# 6 数据模拟与测试

## 6.1 系统用户规模

在做数据模拟之前，首先必须要考虑到系统的用户规模。一般来说，为了保证模拟数据的真实性和可靠性，需要尽可能接近或超过实际系统的用户规模。具体而言，如果系统用户规模为百万级别，那么模拟数据的规模也应该达到百万级别或以上，以便充分测试系统的性能、稳定性和可扩展性。

系统的用户规模影响诸多方面的内容，比如用户关系。用户规模越大，用户关系的复杂程度就越高，数据处理的难度越大，输出的结果可能会变得不可控。

用户规模受到用户信息的限制。本测试内容主要针对用户关系的发现，模拟数据中，尽管用户信息的细致性和真实性并非是重点内容，但是好的用户案例可以更好的帮助测试人员评估预期结果的合理性，因此，在充分保证用户信息的真实意义前提下，模拟的数据不能过高，过高将会导致测试数据失真。

综合上述内容，考虑到模拟数据的生成和处理成本以及数据模拟的实际目的，针对用户发现的数据模拟数据量在十万级别。这个级别的数据量可以保证充分的测试接口，同时兼顾到数据的生成和处理成本，保证测试效果和数据质量。

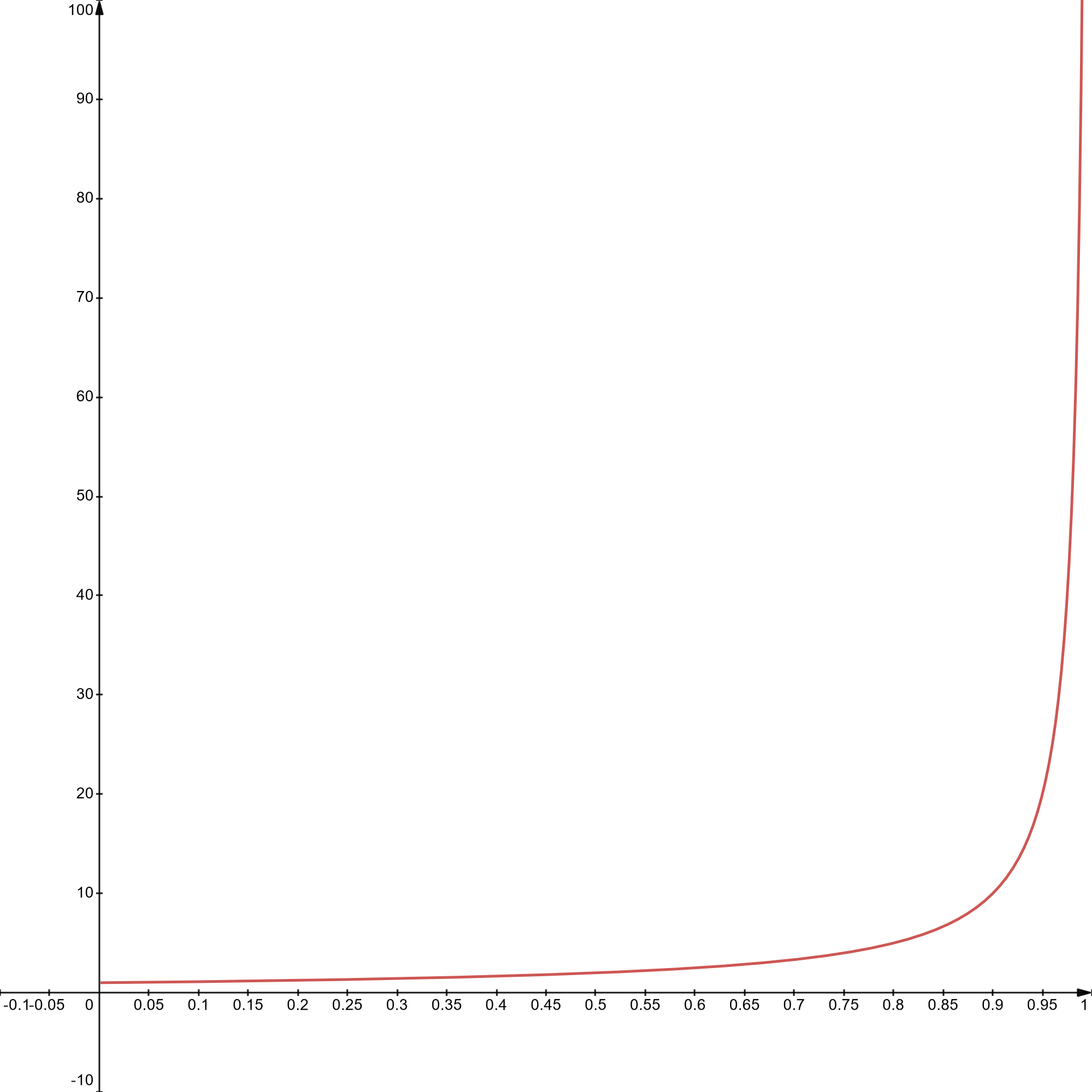
## 6.2 用户关系的模拟

成功模拟一定规模的用户数据后，下一步内容就是用户关系数据的模拟。与用户数据的模拟不同，用户关系数据是驱动用户发现算法的关键。在进行用户关系数据模拟时，需要考虑到用户之间的互动方式、频率和强度等因素，并根据实际情况生成相应的数据。

有一个基本定律，对于内容的访问遵循80/20原则，也就是20%的内容，会占有80%的访问量。对于用户关系的一个描述是，20%的用户，会占据系统80%的粉丝数据。在数学中有一种分布规律可以近似描述这种规律：Zipf定律(Zipf's law)。以下模型能够很好的适合用户关系的模拟：

其中，u是一个[0,1)的随机数，α是幂律分布的指数。当α = 1时，它的分布情况比较理想，分布规律近似相近于函数：

图形工具可以直观的表现它的规律。



【todo做个标记】

系统把上述公式作为用户关系的概率函数，产生了超10万用户关系记录。它们被存入数据库，并形成事务集，用来后续算法处理。



算法的输出是频繁事务集。在最小支持度为1000的参数模型下，共生成了241条频繁项集。这是基于用户关系的用户推荐的核心数据。



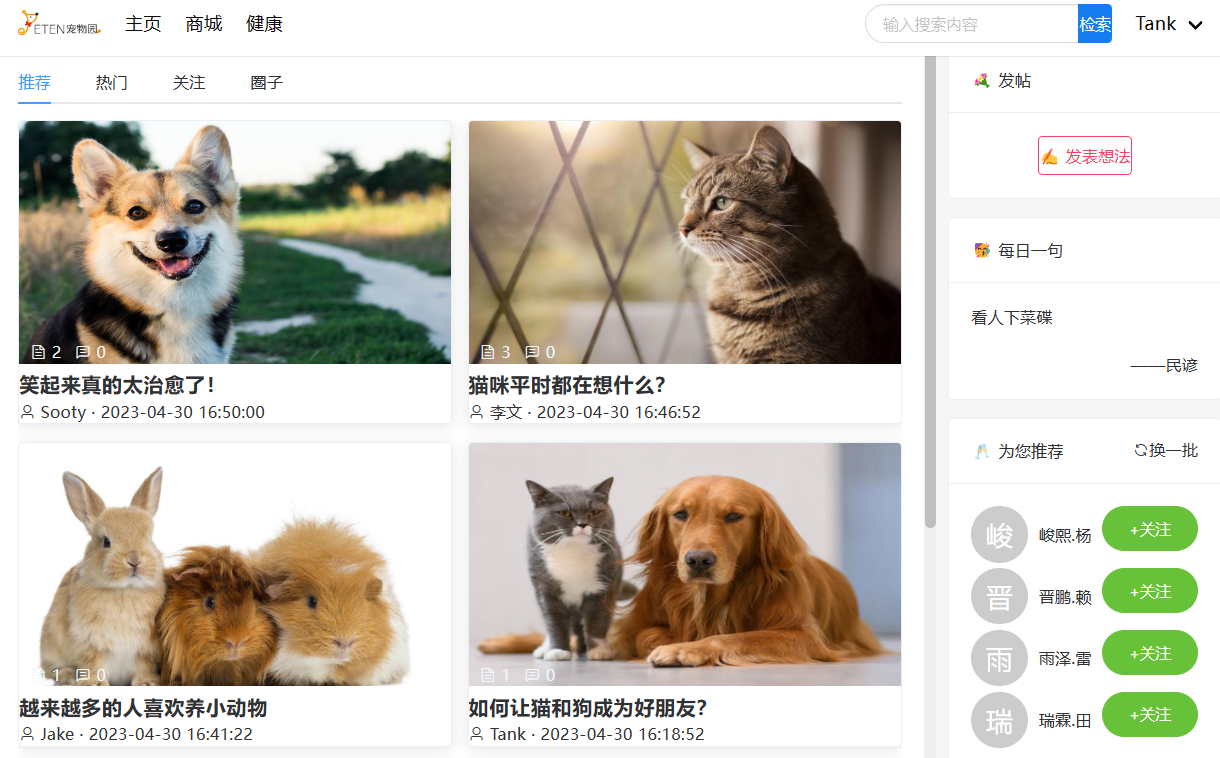
## 6.3 系统主要功能测试

功能测试是黑盒测试中的一种。本系统庞大，功能众多，本章节展示系统主要功能测试。

1、系统标识



2、系统概览



3、内容编辑与发布



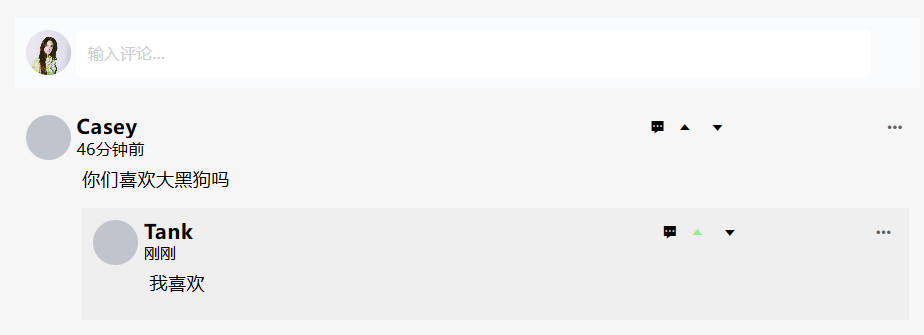
系统采用了富文本编辑器，支持markdown的语法，用户的编辑体验将更加丰富。



4、文章预览与推荐



5、用户评论



6、