中文維基百科
用戶分析

zh. 2006年
中文維基年會
CHINESE WIKIMEDIA CONFERENCE 2006

Snowyowls@zh.wikipedia.org

簡介

- 分析用戶的編輯歷史
- 瞭解不同用戶參與維基百科的特點
- 運用數學工具對用戶的編輯次數數據進行深入分析

數據來源

- 中文維基百科註冊用戶超過8萬
- 選取其中一小部分進行分析
- 將用戶分為5類
- 機器人、管理員、用戶、破壞者、辯論者

數據來源

- **機器人：** 中文維基百科全部閤法機器人
- **管理員：** 中文維基百科管理員
- **用　戶：** 統計頁面中編輯量Top50的註 冊用戶
- **破壞者：** 當前的破壞中列名的註冊用戶
- **辯論者：** 在南京大學和法輪功兩個條目的討論頁發言超過5次的用戶

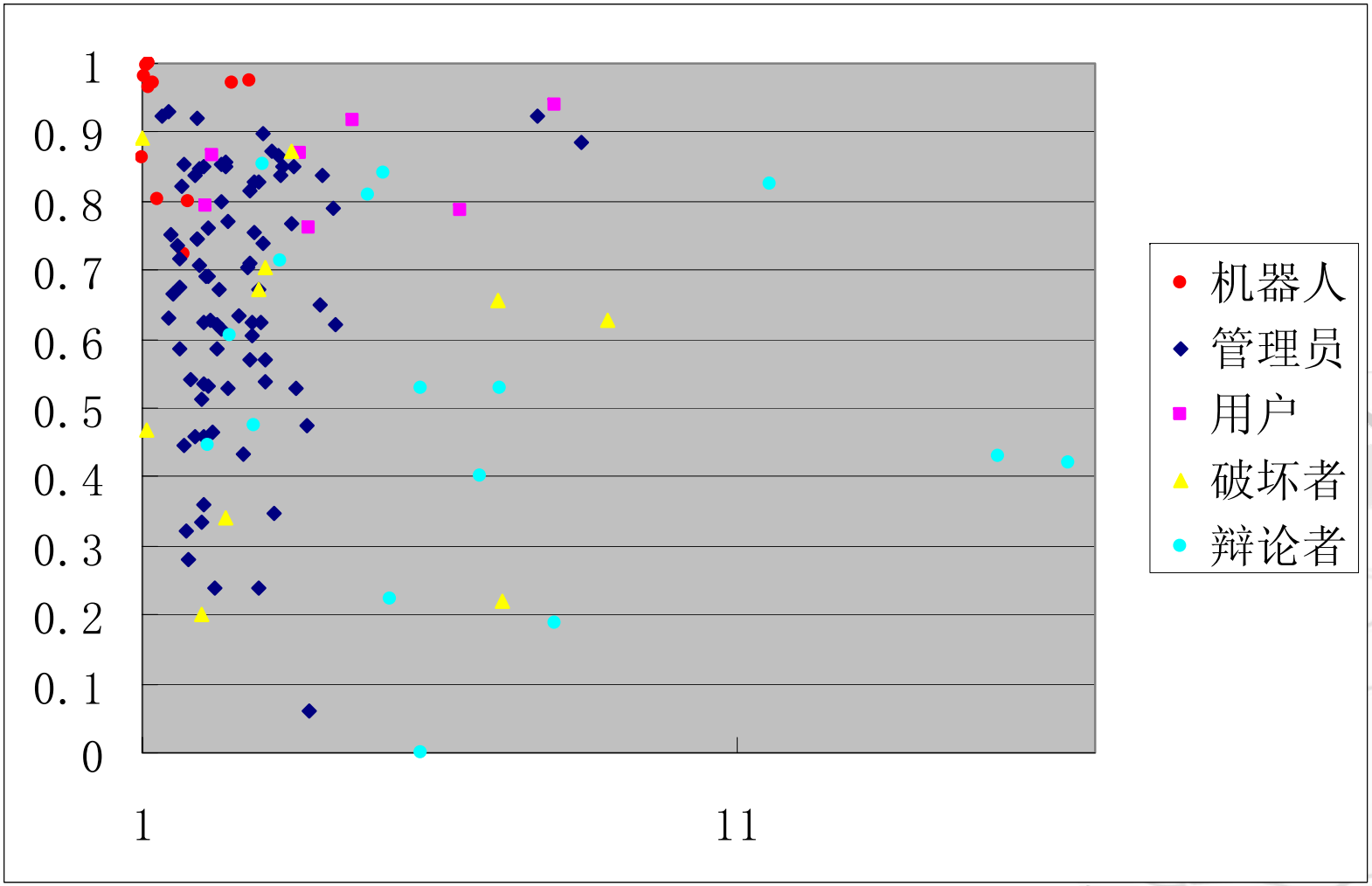
數據來源

- 用戶的編輯數據用kate的工具從數據庫中提取
- 數據包括：Image uploads、cur and old images、Distinct pages edited、Deleted edits、Edits/pages、Articles、Talk、User、User talk、Project、Project talk、Image、Image talk、MediaWiki、MediaWiki talk、Template、Template talk、Help、Help talk、Category、Category talk、Portal、Portal talk、move

方法與結果

- 用戶的平均頁面編輯次數對內容編輯傾嚮作圖
- 內容編輯傾嚮是用戶的Articles、Template、Portal三項編輯次數之和佔總編輯次數的比例
- 辯論者常具有極高的平均頁面編輯次數

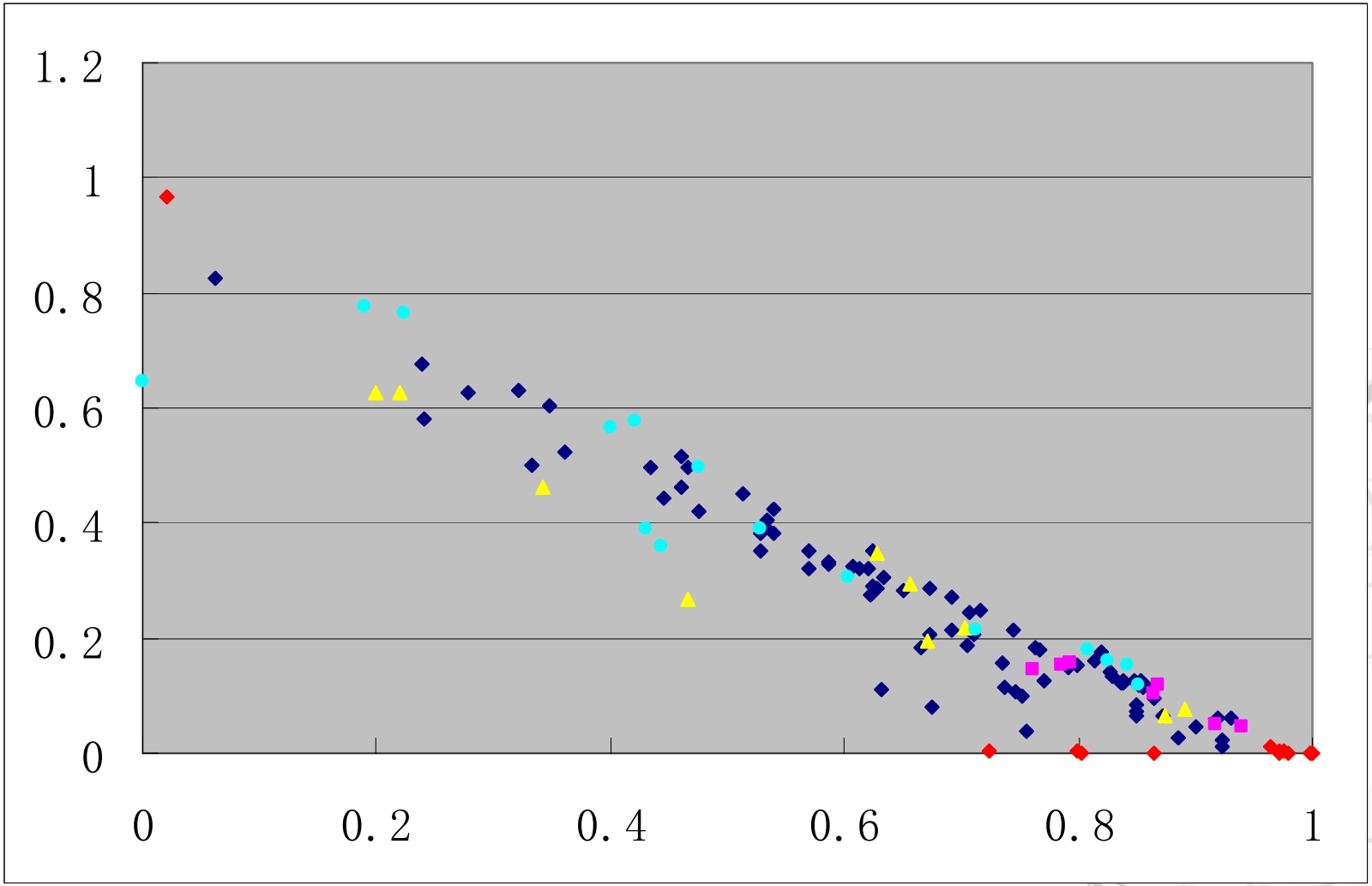

zh.


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
11
机器人
管理员
用户
破坏者
辩论者

方法與結果

- 用戶的內容編輯傾嚮相對討論傾嚮作圖
- 用戶的討論傾嚮是用戶在所有talk頁面的編輯次數以及在Project編輯次數之和佔總編輯次數的比例
- 圖錶顯示這兩項數據之間有良好的負相關性

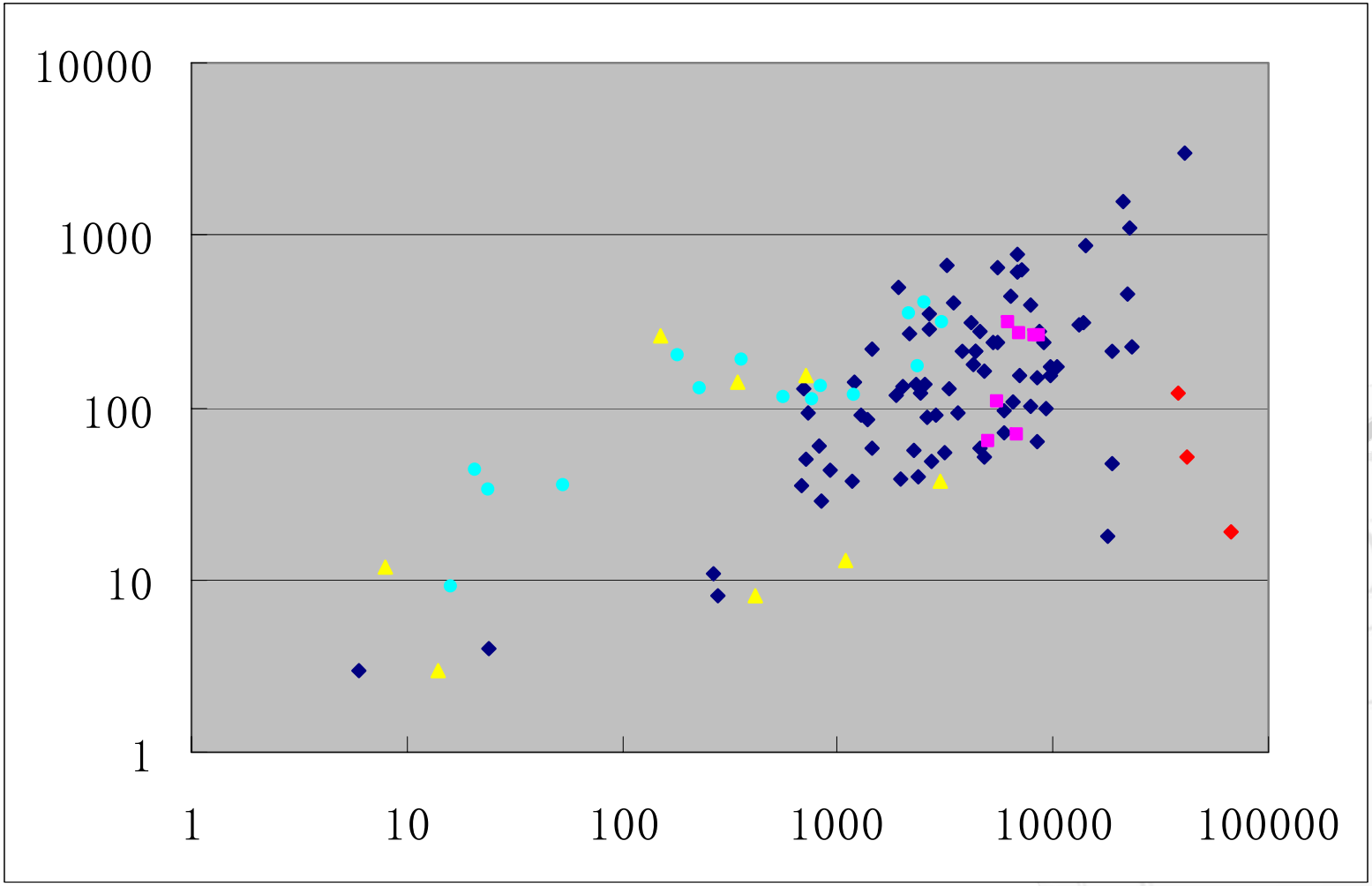

zh.


1.2
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1

方法與結果

- 用戶的條目編輯次數相對討論頁編輯次數作圖
- 坐標均採用對數坐標
- 一部分破壞者和機器人的討論頁編輯次數為0，在對數坐標中沒有體現

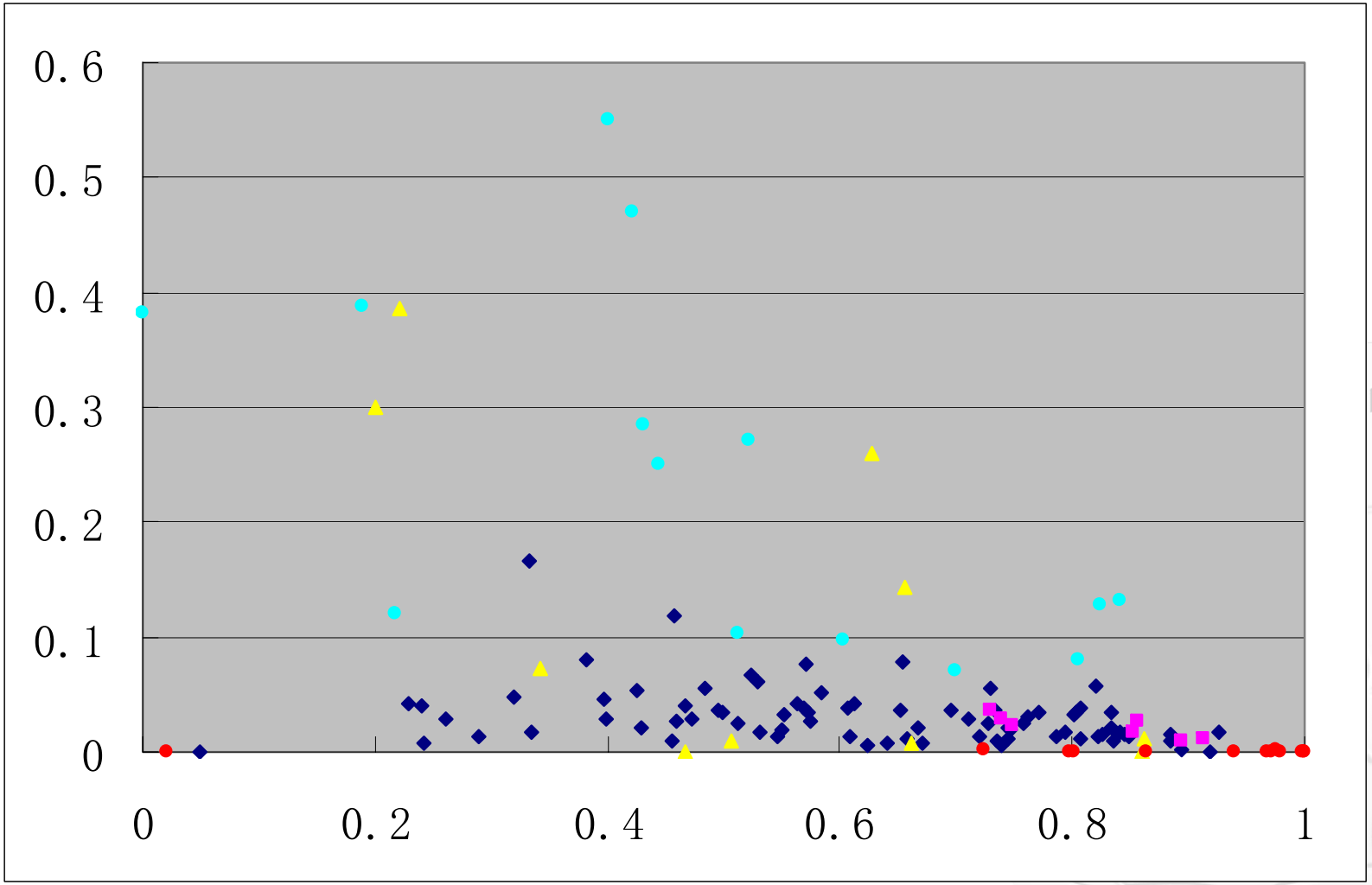

zh.


10000
1000
100
10
1
1
10
100
1000
10000
100000

方法與結果

- 用戶的相對條目編輯次數對相對討論頁編輯次數作圖
- 相對條目編輯次數和相對討論頁編輯次數是用戶的條目和討論頁編輯次數蘸個人總編輯次數之比例
- 在這副圖中，辯論者和破壞者顯示特別的分佈

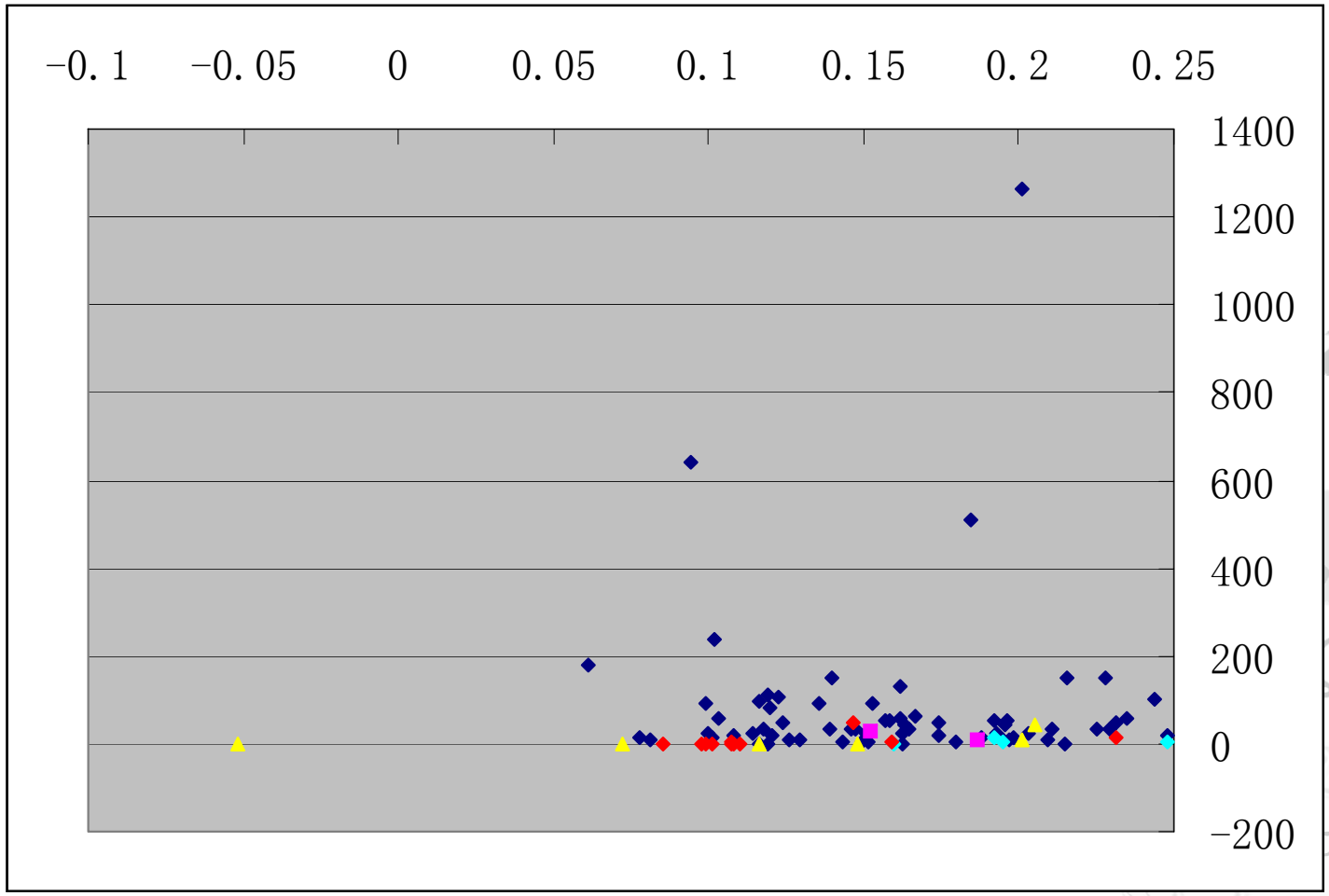

zh.


0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.2
0.4
0.6
0.8
1

方法與結果

- 對用戶進行主成分分析后第一第二第三主成分組成的綜閤主成分相對第四主成分作圖
- 在統計學中，主成分分析是一種簡化數據集的技術。它把數據變換到一個新的坐標系統中，使得任何數據投影的第一大方差在第一個坐標(稱為第一主成分)上，第二大方差在第二個坐標(第二主成分)上，依次類推。主成分分析經常用減少數據集的維數，同時保持數據集的對方差貢獻最大的特徵

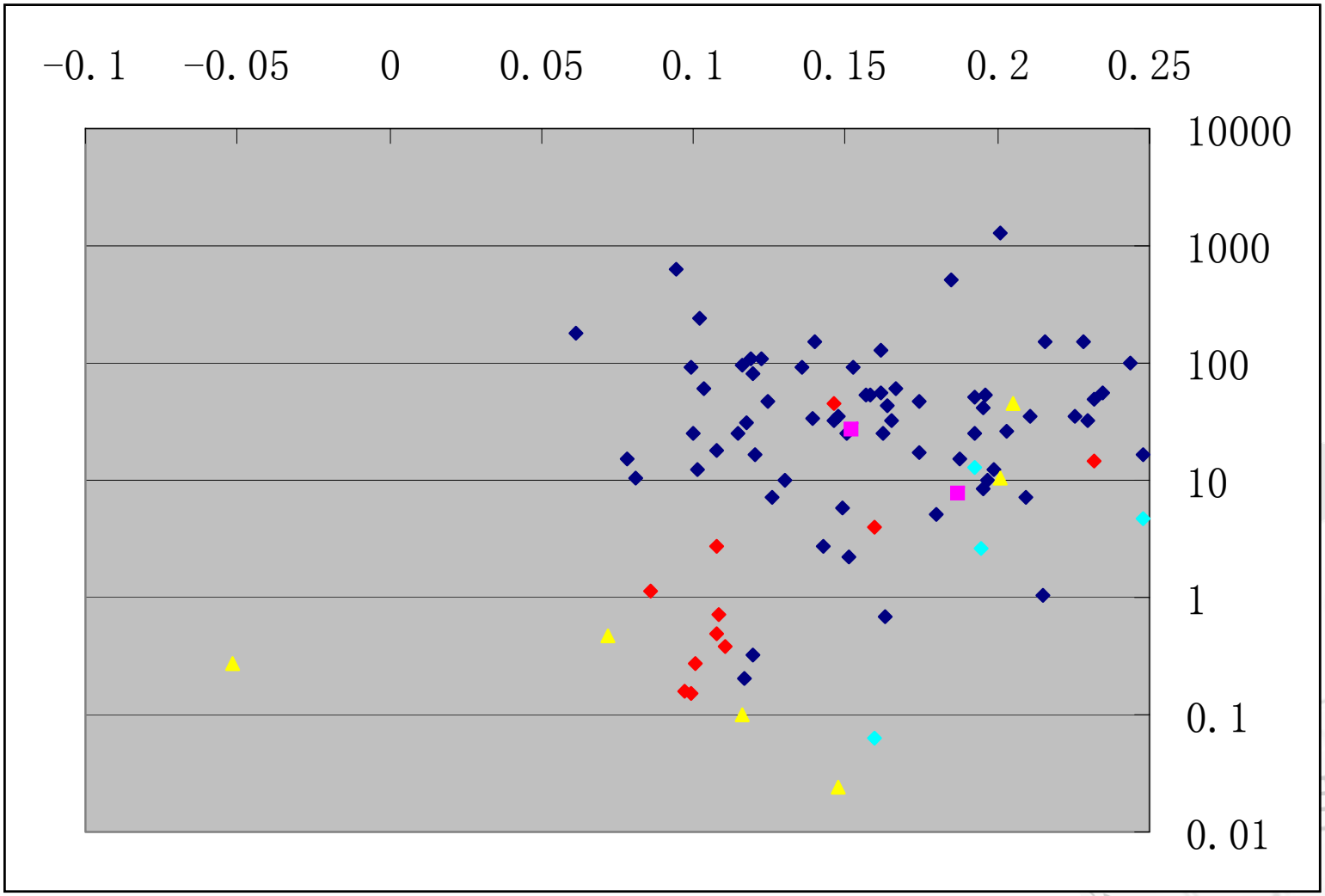

zh.


-0.1
-0.05
0
0.05
0.1
0.15
0.2
0.25
1400
1200
1000
800
600
400
200
0
-200

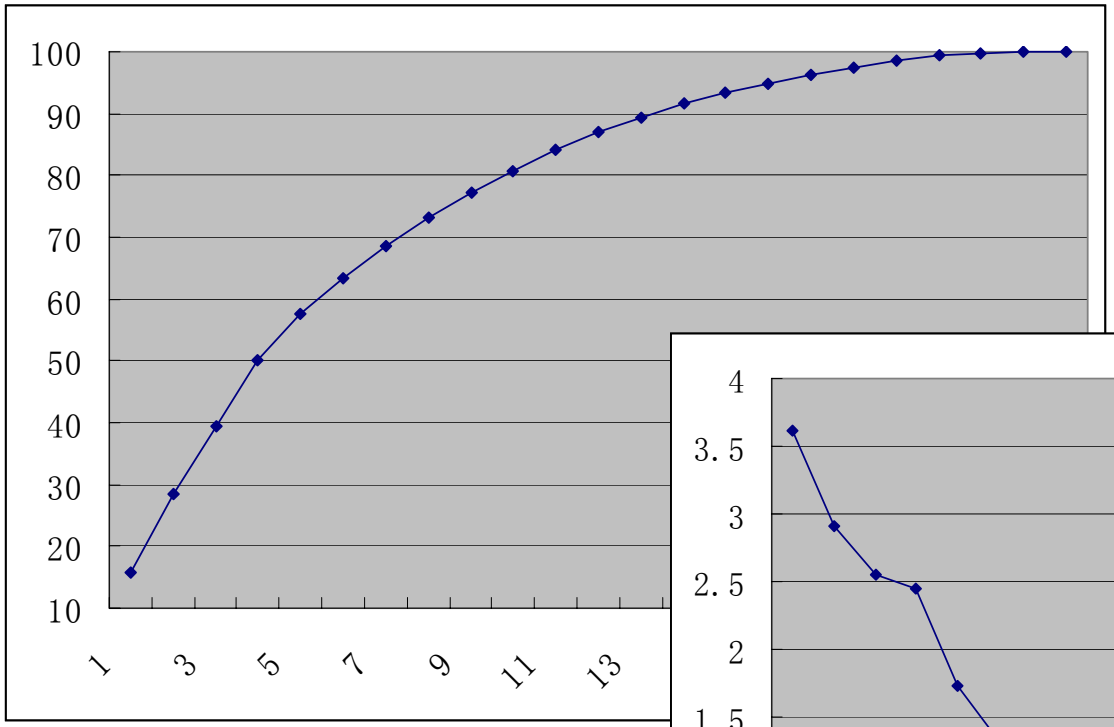
方法與結果

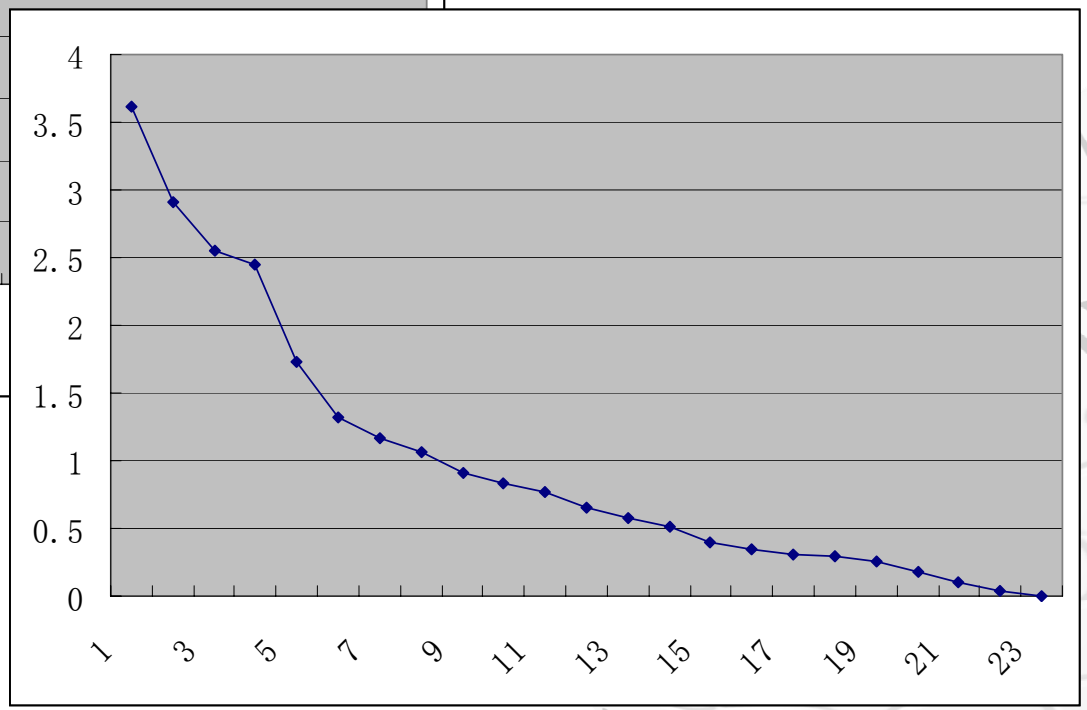
- 這是將縱軸變換為對數坐標軸之后的圖像


zh.


-0.1
-0.05
0
0.05
0.1
0.15
0.2
0.25
10000
1000
100
10
1
0.1
0.01


zh.


100
90
80
70
60
50
40
30
20
10
1
3
5
7
9
11
13


4
3.5
3
2.5
2
1.5
1
0.5
0
1
3
5
7
9
11
13
15
17
19
21
23

一點小小的討論

- 不同類型的用戶，編輯行為會有所差異
- 可以通過挖掘這些數據獲得關于用戶的更多信息
- 辩论者和机器人是编辑行为最具有特点的群体，破坏者有不同的类型
- 数据分析的结果显示管理员并非一个具有共同特征的群体
- 用主成分分析方法處理的結果並不理想，可以尝试其他多元统计分析方法

謝謝