



# 维基百科编辑战可视化

郭宇涵, 韩勤, 姜宇珂, 王益明 (拼音首字母排序)

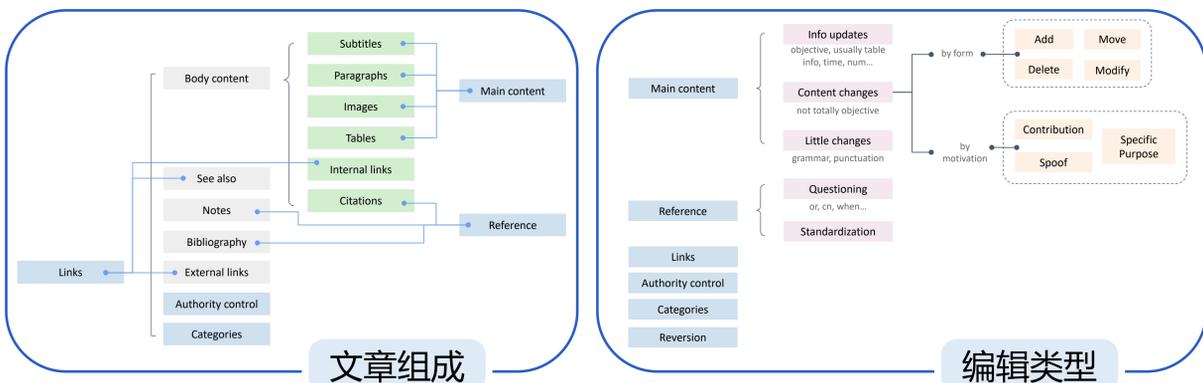
指导老师: 袁晓如

## 项目背景

维基百科是全球网络上最大且最受大众欢迎的参考工具书, 其内容可以被大众用户所修改。开放的系统可以产生高质量、中立的百科内容, 但不可避免的是, 用户针对页面内容的某些方面会有不同的观点。当意见无法达成一致时, 编辑者们可能反复修改文章, 甚至采取挑衅性的编辑行为, 这种情况下会出现一场编辑战。

维基词条通常有成百上千个历史版本, 编辑者的编辑行为也高度复杂, 因此, 我们希望通过可视化的方法还原词条编辑历史中发生的事件。

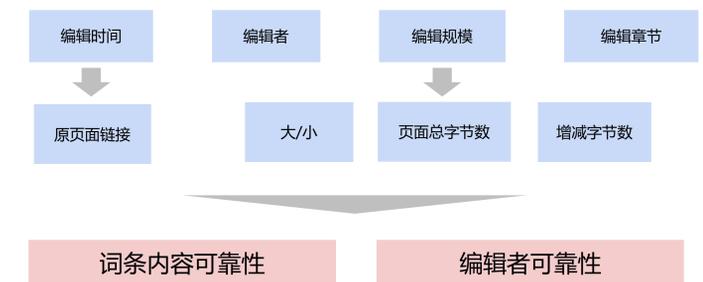
## 维基百科内容分析



## 任务分析

- 时间+编辑: 页面被编辑的频繁程度和修改幅度
- 时间+规模+编辑者: 编辑者对词条的关注情况和贡献情况
- 时间+规模+页面内容: 词条内容演变情况
- 时间+规模+编辑者+页面内容: 更具体地反映不同编辑者的观点, 以及编辑者之间的关系
- 章节: 细化分析粒度

2021年10月27日 (三) 06:42 Sezt0 (讨论|贡献) 小.. (49,675字节) (0) .. (一周周记录) (撤销)  
 2021年10月27日 (三) 06:10 Sezt0 (讨论|贡献) 小.. (49,675字节) (-151) .. (一周周记录) (撤销)  
 2021年10月27日 (三) 06:05 Sezt0 (讨论|贡献) 小.. (49,826字节) (-382) .. (一周周记录) (撤销)  
 2021年10月27日 (三) 05:51 Sezt0 (讨论|贡献) 小.. (50,208字节) (0) .. (一月月记录) (撤销)  
 2021年10月27日 (三) 02:45 Sezt0 (讨论|贡献) 小.. (50,208字节) (0) .. (一周冠军数记录) (撤销)



## 可视化设计

### 词条信息简介

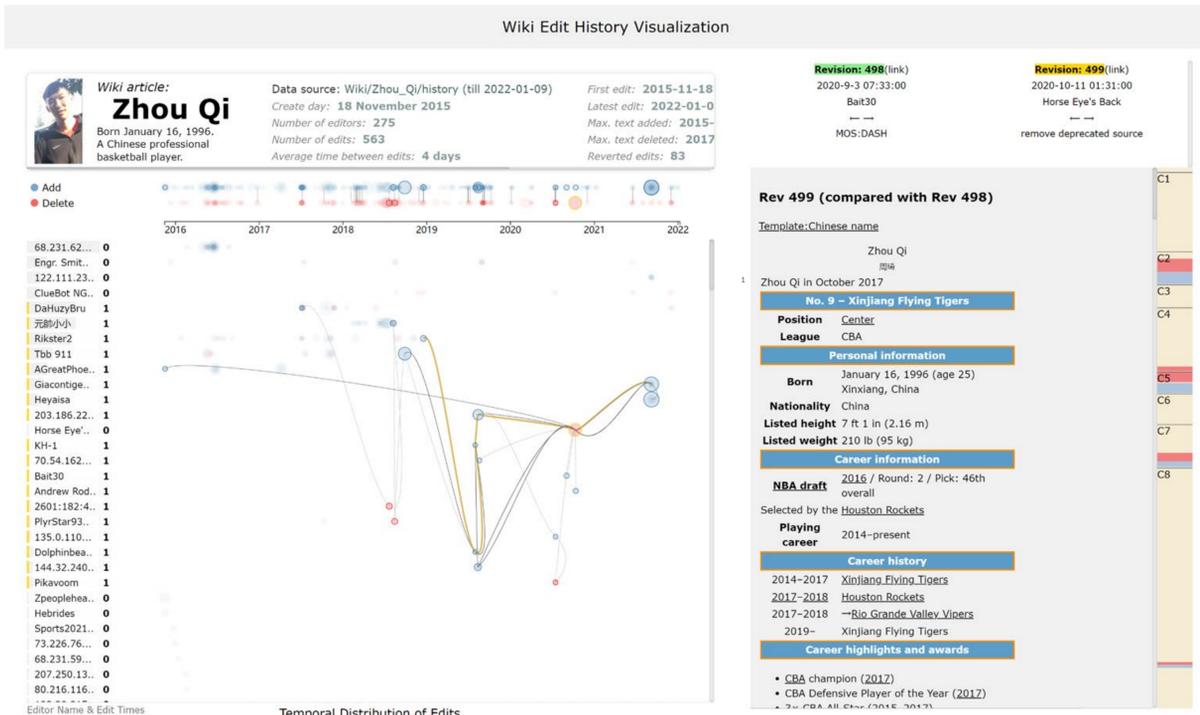
词条信息简介视图展示了词条编辑的概要信息, 包括: 数据来源、词条创建时间、编辑者数量、编辑次数和平均编辑次数等, 并提供到相应维基页面的链接。

### 概览视图

概览视图展示所有编辑历史, 每一透明圆点表示一次编辑, 其圆心横坐标映射编辑时间, 面积映射编辑字节数。我们将所有编辑分为了两行, 分别对应字节数的增删, 同时用直线连接了两个有撤销关系的版本。

### 编辑者视图

每个编辑者的编辑对应编辑者视图中的一行。圆点编码方式与概览视图一致。左侧标注了编辑者ID和编辑次数(灰色矩形)。点击某版本后, 对该版本修改的每个句子, 高亮修改过该句的所有版本并用曲线串联。悬停曲线, 右边滚动到曲线对应的句子。将编辑者重排使得相关版本涉及的编辑者向被选版本编辑者靠近。相关版本数量用金色矩形表示。



### 版本比较

呈现被比较版本的信息, 包括: 版本编号、版本编辑时间、编辑者、编辑评论等信息。并为用户提供选择上/下一版本作为比较版本的交互功能。

### 细节内容

将被比较的两个版本信息整合, 提取并呈现两版本间的编辑内容, 如: 增加、删减和移动, 相应内容用不同颜色呈现。被编辑内容能够与编辑者试图进行交互, 将鼠标移至某一编辑内容能够相应地定位编辑者视图中对应的编辑内容所表达的连线, 以此定位所有修改过内容的编辑者。

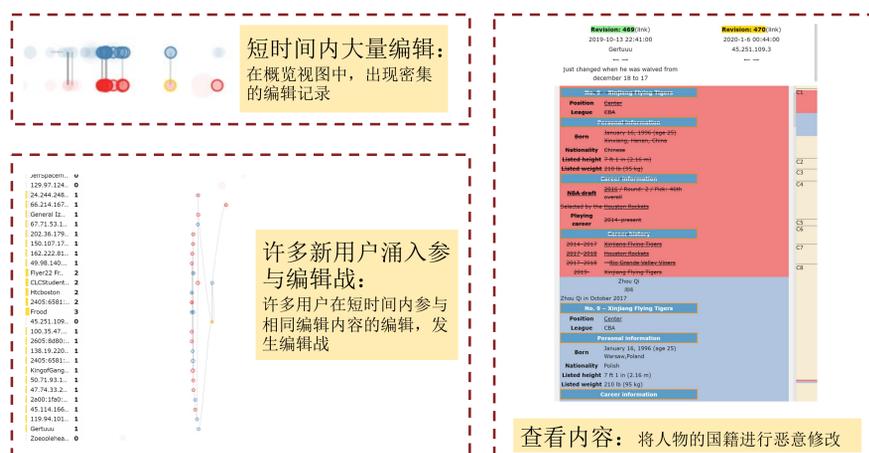
### 缩略导航

缩略导航能够快速定位整合版本中的编辑内容, 红色对应删减内容, 蓝色对应增加内容, 绿色对应移动内容。缩略导航同时能够呈现维基百科文章的章节信息, 方便用户快速定位某一感兴趣的章节。

## 数据处理

- 维基百科编辑历史爬取
- 编辑版本分析
  1. 字段统计
  2. 章节划分: 将内容整理为“章节-字段”层次信息
  3. 编辑信息: 编辑者、编辑时间、编辑注释
- 版本间匹配算法
  1. 根据字段粒度比较不同版本
    - i. 识别内容增加、删除、移动
  2. 将两个比较版本合并为一整合版本
    - i. 标记有改动的字段
    - ii. 计算字段的最小编辑距离, 得到细粒度信息

## 可视化发现



## 未来展望

- 细化编辑类型: 图标设计, 检测算法, 支持用户手动标记
- 文本差异视图: 支持多版本比较; 编码更多细节信息
- 编辑者视图: 图标和曲线编码更多信息, 优化布局(编辑者聚集方式、曲线控制点、减少重叠), 增强交互
- 导入更多词条: 探索不同词条编辑的特点
- 对偶空间: 与第12组内容-时间分布整合, 更多探索空间
- 更多应用: 扩展到共享文档编辑历史、代码版本比较等