配算法，对OntoThesaurus叙词概念的覆盖率达到91%；基于《知网》义原词频的词汇等同关系识别算法，比原算法在准确性上提高了8%，达到了78%的准确率

关键词　叙词表；本体；OntoThesaurus；等同关系；自动构建；《知网》；相似度；模式匹配
中图分类号　TP18；TP 301.6　　**文献标识码**　A

Research on automatic building the equivalent relationship between the concepts in OntoThesaurus

Fang Yuan[1], Zeng Xin-Hong[2,1]
(1. College of Computer and Software, Shenzhen university, Shenzhen 518060, China)
(2. Shenzhen University Library, Shenzhen 518060, China)

Abstract　OntoThesaurus (Chinese-thesaurus-ontology) is a new KOS, and has the characteristics of both thesaurus and ontology. The supporting system OTCSS (OntoThesaurus Co-construction and Sharing System) adopts an WIKI-like co-constructing method. The advantages of the approach are revisions with high accuracy and integrated amendments. But its timeliness is uncertain, and its success relies on a large number of Internet users to participate. This article proposes the scheme of automatic building the equivalent relationship between the concepts in OntoThesaurus, which proposes the pattern matching algorithm based on Web search engine, making the coverage rate of OntoThesaurus concepts reach to 91%; and proposes the recognition algorithm for equivalent relationship between words based on the sememe word frequency of "HowNet", which is increased by 8% on the accuracy, to 78%.

Keywords　Thesauri; HowNet; similarity; pattern match; ontology; OntoThesaurus; the equivalent relationship

0　引　言

面对网络化、数字化、海量分布、复杂的信息资源，当前的检索技术还不能有效地揭示和发现信息之间内在的知识联系，不能帮助用户迅速及时地找到自己所需的知识。因此可以有效组织、高效检索的受控语言（叙词表，分类法，规范文档）便再一次成为大众的焦点，网络知识组织系统（NKOS）便是在这样一种环境下应运而生。

在NKOS的各种应用中，叙词表和本体是最复杂、控制程度最高的应用，已经成为该领域人们研究的热点[1]。本体是“共享概念模型的明确的形式化规范说明”（Borst，1997）[2]，它形式化地定义了领域内共同认可的知识；叙词表在信息检索领域中的定义是“一部受控标引语言的词汇集，它组织方式规范，使得

词汇之间的推理关系(比如上位和下位关系)明确化[3]"。

中文叙词表本体(OntoThesaurus)是叙词表与本体的融合,即一种新型的、同时具备叙词表和本体特征的知识组织系统。其配套系统"中文叙词表本体共建共享系统"(OTCSS)采用的共建方法是依靠一种机器辅助的人工构建方法(类似于 WIKI),标引员、领域专家和一般网络用户可以在使用 OTCSS 提供的网络术语学服务(TS)的同时,对增加同义词、新概念、新关系种类等提出自己的修订意见,修订专家可利用这些意见对词表进行网上修订,实现其动态更新。这种方式的好处是修订意见准确率高、修订内容完整等,但是也存在修订信息时效性不稳定的缺点,其成功还需要依赖网络用户的大量参与。因此我们迫切希望找到一种自动或半自动的方法来完成概念间关系的构建,从一定程度上减轻修订专家的负担。

本文提出了中文叙词表本体中等同关系的自动构建方案,从更新速度和准确性两个方面缓解上述问题。

1 研究背景及方案简介

1.1 概念间关系自动识别研究现状

等同关系识别主要是研究概念的同义词,以及同义词间的相似性问题;对于中文叙词表本体而言,同义词控制质量决定着相同主题概念标引的一致性,文献信息的查准率和同义词扩展检索的查全率。目前国内外对于概念间关系识别(包括等同关系识别),常用方法有以下四种:

基于模式匹配的方法[4]。通过对大规模的文本信息进行归纳分析,找出常见的句法或者语法模式,并将其作为规则,然后分析待测文本中的语句是否符合某种规则类型,如果符合则自动识别这种关系。比如给定一种同义关系的规则"A 也称为 B",那么对于语句"减压病也称为潜水员病",系统可以自动识别"减压病"的一种同义关系为"潜水员病"。某些模式是手工定义的[5],而另外一些是从样本句子或者辞典中学习得到的[6]。

基于词聚类的方法。在同一词簇中,词与词的语义距离比较小,可以利用词的相似性,计算出词间语义距离,对词进行聚类(如层次聚类,就是利用词间上下位关系把目标数据放入少数相对同源的组或"类"(cluster)里)。基于层次聚类自动构建概念间关系的研究比较多,如 Maedche[7]、Aguirre[8]、Khan[9]、Bisson[10]等都提出相关理论。国内对层次聚类的研究也很多,仲云云等[11]利用相似词可替换性得到词族并找出其等级关系。

基于词汇共现模型的方法。词汇共现模型是基于统计方法的自然语言处理研究领域的重要模型之一,根据共现模型,某些经常出现在同一窗口(包括段、句等)的词汇,在一定程度上表明它们之间存在较大相关性,它常用于构建词间非分类关系。曹恬[12]结合 VSM 和词共现方式找到了一种计算文本相似度的方法;Maedche[7]认为词汇在上下文中共现的现象可以使用关联规则方式表达,通过建立关联规则库在 Text-To-Onto 中发现概念间的关系。

基于词典或词汇分类体系方法。词典或词汇分类体系有着良好的结构和大量的语义信息,据此可以使用基于词典的启发式模式匹配[7]来构建概念间关系。国内的相关研究可以归纳为基于"WordNet"[13]、"知网(HowNet)"的词汇语义相似度算法[14]和基于"同义词词林"[15]、"中文概念词典"[16]等的词汇语义相似度算法。

以上方法是当前本体自动构建领域的研究热点,但是它们只是实现构建系统的核心思路,在实际建设过程中,还存在许多细节问题亟待解决。实际上,本体自动构建技术的研究还处于早期阶段,各种算法还很不完善,特别是对于跨领域、非结构化文档的构建技术和本体构建本地化研究刚刚萌芽。

1.2 中文叙词表本体等同关系自动构建方案

我们在已有本体自动构建技术及其实践经验的基础上,根据中文的特点与项目要求,提出了中文叙词表本体等同关系的自动构建方案,体系结构如图 1。

从图 1 可以看出,本文的等同关系自动构建方案可以分为两个步骤:

1) 数据准备与预处理:从中文叙词表本体(如 CCT1_OntoThesaurus)中提取主题词,并利用网络爬虫找到匹配模式;然后利用模式匹配算法在 Web 中得到同义词信息片段;同时,利用英文翻译推导算法,得到同义信息;最后利用同义词推导算法,扩展同义信息检索范围。

2) 候选同义词评价和筛选:应用候选同义词相似度的评价方式对候选同义词进行评价和筛选,并将筛选出的候选词与评价一起发送给 OTCSS 系统,由修订专家最终确认。

该方案的创新点主要有以下两点:

1）提出基于Web搜索引擎的模式匹配算法，较大幅度地提高了抽取同义词的数量。

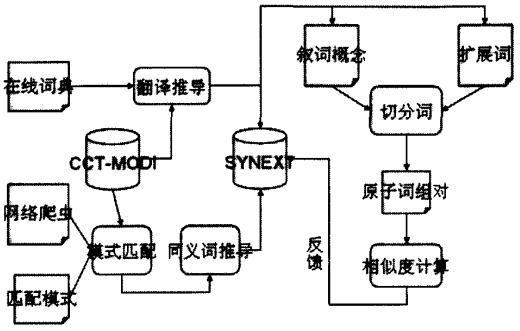

图1 中文叙词表本体的等同关系自动构建体系结构

Fig.1 the architecture of automatic building the equivalent relationship in OntoThesaurus

本文采用网络爬虫结合模式匹配算法在Web上抽取同义关系，并使用种子学习算法寻找匹配模式，使用同义推导扩展同义关系搜索范围。实验证明，新算法对OntoThesaurus叙词概念的覆盖率达到91%，剩下没有被覆盖的词汇大部分属于合成词和生僻词。

2）提出基于《知网》义原词频的词汇等同关系识别算法，在评价和筛选阶段较大幅度地提高了同义词识别的准确率。

原有算法存在若干缺陷：1）没有考虑义原在义原树中的绝对位置；2）第一独立义原权重过大；3）每种义原的强弱程度没有考虑。这些缺点限制了原算法的应用范围和准确性，本文提出的算法可以在一定程度上修正这些缺点，实验证明，新算法比原算法在准确性上提高了8%，达到了78%的准确率。

2 基于Web搜索引擎的模式匹配算法

信息抽取领域中，模式匹配是指通过对大规模文本信息进行归纳分析，找出频繁出现的句法和语法模式，将其作为规则，然后分析待测文本中的语句是否符合某种模式，如果符合则自动识别这种关系。

例如，给定一种同义关系的规则“A也称为B”，那么对于语句“减压病也称为潜水员病”，系统可以自动识别“减压病”的一种同义关系为“潜水员病”。

模式匹配中所定义的关系规则被称为匹配模式，关系两端的实体部分被分别称为匹配词和被匹配词。在上例中，“也称为”是匹配模式，“减压病”是匹配词，而“潜水员病”则是被匹配词。

经典的模式匹配算法分为语料选择、词汇定义、模式获取、模式表示、信息抽取和模式性能评价六个步骤。图2说明了这种算法的具体流程。

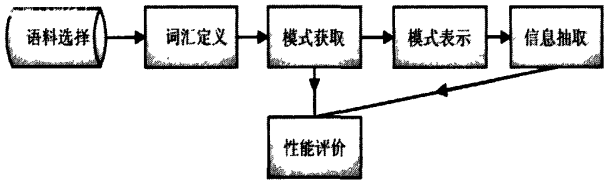

图2 模式匹配算法流程

Fig.2 Pattern matching algorithm processes

模式匹配算法的核心部分是模式获取和信息抽取两个部分。其中，模式获取是指获取在语料中重现度较高的匹配模式；信息抽取是指抽取出用户需要的关系集合。

2.1 基于种子学习的半自动产生提取模式

在Snowball系统[17]中，对于用模式提取的种子句，需要得到种子句的句型结构。针对本文识别同义词关系的要求，举例如下：

[减压病/n][也称为/v][潜水员病/n]

上例中的句型结构是 n-v-n,一般情况下,动词可以作为匹配模式,被识别的同义词为名词,那么这个种子句的模式可以被转换成为"A 也称为 B";同时,"减压病"和"潜水员病"被称为匹配同义词对。而这对匹配同义词又存在其他组合方式,表 1 记录了它们以及它们对应的句子模式:

表 1 模式解析对照

Table 1 Mode analysis comparison

原句子	句型结构	句子模式
减压病又叫潜水员病	[减压病/n][又叫/v][潜水员病/n]	A 又叫 B
减压病(潜水员病)	[减压病/n][(/wp][潜水员病/n][)/wp]	A(B)
减压病,是潜水员病的俗称	[减压病/n][,/wp][是/v][潜水员病/n][俗称/n]	A,是 B 的俗称

从上表可以看出,匹配同义词对存在多种组合句式,因此我们的系统应该尽可能多的帮助寻找不同的匹配句式。图 3 说明了我们的方法。

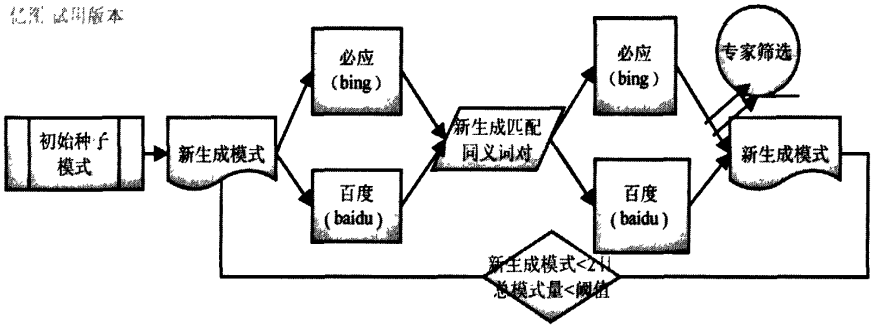

图 3 基于种子学习的半自动提取模式

Fig. 3 Extraction patterns based on semi-automatic seed learning

程序首先定义了 N 个初始种子模式,使用网络爬虫算法进行搜索;在搜索引擎列表中自动挖掘页面中存在种子模式的窗口(包括句子,段落,短语)(窗口大小 < M 个文本距离)。

这时窗口中常会出现大量的修饰性词语,如形容词、副词、语气词等;这些修饰性词语的加入会增加模式长度,而且影响了模式的准确率,因此我们仅保留对关系识别最有判断价值的核心词语,如名词、动词等,而把一般修饰性词语,即对判断实体关系无太大价值的词滤去。

然后将筛选得到的短语中的名词部分作为匹配同义词对,再次使用网络爬虫算法,经过同样的筛选过滤之后,得到的短语存入数据库中,给专家筛选得到新模式,将新提取的模式作为种子,出发算法运行;当新生成模式不再增加或者总模式量达到一个阈值 K 时,算法即可终止。

2.2 结合模式匹配算法从 Web 上抽取同义词

数据抽取是同义词关系抽取的核心步骤,但是其抽取效率与抽取准确度仍和匹配模式以及数据源的选择有着直接的关系,换句话说,只要前期准备工作做得足够充分,抽取过程就会顺利且高效。整个抽取过程分为三个步骤:

(1)组合、检索关键词

匹配模式或者被匹配词不能够单独用来检索同义词关系,因为我们需要找到特定词汇的同义词关系,所以模式和被匹配词应该结合起来。关键词组合方式如式 1:

$$\text{组合关键词} = (\text{模式}) + \text{被匹配词} + (\text{模式}) \quad (1)$$

匹配模式既可以在被匹配词前也可以在被匹配词后。由于中文字符串在 URL 中不能被自动解析,我们将其转换为 utf-8 格式,再使用 java 语言的中 URL 类进行检索。而每一个被匹配词的检索次数取决于模式的数量。

(2)从网页中抽取所需信息片段,存入数据库

在Web搜索引擎中使用匹配模式算法得到的页面包含我们需要的同义词关系,我们使用Htmlparser技术解析页面,同时使用停用词表去除残留格式标记(如 等)和停用词,包括代词、冠词、介词、叹词、拟声词等。过滤这些冗余信息可以提高抽取效率和避免抽取到无用信息。

设定词汇终止符和同义词列表的分隔符成为比较重要的一步,我们通过大量的实验统计得出了这些词汇,其中包括如表2所示的两种情况。

表2 终止符与同义词列表分隔符

Table 2 Terminator and a synonym list separator

并列连词	和,跟,与,同,及,乃至,此外,以及,并且,<,>,|,;,(,),(,),、,“,”,《,》,‘,’,〉,〈,、,\,>,<,(,),、,|,[,],|,|,’,“,”,‘
终止符	",",".","?","!",";","?","。","!",",",";","?"," ·",".","..."

(3)同义关系推导

得到同义词之后,我们通过同义词推导扩展更多同义词信息,表3列出了同义关系的性质和推导规则。利用词汇之间二元关系的性质,可以对词表中已知的等同关系进行推导,发现和挖掘出更多的未知等同关系。

表3 同义关系性质[23]

Table 3 The characteristic of Synonymous relationship

关系	推导规则	说明
对称性	xsy⇒ysx	如果A与B是同义关系,B与A也是同义关系
传递性	xsy,ysz⇒xsz	如果A与B是同义关系,B与C是同义关系,A与C也必定是同义关系

词汇关系的整个推导流程如图4所示:

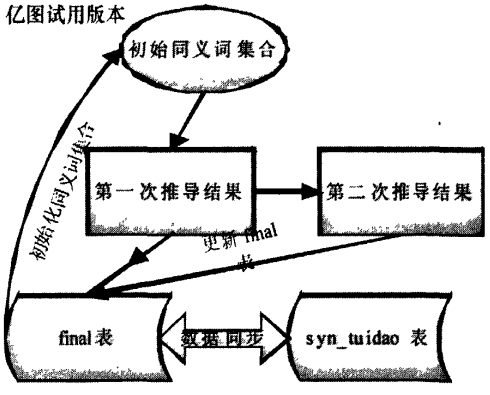

图4 同义词推导流程

Fig.4 Synonym derivation process

注:上图中final表与syn_tuidao表保持数据同步。为了保证数据抽取的效率,本文作两次推导。

最后我们可以从在线词典中抽取词汇的英文翻译,然后利用英文翻译找到对应的中文解释。根据同义词传递性原理,可知这些中文解释在一定程度上与原词汇属于同义关系。

3 基于《知网》义原词频的词汇等同关系识别技术

基于《知网》义原词频的词汇等同关系识别技术是中文叙词表本体等同关系自动构建方案的核心。它用来将已抽取出的大量 Web Data 筛选成更合适的候选集，并对同义词对给出相应的相似度参考值，辅助修订专家的修订工作。采用的识别技术需要准确性要求较高，否则，如果有用的词汇被漏掉，没用的词汇被留下，会让整个工程事倍功半。

《知网》[18]是董振东、董强两位先生于 1988 年开始发起并创建的一部以各类概念为描述对象的知识系统。它以义原为单位来描述现实世界中的对象，因此它只需要有一定规模的义原库就可以描述大量的事物，具有良好的可扩展性和语义表现力；其中义原是最基本的，在语义上不可再分的最小单位。在《知网》中词汇在语义上由义原构成，因此基于《知网》的词汇相似度算法通常分为义原相似度计算和词语相似度计算两个步骤。

3.1 义原相似度的计算

根据《知网》的定义，每一个词汇都由若干义原组成，所以词汇的等同关系识别就等价于词汇相似度的比较，而词汇相似度的比较可以拆分为义原相似度的计算。

Dekang Lin[19]提出的相似度理论告诉我们，任何两个元素的相似程度取决于它们的共性以及它们的个性。使用数学语言表达如公式(2)所示。

$$Sim(a,b) = \frac{common(a,b)}{common(a,b) + diference(a,b)} \tag{2}$$

在《知网》中，存在一个树状义原层次结构，里面存储着每个义原的若干信息：父节点、子节点以及位置等信息。越往义原层次结构的顶层，它所表示的语义信息就越模糊，越往底层，它所表示的语义信息就越具体；因此在计算义原相似度的时候不仅需要考虑两个义原的共性和个性，还应考虑义原所处的绝对位置。但是更好的办法是在共性和个性的定义中用到义原层次位置关系，这样就同时保证了处于上层的元素相似度不会过高，而处于底层的两义原不会因为距离过长而导致相似度降低得很快。

因此，本文结合上述分析并参考文献[20]定义，得到了一个计算义原相似度的公式(见公式(3))。

$$sim(a,b) = \frac{2 * common(a,b)}{difference(a,b) + 2 * common(a,b)} \tag{3}$$

其中 different(a,b)表示义原 a,b 在义原树中的距离；common(a,b)表示义原 a 和义原 b 在义原树中的重合部分，如果没有重合部分，那么 sim(a,b) = φ，其中 φ =0.2 表示最低重合相似度。

$$sim(a,b) = \frac{\alpha}{d + \alpha} \tag{4}$$

其中 d 表示义原 a 与义原 b 之间在义原层次体系中的路径长度，α 表示一个可调节参数。

3.2 词汇相似度的计算

因为词汇可以被拆分为一个或多个义原(称为义原组)，所以在得到义原相似度之后，关键问题变成了如何对应两个词汇的义原组，已经被对应的义原对之间的权重大小如何分配？对此，文献[21]、文献[22]和文献[20]给了我们一定的启示。

文献[21]认为义原根据其描述方式，在词汇义原组中所处位置可以分为四大类别：第一独立义原、其他独立义原、关系义原和符号义原。其中，关系义原使用“关系义原 = 基本义原”或者“关系义原 = (具体词)”来描述的义原；符号义原是用“关系符号 基本义原”或者“关系符号(具体词)”加以描述的义原；独立义原是用“基本义原”，或者“(具体词)”进行描述的义原，第一独立义原一般位于义原组的首位；具体词是指《知网》中使用圆括号()括起来的词汇，不存在于义原树中。下例中描述了一个词汇使用义原组表示的情形，并标注了各义原的类型。

例 词汇“称帝”的义原组可以表示如下：

称帝 peak|说，content = undertake|担任，#official|官，royal|皇

注：“peak|说”是第一独立义原，“royal|皇”是其他独立义原，“content = undertake|担任”是关系义原，“#official|官”是符号义原。

文献[21]对四大类型之间的权重分配依次是：$\beta_1 = 0.5$，$\beta_2 = 0.2$，$\beta_3 = 0.17$，$\beta_4 = 0.13$，而且是固定的分配；对于其他独立义原、关系义原和符号义原等可能同时包含多个义原的情况，采用的方式平均分配。这就导致了第一独立义原权重过大，而其他类型义原即使再相似也不能决定整个词汇的相似度。文献[21]的词汇相似度算法见公式(5)。

$$sim\_word(a,b) = \sum_{i=1}^{4} \beta_i \prod_{j=1}^{i} sim_j(a,b) \tag{5}$$

(注:sim_word 表示词汇 a,b 的相似度,simj(a,b)表示词汇 a,b 第 j 种类型义原相似度。)

文献[22]在文献[21]的基础上做了一定改进,它认为所有独立义原应该放在一起计算,这从一定程度上解决了第一独立义原权重过大的问题,并且可以解决顺序不一致导致相似度下降的问题。

文献[20]提出了弱义原的概念,它认为越往顶层节点,所包含的语义信息就越少,因此它称那些区别能力弱、重复多、语义信息少的义原为弱义原,并将"人,实体,物质,地方,万物,属性,场所,众,良,莠,专"等专门放入弱义原列表中。这种方式的缺点是人为定义弱义原,当《知网》更新之后不能随时变更;同时,只要义原相似度定义得够好,弱义原自然是可以被降低权重的;再者,弱义原概念模糊,没有量化。

综合前人观点,并作出合理分析之后,本文提出了一种基于《知网》义原词频的同义词识别算法。在分析《知网》结构后,我们发现衡量弱义原的因素不仅可以是义原树的层次,而且也可以是义原在《知网》中出现的频率,出现频率越多代表它的共性越大,那么区分度就会越低。比如:义原"属性值"的出现次数达到了 10950,占全部词汇的 1/6;在义原树中深度为 4 的义原"女"出现的频率和处于义原树顶层的"实体"一样多。

因此弱义原定义应该包含义原树深度和出现频率两种因素。而在 4.1 节中提出的义原相似度算法已经将义原深度考虑进去了,因此出现频率就成为在词汇相似度计算中核心影响因素。

此外,针对文献[21]中第一独立义原权重过大,且有顺序不一致导致相似度下降的问题,本文采用文献[22]的做法,将所有独立义原综合起来进行计算。本文所采用的相似度计算方法如公式(6)(7)所示。

$$sim\_word(a,b) = \sum_{i=1}^{3} \beta_i \prod_{j=1}^{i} sim_j(a,b) \tag{6}$$

$$sim_1(a,b) = \frac{\sum_{i,j=0}^{n} (Pab - Pi - Pj) * sim(i,j)}{(n-1) * Pab * \sum_{i,j=0}^{n} sim(i,j)} \tag{7}$$

(注:sim_word(a,b)是指词汇 a 和词汇 b 的相似度;$\beta_1 = 0.7$, $\beta_2 = 0.17$, $\beta_3 = 0.13$; siml(a,b)表示词汇 a,b 的独立义原的相似度;n 表示可匹配义原对数,i,j 表示第 k(0 < k < n)个义原对中对应在词汇 a,b 中的义原序号;Pab 表示词汇 a 和词汇 b 所包含所有义原在《知网》中的词频之和;Pi,Pj 表示词汇 a 的第 i 个义原的词频与词汇 b 的第 j 各义原的词频。)

表 4 给出了文献[21]词汇相似度算法与本文词汇相似度算法在部分词汇对计算结果的比较。

表 4 部分词汇对使用文献[21]和本文算法计算相似度结果

Table 4 the results of similarity by this algorithm using Literature 21

词汇一	词汇二	文献[21]算法	本文算法
教	宗教	0.19	0.6
生效	见效	0.24	0.61
帮腔	支持	0.17	0.6
提倡	倡议	0.17	0.6
炭	木炭	0.04	0.6
政策	策略	0.28	0.6
升降	起落	0.44	0.89
改建	改造	0.44	0.89
红人	宠儿	0.24	0.74

从上表可以看出，本文算法不但可以有效分辨同义词，而且对原算法的改进也是很明显的。

4 实验结果

本实验系统采用的语料有两个：一是从《中国大百科全书·经济卷》和《当代金融词典》中精选的3581个经济词汇，该语料的特点是权威、专业和概念清晰明确[24]；二是国家社科基金05CTQ001项目组建成的CCT1_OntoThesaurus，本文从中选取700多个词汇做实验比较，该语料中词汇涉及领域较广，部分专业词汇语义晦涩，在日常比较少出现。此外，为了对本文词汇相似度算法进行比较，系统采用了《同义词词林》，它是一部广泛包含同义词、同类词的词典，它每一行代表一个词群。本文使用共享版《同义词词林》的改进版作为开放性实验验证数据，过滤掉在《知网》中不存在的词汇以及相关词，最后得到每一行的前两个词汇作为计算对象。

对于语料一，系统不仅与文献[23]进行了抽取数量上的对比，而且对3581个经济词汇进行了同义词、英语翻译和相关词三方面的扩展。这两个实验分别如图5和图6。

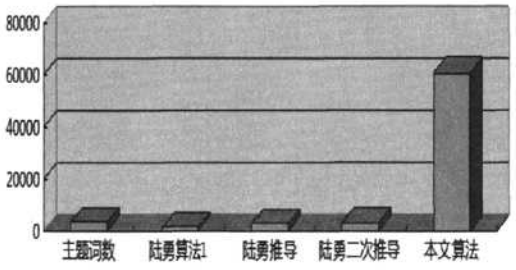

图5 与文献[23]对比

Fig. 5 The contrast of Literature 23

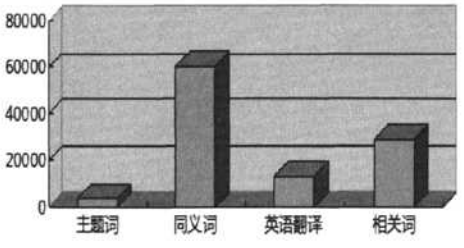

图6 同义词、英语翻译和相关词抽取数量

Fig. 6 The number of Synonyms, related words and English translation

在与文献[23]的对比中，我们发现本文算法比原算法要抽取的数量多很多，造成这种情况的原因是数据源不同：原算法使用《中国大百科全书·经济卷》和《当代金融词典》作为数据源，而本文算法使用百度、必应和雅虎中国三大搜索引擎作为数据源。前者存在数据所属领域单一、数据使用范围狭窄的问题，而本课题要求大规模地选取同义词候选词，然后在这中间找出合适的推荐词以供修订专家选择。对于3581个经济词汇，我们一共抽取了同义词60551个，英语翻译13602个以及相关词29318个。

我们使用手工随机抽样方法对程序形成的同义词进行验证。手工验证的方法是利用随机抽样法，从3581个经济词汇中随机抽样5%（也就是180个），进行手工判定，结果发现这180个词汇产生了3043个同义词，其中有效同义词（包括同义词、近义词或者语义上相关度很大的词汇）为56%（1704个词汇），22%的词汇是上下位词（670个）。

对于语料二，本系统进行了同义词推导的实验，具体数据见图7。

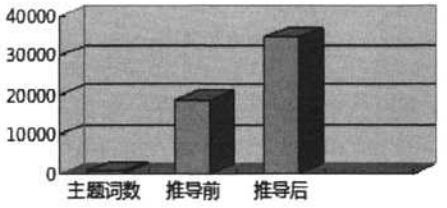

图7 对语料2进行同义词抽取及推导数量对比

Fig. 7 The contrast of quantity by extracting & Deriving Synonyms in corpus 2

在文献[23]中,推导前的同义词数一共2829,经过推导一共新增2925个词,新增比例为103%。在本文算法中对于CCT1_OntoThesaurus中的数据推导前的同义词数一共18680,推导后新增15704,新增比例也达到了84%,但是后者的数量是前者的数倍,因此从绝对数量来说它的扩展规模超过了前者。此外本文算法对于语料2中的数据仅有61个没有推导到同义词,覆盖率达到了91%。

以上两个实验的结果说明,相对于以前的算法,本文提出的基于Web搜索引擎的模式匹配算法在数量上有了较大提高。对于第二个创新点,我们从横向和纵向两方面进行了验证和分析。

一方面我们进行横向比较,将本文算法与文献[21]的算法进行比较,采用同一数据集(《同义词词林》),《同义词词林》共有19352,其中有9995行同义词群,本文抽取其中符合条件的7969对词汇,计算结果如图8。

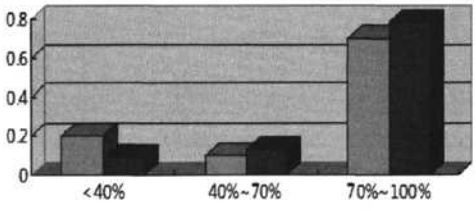

图8 与文献[21]的对比

Fig. 8 The contrast of Literature 21

分析图中数据,通过对比我们发现文献[21]相对于本文算法的误差较大,在相似度[0.0,0.4)区段,文献[21]占了全部的20%,说明文献[21]算法还存在一定的缺陷,而本文算法下降了11%。在相似度[0.7,1.0]区段,本文算法也提高了8%,说明抽取出来的同义词对经过本文算法自动分析已经达到基本一致的水平。

同时相似度区段[0.4,0.7)和[0.0,0.4)还保持22%的比例说明,一方面《知网》语义解释还不完善,还需要进一步的完善知网;另一方面,通过其他语义信息(相关关系、部分-整体关系等)也许可以找到更好的算法,弥补只通过上下位关系计算相似度算法的不足。

另一方面,我们进行了纵向比较,对语料1和2中抽取出的同义词进行自动筛选,将同义词相似度大于某个阈值(经验值为:0.4),并且词性一致的词汇自动推荐给专家。结果发现,推荐出的词汇确实存在较大相关性,但是也有部分词汇被错误推荐。表5给出了部分抽取结果和推荐结果对比。

表5 部分抽取词汇和推荐词汇对比

Table 5 The contrast between some extracted words and recommended words

主题词	抽取词	推荐词
同步卫星	球静止卫星|球自转周期相同人造球卫星|同步光學網路|通讯卫星|球自转周期人造球卫星|球静止卫星轨道距面高度3万6千公里|球静止轨道卫星|通信卫星|同步气象卫星|球同步运转|通讯卫星|定点卫星|相静止卫星|通讯卫星.|动伏卫星|球同步卫星|圆形轨道赤道面重合|下特点|运行周期|球静止轨道|24|同步轨道通信卫星|球|动讫卫星|静止卫星|驻定卫星|轨道面倾角零|球静止|倾角零圆形同步球轨道卫星	球静止卫星|球自转周期相同人造球卫星|定点卫星|球同步卫星|赤道同步卫星|球静止轨道卫星|动讫卫星
巴氏灭菌	巴斯德消毒法|常|液体|法|尿液|分钟|低温消毒法|明显延缓细菌发育|加热至中|进行消毒|种消毒法可杀死大数种类病原菌|杀菌|巴氏杀菌|牛奶|巴氏灭菌法|温度|维持一定时间|用巴氏法灭菌|灭菌|巴氏|巴氏消毒法	巴斯德消毒法|低温消毒法|巴氏杀菌|巴氏消毒法
岩溶作用	称喀斯特化|岩溶酌|喀斯特作用	喀斯特作用
稳定同位素	安定同位体|安定同位素|重氢|安定核素|性|放射性核素|稳定|稳定核种|核素|稳定性核素	安定同位体|安定同位素|安定核素|稳定性核素
妊娠病	妊娠毒血症|免疫妊娠病|产前病|胎前病|疾病英文名	妊娠毒血症|免疫妊娠病|产前病|胎前病
中央革命根据地	中共苏区|中央革命|中央苏区|中央|中央革命|胡百昌	中共苏区|中央苏区
射气测量	氧气测量|射气探勘|射气勘探	氧气测量|射气探勘|射气勘探
痕迹学	足迹化石学|古痕迹学|遗迹学|古遗迹化石学|古遗迹学|化石足迹学|验枪学|生痕学|司法弹道学	足迹化石学|古痕迹学|遗迹学|古遗迹化石学|古遗迹学|化石足迹学
外交学	外交|外交家|外交官|外交人员|外交术|外交手腕|技术外交	外交|外交家|外交官|外交人员|外交术|外交手腕|技术外交

5 结 语

本文针对OTCSS系统存在的修订信息时效性不稳定和修订信息不够丰富的不足，提出了可以缓解上述问题的概念间等同关系自动构建方案。

目前系统还处于实验阶段，很多想法还不够成熟，有待改善和提高。今后研究的重点主要在提高系统运行效率、智能推荐同义词、对抽取词汇进一步分析词汇关系等方面。

此外《知网》收录的词汇数只有6万多个，有很多中文叙词表本体上有但是《知网》上没有的词汇，因此未登录词的同义词识别技术是目前的一个研究方向，文献[20]对此有初步的论述和解决方案。

参 考 文 献

[1] 曾新红. 中文叙词表本体——叙词表与本体的融合 [J]. 现代图书情报技术，2009(1).

[2] P. Borst. Construction of Engineering Ontologies for Knowledge Sharing and Reuse[D]. The Neterlands: University of Twente, Enschede, 1997. Wegner P, Zdonik SB. Inheritance as an incremental modification mechanism or what like is and isn't like. In: Gjessing S, Nygaard K, eds. Proc. of the ECOOP'88. LNCS 322, Heidelberg: Springer-Verlag, 1988. 55-77.

[3] ISO 2788-1986, Documentation - Guidelines for the Establishment and Development of Monolingual Thesauri, 2nd ed[S]. Geneva: International Organization for Standardization, 1986.

[4] Hearst M A. Automated Discovery of WordNet Relations[A]// Christiane F. WordNet : An Electronic Lexical DataBase [M] , MIT Press , 1992:132-152.

[5] 梁建，王惠临. 基于文本的本体学习方法研究[J]. 情报理论与实践，2007(1):112-115,17.

[6] Casado Ruiz M, Alfonseca Em Castells P. Automatic Extraction of Semantic Relationships for WordNet by Means of Pattern Learning from WikiPedia[C]. In: Natural Language Processing and Information Systems: 10th International Conference on Applications of Natural Language to Information Systems, NLDB 2005, Alicante, Spain , june 15-17, 2005:67-79.

[7] Maedche A, Staab S. Ontology Learninf foe the Semmantic Web [J]. IEEE INTELLIGENT SYSTEMS, 2001, 16(2):72-79.

[8] Agirre E, Ansa O, Hovy E et al. Enriching Very Lage Ontologies Using the WWW[C]. In: proceedings of the Ontology Learning Workshop, ECAI2000, Berlin, Germany, 2000.

[9] Khan L, Luo F. Ontology Construction for Information Selection [C]. In: proceeding of 14th IEEE International Conference on Tools with Artificial Intelligence, Washington DC, 2002:122-127.

[10] Bisson G,Nedellec C , Ganamero L. Designing Clustering Methods for Ontology Building - The Mo'K Workbench[C] . In: proceedings of the ECAI Ontology Learning Workshop, 2000:13-19.

[11] 仲云云,侯汉清,杜慧萍.电子政务主题词表自动构建研究[J].中国图书馆学报,2008(3):97-102.

[12] 曹恬,周丽,张国煊.一种基于词共现的文本相似度计算[J]. 计算机工程与科学,2007,9(3):52-53.

[13] 毕玉德,阎艳萍.一种基于 WordNet 的多语种词汇语义网半自动构建方法[J].解放军外国语学院学报,2008,31(5):55-59.

[14] 张玉娟.基于知网的句子相似度计算的研究[D].中国地质大学(北京).2006.

[15] 程涛,施水才,王霞,吕学强.基于同义词词林的中文文本主题词提取[J].广西师范大学学报(自然科学版).2007,25(2).

[16] 郭文宏,范学峰. 面向 Web 结构化信息处理的汉语知识库构建研究[J]. 计算机科学, 2009,(01).

[17] Eugene Agichtein,Luis Gravano. Snowball: Extracting Relations from Large Plain-Text Collections[EB/OL]. http://citeseerx. ist. psu. edu/viewdoc/download? doi = 10. 1. 1. 41. 4658&rep = rep1&type = pdf .

[18] 董振东,董强."知网"[EB/OL]. http://www. keenage. com . 1999.

[19] Dekang L. An Information-theoretic Definition of Similarity[C]//Proceedings of the 15th International Conference on Machine Learning. 1998: 296-304.

[20] 夏天.汉语词语语义相似度计算研究[J].计算机工程.2007,33(6).

[21] 刘群,李素建.基于知网的词汇语义相似度计算[C]//第3届中文词汇语义学研讨会论文集.2002(5).

[22] Eugene Agichtein,Luis Gravano. Snowball: Extracting Relations from Large Plain-Text Collections[EB/OL]. http://citeseerx. ist. psu. edu/viewdoc/download? doi = 10. 1. 1. 41. 4658&rep = rep1&type = pdf .

[23] 陆勇,侯汉清. 基于模式匹配的汉语同义词自动识别[J].情报学报,2006,25(6):720-724.

[24] 王益等.当代金融词典.北京:中国经济出版社,2000.