Suiyuan Zhang

a_dayuanzi@163.com · (+86) 136-5115-6109

Education Background

Institute of Computing Technology, Chinese	Key Laboratory of Network Data	*Master*	2015 – Present
Wuhan University of Technology	Computer Science	*Scholar(2/110)*	2011 – 2015

Awards

Intelligent Traffic Prediction Challenge	1 / 1716		2017.07 – 2016.08
Kaggle Quora Pair Question Match Contest	31 / 3304		2016.05 – 2016.07
Ali Music Trends Forecasting Competition	16 / 5475		2016.05 – 2016.07
Churn Rate Prediction	1 / 200+		2016.07 – 2016.09
Hotel Production Forecast for the Next 30 Days	1 / 200+		2016.07 – 2016.09
National Collegiate Mathematical Modeling Contest	First Prize	top 1%	2013.09

Internship

Didi Chuxing Map Department *Algorithm Engineer Intern* 2017.06 – 2017.08

- Be familiar with the framework of line recommendation, developed by pyspark in experimental data
- The link topological structure of road is established, and the traffic condition is excavated by GCN (graph convolutional neural network model)

Ping An Group of China Commercial Real Estate *Indoor Positioning Engineer* 2014.09 – 2015.01

Project Experience

Distributed Information Retrieval System **Core Member** 2016.09 – Present

Written by C ++, the underlying system of multiple systems

- Invert index data structure, supporting for real-time indexing, a new document can be retrieved for no more than 1 minutes
- Support node hot plug, load balancing, construction index exception handling, support document segmentation
- Update improving, delete efficiency, usie Thift framework for communication for transmission higher efficiency
- Use mmseg segmentation algorithm, based on cosine similarity as a basis to sort, use DBSCAN to cluster results

Ctrip Cloud - Division Tournament Big Data Contest **Captain** 2016.07 – 2016.09

To predict the daily demand of 500 hotels in February of 2016 according to the historical hotel night volume in one year.

- Abstract as two sub-problems, select the hotel attributes, do the of time window statistics features
- The double-split point based random forest algorithm is better than the normal random forest and XGBoost

Trec 2017 OpenSearch **Captain** 2017.04 – Present

A paper sorting system according to the paper search site CiteSeerX feedback information. Obtained the score of 0.133 (logloss), 31/3304 in Quora Match Contest of the same features and model in the just-concluded Kaggle platform

- Build TF-IDF matrix. Do the TSVD, NMF decomposition. Train LDA topic model.
- Train WordToVector, DocToVector and LDAToVector to convert sentences to vector.
- Use wordnet to extract the feature of synonyms, evaluating the similarity by editing distance, cosine distance, etc.
- Total of more than 2000 multi-dimensional features , we use LightGBM、FM、FFM to train.
- Use LSTM and CDSSM. To combine the models, use stacking. DBGD to meet the real-time demand

IT Competence

- Programming language: Python 、C++ 、Java

Others

- blog: http://blog.csdn.net/chedan541300521（XGBoost source reading notes）
- GitHub: https://github.com/dayuanzi（The double-split point based random forest algorithm）