PPT 文档的概念图自动构建①

黄光轮[1,2], 文益民[1], 朱文宇[1], 易新河[1]

[1](桂林电子科技大学 计算机科学与工程学院, 桂林 541004)
[2](中国科学院大学 工程管理与信息技术学院, 北京 100049)

摘 要: 随着教育技术的发展, 越来越多的人在学习过程中使用 PPT 文档. 对 PPT 文档进行概念图的构建, 使得学习者能快速且全面地了解一个PPT文档的知识要点, 有益于学习者加快学习速度, 有益于获取学习者的学习行为. 基于此, 提出了一种利用 Microsoft Office 编程技术、文本挖掘技术和社会网络分析技术自动提取 PPT 文档中的概念术语、概念术语之间的关系及构建概念图的算法. 实验结果表明: 该算法可以计算概念术语的重要性; 算法提取的概念术语具有一定的准确率, 提取到的越重要的概念术语的准确率越高.

关键词: PPT 文档; 概念图; 社会网络分析; 概念术语; 学习行为

Automatically Constructing Concept Map of the PPT Documents

HUANG Guang-Lun[1,2], WEN Yi-Min[1], ZHU Wen-Yu[1], YI Xin-He[1]

[1](School of Computer Science and Engineering, Guilin University of Electronic Technology, Guilin 541004, China)
[2](College of Engineering & Information Technology, University of Chinese Academy of Sciences, Beijing100049, China)

Abstract: With the development of educatinal technology, more and more people use PPT documents in their learning process. By constructing a concept map for PPT doc uments, learners can quickly and comprehensively understand all the knowledge points of the PPT documents, being beneficial for the learners to speed up the learning and collect his learning behaviors. In this paper, we proposed an algorithm using Microsoft Office programming technology, text mining technology and social network analysis technology to automatically extract concept terminologies and the relationships between them from the PPT documents to construct concept map. The experimental results show that the psoposed algorithm can effectively calculate the importance of each concept terminology. And all the concept terminologies extracted by the proposed algorithm have a certain accuracy, the more important the concept terminology, the higher the accuracy.

Key words: PPT document; concept map; social network analysis; concept terminology; learning behavior

概念图作为一种知识表示和知识组织的工具, 是对某一领域中的概念术语及其关系的可视化表达, 最早由教育学家诺瓦克作为教学工具提出[1], 并且已经被引入到知识管理领域, 用于支持知识提取、知识组织、知识评价等活动的完成[2]. 概念图作为支持学习的工具, 在教育教学中有着广泛的应用[3]. 概念图的构建主要包括领域概念术语的提取, 概念关系的抽取. 目前, 概念图的构建往往由领域专家或者相关研究人员人工完成. 显而易见, 针对每个文档或者文档集人工构建概念图的方法具有一定的局限性, 因为建立一个完整准确的概念图必须花费领域专家大量的时间和精力.

随着现代教育技术的快速发展, PPT 文档已经成为人们获取知识的重要来源之一. 当前在 PPT 文档使用中存在两个问题: 不容易实现对 PPT 文档中的概念术语的快速检索; 不容易实现对 PPT 文档内容的选择学习; 不太容易获取学习者在使用 PPT 文档进行学习时的学习行为. 因此, 研究 PPT 文档的概念图自动构

① 基金项目:国家级大学生创新性实验立项项目(201210595003);中国高教学会教育信息化专项课题(2014XXH1205YB);广西区高等教育教学改革工程项目(2014JGZ116)

收稿时间:2014-12-24:收到修改稿时间:2015-03-12

建技术, 使得 PPT 文档使用者能利用概念图进行学习, 提高学习效率, 这是一个非常有意思的问题.

1 相关研究

目前, 针对纯文本文档的概念图自动构建的研究, 国内已有一些成果. Chen 等提出概念图自动构建的关键在于概念术语的自动抽取和概念术语之间的关系确定. 概念术语来源于文献中的关键词或者高频词语, 概念术语之间的关系由它们共现的频率及位置确定[4]. 邓三鸿等针对 CSSCI 数据库中的刊物和论文提出构建学科知识地图, 亦即概念图的方法. 概念术语来源于文献中的关键词, 概念术语之间的关系由它们在同一篇文献中共现的频率确定[5]. 傅骞等针对移动学习提出了一种自动构建移动学习领域的概念图的方法. 概念术语来源于题录信息中的高频词, 概念术语之间的关系通过两术语在句子中的共现频率表示, 而且提出了以下三种研究假设——概念图可由概念术语及术语之间的关系来表示; 如果两个术语在同一句子中出现, 就暗示这两个术语之间存在一定的关系; 概念图中的关系可以用数字来表示, 数值越大, 关系越密切[6]. 张会平等针对文献数据库提出基于词共现的概念图自动构建研究方法, 概念术语来源于文献中的关键词或高频词, 概念术语的关系通过计算概念术语之间的关联强度得到[7]. 孙珠婷等在概念图构建中结合网络爬虫技术和 LSA 方法对领域文本资源进行概念术语的提取, 提取流程包括四步: 获取领域文本资源; 文本预处理; 特征项提取; 利用 LSA 提取概念术语[8]. 李建素等认为关键词不仅能在一定程度上反映文章内容, 而且反映主题内容还存在着程度大小的问题, 获取文档的关键词就是选出最能反应文章主题内容的那些候选项[9].

PPT 文档作为一种与纯文本文档不同格式的文档, PPT 文档中的概念术语不但与其所在的语句有关, 还与其所在的位置有关. 根据作者的查证, 针对 PPT 文档的概念图的自动构建还未曾被人关注.

2 数据获取以及预处理

本文利用 Microsoft Office 编程技术析取 PPT 文档中的纯文本数据. 这些纯文本数据包括标题中的文本内容及正文中各段落的文本内容, 同时还要析取标题及各段落的层次关系. 具体步骤如下:

步骤 1: 生成一个 Package 类的实例 pptPackage. 首先导入 System.IO.Packaging.Package, 然后调用 Package 类的函数 Open(String, FileMode, FileAccess)按给定的文件模式和文件访问设置打开位于给定路径的 PPT 文件的包, 使得程序能与 Office Open XML File Formats 进行交互, 同时将函数 Open(String, FileMode, FileAccess)的返回值赋给 pptPackage.

步骤 2: 调用 pptPackage 下的函数 GetRelation-ships ByType (string relationshipType), 返回与指定的 relationshipType 匹配的包级别关系的集合 relationships.

步骤 3: 在遍历集合 relationshis 的过程中, 调用函数 ResolvePartUri(Uri sourcePartUri,Uri targetUri)获得指定 SourcePartUri 和 targetUri 参数之间解析的目标部件的 URI, 然后调用 GetPart 获得与目标部件的 URI 对应的 PPT 文档部件 documentPart.

步骤 4: 定义并初始化类型为 XmlNamespace Manager 的实例变量 nsManager, 然后调用函数 AddNamespace (string prefix,string uri)将指定的命名空间添加到命名空间集合 nsManager.

步骤 5: 调用函数 Load (XmlReader reader)从指定的 XmlReader 加载 PPT 文档部件的 XML 文档.

步骤 6: 调用函数 SelectNodes(String, XmlNamespaceManager), 选择匹配 XPath 表达式的节点列表 sheetNodes. XPath 表达式中的任何前缀都使用步骤步骤 4 提供的 XmlNamespaceManager 的实例进行解析;

步骤 7: 遍历 sheetNodes, 对每一个 sheetNode, 执行步骤步骤 8 和步骤 9;

步骤 8: 调用 sheetNode 类的 Attributes 方法获得当前节点的 ID, 即为幻灯片的 ID;

步骤 9: 调用函数 SelectSingleNode (String) 分别获得与标题、正文及它们的段落层次相匹配的 XPath 表达式的第一个 XmlNode 类型的节点. 然后从第一个 XmlNode 类型的节点开始, 分别提取节点类型为 XmlNode 的节点“a: t”的文本 text 和属性“lvl”的值 value, text 为标题或正文的文本, value 为段落的层次值.

3 候选概念图的构建

候选概念图由候选概念术语以及它们之间的关系构成. 候选概念术语是指从经过预处理后的文本数据中提取的关键词, 作为构建候选概念图的节点. 本文

采用中国科学院计算技术研究所开发的汉语词法分析系统 ICTCLAS 对从 PPT 文档中提取的所有文本内容进行有词典分词, 其中导入的词典为停用词库[10]. 然后从分词后的纯文本中提取词频大于词频阈值的所有名词词语, 或者是邻接种类大于邻接种类阈值的所有词语[11], 作为候选概念术语. 为了提高候选概念术语提取的准确度, 本文研究需要的数据还包括高频常用词库(用于过滤 PPT 文档中的高频常用词语, 下载地址为 http://www.datatang.com/data/43728).

将文本中出现的所有候选概念术语以及它们之间的共现和层次关系映射为一个网络图 G, 即候选概念图. 根据图论中图的定义 G = {V,E}, 则需要对 G 中的顶点集和边集给出相应的定义. 在网络图中, 顶点表示候选概念术语, 顶点之间的连线表示两个候选概念术语之间的关联程度.

在介绍如何构建候选概念图之前, 先定义段落层次的相关概念. 假设演示文稿某张幻灯片的内容编排如图 1 所示, 则图中有四个层次: 1 和 5 属于一级层次, 2 和 6 属于二级层次, 3、4 和 7 属于三级层次, 标题则属于顶层.

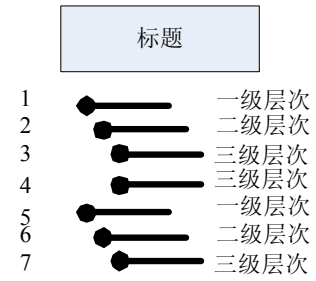

图 1 幻灯片内容编排模式

PPT 演示文稿中的这种布局体现了语义的一种层次关系: 顶层的文本由属于一级及以下层次的文本解释. 一级层次由紧随其后的一个或多个二级及以下的层次解释. 以此类推, 第 m 级层次由紧随其后的一个或多个(m+1)级及以下的层次解释. 例如, 在图 1 中, 一级层次 1 只能由紧随其后的二级层次 2 以及三级层次 3 和 4 解释, 不能由三级层次 7 解释.

在比较规范的 PPT 演示文稿中, 内容编排一般有以下两种模式: 一种是平行模式, 另一种是非平行模式. 两种模式下的候选概念术语网络构建方法如下:

(1)平行模式下的网络构建方法

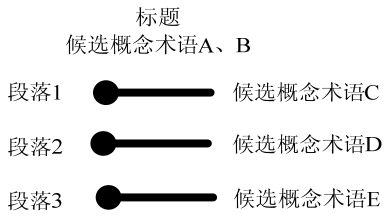

图 2 平行模式

如果演示文稿幻灯片的内容编排如图 2 所示, 则称段落 1、段落 2 和段落 3 中出现的候选概念术语 C、D 和 E 都用于解释标题中出现的候选概念术语 A 和 B. 在图 2 中, 因为 A 和 B 分别与 C、D 和 E 存在解释关系, 所以可以认为与 A 和 B 对应的节点分别和与 C、D 和 E 对应的节点存在较强的联系. 候选概念术语 C、D 和 E 之间不存在解释关系, 但是由于它们出现在同一张幻灯片中, 所以可以认为它们之间存在较弱的联系. 在标题中出现的候选概念术语 A 和 B 一般比较重要, 也更有可能是概念术语, 即使 A 和 B 之间不存在解释关系, 与它们对应的节点之间存在较强的联系也是合理的. 在本文中, 如果两个节点之间存在较强的联系, 则边的权值加 1. 如果两个节点之间存在较弱的联系, 则边的权值加 0.2. 用网络图表示如图 3.

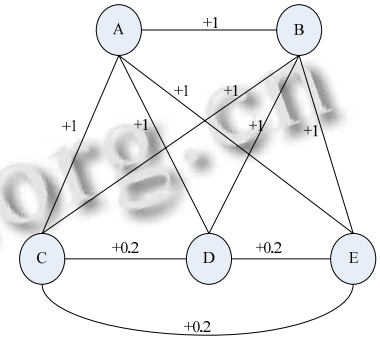

图 3 平行模式下构建的网络图

(2) 非平行模式下的网络构建方法

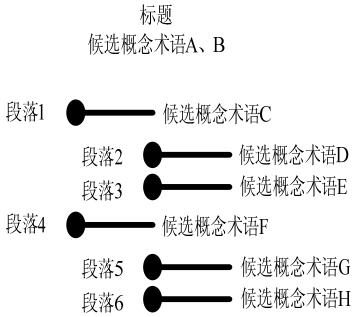

图 4 非平行模式

若演示文稿幻灯片的内容编排如图 4 所示, 则在标题出现的候选概念术语 A、B 跟候选概念术语 C、D、E、F、G 和 H 存在解释关系, 候选概念术语 C 跟候选概念术语 D、E 存在解释关系, 候选概念术语 F 跟候选概念术语 G、H 存在解释关系. 用网络图表示如图 5.

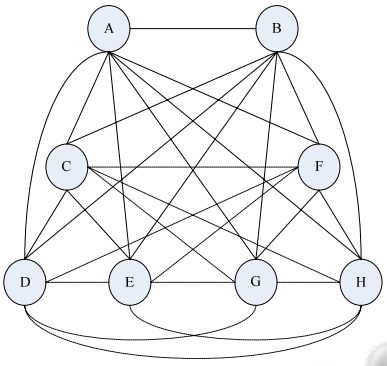

图 5　非平行模式下构建的网络图

图 5 中实线表示边的权值加 1, 虚线表示边的权值加 0.2.

4　算法设计

本文提出了两种算法来实现对 PPT 文档中的概念图自动构建. 两种算法的实施步骤如下:

算法 1.

(1) 提取 PPT 文档中的纯文本数据及其内容编排格式;

(2) 对纯文本数据进行分词;

(3) 设定词频阈值, 提取分词后的纯文本中词频大于词频阈值的所有名词词语作为候选概念术语;

(4) 利用高频常用词库对候选概念术语进行过滤, 剩余候选概念术语用于构建候选概念图;

(5) 运用特征向量中心性算法, 计算候选概念图中每个概念术语的影响力值, 按照影响力值从大到小对概念术语进行排序, 选取影响力值靠前的候选概念术语作为概念术语, 并根据概念术语之间的网络关系构建概念图.

算法 2.

(1) 提取 PPT 文档中的纯文本数据及其内容编排格式;

(2) 对纯文本数据进行分词;

(3) 设定邻接种类阈值, 提取分词后的纯文本中邻接种类大于邻接种类阈值的所有词语作为候选概念术语;

(4) 利用高频常用词库对候选概念术语中的高频常用词语进行过滤, 剩余候选概念术语用于构建候选概念图;

(5) 运用特征向量中心性算法, 计算候选概念图中每个概念术语的影响力值, 按照影响力值从大到小对概念术语进行排序, 取影响力值靠前的候选概念术语作为概念术语, 并根据概念术语之间的网络关系构建概念图.

5　实验结果与分析

为了评估自动构建的概念图中概念术语的准确性, 本文收集了各个学科的规范概念术语集合(下载地址为 http://shuyu.cnki.net/)以及包括材料物理性能、操作系统和分子生物学等 11 门课程的 PPT 课件作为实验数据, 将影响力靠前的 n 个概念术语中出现在规范概念术语集合中的术语个数与 n 之比作为从每一门 PPT 课件中自动提取的概念术语的准确率.

算法 1 中, 当词频阈值被设置为 3 时, 自动提取的概念术语的准确率如表 1 所示.

表 1　词频阈值为 3 时概念术语的准确率

课程名称	n=50 时, 概念术语的准确率(%)	n=100 时, 概念术语的准确率(%)
材料物理性能	66	58
操作系统	86	72
分子生物学	38	34
花卉学	68	55
农林	38	37
人体生理课件	60	50
生物仪器分析	66	64
数据库	82	70
计算机网络	72	68
遗传学	30	22
药理学	38	30

由表 1 可以知道, 提取到的前 50 个概念术语的准确率比后 50 个概念术语的准确率高, 这是因为前 50 个概念术语的影响力值普遍较大, 更有可能是真正的概念术语.

在算法 1 中, 我们直接将词频大于预先设定的词频阈值的名词词语作为候选概念术语. 这种方法的缺点之一就是一些比较长的概念术语由于被分割成较小的片段而不能被准确提取. 例如, 《数据库课程》课件中的概念术语“关系模式”被提取成“关系”和“模式”,

《工艺设计概论》中的概念术语“真空冷却器”被提取成“真空”和“冷却器”. 此外, 一些概念术语往往是通过不同词性组合而成, 而要归纳概念术语的词性组合规则非常不易, 不正确的规则不是导致概念术语不能被准确提取就是导致概念术语的丢失.

在算法 2 中, 当邻接种类阈值被设置为大于 3 的时候, 自动提取的概念术语的准确率如表 2 所示.

表 2 邻接种类阈值大于 3 时概念术语的准确率

课程名称	n=50 时, 概念术语的准确率(%)	n=100 时, 概念术语的准确率(%)
材料物理性能	74	63
操作系统	86	80
分子生物学	58	51
花卉学	72	57
农林	48	43
人体生理课件	70	58
生物仪器分析	76	71
数据库	86	73
计算机网络	84	78
遗传学	42	37
药理学	36	35

根据表 1 和表 2, 可以得到如下两个折线图.

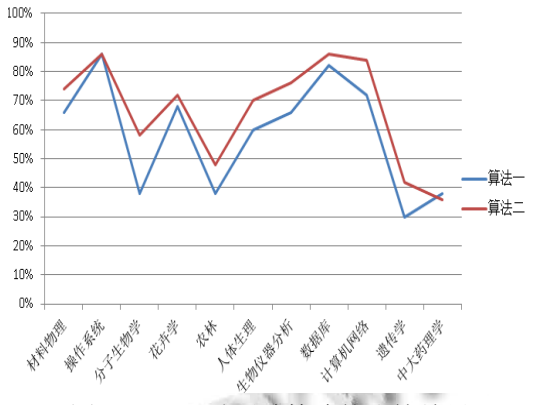

图 6 n=50 时两种算法的比较结果

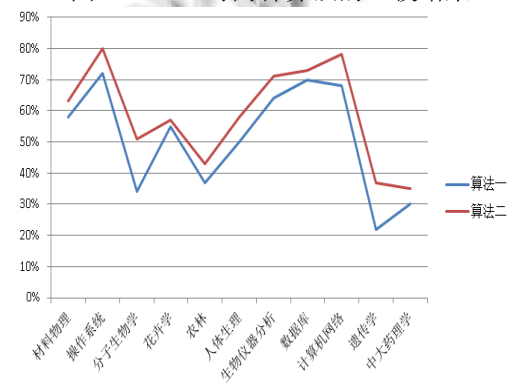

图 7 n=100 时两种算法的比较结果

图 6 和 7 中横轴代表不同的课件, 纵轴代表算法提取到的概念术语的准确率. 从中可以看出, 算法 2 的效果普遍比算法一更优.

从算法 1 和算法 2 不难发现, 在概念术语提取的过程中, 出现了二种类型的误差: 一是概念术语的遗失; 二是非概念术语作为概念术语的提取, 这是最严重的误差, 也是导致某些课件概念术语提取的正确率下降的主要原因.

(1)概念术语的遗失. 在概念术语提取的过程中, 一些概念术语由于在候选概念图中孤立地出现, 导致中心地位下降而被遗失. 比如数据库课件中的“并发事务”、“聚簇码”; 数字逻辑课件中的“计数器”; 生物仪器分析课件中“质子”等. 这说明: 这些课件的制作者没有对以上概念术语给予足够的重视.

(2)非概念术语的提取. 这是在概念术语提取过程中对概念术语提取正确率影响最大的一类误差. 人们在日常写作时, 总会运用到大量的常用词语, 这些常用词语出现频率高, 邻接种类值高, 经常出现在文本的不同语境中. 在某些领域中, 这些词语还可以构成概念术语. 例如, “中间”对一般人而言是常用词, 但是它能构成计算机体系结构领域的概念术语“中间件”. 如果只是简单地将常用词语标记为停用词, 意味着某些领域中某些概念术语不能被正确提取. 况且, 常用词数量成千上万, 需要花费大量的人力物力才能将这些常用词都找出来, 而且随着语言的发展, 有些常用词的含义也会随之发展. 因此, 有些常用词虽然不是概念术语, 但要全部剔除它们还是很困难的.

另外, 从实验过程中还发现, 数据库、数字逻辑、中山药理学、花卉学等课件的概念术语提取正确率较低, 这很大程度是由上述两种类型的误差导致. 另外, 人们在制作 PPT 课件时, 大部分是基于个人知识使用概念术语, 使得概念术语的规范性得不到保证. 即使这些词语的影响力值较高, 但是由于它们不在规范的概念术语表中, 所以在计算概念术语提取准确率时, 这类词语被视为是非概念术语, 导致了概念术语提取准确率的下降. 比如, 数据库课件中的“事务”是数据库领域的一个概念术语, 但由于它不在规范概念术语表中, 使得它被认为是非概念术语.

6 结语

本文提出了一种自动提取 PPT 文档中的概念术语

以及概念术语之间的关系, 以构建概念图的算法. 首先利用 Microsoft Office 编程技术析取 PPT 文档中的纯文本数据及标题和各段落的层次关系; 然后利用中文分词技术对纯文本数据进行分词, 提取候选概念术语; 利用候选概念术语之间的共现关系以及层次关系构建候选概念图; 最后运用社会网络中心性分析算法提取概念术语; 最后根据概念术语之间的网络关系构建概念图. 实验结果表明: 该算法可以计算概念术语的重要性; 算法提取的概念术语具有一定的准确率, 提取到的越重要的概念术语的准确率越高.

该方法可在一定程度上弥补领域专家或者研究人员需要花费大量时间和精力针对不同文档或者文档集构建不同概念图的不足, 并有助于向学习者展示一个文档或者文档集的知识要点. 然而, 在概念术语提取过程中, 本文提出的算法还存在概念术语遗失、非概念术语提取等问题. 未来可考虑从以下角度展开进一步的研究: 一是通过语义处理屏蔽一些使用频率高且能够适应于多种语境的常用词语, 改进候选概念术语的提取效果; 二是深入挖掘概念术语在 PPT 文档中具有的特征, 优化概念术语提取算法; 三是寻找更优的算法提取概念术语之间的关系, 提高自动构建的概念图的质量.

参考文献

1 Novak JD. Learning how to learn. London, Cambridge University Press, 1984.

2 马费成,郝金星.概念地图在知识表示和知识评价中的应用(I)—概念地图的基本内涵.中国图书馆学报,2006,(3): 5–9, 49.

3 陈浩.概念图相关知识及其在教育系统中的应用.学理论,2013,(33):281–282.

4 Chen NS, Wei CW, Chen HJ. Mining e-learning domain concept map from academic articles. Computers & Education, 2008, 50(3): 1009–1021.

5 邓三鸿,金莹,杨建林.学科知识地图的构建—以图书、情报学为例.情报学报,2006,25 (1):3–8.

6 傅骞,魏顺平,贺龙祥.移动学习领域概念图的构建研究.中国电化教育,2007,(10):96–99.

7 张会平,周宁.基于词共现的概念图自动构建研究.情报理论与实践,2008,(6):928–930,903.

8 孙珠婷,顾倩颐.概念图构建中概念术语自动提取的研究与实现.计算机工程与设计,2012,(7):2864–2867.

9 李素建,王厚峰,俞士汶,辛乘胜.关键词自动标引的最大熵模型应用研究.计算机学报,2004,(9):1192–1197.

10 刘克强.2009 共享版 ICTCLAS 的分析与使用.科教文汇(上旬刊),2009,(8):271–280.

11 Feng H, Chen K, Deng X, et al. Accessor variety criteria for Chinese word extraction. Computational Linguistics, 2004, 30(1): 75–93.