Detecting Spam Reviews on the Chinese Online Shopping Site TaoBao

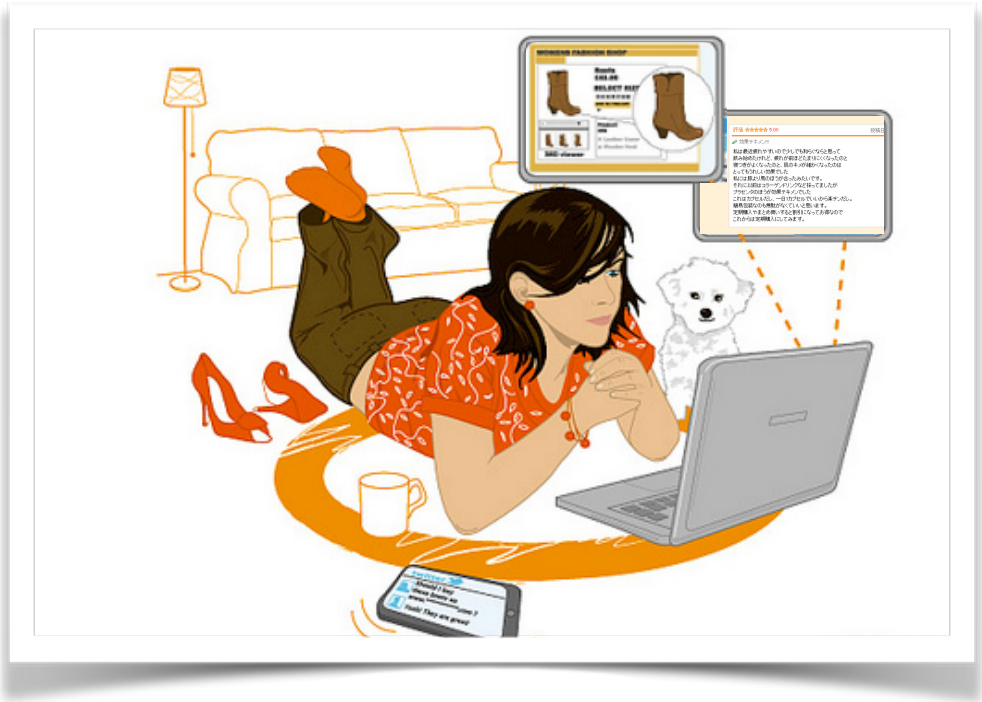
Shinshin Ryu, Fumito Masui, Michal Ptaszynski
Department of Computer Science,
Kitami Institute of Technology, Japan

Outline

- **Introduction**
- Data set
- Review Analysis
- Classification Experiment
- Conclusions

Introduction

- Various online shopping sites
- Spam reviews
- “Internet Water Army” in China



Introduction

- Related Work:

review content

spammer's behavior

- Related Chinese research problem:

→the lack of publicly available Chinese data sets

→the low accuracy and reliability of training data

→the insufficient amount of public user information on Chinese shopping sites

…

Introduction

- In our research:

Chinese fake review detection method
→stealth marketing
→14 product parameters
→SVM (Support Vector Machine)

Outline

- Introduction
- **Data set**
- Review Analysis
- Classification Experiment
- Conclusions

Data set

- Data Set

https://world.taobao.com/



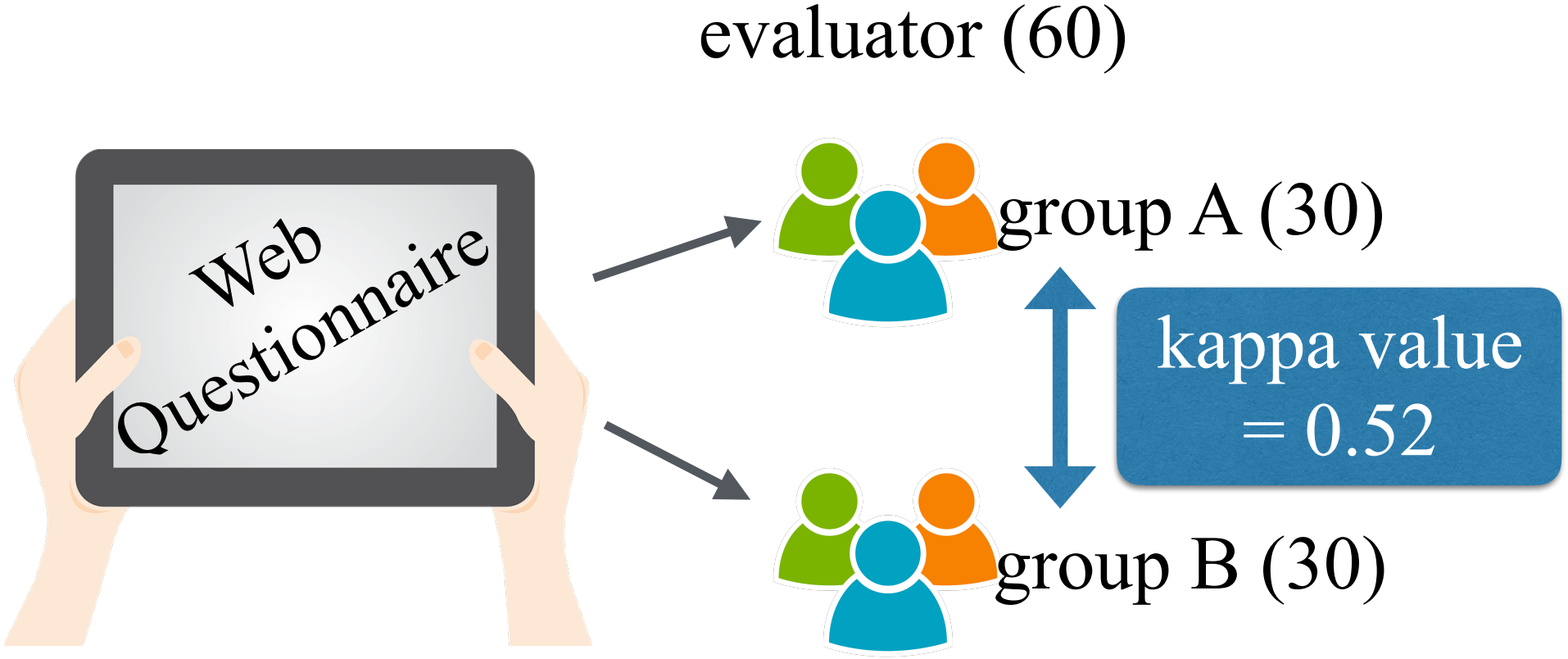
total review: 200 fake review: 31.5%

test data: 20 training data: 180

→10-fold cross-validation

Data set

- Labeling



Outline

- Introduction
- Data set
- **Review Analysis**
- Classification Experiment
- Conclusions

Review Analysis

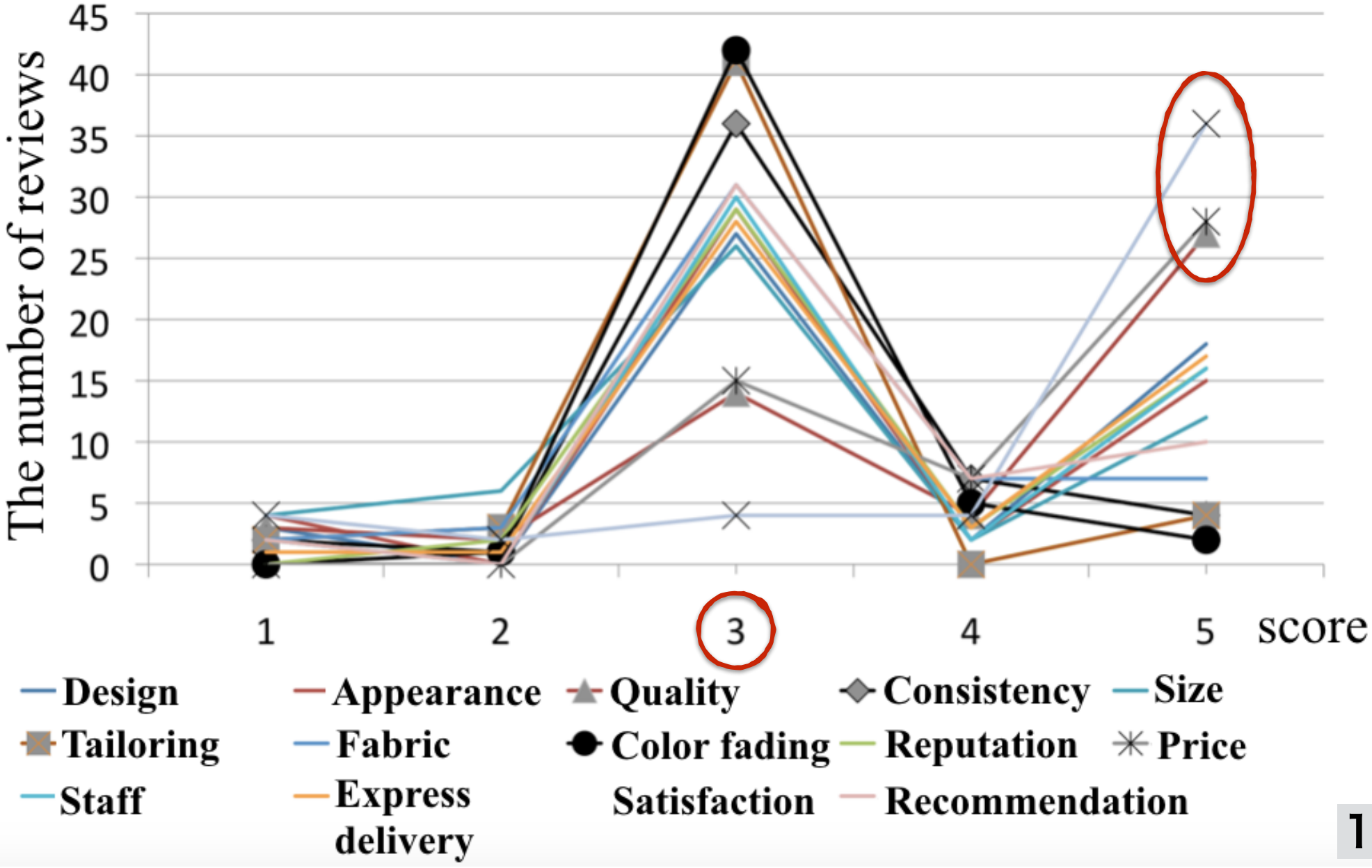
- Product Parameters:

Design, Appearance, Quality, Consistency, Size, Tailoring, Fabric, Color fading, Reputation, Price, Staff, Express delivery, Satisfaction, Recommendation

5	very satisfied
4	satisfied
3	neither
2	dissatisfied
1	very dissatisfied

Review Analysis

- The scoring results on apparel products



Review Analysis

- The ratio of fake reviews in non-confidence interval

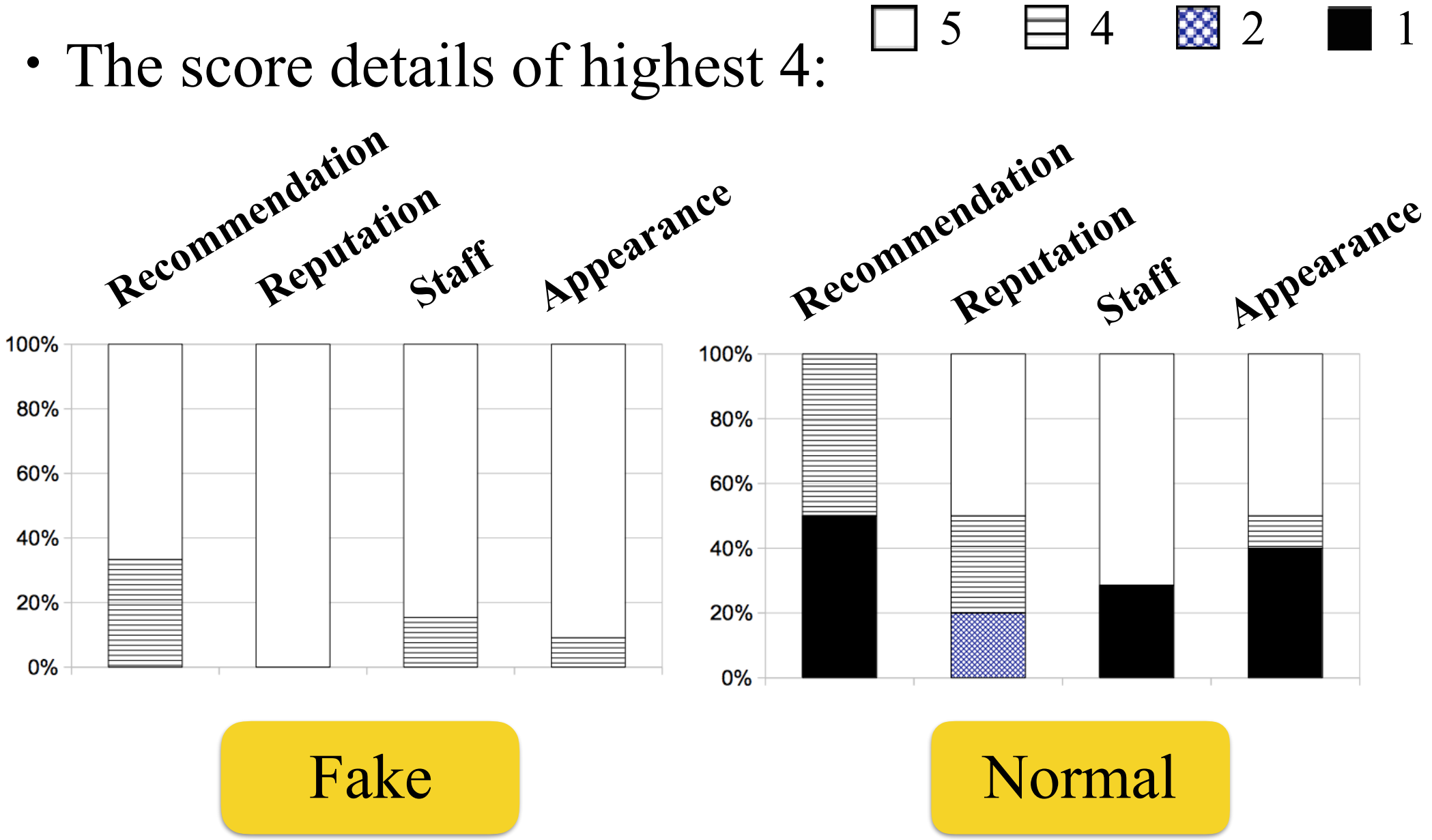
95%

4 highest

Parameter	Proportion
Recommendation	**0.83**
Reputation	**0.61**
Staff	**0.61**
Appearance	**0.53**
…	…
Fabric	0.42
Size	0.36
Price	0.20
Quality	0.00

Review Analysis

- The score details of highest 4:



Fake

Normal

Review Analysis

- The features of product parameters:

A. Recommendation

→“快快入手”(to buy quickly),

→“强烈推荐”(highly recommended)

B. Quality & Price

→“物美价廉”, “物超所值”(really good and cheap)

C. Staff

→“服务态度”(attitude of the staff and services)

D. Size

→the numbers

Outline

- Introduction
- Data set
- Review Analysis
- **Classification Experiment**
- Conclusions

Classification Experiment

- SVM (Support Vector Machine)
- two experiments

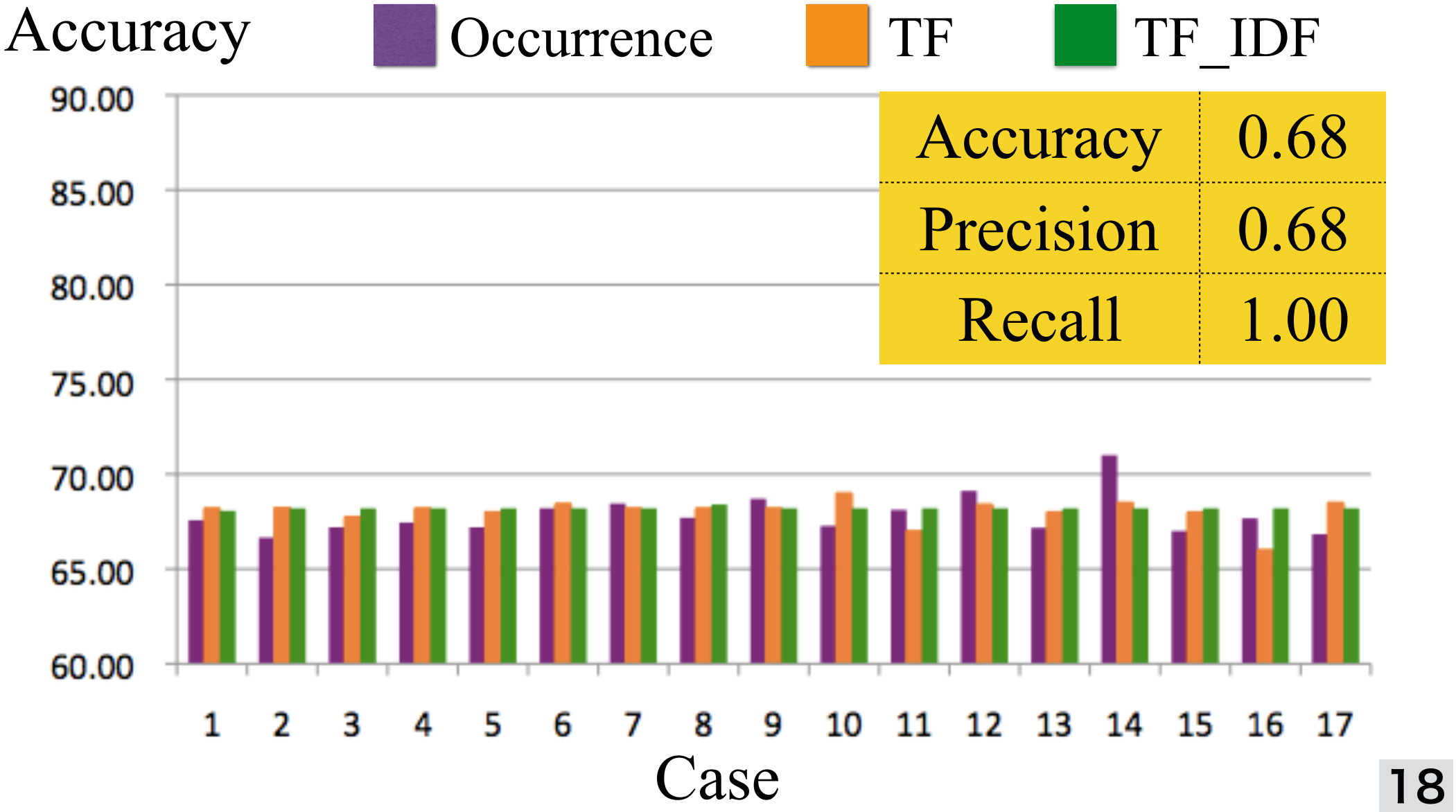
  →experiment 1: word & product parameters
  →experiment 2: product parameters

Classification Experiment

Case	Feature
1	Bag-of-Words
2	Staff, Quality&Price , Size, Recommendation *
3	Staff, Quality&Price, Size, Recommendation
4	Quality&Price, Size, Recommendation
5	Staff, Size, Recommendation
6	Staff, Quality&Price, Recommendation
7	Staff, Quality&Price, Size
8	Staff, Quality&Price
9	Staff, Size
10	Staff, Recommendation
11	Quality&Price, Size
12	Quality&Price, Recommendation
13	Size, Recommendation
14	Staff
15	Quality&Price
16	Size
17	Recommendation

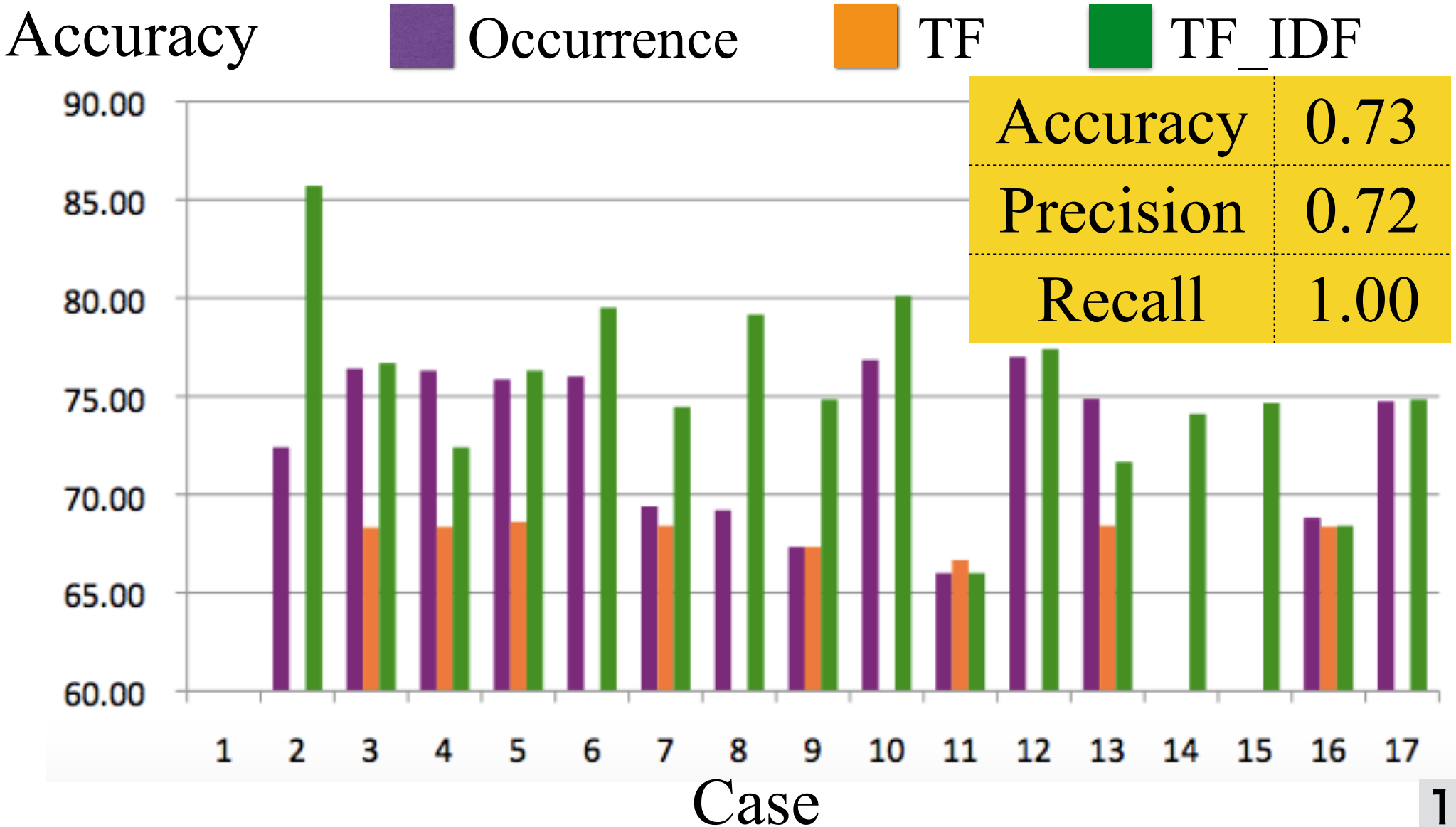
Classification Experiment

- experiment 1(word & product parameter)



Classification Experiment

- experiment 2 (product parameter)



Accuracy	0.73
Precision	0.72
Recall	1.00

Classification Experiment

- Accuracy of experiment 2 in a descending order
→Weight : Occurrence

Case	Feature	Accuracy
12	Quality & Price、Recommendation	77.00
10	Staff、Recommendation	76.85
3	Staff、Quality & Price、Size、Recommendation	76.40
4	Quality & Price、Size、Recommendation	76.30
6	Staff、Quality & Price、Recommendation	76.00
5	Staff、Size、Recommendation	75.85
13	Size、Recommendation	74.87
17	Recommendation	74.73
2	Staff、Quality & Price、Size、Recommendation*	72.40
7	Staff、Quality & Price、Size	69.40
8	Staff、Quality & Price	
16	Size	
9	Staff、Size	
11	Quality & Price、Size	66.00

Recommendation + α

Outline

- Introduction
- Data set
- Review Analysis
- Classification Experiment
- **Conclusions**

Conclusions

- The four features of product parameters:

  →Recommendation, Quality & Price, Staff, Size

- Two experiments:

  →experiment 1 (word & product parameters)
  →experiment 2 (product parameters) showed higher results
  →”Recommendation + α” had higher performance

Conclusions

- Future work

→Automatically increase the data and the feature set
→Considering of the different levels of features

Thank you