 ****

**研究生科研实践报告**

**基于知识点重要性的MOOC学习引导框架**

题目：

**姓 名：**  朱纪乐

**学 号：** 1501214430

**院 系：** 信息科学技术学院计算机系

**专 业：**  计算机软件与理论

**研究方向：** 教育数据挖掘

**导 师：**  张铭教授

**二零一七 年 六 月**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **北京大学信息科学技术学院计算机系 研究生《科研实践》课程考核表** | | | | | | | |
| 学 号 | | 1501214430 | 姓 名 | | 朱纪乐 | 导 师 | 张铭 |
| 专 业 | | 计算机软件与理论 | 研究方向 | | 教育数据挖掘 | 考核时间 | 2016-2017学年第2学期 |
| 所在研究所 | | 网络与信息系统研究所 | 所在实验室 | | 数据库 | 考核成绩 |  |
| 科研实践内容和成果：   1. 调研相关工作研究进展，包括在线学习引导、知识点定义、PageRank算法及其扩展。 2. 设计了一个基于知识点重要性的在线学习引导框架，以字幕文件作为输入，返回涉及到的知识点重要性排序结果。 3. 提出了对基于有序文档的PageRank、重要性数值平滑等方法，对排序算法进行优化。 4. 在MOOC课程数据上进行实验，对比其他基准算法，验证模型有效性并讨论分析实验结果。 | | | | | | | |
| 导师意见：  大规模开放在线课程（MOOCs）提供了丰富的学习资源，成为了学生最主要的自学途径之一。项目的选题具有一定的理论价值和广阔的应用前景。朱纪乐同学提出了基于知识点的学习引导算法框架，并在北京大学《数据结构与算法》课程数据上进行实验。项目工作扎实，组织逻辑清晰，实验结果和数据可靠。部分研究成果已经体现在朱纪乐同学发表在EDM’17的一作论文，在此基础上继续深入研究，将完成硕士研究论文。 | | | | | | | |
|  |  | |  | 签名： | | 2017年 6月 16日 | |
| 考核小组意见： | | | | | | | |
|  |  | |  | 签名： | | 年 月 日 | |

**目 录**

[第1章 科研实践研究背景 - 1 -](#_Toc373835284)

[1.1 MOOC相关背景 - 1 -](#_Toc454479340)

[1.2 MOOC存在的问题 - 1 -](#_Toc454479341)

[1.3 相关措施和解决方案 - 2 -](#_Toc454479342)

[第2章 前沿技术调研报告 - 4 -](#_Toc373835285)

[2.1 在线学习引导 - 4 -](#_Toc454479344)

[2.2 知识点定义 - 5 -](#_Toc454479345)

[2.3 PageRank算法及其扩展 - 5 -](#_Toc454479346)

[第3章 科研实践工作（具体） - 6 -](#_Toc373835286)

[3.1 MOOC学习引导框架概述 - 6 -](#_Toc454479350)

[3.2 知识点抽取 - 7 -](#_Toc454479351)

[3.3 基于有序文档的PageRank - 8 -](#_Toc454479352)

[3.4 计算知识点重要性 - 11 -](#_Toc454479353)

[3.5 实验和分析 - 12 -](#_Toc454479356)

[3.6 讨论和总结 - 15 -](#_Toc454479359)

[参考文献 - 17 -](#_Toc373835287)

[附1：已发表论文 - 20 -](#_Toc373835288)

[附2：申请发明专利、或获得发明专利授权、或取得通过技术鉴定的相关成果 - 21 -](#_Toc373835289)

# 科研实践研究背景

## MOOC相关背景

随着互联网教育的不断发展，大规模开放在线课程（Massive Open Online Course，简称MOOC，中文译为慕课）已经成为学生在课堂之外最流行的自学方式之一。据《纽约时报》[1]的报道，以Udacity、Coursera、edX为代表的MOOC平台在2012年先后与世界一流名校合作，向所有学生免费提供该校最受学生欢迎的若干课程，因此，2012年被真正视为“MOOC的启航之年”。自此，MOOC课程开始了井喷式的发展，吸引了大量学生上网学习，平台的课程也越来越多，课程内容也越来越丰富。至2017年5月为止，Coursera已经与超过150所高校机构展开合作，开设了超过2500门课程，并拥有上千万的注册用户。这种可以容纳几十万学生在同一课堂、允许任何学生在任何时候进行学习的教学方式，成为了推动全世界范围内教育发展的一个具有时代意义的里程碑。

MOOC相比于与传统远程教育，加入了更先进的教育理念。首先，在视频制作方面，MOOC不再是把传统课堂上的教学录像直接上传，而是由教师制作相对简短的视频片段，将原先一个课时的内容拆分为多个10分钟左右的短视频，从而能够让学生安排更自由的学习计划。其次，同伴互评作为MOOC平台解决作业批改问题的重要手段，不仅使学生了解了他人的学习方法、增强批判性思维，还能够在互评时，通过分析、概括、总结等认知过程，进一步提升自己对知识的理解和掌握。最后，MOOC还提供了在线论坛服务，以便让学生与学生、学生与老师之间能够相互沟通交流，同伴之间的督促、交流、合作使得学生之间的思维碰撞更密集，进而互相学习共同进步，师生间的答疑解惑能够让学生及时解决学习中遇到的困惑。

## MOOC存在的问题

虽然MOOC平台为学生提供了丰富的、优质的教学资源，是对传统教育模式的一次革新之举，但是MOOC的高流失率、学习体验不佳等现象一直为人所诟病[2]。这些批判意见反映出MOOC内在的很多问题，例如缺乏对学生在学习过程中的引导和交互、教师自身专业领域的研究经验无法传达给学生、在线教育缺乏科学的学科体系和系统的训练过程、以及学生对知识仅仅停留在记忆和理解阶段而不是实际应用等等[3,4]。

以最为著名的全球首批MOOC课程——麻省理工大学电路与电子 MOOC 课程为例，其最终的课程完成率不足5%，其他大多数类似课程的完成率都不足13%。又比如，北京大学2013年秋季开设的“数据结构与算法”课程总共有13000多人注册，但是老师和助教团队只有十人左右，师生之间只能通过论坛进行交流讨论，但往往老师助教不能及时解答学生的困惑，而当学生怀揣着对困惑的无力感结束这次学习时，他们会比别的学生更容易流失。

因此，我们需要对课程的教学机制进行一些设计，对学生的学习过程进行一些干预，从而能够使MOOC课程主动引导学生进行学习，并提升其学习体验和学习效果[5,6]。

## 相关措施和解决方案

在传统的课堂上，老师可以随时注意到每名学生的学习状况，并根据学生的掌握情况进行有针对的辅导和帮助。而在大规模在线公开课程中，老师无法直接了解到学生潜在的困惑，并且教学团队也很难照顾到每一位学生。因此，智能化的学生引导是当前提升MOOC学习体验的最重要的方式之一。例如，根据学生在论坛上的发帖，分析学生是否对某个知识点存在困惑，从而为其推荐相关的学习资源[7]；利用学生提交测验的答题记录，分析学生是否掌握了某个知识点，以便对学生更准确地评估并予以辅导[8]；通过在视频中间插入一些小问题，一方面重新唤起学生的注意力，另一方面可以让学生自我检验[9]。

然而，上述学习引导方法仅仅只是在学习过程中或学习过程后，通过与学生的交互和学生的反馈，采取相对应的措施。鉴于很多学生仅仅在MOOC课程中观看视频，但却很少登录论坛发帖、完成每周的测验，因此，对于这些缺乏学习反馈信息的学生，系统很难准确地评估其学习状况，从而无法做出准确的引导。所以，如果能够在学生开始学习之前给予一些适当的引导，也许可以帮助到更多的学生，并增加其对课程的归属感。

随着MOOC的发展，传统的教育学观点也在不断变革，Conole在其7C教学设计框架中提及，为了让学生对知识的概念有一个预先了解（Conceptualize），应当在课前先告诉学生这门课程或是这个课时主要讲的是什么、这节课的重点和难点是什么、学生应当在学习完之后掌握些什么。在传统课堂上，老师总是会强调这些重要的知识点，但在MOOC中，这些重复的内容显得有些冗余，往往在制作视频的过程中会被删除。此外，即使老师在视频中反复强调某个知识点，MOOC学生也容易错过这些信息。研究表明，在edX平台上的学生中，即使是那些最终获得证书的学生，平均也只看了每个视频的前三分之一[10]。

本工作主要是分析MOOC的视频字幕文件，找出当前学习模块中一些最重要的知识点。这种自动化的学习引导不仅有助于学生对即将学习的内容有一个大致的了解，还能够为学生在考试前提供复习指导纲要，从而在学习之初有一个良好的学习体验。

# 前沿技术调研报告

本工作主要研究通过告知学生即将学习内容的重要知识点，以引导学生更好地进行学习，其中关于知识点的重要性排序算法是一种基于PageRank的扩展。下面将主要针对在线学习引导、知识点定义和PageRank算法及其扩展三方面介绍相关前沿技术。

## 在线学习引导

学生参与到MOOC学习的途径通常包括观看视频、提交作业、浏览论坛等等。因此，大部分学习引导和干预手段也根据这些在线学习行为所设计的。一些工作研究了关于视频制作和视频内容对学生参与度的影响，例如：单个视频的长度、授课教师的讲课方式、教学工具的设计和准备、视频中是否穿插几个小问题等等[9,11]。还有一些工作研究了学生在提交作业时的动态行为记录，例如：学生对各个知识点的掌握情况、实时反馈意见对学生成绩的影响、推断学生解题的策略并为其他学生生成提示信息等等[8,12,13]。另外，部分研究根据学生在论坛上的发帖内容，分析学生是否对某个知识点存在困惑、是否需要老师和同学的帮助、以及如何给存在困惑的学生推荐视频片段等等[7,14,15]。

然而，上述学习引导方式大都是在学生真正开始学习之后采取的干预。很大一部分学生在看视频遇到困惑时会选择放弃，也不会在论坛上提问，并且以学会某个知识点为目标的学生通常不会尝试提交作业。因此，应当在学生学习知识之前就进行交互、干预、引导，而不是在学习完成之后的反馈阶段。例如，Basu等人提出在学生尝试某个编程题之前，通过向学生提问样例输入的输出结果，从而帮助学生真正理解题意。与这些引导措施的不同之处在于，本工作旨在为学生学习之初提供一些学习引导，通过视频内容的分析，告诉学生即将学习的内容中最重要的知识点是什么。

## 知识点定义

在不同的研究领域中，知识点的定义不尽相同。在传统的智能教育系统（Intelligent Tutoring Systems, ITS）中，知识点是由专家预先定义好的[16]。近年来不断有工作利用学生在测试中的答题情况[17,18]、相关垂直领域知识图谱[19]等工具，自动抽取某个课程的知识点。比如，可以利用深度神经网络和学生的答题记录，训练得到题目的嵌入表达，再通过聚类方法定义知识点。

然而，并非所有MOOC课程都有大量的测试题或是领域知识图谱，因此在MOOC环境下，缺乏训练数据的模型将无法在实际课程中使用。相关的MOOC研究工作中，都简单地将课程中的一个章节或是一道题作为一个知识点[8,20]，忽略了一个章节或一道题中可能包含多个知识点的情况。借鉴于Matsuda等人将LDA（Latent Dirichlet Allocation）模型[21]应用在习题集上，从而自动抽取题目内容中隐含的知识点[17]，本工作将LDA模型迁移到了视频字幕文件中。由此，本工作中的知识点都被表示为关于单词的分布，并且这种以单词分布形式出现的知识点不仅便于之后的重要性计算，还能够以“词云”形式进行展示。

## PageRank算法及其扩展

为了对知识点的重要性排序，首先要求出单词的重要性，然后根据知识点关于单词的分布，加权求和得到知识点的重要性。在信息检索领域，PageRank算法是一种根据网页之间相互的超链接，来衡量网页相对重要性的图模型[22]。在此基础上，又有很多研究对该算法进行扩展，例如加入外部信息从而得到更准确的结果[23]、如何将PageRank应用在无向图上 [24]等等。

在自然语言处理领域中，Mihalcea等人提出的TextRank将PageRank应用于文本信息上[25]，可以用来完成信息抽取、文档摘要等任务。此类研究首先根据单词在文本中的共现情况构建关于词的图，然后采用Random Walk算法求出各个单词的重要性。与上述研究不同的是，本工作将利用MOOC中各章节发布的时序关系，在构建词图时以不同权重加入之前章节的词与词的关系，并在得到单词重要性之后，利用转换函数平滑得到的数值和知识点关于词的分布，求出最终每个知识点的重要性。

# 科研实践工作（具体）

## MOOC学习引导框架概述

本工作的主要目标是：在学生开始学习课程内容之前，通过自动提示学生章节中的重要知识点，从而为其提供学习引导。基于这种引导方式，学生能够对即将要学习的内容有一个大致的了解，从而在学习过程中对这些部分产生更高的集中力。又或是即将进行测验时，学生可以通过浏览这些重要的知识点，检验自身是否掌握了最主要的课程内容。简单来说，本工作提出了一个MOOC学习引导框架，以课程视频的字幕文件作为输入，返回一个按重要性排序的知识点列表。图1展示了这个引导框架的整体结构。

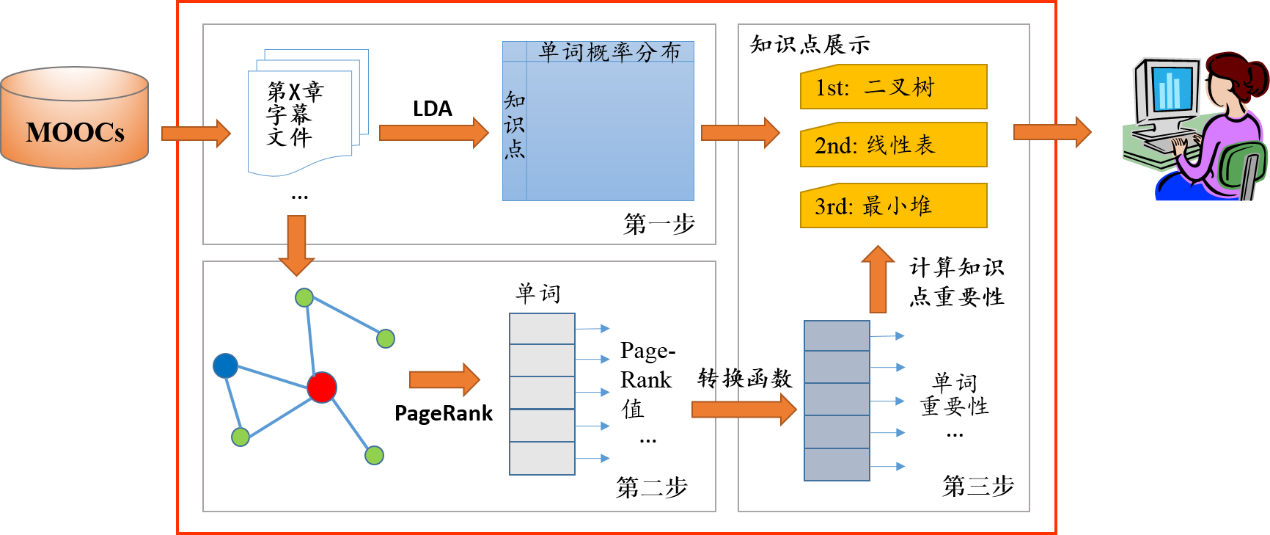


图 1. MOOC学习引导框架示意图

该框架可以大致可分为三步，将在之后的3.2、3.3、3.4节中依次详细介绍。首先，基于输入的字幕文件，使用LDA模型抽取课程内容中的知识点；其次，利用单词在字幕文件中的共现性构建词图，并使用PageRank算法求出每个词的重要性；最后，使用转换函数对单词的重要性分布进行平滑，并最终求得每个知识点的重要性。

## 知识点抽取

本工作首先对字幕文件进行预处理，利用Python词性标注工具，提取出文本中的名词词组，并采用一些启发式规则删除一些停用词、代数符号、问候语等，从而用得到的名词序列来表示视频的内容。然后，借鉴于之前的相关工作[21]在习题集上使用LDA模型抽取知识点，本工作将LDA模型应用于学生即将学习的课程内容上，从而得到更合适、更准确的知识点。

LDA是一种生成式的概率主题模型，也就是说，该模型认为文档的每个词都是通过“在给定先验分布下，以一定概率选择了某个主题，并从这个主题中以该主题关于词的先验分布的选择某个单词作为最终结果”这样的过程得到，该图模型的表示如图2所示，其中α和β是先验参数，θ是主题的多项分布，z是某个选定的主题，w是根据该主题的多项分布以及先验生成的单词。通过训练数据的学习，该模型可以识别语料库中潜藏的主题信息，每个主题都以单词的分布形式呈现，权重越高的单词越能表现该主题的含义。本工作将字幕作为文档，将预处理得到的名词作为单词，从而LDA模型训练得到的隐含主题就可以视为本工作中的知识点，即每个知识点都可以表示成关于单词的概率分布。

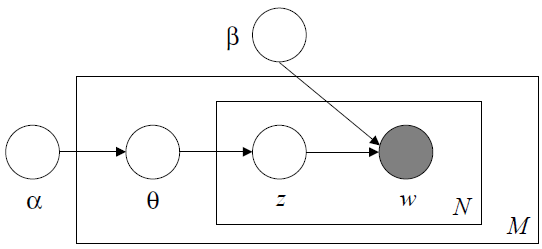


图 2. LDA的图模型示意图

另外，即使课程本身已经有预先定义好了的知识点，也可以将其转化成关于单词的概率分布形式，以便后续重要性的计算。具体来说，如果教师给定了n个单词作为某一个知识点，本工作仍然可以用一个关于单词的概率分布来表示这个知识点，即：将这n个单词在分布中的概率设置为1/n，其余单词的概率设置为0。表1展示了LDA模型应用于课程字幕文件后的实际效果，其中每个知识点都选用其出现概率最高的三个单词来表示。

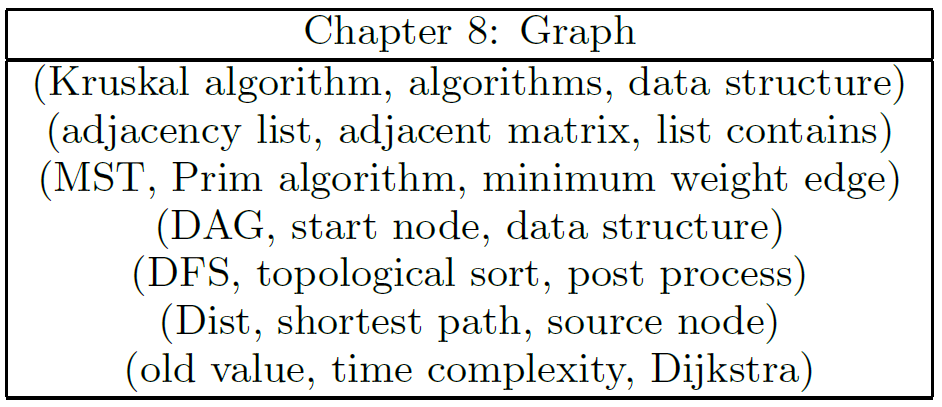


表 1. LDA模型在MOOC字幕文件上应用生成的知识点表示形式

## 基于有序文档的PageRank

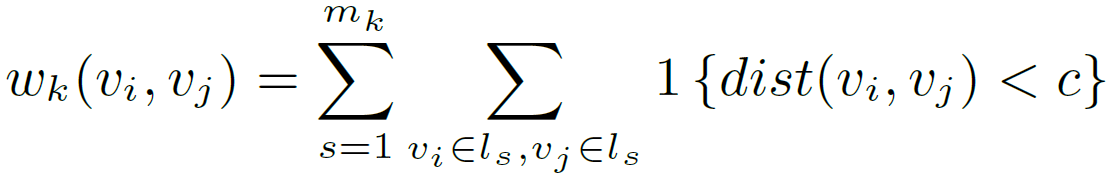
在抽取出各个章节的知识点之后，本工作利用了视频的字幕文件对课程内容进行分析，并求得每个单词在当前文档集合中的重要性。首先，出于直觉和经验上的考虑，越重要的单词越容易在视频中被反复提及，并且老师在讲述一个新的是知识点的时候，往往会拿相关的知识点作比较，以便学生进行记忆。因此，基于词的频数和共现性，本工作主要设计了一个特别的PageRank方法，在构建词图时利用了文档之间的顺序关系，并最终计算各个章节中单词的重要性。

具体来说，算法可以分为三步。首先，基于单词的共现性为每一个章节构建一个词图；然后对于每一个章节，将其之前已发布的章节词图按不同权重相加，得到本章节的一个扩展词图；最后，在该扩展词图上利用随机游走算法计算单词的重要性。这部分算法的最终输出是一个词组的有序列表，并按照单词的重要性进行排序。

### 构建词图

考虑到单词在文档中的“共现性”，即在文本中相近位置出现两个单词会具有相近的含义，而当老师讲解比较重要的知识点的时候，往往会用很多关键的术语，因此重要单词之间的联系会更紧密一些。本工作在构建词图时，将名词词组视作结点，对位置相近的单词之间连边，从而构建一个带权图。具体形式化的定义如下：

给定第k个章节，由若干已经预处理好的字幕文件构成，可以构建一个无向图，其中表示单词集合，表示边集。因此，本工作借鉴于TextRank[25]构图方法，对每一个章节分别构建一个基础的词图，即如果两个名词词组在文档中的位置小于预设的参数c，则将这两个词之间连一条边，边的权重为这两个单词在所有本章节字幕文档中的“共现”次数。上述定义可公式化为：



其中1{Expression}是一个指示函数，dist(vi,vj)表示两个单词vi和vj在文档中的偏移距离。从公式中可以发现，如果两个单词同时出现得越频繁，则这两个单词之间的权重越高。图3给出了一个具体构建词图的样例。

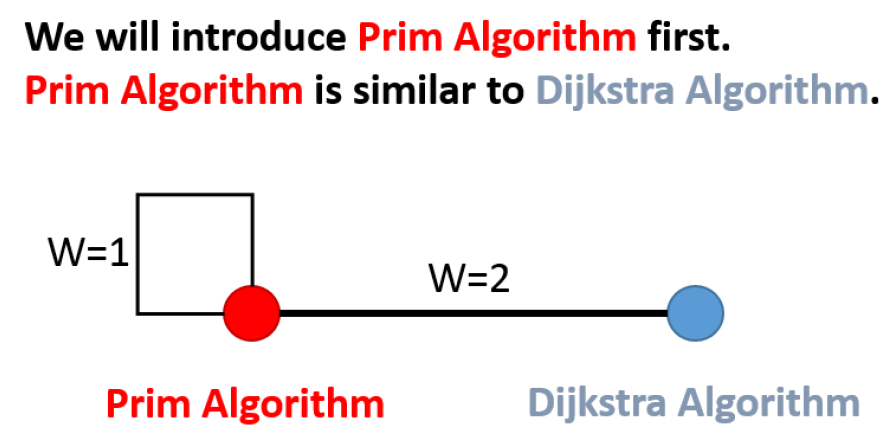
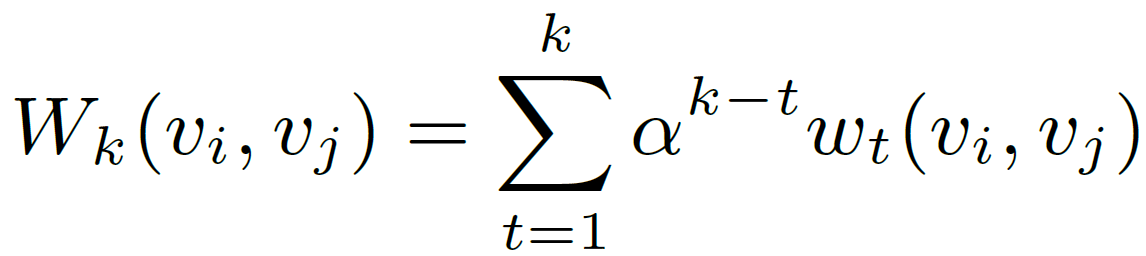


图 3. 一个构建词图的样例：根据给定上面的文本段落，构建一个以名词词组为结点的带权图

### 章节组合

又考虑到老师会避免重复讲解同一个知识点，在第二次提及和应用的时候往往会被省略。因此，如果只是使用本章节的字幕文件，可能词与词之间的关系会不够充分。例如，在如下一个选自“二叉树”章节的段落中，“我们使用队列来实现广度优先搜索，…，二叉链表是一种用来存储二叉树的方法。”，涉及到实际应用的知识点“队列”和“二叉链表”之间可能不存在关联，但如果加入了“栈与队列”章节，就可以对“队列”和“二叉链表”的关系进行补充。因此，当单词之间互相传播重要性的时候，某些单词可能不会直接相连，而通过之前章节的补充，可以更加完善这一章节的词图。

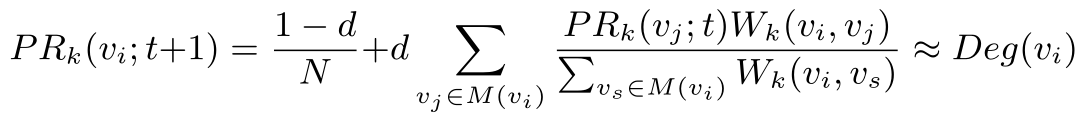
基于上述情况，本工作对不同章节的词图进行组合，使得当前章节的词图能够包含之前章节单词互相之间的关系。因此，本工作提出一种基于有序文档的组合方式，将不同章节的词图组合在一起。具体来说，当需要对某一章节的单词重要性进行排序时，为之前已发布章节的词图赋予一个受衰减因子α影响的权重，距离本章节越近发布的内容权重越高。其边的权重如下所示：



### 计算单词重要性

最后在计算过程中，我们需要求出章节词图中每个单词的重要性。PageRank是根据网络中的相互连接关系，利用随机游走模型，在每一轮计算中，各个结点将自身的值分发到其相邻节点上，经过不断重复的迭代计算，最终趋于一个收敛的状态，而这个状态下各结点的值就可以用来定义其重要性。本工作在实验过程中，将迭代轮数设置为20（足够到达收敛状态），并根据经验规则设置阻尼系数d为0.85。

具体迭代计算的过程可以参考如下公式。首先，本工作在单词重要性初始化时，定义每一个单词具有相同的重要性，即，其中N是词图中总的结点个数。在随后的每一轮迭代中，各个结点的PageRank值变化情况如：



其中表示单词在第k章中经过t轮迭代计算的PageRank值，代表的相邻结点集合。该计算过程保证了每一轮迭代后，各个结点的PageRank值之和恒等于1。

## 计算知识点重要性

基于3.2和3.3节的算法，可以得到以单词概率分布形式表示的知识点，以及每个单词的重要性。如果简单地将单词重要性乘以其在知识点中的权重然后加权求和，固然可以得到关于知识点的重要性，但PageRank方法只考虑了单词之间的相对重要性，会使得排名靠前的单词之间的重要性差别很大，从而个别单词在知识点中的权重会决定整个知识点的重要性，而与单词的重要性无关。为此，本工作尝试利用三种常用的转换函数对单词重要性进行平滑，分别是线性函数、sigmoid函数和正态分布函数。

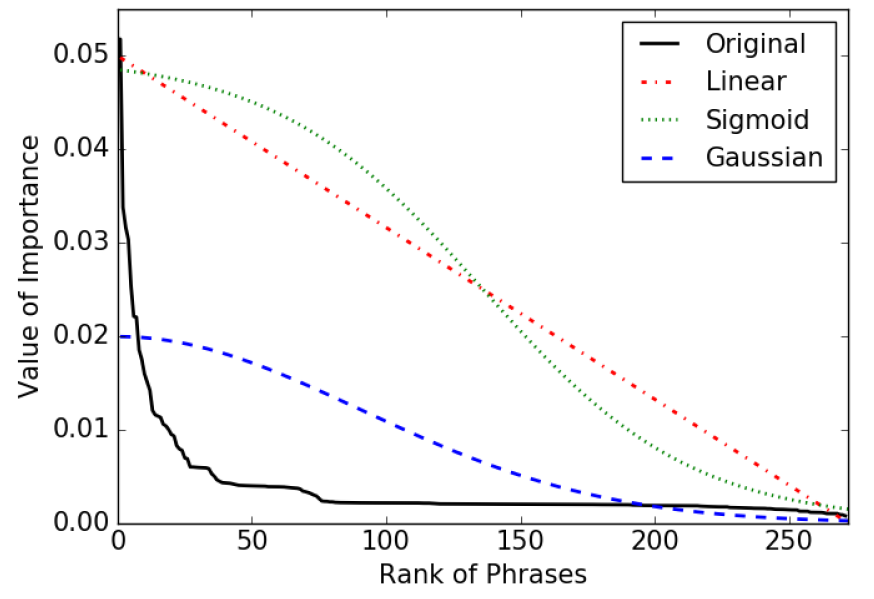
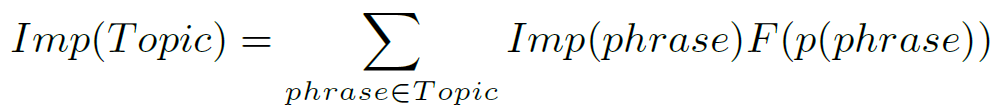


图 4. PageRank以及三种用于平滑的转换函数关于单词重要性分布的对比

这些转换函数主要都有两个共同的性质，一是单调递减函数，二是下降的梯度更平缓一些，从而可以避免在前几个单词中出现的重要性骤减的情况。图4是原始的PageRank值的分布和经过转换函数平滑后的分布的对比。而经过上述转换之后，可以得到一个变化更平缓的单词重要性分布，将该分布与知识点关于单词的分布相乘，得到的乘积即为知识点的重要性。具体的公式如下：



其中，p(phrase)表示某个单词在知识点中出现的概率，F表示某个转换函数。计算得到知识点重要性后，将各个知识点排序后输出，作为最终的结果。

## 实验和分析

为了评估上述算法的有效性，本工作通过对章节知识点重要性进行排序的结果比较。本工作使用了4个常用的基准算法作性能比较，知识点重要性的真实值是通过三名领域专家对习题库数据进行人工标注得到的。本工作用了三种评测指标来衡量排序算法和基准算法的效果。

### 实验设计

实验使用的数据来自北京大学《数据结构与算法》课程，其中包括了课程的65个字幕文件和三位课程助教参与的习题标注，预处理后得到共3964个不同的单词和254道习题。

整个实验流程设计为：1）利用每个章节的字幕文件抽取知识点；2）利用本工作设计的模型算法或基准算法求出该章节内各个知识点的重要性，并得到一个排序结果；3）利用标注数据，得到这些知识点关于重要性的最优排序结果；4）比较这两个排序结果，并计算相关评测指标；5）对各个章节的评测指标求平均，作为该算法的实际效果。注：通过采用不同的转换函数和章节组合方式，本工作设计的算法可以有多个变种版本。

### 基准算法

为了评估算法的性能，本工作采用了4种常用的策略作为基准算法进行对比，分别是：1）随机策略；2）词袋模型；3）TF-IDF；4）PageRank在文本中的扩展应用——TextRank。并且为了可比较性，这些基准算法也会采用3.2节中抽取出的知识点作为排序的对象。其中，随机策略顾名思义，就是知识点排序的先后顺序是随机选择的；词袋模型则是用单词在在某一章节的文档中出现的频数来代表其重要性；但由于词袋模型中，某些在每个章节都出现很多次的单词其实并没有意义，因此，TF-IDF通过对原始的词频乘以逆向文件频率加权，来解决这一问题；TextRank方法[25]主要应用于信息检索领域的关键词抽取任务，在本工作的实验过程中，作为基准算法的TextRank不采用章节组合策略和重要性平滑技术。

### 数据标注及评测指标

从学生的角度来看，考虑到完成MOOC课程学习的学生都会努力完成课后测验和考试。因此，本工作认为学生会更重视那些在考试中涉及更多的知识点，并定义本课题的知识点重要性为“涉及到该知识点的习题数量”。为此，有3位课程助教对习题与知识点之间的关系进行了标注。具体来说，给定一个习题集和抽取出的知识点，为每个知识点匹配所有相关的习题，其中3位助教标注结果的统一性为Kappa=0.535∈[-1, 1]，具有比较强的一致性。考虑到助教对生成的知识点理解的差异，本工作采用三个标注结果的并集作为最终知识点与习题的对应关系。因此，知识点重要性在本工作中就可以数值化为这个标注习题集合的大小。

在得到知识点重要性的标准值后，就可以对知识点进行排序，从而产生最优的排序结果，以供算法评测时作为对比。本工作使用排序模型常用的三种评测指标：nDCG、MAP和Kendall’s τ。

### 结果和分析

表2展示了本工作提出的算法与各基准算法之间的对比，其中本工作的7种具有不同参数的算法主要区别于两个方面，一是是否加入了其他章节的内容进行词图补充，二是是否利用了且利用了何种转换函数。通过表中数据可以发现，本课题提出的所有算法及其变形都能够赶超基准算法，其中，最优的算法变种(α-PR-sigmoid) 比TextRank方法在MAP指标上高出了18.9%。另外，在三种评测指标上，结果都显示出了较强的一致性。最后，可以发现通过章节组合策略和重要性平滑技术，都能显著地提升算法的效果。

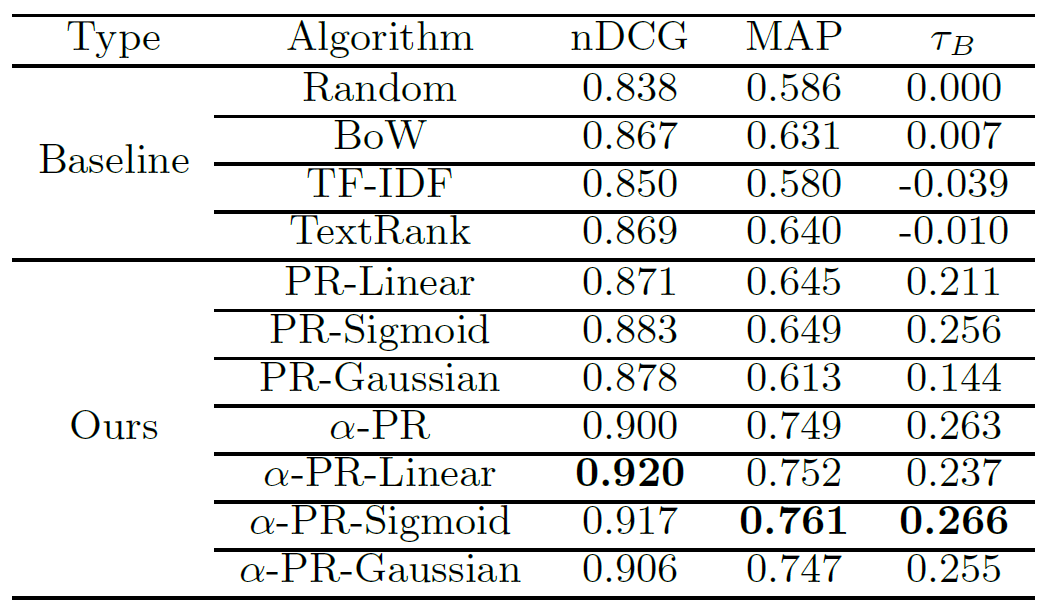


表 2. 实验结果：算法性能比较。（三个指标均为越大越好）

本工作提出的算法能够优于其余基准算法（尤其是词袋模型和TF-IDF）的可能原因有：1）PageRank利用了词与词之间的关系，因此，一个单词即使不是频繁出现，但只要周围都是一些重要的概念，同样可以得到较高的重要性；2）将之前章节的词图与本章节的词图加权组合，能够使词与词之间的关系更完整、更准确，因为某些之前章节已经讲解过的但却重要的知识点在本章节会被一笔带过；3）转换函数减少了PageRank值与真实重要性之间的偏差，这样就能避免某个单词在知识点中出现的概率决定了知识点重要性的情况。

另外，本工作在进行词图的章节组合时，衰减因子α是预先设定的。因此，为了考察不同的参数对算法效果的影响，本工作进一步对参数进行分析，结果如图5所示。图中使用了nDCG作为指标来观测算法性能随参数α改变而产生的变化情况，事实上，在其余两种指标下，变化情况也是类似的。值得注意的是，当α取值为0时，算法就退化为了不进行章节组合的情况。

图5中还可以看到一个有趣的现象，当α从0.05升至1.00时，对于使用了转换函数的三种算法来说，性能呈现递减的趋势，但如果直接使用计算得到的PageRank值，则算法性能基本保持不变，除了在α=1.00时有一个较大的提升。因此，在表2的实验结果展示中，对于使用转换函数的算法，α预设为0.05，对于没有使用转换函数的算法（α-PR），α预设为1.00。可能产生这种现象的原因是，当使用转换函数时，一个较小的α会在不影响本章节词与词之间关系的前提下，适当补充之前章节的内容。但在直接使用PageRank值的情况下，排在前几位单词的重要性会被夸大，因此α取1.00时能够通过加入其他章节的词与词之间关系来“缓释”这种影响。

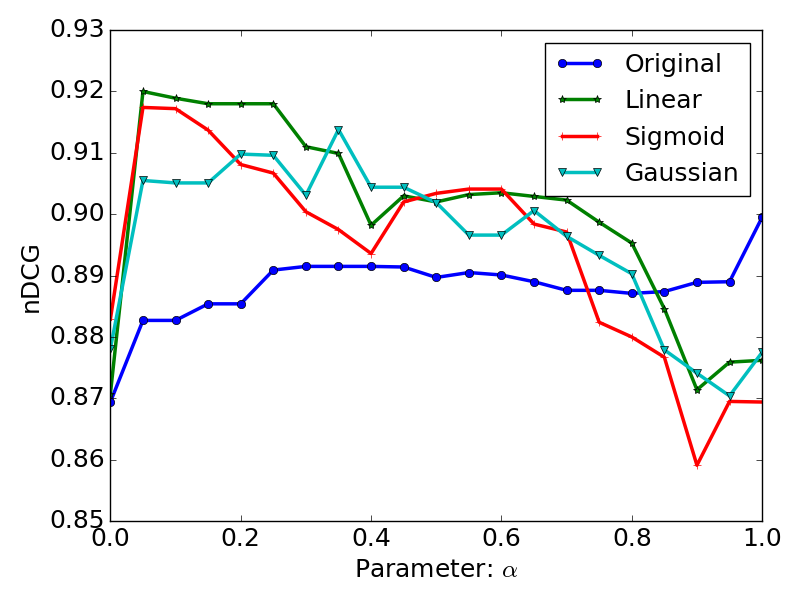


图 5. 参数分析：四种使用不同转换函数的算法随着α变化指标nDCG的变化情况

## 讨论和总结

上述实验展示了本工作提出的算法在对每章的知识点重要性排序任务上的性能，从而可以为学生提供学习前的指导，让学生指导下一章学习内容的重点。不仅如此，本工作实则还能够应用于其他学习场景，例如针对学生的学习状况，分析考前复习重点。具体算法思路与本工作类似，即在构建词图和章节组合的时候，将各章节的词图按照学生章节测验的成绩进行加权组合，从而可以计算出对该学生来说需要着重复习的知识点。

总的来说，本工作提出了一个创新的算法框架，用于为学生提供学前指导，分析章节内容中的重要知识点。算法首先通过LDA模型从字幕文件中抽取知识点，然后基于PageRank的思路求解单词的重要性，最后对单词重要性进行平滑并计算知识点的重要性。实验结果也展示出了该算法的有效性。未来对本工作的进一步扩展将从三方面入手，一是扩展数据集的多样性，考察算法在其他课程上的效果，二是在真实场景中进行个性化学习重点分析，三是针对这些学习重点推荐相关学习资源。

# 参考文献

[1] Pappano, L. (2012). The Year of the MOOC. The New York Times, 2(12), 2012.

[2] Anderson A, Huttenlocher D, Kleinberg J, et al. Engaging with massive online courses[C]//Proceedings of the 23rd international conference on World wide web. ACM, 2014: 687-698.

[3] Conole G G. MOOCs as disruptive technologies: strategies for enhancing the learner experience and quality of MOOCs[J]. Revista de Educación a Distancia, 2015 (39).

[4] Sonwalkar N. The first adaptive MOOC: A case study on pedagogy framework and scalable cloud Architecture—Part I[C]//MOOCs Forum. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA: Mary Ann Liebert, Inc., 2013, 1(P): 22-29.

[5] Basu S, Wu A, Hou B, et al. Problems before solutions: Automated problem clarification at scale[C]//Proceedings of the Second (2015) ACM Conference on Learning@ Scale. ACM, 2015: 205-213.

[6] Kulkarni C E, Bernstein M S, Klemmer S R. PeerStudio: rapid peer feedback emphasizes revision and improves performance[C]//Proceedings of the Second (2015) ACM Conference on Learning@ Scale. ACM, 2015: 75-84.

[7] Agrawal A, Venkatraman J, Leonard S, et al. YouEDU: addressing confusion in MOOC discussion forums by recommending instructional video clips[J]. Educational Data Mining, 2015.

[8] Wang Z, Zhu J, Li X, et al. Structured Knowledge Tracing Models for Student Assessment on Coursera[C]//Proceedings of the Third (2016) ACM Conference on Learning@ Scale. ACM, 2016: 209-212.

[9] Kovacs G. Effects of In-Video Quizzes on MOOC Lecture Viewing[C]// Proceedings of the Third (2016) ACM Conference on Learning@ Scale. ACM, 2016: 31-40.

[10] Fowler G A. An early report card on massive open online courses[J]. The Wall Street Journal, 2013, 8.

[11] Guo P J, Kim J, Rubin R. How video production affects student engagement: An empirical study of mooc videos[C]//Proceedings of the first ACM conference on Learning@ scale conference. ACM, 2014: 41-50.

[12] Kulkarni C E, Bernstein M S, Klemmer S R. PeerStudio: rapid peer feedback emphasizes revision and improves performance[C]//Proceedings of the Second (2015) ACM Conference on Learning@ Scale. ACM, 2015: 75-84.

[13] Piech C, Sahami M, Huang J, et al. Autonomously generating hints by inferring problem solving policies[C]//Proceedings of the Second (2015) ACM Conference on Learning@ Scale. ACM, 2015: 195-204.

[14] Chaturvedi S, Goldwasser D, Daumé III H. Predicting Instructor's Intervention in MOOC forums[C]//ACL (1). 2014: 1501-1511.

[15] Yang D, Wen M, Howley I, et al. Exploring the effect of confusion in discussion forums of massive open online courses[C]//Proceedings of the Second (2015) ACM Conference on Learning@ Scale. ACM, 2015: 121-130.

[16] Baker R S J, Corbett A T, Koedinger K R. The difficulty factors approach to the design of lessons in intelligent tutor curricula[J]. International Journal of Artificial Intelligence in Education, 2007, 17(4): 341-369.

[17] Matsuda N, Furukawa T, Bier N, et al. Machine Beats Experts: Automatic Discovery of Skill Models for Data-Driven Online Course Refinement[J]. International Educational Data Mining Society, 2015.

[18] Piech C, Bassen J, Huang J, et al. Deep knowledge tracing[C]//Advances in Neural Information Processing Systems. 2015: 505-513.

[19] Dai Y, Asano Y, Yoshikawa M. Course Content Analysis: An Initiative Step toward Learning Object Recommendation Systems for MOOC Learners[J]. International Educational Data Mining Society, 2016.

[20] Pardos Z, Bergner Y, Seaton D, et al. Adapting bayesian knowledge tracing to a massive open online course in edx[C]//Educational Data Mining 2013. 2013.

[21] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(Jan): 993-1022.

[22] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: Bringing order to the web[R]. Stanford InfoLab, 1999.

[23] Haveliwala T H. Topic-sensitive pagerank[C]//Proceedings of the 11th international conference on World Wide Web. ACM, 2002: 517-526.

[24] Grolmusz V. A note on the pagerank of undirected graphs[J]. Information Processing Letters, 2015, 115(6): 633-634.

[25] Mihalcea R, Tarau P. TextRank: Bringing order into texts[C]. Association for Computational Linguistics, 2004.

# 附1：已发表论文

Jile Zhu, Xiang Li, Zhuo Wang, Ming Zhang. An Effective Framework for Automatically Generating and Ranking Topics in MOOC Videos [C]. The 10th International Conference on Educational Data Mining, 2017. Accepted.

Ming Zhang, Jile Zhu. A data-driven analysis of student efforts and improvements on a SPOC experiment[C]//Proceedings of the ACM Turing 50th Celebration Conference-China. ACM, 2017: 1.

# 附2：申请发明专利、或获得发明专利授权 、或取得通过技术鉴定的相关成果