

# 第一章 维基百科用户分析

## 1.1 维基百科简介

维基百科 (Wikipedia), 是一个语言、内容开放的网络百科全书计划。英文的“Wikipedia”是“wiki”(一种可供协作的网络技术)和“encyclopedia”(百科全书)结合而成的混成词。

维基百科由来自全世界的自愿者协同写作。自 2001 年 1 月 15 日英文维基百科成立以来, 维基百科不断的快速成长, 已经成为最大的资料来源网站之一, 而以热门度来说, 则为世界第六大的网站, 在 2008 年吸引了超过 684,000,000 的访客, 目前在 272 种的独立语言版本中, 共有 6 万名以上的使用者贡献了超过 1000 万篇条目。截至今日, 共有 314,766 篇条目以中文撰写; 每天有数十万的访客作出数十万次的编辑, 并建立数千篇新条目以让维基百科的内容变得更完整。(请参见维基百科统计)

维基百科一直坚持内容的开放性。这时维基百科取得成功的一个重要原因。维基百科所有的内容, 包括文字、图片等均在“知识共享署名-相同方式共享 3.0 协议”之条款下提供。任何人都可以自由引用维基百科的全部或者部分内容, 仅需要注明其出处即可。维基百科不仅赋予内容使用的开放性, 还奉行参与的开放性。这种开放性同自由软件运动有很大的相似之处。Raymond 将这种开放、自由的参与和使用方式比喻为“集市”似的方式, 区别于传统的那种严谨、刻板、集中的“大教堂”方式。任何人只要愿意遵守维基百科社区的政策, 就可以加入到协同创作过程中来<sup>[2]</sup>。

维基百科的成功证明了“群体智慧”的力量。Surowiecki 指出利用群体的智慧开展协同工作对于当今的政治、经济、商业等各个方面具有重要的影响。日益频繁的人际交互已经超出了时空的阻隔, 正在改变人们的生活方式。维基百科的成功不仅仅是信息技术创新的结果, 更是人们相互协作, 共同应对面临的各种挑战, 利用集体的力量取得成功的最好体现<sup>[2]</sup>。

协同创作是在互联网环境下, 创作者借助信息技术超越时间和空间的限制, 共同合作完成特定主体的内容创作。随着知识分工的日益细化, 仅凭个人所掌握的知识越来越难以适应不断提升的知识创新需求。为了弥补自身知识的不足, 以协同方式共同

完成新知识的产生成了越来越多的人的必然选择。然而，新的知识创新方式尽管能积极应对知识创新所带来的压力和挑战，也随之产生了新的问题。Liccardi 等人将这些问题归纳为五个方面<sup>[2]</sup>：

1. 沟通不足。知识协同是促进知识创新的手段，而持续有效的沟通是知识协同的基础。协同平台不但要保证协同成员间的沟通顺畅，更重要的是能够追溯思想、灵感的产生过程，确保这些内容能够完整真实地记录下来。
2. 内容与讨论脱节。知识创新是一个持续、反复的过程中，需要存于者进行大量的讨论、评价，对有疑问、不明确的地方进行辨析、扬弃，最终达成一致。讨论是最终内容形成的完整链条，串起了内容各个版本间的变化过程，对于理解最终内容非常重要。一旦导论于内容间的关系没有得到紧密关联，那么协同过程中必然会出现大量的误解和矛盾。
3. 对群组讨论缺乏足够的支持。对内容的讨论常常是以群组讨论的方式开展的。当参与协同的成员众多时，会出现大量参与者同时参与多个讨论主体的情况，而讨论本身会进一步衍生出其他的讨论。这就要求协同平台能够支持有效的群组讨论，维系特定的讨论线索，并有效解决讨论过程中出现的各类问题。
4. 缺乏旧有版本的回溯功能。协同参与者需要不时审视以前的内容，并做出决定是否之前某一版本更好，进而应该恢复到那一版本。如果缺乏回溯功能，这些需求将很难得到满足。
5. 解决冲突。当多个协同者对内容进行编辑时，必然会产生冲突。冲突的原因是由于协同者“同时”对某一段文字进行了修改，一般来说这类冲突很难通过自动化的手段解决。好的协同平台除了要及时给出冲突的提示外，还应该根据相应的规则和手段协助协同成员解决冲突。

一个好的协同平台，应该能有效地处理好以上问题，为协同者提供强大的技术支持和保障 Neuwirth 等人归纳了一个优秀的协同平台所应该具有的特征：

1. 提供适合的方法和手段促进创作者在内容创作过程中开展有效的交互。
2. 将内容创作和对内容的讨论，意见等内容有机地结合起来。
3. 提供有效的工具支持协同创作和内容讨论这两种形式的交互。

这几个特征实际上揭示了协同创新的本质。新的内容（或者知识）来源于个体、团队、组织等不同层次群体之间的交互和沟通，在思维的碰撞中产生。协同创作的本意在于激发参与者的沟通欲望，调动参与者的协作热情，共同完成某一主题的知识或者内容的创新。在创新过程中，共享各自的知识已经不再是主要目的，因此，支持协同成员间的社会化交互，提供高效、简洁、易用的交流平台就成为 wiki 平台成功的关键因素。

## 1.2 维基百科社区中的知识协同

维基百科的协同活动主要是协同编辑，是由一群人一起，而非单独一人完成的写作工作计划。协同的目标是针对某一主题给出其“百科式”的解释，不仅包括该主题自身的含义，还可能包括其历史背景和演变过程，其他人的评价，对其他方面的影响等内容。协同的最终结果是一个个具体的条目。维基百科用户可以参与到绝大部分条目的协同协作中。条目被分为不同的种类，维基百科成不同的种类为命名空间。维基百科目前有 20 个命名空间，其中包括 9 个基本的名字空间；此外还有两个虚拟名字空间。表 1 给出了维基百科中基本命名空间的清单：

表 1: 维基百科的命名空间

编号	命名空间	内容
1	条目	条目命名空间” 又称“主命名空间”，包含了维基百科上的所有条目页面，或“百科全书文章”。
2	维基计划	这个命名空间提供了有关维基百科的内容信息，包括维基百科自身的信息、方针、指引、论述，以及维基人的讨论空间“互助客栈”、询问处等。
3	帮助	包含了所有维基百科及 MediaWiki 软件的使用指南信息。有些内容帮助读者更好地使用维基百科，而另一些内容则为编者准备，用来更好地编写维基百科。有些信息亦是来自元维基和 MediaWiki 网站上复制而来的。
4	用户	包含了所有用户的个人页面，以及其个人创建的相关页面。
5	分类	包含了所有的分类页面，内容为该分类之下的页面和子分类列表，以及可选的分类提示信息。

6	文件	包含了图像和声音的文件描述页，以及指向文件本身的链接。
7	MediaWiki	包含了所有的软件界面文字，例如在一些页面上自动生成的信息和链接。这个名字空间用于定制和翻译 MediaWiki 的软件界面。
8	模板	包含了所有的模板。模板是一类特殊的页面，用于嵌入或替换引用进其他的页面，以加入一些标准化的内容，或者信息栏、导航栏等。
9	专题	包含了所有的主题页面。一个主题页面是关于某一方面内容的信息集合，一个相关条目的入口。

命名空间对维基百科中所有创建的内容用途角度进行了划分。尽管几乎所有的内容都是社区成员的协同结果，但是这并不意味着这些内容都会纳入到本文的研究范围。事实上，协同的主要成果是各个条目页面，而其他几个命名空间的内容均是为了更好地编写条目而提供辅助功能的。因此，本文将维基百科社区内的知识协同活动限制为共同编写某一条目内容，而忽略其他内容的协同编写活动。这样的目的是：一方面保留了协同活动的主体，同时有减少了数据分析和处理的难度，突出了研究的重点。协同创作条目还可以分为两个部分，条目的编写和讨论。显然，条目的内容本身是协同的直接结果，条目的编写是直接的协同活动，参与条目创作的用户是协同的直接参与者；而条目的讨论是协同过程的间接结果，讨论本身不是协同的目的而是必要手段，那么参与讨论的用户是否也应该作为协同活动的参与者？

讨论页是特殊的维基百科页面，它包含了所有对主题文章的讨论。任何的问题、疑虑、怀疑、参考文献、有关文章的论战或者评论都可以在相关的讨论页提出来。在讨论页中，协同者可以分享自己的思想和观点，整理内容创作的思路和逻辑，分析内容的取舍，澄清材料的真伪，最大限度地保障协同的质量。维基百科的讨论的目的在于解决协同创作过程中所遇到的一些问题，主要的方式是头脑风暴，汇集各方的思路和意见；而条目的编写在于组织各类材料和内容，完成实际的内容创作，强调“做”而不是“说”。同讨论过程的松散性不同，内容编写是非常正式、严谨的。Viégas 等认为，用户参与讨论本质上是一种合作行为而非协同行为。讨论的内容可以分为以下几个部分：

1. 征集合作者。征集着意识到条目内容本身还有不完善的地方，但是个人又无力完

成，因此号召具有相关知识的人补充完善条目内容。

2. 寻找信息。一些用户试图从条目中寻找相关信息，但是条目内容并未涉及这些信息，因此希望有人能够提供这些信息。
3. 讨论本页面的恶意篡改行为。例如是否暂停页面编辑，或者是否取消对页面的保护等内容。
4. 讨论条目内容是否符合维基百科的编写指南和相应的政策。维基百科条目的编写有许多规定和准则，如果怀疑某一部分内容同这些规定和准则相违背，那么参与者会在讨论页提出自己的质疑。
5. 引用其它维基资源。通过引用其他的条目或者讨论内容来解释自己的编辑行为。
6. 与讨论主题无关的内容。用户在讨论页发布广告、交友等垃圾或恶意信息。
7. 投票。当某一部分内容存在争议，且没有压倒性的证据支持某一方的论点，用户通过投票方式来解决争端。
8. 征集内容审阅着。条目编写者期望提高内容质量，征集熟悉相关领域的用户对内容进行审阅，提出具体意见。
9. 其他讨论。

讨论所涉及的内容是协同创作的辅助手段，是提升知识协同水平，达到预期协同成果的保障。但是讨论本身并不直接产生协同的结果，而仅仅是协同活动的必要补充。因此，本研究所讨论的知识协同活动并不包括讨论的内容。在本文中，将知识协同定义为：用户参与的维基百科条目的协同创作活动。

### 1.3 知识协同的参与者

维基百科的开放性决定了任何人都可以参与到协同创作中来。不论是在维基百科中注册的用户还是未注册的匿名用户，均可以为编写条目贡献自己的力量。维基百科社区将用户分为不同的角色，每种角色有各自的权限。用户的角色还可以进一步划分为普通用户和管理用户。表 3 列出了普通用户的角色和权限。

表 2: 协同与合作

	协调	合作	协同
必要条件	有共同目标；多人参与；知道何时由谁做什么。	有共同目标；多人参与；相互信任与尊重；承认合作是双赢（多赢）。	有共同目标；多人参与；积极的投入；对协同群体有归属感；开放的沟通和交互；相互信任与尊重；互补的知识与技能。
主要目的	避免工作的重叠或者缺失。	在合作过程中各自获得利益。	通过集体的努力完成个体无法独立完成的工作，并取得合作成果。
预期成果	令人满意的工作成果。	取得工作成果的同时还节省了时间和投入。	除了合作产生的结果，还取得了创新性的成果以及完成工作的成就感。
适用范围	应对简单、独立性高的任务，成员角色和进度安排非常明确。	应用于在复杂环境下系统地解决问题。	在复杂环境下，需要彼此理解并认可对方，建立一致的价值观，通力协作解决问题。

表 3: 维基百科用户分类

用户类型	定义	权限
匿名用户	未在维基百科网站注册账户的用户。	浏览所有页面；编辑所有未经保护的页面；在任意命名空间下创建讨论页面。

新注册用户	已经在维基百科网站注册了账户，但是还未确认其电子邮件地址。	创建新页面；给其他已经确认邮件地址的用户发送电子邮件；将某次页面编辑标注为细微改动；删除页面无须确认；定制维基百科界面和账户信息。
自动确认用户	已经在维基百科网站注册了账户，并且已经确认其电子邮件地址；用户状态由系统自动确认：用户注册 4 天以上并且进行过 10 次编辑即成为确认用户。	移动页面；对部分保护页面进行编辑；上传新文件或者上传已存在文件的新版本。

从表中可以看出，协同创作的参与者必然属于某个普通用户角色。即使是权限最低的匿名用户，也可以参与到已创建条目的编写过程中去。而一旦在维基百科进行注册，则可以获得更高级的编辑功能。维基百科的内容主要是有不同用户创建的。

管理用户主要参与到社区成员和内容的管理工作。比如用户权限的授予和回收，管理内容的创建和编辑工作，对争议和冲突进行调解和仲裁，执行特定的任务，以及开发各类辅助工具方便用户使用维基百科，协助管理者和学术研究人员进行数据统计和分析。维基百科中的管理用户主要包括：系统管理员；系统行政员；系统监督员；回退员；IP 封禁例外者；账户创建员；上传者；机器人；程序开发人员等。

管理用户不直接参与到条目的协同创作中，但是管理用户仍然会给条目的内容带来一定的变动，这些变动包括：

- 管理员删除条目。由于条目本身未能达到维基百科自身的要求，或者是条目内容被其它条目所取代或覆盖，则管理员将删除该条目。
- 机器人编辑条目内容。机器人是由用户开发的自动化或者半自动化程序，参与各类内容创建与编辑的辅助性工作，主要用于自动处理繁琐的格式或数据。机器人按照预定的目标和规则对页面内容进行重新编辑，比如调整内容的结构，增加条目间的链接等。

管理用户对条目的变更的主要目的是：保障条目的一致性，清除冗余条目，促进条目

内容更加符合维基百科的编写规范和格式要求。管理用户本身并不创建新的内容，也并不提升已有内容的编写质量。因此，管理用户不视为协同的参与者，其对于条目内容的变更也不视为知识协同活动。在本研究中，协同的参与者仅限于普通用户。

### 1.4 协同贡献度量

#### 1.4.1 已有成果的回顾

在上一部分中，本研究明确定义了维基百科社区的知识协同：维基百科普通用户共同参与条目内容的编写。参与同一条目编写的用户可能会多达数百人，尽管每个人都为内容创建贡献了自己的时间、精力和知识，但是每个人对于条目内容的贡献程度却是各不相同的。为了反应社区成员参与知识协同活动的积极程度和共贡献大小，需要对用户的协同行为进行度量。度量协同行为对于研究维基百科社区知识协同的模式、理解协同行为具有重要的意义。首先，度量用户的协同贡献可以更好地促进虚拟实践社区的发展。对于贡献程度很大的用户，社区可以通过表彰，激励等手段促进其更好地参与社区的发展中去，同时调整社区对用户的支持力度，最大限度发挥这些用户的能力，合理利用社区资源。其次，通过度量协同行为可以帮助社区发现协同平台，协同政策等方面的不足，促进社区对相关的问题进行改进，更好地支持用户的协同。在本研究中，度量协同活动可以揭示维基百科用户的协同模式，区分不同的用户群体，从而为分析用户参与维基百科社区知识协同的动机因素打下基础。

用户的协同贡献可以从多个方面进行度量。Kittur 和 Chi 认为用户的编辑次数 (Edit count) 可以用于衡量用户的贡献。一个用户参与的编辑次数越多，那么意味着他对此条内容的改进就越大，从而作出的贡献也就大于编辑次数少的用户。根据维基百科社区进行的一次针对英文维基百科的统计结果，维基百科的内容并非是由社区所有的用户持续不断地进行小规模改进，最终汇集成现在的内容规模。实际上，维基百科的大部分内容是由一小部分的用户完成的。统计结果表明，在所有针对条目的编辑中，超过 50% 的编辑是由 0.7% 的用户 (524 人) 完成的，而社区中最活跃的 2% 的用户贡献了总编辑次数的 74%。这也就意味着，维基百科的大部分用户仅仅是参与少量的内容修订，真正对协同创作做出主要贡献的仅仅是一小部分核心用户。正是这些核心用户的努力使得维基百科取得了巨大的成功。但是，有学者认为，编辑次数本身只反应了用户参与协同活动的活跃程度，而不是对贡献内容多少 (Text count) 的反应。Swartz 随机选取了一些条目，分别统计出编辑次数最多的 10 位用户，以及贡献内容最



多的 10 位用户，发现两组用户的差异极大：编辑次数最多的 10 位用户均进行了至少上千次的编辑，但是几乎没有人同时成为内容贡献最多的人；贡献内容最多的 10 位用户最多只进行了 25 次编辑，最少的甚至只进行了一次编辑，但是却贡献了绝大部分的内容<sup>[2]</sup>。基于编辑次数以及基于贡献内容这两种方式反映了不同研究着对贡献的不同理解，不同的度量方式所得到的结果也不尽相同。

尽管这两种类型的贡献度量得到了广泛研究，但是其缺点也非常明显。Adler 等批评这两种度量方式均是不稳健的<sup>[2]</sup>。不论是编辑次数还是内容贡献，二者都容易被恶意利用，从而导致最终结果产生偏差。如果根据编辑次数统计用户贡献，那么一个用户很容易将一次规模较大的修订分解成数个小修订，从而增加编辑次数；或者该用户可以进行错误的编辑，然后利用回退功能取消这次编辑，同样可以达到欺骗的目的。由于维基百科的条目众多，使得这种行为很难被发现。更重要的是，这种行为严重扰乱了条目的正常编辑过程，从而降低了条目质量和内容的稳定性，甚至打击其他协同用户的积极性。因此，使用编辑次数来衡量用户贡献在实际中效果并不好。类似地，如果以文字数量作为贡献度量，那么恶意的用户可以在一次编辑中集中添加大量的文本，随后删除这些内容，从而达到欺骗的目的。

这两种贡献度量的方式的根本问题在于：首先，它们不能抑制恶意用户利用度量本身的缺陷进行欺骗，而且欺骗的行为也难于识别；其次，这两种度量仅仅反映了内容贡献的一个侧面，而不是完整、全面的衡量用户的贡献。因此，这两种类型的度量往往会低估用户的实际贡献。编辑次数反映了一个用户参与协同的频率，但是却无法衡量该用户的“生产力”；而内容贡献仅仅以新增的文本内容为统计依据，对于那些重组文章内容，修订文字错误，移除恶意篡改等内容维护工作则忽略不计。因此，迫切需要一种协同贡献的度量，来真实反应维基百科用户对知识协同的实际作用。

编辑次数和文本数量均是协同贡献的数量指标。对于贡献的度量，更重要的是衡量其质量。Alder 等将贡献的质量定义为内容文本从加入到移除的时间长度。由于维基百科的条目编辑对于所有人开放，因此条目的内容会很快发生变动。如果某位用户再一次编辑中所新增的内容质量很高，那么这些内容就会在很长一段时间内，历经多次修订而得以保留，除非有用户用更高质量的内容替代之。反之，如果一段内容的质量很低，那么其实际寿命就会很短，很快就在随后的编辑中被取代。因此，内容的质量可以用一个位于区间  $[-1, 1]$  的常数来表示。基于以上假设，因此一个用户的内容贡献可以根据其文字寿命来度量。文本寿命是指在一次编辑过程中，一个用户实际新增加

的文字在随后的各个修订版本中所存续的时间。文本寿命同新增文本的数量和其存续的时间成正比，因此在度量过程中兼顾了内容的数量和质量两方面的因素。文本寿命的主要缺陷在于：它难于识别恶意的篡改和破坏。不论用户的编辑属于何种类型，最终都会被视为是用户贡献，而不是根据其特征将正常的编辑行为和破坏行为区分开来。因此，可以将回退编辑视为“负”的贡献，一旦某一段文字被撤销或者被新的内容代替，那么即认为该段内容的作者的贡献为“负”。对于恶意用户的编辑行为，由于其篡改和破坏的内容往往能在很短的时间内被修正，因此其贡献可以立即被判定为“负”，从而有效地将恶意用户和正常用户的实际贡献区隔开来。但是，这两种度量方式均不能有效地反映出从事内容维护工作的用户的贡献。

尽管上述方法对于衡量用户仍有不足之处，但是该方法实际上揭示了用户贡献认定的本质：被其他协同互用所认可的内容数量。用户创建内容本身是为了完成协同目标而进行的，其工作必然要被其他协同者所接受。然而，使用基于时间的判定方式来判断协同内容的质量本身是不稳定的。一段文字内容即使在较长的时间段内，尽力数个修订版本后仍得以保留并不意味着内容本身是高质量的。一方面，由于参与内容编辑的用户完全是根据个人兴趣和热情参与进来的，因此不同条目的用户活跃程度各不相同，对于哪些相对不活跃的条目来说，每次变更需要花费更多的时间。另一方面，即使内容经历了多次保本修订仍然保留也并不意味着其质量是受到认可的，有可能是存在着质量更为低劣的内容吸引了协同者的注意。尤其是编辑过程中出现编辑战或者恶意破坏的行为时，参与者主要精力都用于恢复正确的内容而无暇顾及及其他内容。因此，衡量真正的内容质量需要一个更稳健、准确的指标。

### 1.4.2 协同贡献度量的改进思路

度量用户的协同贡献，需要建立合理的度量手段。采用的方法必须要符合以下几个原则：

1. 维护公平性。度量结果必须要客观反应所有参与者的工作成绩，并且保证工作成效高的成员其度量结果也高。
2. 阻止个体的作弊倾向。任何度量方式都有其缺点和漏洞，这些漏洞容易被人利用而打乱整个群体的公平性。尽管不可能完全杜绝作弊，但是好的度量应该能显著提升作弊的成本，加大作弊的难度。

3. 符合评价者的价值取向。度量总本质上讲是主观的，依据评价者的价值取向而设定的。同时度量也是一种引导机制，引导被评价者的行为向符合评价者价值观的方向发展。如果度量错误地引导了被评价者的行为，该度量就是不合适的。

已有的研究成果为改进度量提供了有效的思路：既评价用户的协同贡献应该是依据其协同的质量，而不是协同的数量来进行。但是，很少有人分析过用户的协同贡献是否具有可比性，如果有可比性应该如何比较？本文认为，用户的协同贡献的可比性来自于两个方面：条目内的可比性和条目间的可比性。Adler 等人认为用户的协同贡献可以定义为作者所编写的被接纳的内容的数量。这实际上是条目内的可比性。因为条目的内容质量收到所有参与知识协同的影响，最终形成了一个综合、统一的质量标准，因此质量的可比性就可以直接体现为用户被接纳的内容的数量的比较。然而，这个结果并不能应用于不同的条目之间协同贡献的比较。因为不同条目间奉行了不同的质量标准，即使两个用户被接纳的内容的数量相同，其贡献也不能直接比较。为此，需要设立统一的质量标准，分别评价每一个条目的质量，使不同条目间的质量具有可比性。用户的协同贡献最终可以表示为条目内贡献和条目质量的函数：如果两个条目的质量处于同一水平，则用户的条目内贡献越高则用户协同贡献越高；如果两个用户的条目内贡献相同，则所在的条目质量越高，用户的协同贡献越大。

#### 1.4.3 基于文本相似度的协同贡献度量

文本相似度是两段文本的相似程度。使用文本相似度作为用户协同贡献的度量，可以有效克服传同方法的不足，充分利用个版本的文本信息，有效地反应用户贡献的大小。

基于文本相似度的协同贡献的核心思想是：个体的协同贡献可以经由比较个体编辑的内容与最终版本的文本相似度得到。维基百科的研究者认为，经过一段较长时间的协同编辑，条目最终可以达到一个稳定、高质量的状态。这也就意味着，一个条目的最终版本可以被视为是一个高质量的协同成果，最终版本里的所有内容是经过该条目所有参与者共同认可的，因此可以将最终版本中的内容视为衡量一个用户参与协同活动所做出贡献的标准。一个条目从最初创建开始，经过不断的完善与修订，所有高质量的内容得以保留，而低质量的、错误的内容不断得到替换与改进。最终形成一个稳定的版本。尽管维基百科的所有条目本质上并不存在一个最终的版本，但是经过较长时间的演化，条目的内容基本上已经固定，修订则主要集中在新内容的添加及修正

原有内容的错误等方面，在总体上已经十分接近最终版本。因此，本研究认为，在某个时间点上，对于那些内容十分稳定成熟的来说，可以使用该事件点上的最后一个版本作为条目最终版本的近似，以最后版本为标杆，分析协同用户各自的具体贡献。

记维基百科中所有的内容条目集合为  $E = \{e_i | i = 1, 2, 3, \dots\}$ ，协同用户的集合为  $A = \{a_i | i = 1, 2, 3, \dots\}$ ，则对于任意一个条目  $e_i$ ，在任意一个时间点上必然存在  $n (n > 0)$  个版本，记为  $v_{i,j}, j = 0, 1, 2, \dots, n$ 。其中， $v_{i,n}$  表示条目  $e_i$  当前时间点的最后一个版本，而  $v_{i,0}$  表示条目  $e_i$  在没有任何内容时的初始状态版本，这时条目  $e_i$  实际上为空。每次用户的编辑都会增添或者删改一部分内容，这些内容最后都可能在条目的最后版本中得到保留。因此，用户在每一次编辑过程所做的贡献可以视为究竟有多少内容在最后版本中仍然存在。如果两个版本中有一部分文字相同，则两个版本的文字在一定程度上是相似的，如果存在一个函数  $similarity(v_{i,j_1}, v_{i,j_2})$  能够将相似程度以数值的程度表示出来，则该值可以作为用户在一次编辑中所做出的贡献。为此，可以定义条目  $e_i$  的各个中间版本同版本  $v_{i,n}$  的相似度  $s$ ：

$$s_{i,j} = \begin{cases} 0 & j = 0 \\ similarity(v_{i,j}, v_{i,n}) & j = 1, 2, \dots, n-1 \end{cases}$$

条目的最初始版本  $v_{i,0}$  同其他版本的相似度总为 0。其他版本同最后版本的相似度表示为函数  $similarity$  的函数值，通过该函数可以计算出每个中间版本同最终版本的相似程度，即在中间版本和最终版本中均出现的内容。

#### 1.4.4 文本相似度的确定

确定文本相似度已经有许多成熟的算法。一般来说，判断两段文本内容在多大程度上相似，就是寻找在两端文本中能够完全匹配的文字的长度。Ratcliff/Obershelp 提出的文本匹配算法是一个得到广泛应用的方法。该方法通过寻找在两段文本中均出现的最长文本字串从而得出两段文本的相似程度。设有两段文本  $T_1$ 、 $T_2$ ，其长度分别为  $t_1$ 、 $t_2$ ，两段文本可以完全匹配的最长字文本串为  $S_1$ 、 $S_2$ ，其长度为  $s$ ，则两段文本的相似程度可以表示为：

$$similarity(T_1, T_2) = \frac{2 \cdot s}{t_1 + t_2}$$

例如，有两段文本“abcdef”和“abcd”，其最长完全匹配字串为“abcd”，因此可以计算其相似度为  $s = \frac{2 \times 4}{6+4} = 0.8$ 。当相似度为 1 的时候，说明两段文本完全相同，其最

大完全匹配字符串就是其自身；当相似度为 0 的时候，说明两段文本完全不同，最大完全匹配字符串不存在。

维基百科社区的知识协同活动主要有三种：1) 文字的添加，即在不该变原文内容的情况下增添新内容；2) 文字的删除，将上一版本中的某些内容删除掉；3) 文字的重排，改变文字的顺序，但不涉及内容上的增删。这三种活动在一次编辑过程中可能单独出现，也可能多种活动同时出现。内容的修改可以视为增加和删除同时出现。

Ratcliff/Obershelp 算法的特点是算法效率高，简单易用，但是直接使用 Ratcliff/Obershelp 算法计算相似度会有一些较大的缺陷。首先，算法对内容添加的位置敏感。设有一段文本  $T$ ，该文本可以分为两部分  $T(T_1, T_2)$ ，长度分别为  $t_1$  和  $t_2$ ，不妨设  $t_1 > t_2$ 。另一段文本  $T'$ ，该文本可以分为三部分  $T(T_1, \alpha, T_2)$ 。文本  $T$  和  $T'$  的唯一区别就是在  $T'$  中多增加了一个字符  $\alpha$ 。根据相似度计算公式： $s = \frac{2 \cdot t_1}{2(t_1 + t_2)}$ 。在极端情况下，两段文本的相似度仅为 0.5。这意味着新增加文本的位置对于文本相似度有很大的影响。如果一个用户在某一条目版本的基础上新增加了内容，并且该内容最终被接收，仅仅是由于添加的位置不同，可能会使新版本同最后版本间的相似度小于老版本同最终版本的相似度。这不仅与直觉不符，结果也不合理。

其次，如果在编辑过程中编辑行为是重组文本顺序，则 Ratcliff/Obershelp 算法同样显示出极大的不足。在条目的编辑过程中，用户往往为了使内容更符合维基百科的编写规范，行文更流畅，逻辑更通顺，在不改变条目的内容基础上（或仅作出少量文字性地修改），调整段落和句子的顺序。设有一段文本  $T$ ，该文本可以分为两部分  $T(T_1, T_2)$ ，长度分别为  $t_1$  和  $t_2$ ，不妨设  $t_1 > t_2$ 。用户在一次编辑中变更了两部分文本的顺序，新的文本为  $T'(T_2, T_1)$ 。两段文本的相似度经计算可得： $\frac{2 \cdot t_1}{2(t_1 + t_2)}$ 。另一方面，如果用户在原始文本的基础上将  $T_2$  部分删除，得到新的文本  $T(T_1)$ ，可以计算其与原始文本的相似度为： $\frac{2 \cdot t_1}{2t_1 + t_2}$ 。显然， $\frac{2 \cdot t_1}{2 \cdot t_1 + t_2} > \frac{2 \cdot t_1}{2(t_1 + t_2)}$ 。这就意味着应用 Ratcliff/Obershelp 算法，删除一段文本比改变文本的顺序更接近原始版本。但是，改变文本顺序保留了原始文本的所有内容，直观上更接近原始文本，而删除文本内容使得新的文本同原始文本产生了较大的差异，相似度应该较小。

造成这种结果的原因是，Ratcliff/Obershelp 算法仅仅考虑了最大完全匹配的文本字符串，而对其他完全匹配完全忽略。在上例中，文本  $T_2$  尽管也是完全匹配字符串，但是却不能直接影响相似度的结果。因此原始的 Ratcliff/Obershelp 算法既无法反应用户重排文本所做出的贡献，同时也对新增文本的位置极其敏感。本研究在该算法的基础上，

进一步改进了原有算法，使其即能适应一般的文本操作，也适用于文本的重新组织。改进的方法应该满足以下条件：

1. 若有文本  $T$  以及文本  $T_l$ ，文本  $T$  新增文本  $t$  后形成文本  $T'$ ，若  $t \in T_l$ ，则  $s(T, T_l) < s(T', T_l)$ 。
2. 若有文本  $T(T_1, T_2)$ 、文本  $T'(T_1)$  以及文本  $T''(T_2, T_1)$ ，分别使用 Ratcliff/Obershelp 算法和改进的算法计算相似度  $s(T, T')$  和  $s'(T, T'')$ ，则  $s(T, T'') < s'(T, T'') < 1$ 。

算法改进的思路是：文本的相似度不应仅考虑最大完全匹配子串的长度，而是应该考虑所有的完全匹配字串。两段文本中的每一个完全匹配的字串均可以计算得到一个相似度，这些相似度的线性组合最终成为整段文本的相似度。为此，可以在首次计算相似度之后，将两段文本中的最大完全匹配子串移除，同时使用一个虚拟字符代替该字串，即保持原有文本其他字串的相对位置不变。重新计算两个新的文本段落的相似度，并乘以适当的系数。反复使用该算法进行迭代，直到新生成的文本段落没有完全匹配字串。则迭代过程中所得到的所有相似度的和即为两段原始文本的相似度。详细算法如下：

---

**Algorithm 1** 改进的 Ratcliff/Obershelp 算法

---

输入：文本  $T_1, T_2$

$S = 0$

$c = 1$

**while** 文本相似度  $s(T_1, T_2) \neq 0$  **do**

$S = S + s \cdot s(T_1, T_2)$

$c = c(1 - s(T_1, T_2))$

将  $T_1$  和  $T_2$  的最大完全匹配字串  $T$  从原始文本中移除，用长度为 1 的虚拟字符代替，得到新的文本：  $T'_1 = T_1 - T$ ，  $T'_2 = T_2 - T$

$T_1 = T'_1$ ，  $T_2 = T'_2$

**end while**

**return**  $S$

---

该算法通过反复提取文本的最大完全匹配子串来达到计算相似度的目的，同时满足算法的有效性。每次迭代都会得到新的文本最大相似度和相似度系数  $c$ ，且  $\sum c = 1$ 。

改进后的算法可以满足改进的要求。首先证明使用改进的算法得到的文本相似度仍然满足  $s \in [0, 1]$ ，证明如下：

$$S = s_1 + (1 - s_1)s_2 + (1 - s_1)(1 - s_2)s_3 + \dots + (1 - s_1)(1 - s_2) \cdots (1 - s_{n-1})s_n$$

$$\text{显然 } S \geq s_1 \geq 0$$

$$1 > S \Leftarrow$$

$$1 > s_1 + (1 - s_1)s_2 + (1 - s_1)(1 - s_2)s_3 + \dots + (1 - s_1)(1 - s_2) \cdots (1 - s_{n-1})s_n \Leftarrow$$

$$1 - s_1 > (1 - s_1)s_2 + (1 - s_1)(1 - s_2)s_3 + \dots + (1 - s_1)(1 - s_2) \cdots (1 - s_{n-1})s_n \Leftarrow$$

$$1 - s_2 > (1 - s_2)s_3 + \dots + (1 - s_2) \cdots (1 - s_{n-1})s_n \Leftarrow$$

$$\vdots$$

$$1 > s_n$$

证毕

对于文本的重排，算法有效性也可以证明。设有一段文本  $T(T_1, T_2)$ ， $T_1$  和  $T_2$  长度分别为  $t_1$  和  $t_2$  且  $t_1 > t_2$ 。现对其进行两类不同的编辑，分别得到文本  $T'(T_1)$  和  $T''(T_2, T_1)$ ，可以证明  $s(T, T') < s(T, T'')$ 。证明如下：

$$s(T, T') = \frac{2t_1}{2t_1 + t_2}$$

$$s(T, T'') = \frac{2t_1}{2(t_1 + t_2)} + \left(1 - \frac{2t_1}{2(t_1 + t_2)}\right) \frac{2t_2}{2(t_2 + 1)}$$

$$= \frac{t_1}{t_1 + t_2} + \frac{t_2^2}{(t_1 + t_2)(t_2 + 1)}$$

$$s(T, T') < s(T, T'') \Leftarrow$$

$$\frac{2t_1}{2t_1 + t_2} < \frac{t_1 t_2 + t_1 + t_2^2}{(t_1 + t_2)(t_2 + 1)} \Leftarrow$$

$$2t_1 t_2^2 + 2t_1^2 t_2 + 2t_1 t_2 + 2t_1^2 < 2t_1^2 t_2 + 2t_1^2 + 2t_1 t_2^2 + t_1 t_2^2 + t_1 t_2 + t_2^3 \Leftarrow$$

$$t_1 t_2 < t_1 t_2^2 + t_2^3 \Leftarrow$$

$$t_1 \geq 1, t_2 \geq 1$$

证毕

为了进一步验证算法的有效性，本文随机选取了 5000 个维基条目，每个条目中个选取两个版本，应用改进算法同 Ratcliff/Obershelp 算法进行比对。同时，本文将编辑行为进行了分类，归纳为：文本的增减、文本重排、文本增减/重排三种类型。通过比对实验发现：对于纯粹的文本增减，两个算法所得到的相似度值基本一致，波动幅度

小于 1%；对于纯粹的文本重排，改进算法所得到的相似度值相对传统算法有明显提升，平均提升幅度为 12%；而对于文本增减和重排混合类型的编辑，改进算法所得到的相似度值相对传统算法有小幅提升，平均提升幅度为 4%。试验结果表明，改进算法即保留了传统算法的优点，同时对于传统算法所不适合处理的文本重排类型的编辑有较大幅度的改进。

改进的 Ratcliff/Obershelp 算法是一个理想的计算用户协同贡献的方法。通过计算各个版本同最后版本之间的相似度，可以进一步得到用户在每个版本中的实际贡献。对于条目  $e_i$ ，用户在版本  $v_{i,j-1}$  基础上进行编辑，得到版本  $v_{i,j}$  所做出的实际贡献可以表示为：

$$c_{i,j} = v_{i,j} - v_{i,j-1}, \quad 1 < j \leq n$$

图 1.1 显示了一个典型的条目编辑演化过程，以及用户在各个版本的贡献。

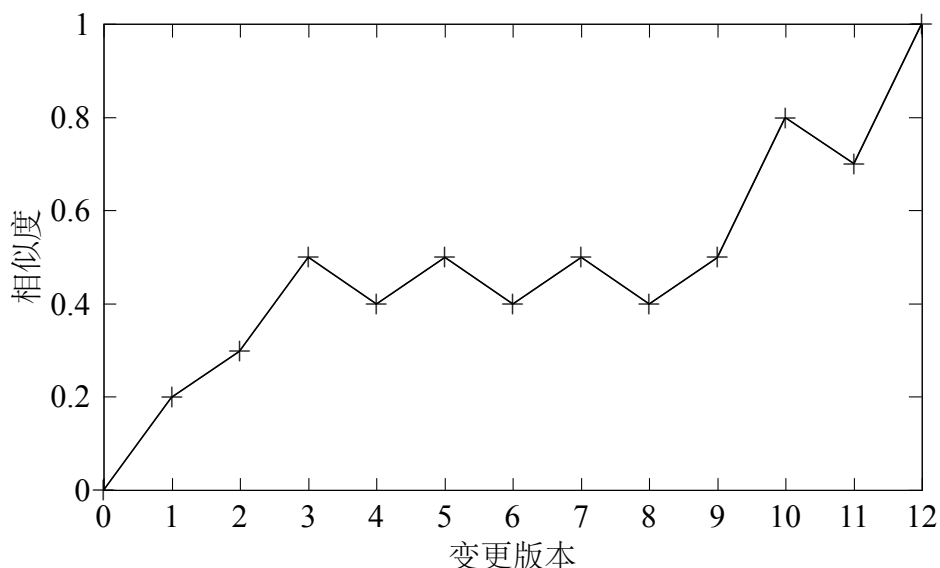


图 1.1: 不同版本的用户贡献

由图中可以看出，条目从初始没有任何内容的状态，经历了 3 个版本的编辑，形成了一定质量的内容。从第 4 个版本到第 9 个版本经历了一次编辑战，双方围绕某一内容反复进行修改和回退。当编辑战最终得到解决后，从第 10 个版本开始，条目内容进一步得到改进。第 11 个版本所编辑的内容未收到认可，在随后的第 12 个版本中，不但修正了第 11 个版本的内容，还进一步提升了条目的内容质量，形成了条目的最后版本。

但是直接使用相邻版本的文本相似度的差作为用户贡献不能有效抵制用户的破



坏行为。恶意用户也可以利用算法的漏洞骗取高贡献度。例如，一个用户可以利用两个账户，分别进行不断的删除内容，恢复内容的工作。每一次恢复都会使账户的贡献值增加，甚至可以使总贡献值超过 1。为解决这个问题，应该严格定义哪些协同行为可以获得正的贡献值。因为维基百科中知识协同的目标是编写更多、质量更高的条目，只有满足此目标的行为才可以视为做出贡献，所以真正能获得贡献的行为包括：1) 用户添加新内容并被其他用户所认可；2) 用户删除不合适的内容并被其他用户认可；3) 用户重排内容并被其他用户所认可。相应的，用户删除被认可的内容以及恢复已有的内容是不能获得贡献的。为此，在每一次用户编辑后，可以提取之前条目所有版本中已获得确认的内容形成一个虚拟版本，该版本实际上是当前编辑所取得的最大成就。设条目  $e_i$  经过了  $n$  次编辑，对于版本  $v_{ij}, 0 < j < n, \exists V, V$  包括了  $v_{i,1}$  到  $v_{i,j}$  各个版本中被接受的内容和其相对位置，则对于任意版本  $v_{i,m}, m \leq j$  都有  $s(v_{i,m}, v_{i,n}) \leq s(V, v_{i,n})$ 。用户的贡献可以表示为  $s(v_{i,j+1}, v_{i,n}) - s(V, v_{i,n})$ 。这样，只有当  $v_{i,j+1}$  版本的信息量超过  $V$  时，用户才可以获得正贡献，这样就有效地抵制了破坏和欺骗行为。

由此，可以将用户的协同贡献量化为具体数值  $c$ 。显然， $c$  的取值范围在  $[1, -1]$  之间。如果  $c_{i,j} > 0$ ，则意味着用户在当前版本的编辑活动为条目的完善作出了正的贡献；反之如果  $c_{i,j} < 0$ ，意味着用户在当前版本所做的编辑未得到承认，做出了负的贡献。用户的每一次编辑都可以计算得到贡献值，用户在一个条目中获得的所有贡献值的代数和即为该用户为该条目内容所做的总的贡献值。显然，所有用户的贡献值的代数和为 1，即  $\sum_{j=0}^n c_{i,j}$ 。用户的贡献也可能会小于 0，说明用户所贡献的内容并不为其他协同者所认可，要么被回退，要么在随后的编辑过程中被舍弃。

应用本文所提出的用户协同贡献的度量方法，克服了已有方法的不足之处，有效地区分了不同用户的协同贡献，主要表现为：

1. 协同编辑的每一次版本变更均可以计算用户在该次编辑中所做的实际贡献，贡献值即考虑了编辑内容的数量，也体现了内容质量的优劣。用户的贡献同时还体现了他人对协同工作的反馈情况，正贡献意味着外界的正面反馈，而负的贡献则反映了他人的负面反应。
2. 用户对于某一条目的贡献是其在每一个版本中的贡献的总和，这对于那些以维护条目内容为主要工作的用户来说，能够恰当地体现其贡献。这些用户的特点是每

次内容编辑的数量不大，但是编辑次数很多；因此，尽管其每次编辑的贡献值可能并不高，但是经过不断地参与编辑，贡献值的总和仍有可能达到比较高的值，体现其不可或缺的作用。

3. 同传统算法相比，改进的算法对于文本的重组给出了贡献的度量，体现了参与此类工作用户的价值。同时，文本重组的贡献远远小于内容贡献的贡献，这保证了度量的公平性。
4. 对于明显的恶意编辑行为具有一定的识别能力，并且用户贡献能对故意欺骗做出正确的反应。恶意行为对于研究正常用户的协同行为有很大的负面影响。应用本文提出的方法可以判断编辑过程中的“编辑战”和明显的恶意破坏行为，对于正确分析用户的协同行为具有重要作用。

### 1.5 条目质量的评价

上文提出的用户贡献的计算方法对于分析同一条目的协同者之间的贡献大小具有良好的效果，但是该方法对于不同条目之间的协同者之间如何比较协同贡献是不适用的。事实上，不论是基于内容字数（word count）的方法，还是本文提出的基于相似度的方法均无法应用于不同条目间协同者的贡献比较。例如，如果有两个条目  $e_1$  和  $e_2$ ，均由一个作者独立完成，且两个条目的长度基本一致。但是，条目  $e_1$  因为编写质量较高而被认定为“特色条目”，而条目  $e_2$  仅仅是一般质量的条目。根据现有的用户贡献计算方法，均可以得到两个作者的贡献是基本相同的。显然，这个结论并不符合虚拟社区和研究人员对于协同贡献的定义。即使两位作者参与编写的内容在数量上相同，贡献较高质量的内容的作者理应获得更高的评价，既他的贡献应该大于内容质量一般的作者。

这个问题说明衡量一个用户的协同贡献不仅应该考虑该用户在参与同一条目编写的群体中所做的贡献，还应该考虑该条目自身的编写质量。事实上，不同的条目编写者对于条目自身的质量认可程度是不同的，这就造成了条目内容的质量千差万别。在英文维基百科社区，条目的质量被分为 7 个等级，从质量最高的特色条目到质量最低的小作品<sup>1</sup>。中文维基百科目前的评级制度并不完善，仅仅涉及到了编写质量较高的特色条目。根据统计，截止到 2010 年 8 月，中文维基百科在所有 320,510 条目中共有

---

<sup>1</sup>[http://en.wikipedia.org/wiki/Wikipedia:Version\\_1.0\\_Editorial\\_Team/Assessment](http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment)

152 修改篇特色条目，平均每 2,108 条条目有一条特色条目<sup>2</sup>。这些数据说明，即使条目的内容经过了本条目的其他协同用户的认可得以保留，也不代表其质量可以获得社区其他用户和读者的认同。因此为了比较协同用户在社区中所做的实际贡献贡献，必须要考虑条目的编写质量。

维基百科社区为高质量的条目制定了一系列的评价准则，包括：详实的内容；恰当的遣词；观点中立客观；引用外部资料准确；内容结构编排合理；适当添加图片说明；符合格式指南；无错别字，且标点符号应用得当；链接适当，没有多余的链接<sup>3</sup>。这些评价标准中一部分是可以通过定量指标来描述，一部分则几乎无法使用定量指标来描述（如观点中立客观）。这就给评价条目内容的质量带来了很大的困难。当前虽然有一些学者针对评价内容质量开展了初步的定量研究研究，试图通过自动化的方式使用机器进行评级，但是这些研究均只是反映了质量评价的一个侧面，不可能完整的反应条目质量，更不可能代替人工的审核和评价。但是，应用自动化的评价方法对于确定条目内容质量仍然具有重要的作用。首先，维基百科中条目的数量巨大，英文维基的条目数量已经超过 300 万，中文维基的条目数量也已经超过 30 万，并且还在持续快速增长。如此巨大的数量使得严格的人工评级远远赶不上内容的增长。其次，条目内容是处在不断演进的过程中。即使条目经过评级，其内容还是会不断变化。尤其是当一个条目被评为特色条目时，会吸引更多的用户参与到内容的编辑中<sup>[9]</sup>。内容的动态性意味着评价也应该是动态的，及时反应变更的内容对条目质量的影响。第三，维基百科条目的内容涉及非常广泛，需要不同领域的专业人士对内容质量进行评价，这也为人工评价内容带来了很大的困难。最后，由于参与评价的人员可能具有不同的背景、阅历、和主观倾向，因此评价结果不可避免地带有一定程度的主观性。如果一个条目有两组具有不同价值取向的成员进行评价，其结果可能会大相径庭。也就是说，评价标准一致性难以保证。基于以上原因，一些学者试图利用一些定量指标，构建一个客观的评价标准。

评价条目内容质量的指标大致可以分为两类：基于条目本身的属性和基于用户的属性。如果一个条目自身的质量比较高的话，那么该条目一定会表现出一些异于其他低质量条目的特点。Lih 认为参与条目编写的人员数目和条目的编辑次数是影响条目质量的重要指标。这两个指标实际上反映了维基用户的参与程度，参与成员和内容变更

<sup>2</sup><http://zh.wikipedia.org/zh-cn/Wikipedia:特色条目>

<sup>3</sup><http://zh.wikipedia.org/zh-cn/Wikipedia:特色条目标准>

越多,意味着更多的内容被囊括到条目中,更多的错误被发现并得到纠正,不同角度的观点得到充分阐述,最终提升了整个条目的质量。这个观点在后续的研究中陆续得到支持。**Lim** 等基于条目内容的长度,条目的编辑次数,以及每个用户各自编辑的次数构建了一个评价模型。利用该模型可以针对条目内容评价其质量等级。**Zeng** 等人基于动态贝叶斯网络提出了一个条目内容的信任模型。他们认为:如果一个条目的内容被某一用户所修改,那么修改后的内容的可信度取决于三个因素,条目以前版本的可信度,当前版本作者的声誉和当前版本所修改的内容数量。三个因素同内容质量的关系为:如果之前版本的内容可信度高,那么修改后的内容可信度仍然会比较高;如果当前版本作者的声誉比较好(即经常贡献高质量的内容),那么他修改的内容也应该有较高的可信度;当前版本修改的内容越少,越有可能维持内容的可信度。**Stivilia** 等提取了七个指标作为评价条目内容质量的度量,包括:1) 内容涉及的范围;2) 内容的格式;3) 内容的独创性;4) 内容的权威性;5) 内容的准确程度;6) 内容的时效性以及7) 内容的可访问性。**Stivilia** 等还进一步将这些指标量化,通过条目长度、编辑次数、回退频率、外部链接等数量指标将上述指标转换为可计算的指标。**Dondio** 等从十个角度提出了内容质量的评价指标。随着越来越多的数值属性被引入到评价体系中来,**Blumenstock** 针对近 100 个指标,例如内容长度,句子长度,内部链接和外部链接的数量等分别进行研究,考察其能否有效区分特色条目和非特色条目。实验表明,单纯利用内容长度的值可以达到 97% 的正确率。

根据用户的特征来判断条目内容的质量是另一类广泛采用的方法。用户的特征一般包括用户的注册状态、参与协同的活跃程度以及用户的声誉。其中,用户的声誉对于内容的质量呈现出显著的正相关性。**Adler** 等通过对意大利语维基百科的研究发现:由声誉较低的用户所贡献的内容质量低下的可能性要远远高于由声誉高的用户所贡献的内容,并且这些其存在的时间也非常短暂,很快会被其他用户回退或者删除<sup>[7]</sup>。用户的声誉来自于其以往的编辑行为,如果其编辑的内容越能得到其他人的认可,存在的时间越长,那么该用户就越容易在社区中积累良好的声誉。因此,如果一个条目由声誉较高的用户进行编辑的话,那么该条目的内容质量较高的可能性就较大。

上述两种思路在本质上是一致的。用户的权威程度(声誉)决定了其参与的条目质量的,而用户的声誉也是由条目文本自身的特征计算得到的。区别在于,前者是对当前条目特征的分析,而后着是根据其他条目的文本特征预测当前条目的质量。本研究认为,利用条目内容自身的属性来评价其内容质量更符合质量评价的准则,利用条

目内容自身的属性完全可以用来进行条目质量的评价。Blumenstock 总结了现有研究关于文本自身的特征属性，由于这些特征是针对英文维基百科的，有相当一部分并不适合中文维基（如音节、分词等）。本研究在综合分析的基础上，去除了不适合的文本特征，总结归纳了适用于中文维基的文本特征，如表 4 所示。

表 4: 内容条目的文本特征

文字特征	结构特征	演化特征	其他特征
条目内容的长度	条目划分的段落 内部链接数量 外部链接数量 条目中图、表的数量 参考文献的数量	条目的编辑次数 参与编写的用户数量 条目编辑的频率	该条目被其他内容引用次数

表 4 中所列出的文本特征，均为可以直接获取的数量指标。其中，条目内容的长度同内容质量关系最为密切。Blumenstock 认为：条目越长，则其质量越佳。尽管维基百科对于特色条目的评价准则中并未对条目的长度作出明确的要求，并且明确指出“较短的条目也有可能入选”，但是实际上目前所入选的特色条目的长度都是较长的。维基百科自身的性质决定了较短的条目很难做到内容翔实、丰富，兼收并蓄各种思想和观点。但是，内容的长度对于内容质量来说本身只是一个必要条件而非充分条件。高质量的特色条目一般会比较长，但是内容较长的条目却不一定都具有较高的质量。这是由于条目长度本身是一个不够稳健的指标，任何人都各以通过纯粹的复制粘贴来扩展条目，而不论这些内容是否可靠、准确、合法。因此，只有综合考虑其他因素，才能准确地对条目质量做出评价和判断。

如果将条目内容质量作为因变量，表 4 中所列出的因素作为自变量，则可以构建一个回归模型，来表现各个因素同内容质量的关系。如果该模型能够很好地拟合现有数据，则可以利用该模型来对维基百科中的其他条目的质量进行评价。本文采用一元线性回归模型来分析自变量和因变量的关系。回归模型如下：

$$\begin{aligned}
 \text{内容质量} = & \alpha + \beta_1 \text{内容长度} + \beta_2 \text{条目段落数} + \beta_3 \text{内部链接数} + \beta_4 \text{外部链接数} \\
 & + \beta_5 \text{图表数量} + \beta_6 \text{参考文献数} + \beta_7 \text{编辑次数} + \beta_8 \text{用户数量} \\
 & + \beta_9 \text{编辑频率} + \beta_{10} \text{其他内容引用数} + \epsilon
 \end{aligned}$$

模型中的  $\alpha$  和  $\beta$  为待定系数， $\epsilon$  为随机误差，服从均值为 0，方差为  $\sigma^2$  的正态分

布。

### 1.5.1 数据集的选择

中文维基百科目前并没有非常完善的分级制度，仅有的两类级别分别是特色条目和小作品。小作品是比较短的文章，通常只有一段或更少，大多数小作品除了一点儿微不足道的主题之外，就没有其他资讯了<sup>4</sup>。小作品是所有条目中质量最低的一类条目。与之相对的是特色条目。其他条目的质量介于这二者之间。

截止到 2010 年 8 月，中文维基百科共有特色条目 152 条，小作品 173 个，这些条目都将纳入到数据集中。质量尚可的条目将从其他条目中选取。对于其他未评级的条目，可能既包括质量很高，但是因为各种原因还未进入到特色条目中的跳木；也包括一些质量同小作品相似，但是未被表示为小作品的条目。这就给选择质量一般的条目带来了一定的困难。本文所采取的方式是，选择一定长度的条目进入数据集。当前特色条目的内容长度至少达到 20000 字，而小条目则规定长度不超过 3000 字。因此本文选取内容长度在 5000–10000 的条目。这个长度的条目一方面有一定数量的内容支撑，同时又存在各种缺陷难以达到特色条目的条件，条目质量属于“一般”等级的概率很高。在满足条件的条目中，随机选取了 500 个条目，作为一般质量的条目。在选取的数据集中，随机抽取 100 个特色条目、100 个小作品和 400 个一般条目作为模型拟合的数据，其余数据作为模型的验证数据。

通过使用最小二乘法拟合数据，可以求得线性回归模型的各项系数，具体值如下：

从结果可以看出条目内容的长度、内部链接数、参考文献数、编辑次数、用户数量这几个因素均通过  $t$  检验对条目内容的质量具有显著影响。模型总体的效度  $R^2 = 91.34\%$ ， $F$  检验的结果  $F = 22.85$ ,  $sig. = 0.000$ ，说明模型拟合优度很好。条目内容的质量可以通过计算这几个文本特征的线性组合而得到。

利用得到的回归模型，通过测试数据可以检验模型的实际效果。验证结果如表 6 所示：可以看到，模型对于条目内容质量的评价具有较高的正确率和可靠性。相对于特色条目来说，对于小作品和一般条目的评价的准确度更高，说明判断一个条目内容“不够好”要比判断条目内容是否“足够好”要容易的多。这也只是模型可以继续改进的方向。

<sup>4</sup><http://zh.wikipedia.org/zh-cn/Wikipedia:小作品>

表 5: 回归模型系数

回归模型系数	系数取值	$ t  > p$
$\beta_1$	1.338	0.000
$\beta_2$	0.267	0.34
$\beta_3$	1.214	0.01
$\beta_4$	0.633	0.22
$\beta_5$	0.491	0.54
$\beta_6$	1.004	0.1
$\beta_7$	0.948	0.01
$\beta_8$	1.33	0.000
$\beta_9$	0.348	0.37
$\beta_{10}$	0.259	0.57
$\alpha$	1.244	0.01

表 6: 模型检验结果

	小作品	一般条目	特色条目
识别正确率	97.14%	94.57%	91.33%

通过评价条目内容的质量，使得用户贡献有了一致的比较基准。为了同用户贡献一致，也可以将条目质量归一化。用户贡献最终可以定义为：

$$\text{用户对某条目的协同贡献} = \sqrt{\text{条目质量}} \times \sum \text{用户每次编辑贡献}$$

将条目质量开平方实际上反映了这样一个事实：条目质量对用户贡献度的影响不是线性的。一个用户编写的内容越多（可以近似认为用户贡献度越高），则内容体现的条目质量特征也就越明显，二者间的关系也就越强。反之，如果用户进编写了几个字，则这些内容几乎与条目质量没有关系。而本文所采取的定义形式则反映了这一点。通过定义用户在社区中的贡献度，就可以根据用户对协同的贡献进一步分析其协同行为和动机，为下一步的研究打好基础。

## 本章小结

本章首先分析了维基百科社区中用户的各种行为，在此基础上明确地定义了用户的协同行为。随后提出计算用户的协同贡献是分析用户行为模式，进行用户分类的基本前提。本章分析了现有对用户贡献计算的文献，提出用户对某一条目的贡献应该由

用户在该条目中的贡献度和条目自身质量两个方面决定，并给出了相应的计算方法。通过本章的研究，为后序章节的研究工作打下基础。