云存储技术在学术写作教学中的应用研究*

李振华，李星遥
（清华大学 软件学院，北京 100084）

摘　要：学术写作是高校研究生和本科生教学体系的重要内容，其教学效率与效果受益于计算机技术的发展。从传统的电子邮件、Web 编辑器、版本控制系统到新出现的云存储系统，为学术写作的教学过程提供了更具实时性、便捷性与可靠性的工具支撑。作者结合高校师生学术写作的应用特点和具体场景，深度分析典型云存储系统的技术特点、实用效果和普遍问题，创新设计并系统实现优化解决方案，达到快同步、低冲突、准定位、可追溯的互动教学效果。

关键词：学术写作；协同编辑；云存储；隐式操作；三路合并
中图分类号：T393　　**文献标志码：**A　　**文章编号：**1673-8454(2021)04-0093-04

学术写作是高校研究生和高年级本科生需要学习和掌握的一项基本技能，是教学体系中既具备关键意义（学生从被动输入到主动输出）又极富个性化特点（学生写作水平的高低客观存在却又很难刚性量化）的重要内容。具体形式上，高校各专业师生通常采用计算机软件工具在线协同编辑学术文档（最常见的形式是论文），以实现实时可达的教学效率和及时互动的教学效果，师生双方能够便捷、清晰地看到彼此对论文的修改内容和编辑轨迹，有助于深入细致的写作水平感知和写作技巧领悟，达到教学目的。

一、计算机辅助学术写作教学现状

伴随计算机技术的不断发展，学术写作所依托的教学软件工具在效率与效果两个方面也相应不断提升。从时间维度看，这些教学软件工具可大致分为四类，包括较为传统的电子邮件、Web 编辑器（如 Google Docs）、版本控制系统（如 SVN 或 Git）以及近年来新出现的云存储系统（如 Dropbox、OneDrive 或坚果云）。其中，电子邮件方式虽然最为简单，但其效率很低、只能串行协作；Web 编辑器最为易用、无需安装任何软件，但其功能较弱且仅支持特定文档类型；以 SVN 或 Git 为代表的版本控制系统通过给共享文档明确加锁能够规避编辑冲突等诸多问题，但使用复杂、操作繁琐，一般只适合计算机专业师生协作编辑程序代码。

云存储系统是近年来国际上诸多学术团体和研发团队普遍采用的协同编辑工具，它具备传统网盘的简单易用性和自动化可靠备份等优点，又额外提供文档版本控制功能。相比于较早出现的电子邮件、Web 编辑器和版本控制系统等协同编辑方式，云存储系统具有两个突出优势：

1.允许合作者自由并行编辑

允许合作者自由并行编辑，源于“自动化文件同步”模式在云存储系统中的普遍采用[1]，每个用户对共享文档做出的修改都会被客户端自动同步到云端，再由云端自动同步到所有协作者。因此，师生之间通常不需要做无意义的互相等待，有效节约互相协调写作顺序的工作量。

2.节省协作者网络流量

节省协作者网络流量，源于“数据消重”和“差分同步”两项关键技术在云存储系统中的广泛实施[2]，能够自动识别重复内容从而避免协作者之间冗余的数据传输，基本上能做到只传输每个协作者真正修改的“增量”部分，不需要把一个较大的文件整体反复传来传去。特别是 2020 年新冠疫情期间，某些师生所在的地方没有宽带，只能使用宝贵的手机蜂窝（4G/5G）流量，此时云存储协同编辑的节省流量优势就显得格外重要。仅以 Dropbox 这一代表性云存储系统为例，目前全世界有超过 30 万个项目团队使用 Dropbox 进行协同编辑，每秒钟提交高达 4000 次数据更新到云端[3]，如此繁多的编辑操作如果以非云存储的传统方式进行，所消耗的数据流量将是目前的数十倍。

因为云存储系统“比较”匹配高校学术写作教学的各方面需求，并且适合于各专业师生快速掌握（无需精通计算机技术），所以最适合在学术写作教学中使用。不过，根

*基金项目：国家自然科学优秀青年基金项目“云计算内容分发的基础结构与关键技术”（编号：61822205）。

据笔者多年的实际教学经验，目前国际国内各主流云存储系统所提供的协同编辑服务，相比支撑学术写作教学的理想环境还有相当大的差距。

二、云存储在学术写作教学过程中面临的现实问题

云存储系统看似是一个理想的学术写作教学工具，在支持协作编辑的同时也足够简单易用，但是其也存在一个严重影响体验的冲突问题亟待解决：在实际教学过程中，肯定会出现多个师生用户并行编辑一份文档的情况，即使是串行发生的编辑操作，也会因为网络延迟的存在导致在云端视角是并行发生的情况出现，而并行发生的编辑就不可避免地会产生冲突。

对于冲突的处理，主流云存储系统都将操作透明性和用户友好性放在第一位，通过各种技术性的努力尽量避免冲突的发生或者自动解决已发生的冲突，但这些努力似乎效果并不尽如人意。根据笔者的实际教学经验，在日常使用中依然会碰到各种出乎意料的冲突问题，这些冲突问题会对本应该快速、准确的合作学术写作过程带来非常不好的体验，典型情况如下：

1.冲突概率大

即使协作者之间的编辑行为串行发生，冲突问题依然无法避免。此时，师生之间不得不就串行写作的时间进行更细粒度的控制，降低教学效率，影响教学效果。

2.同步时延长

一个协作者的单次编辑行为经常不能及时同步到其他协作者。实际上，该情况往往是造成上一情况（冲突概率大）的根本原因。

3.数据更新丢失

一个协作者的编辑行为有时会被另一个协作者的编辑行为完全覆盖。此时，师生之间将不得不停止整个互动写作过程，以处理数据丢失的严重后果。

4.文件读写死锁

一个协作者仅仅打开共享文档就有可能阻止所有协作者的编辑意图。这种情况会对教学过程造成很大的困扰，师生之间往往需要多次主动交流才能继续写作。

鉴于以上情况，需要构建一个可以合理高效解决上述冲突问题的云存储系统，减少师生的额外交流和主动干预，提升编辑效率，优化教学效果，充分发挥云协作编辑的优势。

三、云存储协同编辑问题的深度理解

为了搞清楚上述协同编辑冲突问题的发生机理，笔者结合高校师生学术写作的应用特点和具体场景，深入分析了国际国内八个最具代表性的云存储系统：Dropbox、Google Drive、OneDrive、iCloud Drive、Box、SugarSync、Seafile、坚果云，其中最后两个是国内企业自主研发的面向协同编辑的云存储系统，在我国高校师生中十分流行。

笔者收集了多个协同编辑真实数据集，具体信息见表1，均来自笔者所在实验室的教师和学生使用云存储协同编辑学术论文的真实行为。每个数据集均包含数据收集期间所有用户上传到云端或服务器端的所有共享文件版本。

表1 在线协同编辑真实数据集统计信息

数据集名	时间段	协作者数	文件数	版本数	主要文件类型
Dropbox-1	11/2/2018 - 2/6/2019	5	305	3527	tex, pdf, Matlab
Dropbox-2	4/3/2019 - 5/14/2019	6	216	2193	tex, pdf, Matlab
OneDrive	3/15/2019 - 5/31/2019	5	253	2673	tex, pdf, Matlab
iCloud Drive	2/1/2019 - 4/30/2019	6	301	3211	tex, pdf, Matlab
Box	3/21/2019 - 5/2/2019	8	273	2930	tex, pdf
SugarSync	4/11/2019 - 5/26/2019	9	325	3472	tex, pdf, Matlab
Seafile	2/17/2019 - 4/30/2019	7	251	2823	tex, pdf, Matlab

使用八个云存储系统和虚拟用户设备（租自亚马逊EC2公有云平台）重放上述数据集的协同编辑行为，以最大程度地复现各类冲突问题。重放的同时，使用Wireshark软件在用户端记录所有IP层网络流量，并使用Charles软件尽量解密应用层数据；对于应用层数据极难解密的云存储系统，通过阅读官方技术文档或者系统源代码搞清楚其工作原理。

基于上述努力，深度分析八个代表性云存储系统的技术特点和性能优劣，从量化角度理解到它们所提供的协同编辑服务相比支撑学术写作教学的理想环境确实还存在较大的差距。对于造成这些差距的关键技术原因，具体总结为如下五个方面：

1.编辑冲突的发生概率

从表1中可以计算出，编辑冲突的发生概率位于0到4.8%之间，Google Drive最低，iCloud Drive最高。更糟糕的是，绝大多数冲突都属于“伪冲突”，因为用户虽然同时修改了同一个文件，但是修改的位置并不相同。

2.文件锁机制

八个云存储系统中有五个未使用任何锁机制，因此无法主动避免冲突。剩下三个系统虽然使用了锁机制，

但因为粒度很粗（完整文件级别）而且无法实时监控用户对文件的编辑行为，所以经常出现过早或者过迟锁文件的情况，导致冲突问题的产生。

3.冲突解决策略

当冲突已经发生时，八个云存储系统中有六个采用的策略是云端保存所有冲突版本同时推送到所有协作者，由协作者自行处理，工作量很大；与此不同，iCloud Drive 让用户选择具体版本，工作量依然不小。Google Drive 的策略最为简单——云端和客户端仅保存共享文件的最新版本，协作者工作量为零，但非常容易丢失数据更新。

4.同步时延和消息队列

八个云存储系统中有五个使用了共享式消息队列来同步协同编辑的数据更新，导致同步时延最差时可以高达数十个小时，明显加剧了冲突问题。

5.数据更新方式

八个云存储系统中，有的使用简单全文同步技术，有的使用复杂但节省流量的差分同步技术。坚果云两者兼用，在同步流量和同步时延方面实现了很好的折中。

四、适配学术写作教学场景的优化方案

本文第一作者曾经指导过数十名研究生及高年级本科生的学术写作完整过程，深刻体会到学术写作支撑工具对教学效果的重要意义和其现阶段的现实问题。为填补云存储协同编辑实际性能和理想效果之间的差距，优化学术写作教学过程的实施效率和实际效果，基于第二部分的深度测量分析，总结出“隐式操作”这一基本概念来重新刻画学术写作过程中云存储用户间的协同编辑行为[4-5]，创新设计轻量高效的编辑冲突检测方法和自动化“三路合并”解决方案[6]。期望能够有效避免协同编辑冲突，并且整个过程中完全不对共享文档加锁，从而保证操作透明性和用户友好性，给师生带来更为顺畅便捷的互动使用体验。

1.协同编辑中的显式和隐式操作

“操作”是协同编辑的基本行为单元，在学术写作过程中是某个教师或学生的一次最小编辑行为。如果将一个共享文档看成一个字符串序列，那么“显式操作”就是用户真实的编辑行为，比如插入或删除一个字符；与之对比，“隐式操作” 更注重用户编辑行为的实际效果，比如用户先插入两个字符再删除其中一个字符，那么隐式操作就是仅仅插入一个字符。在云存储协同编辑的场景中，隐式操作是从云端角度看到的同一用户编辑同一文档后的两个连续版本之间的差异，相比显式操作更注重最终结果，而且无需系统在客户端实时监控用户编辑行为，非常适合用来刻画云存储系统中师生间的互动写作行为。

2.轻量高效的操作推断和编辑冲突检测方法

如图 1 所示，当两个不同的用户并行修改同一共享文档的 V_0 版本之后，分别产生了版本 V_1 和 V_2 上传到云端，此时云端首先需要完成的任务是推断出从 V_0 到 V_1 的隐式操作序列 S_1 以及从 V_0 到 V_2 的隐式操作序列 S_2。操作推断的常见方法是使用动态规划算法，但在云存储场景中会产生很大的计算量，比如 500 KB 的文档在普通虚拟机服务器上需要计算长达 30 秒。为提高计算效率，使用编辑图[7]的方法重新组织 V_0、V_1 和 V_2，从而极大缩短操作推断的时间开销，比如 500KB 的文档在普通虚拟机服务器上只需要计算 0.2 秒。有了隐式操作序列 S_1 和 S_2，能够以 O(N)的时间复杂度快速检测出是否真的存在编辑冲突（而不是伪冲突），其中 N 表示隐式操作序列 S_1 和 S_2 的长度之和。

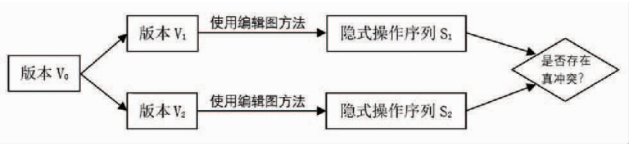

图 1 操作推断

基于上述方法，既实现了师生互动写作教学过程的快速数据同步（“快同步”），又达到了对任何一个参与写作者的每一次基本编辑行为的准确定位（“准定位”）。这样，无论对于教师还是学生，都能够快捷清晰地看到彼此对论文的细节修改内容，教师可以细致地感知到学生当前的写作水平，学生也可以深入地领悟到教师的写作技巧，有效提升了教学效率、优化了教学效果。

3.用户友好的操作转换和三路合并解决方案

如果隐式操作序列 S_1 和 S_2 并不存在真冲突（只有伪冲突），那么直接使用经典的操作转换[8-9]技术就可以将 S_1 和 S_2 合并为结果操作序列 Sr。否则，需要设计智能化操作转换技术以最小化用户干预程度、最大化用户友好性。具体来说，只发送一个最终合并版本 $V_{1,2}$ 给两个用户，并最大程度保留每个用户的编辑行为意图，同时最小化两个用户所要人工处理的冲突数量[10]。为顺利达成上述目标，如图 2 所示，使用拓扑图[11]和拓扑排序重新组织 S_1 和 S_2，将真冲突操作的转换提前到伪冲突之前，从而最小化真冲突操作对整个转换过程的影响。最

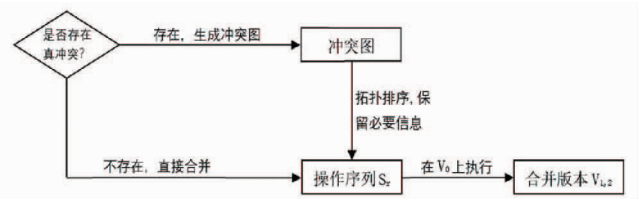

图 2 操作转换并生成合并版本

后，云端在 V_0 上执行 Sr 生成 V_{12} 同时发送给两个用户，完成三路合并整个过程。

三路合并的一个明显优势是赋予每一名参与互动学术写作的师生随时追溯教学过程中任意共享文档的任意版本的能力（“可追溯”），以使教师能够帮助学生在特定阶段宏观“复盘”写作过程的整体编辑轨迹，让学生清楚地看到每一部分内容不断进化提升的状态演变，这对于学生写作能力的提高具备总结性的整体意义。

五、优化方案的实践效果

基于亚马逊 EFS（弹性文件系统）和 S3（简单存储服务）搭建了原型系统 UFC2（User-Friendly Collaborative Cloud），于笔者所在实验室的 40 名师生中实际使用（包含 4 名教师及 36 名学生）开展学术写作教学工作，以评估上述优化方案的真实效果。同时开源 UFC2 系统的所有代码 https://ufc2.github.io/，以便于其他研究者复现实验结果。

实验室 40 名师生实际使用（前后长达 7 年时间）的测量结果表明：UFC2 平均能够减少 98%的编辑冲突，实现了“低冲突”的教学前提。同时，如图 3 所示，UFC2 解决冲突带来的额外时间开销通常介于 10 毫秒到 80 毫秒，平均来说只占文件同步时延的 2%，实现了“快同步”的教学前提。UFC2 系统中单次文件更新所消耗的网络同步流量一直低于 60 KB，即使当文件较大（超过 100KB）时也只需要 10 KB 左右，这对于 2020 年新冠疫情期间使用宝贵手机流量开展学术写作教学的师生来说十分重要、受益匪浅。同时，由于每个用户的编辑行为意图被最大程度保留，因此师生间人工协调解决的开销很低，大多数情况下只需要一条微信消息即可达成一致，达成了“准定位”的教学目标。此外，每个用户每次编辑行为后的文件版本都被云端可靠地保存，达成了“可追溯”的教学目标。学术写作过程的“低冲突”“快同步”“准定位”“可追溯”结合在一起，形成了良好的互动教学效果。

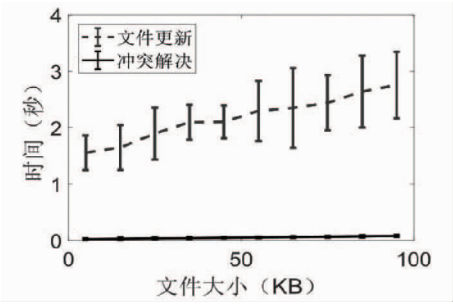

图 3 冲突解决和文件更新所需时间随文件大小变化趋势

作为云计算技术的一种新应用形式，云存储在线协同编辑近年来快速流行，有效提升了高校计算机教育和科研工作的效率，特别是研究生和高年级本科生学术写作的互动教学效果，但在编辑冲突这一关键问题的解决上依然存在诸多缺陷、影响教学体验。本文从隐式操作的新角度提出并实现基于云存储协同编辑的学术写作优化教学方案，达成了快同步、低冲突、准定位、可追溯的良好互动教学。下一步计划吸引更多的高校师生用户以充分检验设计方案的扩展性与完备性，以充分考虑不同类型师生的教学需求，进一步丰富和提升教学效果。

参考文献：

[1]Li Z,Wilson C,Jiang Z,et al.Efficient Batched Synchronization in Dropbox-like Cloud Storage Services[C]. Proceedings of the International Middleware Conference (Middleware).Beijing,China:ACM,2013:307-327.

[2]Li zhenhua,JIN cheng,XU tianyin,et al.Towards Network-level Efficiency for Cloud Storage Services[C].Proceedings of the Conference on Internet Measurement Conference(IMC).New York,USA:ACM,2014:115-128.

[3]Smith C.33 Staggering Dropbox Statistics and Facts [EB/OL].https://expandedramblings.com/index.php/dropbox-statistics/.

[4]何发智，吕晓，蔡维纬，等.支持操作意图一致性的实时协同编辑算法综述[J].计算机学报,2018(4):840-867.

[5]张晓杰，刘杰，马志柔，等.基于操作日志的云存储服务多终端同步算法[J].计算机工程与设计,2013,34(11):3894-3899.

[6]Sousa M, Dillig I, Lahiri S K.Verified Three-way Program Merge[C].Proceedings of the Conference on Object-Oriented Programming, Systems, Languages & Applications(OOPSLA).Boston,MA,USA:ACM,2018:1-29.

[7]Myers E W. An O(ND) Difference Algorithm and Its Variations[J].Algorithmica,1986(1):251-266.

[8]Ellis C, Gibbs S. Concurrency Control in Groupware Systems[C]. Proceedings of the International Conference on Management of Data (SIGMOD).New York,USA:ACM,1989:399-407.

[9]廖斌，何发智，荆树旭.实时协同工作系统中操作转换算法综述[J].计算机研究与发展,2007(2):326-333.

[10]高丽萍，陶长青.文件管理中一种新颖的冲突检测和解决方法[J].小型微型计算机系统,2019,40(6):1227-1235.

[11]Marx D.Graph Colouring Problems and Their Applications in Scheduling[J].Periodica Polytechnica Electrical Engineering (Archives),2004(48):11-16.

（编辑：鲁利瑞）