基于异构中文在线百科的层次话题构建

王煦中[1] 刘 琰[1] 胡琳梅[2] 陈 静[1]
(数学工程与先进计算国家重点实验室 郑州 450002)[1] (清华大学计算机科学与技术系 北京 100084)[2]

摘 要 中文在线百科包含大量有价值的信息,很多工作成功地将其用于各类知识获取任务。例如,拥有相似话题的文档可以被归为一个概念。从这些在线百科中构建出的针对某一概念的层次话题对于搜索与浏览、信息组织和检索等应用都有很大的帮助。然而,目前尚未出现对在线百科中某一概念层次话题构建的研究。针对中文在线百科的异构性与粗糙性的问题,提出了一种基于贝叶斯网络的话题层次构建方法。该方法同时综合文档的结构化目录信息和非结构化文本信息,采用最大树形图算法自动地在文档所属概念的贝叶斯话题网络中建立层次话题。实验证明,与原有的百科话题结构相比较,所提方法在保持75%的准确性的同时扩充了4倍的内容。

关键词 中文在线百科,层次话题,结构化目录信息,非结构化文本信息
中图法分类号 TP301 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.05.040

Building Hierarchical Topic Based on Heterogeneous Chinese Online Encyclopedia

WANG Xu-zhong[1] LIU Yan[1] HU Lin-mei[2] CHEN Jing[1]
(State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450002, China)[1]
(Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)[2]

Abstract Chinese online encyclopedia carries a huge amount of high quality information. Previous studies have utilized it for different knowledge acquisition tasks. For instance, the articles with similar subjects are grouped together into categories. Constructing a certain category topical hierarchy from the online encyclopedia is significantly beneficial for many applications such as search and browsing, information organizing and information retrieval. However, no attempts have been made to explore topic hierarchy of given category in online encyclopedia. Considering most of the online encyclopedia is heterogeneous and rough, this paper proposed a novel scheme of constructing topic hierarchy based on the Bayesian network. This scheme will incorporate both the structured contents table and unstructured text descriptions in the articles of the same category into automatic topic hierarchy learning for the online encyclopedia category using the algorithm of maximum spanning tree on the Bayesian topic network. Experimental results show that, compared with the existed encyclopedia topical hierarchy, our approach expand the content of 4 times while maintaining the accuracy of 75%.

Keywords Chinese online encyclopedia, Topic hierarchy, Structured contents table, Unstructured text description

1 引言

随着信息化的发展,越来越多的人们选择从在线百科上获取知识。作为一种自由开放并由世界各地的人们通过互联网合作编写的百科全书,中文在线百科提供了大量的信息,数以千万计的文档涵盖了各种事物和概念。如今,中文维基百科、百度百科和互动百科作为最流行的中文在线知识库,收录了累计将近3000万条中文条目[1)]。这些条目不仅包含了很多抽象概念(如"地震"),而且包含了成百上千的关于这个概念的实例。图1展示了概念"地震"的一个实例——"唐山大地震"。

人们通过总结这些大量的实例内的标题(如图1中的"概要"、"地震破坏"和"损失情况"等)及其描述文本,可以方便地获取这个概念的相关知识:当一个用户搜索一个抽象概念时,他/她会看到关于这个概念的百科层次目录。图2展示了概念"地震"在百度百科中的层次目录。本文将这些层次目录中的每一条称为关于这个概念的话题。从这个层次目录中用户可以方便、快速地了解到有关这个概念的带层次结构的话题

1) http://zh.wikipedia.org/wiki/网络百科全书

到稿日期:2016-04-27 返修日期:2016-06-18 本文受国家自然科学基金项目(61309007),国家"八六三"高技术研究发展计划基金项目(2006AA01Z409)资助。

王煦中(1991—),男,硕士生,主要研究领域为知识工程与新闻媒体挖掘、社会网络分析,E-mail:Observerspy@hotmail.com;刘 琰(1979—),女,博士,副教授,主要研究领域为网络数据智能分析、社会网络分析;胡琳梅(1991—),女,博士生,主要研究领域为知识工程与新闻媒体挖掘;陈 静(1990—),女,硕士生,主要研究领域为网络舆情监控。

信息。但不幸的是，建立在线百科层次话题是很复杂的一项工作，通常由专家们人工进行构建。这样不仅费时费力，而且所构建的话题也十分抽象，同时不够完整，例如图 2 中的层次话题只是简单介绍了地震的分布、烈度等，缺乏大量的其他相关详细话题（例如地震成因中的“火山地震”和“人工地震”等）。

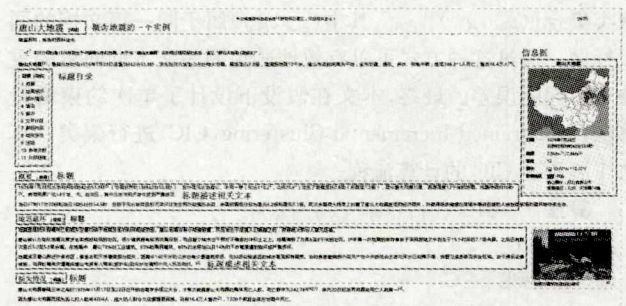


图 1　概念“地震”的一个实例——“唐山大地震”

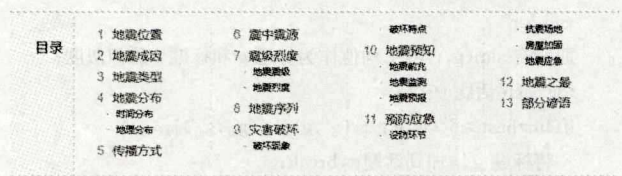


图 2　概念“地震”在百科中的部分层次话题结构

因此构建一个完整的综合性层次话题具有广泛的作用，它不仅对这一概念中的所有话题进行了总结，而且对搜索与浏览、信息组织和检索等应用也有很大帮助，同时也可以作为先验知识来提升文档组织和对新的语料集进行话题抽取的效果[1]。但是由于在线百科是一种新兴的媒体，这方面的研究还比较少。一些研究人员从维基百科的信息框和文本描述入手[2-3]，使用维基分类系统进行分类构建[4-5]；也有部分研究通过学习文本信息来构建层次话题[6-8]；还有一些研究使用传统层次聚类方法来试图构建层次话题[9]。现有的层次构造方法主要分为层次主题模型和层次聚类，但这两种方法并不适合本文的工作。首先，层次主题模型方法例如 hLDA，hPAM，hHDP 等，主要通过“词袋模型”将一个主题用多个词的分布进行描述，并进一步生成层次结构。这类方法大部分都是无监督的，同时得到的结果的语义都必须经过人的解读。而本文的工作需要每一个话题有其明确的语义含义。其次，层次聚类方法例如自上而下和自下而上的方法等，主要是通过计算文本相似度进行划分聚类，与主题模型方法类似，聚类结果的语义无法得到有效的控制，而且大多数聚类算法要事先设定聚类层数。而本文的工作是无法事先得到这一先验知识的。同时，与主要研究概念层次结构的本体构建不同[10-12]，本文研究的是话题层次结构的构建。例如，在给定概念“动物”下，本体构建工作将会把“猫”、“狗”等概念归于其下；而本文的工作将会把“动物保护”、“动物灭绝”等话题归于其下。

本文通过利用中文维基百科、百度百科和互动百科文档，提出了一种基于贝叶斯的话题层次网络来计算最优树结构的方法，并用于建立某一概念的综合性话题的层次结构。本文主要贡献有：

1）提出了一种新的利用异构在线百科中的文档目录结构来构建概念层次话题的框架。

2）提出了一种灵活的方法对话题层次进行自动构建。首先建立了具有结构信息的贝叶斯网络，其次通过最大树形图算法来计算话题层次关系。

3）通过在真实数据集中的大量实验表明，与原有的百科话题结构相比较，本文方法在保持 75% 的准确性的同时扩充了 4 倍的内容。

2　异构在线百科层次话题构建问题描述

本节给出层次话题构建的相关概念，并对异构在线百科中层次概念的提取给出形了式化描述。文中所用到的符号及其意义如表 1 所列。

表 1　本文所用到的符号

符号	意义
c	概念
a_i	概念相关百科文档
A_c	概念相关百科文档集
g	标题
d_g	标题相关描述文本
G	标题集
t	话题
T	话题集
H_c	层次话题结构
R	话题偏序关系集

定义 1（概念）　每一个概念 c 都是对一个问题的抽象描述。

定义 2（概念相关百科文档）　所有有关概念 c 的异构在线百科的文档本文被称为概念相关的百科文档。定义文档集合 $A_c=\{a_i\}_{i=1}^{N}$，其中 a_i 表示概念 c 下的一篇文档。

定义 3（标题及标题集）　每一个标题 g 都是由文档中与其相关的描述文本 d_g 来表示。标题集 $G=\{g_1,g_2,\cdots,g_{|G|}\}$ 是这个概念相关的所有标题 g 的集合，其中 $|G|$ 为标题个数。

定义 4（话题及话题集）　每一个话题 $t=\{g\}$ 是由具有一类相同意义的标题 g 所构成的集合，并将这些标题的描述文本合并为 $\{d_g\}$ 来表示这个话题。话题集 $T=\{t_1,\cdots,t_{|T|}\}$ 是这个概念相关的所有话题 t 的集合，其中 $|T|$ 为话题个数。

定义 5（话题层次）　话题层次是以给定概念 c 为根节点的一棵树 $H_c=(T,R)$。这是一棵逐渐细分的多层话题树，T 是话题集，$R=\{\langle t_i\prec t_j\rangle\mid t_i,t_j\in T\}$ 是话题偏序关系集，描述了 T 的层次结构。这个偏序关系 $\langle t_i\prec t_j\rangle$ 代表 t_i 是目录表中 t_j 的子话题。

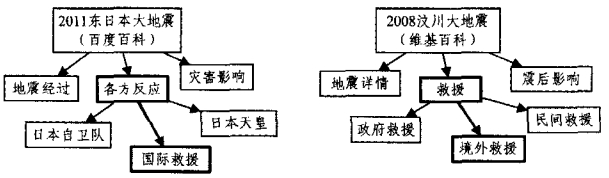
问题描述：给定某个概念 c，搜集相关的多个在线百科文档得到文档集 A_c，从其中抽取全部文档标题集 $G=\{g\}$，然后通过聚类合并相同文本意义的标题得到话题集 T，最终通过计算话题偏序关系集 R 得到一棵综合全面的话题层次树 $H_c=(T,R)$，从而形成对所给概念 c 的整体描述。

本文所面临的挑战如下：

1）由于编写在线百科的人们具有不同文化背景，导致了不同在线百科文档中目录标题具有多样性，文档中可能有很多表述不同但是意义一致的标题，例如图 3(a) 中的“国际救援”和图 3(b) 中的“境外救援”。层次话题构建的第一个关键问题就是如何构建一个能够正确反映这一话题的标题集合。

为了解决该问题，本文提出了基于描述文本相似度的一种单次带约束的增量聚类算法来寻找标题簇，然后将每一个簇映射到对应话题中形成话题集。

2)对于每一篇文档，从目录表中可以看到这些话题都有相应的文字描述以及对应其他话题的结构关系。但是，不同文档中的上下级关系可能是不一致的，例如图 3(a)中的“各方反应→国际救援”和图 3(b)中的“救援→境外救援”。层次话题构建的第二个关键问题就是如何确定子话题的上下级关系。为解决该问题，本文根据目录结构信息建立了话题层次的有向图(也可以看作是贝叶斯网络)，然后使用基于有向图的最大树形图算法来确定话题间的上下级关系，最终得到话题层次结构。



(a)2011 东日本大地震 (b)2008 汶川大地震

图 3 两个“地震”实例的部分话题层次结构

下面将具体介绍如何解决这两个问题。

3 异构在线百科层次话题构建的步骤

由问题描述可知，中文在线百科层次话题的构建问题可以通过分别解决话题集聚类问题和层次话题构建问题来解决。

1)话题集聚类：抽取概念 c 相关的文档集 A_c 的所有标题构成标题集 G，根据标题和相关描述文本 d_g 的相似度对标题集 G 聚类，最终构成话题集 T。

2)层次话题构建：将话题集 T 中的结构化信息引入到贝叶斯网络中，构建话题层次结构，最终通过最大树形图算法来得到预测的最优树结构 $H_c=(T,R)$。

3.1 基于文本的约束增量聚类算法的话题集识别

本文首先在给定概念 c 后，通过在线百科中有关概念的列表(例如：中文维基中“地震列表”这一条目[1])，或者人工总结相关的在线百科链接，使用爬虫技术爬取文档集 A_c。文档集 A_c 经过预处理后得到标题集 G。然后通过分别对标题集 G 的描述文本 $\{d_g\}$ 计算相似度 $sim(g_i,g_j)$ 并聚类来得到话题集 T。

然而，传统的文本聚类算法如 K-means，LDA 等，通过只使用词向量的方式得到的聚类结果在语义上并不明确，通常需要进行人为的解释；且在线百科的宽泛性即话题内容之间相互交叉包含，更使得单纯的词向量不能很好地反映话题明确的语义信息。然而在传统方法里，类似“地震经过”和“受灾情况”的标题极有可能出现在一个聚类结果中，它们都大量描述了地震所造成的破坏场面。另外，这些算法还需要求一些先验参数，例如聚类个数。由于本文的工作无法确定每一个概念的话题个数，因此这些传统算法并不适用于本文。

由于人们在写作时往往通过标题来划分文章内容，因此本文提出以下假设：同一篇文档中同级标题具有互斥性，即来自同一篇文档的同级标题需要聚类到不同的话题中。如图 3(a)中的“地震经过”、“各方反应”和“灾害影响”3 个同级标题显然不应被聚类到一个话题中。因此，本文建立了一个不可聚类集合 $\{(g_{ai},g_{aj})\}_{a\in A_c}$，其中 (g_{ai},g_{aj}) 属于同一篇文档的同级标题。通过这种方式可以有效地避免因为文本词向量的重叠所引起的误差。最终，本文在假设下设计了单次约束增量算法(Constrained Incremental Clustering，CIC)进行聚类。算法 1 展示了 CIC 的计算流程。

算法 1 CIC $(G_t,\{(g_{ai},g_{aj})\}_{a\in A_c})$

```
1. 设定 T=∅;
2. for each 标题 g∈G_t
3.     for each 话题 t∈T
4.     计算相似度 sim(g,t);
5.     end for
6.     取所有 sim(g,t)的平均值作为标题 g 和标题 t 的相似度;
7.     for each 话题 t∈T
8.     if(furthest>δ&&∀g't,(g',g)∉{(g_ai,g_aj)}_{a∈A_c})
9.       将标题 g 增加到话题 t,break;
10.      end if
11.    end for
12.    if(标题 g 未分配)
13.      将标题 g 加入新建的话题 t 中;
14.    end if
15. end for
16. return T.
```

其中，相似度计算 $sim(g,t)$ 由两个部分线性组成。

1)词法相似度。每一个标题 $g=\{w\}$ 是由一些单词 w 构成的序列所组成的短语，当且仅当两个词完全匹配时才意味着相等。本文通过如下方法来计算词法相似度：

$$WordSim_{g,g'}=\frac{|g\cap g'|}{0.5|g|+0.5|g'|}$$

即，两个标题的匹配词数除以两个标题的平均词长度[13]。

2)上下文相关度。对于每一个标题 g，计算它的描述文本 d_g 上的所有单词词频分布 ϕ_g，然后使用两个标题的词频分布来计算余弦相似度作为上下文相关度：

$$TextSim_{g,g'}=\cos(d_g,d_{g'})$$

通过引入不可聚类假设和词法相似度，这种增量聚类的方法比传统聚类方法更适合本文的问题。最终形成的话题的表现形式即为聚到同一类别的标题的集合。实验部分将以 LDA 作为基线方法与 CIC 算法加以对比。

3.2 基于贝叶斯网络的层次话题构建

首先提出一种简单层次话题构建(Simple Topic Hierarchy Induction，STHI)方法。先将话题集 T 中的每一个元素 t_i 视为节点，建立有向无环图；然后将统计出的话题关系对 $\langle t_i\prec t_j\rangle$ 中的全部标题关系对 $\langle g_i\prec g_j\rangle$ 在文档集 A_c 中出现的次数 $n_{i,j}$ 作为该图边上的权重；最终通过最大化该有向无环

[1] http://baike.baidu.com/link?url=ODoJVCyWsc83NJ0WLGJR3rAj6PrYN3MG9u0m1kRtbkOJo8waKfOVPHKykreGQ1ugyAoOjVNkQyveR4U5XnMv3q

图上的权重来得到一棵最优的层次话题结构树。这棵树的根节点代表目标概念 c。

然而，这样的结构将产生一个问题：所有节点都倾向于和根节点相连接。例如："＃地震烈度＃各地最大震度＃震级＃"这一话题和父话题"＃地震＃"在文档集中出现的次数为 4，而它与父话题"＃地震特性＃"出现的次数为 2。在上文的算法中，由于选择最大的权重，因此它将链接到根话题"＃地震＃"中(4>2)。然而由于"＃地震＃"在文档集中总共出现 48 次，而"＃地震特性＃"总共出现 6 次，因此"＃地震烈度＃各地最大震度＃震级＃"这一话题占"＃地震特性＃"话题的比例更大(2/6>4/48)，说明这一话题在"＃地震特性＃"中的作用更大。这个比例实际上可视为一种条件概率。因此，下面将提出一种基于贝叶斯网络的方法(Category Topic Hierarchy Induction，CTHI)来有效解决该问题。

本文对每一个话题对$\langle t_1 \prec t_2 \rangle$上的结构条件概率进行如下定义：

$$P^{struc}(t \mid par_H(t)) = \frac{\#\langle t \prec par_H(t)\rangle}{\# par_H(t)} \tag{1}$$

其中，$\#\langle t \prec par_H(t)\rangle$是$\langle g_i \prec g_j \rangle$关系对在标题关系集 R_a 中出现次数的总和；$\# par_H(t)$是上级标题 g_j 在全局标题集 G 中出现次数的总和。

本文将每一个话题 $t \in T$ 视为该概念的贝叶斯网络中的一个节点变量，因此给定 N 个节点的层次话题结构 H 的联合概率分布为：

$$P(N \mid H) = P(root) \prod_{n \in N \setminus root} P(n \mid par_H(n)) \tag{2}$$

$P(n \mid par_H(n))$是在给定父节点 $par_H(n)$下的条件概率，即上文中所计算的比例。显然，式(2)就是 H 的似然函数。因此，可以通过最大化其似然函数来建立概率最高的贝叶斯网络结构。

$$\begin{aligned} H^* &= \arg\max_H P(N \mid H, \theta) \\ &= \arg\max_H P(root) \prod_{n \in N(r)} P(n \mid par_H(n), \theta) \\ &= \arg\max_H \sum_{n \in N} \log(P(n \mid par_H(n), \theta)) \end{aligned} \tag{3}$$

最终本文得到一个基于结构条件概率的贝叶斯网络，再对其采用 Chu-Liu/Edmonds[14] 最大树形图算法计算最优树结构[15]。该算法首先寻找最大入边，此过程可能会形成环；然后递归地打破该环：每一次将环视为一个压缩的新节点，然后重新选择进入环的最大入边；在回溯过程中，每一个压缩的节点将重新展开，最终形成层次话题树结构 H。

4　实验结果与分析

目前还没有对在线百科层次话题构建的工作，而与本文最相近的研究内容属于层次聚类。关于这些层次聚类的研究已经很丰富了[16-17]，然而与传统文本聚类一样，它们的结果都很难从语义上理解——很多情况下需要人工解释。而本文的每一个节点都是一个明确的语义话题。除此之外，层次聚类还需要指定聚类个数，这是本文工作中无法获取的一个先验条件。因此本文并没有选取与层次聚类算法进行对比。下面首先在话题集识别步骤实验 CIC 算法，然后对层次话题构建步骤进行实验。实验代码已经公开在 GitHub 上[1]。

4.1　实验数据

实验在相关的中文在线百科页面中爬取了两个真实的数据集。由于在线百科数据本身不是很大，而且过大的数据集不便于人工标注和展示，因此本文选择了其中两个数据集。1)该数据集包含48 篇地震相关的文档(中文维基 24 篇、百度百科 15 篇、互动百科 9 篇)，主要描述近年来的重大地震和历史上著名的地震。其中中文维基搜索前 100 篇"地震"相关条目，维基字数统计为 182637，实验选取的 24 篇维基字数统计为 129181，占前 100 篇相关条目的 70.73%，证明实验数据集基本包含了主要的地震相关话题。百度百科和互动百科按照维基筛选的 24 篇进行搜索，去掉基本重复的条目。2)该数据集包含 37 篇总统选举相关的文档(中文维基 20 篇、百度百科 10 篇、互动百科 7 篇)，涉及 10 个国家，时间跨度从 2000 年到 2014 年。其中中文维基搜索前 50 篇"选举"相关条目，维基字数统计为 51559，实验选取的 20 篇维基字数统计为 38414，占前 50 篇相关条目的 74.51%，证明实验数据集基本包含了主要的相关选举话题。

对于每一篇文档，本文保留其带文本描述的标题以及目录表信息。由于这些文档是 HTML 格式组成的，因此可以通过正则表达式提取文档中内容有关的信息，从而将标签、表格等信息过滤掉，然后使用 ICTCLAS 对处理后的所有文档进行中文分词。根据 ICTCLAS 分词结果，保留名词(除去人名、地名、组织机构名和专有名词)、动词、形容词、区别词和状态词并去掉了常见的中文停用词作为处理后的文档内容，然后按照定义 2 的三元组结构将所有异构文档统一格式化储存。同时，实验去掉了文档中的常见停用词以及实体(如人物、组织、地点等)来建立层次模型。经过上述预处理后，实验的数据集词表大小为 4419 和 1214，分别包含 426 个和 212 个不同标题。

4.2　评估基于文本的约束增量聚类算法的话题集识别

本节对 3.1 节中提出的 CIC 算法进行评估。实验通过对 3 个在线百科已知的话题层次结构进行人工分析，最终分别构建两个标准测试集。其中地震集话题数量为 120 个，涵盖了"救灾"、"地震特性"、"次生灾害"等多个话题及对其进一步描述的子话题，如隶属于"救灾"话题之下的"政府救援"、"民众自救"等子话题，子话题最大层级为 5 层；选举集话题数量为 68 个，涵盖了"选举情况"、"选举方式"、"选举中的问题"等多个话题，子话题最大层级为 3 层。

实验使用 Adjusted Rand Index (ARI)来评估聚类结果。ARI 是 Rand Index 的一种变形[15]，用以度量 2 个聚类结果的相近程度，定义如下：

$$t_1 = \sum_{i=1}^{K_A} C_{N_i}^2,\ t_2 = \sum_{j=1}^{K_B} C_{N_j}^2,\ t_3 = \frac{2t_1 t_2}{N(N-1)},$$

$$ARI(A,B) = \frac{\sum_{i=1}^{K_A}\sum_{j=1}^{K_B} C_{N_{ij}}^2 - t_3}{\frac{t_1+t_2}{2} - t_3} \tag{4}$$

[1] https://github.com/Observerspy/CH_Topic_Hierarchy

其中，A 和 B 是两个聚类结果，分别有 K_A 和 K_B 个簇；N_i，N_j 分别表示 A 中第 i 个簇的数据数量和 B 中第 j 个簇的数据数量，N_{ij} 表示 A 中第 i 个簇同时在 B 中第 j 个簇的数据数量。由此可知，对于两个随机聚类，$ARI(A,B)=0$；反之，两个聚类结果相同，$ARI(A,B)=1$。

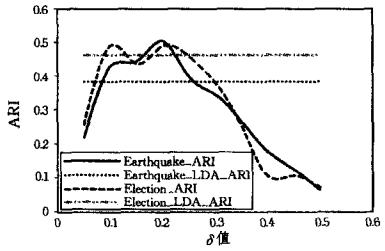


图 4 不同 δ 的聚类效果

如图 4 所示，将其与 LDA 算法进行对比，得到了对于不同的阈值 δ 两个数据集的聚类效果。其中，LDA 算法由于不受阈值 δ 变化的影响，因此是一条与横坐标平行的线。对于地震数据集，当 $\delta=0.2$ 时 ARI 指标最大；对于总统选举数据集，同样是 $\delta=0.2$ 时 ARI 指标达到最大。因此，本文选择 0.2作为两个数据集的阈值。由于 3.1 节中所述传统方法存在一定的缺点，因此图 4 中本文将实验结果与 LDA 聚类进行了对比。实验的 ARI 指标反映了本文的算法优于 LDA 聚类算法。其中 K 的取值为人工建立的标准话题集大小。

4.3 基于贝叶斯网络的层次话题结果分析

由于篇幅有限，本文在图 5 中只展示部分地震集的层次话题结构。

对于一个想了解“地震”这一概念的人来说，这样一个树形结构将能大大方便其获取相关知识：其可以快速地了解“地震的成因”——“气候变暖”、“人工地震”和“火山地震”等；另外，也会对地震所造成的种种“次生灾害”有一个全面的认识。同时这个结构也将对信息检索起到帮助：当检索“海啸”时，不仅能够获得有关“海啸”这一概念的知识，而且能够获得在“地震”->“次生灾害”这一分类下的“海啸”话题。这样丰富的信息检索结果将使得人们在学习一个新的概念时能够与其他概念相联系，更加促进对这一概念的认识。

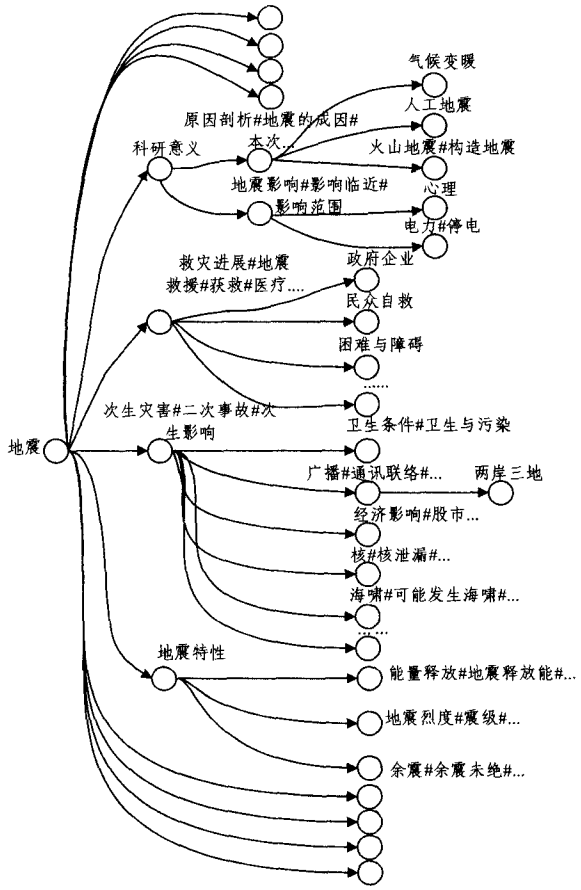


图 5 地震集层次话题构建部分结果展示

4.4 评估异构在线百科层次话题的准确度

本节从准确率、召回率和扩充率 3 方面对基于贝叶斯网络的话题层次构建进行评估。考虑到因目前缺乏相关评价标准，导致人工建立的标准层次话题测试集带有一定程度的主观性，在本节实验中，CTHI 的实验结果分别与 4.2 节中人工建立的标准层次话题测试集和在线百科中已有的层次话题结构进行比较。实验通过如下方式定义准确率 p、召回率 r 和扩充率 e：

$$p=\frac{\#\text{在标准层次话题结构中存在的由 CTHI 计算得到的话题对}}{\#\text{CTHI 计算得到的话题对}}$$

$$r=\frac{\#\text{在原有在线百科层次话题结构中存在的由 CTHI 计算得到的话题对}}{\#\text{原有在线百科层次话题结构中的话题对}}$$

$$e=\frac{\#\text{CTHI 计算得到的话题对}-\text{原有在线百科层次话题中存在的 CTHI 计算得到的话题对}}{\#\text{原有在线百料层次话题结构中的话题对}}$$

表 2、表 3 列出了 CTHI 方法构建的层次话题结构的准确率和在 3 个在线百科上的召回率与扩充率。其中地震数据集达到了 77.88%的准确率与 71.69%的平均召回率，总体比原层次话题规模扩大了 3.3～5.6 倍；选举数据集达到了 71.43%的准确率与 55.29%的平均召回率，总体比原层次话题规模扩大了 3.9～8.4 倍。其中由于“选举”(“民主选举”)这一概念在百度百科和互动百科上原有的层次话题结构过于简单粗糙，话题数量少，且内容质量低于中文维基，因此这两者的召回率偏低而扩充率偏高。如果忽略这两者在召回率和扩充率上的表现，实验在上两个数据集上的平均准确率为 74.66%，平均召回率为 67.99%，平均扩充率为 407.37%，充分说明了本文方法的有效性。

表 2 层次话题构建评估/%

数据集	p	中文维基 r	百度百科 r	互动百科 r
地震集	77.88	72.22	75.00	67.86
选举集	71.43	64.29	44.44	57.14

表 3 层次话题构建的扩充率/%

数据集	中文维基 e	百度百科 e	互动百科 e
地震集	555.56	395.83	335.71
选举集	385.71	655.56	842.86

结束语 本文提出了通过使用中文在线百科文档构建一个针对概念的综合性层次话题结构。本文的主要方法是将不

同的标题根据相似度聚类为话题，然后利用结构信息建立话题层次的贝叶斯网络，最后通过 Chu-Liu/Edmonds 算法计算最优树，获得最终的层次话题结构。实验结果证明了本文方法的有效性。在今后的研究中，将进一步研究如何提高聚类准确度以及尝试将跨语言的在线百科合并到已有的层次话题结构中。

参 考 文 献

[1] TED P, SIDDHARTH P, JASON M. Wordnet: Similarity-measuring the relatedness of concepts[C] // HLT-NAACL 2004. Association for Computational Linguistics, 2004: 38-41.

[2] WU F, WELD D S. Automatically refining the wikipediainfobox ontology[C] // Proceedings of the 17th International Conference on World Wide Web. ACM, 2008: 635-644.

[3] WU F, HOFFMANN V, WELD D S. Information extraction from wikipedia: Movingdown the long tail[C] // Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2008: 731-739.

[4] LI R, BAO S H, YU Y, et al. Towards effective browsing of large scale social annotations[C] // Proceedings of the 16th International Conference on World Wide Web. ACM, 2007: 943-952.

[5] NASTASE V, STRUBE M. Decoding wikipedia categories for knowledge acquisition[C] // AAAI. 2008: 1219-1224.

[6] DMBTL G, MIJJB T. Hierarchical topic models and the nested chinese restaurant process[J]. Advances in Neural Information Processing Systems, 2004, 16: 17.

[7] MIMNO D, LI W, MCCALLUM A. Mixtures of hierarchical topics with pachinko allocation[C] // Proceedings of the 24th ICML. ACM, 2007: 633-640.

[8] ZAVITSANON E, PALIOURAS G, VOUROS G A. Non-parametric estimation of topic hierarchies from texts with hierarchical dirichlet processes[J]. The Journal of Machine Learning Research, 2011, 12: 2749-2775.

[9] CHUANG S L, CHIEN L F. A practical web-based approach to generating topic hierarchy for text segments[C] // Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management. ACM, 2004: 127-136.

[10] TANG J, LEUNGH F, LUO Q, et al. Towards ontology learning from folksonomies[C] // IJCAI. 2009: 2089-2094.

[11] ZHU X W, MING Z Y, ZHU X Y, et al. Topic hierarchy construction for the organization of multi-source user generated contents[C] // Proceedings of the 36th International ACMSIGIR Conference on Research and Development in Information Retrieval. ACM, 2013: 233-242.

[12] NAVIGLI R, VELARDI P, FARALLI S. A graph-based algorithm for inducing lexicaltaxonomies from scratch[C] // IJCAI. 2011: 1872-1877.

[13] MONGE A E, ELKAN C, et al. The field matching problem: algorithms and applications[C] // Proceedings of the 2nd ACM SIGKDD. 1996: 267-270.

[14] CHU Y J, LIU T H. On shortest arborescence of a directed graph[J]. Scientia Sinica, 1965, 14(10): 1396.

[15] VINH N X, EPPS J, BAILEY J. Information theoretic measures for clusterings comparison: is a correction for chance necessary? [C] // Proceedings of the 26th Annual International Conference on Machine Learning. ACM, 2009: 1073-1080.

[16] LIU X, SONG Y, LIU S, et al. Automatictaxonomy construction from keywords[C] // KDD. 2012: 1433-1441.

[17] WANG C, DANILEVSKY M, DESAI N, et al. A phrase mining framework forrecursive construction of a topical hierarchy[C] // KDD. New York, NY, USA, ACM, 2013: 437-445.

（上接第 198 页）

[9] ZHU P, HAN J. Asynchronous stochastic Boolean networks as gene network models[J]. Journal of Computational Biology, 2014, 21(10): 771-783.

[10] FARYABI B, VAHEDI G, CHAMBERLAND J F, et al. Intervention in Context-Sensitive Probabilistic Boolean Networks Revisited[J]. Eurasip Journal on Bioinformatics & Systems Biology, 2009, 2009(1): 1-13.

[11] SHMULEVICH I, DOUGHERTY E R, ZHANG W. Gene perturbation and interventionin probabilistic Boolean networks[J]. Bioinformatics, 2002, 18(10): 1319-1331.

[12] PARKER D. Implementation of Symbolic Model Checking for Probabilistic Systems[D]. University of Birmingham, 2002.

[13] KWIATKOWSKA M, NORMAN G, PARKER D. Stochastic Model Checking[M]. REMKE A, STOELINGA M, eds. Berlin Heidelberg: Springer, 2014.

[14] DATTA A, CHOUDHARY A, BITTNER M L, et al. External Control in Markovian Genetic Regulatory Networks[J]. Journal of Chemical Physics, 2003, 109(11): 4569-4575.

[15] TOURNIER L, CHAVES M. Uncovering operational interactions in genetic networks using asynchronous boolean dynamics [J]. Journal of Theoretical Biology, 2009, 260(2): 196-209.

[16] KOBAYASHI K, HIRAISHI K. Symbolic approach to verification and control of deterministic/probabilistic Boolean networks [J]. IET Systems Biology, 2012, 6(6): 215-222.

[17] FENG J, YAO J, PENG C. Singular Boolean networks: Semi-tensor product approach[J]. Science China(Information Scien-ces), 2013, 56(11): 1-14.

[18] LANGMEAD C J, JHA S K. Symbolic approaches for finding control strategies in Boolean Networks[J]. Journal of Bioinformatics & Computational Biology, 2009, 7(2): 307-319.

[19] KOBAYASHI K, HIRAISHI K. Verification and Optimal Control of Context-Sensitive Probabilistic Boolean Networks Using Model Checking and Polynomial Optimization[J]. The Scientific World Journal, 2014, 2014(2): 295-318.

[20] LIU Q, GUO X, ZHOU T. Optimal control for probabilistic Boolean networks[J]. IET Systems Biology, 2010, 4(2): 99-107.