文章编号：1003-0077(2015)03-0044-08

HowNet 与 CCD 映射方法研究

向春丞[1]，穗志方[1,2]，詹卫东[1]

(1. 北京大学 计算语言学教育部重点实验室，北京 100871；
2. 语言能力协同创新中心，江苏 徐州 221009))

摘 要：本体映射是解决本体异构问题的关键方案。该文以 HowNet 和 CCD 中的名词性概念为例，首先利用机器学习技术发现初始映射关系，主要包括特征选择、样本集合划分、分类器选择等步骤；然后考虑本体的整体结构信息，利用相似度传播算法，对初始映射关系进行全局调整。实验表明，最终的一对一和一对多映射关系的准确率分别达到了 94%和 87.5%。

关键词：本体映射；机器学习；分层抽样；相似度传播算法

中图分类号：TP391 **文献标识码**：A

On Mapping between HowNet and CCD

XIANG Chuncheng[1], SUI Zhifang[1,2], ZHAN Weidong[1]

(1. Key Laboratory of Computational Linguistics(Peking University),
Ministry of Edacation, Peking University, Beijing 100871, China;
2. Collaborative Innovation Center for Language Ability, Xuzhou, Jiangsu 221009, China)

Abstract: Ontology matching is the key solution to the semantic heterogeneity problem. Focusing on the Noun concept of HowNet and CCD, this paper applies machine learning to identify the initial mapping relationships, disicussing the the feature selection, sample collections division and classifier selection. Further, employing the overall structure of the ontology, the similarity propagation algorithm is introduced to adjust the initial mapping globally. Experiment result shows that the precision of 1:1 and 1:n mapping relationships reaches 94% and 87.5%, respectively.

Key words: ontology matching; machine learning; stratified cross sampling; similarity propagation algorithm

1 前言

本体作为一种能在语义和知识层面上描述领域概念的建模工具，近年来在人工智能、信息检索、语义 Web 等领域受到了极大关注，本体数量在其研究和运用领域呈爆炸式增长。然而，独立地设计和开发导致了大量描述同一领域知识的本体之间存在严重的异构问题，极大阻碍了本体之间知识的共享和融合。本体映射能够在异构本体之间发现语义相似的实体，是解决本体异构问题的关键方案[1]，已成为当前语义 Web 领域中的一个研究热点。

目前，研究者们已在本体映射方面做了大量工作，提出了许多映射方法[2-4]，如基于实体名称、基于本体结构、基于背景知识以及基于语义的方法等。通常，大部分本体映射系统[5-6]都将多个基本匹配器进行线性综合，然后使用一些优化策略，发现映射关系。然而，手动地设置组合时的参数很难获得最佳映射关系，于是研究者们将机器学习技术[7-9]引入本体映射任务，自动地对基本匹配器进行组合。

中文本体映射方面的研究相对薄弱。文献[10]尝试将知网与同义词词林进行融合，首先利用知网中的义原对词林中的每个原子词群给出一个 DEF 描述；然后在该特征上定义两种形式的相似度计算，并将它们结合起来，通过反复试验确定阈值，实现分类的目的。其相似度计算过程中仅考虑了本体本身

收稿日期：2013-04-08 定稿日期：2013-07-28
基金项目：国家重点基础研究发展计划(2014CB340504)，国家自然科学基金(61375074)。

的词汇信息，缺乏对本体结构以及外部词典或互联网资源的利用，对词汇语义信息的利用也不够。

本文初步探索了知网（HowNet）与中文概念辞书（Chinese Concept Dictionary，CCD）两部词典的映射方法。首先利用两者的词汇信息、语言信息以及语义信息定义映射特征；然后给出将样本集划分成正例集、负例集以及测试集的策略，接着利用机器学习技术发现映射关系；最后考虑本体的整体结构信息，利用相似度传播算法对初始映射结果进行调整。实验表明，最终的概念之间的一对一和一对多映射关系准确率可达到 94％和 87.5％。

2　术语及相关介绍

本节给出相关的术语和介绍，包括本体和本体映射的定义、HowNet 与 CCD 的简介以及本文中待映射本体的说明。

2.1　本体和本体映射

在计算机科学的不同领域，有很多的数据和概念模型都可以被称为本体，例如，普通的分类、数据库模式、UML 模型、字典、主题词表、XML 模式以及正式化的本体等。根据文献[11]的描述，本体（Ontology）主要包括概念（Concepts）、属性（Properties）、实例（Instances）以及公理（Axioms），可形式化地表示为：

$$O = \{C, P, I, A\}$$

其中，C 表示概念或类（Classes）的集合；I 表示概念的实例或个体（Individuals）的集合；P 表示属性集合，分为对象属性（Object Properties）和数据属性（Datatype Properties），前者用来表示概念之间或实例之间的关系，后者用于描述概念或实例的固有特征；A 表示公理集合，用来对概念或属性进行约束。

本体映射（Ontology Matching）是发现不同本体的实体之间的关联关系（relationships）或对应关系（correspondences）的过程[1]。所谓本体的实体，主要指本体中的概念、实例或者属性。可将本体映射形式化为：

$$A' = F_{(O_S, O_T, A, p, r)}$$

其中，函数 F 表示映射过程，O_S 和 O_T 分别表示源本体和目标本体，A 表示 O_S 与 O_T 之间可能已存在的映射关系，p 表示映射过程中用到的权值或阈值等参数，r 表示映射过程中用到的外部资源，A' 表示映射结果，可理解为由具有映射关系的实体对组成的集合。实体之间的映射结果可以是一对一、一对多、多对一以及多对多的映射情况。

2.2　HowNet① 与 CCD 的简介

知网（HowNet）是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库[12]。知网的规模主要取决于双语知识词典数据文件的大小，包含 194 302（2011 版）条义项记录。

中文概念辞书是一个基于 WordNet 框架的汉英双语语义知识库[13]。它将代表概念的词语分为名词、动词、形容词以及副词四种，目前收录了近十万个概念。

图 1 给出了一条 HowNet 记录的例子。其中，NO. 表示记录的编号、W_C 表示概念的中文表述、W_E 表示概念的英文表述、DEF 是对概念的规格化描述。DEF 中第一位置的义原“Circuitances|境况”称为主要特征，它是概念“CONDITION|状况”的直接上位概念。

图 2 为 CCD 中描述名词性（POS＝n）概念“{态势 情形 状况 状态}”的主要格式，其中 Definition 和

```
NO.=180648
W_C=状况
G_C=noun [zhuang4 kuang4]
S_C=
E_C=
W_E=condition
G_E=noun
S_E=
E_E=
DEF={Circumstances|境况:host={
event|事件}{thing|万物}}
```

图 1　HowNet 记录举例

```
Offset=00016185
POS=n
Category=03
Synset=state
CSynset=态势 情形 状况 状态
Definition=the way something is with respect to its main attributes
CDefinition=事物固有的方式
Note=the current state of knowledge; his state of health;in a weak financial state
CNote=现有的知识状态；他的健康状况；在不利的经济状态下
Hypernym=
Hyponym10058077100602591006346910063667100639 06…
```

图 2　CCD 概念及其描述举例

① 出于表述简便，本文中所谓的“HowNet”主要指知网系统中的双语知识词典数据文件。

Note分别表示概念的释义(定义)和使用举例,Hypernym和Hyponym表示该概念的直接上、下位概念的编号。通常,一个概念的直接上位概念只有一个,而直接下位概念有多个。

2.3 待映射本体

HowNet和CCD都是一部体现了对客观世界的认识与把握的中英文词汇概念语义词典,因此将其所描述的概念进行映射是合理的。本文映射任务中,源本体O_S中的概念为HowNet中的名词性概念,目标本体O_T中的概念为CCD中的名词性概念。

由于HowNet和CCD的编纂时期、概念划分粒度以及应用目标等方面存在一定的差异,因此两部词典中收录的名词性词语的数量差别较大,其统计结果如表1所示。

表1 待映射本体初步统计表

	中文词语数	英文词语数	义项数(记录/同义词集)
O_S	57 072	55 399	87 393
O_T	103 735	92 716	64 895
共现词语数	25 870	26 370	—

由于本文的映射策略还考虑了概念的分类体系对映射关系的影响,因此我们将描述HowNet概念的实体类、属性类以及属性值类义原以HowNet记录的形式加入到了原来的记录集合中,其中实体类义原的DEF不变,属性类和属性值类义原的DEF定义为其上位概念。

3 利用机器学习技术发现映射关系

本节主要介绍将机器学习技术用于HowNet与CCD的映射任务。将映射问题看作二分类问题,首先进行映射特征的选择;然后给出将样本集自动划分成训练集和测试集的策略;最后介绍分类器的选择和预测过程。

3.1 选择映射特征

文献[10]中提出的知网与同义词词林的融合特征,为CCD中的每个同义词集定义一个DEF描述,得到映射特征$F3$-$F6$(表2)。另外,利用CCD概念的Note和Definition属性,定义映射特征$F7$和$F8$。映射特征$F1$-$F8$的具体描述如表2所示。

表2 HowNet与CCD的映射特征

映射特征	特 点	描 述
No.	—	实例编号,由HowNet的概念编号和CCD的概念编号组成
$F1$	词汇级别	CSynset, ESynset是否分别包含W_C, W_E
$F2$	词汇级别	CSynset, ESynset是否分别不包含W_C, W_E
$F3$	词汇级别	C_H主特征向量与C_C主特征向量的重合度
$F4$	词汇级别	C_H次特征向量与C_C次特征向量的重合度
$F5$	词汇级别	C_H主特征向量与C_C主特征向量的相似度
$F6$	词汇级别	C_H次特征向量与C_C次特征向量的相似度
$F7$	语言级别	概念的释义之间的相似度
$F8$	语义级别	概念的举例之间的相似度
C.	类别	是否具有映射关系

其中,特征$F3$-$F6$的计算方法与文献[10]相同。用W_C表示HowNet中某个概念的中文词条,为了计算特征$F7$,首先从新华字典中获取W_C的名词性释义;如果该名词性释义有多项,则说明W_C为多义词,此时利用其相应的DEF中的主要特征和次要特征进行排歧、选择;如果字典中未给出W_C的名词性释义,则取其基本释义代替。然后再计算W_C的释义与CCD中概念的释义之间的余弦相似度。

对于特征$F8$,首先利用互联网语料训练得到Bigram语言模型,然后将CNote中出现的CSynset中的词语用W_C替换,将替换后的CNote的概率作为特征$F8$的值(采用加一平滑技术处理数据稀疏问题)。如果$F8$的值较大,则说明两个概念之间

的语义相似度越高。

3.2　划分样本集合

将一个 HowNet 概念和一个 CCD 概念组成的概念对$<C_H, C_C>$称为一个映射样本，它由表 2 中定义的映射特征来描述。如果 C_H和 C_C之间存在映射关系，则将该映射样本称为正例，否则称为负例。

对包含 87 393 个概念的 O_S和包含 64 895 个概念的 O_T进行统计，其中使得特征 $F1$ 的值为真的映射样本的个数为 37 021 个，涵盖了 29 086 个 HowNet 概念和 18 283 个 CCD 概念。从这些映射样本中随机选取 200 个人工进行观察，发现其中有 187 个可以被看作正例。也就是说，如果把使得特征 $F1$ 值为真的映射样本作为正例，其可信度能够达到 98%。这主要是由于 W_C 和 W_E 之间具有相互排歧的作用。我们允许一定的误差存在，利用特征 $F1$ 和 $F2$ 对样本集合进行划分，即将特征 $F1$ 和 $F2$ 的值为真的映射样本分别作为正例和负例，其他包含了 49 697 个 HowNet 概念和 29 503 个 CCD 概念的大约 16.5 万个映射样本组成测试集。

3.3　分类器的选择和预测

目前，能够解决二分类问题的机器学习算法有很多，因此需要根据实际任务的特点进行选择。首先，利用分层抽样方法从负例集中抽取与正例集规模相当的样本，并将其与所有正例组成训练集；然后，对多个分类器在训练数据集上采用交叉验证的方法进行训练，选择 F 值最高的一个作为最终的分类器对测试样本进行预测，从而发现测试样本中的映射关系。由于特征 $F1$ 和 $F2$ 已被用于样本集合的划分，因此，在分类器的训练和预测阶段均不考虑样本的这两个特征。

4　基于相似度传播算法的映射关系调整

相似度传播算法[14](Similarity Flooding Algorithm，SF)是一种图匹配算法，它将图中的节点看作概念，节点之间的连边看作概念之间的关系，认为两个概念之间的映射结果不仅跟它们各自的特征有关，还跟其邻近概念，甚至图中所有其他概念的映射结果也有关。概念之间的相似度通过图中的连边在整个图上进行迭代传播。

本文不把待映射本体的分类结构 H_S和 H_T(如图 3 所示)按照文献[14]中的方法进行合并，因为这样会急剧增加节点个数。例如，对 H_S和 H_T中分别以节点 A 和 B 为根的子树进行合并，节点个数将由 m+n+2 个变为 m*n+1 个，而 CCD 和 HowNet 中很多概念都有几十甚至上百个子概念。

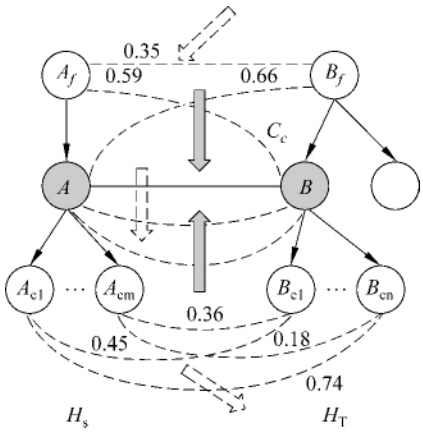

图 3　相似度传播算法示意图

定义对概念对<A，B>的映射关系有影响的邻近环境为：{<A_f，B_f>，<A_f，B>，<A，B_f>，<A，B_c>，<A_c，B>，<A_c，B_c>}，其中<A_f，B_f>表示“如果概念 A 和概念 B 的父概念相似，那么概念 A 和概念 B 也可能相似”；<A，B_c> 表示“如果概念 A 与概念 B 的子概念相似，那么概念 A 与概念 B 也可能相似”。

另外，在执行相似度传播算法之前，还需要为每个概念对设置初始相似度值。以概念对<A_f，B_f>为例，如果它对应正例集中的某个样本，那么它的初始相似度值为 1；如果对应测试集中的某个样本，那么它的初始相似度值为分类后的置信度值；否则为 0。以图 3 为例，将上述过程形式化如式(1)所示。

$$\sigma(i+1)=\frac{1}{\max(\sigma(i+1))}*(\sigma(0)+\sigma(i)+\varphi(i)) \tag{1}$$

其中，$\sigma(i+1)$ 表示概念对<A，B>在第 i+1 次迭代后的相似度；函数 $\varphi(i)$ 表示其邻近环境在第 i 次迭代时产生的影响，它由 $\varphi(i)_f$ 和 $\varphi(i)_c$ 两部分组成，即$\varphi(i)=\varphi(i)_f+\varphi(i)_c$，表示为式(2)、式(3)。

$$\varphi(i)_f=\frac{1}{1+|A_f|}f(A_f,B)+\frac{1}{1+|B_f|}f(A,B_f)+\frac{1}{1+|A_f|*|B_f|}f(A_f,B_f) \tag{2}$$

$$\varphi(i)_c=\frac{1}{1+m}\sum_{i=1}^{m}f(A_{ci},B)+\frac{1}{1+n}\sum_{j=1}^{n}f(A,B_{cj})+\frac{1}{1+mn}\sum_{i=1}^{m}\sum_{j=1}^{n}f(A_{ci},B_{cj}) \tag{3}$$

其中，$|A_f|$ 表示概念 A_f 的子概念的个数。上述计算公式基于这样的思想：如果某个概念的子概念个数越多，那么认为由它带来的影响就越小。

利用上述定点计算公式对测试样本的相似度值在整个图上进行迭代修正，达到基于相似度传播算法调整映射关系的目的。

5 实验及结果分析

5.1 负例选择实验

由于待映射本体中的每一对概念之间都有可能存在映射关系，因此样本集的大小为 87 393 × 64 895，其中除了 37 021 个正例样本和约 16.5 万的测试样本外，剩下的均为负例样本。因此，必须对负例样本集进行压缩，使其规模与正例个数相当，且压缩后得到的负例样本的统计特性应与压缩之前比较接近。

本文的做法是，首先从所有负例集中随机选取 1 亿个样本得到样本集 M；然后利用分层抽样方法从 M 中抽取与正例数量相当的样本，得到样本集 $\mathbb{N}$。对 M 和 $\mathbb{N}$ 中所有样本的特征的取值进行统计分析，其结果如图 4 所示（实验通过调用 Weka API 实现）。

图 4 中，F3-F8 对应表 2 中定义的特征，Mean_0 和 Mean_1 分别表示负例集在压缩之前和之后的特征值的均值，StdDev 表示标准差，$\mathbb{N}$ 的大小为 37 038。假设 M 中样本的分布情况与整个负例集中的一致，那么由上图可知，通过分层抽样方法得到的 $\mathbb{N}$ 中样本的统计特性与 M 中的非常相似，因此我们可以认为样本集 $\mathbb{N}$ 可以代表整个负例样本集。

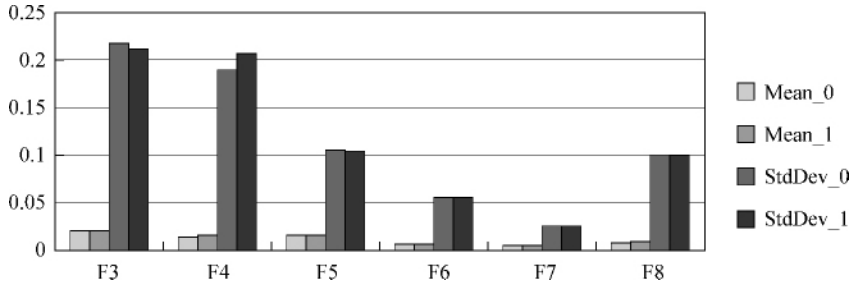

图 4 负例集压缩前后的统计特性对比

5.2 分类器的选择实验

现在我们已经得到了一个包含 37 021 个正例和 37 038 个负例的训练数据。此时我们希望找到一个在该训练集上表现较好的分类器，以对测试样本进行分类，从而发现更多的映射关系。我们分别对朴素贝叶斯（Naive Bayes，NB）分类器、决策树（Decision Tree，DT）分类器以及最大熵（Maximum Entropy，ME）分类器进行了实验和比较，分类器训练时均采用 10 折交叉验证方式，实验通过调用 Weka API 完成，其训练结果如表 3 所示。

表 3 分类器训练效果比较

Classifier	*Precision*/%	*Recall*/%	*Feature*/%
NB	87.60	86.50	86.40
DT	88.50	88.10	88.10
ME	88.00	87.10	87.10

上表中，*Precision*、*Recall* 以及 *Feature* 表示分类器在训练集上的查准率、查全率以及 F 值。从表 3 可以发现，决策树在本文的训练集上表现最好，于是我们将其作为最终的分类器。图 5 为该决策树分类器的部分结构。

从图 5 可知，次特征向量（secdf_cos_sim，即特征 F6）的相似度对类别的区分能力最强，被选定为根节点。完整的决策树模型共有 49 个节点，其中包含 25 个叶节点。

5.3 映射关系发现实验

依次利用以下三种方法发现从 HowNet（O_S）到 CCD（O_T）的映射关系。

（1）基于特征频度统计和特征向量计算结合的方法[10]

该方法主要参考文献[10]，它首先通过反复试验设定阈值，然后执行多步判断，实现对概念对的分类。该过程可以看作是人工制定分类规则来判定概念之间是否存在映射关系。本文并未使用这样反复尝试的方法选定所需阈值，而是将阈值设定为相应特征值的均值。测试集中所有样本的特征的统计特性如表 4 所示。

```
secdf_cos_sim <= 0.083333
|   mainf_cos_sim <= 0.123203
|   |   defSim <= 0.03849
|   |   |   mainf_ch_sim <= 0.065789
|   |   |   |   secdf_ch_sim <= 0.054348: no (39392.0/5614.0)
|   |   |   |   secdf_ch_sim > 0.054348
|   |   |   |   |   secdf_cos_sim <= 0.021439: no (146.0/71.0)
|   |   |   |   |   secdf_cos_sim > 0.021439: yes (122.0/26.0)
|   |   |   mainf_ch_sim > 0.065789
|   |   |   |   mainf_ch_sim <= 0.1125: no (42.0/18.0)
|   |   |   |   mainf_ch_sim > 0.1125: yes (390.0/107.0)
|   |   defSim > 0.03849
|   |   |   mainf_ch_sim <= 0.071429
```

图 5　决策树分类器部分结构

表 4　测试集样本的统计特性

	F3	F4	F5	F6	F7	F8
Mean	0.354	0.585	0.236	0.233	0.043	0.005
SedDev	0.894	1.738	0.346	0.323	0.097	0.073

表 5　HowNet 到 CCD 的映射结果统计表

	方法 1	方法 2	方法 3
1∶1	11 504/95.00%	18 551/91.50%	18 546/94.00%
1∶n	7 913/92.50%	19 684/85.50%	19 745/87.50%

根据表 4 的统计结果，对方法 1 中的相关阈值进行设定。其中主特征和次特征的重合度阈值分别设定为 0.354 和 0.585；主特征和次特征的向量相似度阈值分别设定为 0.236 和 0.233。

(2) 基于统计决策树的方法

文献[15]中也利用了决策树的方法进行本体映射，但是其决策规则均由人工进行构造，其分裂节点的阈值通过反复试验选定，这样的阈值选定策略不仅费时费力，而且对训练数据的适应能力较差。本文中的决策树模型则是通过机器学习方法自动训练得到，从而能有效地发现训练数据中所蕴含的分类规律。

(3) 基于相似度传播的方法

方法 1 和 2 仅考虑了概念的局部特征，没有充分利用本体固有的结构信息。本文中的相似度传播方法主要是在方法 2 的基础之上，利用本体的整体结构信息来对映射结果进行调整，使其更为合理，另外，该方法还可以发现测试集之外的映射关系。与方法 1 类似，该方法中映射阈值取为算法迭代一定次数(n=100)后相似度值的均值，即 0.43。

从对测试集的映射预测结果中随机选取 200 个进行人工评价，以上三种映射方法的映射发现结果如表 5 所示(观察从 HowNet 到 CCD 的映射情况)。

其中，方法 1 中的“1∶1”(一对一)映射结果“11 504/95.00%”表示：测试集中有 11 504 个 HowNet 概念，每个仅能映射到一个 CCD 概念上，映射准确率为 95.00%。“1∶n”(一对多)表示：一个 HowNet 概念与多个 CCD 概念具有映射关系。

5.4　实验结果和错误分析

方法 1 主要基于概念词语的 DEF 描述的词汇级匹配特征，即如果两个概念的主要特征和次要特征具有较高的相似度，那么这两个概念可能具有映射关系。但就 HowNet 与 CCD 的映射任务而言，该方法仅能发现部分映射关系。

方法 2 在方法 1 的基础之上还考虑了其他一些特征，并利用机器学习技术自动的对基本匹配器进行组合，能够发现测试集中其他大部分映射关系。例如，HowNet 概念“WC=丹麦首都，WE=capital_of_denmark”，其主要特征为“place|地方=1.0”，次要特征为“capital|国都=1.0，Denmark|丹麦=1.0，ProperName|专=1.0”；CCD 概念“csynset={丹麦首都，哥本哈根}，esynset={copenhagen，kobenhavn，danish_capital}”，其主要特征为“country|国家=2.0，ProperName|专=2.0，Denmark|丹麦=2.0，politics|政=2.0，capital|国都=2.0”，次要特征为“place|地方=2.0”；通过计算，其主、次要特征的重合度和相似度均为 0，因此无法利用方法 1 判断这对概念具有映射关系。但其特征 $F7$、$F8$ 的值分别为 0.306 186、1.326 442E-9，即这两个概念的释义和举例之间具有较高的相似度，从而在方法 2 中被认为具有映射关系。

方法3将方法2的分类结果的置信度值作为初值,利用概念的上下位关系,在整个分类结构上对初始映射结果进行迭代地调整。例如,HowNet概念"WC=仓促,WE= precipitation"与CCD概念"csynset = {意外,突然,突如其来},esynset = {abruptness, precipitateness, precipitance, precipitancy, suddenness}",根据样本集划分原则,由这两个概念构成的样本将被视为负例,但在方法3的实验结果中却认为它们之间存在映射关系,这与人的判断结果是一致的。因此,方法3能够发现测试集之外的映射关系。

前两种方法的映射错误主要来自单字多义概念之间的映射。例如,HowNet中由"阵"字表示的概念的义项有"WC=阵,WE=spell"、"WC=阵,WE=position"、"WC=阵,WE=battle_array"以及"WC=阵,WE=front"等,方法1、2都认为它们与CCD概念"csynset={阵,一阵,冲动,发作,爆发,一阵子},esynset={burst, fit}"具有映射关系。

6 结语

本文利用机器学习技术和相似度传播算法对HowNet和CCD中名词性概念之间的映射作了初步探索并取得了较好的效果,但由于两部词典对概念粒度划分、属性定义的差异,还是未能对一部分概念进行映射。

本体映射是一项复杂的任务,本文就映射训练集缺乏、负例集压缩以及映射关系的全局调整给出了初步解决策略。但还有很多方面值得进一步考虑,例如,(1)用于划分样本集的假设限制太严,致使测试集规模偏小;(2)相似度算法在实现时的效率问题等。我们将在后续论文中对这些情况进行更深入的研究。

参考文献

[1] Jerome E, Pavel S. Ontology matching[C]//Proceedings of the Springer-Verlag, Heidelberg (DE), 2007.

[2] Qu Y, Hu W, Chen G. Constructing virtual documents for ontology matching[C]//Proceedings of the 15th International World Wide Web Conference (WWW). Edinburgh (UK), 2006: 23-31.

[3] Gligorov, Risto, et al. Using Google distance to weight approximate ontology matches[C]//Proceedings of the 16th international conference on World Wide Web (WWW). Beijing, China, 2007: 767-776.

[4] Atencia M, Borgida A, et al. A formal semantics for weighted ontology mappings[C]//Proceedings of the Semantic Web-ISWC 2012: 17-33.

[5] Nagy M, Vargas-Vera M. Towards an automatic semantic data integration: Multi-agent framework approach[C]//Proceedings of the Chapter in Sematic Web. In-Tech Education and Publishing KG, 2010.

[6] Li J, Tang J, Li Y, et al. Rimom: A dynamic multistrategy ontology alignment framework. Knowledge and Data Engineering [C]//Proceedings of the IEEE Transactions on 21, 2009: 1218-1232.

[7] Zhang D, Lee W S. Web taxonomy integration using support vector machines[C]//Proceedings of the 13th international conference on World Wide Web (WWW). New York, 2004: 472-481.

[8] Rong S, Niu X, et al. A Machine Learning Approach for Instance Matching Based on Similarity Metrics [C]//Proceedings of the Semantic Web-ISWC 2012: 460-475.

[9] Nezhadi A. H, Shadgar B, Osareh A. Ontology alignment using machine learning techniques[J]. International Journal of Computer Science & Information Technology (IJCSIT), 2011,3(2):139.

[10] 梅立军,周强等. 知网与同义词词林的信息融合研究[J]. 中文信息学报. 2005,19(1):63-70.

[11] Matthew H, Simon J, Georgina M. A Practical Guide To Building OWL Ontologies Using Protégé 4 and CO-ODE Tools(1.)[J]. (2007-10-16)[2008-02-27]. http://protege.stanford.edu,2001.

[12] 董振东. 语义关系的表达和知识系统的建造[J]. 语言文字应用,1998,(3):76-82.

[13] 刘杨,俞士汶,于江生. CCD语义知识库的构造研究[J]. 小型微型计算机系统. 2005,26(8):1411-1415.

[14] Melnik S, Garcia-Molina H, Rahm E. Similarity Flooding: A Versatile Graph Matching Algorithm and Its Application to Schema Matching[C]//Proceedings of the 18th International Conference on Data Engineering (ICDE), 2002: 117-128.

[15] Duchateau F, Bellahsene Z, Coletta R. A flexible approach for planning schema matching algorithms[M]. On the Move to Meaningful Internet Systems: OTM 2008. Springer Berlin Heidelberg, 2008: 249-264.

向春丞(1988—)，硕士研究生，主要研究领域为计算语言学。
E-mail：ccxiang@pku.edu.cn

詹卫东(1972—)，博士，教授，主要研究领域为现代汉语语法、计算语言学、语言知识工程。
E-mail：zwd@pku.edu.cn

穗志方(1970—)，通讯作者，博士，教授，主要研究领域为计算语言学、文本知识工程。
E-mail：szf@pku.edu.cn

中国中文信息学会语音专委会举办“见证言语工程(二)”纪念册发布会

2015 年 4 月 18 日，中国中文信息学会语音信息处理专委会在清华大学 FIT 大楼举办“见证言语工程(二)”纪念册发布会。

我国音韵学和语言学的研究有较长的历史，但言语工程、实验语音学的研究只有几十年历史。面对世界高技术蓬勃发展、国际竞争日益激励的严峻挑战，国内一批专家开创了言语相关的研究。“见证言语工程”纪念册(第二册)收录了中国社会科学院鲍怀翘研究员、同济大学计算机系柴佩琪教授、中国科学院声学所李昌立研究员、中国科学院自动化研究所陈道文研究员、清华大学计算机科学与技术系吴文虎教授、中国科学院声学研究所吕士楠研究员和中国社会科学院语言研究所曹剑芬研究员等七位 80 岁以上老专家的事迹，内容包括老专家自述语音研究历程、科研成果、学术论著和个人感悟等，是我国言语和语音信息处理珍贵的历史记录和见证。

纪念册收录的 80 岁以上言语工程领域的老专家们齐聚发布会，共同见证我国言语工程前进的风雨历程，一一讲述了“见证言语工程(二)”产生的经过，撰写时的感触。

此次发布的《见证言语工程(二)》是 2013 年 4 月发布的《见证言语工程(一)》纪念册的续册。《见证言语工程(一)》收录了方棣棠、张家騄、袁保宗、徐近霈、黄泰翼和林茂灿等六位时年 80 岁以上老专家为我国言语工程领域所做的开创性的工作。该系列的纪念册“前言”由中国科学院院士、清华大学教授张钹撰写；“题字”有中文信息学会理事长、哈尔滨理工大学教授李生提写；由蔡莲红教授整理完成。Dolby 公司赞助了该系列纪念册的出版及发布。袁保宗教授作为第一册的代表，参加了本次发布会。

参加此次发布会的人员包括学会副秘书长杨尔弘教授、专委会主任清华大学郑方教授、专委会前主任清华大学蔡莲红教授、专委会副主任中科院自动化所陶建华研究员、哈尔滨工业大学韩纪庆教授、专委会秘书长清华大学贾珈副教授，全国人机语音通讯学术会议常设机构委员会主席团成员北京交通大学朱维彬教授、清华大学徐明星副教授和王东博士，以及 30 余位师生代表，蔡莲红教授主持了发布会，杨尔弘教授代表中国中文信息学会、郑方教授代表语音专委会分别致辞。