浙江大学 ZHEJIANG UNIVERSITY

PAKDD 2016

Auckland, New Zealand April 19- 22 2016

Incorporating Heterogeneous Information for Mashup Discovery with Consistent Regularization

Yao Wan[1], Liang Chen[2], Qi Yu[3], Tingting Liang[1] and Jian Wu[1]

[1]Zhejiang University, Hangzhou, China

[2]School of Computer Science & Information Technology, RMIT, Melbourne,

[3]Rochester Institute of Technology, Rochester, USA

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

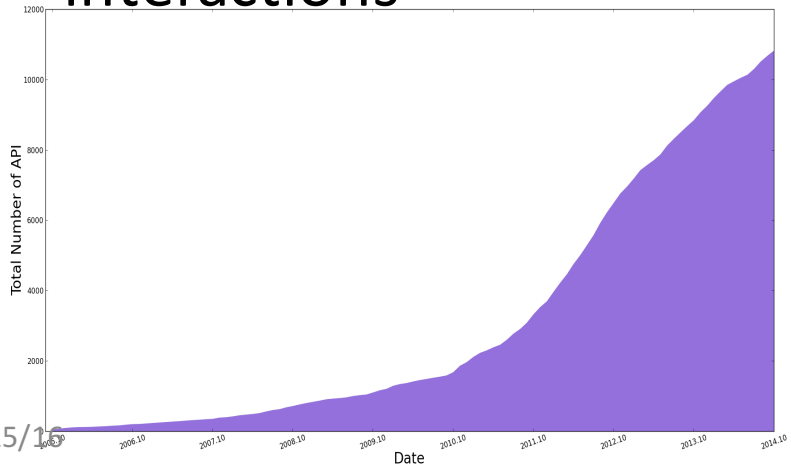
April 21, 2016

Outline

- Introduction
  - Background
  - Motivation
  - Challenge
- Model
  - A probabilistic model
  - Modeling the heterogeneous network
- Experiments
- Conclusion and Future Work

Background

- What is web API?
  - Application Programmable Interface make web programmable
  - RESTful service
  - 10,634 APIs and 6,049 mashups in ProgrammableWeb until Nov. 2014
  - Increase economic transactions from web browers to API-driven interactions

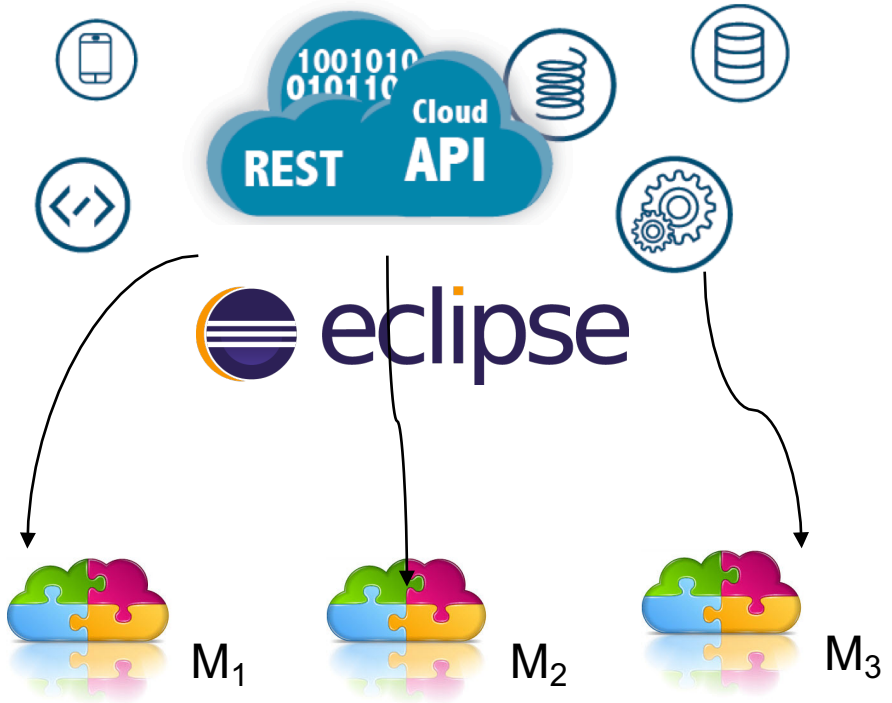


Background

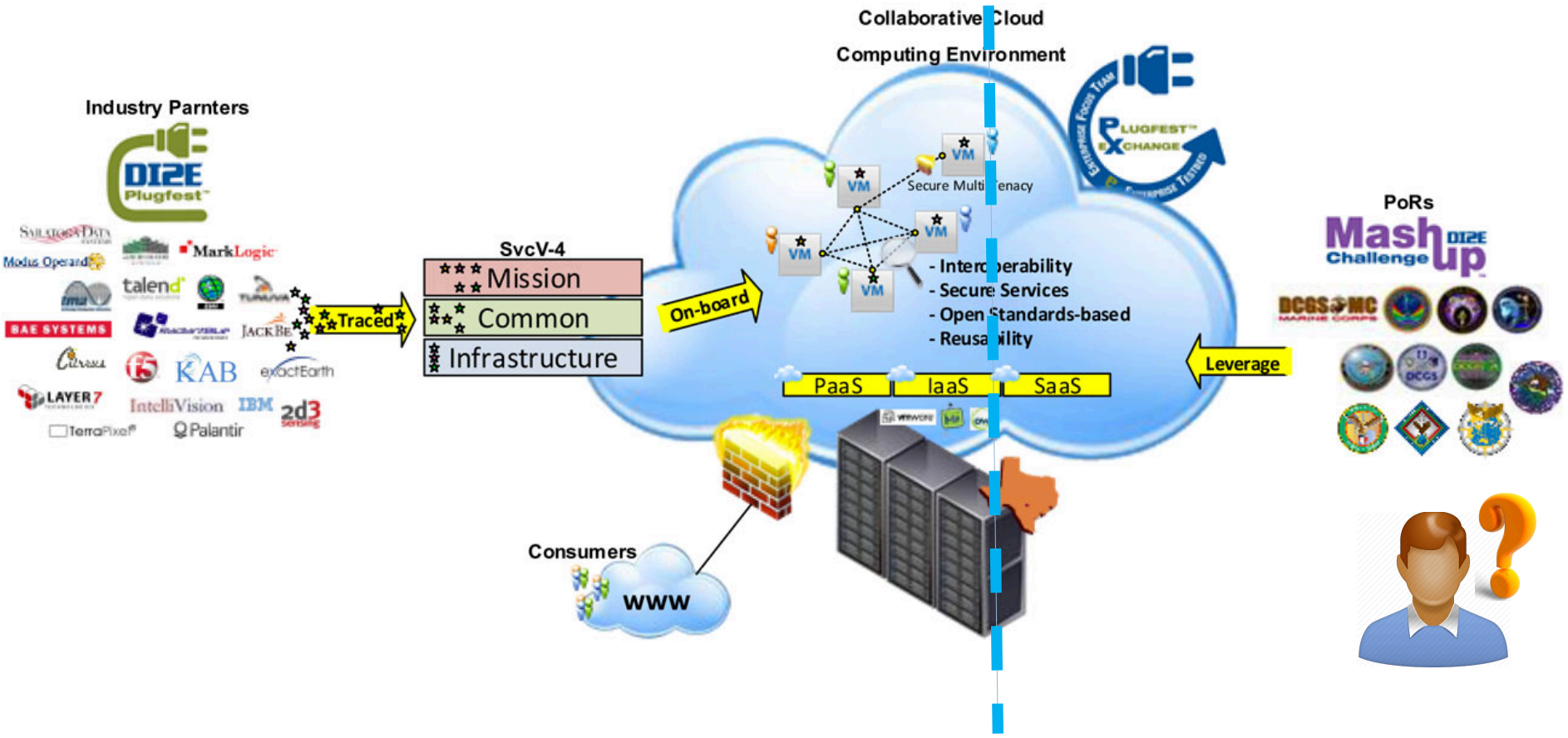
- What is Mashup? (Component Service)

  A mashup is lightweight web application created by combining capabilities from many other services

  - Rapid creation (days not months)
  - Reuses existing capabilities, but delivers new functions
  - Request less technical skills



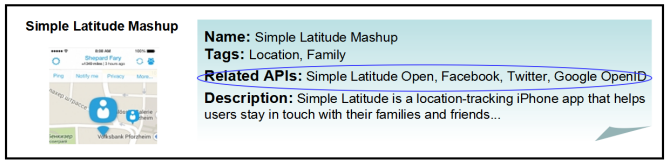
Motivation

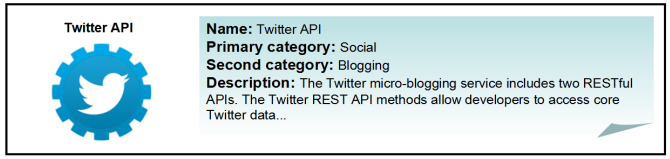


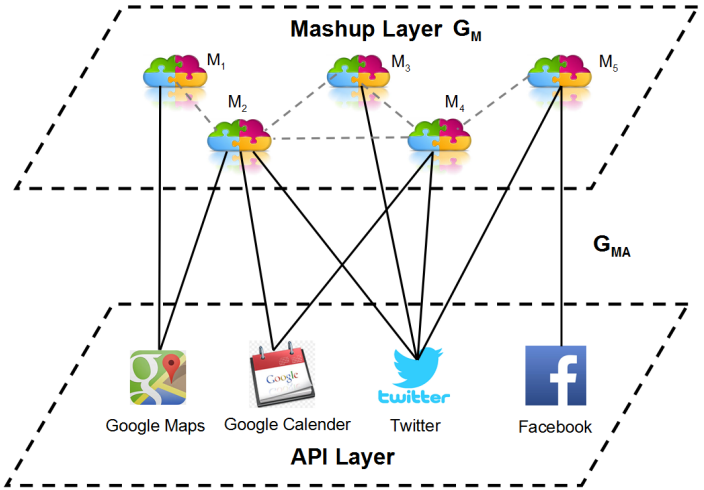
Task: discove the component services - mashup

Challenges

- Conventional methods
  - just utilize the semantic information of service itself
  - ignore the connection between services
- Challenge
  - Design a rank algorithm
  - Integrate the rank algorithm with the network structure





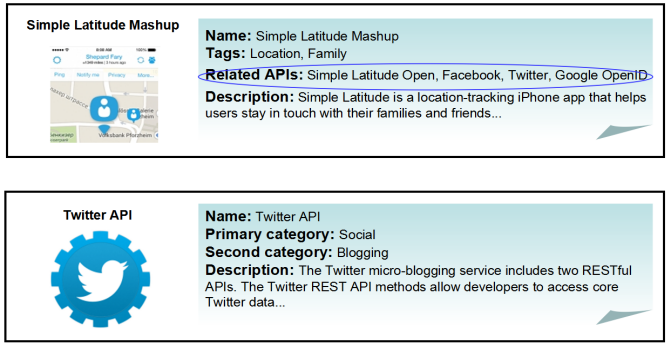


A Probabilistic Model

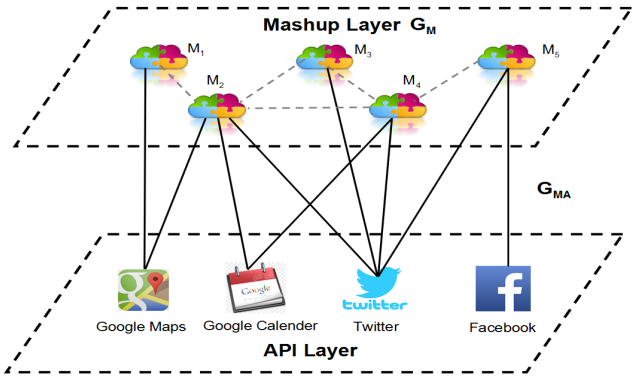
document-centric probabilistic model: estimate the expertise of a candidate by summing the relevance of its associated documents

$$p(m_i|q) = \lambda \sum_{a \in \mathcal{A}_{m_i}} p(m_i|a)p(a|q) + (1-\lambda)p(m_i|q)$$

$$\propto \lambda \sum_{a \in \mathcal{A}_{m_i}} p(m_i|a)p(q|a)p(a) + (1-\lambda)p(q|m_i)p(m_i)$$



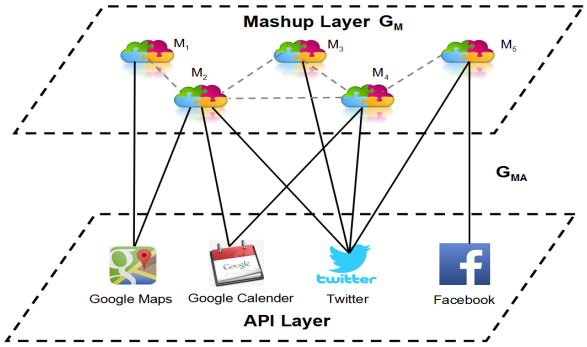
- p(mi|a): the probability of API a belongs to mashup mi
- p(q|a) and p(q|mi): the semantic similarities of API and mashup
- p(a) and p(mi): the quality of API a and mashup mi



$$\mathbf{z} = \lambda \mathbf{P}_{MA}\mathbf{Q}_A\mathbf{x} + (1-\lambda)\mathbf{Q}_M\mathbf{y}$$

Modeling the Heterogeneous Network

- Mashup Consistency Hypothesis:
  - If two mashups share many common services with respect to a given query, then their relevance score in the queried field should be similar in some sense.



- Objective function

$$\Omega(\mathbf{z}) = \mathbf{z}^T(\mathbf{I} - \mathbf{S}_M)\mathbf{z} + \mu \left\|\mathbf{z} - \mathbf{z}^0\right\|^2$$

$$s.t. \quad \mathbf{z}^0 = \lambda \mathbf{P}_{MA}\mathbf{Q}_A\mathbf{x} + (1-\lambda)\mathbf{Q}_M\mathbf{y}$$

- Similarity matrix

$$\mathbf{S}_M = \mathbf{\Pi}^{-1/2}\mathbf{W}\mathbf{\Pi}^{1/2} \qquad \mathbf{\Pi}_{ii} = \sum_j \mathbf{W}_{ij}$$

Optimization

$$\Omega(\mathbf{z}) = \mathbf{z}^T(\mathbf{I} - \mathbf{S}_M)\mathbf{z} + \mu \left\|\mathbf{z} - \mathbf{z}^0\right\|^2$$

$$s.t. \quad \mathbf{z}^0 = \lambda \mathbf{P}_{MA}\mathbf{Q}_A\mathbf{x} + (1-\lambda)\mathbf{Q}_M\mathbf{y}$$

- Setting $\partial\Omega(\mathbf{z})/\partial\mathbf{z} = 0$

$$(\mathbf{I} - \alpha\mathbf{S}_M)\mathbf{z}^* = (1-\alpha)\mathbf{z}^0 \quad \alpha = 1/(1+\mu)$$

Variant Graph Laplacian using as Sm is the adjacency matrix

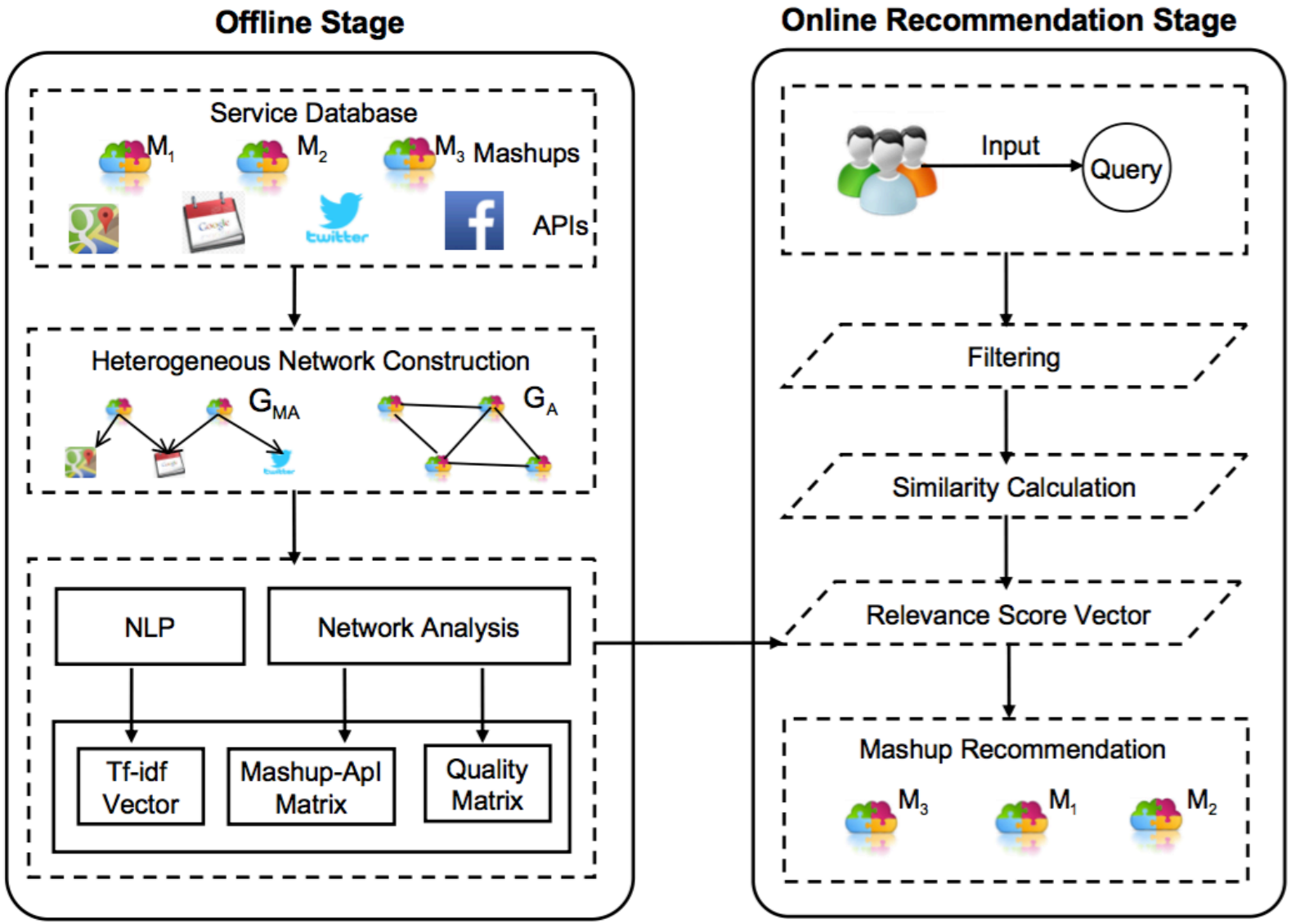
- Since $\mathbf{S}_M$ is usually sparse

$$\mathbf{z}^{t+1} \leftarrow \beta\mathbf{S}_M\mathbf{z}^t + (1-\beta)\left[\lambda\mathbf{P}_{MA}\mathbf{Q}_A\mathbf{x} + (1-\lambda)\mathbf{Q}_M\mathbf{y}\right]$$

$$\beta = 1/(1+\mu),\ \mathbf{z}^* = \mathbf{z}^\infty \text{ is the solution.}$$

Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Scho¨lkopf, B.: Learning with local and global consistency.

Implementation



Experiments

- Dataset

Programmableweb.com

# of mashups	4,699	# edges in G_M	3,760,923
# of APIs	937	# edges in G_{MA}	8,127

- Evaluation metrics

$$P@K = \frac{\#\ relevant\ in\ top\ K\ results}{K}$$

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \qquad \alpha\text{-}DCG_K = \sum_{i=1}^{K}\frac{G_i}{log_2(i+1)}$$

Clarke, C.L., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Bu ̈ttcher, S., MacKinnon, I.: Novelty and diversity in information retrieval evaluation.

Experiments

- Ground Truth

Category	Sample queries
Music	audio, band, lyrics, sound, music
Travel	hotel, tourism, flight, holiday, voyage, travel
Financial	bitcoin, stock, investment, price, finance, economics
eCommerce	shopping, eCommerce
Mapping	GIS, geography, GPS, mapping, traffic
Sports	NFL, fitness, sport, running, athlete
Photos	photo, image, picture, camera
Government	policy, congress, department, government, law
Game	player, game
Education	education, training, student, school
Enterprise	sales, hr, leader, enterprise
Social	media, social

The idea of binary judgment in the evaluation is: a result for a given query can be judged as relevant once its category is identical with the category of the query.

Experiments

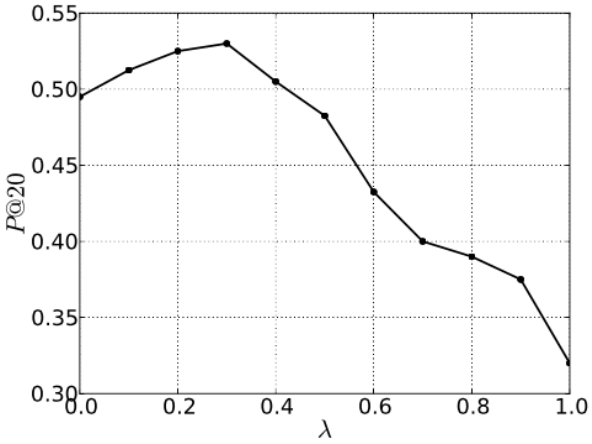
- Experimental results

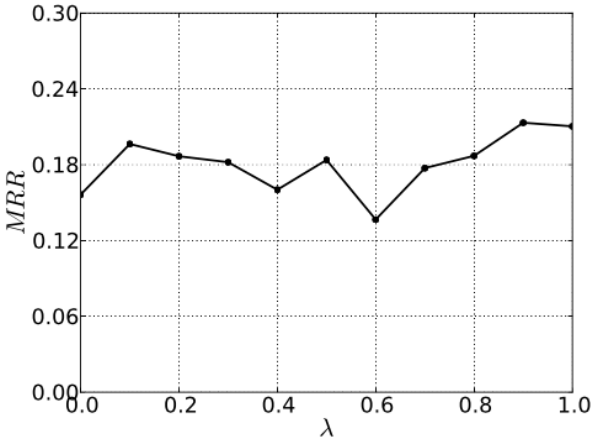
	P@10	P@20	P@50	MRR	α-DCG
PW	**0.595**	0.493	0.431	-	-
MD-Sim	0.555	0.488	0.553	0.121	2.920
MD-Sim+	0.575	0.495	0.534	0.137	2.916
(vs **MD-Sim**)	+3.60%	+1.43%	-3.44%	+13.22%	-0.01%
MD-HIN	0.555	**0.53**	**0.537**	**0.160**	**3.027**
(vs **PW**)	-6.72%	+7.51%	+24.59%	-	-
(vs **MD-Sim**)	0.00%	+8.61%	-2.89%	+32.23%	+3.66%

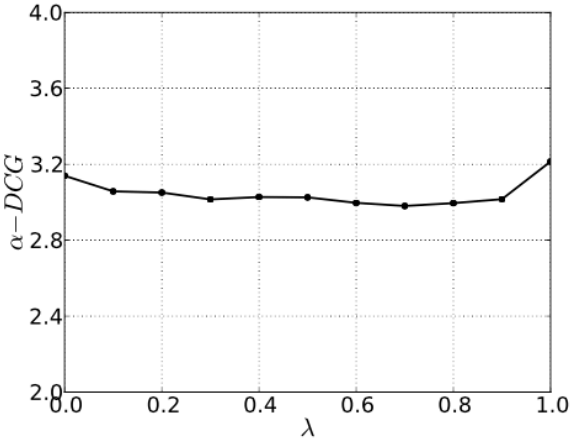
- **Observations**
  - 1. when the value of K is large, our approach has a great advantage over ProgrammableWeb search engine.
  - 2. From the perspective of ranking of results, our extended approach (MD-HIN) achieves better performance than the probabilistic approach
  - 3. Among all the discovery methods, our proposed method generally achieves better performance on both metrics, indicating the effectiveness of our approach.

Parameter Analysis

- Impact of lambda





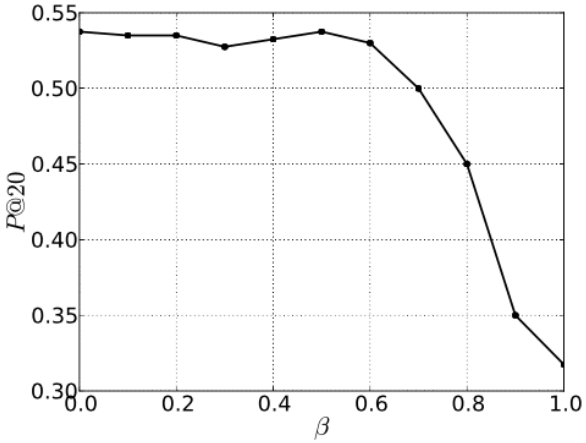


(a) Impact of λ on $P@20$ (b) Impact of λ on MRR (c) Impact of λ on α-DCG

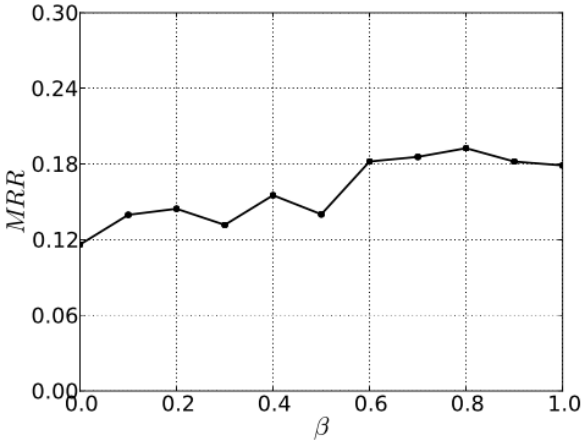
- **Observations**
- 1. the value of λ really has an significant impact on the performance of mashup discovery.
- 2. As λ increases, the P@20 value increases at first, but when λ surpasses a certain threshold, the P@20 value decreases with further increase of the value of λ
- 3. This phenomenon confirms the intuition that purely using the semantic information of mashups or purely employing the semantic information of their related APIs cannot generate better performance than fusing these two factors together.

Parameter Analysis

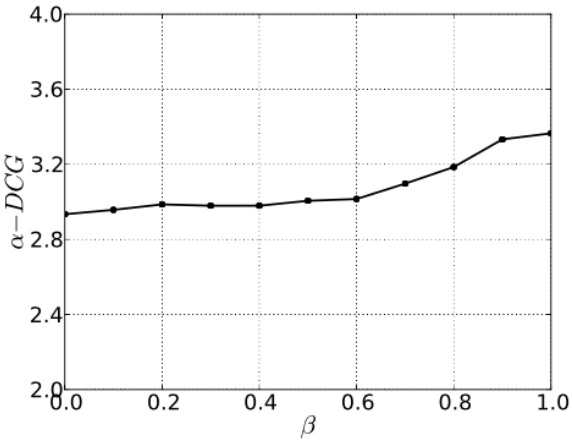
- Impact of beta



(a) Impact of β on $P@20$



(b) Impact of β on MRR



(c) Impact of β on α-DCG

- **Observations**
- 1. β has a significant impact on the precision of the discovery results.

Conclusion and Future Work

- Conclusion
  - Proppose an approach to improve mashup discovery by integrating the semantic information of mashups and their related APIs
  - Similarity consistency is proposed and a regularization framework is implied
  - Compremrehensive experiments on a real-world dataset
- Future work
  - Extend our ground truth with more queries
  - More evaluation metrics will be introducted

Q & A

Thank you!

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