维基百科编辑战可视化

郭宇涵 , 韩勤 , 娄宇珂 , 王益明

2022 年 1 月 23 日

1 背景介绍

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，这种情况下就会出现一场编辑战。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。

2 数据描述与任务分析

维基百科保存了词条的所有历史编辑记录。如图1所示，维基百科的编辑历史界面列出了每次编辑的简要信息，包括编辑时间、编辑者 ID、编辑规模和编辑总结。其中，编辑规模包括是否为小修改、编辑后页面增减字节数、编辑后页面总字节数。编辑总结则可能包含被编辑的章节、撤销信息和编辑者自己的总结。通过这些属性，我们可以了解页面被编辑的频繁程度、修改幅度，以及不同编辑者对词条的关注情况和贡献程度。

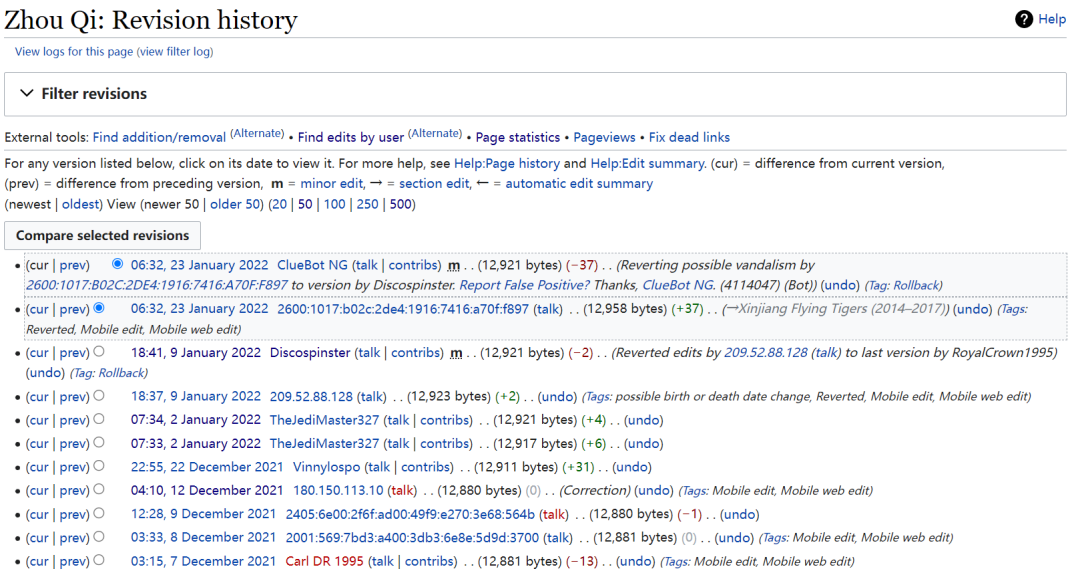

图 1: 维基百科的编辑历史界面

从编辑时间可以链接到该时刻的历史版本，从而获得每个历史版本的具体内容。通过比较相邻两次版本内容，可以得到每次编辑的修改内容。

我们对词条内容和编辑的修改内容做了进一步分析。如2所示，维基词条主要由主体内容、参见内容、注释和引用、外部链接、权限控制和词条类别这些部分组成。主体内容包括章节标题、段落、图片、表格、内部链接和引用。我们将内部链接和外部链接统一归于链接一类，引用和注释统归为参考资料（reference）一类，最终得到主要内容、链接、参考资料、权限控制、所属类别这五个部分。这五个部分在 wiki 语言或 html 语言中有着不同的语法或标签，便于算法区分。

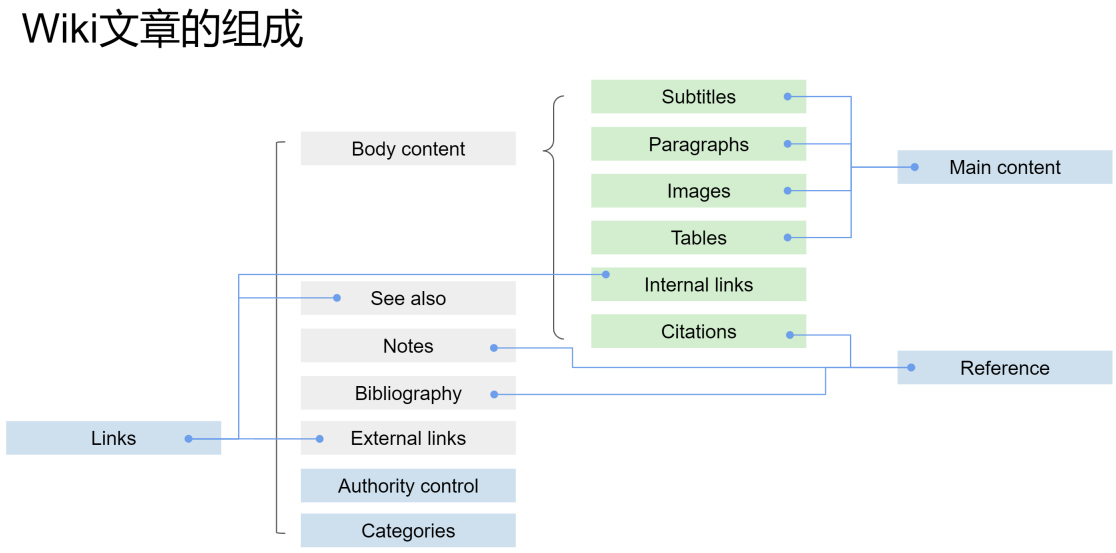

图 2: 维基词条结构

在对词条内容分类的基础上，我们对编辑行为进行了分类。如3所示，我们认为对不同部分的编辑在类别上是不同的。对主要内容的编辑影响着词条呈现的内容，可细分为信息更新（一般是对表格类数据、数字、时间等的更新，较为客观）、内容修改（指对没有“标准答案”的内容的修改，往往掺杂编辑者的主观看法，也容易引起编辑战）、小修改（修改语法、标点等，并不影响内容含义）。其中，内容修改按形式又可细分为增加、删除、移动、句内修改；按动机可分为贡献、恶搞和特殊目的编辑。内容是否有资料来源往往反映其是否客观可靠，因此引用也常常与编辑战相关。我们将引用的修改分为质疑和修正两种类型，当编辑者质疑另一编辑者增加的内容时，往往会在该内容后添加 or（原创研究）、cn（需要引用）、when（时间不明确）等注释，确有依据的编辑者看到后会进行修正，而主观性编辑的编辑者则忽略或删除质疑。链接、权限控制、分类这三类则没有进一步细分。

由于目前算法的限制，在实际实现中，我们没有特别区分修改内容所属的部分，只是将修改分为增加、删除、移动三类。

通过这些属性，我们可以探索词条的演变情况、检测编辑战等编辑事件。我们可视化的任务正是展现词条的整体编辑情况以及编辑过程中的事件，以帮助用户和编辑者了解词条的编辑过程，从而了解词条的争议情况、是否可信任，也帮助维基百科的管理者识别出恶意编辑的编辑者。

3 可视化设计

我们的界面设计如图4所示。左上角是词条信息的简介，左边剩余部分是主视图，即编辑的时间分布视图，右边是版本对比视图。信息简介给出了词条名称、词条简介、编辑次数、参与人数、编辑频度等信息。编辑的时间分布图编码了编辑的时间、增减的字节数、编辑者贡献、撤销行为、编辑的相关性等信息，

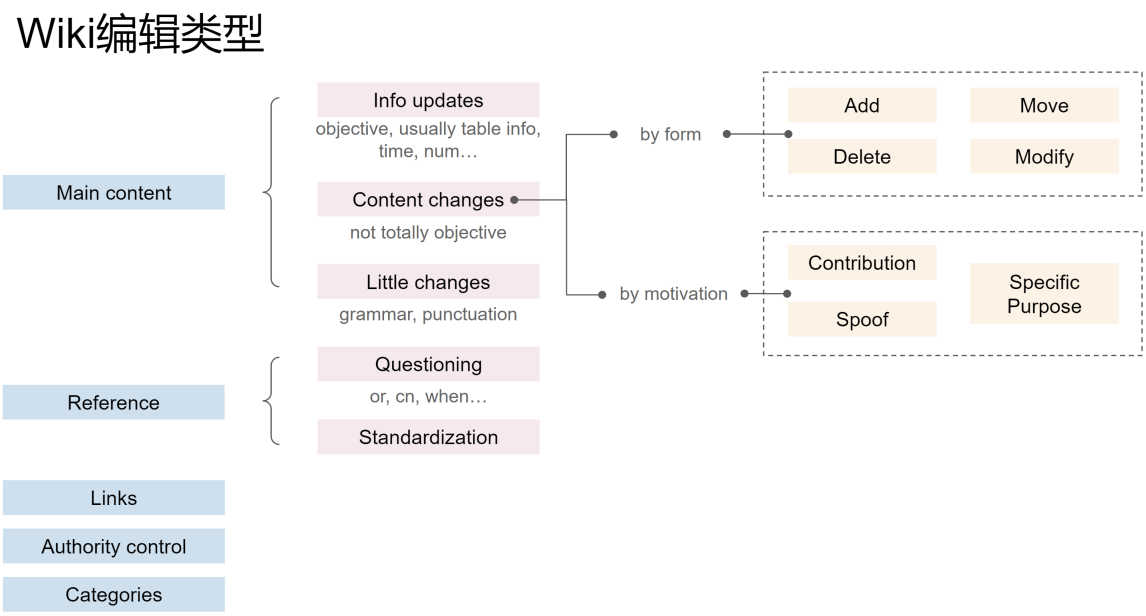

图 3: 维基编辑类型

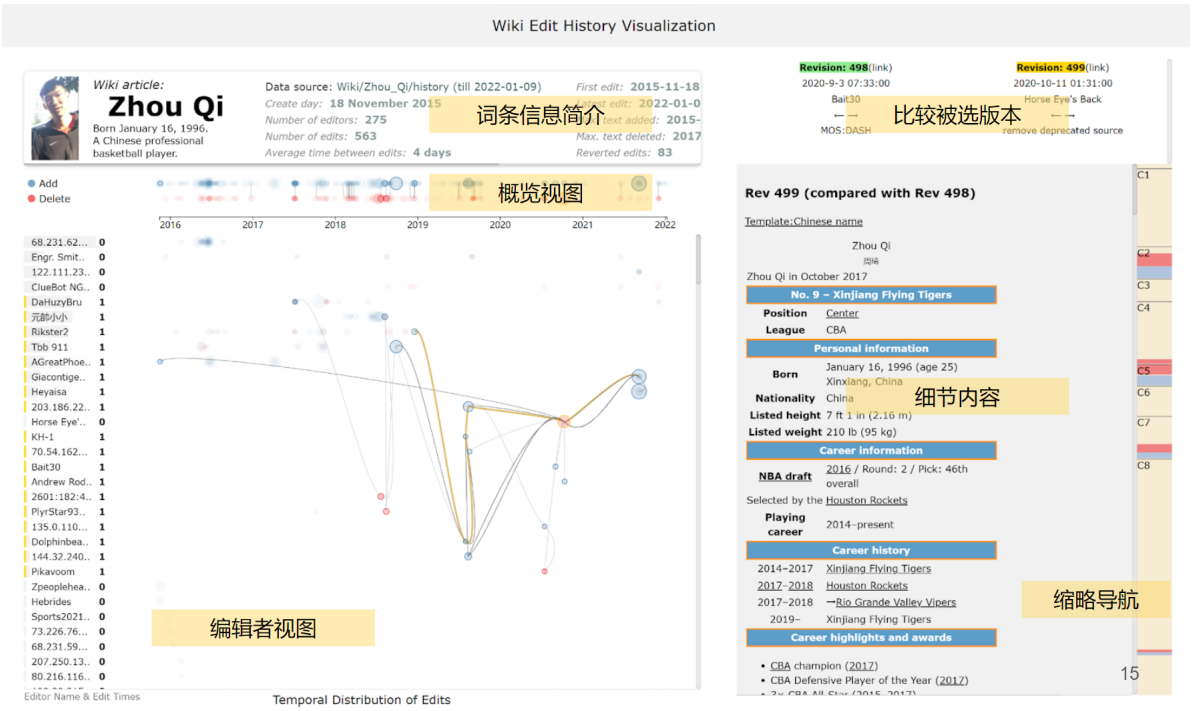

图 4: 可视化设计

用户可以通过该视图发现一些异常编辑行为。版本比较视图展示了两个版本内容的差异，用户可以通过选择上一个/下一个版本自由比较感兴趣的版本。左右视图的交互可以让用户在发现异常编辑行为后，定位到异常编辑所涉及的句子，并通过右侧视图呈现的具体修改内容获知异常的具体内容。

4 系统实现

在本次课程设计的实现中，我们选取了英文词条“Zhou Qi（周琦）”展示。

4.1 数据处理

我们首先爬取了当前所选词条各个版本的网页内容，作为原始数据。

因为我们关注的是每次编辑后词条发生的变化，考虑比较相邻版本提取变化的内容。首先利用 Python 中的 Beautiful Soup 库解析 html 文件，提取出每个页面上的所有标签。以标签为最小划分单位，我们将每个版本的内容划分成由句子组成的列表。这个列表将被用于进行版本比较。

4.2 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。第一栏是对词条的介绍，包括周琦的照片、周琦的简介，可以看到，周琦出生于 1996 年，是一位篮球运动员。词条名“Zhou Qi”链接到相应的维基页面。第二栏列出了数据来源、词条创建时间、编辑者数量、编辑次数和平均编辑次数。“Zhou Qi”这一词条创建于 2015 年，截至 2022 年 1 月 9 日，共有 275 位编辑者参与过编辑，总编辑次数为 563 次，平均 4 天编辑一次。第三栏展示了最早编辑、最晚编辑、最大编辑、最小编辑以及被撤销的编辑数。

4.3 编辑的时间分布

编辑的时间分布视图由概览视图、和编辑者视图组成，两者共用一个时间轴。

在两个子视图中，都用透明圆点表示一次编辑。圆点的圆心横坐标映射编辑完成的时间，圆点的面积映射编辑所修改的字节数（绝对值），蓝色表示字节数增加，红色表示字节数减少。所有圆点的透明度一样，视图中颜色较深的部分是多次编辑重合而成。

如图5，在概览视图中，我们将所有编辑分为了两行，第一行是编辑后页面总字节数不减的，第二行是编辑后页面总字节数减少的。我们通过编辑总结属性提取了撤销行为，共提取到 46 次，并在概览视图中用直线连接了两个有撤销关系的版本。

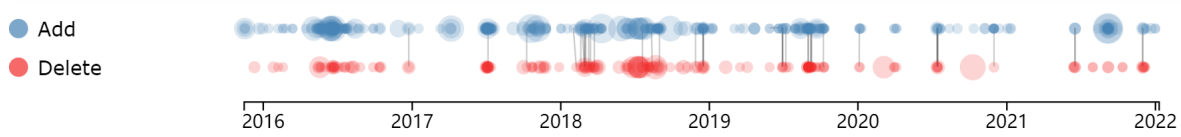

图 5: 编辑时间分布的概览视图

在图6所示的编辑者视图中，我们以每个编辑者为一行。每行最左边标注了编辑者 ID（因为有些编辑者 ID 过长，所以将超出 10 字符的部分省略），编辑者 ID 的背景矩形宽度映射了该编辑者编辑的总次数，在矩形右边标出了编辑次数的数值。行中的圆点分布反映了相应编辑者参与编辑的时间分布情况。

在数据处理时，我们为所有出现在历史版本中的句子生成了一个字典，并记录了每个句子的修改信息（被哪些版本添加、被哪些版本删除）。这样，每个句子，可以对应到一个与之相关的版本序列。对于每个

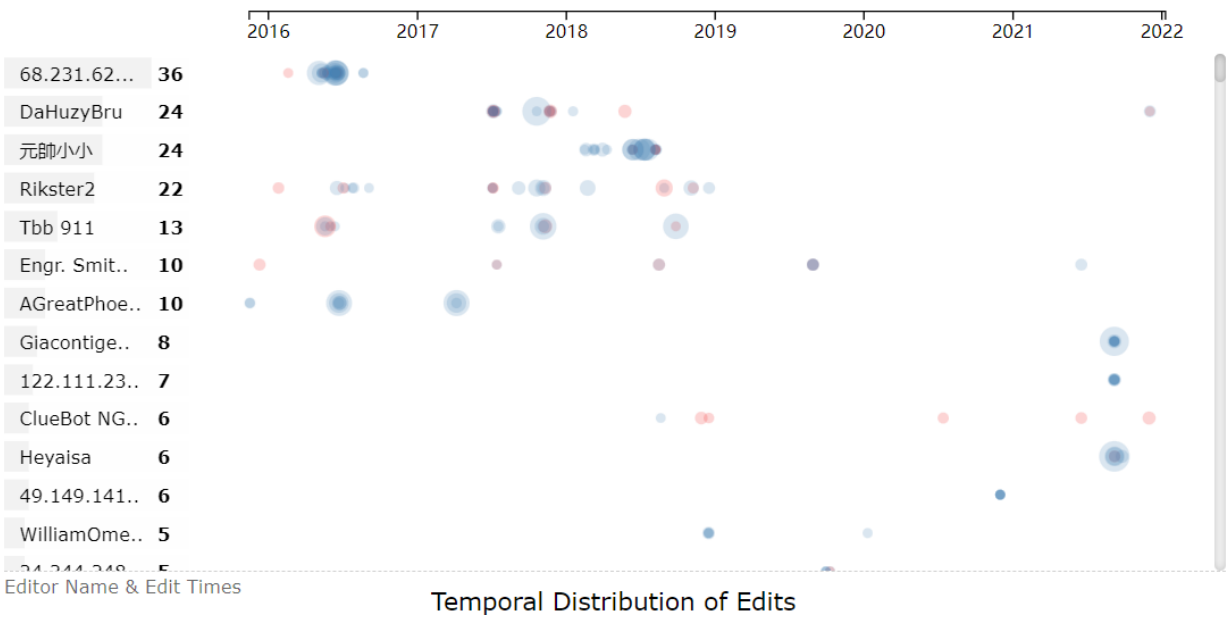

图 6: 编辑时间分布的编辑者视图

版本，我们记录了该版本和前一版本的差异，即该版本所修改的句子。我们将每个句子对应的相关版本序列高亮并用曲线连接，并在悬停或点击时展示该曲线。

图7展示了悬停圆点的交互效果。上下两视图中，悬停的版本和相关的版本都会高亮显示。悬停的点边框为金色，其他相关点边框颜色加深。悬停时，前文所述的曲线出现。此外，我们绘制了一个提示框，展示该版本的简要信息，包括编辑时间，版本编号（括号内为维基页面 ID），编辑者 ID，编辑规模和增、删、移动的句子数。概览视图和编辑者视图的悬停行为是一致的。

图8展示了点击圆点的交互效果。在悬停时视图变化的基础上，点击后，与被选版本相关的版本所涉及的编辑者会向被选版本聚拢，且相关性越高的距离越小。在左侧的灰色矩形上绘制了金色的前景矩形，其宽度映射该行编辑者与所选编辑者的相关版本的数量。概览视图和编辑者视图的点击行为也是一致的。

在概览视图中，编码撤销行为的连线也支持悬停和点击交互，悬停时撤销和被撤销版本都会高亮，点击连线则相当于点击做出撤销的版本。（此处有待改进，点击连线应同时显示出撤销版本与被撤销版本）

在编辑者视图中，我们增加了拖拽交互，支持矩形框选择若干版本。如图9所示，拖拽选择的效果相当于单个版本点击效果的叠加。

在原始时间轴尺度下，编辑点重合严重，密集处难以区分，因此我们设计了时间轴放大交互。如图10所示，在时间轴上拖拽感兴趣的时间范围后，时间轴会随之放大，原本重合的点得以分开。点击空白处，时间轴复原。

4.4 版本比较

版本比较的视图呈现被比较两个版本的信息以及版本整合内容。其中版本信息包括：版本编号、版本编辑时间、编辑者、编辑评论等。在版本整合中，我们通过匹配算法识别出两个版本间的编辑内容，包括：删减、增加与交换内容，并利用基于动态规划的最小编辑距离算法细粒度定位编辑内容，最后将两版本实现整合。

算法的细节呈现中，我们对所有版本中出现的字段进行标号，对于版本的存储由大量字符转为相应标号，节省了大量重复、冗余信息的存储空间。同时我们依据标号匹配实现了轻量级、在线的匹配算法，使得我们的可视化设计能够在线计算与呈现任意两版本的整合、编辑信息，以此节省数据存储空间。

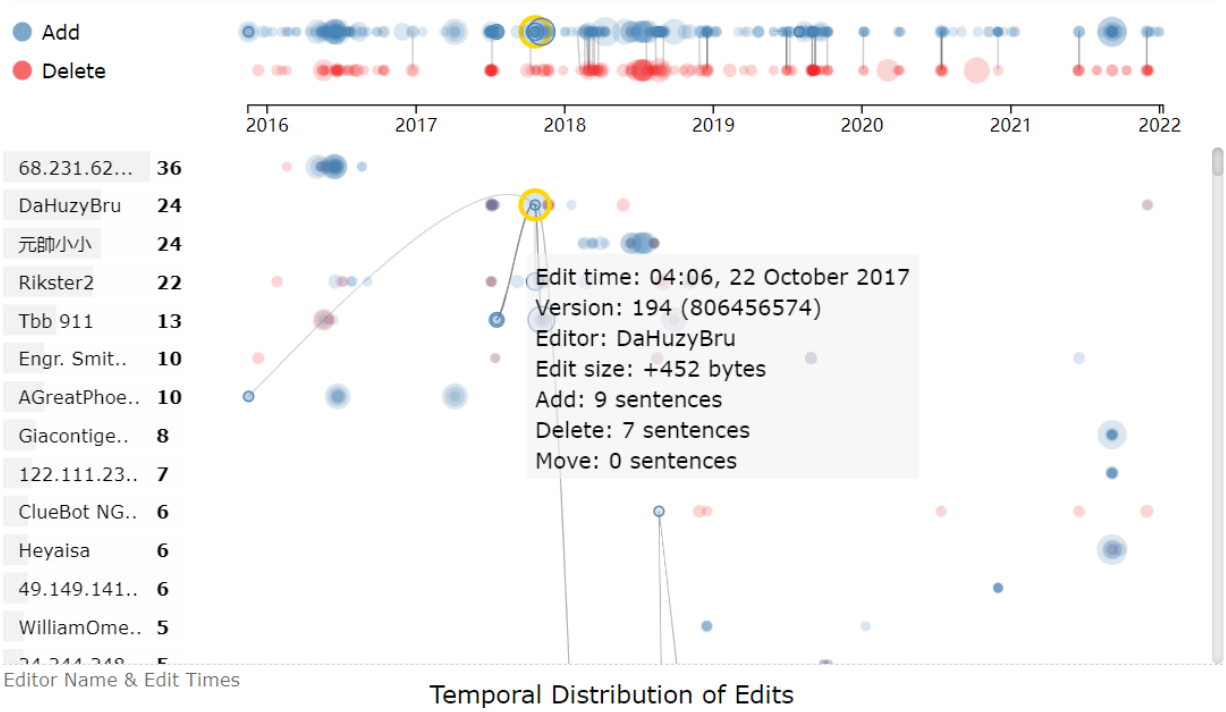

图 7: 编辑时间分布视图的悬停交互

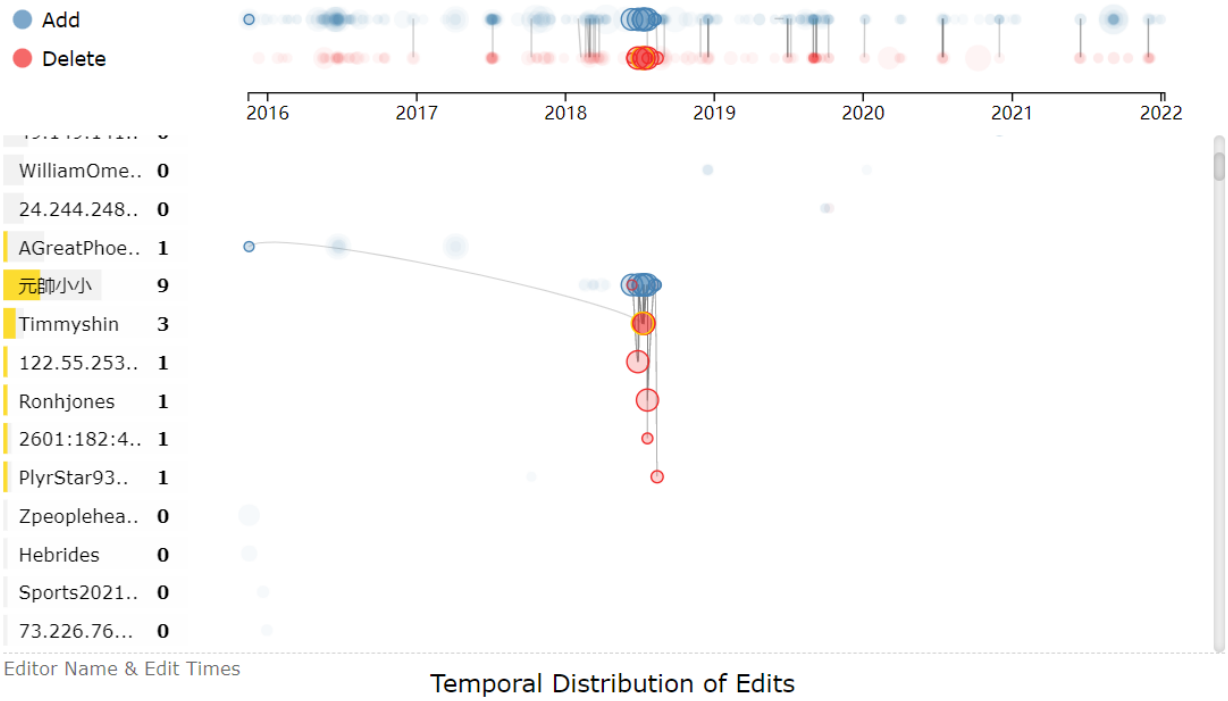

图 8: 编辑时间分布视图的点击交互

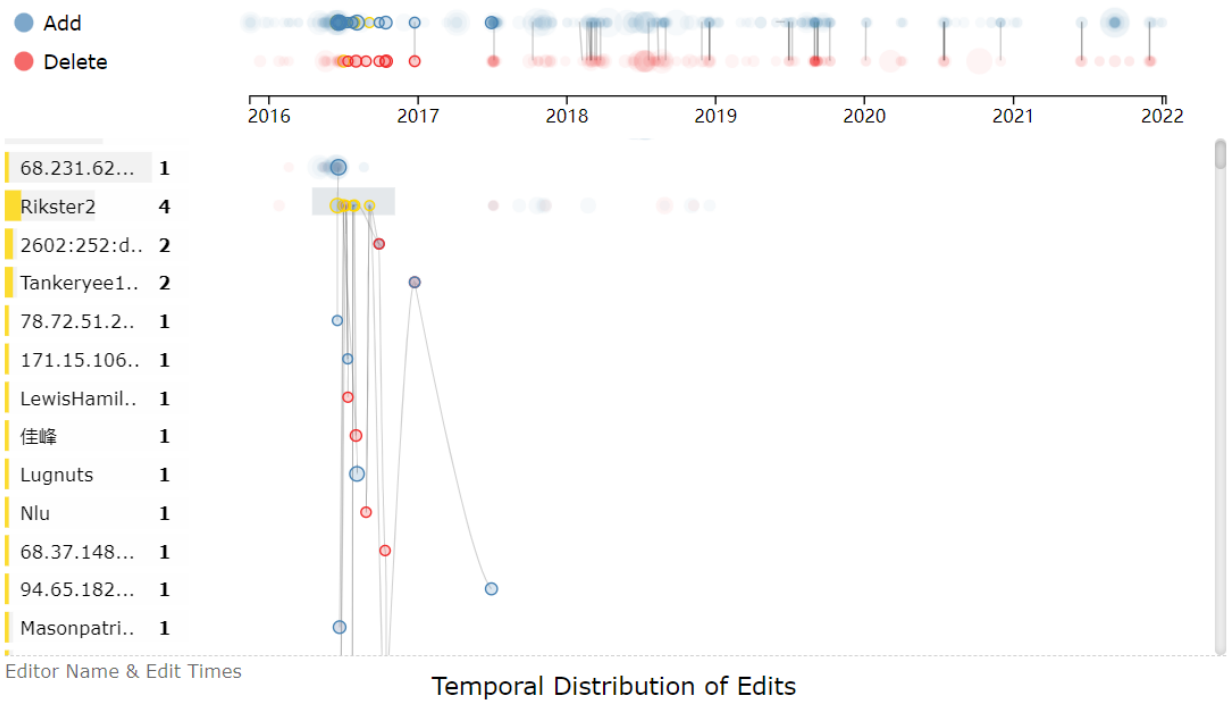

图 9: 编辑时间分布视图的拖拽交互

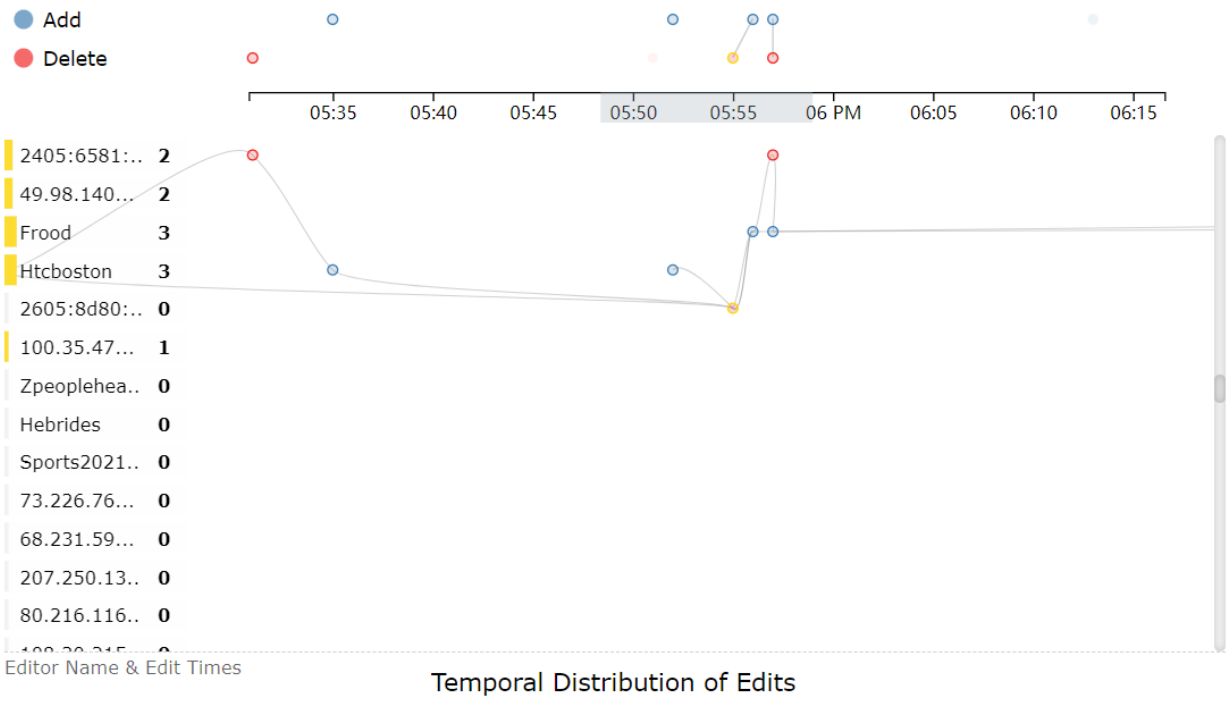

图 10: 编辑时间分布视图的时间轴放大

在版本信息介绍中，我们还为用户提供选择上/下一版本作为比较版本的交互功能。在细节内容板块，我们将编辑内容与编辑者视图中的编辑字段信息进行交互，当光标移动到编辑字段时，编辑者视图中相应的字段连线会高亮，以此定位所有修改过改内容的编辑者。同时，我们也提供了整合版本的缩略导航的功能，使得用户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，并在导航中呈现整合文章的章节信息，方便用户快速定位某一感兴趣的章节。

4.5 联动交互

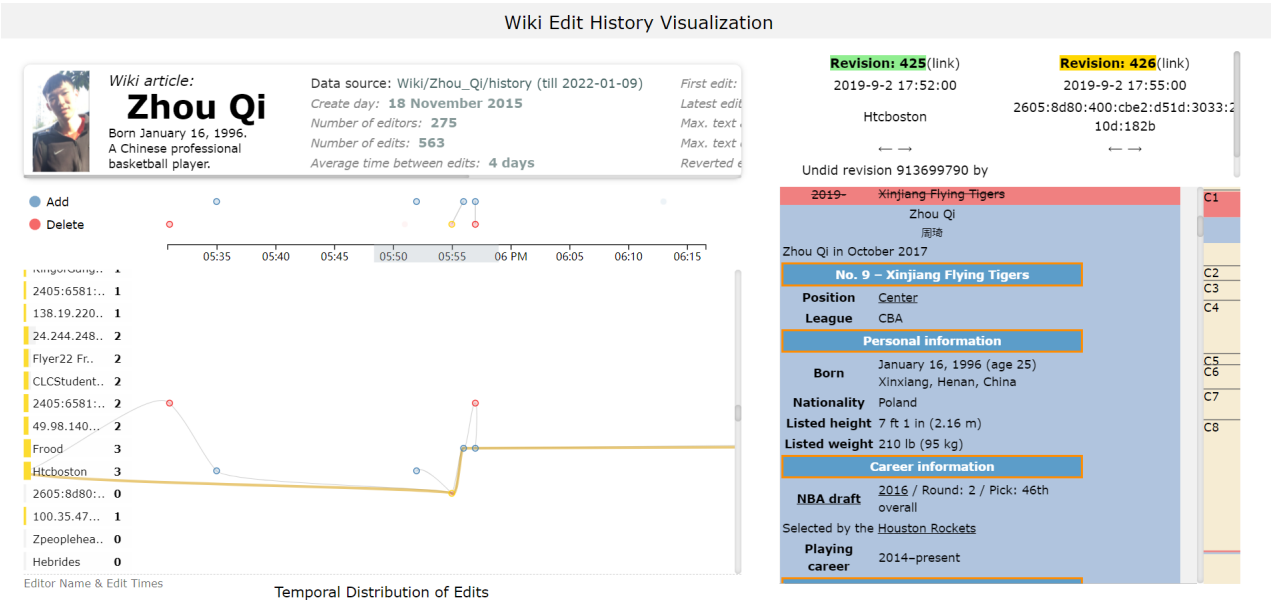

图 11: 左右视图的联动交互

如图11所示，点击某版本后，右侧版本比较视图所比较的版本切换为被选版本与其前一版本。如编辑者视图部分所述，视图中的每条曲线对应一个被修改的句子。鼠标悬停在曲线上时，该曲线高亮，右侧比较视图自动滚动到曲线对应的句子处。鼠标悬停在右侧被修改的句子上时，左侧对应的曲线也会高亮。此外，在右侧视图切换比较的主版本时，左侧视图的被选点也会切换为该版本。

5 发现结果

通过我们的视图以及交互，可以发现词条“Zhou Qi”编辑过程中的许多事件。

5.1 关于 Rumor 的编辑战

在查看编辑者视图时，我们发现编辑者 Timmyshin 的编辑形成了一个巨大的深红色圆点（见图12），也即在 2018 年中的一小段时间内，编辑者 Timmyshin 进行了连续多次大幅删减。

点击其中的一个红点，我们得到了与 Timmyshin 的删减行为相关的编辑。如图13，可以看到，Timmyshin 的删除与元帅小小的增加行为密切相关，元帅小小与此次编辑的相关版本数量达到了 9 个之多。鼠标移至曲线上方，我们发现 Timmyshin 删去了由元帅小小添加的 Rumor 章节，增加了最初版本中带有的 Statistics 章。参与删除的不止 Timmyshin 一人，还有另外四个编辑者。不难推测，在这段时间里，元帅小小反复添加谣言，但随即被 Timmyshin 等编辑者删除。进一步查看这些相关的编辑，发现确实如此。

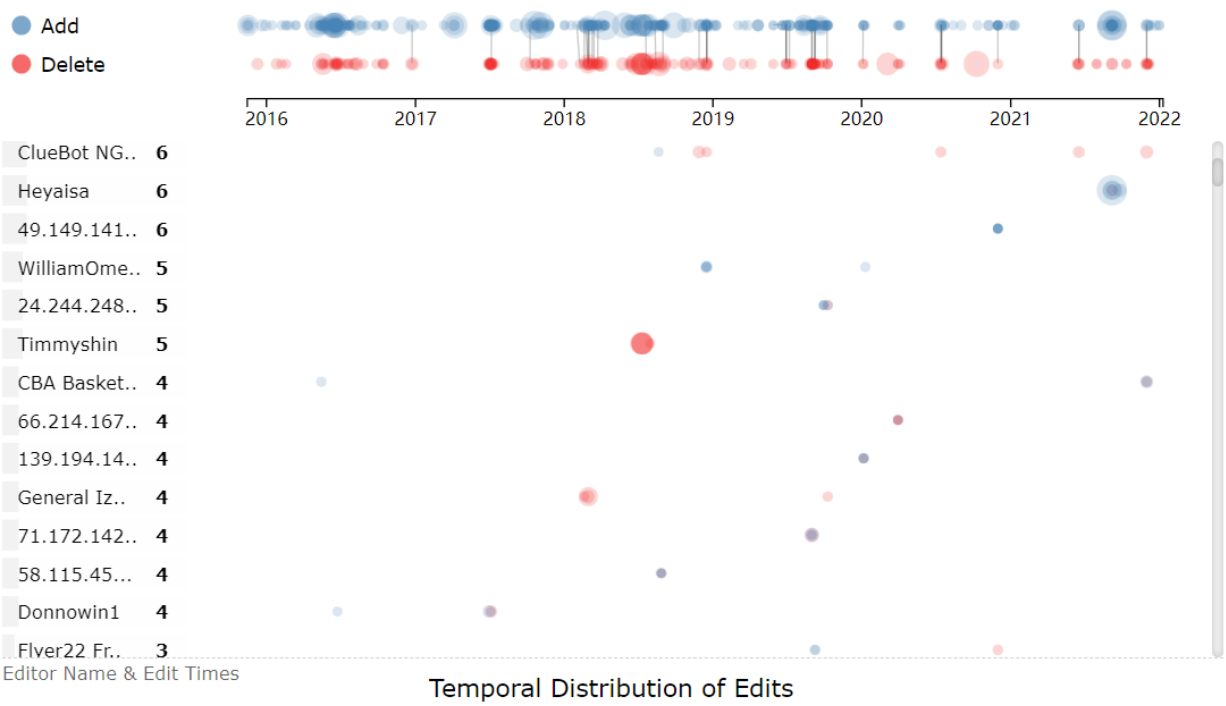

图 12: Timmyshin 的连续大幅删减行为

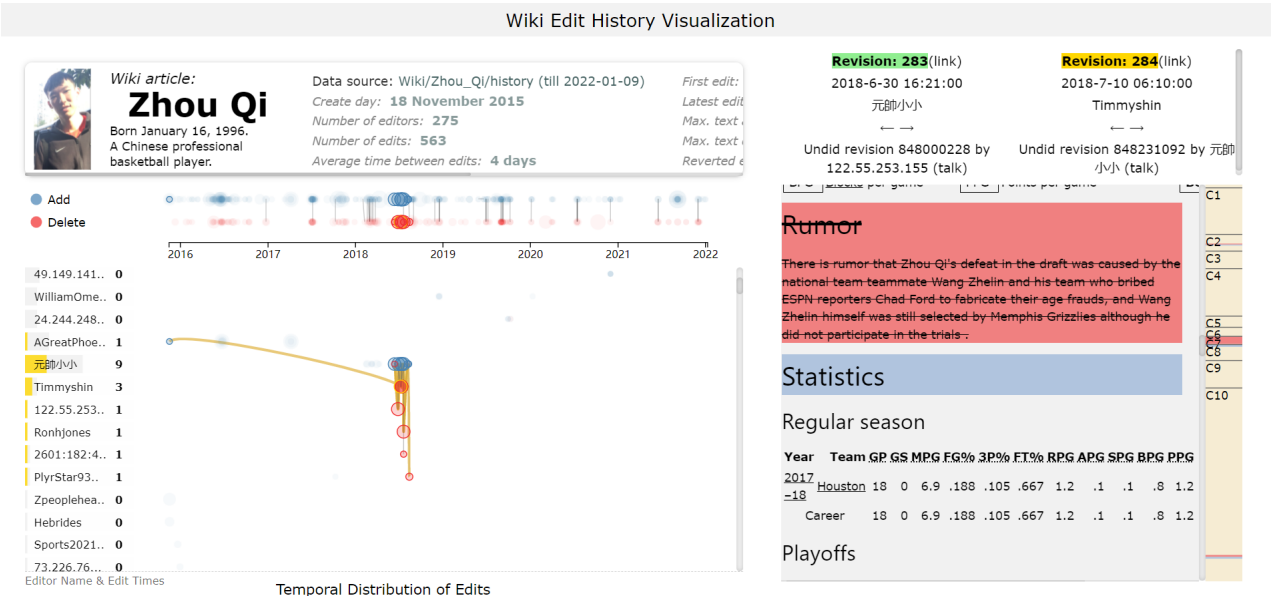

图 13: 关于 Rumor 的编辑战

5.2 关于周琦国籍的编辑战

在概览视图中，我们看到 2019 年下旬发生了多次撤销行为，表明可能发生了编辑战。但这些编辑修改的规模都不大。如图14，点击其中一个版本查看后，发现短时间内这一词条涌入了大量的新用户参与编辑工作。他们的 ID 大多是匿名 ip 地址，而且多数只在这段时间内参加过一两次编辑。鼠标移到曲线上，我们发现这些编辑都与周琦的国籍有关。周琦的国籍被反复修改成 “Poland”，又被改回 “China”。对篮球赛事较为熟悉的组员指出，这次编辑战正发生在周琦在一次重要的比赛中发球失误之后，许多情绪激动的球迷因此在维基百科中将周琦的信息做了许多恶意的篡改。

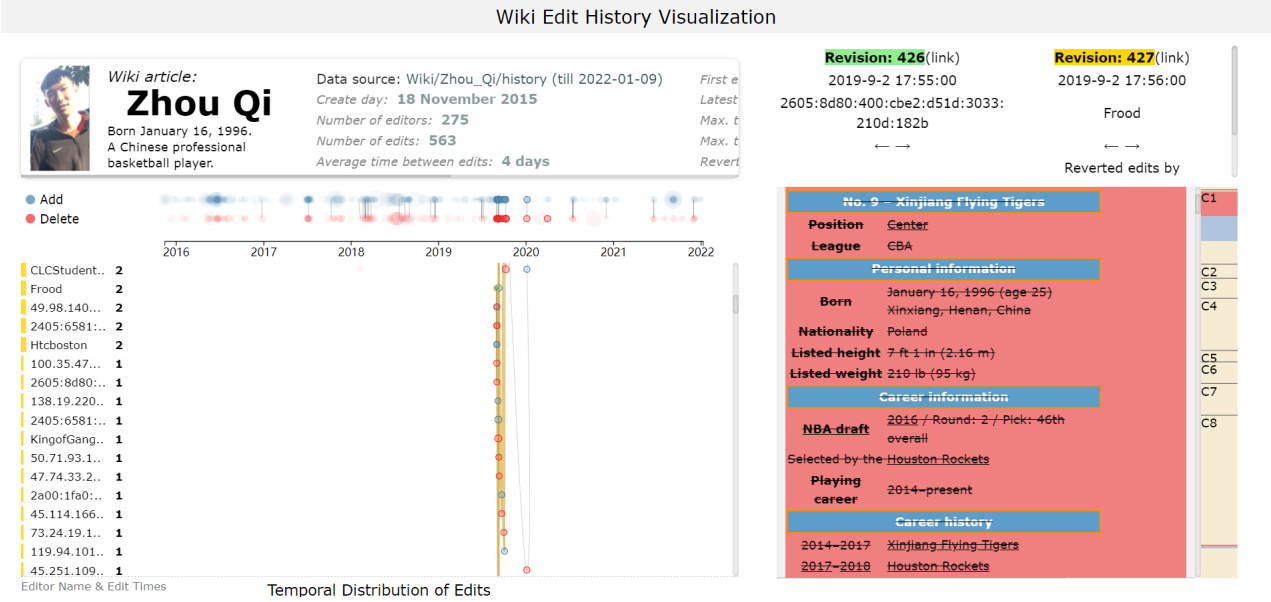

图 14: 关于周琦国籍的编辑战

6 讨论与展望

关于维基百科的编辑历史，已经有许多可视化方法。而我们的设计与已有的方法不同，以展现编辑历史中的事件为任务，支持细节性的分析，帮助用户还原在编辑历史中发生的各种事件。

在已有的视图中，我们可以通过选择撤销连线或是异常的编辑点，或许该次编辑的具体改变内容与相关的编辑版本来发现事件的细节。在未来的工作中，我们希望能更直接地展现出编辑中的事件：

- 事件检测：通过编辑在时间上的异常分布、撤销行为等自动计算出可能的事件；
- 事件图标设计：为检测出的事件设计编码方式；
- 设计条件：更清楚地总结出设计所要满足的需求与条件；
- 文本缩略图设计：在文本缩略图上编码更多版本的更多信息。

7 分工

- 娄宇珂：数据处理、匹配算法及报告相关部分，demo 视频制作
- 王益明：数据处理、匹配算法及报告相关部分，demo 视频制作

- 韩勤：版本匹配视图实现及报告相关部分，海报制作
- 郭宇涵：编辑时间分布视图实现及报告其余部分

数据、任务分析和可视化设计由组员共同完成。