CIKM Competition 2014
Second Place Solution

Team: FAndy

Zhanpeng Fang, Jie Tang

Department of Computer Science

Tsinghua University

Task

- Given a sequence of query sessions
  - Example

    Class1 Query1 –

    Class1 Query1 Title1

    Class2 Query2 –

    Class2 Query2 Title2

    Class2 Query2 Title3

- Classify the class label of test queries

Challenges

- Encoding character
  - Only little prior knowledge can be used
- Heterogeneous data
  - Query, title, session information
- User search behavior
  - How to incorporate user search behavior to help the classification task?
- Unlabeled data
  - How to utilize the large scale unlabeled data?

Result

- 0.9245(public score)/0.9245(private score)
- 2nd place winner
- Achieve in 4 days, from Sep. 27th to Sep. 30th EST

Final LeaderBoard

Rank	Name	Best Quiz Score	Best Submit Time
1	topdata	0.9296	Sep 30 2014 23:59:15 (PDT)
2	FAndy	0.9245	Sep 30 2014 23:15:04 (PDT)
3	adfr	0.9222	Sep 30 2014 03:44:32 (PDT)
4	yingwei_xin	0.9220	Sep 30 2014 23:57:42 (PDT)

Our Approach

- Feature extraction
  - Bag of words
  - User search behavior
- Learning models
  - Logistic regression
  - Gradient boosted decision trees
  - Factorization machines
- Ensemble

Feature Extraction – Bag of Words

- Given a query Q
- One gram, two grams, last gram of Q
  - 0 -> 0.8452

Feature Extraction – Bag of Words

- Given a query Q
- One gram, two grams, last gram of Q
  - 0 -> 0.8452
- One gram, two grams of the clicked titles
  - 0.8452 -> 0.9091, top 12 in the leaderboard!

Feature Extraction – Bag of Words

- Given a query Q
- One gram, two grams, last gram of Q
  - 0 -> 0.8452
- One gram, two grams of the clicked titles
  - 0.8452 -> 0.9091, top 12 in the leaderboard!
- More bag of words features?
  - Queries in the same session of Q?
  - Titles in the same session of Q?

Feature Extraction – Bag of Words

- Given a query Q
- One gram, two grams, last gram of Q
  - 0 -> 0.8452
- One gram, two grams of the clicked titles
  - 0.8452 -> 0.9091, top 12 in the leaderboard!
- More bag of words features?
  - Queries in the same session of Q?
  - Titles in the same session of Q?
  - Performance decreases, 0.9091 -> 0.89x
  - How to use the session information?

Feature Extraction – Search Behavior

- Given a query Q
- Macro features
  - #total search, average length of clicked titles, length of the query
  - 0.9091 -> 0.9105

Feature Extraction – Search Behavior

- Given a query Q
- Macro features
  - #total search, average length of clicked titles, length of the query
  - 0.9091 -> 0.9105
- Session class features
  - For each potential class C, calculate:
    - #class C queries in the same session
    - #class C queries in the next/previous query
  - 0.9105 -> 0.9145

Feature Extraction – Search Behavior

- Same session's queries can help but might contain noises

Feature Extraction – Search Behavior

- Same session's queries can help but might contain noises
- Only use similar queries!
- Same session's queries feature
  - Bag of words feature for same session's queries that are similar to the query Q
  - Use Jaccard to measure similarity between queries
  - 0.9145 -> 0.9182, utilizing the large scale unlabeled data!

Feature Extraction – Search Behavior

- Further add clicked titles of same session's similar queries
  - Performance decrease, 0.9182 -> 0.9176

Learning Models

- Logistic regression
  - Use the implementation of Liblinear
- Factorization machine
  - Use the implementation of LibFM
- Gradient boosted decision trees
  - Use the implementation of XGBoost

Method	Implementation	Score on leaderboard
Logistic Regression	Liblinear	0.9182
Factorization Machine	LibFM	0.9151
GBDT	XGBoost	0.9225

Ensemble

- Ensemble prediction results from different models by logistic regression

Method	Implementation	Score on Validation
Logistic Regression	Liblinear	0.9182
Factorization Machine	LibFM	0.9151
GBDT	XGBoost	0.9225
Ensemble	Liblinear	0.9245

- Ensemble can significantly improves the performance

Summary

- "Tricks" on how to win 2nd place
  - Use unlabeled data
  - Train multiple models
  - Ensemble different results

Thank you!

Questions ?