Sungkyunkwan University

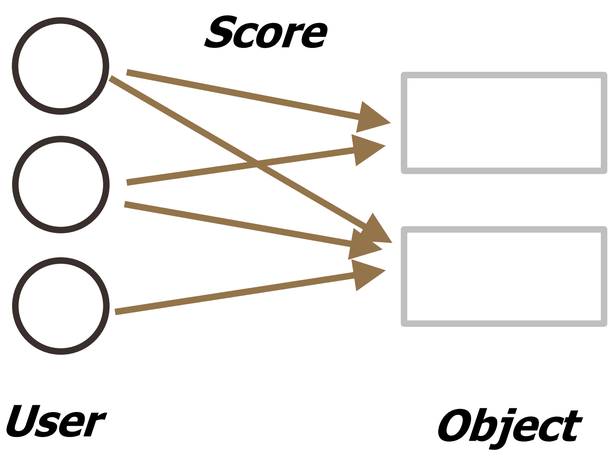
Deviation-based spam filtering method in online ranking system

Daekyung Lee[1], Beom Jun Kim[1]

[1] *Department of Physics, Sungkyunkwan University, Suwon 440-746, Republic of Korea*

Background

Raiting system



- Users evaluate objects and assign discrete scores (1~5).
- Quality of object is determined by average score.
- Vulnerable to distortion by spammer.
- Reputation-based method to filter out spammers.

Group-based Ranking system[1)] (GR method)

- i : User index
- w_{ia}: Score of object a by i
- a :Object index $\{\alpha, \beta, \gamma, \cdots\}$
- k_i: Number of objects evaluated by i

Fraction table of User i

w \ a	α	β	γ
1	0.2	0.1	0.3
2	0.3	0.25	0.1
3	0.1	0.15	0.1
4	0.15	0.4	0.5
5	0.25	0.1	0

$$\vec{f_i} = (f_{\alpha 2}, f_{\beta 4}, f_{\gamma 4}) = (0.3, 0.4, 0.5)$$

$$R_i = \frac{\langle \vec{f_i} \rangle_a}{\Delta_a \vec{f_i}} = \frac{0.4}{0.1} = 4$$

- f_{aw}: fraction of users who gave score w to object a.
- $S_{ia} = \sum_w f_{aw} \delta_{w,w_{ia}}$: f_w that user i obtained in object a
- $R_i = \frac{\langle S_{ia} \rangle_a}{\Delta_a S_{ia}}$

Deviation-based Ranking method(DR method)

Fraction table of User i

w \ a	α	β	γ
1	0.2	0.1	0.3
2	0.3	0.25	0.1
3	0.1	0.15	0.1
4	0.15	0.4	0.5
5	0.25	0.1	0

$$R_i = \frac{(0.84+0.83+0.135)}{3} = 0.512$$

$$R_{i\alpha} = \frac{0.3 - \langle f_{a2} \rangle_a}{\Delta_a f_{a2}} = 0.84$$

$$R_{i\beta} = \frac{0.5 - \langle f_{a4} \rangle_a}{\Delta_a f_{a4}} = 0.83$$

$$R_{i\gamma} = \frac{0.1 - \langle f_{a5} \rangle_a}{\Delta_a f_{a5}} = -0.135$$

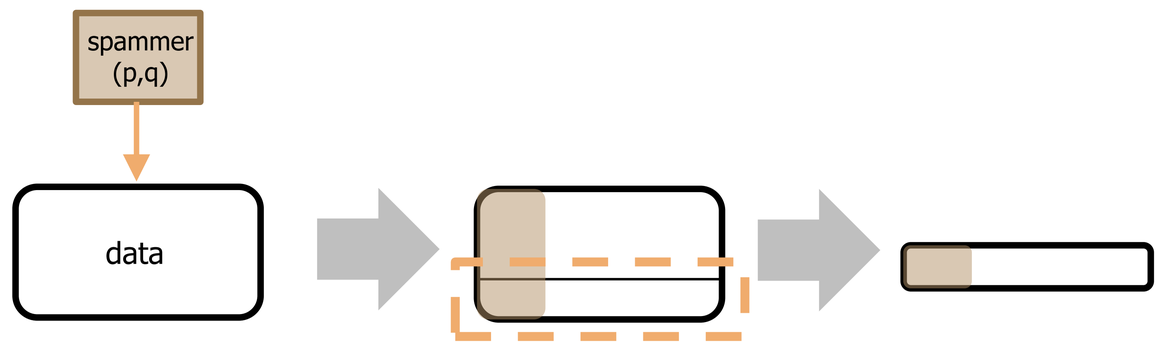
- $S_{iaa'} = \sum_w f_{aw} \delta_{w,w_{ia'}}$
- $R_{ia'} = \frac{S_{ia'a'} - \langle S_{iaa'} \rangle_a}{\Delta_a S_{iaa'}}$
- $R_i = \langle R_{ia'} \rangle_{a'}$

Method of verification

Spammer

- Malicious: always give 1(minimum) or 5(maximum) score 50:50.
- Random: always give random score 1~5.
- p = fraction of spammer in total users.
- q = fraction of spammer evaluation in total object.

Recall



Generate spammer in real data

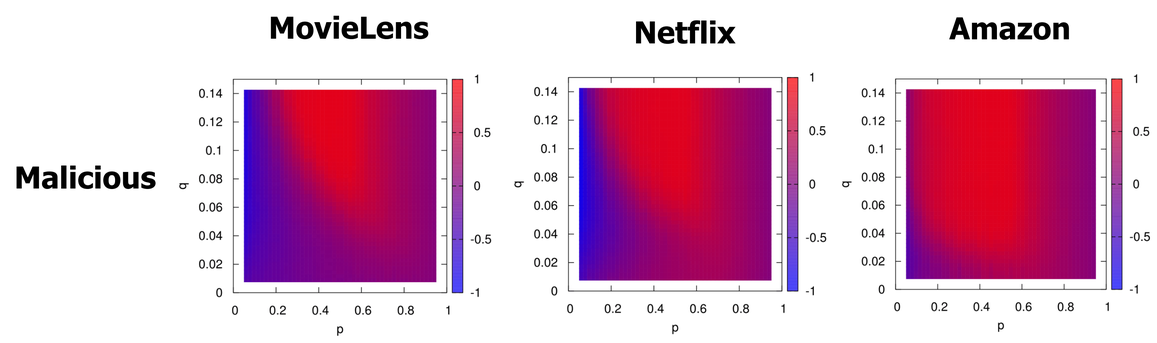
Calculate R_i and remove fraction p of low R_i users.

$$R_c = \frac{removed\ spammers}{removed\ users}$$

Results

Data Set	# of Users	# of Objects	$\langle k_i \rangle$
MovieLens	943	1682	106
Netflix	1038	1215	47
Amazon	662	1500	36

R_c in DR - R_c in GR



Summary&Conclusion

- We suggested a new spam-filtering method (DR) based on GR method.
- DR method shows better performance than GR method in large p region.
- Our next goal: more precise finding of statistical properties of spammers

1) Gao J., Dong Y.-W., Shang M.-S., Cai S.-M. and Zhou T., EPL, 110 (2015) 28003.