历时语料库阶段报告

吴先 wuxian94@pku.edu.cn

# 概述

语料库，一般经科学取样和加工的大规模电子文本库。借助计算机分析工具，研究者可开展相关的语言理论及应用研究。利用大规模语料，对语言现象进行观察，可以为词典编纂、语言教学、传统语言研究、自然语言处理中基于统计或实例的研究方法提供帮助。

历时语料库是在普通的语料库的基础上，加入时间轴，在时间维度上对语料进行划分并分析的语料库。借助历时语料库，语言学家们可以快速沿着一个时间段的切面，从真实的语料出发观察语言现象并加以横向对比，从而挖掘出词语语义、语言风格、词语用法甚至语法等随时间的变化并举出例证。

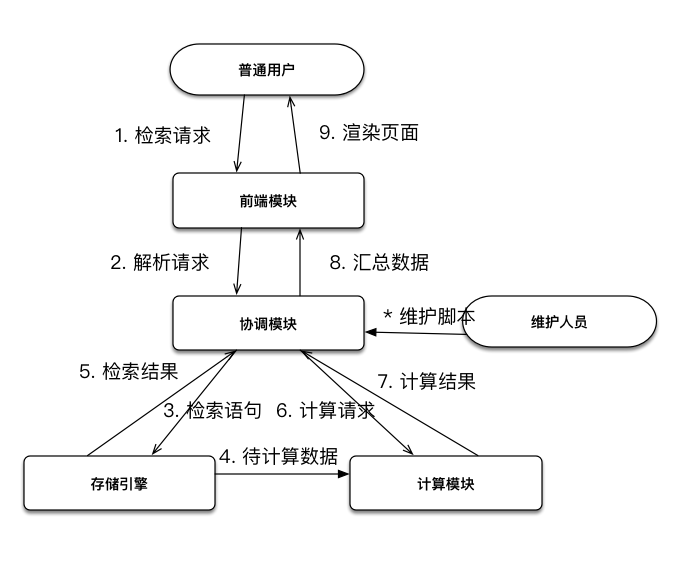
本阶段我们完成了一个历时语料库的可扩展软件结构，实现了一个具备基本语料库功能的语料库Web原型。目前正在进行的工作是设计并实现一个交互式查询系统，利用一个简化的命令式语法与语料库交互，将查询结果显示在一个通用笔记本上，使研究者的研究过程在一个连贯的环境下进行；同时，允许研究者对查询出的内容进行修改与保存，得到自己个性化的结果；这种人工介入的行为也可能为语料库数据带来监督式的修正。

# 目前进展

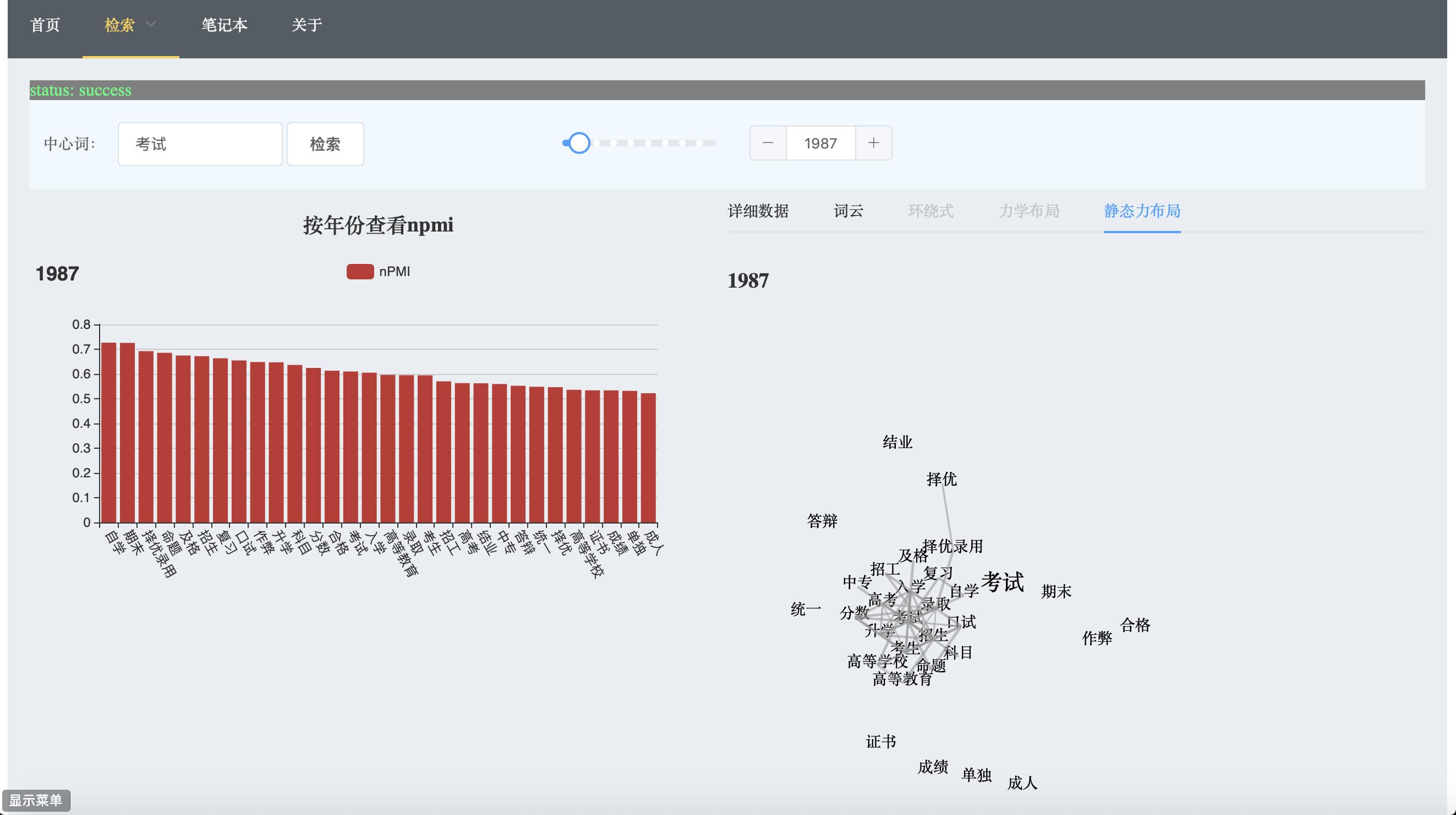
## 总体结构与模块细节

平台共分为4个模块，分别是前端模块、协调模块、存储引擎和计算模块。

* 协调模块，用于协调各模块的运作。是后端服务器的主要组成部分，接受前端的请求后，将请求分发给计算模块，或是直接向存储引擎发起查询，对返果进行筛选和组织后予以返回。同时还要负责执行维护脚本，扩充数据等。
* 前端交互模块，通过web服务，为普通用户提供数据库的全功能、友好的检索接口/页面。输入为检索要求，输出为检索的结果、数据统计及相关图表和下载链接。可视化方面，通过力学模型对词语的相关词进行了随时间连续变化的动态展示，可以反映一个词语用法和语义上的变化趋势。越相关的词会越靠近中心，同时近义词之间也会发生聚集。
* 存储和检索模块，存储和检索语料数据，并加入基本的数据统计功能。利用关系型数据库进行存储，通过SQL语句进行简单统计，并切对语料建立了索引和缓存，提高了数据查询的响应速度。
* 计算模块，用于进行计算密集型任务的处理，比如同义词、同义句检索等。通过RPC框架与协调模块进行通信，降低了系统内部的通信延迟，提高吞吐量。



图表 历时语料库体系结构。扁圆代表操作人群，圆角矩形代表软件模块，箭头代表请求或数据流向，数字代表一次完整请求的请求/数据流。



图表 历时语料库 PMI检索可视化效果

## 不足之处

原型完成后，试用发现与本系统目前市面上的语料库并无明显差异，很多功能是可以被其他语料库配合简单的处理取代的。在开发过程中，BCC语料库也开放了历时查询功能（虽然目前只有词频查询），进一步降低了本系统的竞争力。受Jupyter项目的启发，我们希望能够从交互模式上加以创新，于是启动了下面一个阶段的工作。

# 笔记本式交互系统

## 设计思路

我们的目标是为语言研究者提供好的工具平台，从语言学研究者的角度出发，尝试了解进行相关方面研究的一般流程。钱厚生（2002）介绍了语料库在词典编纂方面的应用，并在其中设想了很多语料库分析功能：词频统计、索引生成、语法分析、语体分析、搭配分析和语义分析1，这些在如今都已经得以广泛应用。我们主要要做的是把词典编纂的工作流进行优化，以单词为核心，尽量多地展示其信息，以及方便地扩展出更多查询。

白荃，等（2007）和孙德金（2010）则是展示了如何利用语料库进行词义分析2,3。研究者们主要通过对核心词所处的句子类型、上下文搭配（词性搭配、具体词语搭配、甚至具体词语在前与在后的情况）、担当的句子成分进行分析，进而分析出其表达功能、使用习惯等更感性的结论。

语料库语言学的一个常见的研究方法，是通过构造各种查询语句，得到研究对象词语的用例，总结归纳得出结论，并且可以以例句加以佐证。这种查询模式在现有的语料库中也有做到，比如BNCWeb的查询语法，是通过类似通配符和大括号匹配的方式进行复杂查询的（比如{eat} ++\* up，表示up在eat后3~4个词位）；IntelliText则是提供了一个表单来辅助构建这个复杂的字符。而且一次会话只能处理一种查询，想要对结果进行比较或是并排查看，是需要自己做额外的工作的。

## 体系结构

笔记本式的交互系统参照Jupyter Notebook项目的体系，保留了其简化结构：Web应用程序、笔记本文件和运行核心。

后端服务器

Web应用程序

运行内核

笔记本文件

笔记本管理

用户交互

图表 3 笔记本式交互系统体系结构

笔记本文件保存了显示内容所需要的全部信息。运行核心位于服务器上，与笔记本内容相互独立，但是通过“连接”的概念与之关联。笔记本中所有命令的执行是依赖于内核的。Web应用程序是用户操作的界面，也是修改笔记本内容的唯一途径，所有命令执行后的结果会返回给Web，并修改当前笔记本文件。

# 目前进展

基础功能方面，系统支持笔记本文件的分用户管理，用户可以在任意终端上访问笔记本文件。除了运行代码的功能，笔记本还支持用轻便的markdown语法记录下文本或是富文本内容。

内核所支持的语法是基于函数的简化语言，目前可以通过命令查询一个词组的例句、词频等信息，并将查询结果用多种格式展示与比较。通过对查询结果集类型的区分，我们得以使前端使用可复用组件显示需要展示的信息，降低用户的学习成本。

Wiki类型是当前重点开发的类型，这是一个包含了词组多维度信息的查询集。我们不仅为了显示这种复杂的数据单独设计了展示用组件，还为其设计了独特的交互方式，使用户可以编辑属于自己的词语Wiki，并像操作其他变量一样复制、保存它们。语言学研究者可以一边使用笔记功能记录下自己的思路，一边使用命令查询自己需要的资料，若是进行词典编纂工作，还可以再一边用wiki的编辑功能撰写词条，使工作环境具有连贯性。

图表 笔记式交互系统Wiki数据查询与修改。

# 技术难点

我们希望wiki能够返回更多有意义的数据，为研究者完成更多辅助工作。目前的难点在于，在历时语料上自动挖掘不同的词义，并将上述信息，按词义分组后进行展示。若是挖掘出的词义位于同一时期，则是多义词；处于不同时期，则可以认为是词义的变迁。虽说用户后续对这个结果可以进行二次编辑，所以不需要很高的精确度，但是自主挖掘词义变化依然碰了很多壁，没能将这种主观上的“词义变化”转化成客观可以观察，或是统计出的事实上来。

参考文献

1. 钱厚生. 语料库建设与词典编纂[J]. 辭書研究, 2002, 2002(1): 58-68.
2. 白荃, 岑玉珍. “轻易” 的语义及用法分析[J]. 语言教学与研究, 2007 (5): 76-81.
3. 孙德金. 现代汉语书面语中的代词 “其”[J]. 语言教学与研究, 2010 (2): 55-62.