程序设计 II, 2013 年秋
大作业：社交应用的数据存储与实现
发布日期：13 年 10 月 08 日
阶段提交：13 年 11 月 17 日
期末提交：13 年 12 月 22 日
截止日期：14 年 02 月 28 日

1 概述

程序设计 2（数据结构）课课程大作业，强调数据结构的选择和对性能的分析。涉及排序、查找、树、散列等操作。目的在于让大家利用所学知识，进行一个中型系统级程序的设计，进一步理解和掌握基本的数据结构和算法。

本课程的课程大作业需要大家设计并实现一个系统，模拟一些简单的社交应用所具备的功能。

2 前台功能要求

本章介绍该系统需要实现的所有前台的功能要求。文中被特别标出的功能为扩展功能，此外为基本功能，你可以从扩展功能中选择一些功能进行实现，或设计自己的扩展功能。提交作业时，你需要详细介绍你实现了哪些前台功能。

2.1 用户

所有用户必须注册或登录后才能使用该系统。注册需要提供登录用的用户名（要求是唯一的，且将被用于在各种用途显示给其他用户）、密码和基本个人信息。密码是一个 8～32 位的字串。

2.2 基本个人信息

基本的个人信息包括姓名、性别和生日。用户必须提供这些信息才能注册。个人信息中的相关数值要求符合特定的格式规定。

2.3 附加个人信息

除了基本的个人信息外，用户还可以选择性填写一些其他的个人信息，包括联系电话、家乡。个人信息中的相关数值要求符合特定的格式规定。

2.4 修改密码

用户登录后可以修改密码，修改时需要提供原密码以进行核对。

2.5 修改个人信息

用户在登录后可以修改自己的个人信息。其中基本信息不允许在修改时被清空，附加个人信息可以任意修改。个人信息中的相关数值要求符合特定的格式规定。

2.6 删除用户

登录后可以选择删除用户，删除用户时需要验证密码。删除后该用户名不能继续登录，也不能再被用来注册新的用户。

2.7 关注

系统采取关注（Follow）机制。任何用户可以关注除自己外的任何用户。

2.8 用户查找

每个人可以根据用户的信息查找其他用户。可以通过用户的用户名找到唯一的用户（或不存在）；或者可以通过用户具有姓名为某特定值，性别为某特定值，生日在某个区间段内，电话为某特定值和（或）家乡为某特定值（若用户未填写特定项目则不会被查找到），这些条件中的一个或特定的几个项目的组合甚至是任意的若干个的组合进行筛选和查找。（找到的用户可以以任何你认为方便的方式进行排序。）注意，进行查找操作时不应当找到自己，但可能找到已经关注的用户。

2.9 关注推荐

你可以根据各用户的性别、生日、家乡，关注了哪些人，哪些人关注了该用户，甚至是发表的消息的关键字等任何你可以得到的信息，给该用户推荐一些他可能认识或感兴趣的人。这些人应当当前不在该用户的关注列表内，且不能是用户自己。

2.10添加关注

在查找某个用户或得到关注推荐后，若当前尚未关注该用户，可以添加对方到自己的关注列表中。

2.11关注列表

用户可以查看自己关注了哪些人（Following），以及那些人关注了自己（Follower）。所有这些人的按照用户名的字典序进行排列。

2.12取消关注

用户可以从自己的关注列表中取消对某些人的关注。若某用户执行了删除用户的操作，则所有与他相关的关注和被关注的关系同时被解除。

2.13消息

用户可以发布一些消息，所有人都可以看到这些消息。

2.14发布消息

每条消息的长度应不大于 140 个字母或汉字。消息在发表后不能进行编辑。

2.15消息列表

用户可以在一个界面下同时查看到自己和他关注的人发表的所有消息；如果消息过多，需要分页显示，并优先显示最新的若干条。这些消息应当以发布时间的倒序排列，即最新的应当最先出现。

2.16用户消息

用户除了查看自己订阅的所有用户的消息外，也可以单独查看某个用户的消息。用户可以通过指定用户名或在好友列表、用户查找、关注推荐时列出的用户中选择用户来查看其消息。

2.17转发消息

用户可以转发一条消息，转发后关注了该用户的人可以在自己的消息列表中看到这条消息。转发的同时可以对消息添加自己的评论构成一个整体，但评论与消息的总长度仍受到与发布消息相同的限制，总长度超过限制时原消息会被截断。转发的消息应当标明原作者。

2.18折叠消息

如果消息转发中没有评论功能，系统应当在显示消息时将所有同一来源的消息显示在一起，省略掉重复的消息。这些消息按照该用户最早看到的一条消息计入排序。

如果消息转发中有评论功能，用户在查看消息时，将所有原消息相同的消息放在一起显示，并省略公共的部分。此时按照这些消息中最新的一个进行排序。

2.19转发计数

显示消息时同时显示消息被转发的数量，间接的转发仍然会计入到该计数中。例如用户甲的消息被用户乙、用户丙转发，而用户丁又从用户乙处转发了该消息；则用户甲的消息的转发计数为 3，而用户乙的该消息的转发计数为 1，其他人那里为 0。

2.20删除消息

用户可以删除已经发表的消息，删除消息的同时，从该消息转发而来的消息同时被删除。

3 文件存储

为了完成上一章所介绍的功能，你需要将数据保存到文件中，并对数据进行索引。这一章介绍你需要完成的系统中对文件存储的后台的要求。这一部分具有较大的灵活性，你需要自己选择合适的数据结构并进行实现。所以你要在最后提交的文档中详细描述你使用了怎样的数据结构，效果如何；以及磁盘文件中数据和索引的组织格式。

3.1 数据文件

用户、关注关系、消息及评论等内容应当被保存到文件中，保证下一次运行程序时可以从文件中读取。数据文件的格式可以自行设计，你可以设计自己的文件头，自行设计对不同类型数据的表达方式。你的后台系统应当提供对数据进行管理（增删改查）的相关接口。

3.2 索引文件

除数据外，还要考虑对数据进行索引，将索引保存到文件中，这样每次运行程序时不需要将所有内容读取到程序中，可以通过索引快速定位数据在文件中的位置并获取相关数据。

你可以在磁盘上采用树、散列表等数据结构保存内存信息，具体使用什么样的索引方式由你自行选择和设计。

你可以自行选择索引文件的具体实现。你可以将数据和索引保存到同一个文件中，也可以将它们保存到不同的文件中；不同的数据表可以使用同一个文件也可以使用不同的文件。但注意不能以用户和（或）消息为单位建立文件或文件夹存储。

3.3 一致性刷新方法设计

一致性指，由于数据量的缘故，不能将所有数据全部在内存中保存。这些数据信息会在内存和硬盘中出现两份，你需要考虑何时对文件进行读写，保持内存和文件数据的相对一致，且保证程序的运行效率。

4 系统测试

4.1 正确性测试

正确性既包括结果正确没有引发程序的异常退出或内存泄露，也包括符合预期的性能要求。你需要自行书写测试程序测试自己对数据存储、查询和删除的操作行为是否正确。测试程序是一个独立于应用的，针对数据管理的相关函数（或类）进行编写的程序。测试程序可以直接调用对应的函数（或方法）进行测试，而不需要通过标准输入输出（或其他的方式）测试完成的程序。

你需要对数据表和索引的插入、删除（如果有）、修改（如果有）和查找操作自动生成随机数据进行测试，此外还要对系统中常见的操作进行测试。测试应当保证数据量不少于 100 万条，以保证可以测试到程序各方面的问题。

4.2 性能测试

性能测试与正确性测试的要求类似，但注重对性能的考察。要求针对常见的操作进行不少于 100 万次的操作。在测试过程中，每隔一定的数据量检查已耗时间，并使用相关软件绘制时间-数据量的散点图（或拟合函数）。具体的测试指标可以自行设计。

注意，由于每个人机器的配置与性能不同，和其他人比较你的程序的效率是没有意义的。我们也不会根据你具体测量的数值进行评分。所以我们关心的是你通过测试可以证明程序的性能符合预期的函数模型，以及通过测试，比较在使用了索引和不使用索引，或使用不同的数据结构的情况下系统的性能的差异。通过比较改进自己的算法或选取更适合的数据结构。

提交作业时，需要提交测试程序，并需要在提交的文档中详细介绍测试的方法和结果，包括上文提到的针对关键操作的耗时-数据量的图表。

5 其他要求限制

5.1 自行实现关键数据结构

数据结构大作业考察大家对数据结构的选择、设计和实现。程序中涉及的关键的数据结构和将内存刷新到文件的操作等应当自行实现，不能使用已有的库，或任何数据库系统。

5.2 函数库使用

除了关键的数据结构外，你可以使用任何你喜欢的库进行开发。但应尽量保持程序的简单，切忌实现了过多无用的功能而忽略了本文提到的你必须完成的功能，否则对你的成绩不会有任何有益的影响。

5.3 开发和编译环境

程序要求使用 C/C++实现。要求给用户提供字符界面使用该系统。答辩时需要大家演示自己的系统。程序为单机程序，不需要实现多线程，网络通讯等高级功能。

6 阶段检查

阶段检查时，前台功能上，查找用户功能可以只能通过用户名进行查找，而不必完成更为高级的查找方式；其他前台功能的要求全部完成。后台存储上，你至少应当能够将数据写入到文件中，并每次运行程序的时候从文件中读取已有的数据。

阶段检查不要求大家实现索引以及基于索引的查找功能。这些要求不是强制性要求，如果你有更好的对系统开发的规划，你可以在中期检查时完成其他的部分（如完成了数据的保存和索引，没有前台功能），但我们要求完成的工作量不少于我们中期检查的工作量要求。

这里规定的检查的内容是最低的限度，我们鼓励你在阶段检查时完成更多后面的功能，阶段检查的目的之一在于检查大家的作业进度。

7 期末检查

期末检查要求至少实现上文所叙述的所有基本要求；完成测试程序；完成文档介绍自己所使用的数据结构和算法，性能测试的结果，及系统的架构和实现的功能的介绍。

对因为期末检查时没有完成而放弃检查的，期末检查未通过的，或者在寒假假期继续改进了系统的同学，可以在 14 年春季学期开学时参加第三轮的检查。第三轮检查与期末检查标准相同。

8 评分

大作业占这门课程总成绩的 20%。满分 20 分中，基本功能实现（5 分），后台实现（5 分），扩展功能实现（5 分），测试与文档（5 分）。

其中扩展功能部分你可以完成前台功能说明中提到的部分，也可以自行设计扩展功能。除此之外你还可以在完成前台的基本功能的基础上完成一些较为复杂的后台索引结构，同样可以视为扩展功能。对于扩展功能的评分，如果你基本完成了本文中所有前台的扩展功能，

或在后台实现了 B 树等较为复杂的数据结构，或者对某一方面有较为高级和复杂的算法（如结合自然语义的关注推荐），都可以在这部分拿到满分。

9 提示和建议

一般来说，测试和修复错误的时间会占到开发总时间的一半以上，尤其涉及到一些较为复杂的算法时测试所占的比重往往会更多。所以请尽早着手完成作业，注意合理安排好自己的时间。

考虑使用二进制文件读写而非文本文件读写，这样既可以提高程序效率，又可以降低开发难度。

虽然我们不允许你直接使用已有的库来实现系统中关键的数据结构，但建议你查阅相关资料，参考已有的数据结构而非从头开始自己的设计。例如你可以去了解数据库系统都是用了怎样的方式管理数据，虽然数据库系统可能有很多较为复杂而我们不需要关心的高级话题（如并发控制），但对于数据的存储和索引，这些设计对我们很具有参考价值。

建议你在开始书写程序之前就对你要使用的数据结构有所设想，以避免在开发过程中，因为要使用某种数据结构而出现对已经完成的代码大量的返工。

10 参考及附录

1. 数据库函数库
2. SQLite: http://www.sqlite.org/
3. Oracle Berkeley DB: http://www.oracle.com/technetwork/database/berkeleydb/index.html
4. Google C++ Style Guide: http://google-styleguide.googlecode.com/svn/trunk/cppguide.xml
5. JPL Coding Standard
6. 华为编程规范