领域 Ontology 的自动丰富——基于 ADL 地名表的实例研究*)

葛 宁 王 军
(北京大学信息管理系 北京 100871)

摘 要 本文以一个地理特征词表(Feature Type Thesaurus,FTT)为研究实例,提出了一种对领域 Ontology 进行自动丰富的方法。FTT 描述了 200 多种地理特征类型,依照等级结构组织,用于标引和组织美国亚历山大数字图书馆地名表(ADL Gazetteer)中的 6 百万个地名。为了对 FTT 进行自动丰富,(1)首先从地名中抽取和发现有检索价值的、表示地理特征类型的通用词;(2)根据它们和标引主题词间的同现关系,在相同词族词汇的聚类过程中,确定与之相对应的主题词,进而将提取出的通用词定位到 FTT 的等级结构中。充分利用已经存在的大量标引语料,实现通用词的定位分析是核心内容,并且实验结果证明有效性达到 82.7%。这项研究的实质是从 Ontology 标引的语料库中自动提取领域知识和标引知识,达到对 Ontology 的自动丰富。这一方法可以应用到类似的语料库和知识库上,实现新术语的发现、Ontology 自丰富及其互操作。

关键词 领域 Ontology,自动丰富,词汇抽取,通用词,地名词典

Automatic Enrichment of Domain Ontology——A Case Study on ADL Gazetteer

GE Ning WANG Jun
(Department of Information Management, Peking University, Beijing 100871)

Abstract The utility of domain ontologies has been increasing in recent years. A critical issue for their wider applications is the automatic enrichment of domain ontology, i. e. to enrich them with new terminologies and relationships to reflect the ever advancing of domain knowledge. Based on the large scale experiment on ADL Gazetteer, one of the biggest digital gazetteers over the world, a solution for this problem is proposed. ADL Gazetteer is a digitalized worldwide gazetteer developed in the Alexandria Digital Library (ADL) Project, which contains millions of geographic names (place names). The place names are indexed with type terms from the ADL Feature Type Thesaurus (FTT), a hierarchical category scheme, which is the ontology study case in this research. To discover generic terms from place names and to set up correlation between the extracted generic terms and the corresponding type terms in the FTT, a pointwise-mutual-information motivated method is used to extract frequent words/phrases from the place names, and a hierarchical clustering algorithm is used to identify the generic terms from the extracted and to determine the concepts in the FTT to which the generic terms are correlated. The effectiveness of the experiment reached 82.7%. The proposed approach can be applied upon other similar ontology-indexed corpora, such as dictionaries and catalogs, and served as an assistant tool for end-user search, terminology discovery, ontology enrichment, and interoperation among different ontologies.

Keywords Domain ontology, Automatic enrichment, Term extraction, Generic term, Digital gazetteer

1 引言

随着语义网络(Semantic Web)的提出和深入研究,Ontology(本体)得到了普遍的重视和应用。虽然描述人类基本知识框架的通用 Ontology 尚在发展阶段,领域 Ontology 在信息组织、信息集成、智能系统等领域已经得到了应用,并且取得了较好的效果。领域 Ontology 是对某个领域关键知识的形式化描述,通常是由领域专家制定的。Ontology 在应用中要解决的一个关键问题是如何自动发现领域中新出现的概念和关系,对 Ontology 进行自动丰富[1]。为解决此问题,我们以一个地理特征词表(Feature Type Thesaurus)为研究对象,在深入研究和大规模实验的基础上,提出了一种基于已标引语料库的、现实可行的 Ontology 自动丰富方法。

目前见到的 Ontology 自丰富方法基本上可划分为 4 类:集成[2,3]、语义映射[4]、关系扩展[5,6]和词汇丰富[7]。集成是指将属于同一领域的不同分支或者不同抽象层次的 Ontology 连接在一起,文[2]提出了知识集成的两种方式和全局知识模型的概念。文[3]提出利用 CORBA(公共对象请求代理结构)技术,实现不同领域、不同语种、分布在不同数据库中的词表的互操作,来联合标引和检索数字对象。这一方法的优点是可充分利用现有 Ontology 资源,缺点是不能及时发现新的术语、反映领域的最新发展。语义映射的方法需要针对领域编制特定规则,本质上还是一种手工的方式。文[4]从词表创建的初衷来考查不同词汇及它们之间关系的差别,在词表等级关系基础上,提出不同词表间最理想的 4 条映射规则,从而建立由一个词表到另一个词表明确的映射。关系扩展和词

*)国家自然科学基金项目(70303002)。葛 宁 硕士,主要研究领域为数字图书馆、知识组织、信息检索、文本挖掘;王 军 博士,副教授,主要研究领域为数字图书馆、知识组织、文本挖掘。

汇丰富是 Ontology 自丰富较常用的方法。文[5]作为 OASIS (Ontologically Augmented Spatial Information System,"走向 Ontology 的空间信息系统")项目的一部分,讨论了在地名词典和地理词表的等级上扩展相关关系的优势和可能性,以及对相关关系的语义距离度量。文[6]从中文文档里提取出关键词,然后利用词与词之间的同现关系,计算它们的相似度,最后自动生成具有一定关联关系的词表。文[7]在语料库中根据语法规则抽取领域术语,用同义词词典合并同义词,并识别具有相同句法结构的词语,分别形成相应的小类,再利用语义词典进行整合,最终建立词间关系。

本文提出一种基于被标引语料库的方法。这种方法与上述已有方法相比,优势在于充分利用了现实标引语料的两个作用:(1)作为发现新术语、新概念的来源,即领域知识;(2)作为发现的新术语/概念和 Ontology 中规范的术语/概念之间映射的中介,即标引知识。由于被标引语料的广泛存在(例如元数据、类目表、索引、产品分类等),这一方法可被广泛地应用于词表、分类表、类目结构等 Ontology 的自动丰富和更新。

在现实语料基础上,本研究的目标是从地名中抽取有检索价值的通用词,并分析它们和标引主题词间的同现关系,据此在等级层次结构上对地理特征词表进行词汇的丰富与扩充。其中,术语抽取已经有比较成熟的方法[17,18]可以借鉴;如何对标引结构进行分析、将经过抽取和筛选的通用词定位到词表中,是本研究的核心内容。

下面,首先介绍本文的研究背景和数据源,第 3 部分是算法分析与设计,第 4 部分是实验和结果分析,最后是对进一步研究的介绍和应用展望。

2 研究背景

亚历山大数字图书馆(Alexandria Digital Library, ADL)是美国最大的、持续时间最长的数字图书馆研究项目之一,自 1995 年起连续得到了美国数字图书馆倡导计划(Digital Library Initiative)长达 10 年的资助。作为 ADL 项目的重要组成部分,ADL 地名表(ADL Gazetteer, ADLG, http://www.alexandria.ucsb.edu/gazetteer/)收录了 4,437,405 个覆盖全球的地理实体和所涉及的 5,947,611 条地名、别名或历史地名以及与之相关的地理空间信息,形成了一个庞大的地理信息系统[8]。ADLG 的一大特色是"它的每一个地名都被赋予某种地理类型,以便对地名进行归类和检索、区分同名的地理实体。为此,ADL 专门设计了一个描述地名类型的主题词表——《地理特征主题词表》(Feature Type Thesaurus, FTT)。FTT 合并了美国图像地图机构(National Imagery and Mapping Agency, NIMA)和美国地质调查部门(U. S. Geological Survey, USGS)的两个地理词表,并进一步扩充而成。FTT 是严格按照树形等级结构组织排列的,最深处有 5 层。它收录了 210 个正式词和 1046 个非正式词,分属行政区划(administrative areas)、水文特征(hydrographic features)、陆地(land parcels)、人工特征(manmade features)、地文学特征(physiographic features)和区域(regions)6 个词族[9]。词族是指一组具有等级关系的主题词集合,6 个族首词分别作为树形结构的根结点,其他所有主题词都依据属分的上下位等级关系,组织到父结点(上位词)和子结点(下位词)中。表 1 给出了一个 FTT 主题词的例子。

表 1 FTT 主题词"cities"(城市)

cities	
Used for (非正式词):	muniipalities(都市), towns(城填), urban aread (市区)
Broader Terms (上位词):	populated places(人口密集地区)
Narrower Terms (下位词):	Capitals(首都)
Related Terms (相关词):	Metropolian Statistical Areas(都市统计区)
Scope Note (范围注释):	For smaller, less formally established communities, use'populated laces' For independent cities, use'countries, 2nd order divisions'
Definition (定义):	Incorporated populated places. [Adapted from USGS Circ 1048]

作为支撑 ADLG 的语义框架,FTT 在应用中遇到了一些问题:

(1)用户查询时,为使用正确的主题词来检索,需要预先熟悉词表内容,给用户造成了额外的负担;

(2)标引员对地名进行标引和归类时,从 FTT 中选择恰当的主题词是一项十分繁重的任务;

(3)FTT 词表本身也需要不断维护和更新,在应用中随时更新和增补新的主题词。

为解决上述问题,北京大学信息管理系和美国亚历山大数字图书馆项目组结成联合研究小组,由美方提供原始数据并验证实验结果,中方研究、设计算法并进行大规模实验。历时半载,为上述问题提供了良好的解决方案:

(1)大多数地名都属于名词性词组,通常有一个专有的修饰语和一个表达类型的通用词汇构成,例如"Santa Barbara High School"。可利用这一特点从地名中提取通用词汇。

(2)通过分析提取出的词汇和用于标引地名的主题词之间的同现关系,从抽取出的短语集合中挑选有价值的描述地理类型的通用词,并确定与它相关联的主题词。例如:"High School"(高中)——"educational facilities"(教育设施,FTT 正式主题词)。

(3)根据 FTT 的等级层次结构,最终筛选出有价值的表示地理特征类型的通用词,作为对应主题词的同义词或下位词,定位在 FTT 词表中,实现对 FTT 的自动更新和丰富。通用词和主题词间的关联关系,既可为用户查询 ADLG 和 FTT 提供词汇帮助,又可用来分析地名的构成成份,通过从地名的通用成份到正式主题词的自动映射,实现对地名的自动标引。

在 ADLG 的 5,947,611 个地名上所进行的实验,共提取出 1036 个通用词。排除评测人员无法识别的非英语地名成份和关联到"populated places"的词汇外,共评估了 446 对通用词与主题词的关联关系。经评测,正确率达到 82.7%,证明上述方法是可行的、有效的。

3 算法设计

通常地,一个地名由专有名词和通用名词组合而成,比如"Great Makalakari Lake"、"Grantwood Memorial Park",以及"Lodi Marsh State Wildlife Area"等。像"Lake"(湖泊)、"Memorial Park"(纪念公园)、"State Wildlife Area"(国家野生区域)等通用地理名词在 ADLG 中很常见,与某些特定主题词间的关联也比较强,甚至很稳定。在此情形下,完全可以从

ADLG中挖掘出通用词，达到自动丰富主题词表的目的。事实上，USGS的地理名称信息系统(Geographic Name Information System，http://geonames.usgs.gov/)，已经将地名中的通用词与特定类别相映射，作为标引工作方针之一[8]。

根据研究目标，实验着力要解决两大障碍：(1)抽词：从地名中抽取典型的通用词；(2)定位：把抽取出来的通用词作为某个主题词的同义词或下位词放到现有FTT词表中。定位是整个算法的核心。

3.1 抽词

因为地理主题词是对地名中具有相对独立语义的词汇的再确认，所以要求抽取出来的词汇满足条件：(a)是独立的语义单元；(b)语义上相对完备。从一个地名中抽取出来且符合这两个要求的地名词片断，以下称之为候选词。例如，地名"Santa Barbara High School"中，同时满足上述条件的候选词是"High School"(高中)，应当用FTT的主题词"educational facilities"(教育设施)标引。

考虑到ADLG地名库规模十分庞大，不失一般性，若抽取出来的词或词组W在整个地名库中出现的频次$frequency(W)$大于某个初始阈值θ，则该结果具有一定的语义独立性，即：$frequency(W)\geqslant\theta$，是$W$为候选词的必要条件。

单纯频次验证的计算代价太高，我们利用"文本互信息"(Pointwise Mutual Information)[10,11]来辅助寻找地名中名词性词组的边界。文本互信息记作PMI：

$$PMI(w_i,w_{i+1})=\log_2\frac{P(w_iw_{i+1})}{P(w_i)P(w_{i+1})},\quad P(w_i)P(w_{i+1})\neq 0 \tag{3.1.1}$$

其中，$P(w_iw_{i+1})$是相邻单词w_i、w_{i+1}共现的概率，$P(w_i)$、$P(w_{i+1})$是w_i、w_{i+1}各自出现在地名中的概率。采用最大似然估计(MLE)计算概率，上式变为

$$PMI(w_i,w_{i+1})=\log_2\frac{frequency(w_iw_{i+1})/N}{frequency(w_i)/N\cdot frequency(w_{i+1})/N} \tag{3.1.2}$$

其中，$frequency(w_iw_{i+1})$表示词串w_iw_{i+1}在整个地名库中出现的频次，N为地名库中的单词总数量。

PMI在中英文语料上的表现都令人满意。文[12]在中文语料上考查了频次、PMI、似然比对数(Log-Likelihood)、χ^2、z-score等9种单个统计量的抽词性能，认为PMI表现最佳。文[13]指出PMI良好地度量两个单词相互间的独立性。PMI值越大，w_i、w_{i+1}之间邻接结合的可信程度越强；反之亦然。短语边界的划分使得提取出的词汇的语义更加完备。

具体应用PMI的方法比较多样：文[14]为解决抽词歧义将PMI与t-检验指标结合；文[15]将它与似然比对数(Log-Likelihood Ratio)结合发现新词，克服在稀疏数据上的应用问题；文[16]将它与朴素贝叶斯方法结合，同时考虑到"词频-逆文献频次"加权指标(tf^*idf)、词组第一次出现距离文首的单词数、关键词在标引中被使用的频次，借助搜索引擎反馈的结果来估计PMI值。考虑到ADLG语料有别于全文和地名通用词出现频次的特点，我们采用了比较简单而有效的方法——词频验证同时辅以基于PMI的短语边界划分的抽词方法。与文[17,18]类似，实验中将地名中PMI值最小处作为词组划分的边界。在大多数语言中，偏正结构的名词性词组使用得较多，故可以选取文本序列左起第一个最小的PMI值作为起始切分点。

综上，算法3.1采取依次从词汇序列中的PMI值最小处切分地名，寻找出现频次大于阈值θ的部分，完成抽词。不难发现，θ越高，候选词长度越短；词与词间的PMI值越高，得到的候选词会越长。同时，设置最低阈值θ也有助于克服PMI在稀疏数据上应用的缺陷[19]。

算法3.1 将原始地名切分成若干候选词

(抽词算法)

输入：一个原始地名字符串；根据式(3.1.2)计算得到的该地名中所有相邻两个单词的PMI值。
输出：切分地名后得到的若干候选词token。
string：字符串类型，用来存储切分的中间结果。
frequency(string)：函数，返回string在整个地名库中出现的频次。
PMI[1,…,k]：有序数组，PMI[i]用来存储PMI(w_iw_{i+1})的计算结果，数组按w_i在string中的出现从左到右依次排序。
breakPoint：整型，表示string中第i个和第(i+1)个单词之间作为切分点。

```
BEGIN
1 设定初始阈值=θ;
2 变量 string = 原始地名;
3 if (frequency(string)>=θ)
4   return token=string;
                  /* token 是候选词 */
5 else if (string 的词长>=2)
6  {k=string 的词长-1;
7    将 string 中所有相邻两个单词的 PMI 值按序存入 PMI[1,…,k];
8    breakPoint=1;
9    for (i=2 to k)
10    {if (PMI[breakPoint]>PMI[i])
11          breakPoint=i;
12    }
13   在 breakPoint 处把 string 分成两个子字符串 string1、string2;
14   string=string1; goto 第3行;
15   string = string2; goto 第3行;
16 }
END
```

对于词组W，W作为候选词抽取出来的次数$TokenCount(W)$应小于或等于W在地名中所有出现的频次$frequency(W)$。因此，算法3.1运行后有可能出现$TokenCount(W)<\theta\leqslant frequency(W)$的情形，需要将算法3.1中的$frequency$函数替换成$TokenCount$函数再次运行，直至所有候选词的$TokenCount$均不低于$\theta$。

3.2 定位

定位作为基于语料库实现词表自动丰富的关键步骤，是考察从地名里抽取出来的候选词与标引这个地名的主题词之间的关联强度，确定与该候选词关联最密切的主题词。据此，筛选有价值的候选词并将它们定位在FTT中。定位的好坏，关键是看候选词与定位到的主题词之间：(a)关联的稳定性、(b)关联的强度和(c)关联的性质。

所有抽取出来的候选词定义了定位所涉及的论域U。对于FTT词表的每一个主题词，都可以在U上定义一个模糊集合T。设k为U上的任意一候选词，T是某一地名主题词，所有被T标引的候选词形成集合T，则k对T的隶属度可通过如下隶属函数得出：

$$\mu_T(k)=\begin{cases}\dfrac{IndexedCount_T(k)}{TokenCount(k)} & (\forall k\in U,k\text{ 被 }T\text{ 标引})\\ 0 & (\forall k\in U,K\text{ 没有被 }T\text{ 标引过})\end{cases} \tag{3.2.1}$$

其中，$\mu_T(k)$称为k对T的原始隶属度，$TokenCount(k)$表示k在整个地名库中被抽取出来的次数，$IndexedCount_T(k)$表示那些抽取出k且被T标引的地名的个数。

原始隶属度$\mu_T(k)$非常小的k出现在集合T的边缘附近，我们认为这时k与T关联有极大的偶然性、不稳定。为消除T接近边缘部分的噪音干扰，设定阈值$\rho(0<\rho<<1)$，将低于ρ值的隶属度规定为0，即：

$$\mu'_T(k)=\begin{cases}\mu_T(k) & (\mu_T(k)\geqslant\rho)\\ 0 & (\mu_T(k)<\rho)\end{cases} \tag{3.2.2}$$

现在给出置信水平 $\lambda(0<\lambda\leqslant 1,\lambda>\rho)$，便可以得出 T 的 λ-截集 $T_\lambda(T_\lambda=\{k|k\in U,\mu'_T(k)\geqslant\lambda\})$。参数 λ 规定了候选词 k 与主题词 T 建立关联的最低强度阈值，它对集合 T 做出了划分，用于圈定 T 的核心区域 T_λ。若 $k\in T_\lambda$，则称把 k 成功定位到主题词 T 上。

如果一个候选词被多个有着共同的直接上位主题词所标引，那么定位到它们的直接上位词也是合理的。因为上位词的内涵小、外延大，在语义上包含了下位词，所以人们总是可以用上位词的概念来指称下位词所表达的事物，并且在组织得当的概念体系下不会出差错或误解。在实践中，对于多主题的标引也往往是这样处理的。像《杜威十进分类法》(Dewey Decimal Classification，DDC)就规定标引 3 个以上主题时使用概括性的上位类(Rule-of-Three)[20]。《美国国会图书馆标题表》(Library of Congress Subject Headings，LCSH)，也遵循更倾向于运用概括性标引而非深度标引的策略[21]。

因此，可以构造一个更加符合标引实际的隶属函数。主题词 R 是主题词 T 的一个下位词，记作：$R=narrow(T)$，类似地，T 的上位词记作 $broad(T)$。若 λ 被认为是一个可接受的概念关联强度，那么可以对(3.2.2)式进行扩充，对抽出来的词 $k\in U$，递归定义一个优化后的 k 对 T 的隶属函数(即关联强度)和定位算法：

$$\mu^*_T(k)=\begin{cases}\mu'_T(k) & (narrow(T)=\Phi)\\ \mu'_T(k)+\sum_{\forall R}\mu^*_R(k) & (R=narrow(T)\neq\Phi)\end{cases}\qquad(3.2.3)$$

特别地，若 $\mu^*_T(k)>1$ 则令 $\mu^*_T(k)=1$。其中，$\mu^*_T(k)$称为 k 对 T 的优化隶属度；$narrow(T)$表示主题词 T 的一个下位词，$narrow(T)=$空集 Φ 表示 T 没有下位词。同样地，$broad(T)=\Phi$ 表示 T 没有上位词，即 T 是一个族首词。优化隶属度 $\mu^*_T(k)$表达了候选词 k 与主题词 T 的关联强度，它是在排除所有原始隶属度小于 ρ 的干扰关联后，将以 T 为根的子树的所有叶子结点的原始隶属度累加得到的。

算法 3.2.1 递归给出候选词对某主题词的优化隶属度

```
输入:一个候选词 k;某主题词 T。
输出:候选词 k 对主题词 T 的优化隶属度 μ*T(k)=Membership(T,k)。
T.weight(k):函数,根据式(3.2.1)返回 k 对 T 的原始隶属度 μT(k)。
T.narrowSet:集合,用来存储 T 的所有下位词。
BEGIN
1 Membership(T,k)/*根据式(3.2.3)*/
2 {if (T.weight(k)<ρ)
                    /*根据式(3.2.2)*/
3 T.weight(k)=0;
4  if (T.narrowSet=null)
5     return T.weight(k);
6  else
7  {for each R in T.narrowSet
8   {T.weight(k)= T.weight(k)+Membership(R,k);/*递归到
    第 1 行*/
9   }
10   return T.weight(k);
11 }
12 }
END
```

算法 3.2.2 将候选词定位到主题词(定位算法)

```
输入:一个候选词 k;所有主题词的集合 T_Set。
输出:定位到的主题词集合 resultSet。
T.weight(k):函数,根据式(3.2.1)返回 k 对 T 的原始隶属度 μT(k)。
T.broadSet:集合,用来存储 T 的所有上位词。
T.narrowSet:集合,用来存储 T 的所有下位词。
Membership(T,k):函数,返回候选词 k 对某主题词 T 的优化隶属
                度,见算法 3.2.1。
BEGIN
1 for each T in T_Set
2 {if(T.broadSet = null)
3  {if(T.weight(k)>=λ)
4   将 T 加入 resultSet;
5  }
6  else if(Membership(T,k)>=λ)
7  {for each R in T.narrowSet
8   {if (Membership(R,k)>=λ)
9      取下一个 T 并 goto 第 2 行;
10  }
11  将 T 加入 resultSet;
12 }
13 }
14 return resultSet;
END
```

综上，对候选词 k 和主题词 T，如果满足下列条件之一：

(1)$\mu_T(k)\geqslant\lambda$；并且 $broad(T)=\Phi$ 或者

(2)$\forall R=narrow(T)\mu^*_R(k)<\lambda\wedge\mu^*_T(k)\geqslant\lambda$，并且 $broad(T)\neq\Phi$；

则 k 成功定位到主题词 T 上。

算法 3.2.2 在每个词族中，试图选出一颗子树，使得这棵子树根结点的关联强度 $\mu^*_T(k)\geqslant\lambda$，且它在词表中的层次最深，那么所选子树的根结点就是应定位到的主题词；例外的是，若子树以族首词为根结点，那么要求没有累加过的原始隶属度 $\mu_T(k)$也$\geqslant\lambda$。

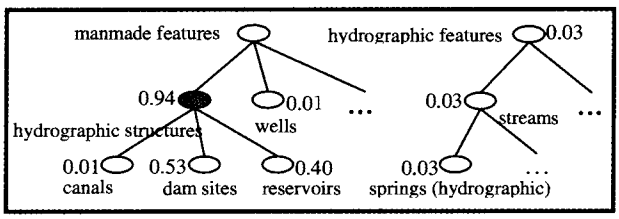

图 1 "Water Supply"对各 FTT 主题词的优化隶属度

图 1 简单说明了这一过程。首先，所有与候选词"Water Supply"(淡水贮藏)相关的主题词依据 FTT 词表结构被分在了不相关的两棵树上。在 $\rho=0.010$、$\lambda=0.9$ 的情况下，在第一树中由"hydrographic structures"(水文建筑结构)、"canals"(水渠、运河)、"dam sites"(坝址)、"reservoirs"(蓄水库)、"wells"(水井)形成的子树其根结点的优化隶属度值 0.01+0.53+0.40=0.94 最先高于 λ，且位置最深。又，"hydrographic structures"不是 FTT 的一个族首词。因此，"Water Supply"被定位到主题词"hydrographic structures"。

为确保关联的准确性以及在合理的语义距离偏差范围内，算法 3.2.2 一方面尽可能地利用 FTT 词表的树形结构，另一方面又控制了隶属度向上簇聚累加的层次，以期最大限度地将抽词得到的候选词定位到关联密切的主题词上。鉴于 FTT 词表结构设计中，族首词的概念过于宽泛、与其下位词的语义跨度太大，因此算法 3.2.2 没有把隶属度簇聚累加到族首词，即 FTT 词族的根结点。

需要指出的是，并非所有的候选词都能定位到 FTT 上。像"Saint"、"San Jose"、"United States"(美国)这样抽取出来的在 ADLG 中较为常见的专有名词，与能指示地名类型的通用词不同，它们出现在各种类型的地名中。例如，"San Jose"出现在 2998 个地名中，包括"San Jose Plaza"、"San Jose Gun Club"、"San Jose Mill"、"San Jose Ranch"、"San Jose River"、"San Jose Speedway"等等。使用"San Jose"的地名分散于各个词族，因此在 λ 相对较高的情况下，很难找到与之足够关联的主题词。而如果 λ 较低，那么会增加选出专有名词的风险。一般来说，λ 至少应大于 0.5，实验中选定在 0.9。定位的过程起到了对候选词中专有名词的过滤。

考虑到定位算法是遵循词表的等级结构逐步向上簇聚、收敛的过程，被定位的通用词大多数应当是所关联主题词的

下位词或者同义词。4.4 节的专家评估结果也印证了这一点。从概念性质上讲，丰富后通用词可以认为是相应主题词所代表的抽象概念的一个具体实例，作为主题词表的入口词。

抽词和定位虽是两步，但是二者又相互弥补，是一个整体：抽词是定位的前处理，定位还包含通用词的筛选，完成抽词未竟的功能。

4 实验及分析

实验在一台高性能 PC 机和关系数据库中完成。测试语料库是包含 5,947,611 个地名的 ADLG 以及具有 210 个正式主题词的 FTT 词表。大致分为 3 步：(1)数据规范化，(2)候选词抽取，(3)同现关联构建。下面，逐一讨论这 3 个步骤以及实验结果并分析错误。

4.1 数据规范化

ADLG 中将近 15%的地名都采用了倒置的表达方式。所谓“倒置”，就是在一个由修饰词和名词性词组构成的地名中，为了使位置靠后的名词性词组在数据库存储时获得索引入口，将修饰词和名词性词组逆序排列的做法(中间使用逗号“，”分隔)，例如：“Beach，Santa Monica”、“Angola，Republica de”。实验为考查纯统计方法的有效性，方便 *PMI* 的计算，需要将它们恢复成正常语序。

4.2 候选词抽取

为便于算法 3.1 的执行，建立了所有二元词对(相邻的两个单词)及其 *PMI* 值的索引表。ADLG 5,947,611 个地名总共使用了 2,141,805 个不同单词；其中半数以上单词的频次为 1。地名中所有的二元词对总共有 2,491,053 个，其中仅出现 1～2 次的占到 90.3%。图 2 体现了二元词对出现频次与数量分布的关系。通过分析发现，图 2 中拟合的幂函数曲线在频次 90 到 100 的区间内存在曲率($D=\left|\frac{y''}{(1+y'^2)^{3/2}}\right|$)最大值(见图 3)。考虑到频次特别高的二元词对的出现具有一定偶然性、曲线拟合时存在误差，为降低抽出专有名词的风险，故选取了较保守的值 100 作为识别低频词的界限，即令 $\theta=100$。

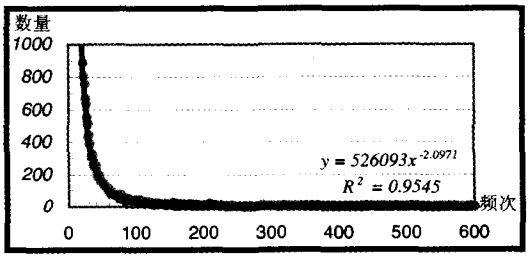

图 2 二元词对数量对词频的分布

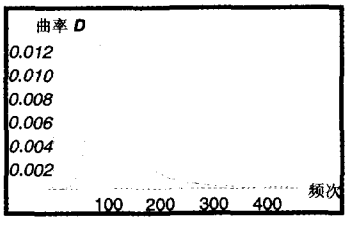

图 3 曲线 $y=526093x^{-2.0971}$(图 2 中)的曲率变化

对于那些含有 of、and、or、else 等连词或 at、the 等停用词(stop word)的词对，以及频次 *frequency*(w_1w_2)低于初始阈值 θ 的词对，它们的 *PMI* 值被预先设为 $-\infty$(实际计算中取 −1024)，大大减少了计算量。在 $\theta=100$ 时，最终抽取出 8805 个候选词，80%的频次在 100～500 之间。绝大多数候选词由 1、2 个单词组成(见图 5)。

4.3 同现关联构建

接下来构建同现关联，就是在每个候选词与一个主题词之间利用优化隶属度公式(3.2.3)建立起映射对，把候选词中的通用词定位到关联密切的主题词上。参数 ρ(算法 3.2.1 中)和 λ 的取值对算法 3.2.2 十分重要。

候选词对主题词的原始隶属度小于 ρ 的映射对，被认为是干扰，应当将其隶属度设为 0，排除在关联计算外。图 4 显示了不同 ρ 值下应当排除的映射对总数量。可以看到，$\rho=0.010$ 是分界：$\rho<0.010$ 时，映射对数量明显呈线性增长；$\rho>0.010$ 时呈对数缓慢增长。于是，实验取 $\rho=0.010$。

表 2 值及相应生成的映射对数量

λ	映射对数量
0.5	4787
0.8	1859
0.9	1040

表 2 列举了在 $\rho=0.010$ 时不同值 λ 下算法 3.2.2 所生成映射对的数量。在 $\lambda=0.9$ 的情况下，总计从候选词中挑出 1036 个通用词定位到 1040 个主题词上，这与 *FTT* 非正式词的数量 1046 相当。为了便于与 *FTT* 词表的非正式词比较，同时考虑到评估的工作量，实验只对 $\lambda=0.9$ 时的结果进行了人工评判。在这 1036 个通用词中，不经簇聚累加、直接定位到 *FTT* 上的有 958 个。

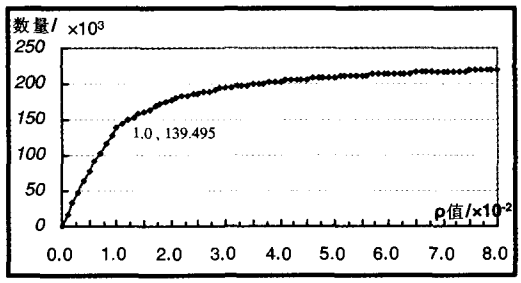

图 4 不同 ρ 值下应当排除的映射对数量

图 5 显示了不同词长通用词的数量分布情况，同时与所有抽取出来的候选词及 FTT 主题词的数量做了比较。从图中可见，通用词的数量与相应词长的 FTT 主题词数量基本成正比。

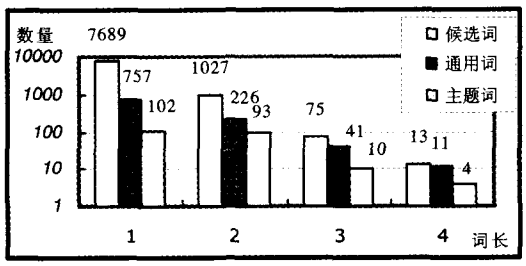

图 5 所有候选词、通用词和 FTT 主题词的数量对词长的分布

4.4 结果评估

在评估通用词定位结果时，遇到了两大障碍：

(1)“populated places”(人口密集地区)：有 407 个通用

词，超过1/3，被定位到“populated places”。这些通用词大多来自于非英语语言国家，比如“Casale”（意大利语的“小村庄”），“Kampong”（马来西亚语的“村庄”），“Ikot”（尼日利亚语的“城镇”）等。此现象与ADLG中45.11%的地理实体——多在美国本土以外——都被归类到“populated places”不无关系。

(2)西文地名：ADLG收录的地名来源于165个国家，不同语言的西文地名给评估工作带来了很大困难。尽管我们使用了包括15种语言（英语、法语、德语、荷兰语、意大利语、西班牙语、葡萄牙语、丹麦语、印度尼西亚语、塞尔维亚语、克罗地亚语、汉语拼音、梵语、瑞典语、阿尔巴尼亚语）在内的Babylon多语言在线电子词典（http://www.babylon.com）对每个通用词进行查找，仍有187个通用词不能确定含义。

以上两类通用词定位的情形被排除在最终评判之外。实际上，真正评估的通用词与主题词映射对数量为（1040－407－187＝446）对。

映射结果正确的评价标准是：(1)选出的通用词应为名词性词组；(2)建立的关联符合FTT词表的选词要求及该词表所描述的关系定义。依据此原则，由美方相关专家来负责评估。他们认为，通用词选取和定位主题词正确的映射对有369对，有效率达到82.74%，并且其中大多数都是下位词或同义词。表3给出了一些正确和错误的例子。

4.5 错误分析

正如在表3中看到的那样，错误的映射对大致上分为两类：

(1)常用修饰语：除了通用的名词性词组外，一些出现频次比较高的常用修饰语也被选出。例如，从地名中总共抽取“Adult”（成人的）110次，其中的101次都与“educational facilities”（教育设施）关联。这是因为由“Adult”构成的词组，比如“Adult School”（成人学校，31次）、“Adult High School”（成人高中，4次）、“Adult Learning Center”（成人学习中心，3次）、“Adult Vocational Center”（成人职业中心，2次）等，频次较低造成（括号中的所注数字为出现频次）。类似现象也发生在出自于“Sacred Heart Church”（圣心教堂）、“Sacred Heart School”（圣心学校）、“Sacred Heart Seminary”（圣心神学院）的候选词“Sacred Heart”（圣心）上，等等。

表3 实验结果举例

映射对		隶属度	是否*正确	地名举例**
通用词	主题词			
Access Area	rereational facilities	93.67	Y	Bluestone～
Childrens Home	buildings	95.31	Y	North Alabama～
Christian Methodist Episcopal Church	religious facilities	100.0	Y	Greater Saint Paul's～
Escarpment	cliffskl	98.66	Y	San Pedro～
Exit	transportation features	98.05	Y	Exit 132
Fish Hatchery	fisheries	90.82	Y	Valley City National～
FM(调频)	towers	99.98	Y	KA2XXZ-FM(Columbia)
Forest Resere	forests	96.60	Y	Marama Hill～
Forest Serice Station	buildings	99.76	Y	Michigan River～
Glacier	glacier features	90.15	Y	North Grant～
High School	educational facilities	99.76	Y	Jefferson～
Jiang(江)	streams	92.65	Y	Chang～(长江)
Memorial Gardens	cemeteries	98.56	Y	Sunset View～
Post Office	post office buildikngs	99.70	Y	Allen～(historical)
Shoping Center	commercial sites	98.93	Y	98th Avenue～
Stadium	sports facilities	93.70	Y	Athens～
State Recreation Area	parks	95.45	Y	Holiday Lake～
Stratit	channels	93.16	Y	Franklin～
UTC(协调世界时)	earthquake features	99.56	Y	Long Beach 1993-03-11 01:54 UTC
Water Supply	hydrographic structures	94.63	Y	～Number 3 Dam
Wildlife Management Area	reserves	95.35	Y	McClellan-Kerr～
Adut	educational facillties	91.82	N	Berkeley～School
Education	building	95.44	N	Wakeley Special～Center
Location	administrative areas	94.15	N	East Kano～;Kilome Sub-Location
Sacred Heart	building	92.52	N	～School ～Sanitarium

*Y代表正确，N代表不正确；**～替代相应的通用词。

(2)标引不一致：ADLG使用FTT词表时，存在对同一类型地名用不同的主题词标引的情形。举例来说（见表4），83.3%包含“Education”（教育）的地名被标引到“educational facilities”（教育设施）上，12.2%被归类到“buildings”（建筑物）。由于“building”是“educational facilities”（教育设施）的上位词，在λ=0.9时“Education”被定位到与之关联看起来并不是很紧密的“building”上。还有，像包含“Location”（地点）或“Sub-Location”的地名，2499个中有2349个标引到族首词“administrative areas”（行政区划）。因此，“Location”也不可避免地关联到与之差别比较大的主题词。

表4 含有“Education”（教育）地名的一些例子

地名	标引的主题词
Southwest Regional Educational Center	buildings(建筑物)
Cenrer for Continuing Education	buildings
Deartment of Education	buildings
Everett Special Education Center	educational facilities
Headland Secial Education school	(教育设施)
Eastern Washington College of Education	educational facilities
Western Eeducational Institute	educational facilities
	educational facilities

从上面可以看出，尽管根据标准被评判为错误的这些关联，在此特定领域和 ADLG 的应用中还是有意义的，对标引员和用户选择主题词也是有帮助的。

结论和展望 本研究从地名中抽取出有检索价值的通用词，然后依据关联统计分析将其中的通用词与 FTT 主题词建立映射。实验从已有的大量语料出发，围绕定位分析问题设计了整套算法，充分利用了 FTT 词表的结构来筛选和关联通用词，同时也考虑到原词表语义上的特点，具有中西文跨语言处理的通用性。实验结果显示，该方法的有效性达到 82.7%。

注意到 ADLG 使用了多个词表标引，包括 FTT、NIMA、USGS 三个词表。基于本研究的成果，可进一步解决不同 Ontology 的互操作问题。解决办法是：应用文中的方法对 NIMA 和 USGS 进行丰富，然后以通用词的关联映射为中介，实现不同地名特征词表中正式主题词的对应和转换；进而为没有结构的词表建立概念等级关系，甚至用来反映各类用户对地理、地质分类认识的差异。

本文提出的分析方法还可以应用到其它类似的语料库上，例如词典、百科全书、分类表、产品目录等，起到发现新术语、丰富词汇辅助系统、更新词表、帮助用户更友好地使用系统等作用。也可应用在数字图书馆、网页博物馆、电子商务网站、企业知识库等各种信息资源组织系统中。

参 考 文 献

1 Velardi P, Fabriani P, Missikoff M. Using text processing techniques to automatically enrich a domain ontology. In: Proceedings of the International Conferenc on Formal Ontology in Information Systems. New York: ACM Press, 2001. 270～284
2 和延立，杨海成，何卫平，等. 信息集成与知识集成. 计算机工程与应用，2003，4：38～41
3 Kramer R, Nikolai R, Habeck C. Thesaurus Federations: Loosely Integrated Thesauri for Document Retrieval in Networks based on Internet Technologies. International Journal of Digital Libraries, 1997, 1(2): 122～131
4 Doerr M. Semantic Problems of Thesaurus Mapping. Journal of Digital Information, 2001, 1(8). http://jodi. ecs. soton. ac. uk/Articles/v01/i08/Doerr/ (accessed Sep 5, 2005)
5 Tudhope D, Alani H, Jones C. Augmenting Thesaurus Relationships: Possibilities for Retrieval. Journal of Digital Information, 2001, 1(8). http://jodi. ecs. soton. ac. uk/Articles/v01/i08/Tudhope/ (accessed Sep 5, 2005)
6 Tseng Y H. Automatic Thesaurus Generation for Chinese Documents. Journal of the American Society for Information Science and Technology, 2002, 53(13): 1130～1138
7 裴炳镇，陈晓明，胡熠，等. 一种建立中文概念分类关系的新算法. 计算机工程与应用，2004，36：18～21
8 Hill L L, Frew J, Zheng Qi. Geographic Names - The Implementation of a Gazetteer in a Georeferenced Digital Library. D-Lib Magazine, 1999, 5(1). http://www. dlib. org/dlib/january99/hill/01hill. html (accessed Sep 5, 2005)
9 Hill L L. Metadata for the ADL Feature Type Thesaurus, 2004-11-15. http://www. alexandria. ucsb. edu/gazetteer/FeatureTypes/FTT_metadata. htm (accessed Sep 5, 2005)
10 Church K W, Hanks P. Word Association Norms, Mutual Information and Lexicography. In: Proceedings of the 27th Annual Conference of the Association of Computational Linguistics. New Brunswick, New York: Association for Computational Linguistics, 1989. 76～83
11 Church K W, Gale W A, Hanks P, et al. Using statistics in lexical analysis. In: Uri Zernik. Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1991. 115～164
12 罗盛芬，孙茂松. 基于字串内部结合紧密度的汉语自动抽词实验研究. 中文信息学报，2003，17(3)：9～14
13 Church K W, Gale W A. Concordances for parallel text. In: Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research. Using Corpora. Berkeley, California: Association for Computational Linguistics, University of California, 1991. 40～62
14 孙茂松，黄昌宁，邹嘉彦，等. 利用汉字二元语法关系解决汉语自动分词中的交集型歧义. 计算机研究与发展，1997，34(5)：332～339
15 刘建舟，何婷婷，骆昌日. 基于语料库和网络的新词自动识别. 计算机应用，2004，24(7)：112～134
16 Turney P D. Coherent keyphrase extraction via Web mining. In: Proceedings Eighteenth International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 2003. 434～439
17 Yang C C, Luk J W K, Yung S K, et al. Combination and Boundary Detection Approach for Chinese Indexing. Journal of the American Society for Information Science (Special topic issue on digital libraries), 2000, 51(4): 340～351
18 张国煊，郁梅，王小华. 基于互信息的汉语短语边界划分. 杭州电子工业学院学报，1995，15(1)：1～5
19 Manning C D, Schütze H. Foundations of Statistical Natural Language Processing. London: The MIT Press, 1999. 182
20 Chan L M, Comaromi J P, Mitchell J S, et al. Dewey Decimal Classification - A Practical Guide. Second Edition. New York: Forest Press, 1996. 55
21 Chan L M. Library of Congress Subject Headings - Principles and Application. Third Edition. Colorado: Libraries Unlimited, 1995. 30

(上接第 127 页)

其中 CDDS 的平均纯度 0.233，CMTC 的平均纯度 0.228。可以认为没有优势。Text(O)的比较结果如图 4 所示。

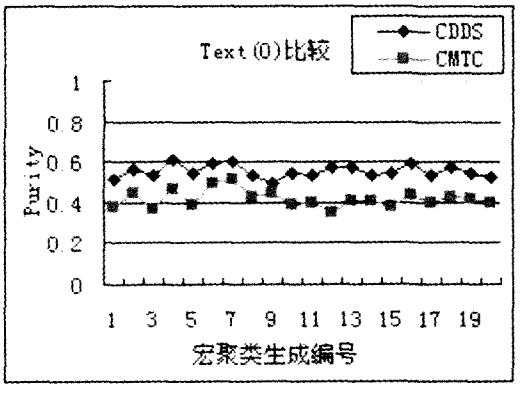

图 4 Text(O)实验数据的结果比较

其中 CDDS 的平均纯度 0.53，CMTC 的平均纯度 0.39，有 0.14 的优势。可以看到，当存在孤立点的时候，该方法具有更大的比较优势。

结束语 本文提出了一种文本流聚类算法，该方法能够在具有演化特征文本流中进行聚类，并且对孤立点不敏感。新的在线微聚类结构提高了聚类的性能；将微聚类分为潜在微聚类和异常微聚类，提高了算法对孤立点的适应能力。实验表明，文中的方法比已有的方法有更好的聚类质量。

参 考 文 献

1 Guha S, Mishra N, Motwani R, et al. Clustering Data Streams. In: IEEE FOCS Conference, 2000
2 O'Callaghan L, et al. Streaming-Data Algorithms For High-Quality Clustering. Wiley Series in Probability and Math. Sciences, 1990
3 Guha S, Rastogi R, Shim K. CURE: An Efficient Clustering Algorithm for Large Databases. In: ACM SIGMOD Conference, 1998
4 Aggarwal C C, Han J, Wang J, et al. A Framework for Clustering Evolving Data Streams. In: VLDB Conference, 2003. 81～92
5 Aggarwal C C. A Framework for Diagnosing Changes in Evolving Data Streams. In: ACM SIGMOD Conference, 2003. 575～586
6 O'Callaghan L, Mishra N, Meyerson A, et al. Streaming-Data Algorithms For High-Quality Clustering. In: ICDE Conference, 2002. 685～696
7 Ordonez C. Clustering Binary Data Streams with Kmeans. DMKD'03, San Diego, CA, USA, June, 2003
8 Aggarwal C C. A Framework for Projected Clustering of High Dimensional Data Streams. In: Proceedings of the 30th VLDB Conference, Toronto, Canada, 2004
9 Aggarwal C C. A Framework for Clustering Massive Text and Categorical Data Streams
10 Cao Feng. Density-Based Clustering over an Evolving Data Stream with Noise. In: Proceedings of the 2006 SIAM Conference on Data Mining (SDM'2006)
11 朱蔚恒，等. 基于数据流的任意形状的聚类算法. 软件学报，2006(3)：379～388