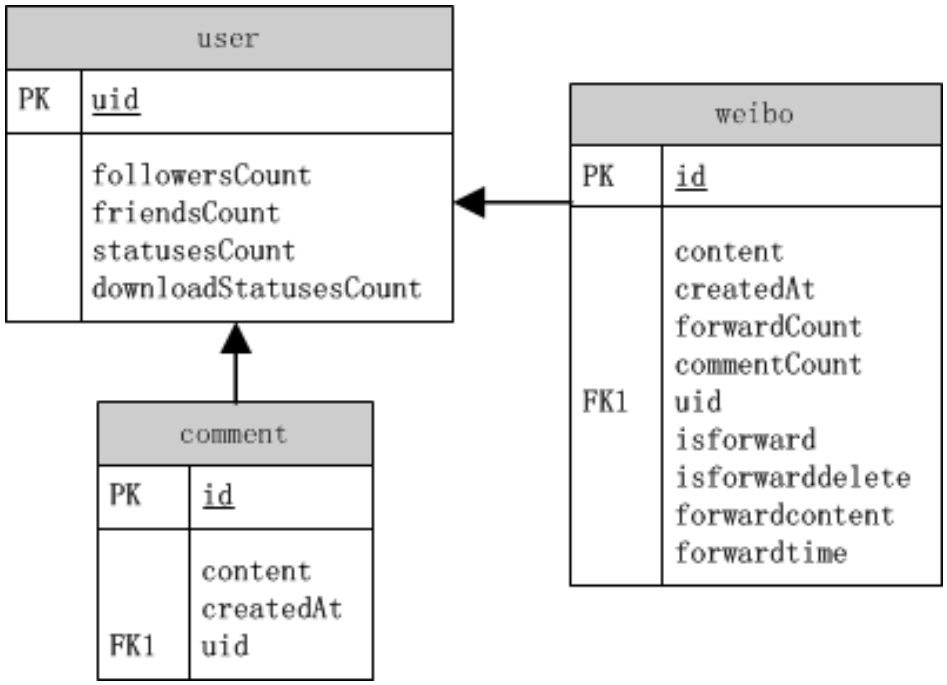
Implementation of Micro-blogging User Intention Modeling

Based on Behavior Analysis

Purpose: Determine who the real users are, versus the fake ones.
Data Source: The data came from Sina Weibo.
Method for Analysis: Classification.



What I did in this project:

1. Labeled the users according to these categories:
   (1) Users who like to share stories, mottos.
   (2) Users who use twitter to do marketing.
   (3) Users created to influence the public opinion.
   (4) Users created as fake fans.
   (5) Users who like to post their own ideas.
   (6) Users who are lazy at posting own ideas. Instead, they love to retweet or make comments.
   (7) Silent users.
   (8) Users own the merits of (5) (6).
   I labeled 506 users in total.
2. Gather statistical data by keyword matching and other calculation.
   (1) Total number of posts.
   (2) Percentage of original posts.
   (3) Percentage of retweets.
   (4) Number of comments / Number of posts.
   (5) Percentage of posts that shares stories, mottos, etc.
   (6) Percentage of marketing posts.

(7) Number of followers.
(8) Number of followings.
(9) How often the user retweet same content.

3. Divide the labeled user into training set (323) and testing set (183).

Training Set

label	Number of users	Percentage
1	25	7.740%
2	17	5.263%
3	21	6.502%
4	40	12.384%
5	33	10.217%
6	22	6.811%
7	136	42.105%
8	29	8.978%

Testing Set

label	Number of users	Percentage
1	8	4.372%
2	16	8.743%
3	9	4.918%
4	20	10.929%
5	16	8.743%
6	11	6.011%
7	83	45.355%
8	20	10.929%

4. SVM: Overall precision is 78.69%.

label	**True positive rate(TP/P)**	**Precision (TP/(TP+FP))**
1	4/8=50.00%	4/15=26.67%
2	7/16=43.75%	7/7=100.00%
3	8/9=88.89%	8/14=57.14%
4	18/20=90.00%	18/20=90.00%
5	12/16=75.00%	12/18=66.67%
6	5/11=45.45%	5/11=45.45%
7	82/83=98.80%	82/83=98.80%
8	8/20=40.00%	8/15=53.33%

5. C5.0 Decision Tree: Overall precision is 78.69%.

label	**True positive rate(TP/P)**	**Precision (TP/(TP+FP))**
1	3/8=37.50%	3/14=21.43%
2	11/16=68.75%	11/15=73.33%
3	9/9=100.00%	9/10=90.00%
4	20/20=100.00%	20/25=80.00%
5	13/16=81.25%	13/20=65.00%
6	7/11=63.64%	7/8=87.50%
7	79/83=95.18%	6/13=46.15%
8	6/20=30.00%	82/83=98.80%

6. I also develop a Java demo, which is responsible for transforming the data into the input data and analyze those data by Libsvm.