浙江大學 ZHEJIANG UNIVERSITY

ICWS 2015

New York, USA June 27- July 2 2015

Time-aware API Popularity Prediction via Heterogeneous Features

Yao Wan[1], Liang Chen[1], Jian Wu[1] and Qi Yu[2]

[1]Zhejiang University, Hangzhou, China

[2]Rochester Institute of Technology, Rochester, USA

{wanyao, cliang, wujian2000}@zju.edu.cn, qi.yu@rit.edu

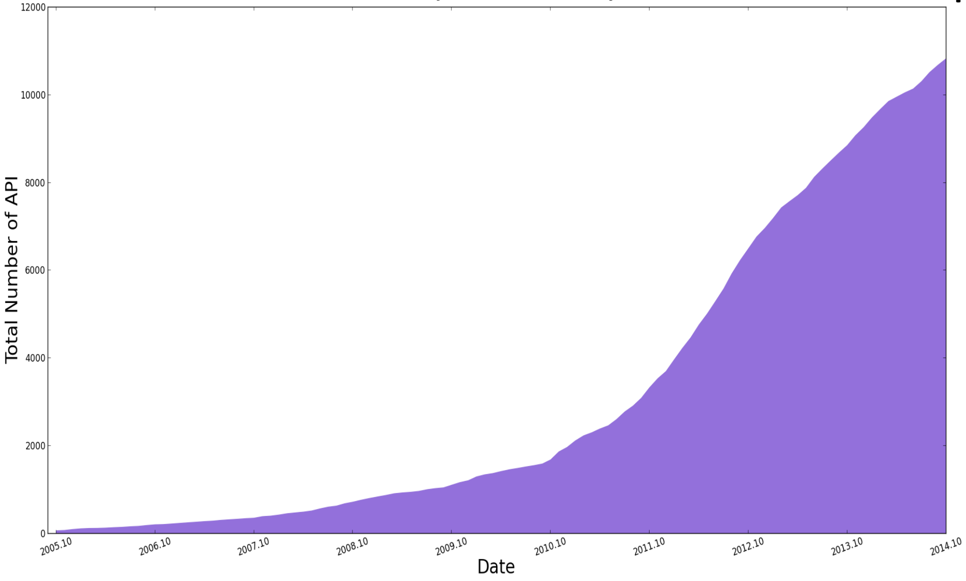
June 28, 2015

Outline

- Introduction
- Dataset Characteristics
- Popularity Prediction Models
- Experiments
- Conclusion & Future Work

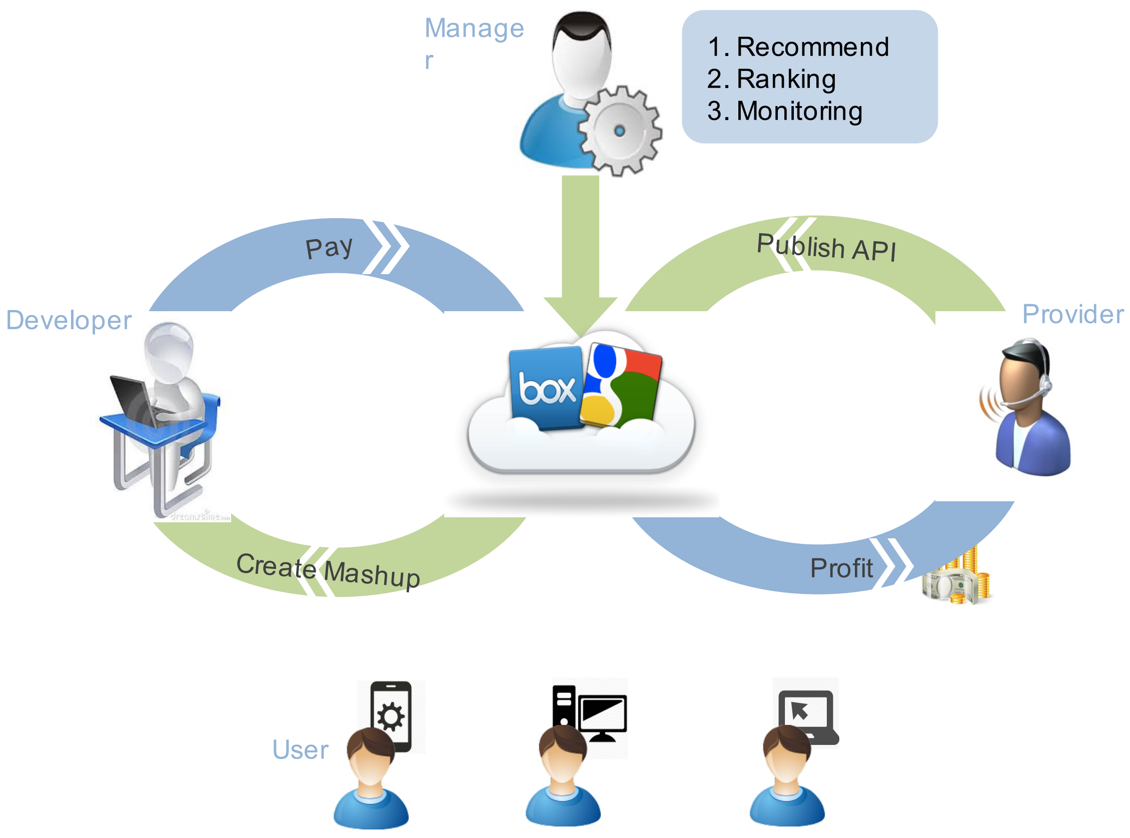
Introduction

- The spring up of web APIs
  - 10634 APIs and 6049 mashups in ProgrammableWeb until Nov. 2014 (www.programmableweb.com)
- The advantages of web APIs
  - Make web programmable
  - Increase ecnomic transactions from web browsers to API-driven interactions
  - eBay, APIs drive over 60% of its sellers; Twitter, 13 billion calls per day through APIs



Introduction

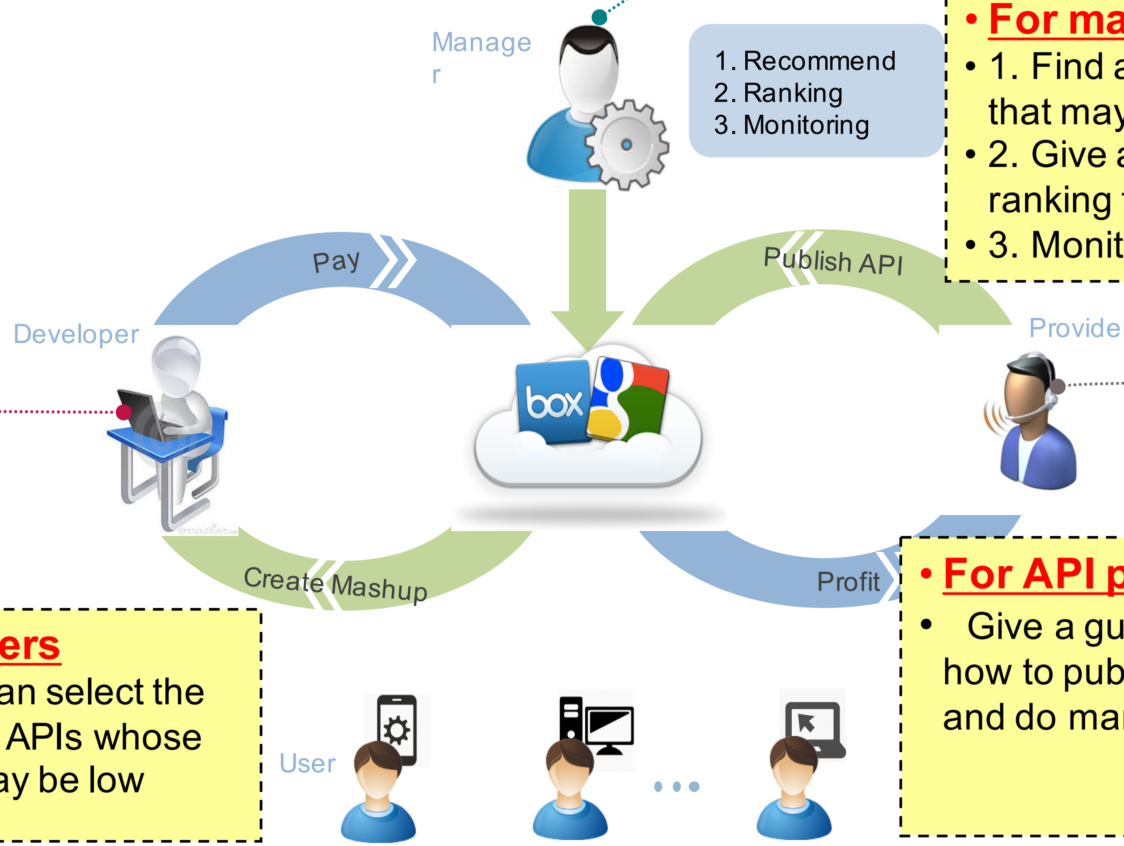
- **Framework of API market:** more and more Web sites dedicated to API market are emerging (e.g.Mashape)



Introduction

Motivation

1. With the increasing APIs, the need to search and manage APIs becomes urgent
2. Programmable manage APIs through the followers number at one monent



- **For manager**
- 1. Find and recommend the APIs that may be popular in the future
- 2. Give a more reasonable ranking to APIs
- 3. Monitor the trend of API growth

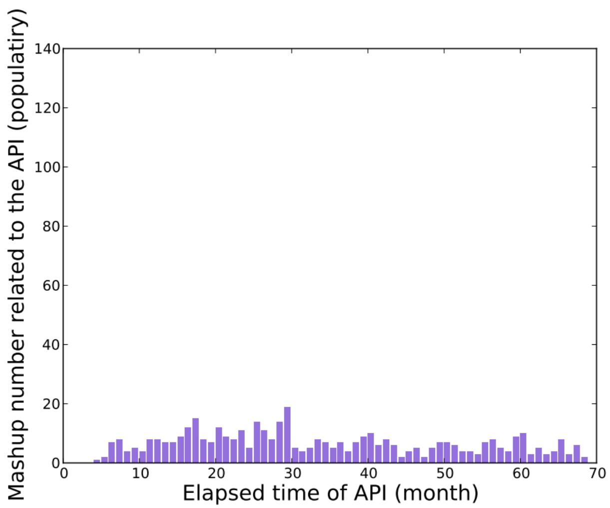
- **For developers**
- Developers can select the latest released APIs whose popularities may be low

- **For API providers**
- Give a guide to API providers that how to publish their APIs effectively and do marketing

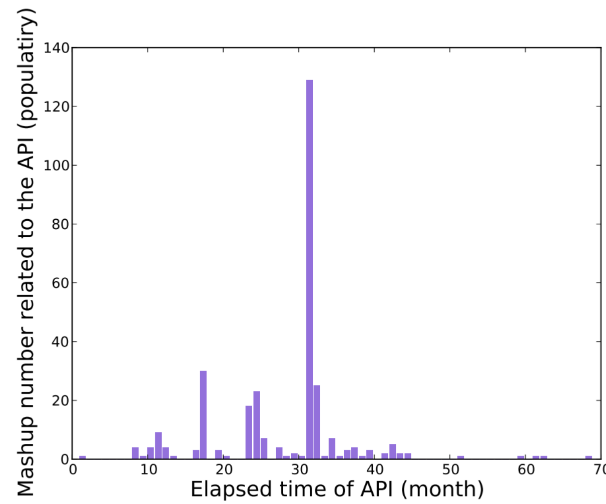
Introduction

Challenges

- The growth patterns of two APIs



Popularity of Youtube API



Popularity of Twilio API

Note: In our paper, the popularity of API is defined as the number of mashups that are composed of this API.

- Heterogeneous features

Outline

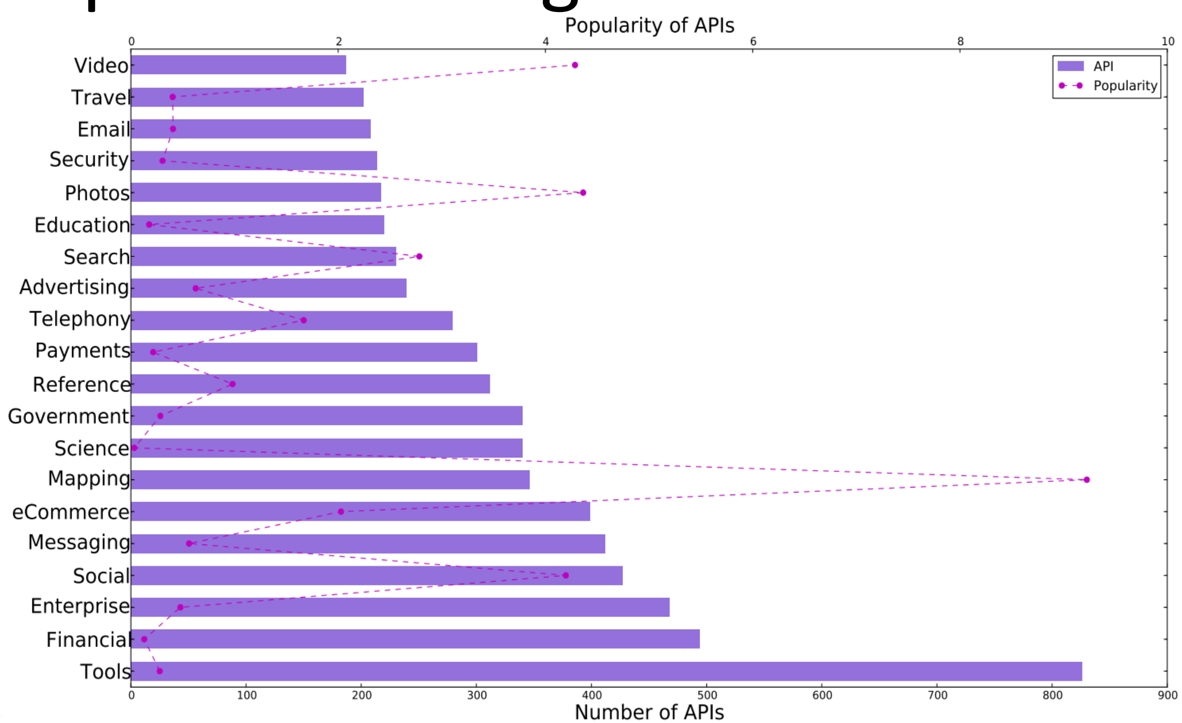
- Introduction
- Dataset Characteristics
- Popularity Prediction Models
- Experiments
- Conclusion & Future Work

Dataset Characterstics

- An overview of ProgrammableWeb data

Number of APIs	10,634
Number of mashups	6,049
Number of users	52,512

- Top-20 API categories

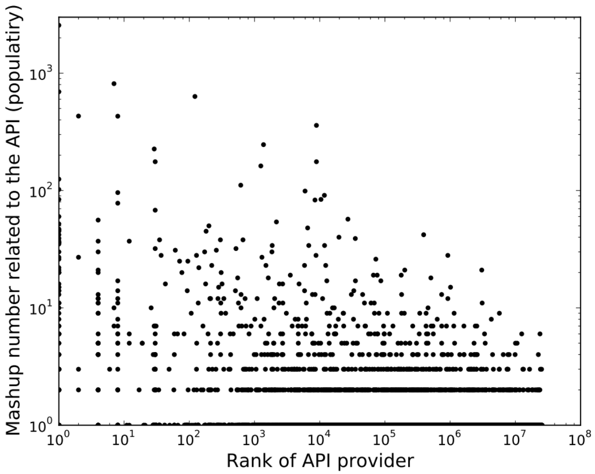


- **Popularity Ranking**

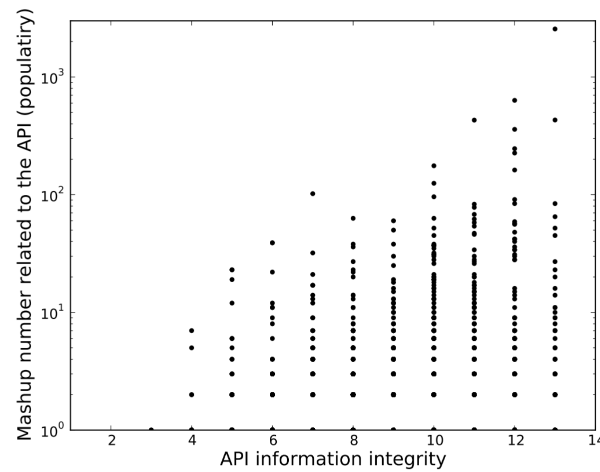
- 1. Mapping
- 2. Photos
- 3. Video
- 4. Search
- 5. Social

Dataset Characterstics

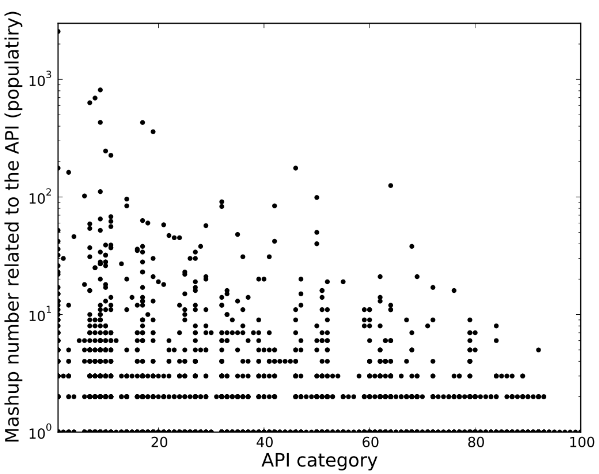
Statistical analysis



Rank of provider is obtained from Alexa.com



API information integrity means the degree of details providers describe the API



- **Observations**
- 1. The lower the rank of provider the lower of its popularity
- 2. The provider of API that with a higher information integrity is more prudent to release the API, the API may be of higher quality
- 3. The category of API has an effect on it's popularity

Outline

- Introduction
- Dataset Characteristics
- Popularity Prediction Models
- Experiments
- Conclusion & Future Work

Popularity Prediction Models

- Evaluation metrics

$$mRSE = \frac{1}{|C|}\sum_{a \in C}\left(\frac{\hat{N}(a, t_t|t_r)}{N(a, t_t)} - 1\right)^2$$

- Szabo-Huberman (S-H) model
  - Szabo and Huberman found a strong linear correlation between the early and future popularity

$$\hat{N}(a, t_t|t_r) = \alpha_{t_r,t_t} \cdot N(a, t_r)$$

$$\alpha_{t_r,t_t} = \frac{\sum_{a \in C} \frac{N(a,t_r))}{N(a,t_t)}}{\sum_{a \in C}\left(\frac{N(a,t_r))}{N(a,t_t)}\right)^2}$$

Popularity Prediction Models

- Linear regression (LR) model
  - Feature vector of API $X_{t_r}(a) = (x_1(a), x_2(a), \cdots, x_{t_r}(a))^T$
  - Optimization problem

$$\underset{\Theta_{(t_r,t_t)}}{argmin} \frac{1}{|C|} \sum_{a \in C} \left(\frac{\Theta_{(t_r,t_t)} \cdot X_{t_r}(a)}{N(a, t_t)} - 1 \right)^2$$

  - Let $X_v^* = \frac{X_{tr}(a)}{N(a,t_t)}$

$$\underset{\Theta_{(t_r,t_t)}}{argmin} \frac{1}{|C|} \sum_{a \in C} \left(\Theta_{(t_r,t_t)} \cdot X_v^* - 1 \right)^2$$

  - Can be solved by ordinary linear squares

Popularity Prediction Models

- Linear regression with heterogeneous features
  - Heterogeneous features
    - Time features
    - Numerical features
    - Categorical features
    - Textual features
- Optimization problem

$$\underset{\Theta_{(t_r,t_t)}}{argmin} \frac{1}{|C|} \sum_{a \in C} \left(\Theta_{(t_r,t_t)} \cdot X_v^* - 1\right)^2 + \lambda \left\|\Theta_{(t_r,t_t)}\right\|_2^2$$

where X_v^* is the feature vector and λ is the tunning parameter, $\|\Theta_{(t_r,t_t)}\|_2^2$ is the L_2 penality item.

Outline

- Introduction
- Dataset Characteristics
- Popularity Prediction Models
- Experiments
- Conclusion & Future Work

Experiments

Setup

- Dataset extraction
  - The target date t_t varies from 24 months to 48 months.
  - We extract the APIs whose releasing date is between 2005 and Nov. 2010
  - Discard those APIs whose popularity are alway zero
  - We get 613 APIs
- Experimental datasets
  - Full dataset
  - Popular dataset: APIs whose popularities are greater than 5
  - Junk dataset: APIs whose popularities are smaller than 5

Experiments

- Feature extraction
  - Numerical feagure
    - Normalization
  - Category feature
    - Encoded into binary code
  - Textual feature extraction
    - Case-folding and tokenization: tokenized by white space
    - Pruning: filter stopwords (e.g. *is, very, should*), keep adjectives using a part-of-speech tagger
    - Stemming: strip word to obtain the stem word
    - Spell correcting
- Training and testing
  - KFold cross validation
  - 50% traing + 25% crossvalidation +25% testing

Experiments

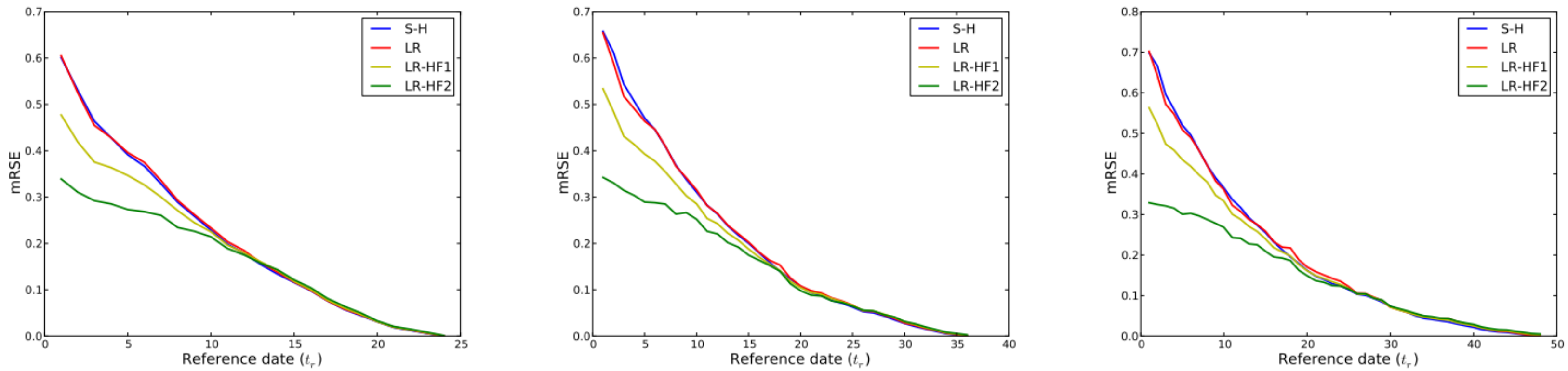
- Performance evaluation on full dataset

TABLE II: Model prediction errors (mRSE) with various t_r and t_t on the full dataset.

mRSE	Methods	$t_r = 1$	$t_r = 3$	$t_r = 5$	$t_r = 7$	$t_r = 9$	$t_r = 11$	$t_r = 13$
$t_t = 24$	S-H	0.3987	0.3236	0.2750	0.2283	0.1867	0.1525	0.1183
	LR	0.3944	0.3240	0.2783	0.2320	0.1894	0.1540	0.1195
	LR-HF1	0.3121	0.2684	0.2364	0.2049	0.1698	0.1448	0.1166
	LR-HF2	**0.2151**	**0.2010**	**0.1886**	**0.1732**	**0.1523**	**0.1326**	**0.1104**
$t_t = 36$	S-H	0.4483	0.3824	0.3404	0.2933	0.2537	0.2214	0.1919
	LR	0.4472	0.3838	0.3434	0.2976	0.2569	0.2217	0.1931
	LR-HF1	0.3792	0.3347	0.3017	0.2679	0.2313	0.2081	0.1875
	LR-HF2	**0.2470**	**0.2366**	**0.2228**	**0.2104**	**0.1933**	**0.1756**	**0.1605**
$t_t = 48$	S-H	0.4577	0.3990	0.3556	0.3153	0.2785	0.2494	0.2218
	LR	0.4584	0.4005	0.3564	0.3186	0.2835	0.2527	0.2265
	LR-HF1	0.3912	0.3553	0.3195	0.2905	0.2588	0.2377	0.2179
	LR-HF2	**0.2607**	**0.2507**	**0.2350**	**0.2241**	**0.2123**	**0.1973**	**0.1858**

Experiments

Performance evaluation on popular dataset



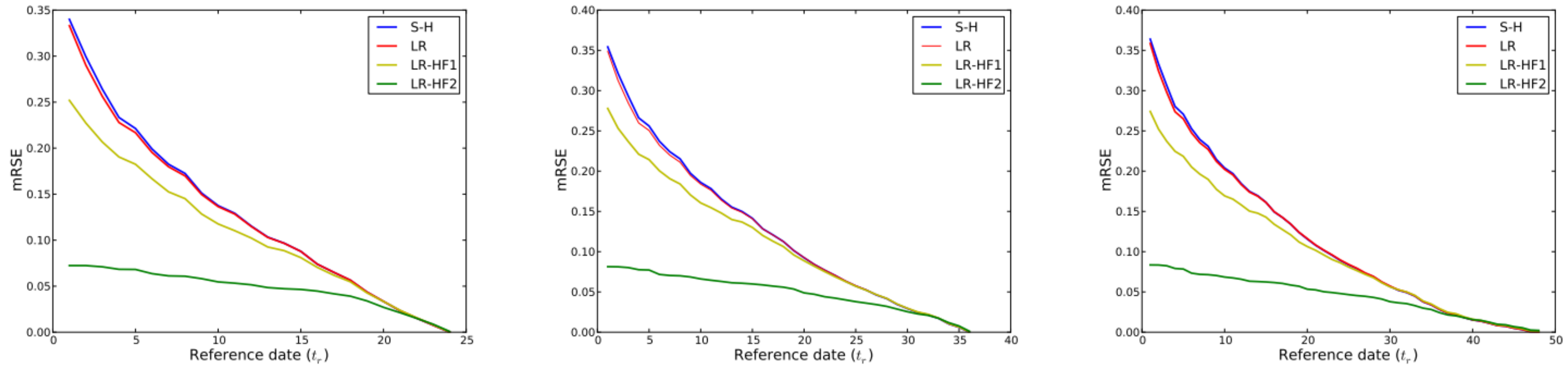
(a) t_t=24 months (b) t_t=36 months (c) t_t=48 months

•Observations

- 1. Our model has and obviously better performance over S-H and LR model, especially when the reference date is small
- 2. With the reference time becoming closer to the target time, the mRSE decreases for all models
- 3. With the reference time becoming closer to the target time, the advantages of our model shrink, S-H and LR model is enough

Experiments

Performance evaluation on junk dataset



(a) t_t=24 months (b) t_t=36 months (c) t_t=48 months

•Observations

- 1. Our LR-HF model still has an better performance
- 2. The improvement is small when just introducing API's textual features
- 3. Our model is particularly suitable for predicting the popularity of APIs that not changes frequently

Experiments

Impact of lambda

TABLE III: Impact of λ

mRSE	$\lambda = 0.0$	$\lambda = 0.2$	$\lambda = 0.4$	$\lambda = 0.6$	$\lambda = 0.8$	$\lambda = 1.0$
$t_r = 1$	0.25131	0.24748	**0.24746**	0.24776	0.24812	0.24846
$t_r = 5$	0.23292	**0.22910**	0.22921	0.22970	0.23026	0.23084
$t_r = 10$	0.19343	**0.18926**	0.18962	0.19081	0.19226	0.19379

- **Observations**
  - 1. The value of lambda really has an impact on the prediction accuracy
  - 2. The most appropriate value of lambda is between 0.2 and 0.4

Outline

- Introduction
- Dataset Characteristics
- Popularity Prediction Models
- Experiments
- Conclusion & Future Work

Conclusion& Future Work

- Conclusion
  - Crawl the APIs and mashups until Nov. 2014 from ProgrammableWeb.
  - Analyze some factors that may have an effect on the popularities of APIs
  - Propose an approach for predicting future popularities of APIs by integrating heterogeneous features
- Future work
  - Collect more time series records of APIs and find more features that may effect the popularity of APIs
  - Discove the popularity trends of APIs
  - Infer the impacts of other data sources on APIs, such as social media

Thank you!

wanyao@zju.edu.cn
http://www.wanyao.me