The 8th Chinese Workshop on Machine Learning and Applications
第八届中国机器学习及其应用研讨会

The Need of Machine Learning and Other Computational Technologies for Monetization in Internet Industry

Aden Yue 岳亚丁，adenyue@tencent.com
Tencent Inc.

November 7, 2010

腾讯网
QQ.com

Motivation

- Gap between the academia and industry.

- Practitioners ask:
  - What technologies are available for the current problems?
  - Can the paper's results be used directly?
- Researchers ask:
  - Academic value in the problem/topic
  - Data

Contents

- Current Situation
  - Monetization
  - Problems (Computational Tasks)
  - Usual Practices: Tencent Case
- Need
  - Solution Framework
  - Parallelized Algorithms
  - Others
- Summary

Revenue Model in Tencent

Free services attract users, paid services suit their personal needs.

（QQ, QQ.COM, Qzone, QQ Game, 3G.QQ.com,QQlive, Paipai......）

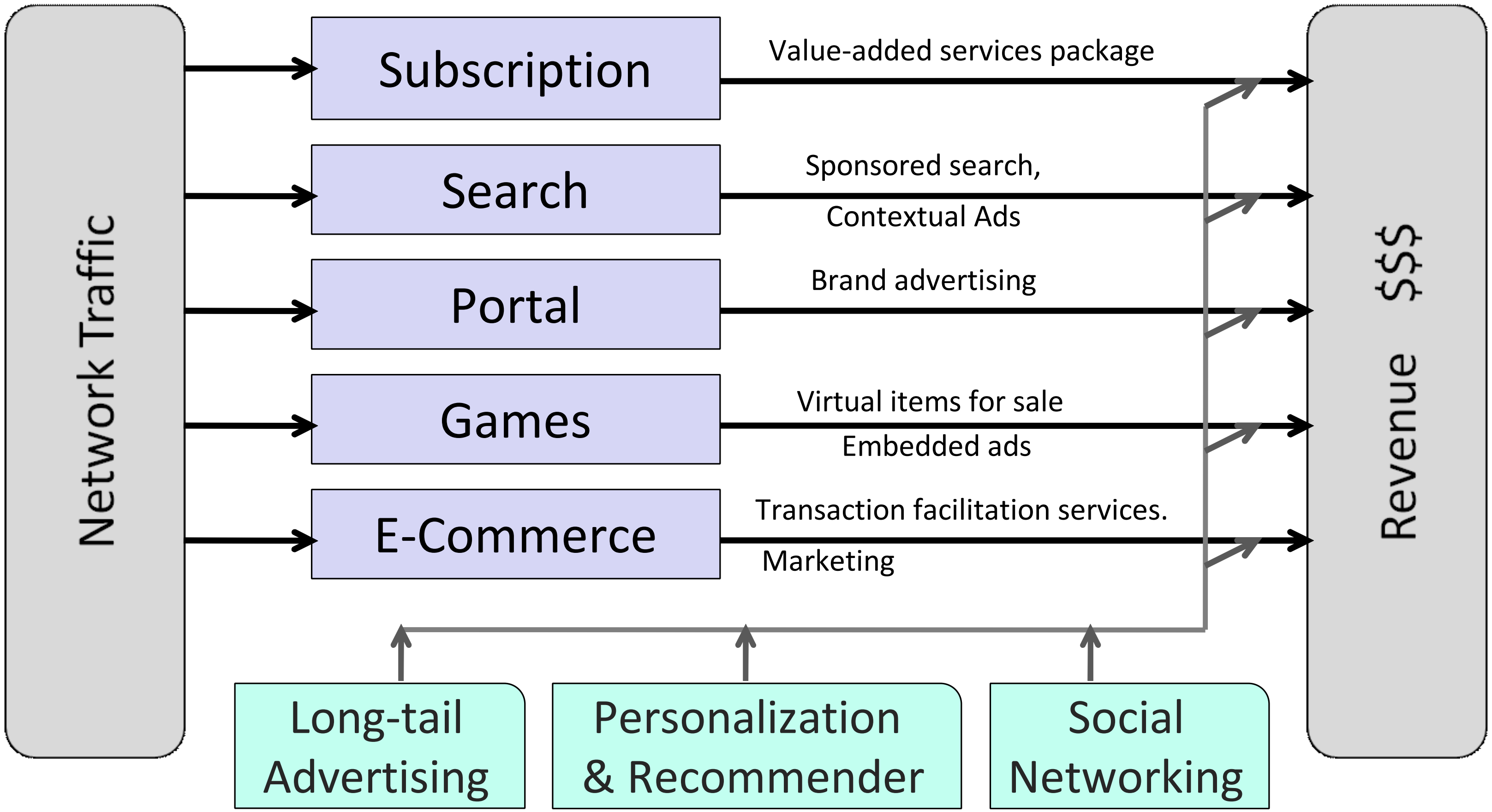
Value-added on Internet (78%)*	Value-added wireless (13%)	Advertising (9%)
Instant Messaging (QQ 会员)	Mobile IM (超级QQ、手机QQ)	IM Ads (客户端、 AIO, Tips)
Online Social Networks (QQ空间、QQ秀、QQ宠物、QQ音乐)	SMS Services (资讯、交友、星座 等)	Web Ads (qq.com, 手机腾讯网)
Online Games (QQ游戏, QQ三国、QQ飞车、DNF 等)	2.5G Services (彩信、WAP、游戏 等)	Paid Search (搜搜www.soso.com)
Others (财付通、QQ交友、QQ相册、Q吧)	3G Services (3G影院, 3G图书室, 手机游戏)	Others (视频广告、社区嵌入广告)

- Up to 2009Q3, There were 1.057 billion QQ accounts, among which 0.4849 billion active, market share > 80%

* Revenue proportion was calculated from Financial Report of Quarter 3, 2009.

Source: 《腾讯介绍100104(精简版)》

Monetization on Internet



- Current Situation
  - Monetization
  - Problems (Computational Tasks)
  - Usual Practices: Tencent Case
- Need
  - Solution Framework
  - Parallelized Algorithms
  - Others
- Summary

Computational Tasks —— ‘Subscription’

Task	Notes
Service Positioning	‘Product – market demand’ matching Product differentiation, cannibalism avoidance
Marketing for User Acquisition	Market share prediction Marketing strategy simulation Interactive marketing / event- driven marketing Fatigue on marketing information (harassment)
Customer (User) Care	Users’ Life-time Value modeling and augmentation: cross-sell, up-sell User Churn Prediction and Retention Understanding users: segmentation, interests and intent
Syndication	Content / items Aggregation and Bundling
Yield Management	Pricing: price differentiation, dynamic pricing KPIs prediction, abnormal variation detection

Computational Tasks —— ‘Search’

Task	Notes
Performance	Crawl and indexing: scheduling, efficiency Ranking and relevance refinement: ranking functions, feedbacks Query understanding: expansion, tagging, reformulation Social search: community interests, social annotation
Sponsored Search & Contextual Advertising	Relevance: ad – query, ad – context (page), user – ad Ad selection: Database-based, IR-based, recommendation-based Keyword suggestion Factors influencing conversion rate Click-through rate prediction Intention recognition, sentiment detection Auction mechanisms: GSP, VCG, Myerson Content analysis: topic modeling on landing pages Click fraud

Computational Tasks —— 'Portal'

Task	Notes
Brand Advertising (Banner Ads)	Combinatorial ad allocation for revenue maximization: periodic optimize and dispatch, scalable, pricing Expressiveness: auctions where bidders' are free to specify preferences (demographics, websites, etc.) in greater details. Inventory packaging: a collection of host web pages for one winner bidder Guaranteed impression: predict supply &demand More types of targeting: behavioral targeting, geo-targeting, re-messaging, retargeting Audience extension: users who visited a publisher's website will see ads on another advertising media/network. Media aggregation: ad exchange, ad networks Which ad creative, which landing page?

Computational Tasks —— ‘Games’

Task	Notes
User Acquisition	Potential users identification
Fraud Detection	Multiple players’ offline cheating on online games Cheating programs (unauthorized servers) ID and property theft
Virtual Items	Pricing and recommendation
Intelligent Avatars, Pets, Virtual Life	Self-learning, ‘genetically’ evolving, adapting to complex environment for more fun (NOT SO URGENTLY NEEDED AS PER INTERATIVE ENTERTAINMENT)
Ecosystem and Economy	Equilibrium for skills and grades settings ‘Currency’ depreciation control and other financial issues

Computational Tasks —— 'E-Commerce'

Task	Notes
Security	Phishing and troyan virus: detection and prevention Fraud prevention: keyword abuse, false prices and description, collusion
Support to Sellers and Hosting Websites	Predict potential buyers Optimize users' experience during their visit session Predict user's commercial interests (short- or long-term) Enhance the transaction volume and users' satisfaction Use the peer reviews from user's friends for users' purchase decision Trustworthiness: rating system, automation, social interaction
Support to Buyers	Recommendation: save search and comparison time Agile decision agent for each user, example-critiquing Understanding users' needs for better matching

Computational Tasks —— ‘Long-tail Advertising’

Task	Notes
Response Enhancement	Prediction of click-through rate, turnover rate Micro mechanism, users’ experience modeling Use of users’ profile, real-time behavioral data, social networking data, geographical information Prediction of users’ interests, interest taxonomy Freer targeting conditions specification
Making More Use of Data Sources	Greedy approach is not optimal Evaluation of data sources Users’ privacy
Analysis Tool	Support to advertisers’ campaign design
Further Innovations	Relevance among users, context, and ads Auction mechanism Extraction of ad attributes from images/videos

Computational Tasks ——
‘Personalization and Recommender’

Task	Notes
Improvement on Current Algorithms	Definition and computation cost of similarity measures Hybrid algorithms by integrating many ones Construction of new attributes of items and users Chain of intent Use of intermediate summarized data Cold start, ‘surprise me’ recommendation
New Problems	Framework for a universal solution Fusion of multiple diversified data sources Stepwise refined recommendations Recommendation activated by users’ behaviors or other events Performance measures, e.g., revenue (other than prediction accuracy) Transient, aging factors

Computational Tasks —— 'Social Networking'

Task	Notes
Understanding Users	Clustering: with more social interaction attributes Behavioral prediction based on linked nodes Interest derivation: influence of users' social ties Users' network value and influences: opinion leaders, influencers Trust among users: authoritativeness
Profitable Applications	Recommender systems: using users' SNS features Target advertising: if one user's friends have viewed and clicked a certain ad, it is very reasonable to show the ad to the user. Location Based Services (LBS) Viral marketing: campaign designs Social search: correlations between preferences of Web search results and similarities among users

- Current Situation
  - Monetization
  - Problems (Computational Tasks)
  - Usual Practices: Tencent Case
- Need
  - Solution Framework
  - Parallelized Algorithms
  - Others
- Summary

Data We Have

- Data warehouse (hundreds of TB)
- Distributed Processing and Computing Platform, based on Hadoop

Type	Product Line	Data
Core Platform	Instant Messenger (IM), Business Operation Support System (BOSS), Marketing Platform	QQ Users' demographics, friends, groups, interests, subscriptions, spending, recharging, message (tips) pushes, arrivals, and clicks.
Key Data in Generated in Use (Operation)	Wireless： 超级QQ、手机游戏、手机腾讯网、手机QQ Internet： Qzone空间、校友、城市达人、QQMUSIC、QQLIVE、QQshow、QQ会员、QQ农场、QQ牧场 Interactive Entertainment： minigame、幻想、QQ宠物、QQ三国、QQ飞车、CF、音速、DNF、X5、寻仙、华夏、大明龙权 Others： 广告、拍拍、soso、问问	Registration info, items and gears, expenditures, operation and behaviors (log, role, grades, activeness, functions used), etc.
Behaviors Online	qq.com, websites of online products, websites of games, paipai.com, mobile qq website	PV\UV, clicks\path\source, source information to clients (QQ, games), etc.

Computational Procedure

Problem Formulation (Computational Tasks)

- Goal
- Value Proposition
- Feasibility
- Data Availability
- Resources
- Priority
- …

Solution Design (Modeling)

- Modeling as classification, regression, clustering, optimization or a hybrid problem?
- Pre- and post-processing

Algorithm

- Maturity and robustness
- Complexity
- Code Availability
- Scalability
- …

System Design & Implementation

- Project Mgt
  - Quality
  - Timeliness
  - Cost
  - …
- Verification
- Validation
- …

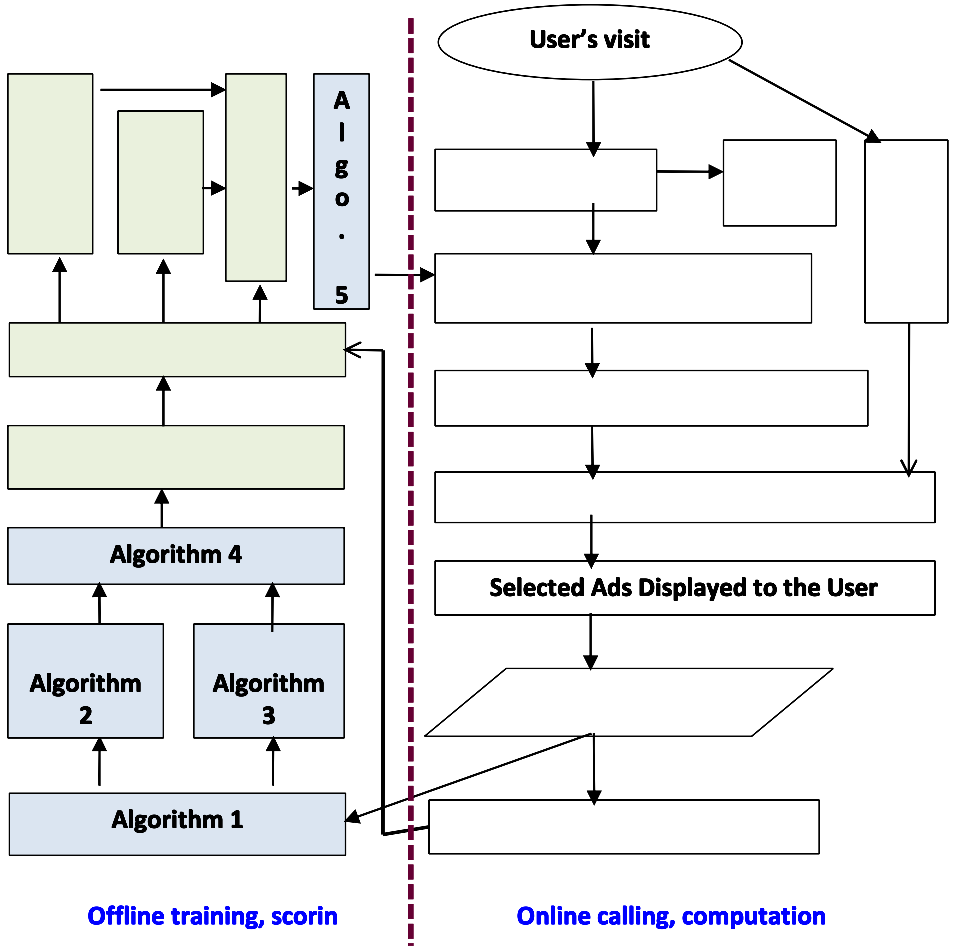
Roll-out

- Benefits reporting
- Performance monitoring
- Continuous improvement with new data sources and updated algorithms

Issues When We Design the Solution

- Available Data
  - Amount, quality, overlap among sources, sparsity, unbalanced label
  - Indicativeness of the goal
  - Making full use of historic data, less dependent on new tests
- Criteria (tentative)
  - Performance oriented
  - Modular components additive and cumulative to the goal, conflict – free, redundancy - free
  - Continuous self-refining
- Preprocessing and Post-processing
  - Feature selection/ extraction / transformation, dimension reduction
  - Ensemble (bagging, boosting), pruning, incremental learning
- Complexity
  - E.g., scoring time for 10^9 users' response on 10^6 items.

Solution Example: Long-tail Advertising



- Multiple algorithms/ components integrated for one goal
- Plenty of preprocessing and post-processing
- Performance is critical
  - System response
  - Profitability

Issues When We Select Algorithms

- Performance Estimation Beforehand
  - Previous tests
  - Published results
  - Experience, best guess
- Implementation
  - Available source code
  - Development cost
- Efficiency and Effectiveness
  - Parallelism
  - 'Sometimes a highly efficient single-machine implementation with sparse matrix operations yielded performance the same order of magnitude as a parallel implementation.' [1]
- Which Variation to Use for a Given Algorithm

Technologies More Used …

- Classification / Regression
  - Bayesian, Random Decision Tree, Gradient Descent Decision Tree, Logistic Regression, …
  - Feature extraction
- Clustering
  - Spectral clustering
- Co-clustering, Bipartite Partitioning
  - Soft, hard
  - Graph based, information theoretic based
- Optimization
  - Direct search: PSO, ACO
  - Linear programs with less over-targeting
- Large Scale Matrix Computation
  - SVD, LSI, pLSI, LDA

Technologies Currently Less Used …

- Reinforcement Learning
- Evolutionary Approaches
  - Symbolic regression with GP or its variants (GEP, MEP, etc.)
  - Automatic product design
- Use of Unlabeled Data, and Less Instances
  - Semi-supervised learning, transfer learning, self-taught learning, active learning
- Self-Organizing Models
  - Group method of data handling, multi-agent system
- Microscopic Mechanism Modeling
  - Diffusion process in SNS
  - Cellular automata, petri net, event graph
- Neuro-Fuzzy
  - Universal approximators with interpretable IF-THEN rules

- Current Situation
  - Monetization
  - Problems (Computational Tasks)
  - Usual Practices: Tencent Case
- Need
  - Solution Framework
  - Parallelized Algorithms
  - Others
- Summary

Need for Solution Frameworks

- E.g., recommender system:
  - Matrix factorization [2]

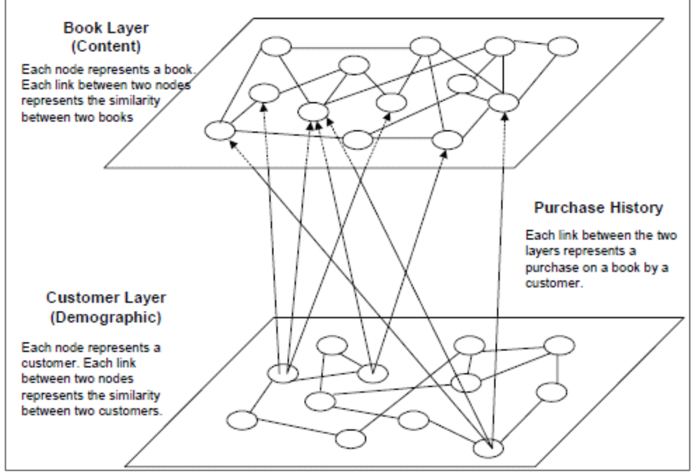
    User u's rating of item i is represented as:

$$\hat{r}_{ui} = \mu + b_i + b_u + q_i^T [p_u + |N(u)|^{-0.5} \sum_{i \in N(u)} x_i + \sum_{a \in A(u)} y_a]$$

  - Graph based [3]

    Exploit high-degree book-book, user-user and book-user associations. →

FIGURE 2. Two-layer graph model of books, customers and purchases

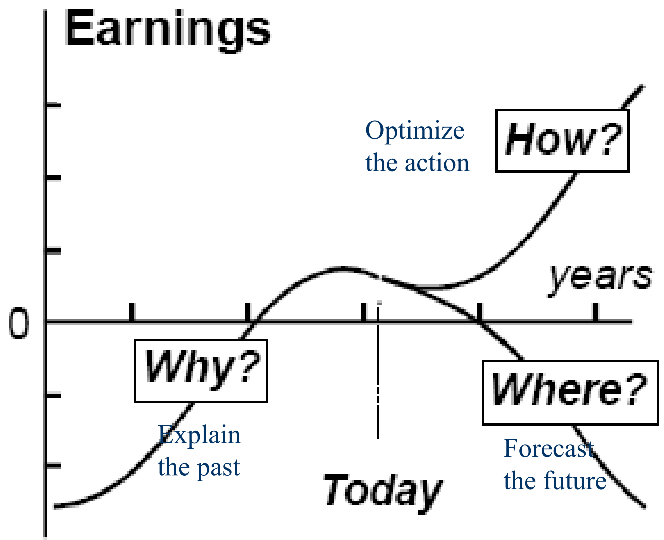


- Unified Framework

  - all components (data, processing algorithms, etc.) contribute additively.

Need for Blueprint for Each Problem Type

- Solution Designs Closer to Real Context
  - E.g., in online brand advertising, multiple ads displayed on one position are selected according to a pre-calculated time proportion or each user's click-through rate, instead of one ad on one position for a whole time slot.
- Profit-oriented, Actionable Insight



What are the fundamental underlying drivers for the business performance and how do we make use of them ?

- Current Situation
  - Monetization
  - Problems (Computational Tasks)
  - Usual Practices: Tencent Case
- Need
  - Solution Framework
  - Parallelized Algorithms
  - Others
- Summary

Parallelized Algorithms

E.g.,

Under MapReduce framework (Hadoop platform preferred)

	Name	Data & Condition
1	Classification / Regression	10^9 instances * 10^4 attributes, unbalanced labels.
2	Co-clustering	$10^{6\sim7}$ rows * 10^5 columns.
3	User Clustering	$5*10^8$ users (each having dozens of attributes) clustered into approximately 10^5 groups
4	Feature Extraction	Not dependent on specific classifiers; universal as similar to nonlinear independent component analysis or genetic programming

Parallelized Algorithms – cont.

- Basic Ones with Excellent Performance
  - Tested on large datasets
  - Robust on stated data conditions
  - Performance (complexity) compared with other latest algorithms (not the old original ones)

Other Challenges (1): Strategic Modeling

- What New Product/Services Will Emerge Next?
  - Which trends assertions are correct?
  - How to assemble multiple qualitative inferences?
  - Quantitative business model evolution
  - New product/innovation prediction
- Strategic Decision Modeling for SNS Development [3]
  - Where are the current SNS heading
  - Monetization opportunities in next generation of SNS
  - Sustainable attraction of SNS platforms
  - Is the offline real relationship critical in SNS development? What extent?
  - How do users' needs and SNS functions co-evolve?
  - What impact does the openness have on users and developers?
  - What applications to launch on mobile systems (cell phones)?
  - Natural or stimulated and optimized structural evolution of SNS for business benefits.

Other Challenges (2): User Again

- Determine Users' Needs at Any Moment
  - E.g., "Tell me: what will user A need and will do on next Tuesday morning?"
  - Suppose we have all the necessary data, including users' demographics, behaviors (original log and aggregated), interests and other derived attributes, …
- User Experience Modeling
  - Quantitative modeling of users' experience and product usability for product improvement
  - Automatic GUI design and product functionality configuration based on evolutionary computation, at least for providing inspiration and innovative ideas to designers.

Other Challenges (3): User's Personal Assistant

- Generalized Recommender Systems
  - Every user has an agent collecting info from the web, and learn the user's interests and instant needs, recommend on daily activities: read, play, purchase, travel, health care, financial planning, etc.
  - Intelligent agent for micro-decision support
  - Personal info assistant, e-commerce decision advisor
  - Technologies involved (probably): NLP, ILP, Recommender Systems, AI, Semantic Web, etc.
- Light weighted, loosely coupled, easily built Decision Support Systems (DSS) for individuals
  - Natural language understanding
  - First order logic extended for accommodating quantities?

Other Challenges (4): DSS for Business

- Decision Support System (DSS) with Simulation
  - The system can tell the product manager when there should be launched what type of marketing campaign, when what functionality should be added to the product, and give the expected effects of the advices (KPIs, future market share, financial benefits, etc.)
  - Further, the system can output multiple action options, recommend the best action plan.

Other Challenges (5): Decision Making Model

General Decision Making Model

- Identity State Eq. :

  $X = f(X, U, t)$

  where state vars: $X \in R^m$,

  control vars: $U \in R^n$
- Measurables:

  $Y = g(X, U, t)$

  where $Y \in R^p$
- Find Optimal Operation:

  $$U^* = \arg\max_{u} [h(Y)]$$

- Should we always fix the numbers m and n ?
  - Yes, for simplicity.
  - No, for real situations.
- What about f?
  - Let the structure of f becomes more and more complicated, like the evolving process of baby's brains?
  - Cascade neural network?
- Alternative?
  - New equation: $X = f(U, t)$

    But, X's historic data are not used then, learning becomes difficult.

- Current Situation
  - Monetization
  - Problems (Computational Tasks)
  - Usual Practices: Tencent Case
- Need
  - Solution Framework
  - Parallelized Algorithms
  - Others
- Summary

Problem-Oriented Researches for Fun

Academic Value

- New research topics arise.
- Methodologies improved.
- New conception and technologies invented.
- Theory tested on real cases.

Industrial Value

- Valuable: solution framework, basic algorithms with excellent performance and their parallelized implementation.
- Even more valuable: industry trend prediction, new product emergence prediction, product-user co-evolution for automatic product designs, users' need at any moment, generalized recommender systems, strategic decision simulation, etc.

References

- [1] Ye Chen, et al., *Practical Lessons of Data Mining at Yahoo!,* CIKM'09, November 2–6, 2009, Hong Kong, China., pp. 1047-1055. (2009)
- [2] Koren, Y.; Bell, R.; Volinsky, C., *Matrix factorization techniques for recommender systems*, IEEE Computer, Volume 42, Issue 8, p.30-37 (2009)
- [3] Zan Huang, Wingyan Chung, Thian-Huat Ong, Hsinchun Chen , *A Graph-based Recommender System for Digital Library*, JCDL'02, July 13-17, 2002, Portland, Oregon, USA. (2002)
- [4] Gordon Sun, Rick Zhuang, Aden Yue, Online Social Networks: *Insight, Commercial Value, and Computational Challenges*, Keynote Speech, CIKM'10, October 25-29, Toronto, Canada. (2010)

Thank You

Q & A …