CROWD COMPUTING

Ling Liu
Professor
College of Computing
Georgia Institute of Technology

Materials from PAKDD 2014 tutorial, Reynold Cheng, Yudian Zheng

Lecture Outline

2

Today

- Machine v.s. Human
- Crowd Sourcing / Crowd Computing: An Introduction

Other interesting topics (Future lectures)

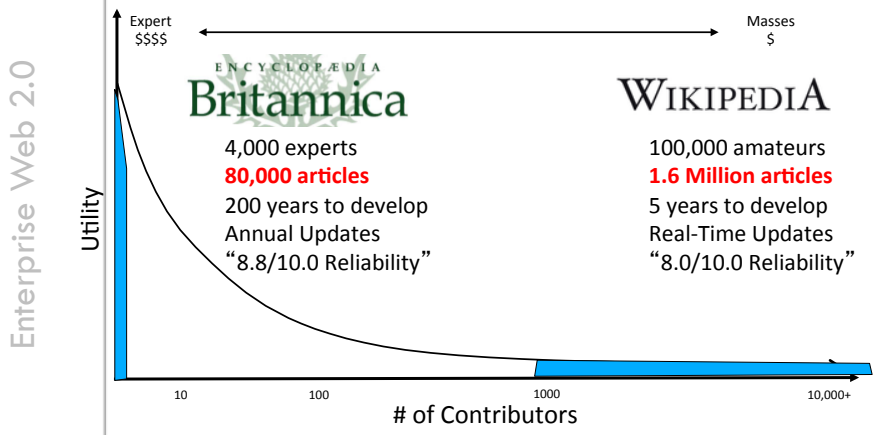
- Crowd Sensing
- Crowd Assisted Data Analytics + Crowd Powered DM

What is crowdsourcing?

- Crowdsourcing is an online, distributed problem solving and production model.
- Users – known as the crowd
  - typically form into online communities based on the Web site, and
  - the crowd submits solutions to the site or produce its contents.

Economics & Wikinomics

Wisdom of the Crowds

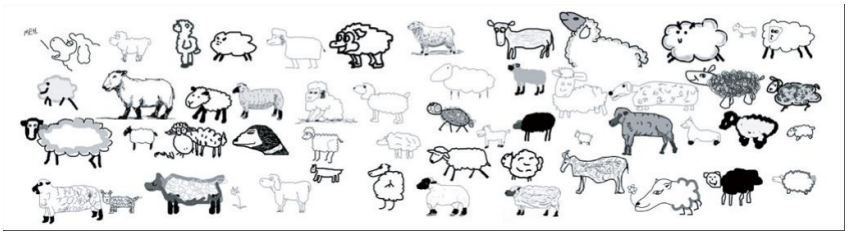


Sheep Market

6

- Draw a Sheep facing left

Pay each worker $0.02
Collect 10,000 drawing sheep



http://www.thesheepmarket.com/

No ground truth

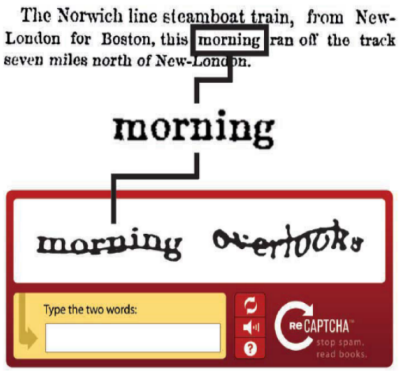
Optical Character Recognition (OCR)

7

- CAPTCHA

Completely Automated Public Turing test to tell Computers and Humans Apart

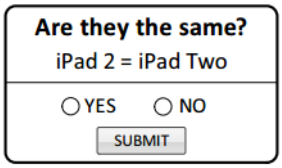
- ReCAPTCHA

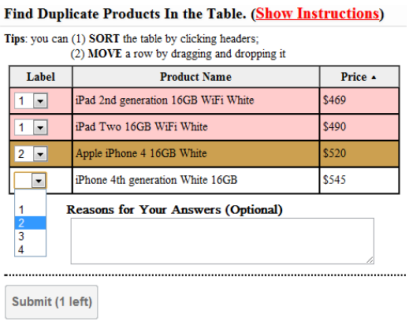


Luis von Ahn, Benjamin Maurer, Colin McMillen, David Abraham and Manuel Blum.
ReCAPTCHA: Human-Based Character Recognition via Web Security Measures.
Science, 321: 1465-1468, 2008

Entity Resolution (ER)

8



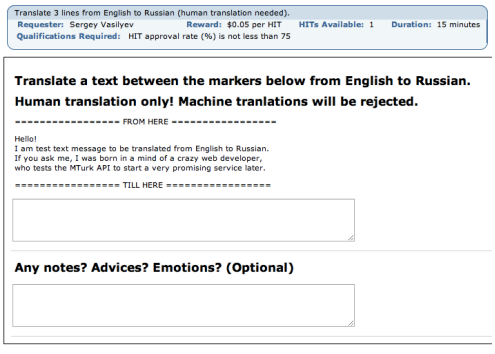


J. Wang, T. Kraska, M. J. Franklin, and J. Feng. Crowder: Crowdsourcing entity resolution.PVLDB, 5(11):1483-1494,2012.
J. Wang, G. Li, T. Kraska, M. J. Franklin, and J. Feng. Leveraging transitive relations for crowdsourced joins. In SIGMOD Conference, pages 229-240, 2013.

Natural Language Processing (NLP)

9

- Translation



[1] C. Callison-Burch. "Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk", EMNLP 2009.

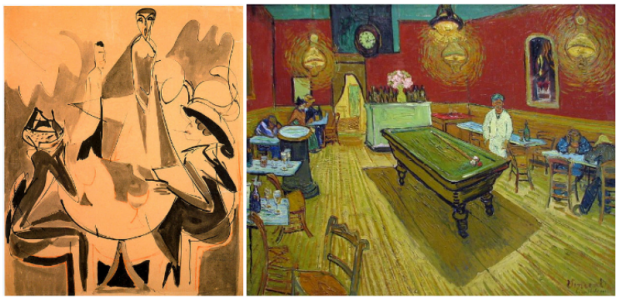
[2] B. Bederson et al. Translation by Iter active Collaboration between Monolingual Users, GI 2010

No ground truth

Computer Vision (CV)

10

- Painting Similarity



How similar is the artistic style in the paintings above?

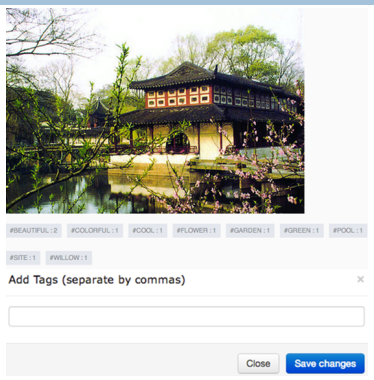
- Very similar
- Somewhat similar
- Neither similar nor dissimilar
- Somewhat dissimilar
- Very dissimilar

A gradient based weighted averaging method for estimation of fingerprint orientation fields. Yi Wang et al. DICTA'05.

No ground truth

Collaborative Tagging

11



User interface for providing "tag" keywords

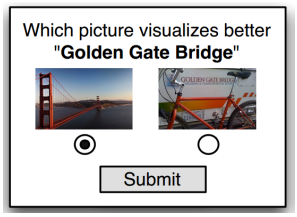
X. S. Yang, D. W. Cheung, L. Mo, R. Cheng, and B. Kao. On incentive-based tagging. In Proc. of ICDE, pages 685-696. 2013

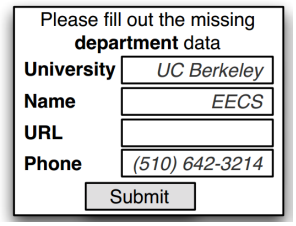
Siyu Lei, Xuan S. Yang, Luyi Mo, Silviu Maniu, Reynold Cheng iTag: Incentive-Based Tagging. ICDE 2014 demo.

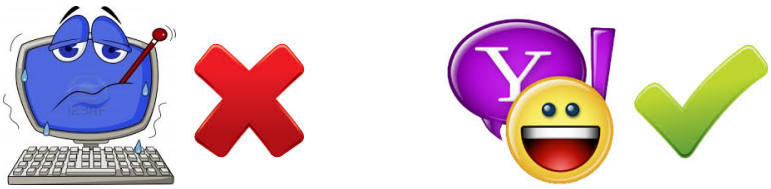
No ground truth

Why crowdsourcing?

12





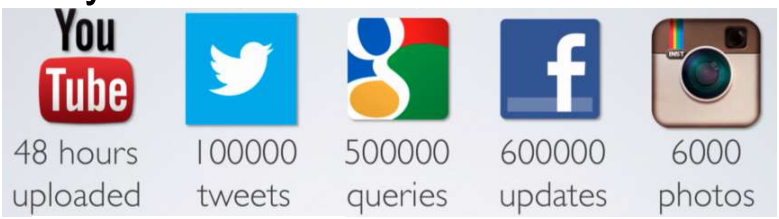


M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin. Crowddb: answering queries with crowdsourcing. In SIGMOD Conference, pages 61-72, 2011.

Why crowdsourcing?

13

- Every Minute



Aditya Parameswaran ,Human-Powered Data Management , http://msrvideo.vo.msecnd.net/rmcvideos/185336/dl/185336.pdf

Crowdsourcing Platforms

14

- Voluntary

- Incentive-based

Crowdsourcing Model

15

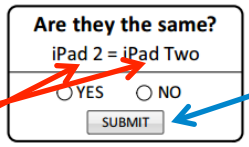
A requester asks a crowd to do Human Intelligence Tasks (HITs) to solve problems.

problem: entity resolution (ER)

HIT: comparison questions

crowd: Internet users

ID	Object
o_1	iPhone 2nd Gen
o_2	iPhone Two
o_3	iPhone 2
o_4	iPad Two
o_5	iPad 2
o_6	iPad 3rd Gen

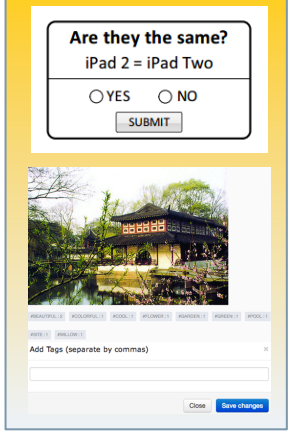


Amazon Mechanical Turk (AMT)

16

- Requesters

- HITs

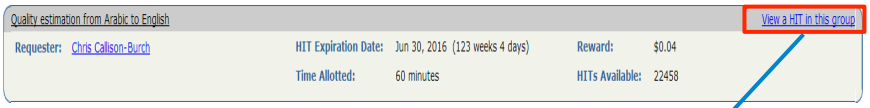


- Workers

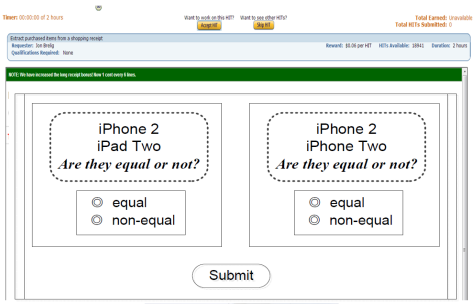
Human Intelligence Tasks (HITs)

17

- HIT group on AMT



- A HIT



AMT Statistics

18

- New York Times (March, 2007)

  *Today, there are more than **100,000** "Turk Workers" in **more than 100 countries** who earn micropayments in exchange for completing a wide range of quick tasks called HITs, for human intelligence tasks, for various companies.*

- Official Amazon Mechanical Blog (August, 2012)

  *more than **500,000 workers** in **190 countries***

http://www.nytimes.com/2007/03/25/business/yourmoney/25Stream.html
http://mechanicalturk.typepad.com/blog/2012/08/mechanical-turk-featured-on-aws-report.html

AMT Statistics (HITs)

19

- Types of tasks

*From **January 2009 till April 2010**, we collected 165,368 HIT groups, with a total of **6,701,406 HITs**, from **9,436 requesters**. The total value of the posted HITs was **$529,259**.*

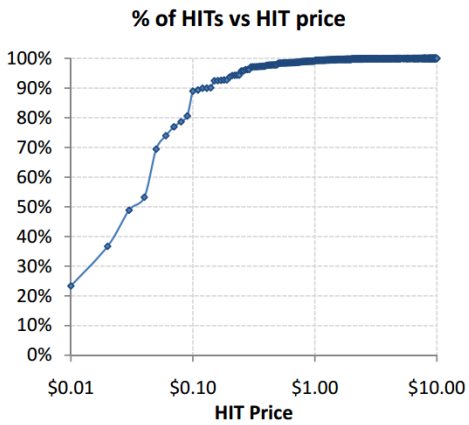
Requester ID	Requester Name	#HIT groups	Total HITs	Rewards	Type of tasks
A3MI6MIUNWCR7F	CastingWords	48,934	73,621	$59,099	Transcription
A2IR7ETVOIULZU	Dolores Labs	1,676	320,543	$26,919	Mediator for other requesters
A2XL3J4NH6JI12	ContentGalore	1,150	23,728	$19,375	Content generation
A1197OGL0WOQ3G	Smartsheet.com Clients	1,407	181,620	$17,086	Mediator for other requesters
AGW2H4I480ZX1	Paul Pullen	6,842	161,535	$11,186	Content rewriting
A1CTI3ZAWTR5AZ	Classify This	228	484,369	$9,685	Object classification
A1AQ7EJ5P7ME65	Dave	2,249	7,059	$6,448	Transcription
AD7C0BZNKYGYV	QuestionSwami	798	10,980	$2,867	Content generation and evaluation
AD14NALRDOSN9	retaildata	113	158,206	$2,118	Object classification
A2RFHBFTZHX7UN	ContentSpooling.net	555	622	$987	Content generation and evaluation
A1DEBE1WPE6JFO	Joel Harvey	707	707	$899	Transcription
A29XDCTJMAE5RU	Raphael Mudge	748	2,358	$548	Website feedback

Ipeirotis, P. G. (2010a). Analyzing the Amazon Mechanical Turk marketplace. ACM XRDS, 17, 16–21.

AMT Statistics (HIT price)

20

- HIT price



70 % HITs cost less than $0.05

Percentage	Price
25%	[$0 , $0.01]
45%	[$0.01 , $0.05]
20%	[$0.05 , $0.10]
10%	[$0.10 , $10.00]

Ipeirotis, P. G. (2010a). Analyzing the Amazon Mechanical Turk marketplace.ACM XRDS, 17, 16–21.

AMT APIs

21

- Application Programming Interfaces

Official: C# Java Perl Ruby
Open-source: Python (boto)

GitHub

PUBLIC boto / boto

- Java example (create a HIT)

```
HIT hit = service.createHIT
(
     title,
     description,
     reward,
     RequesterService.getBasicFreeTextQuestion(
     "How many movies have you seen this month?"),
     numAssignments);
)
```

https://github.com/boto/boto
http://docs.aws.amazon.com/AWSMechTurk/latest/AWSMturkAPI/Welcome.html

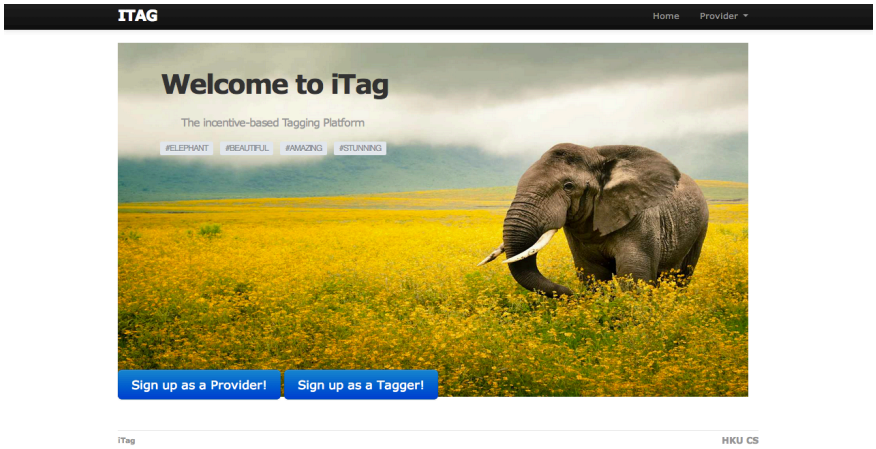
Crowd Sourcing API

22

- **Consumers**
  - Providing Problems to be solved by Crowd
  - Generating Human Intelligent Tasks
    - Consumer-generated
    - Crowd sourcing service providers assisted (manual, semi-automated, fully automated)
- **Contributors** (Crowd)
- **Crowd Sourcing Service Providers**
  - how many workers does a HIT needs (Plurality)
  - how to integrate crowd contributions to meet the consumers' requirements
  - How to create incentives while lower the cost to make it a profitable business model

iTag (1) [ICDE'14 demo]

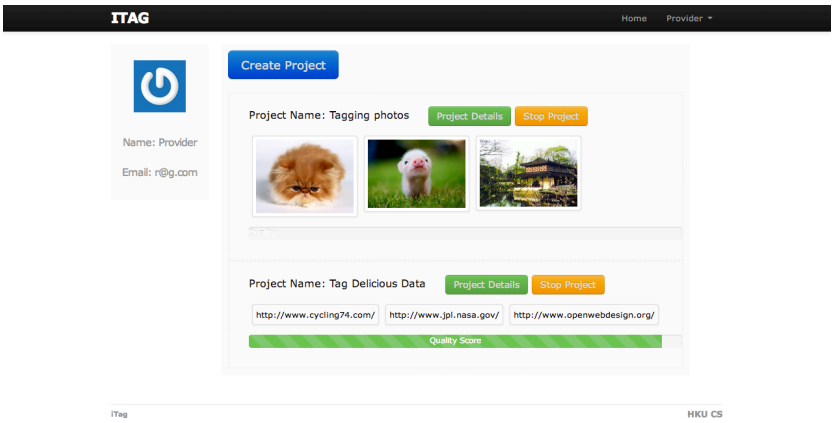
23



- Welcome page for provider

iTag (2)

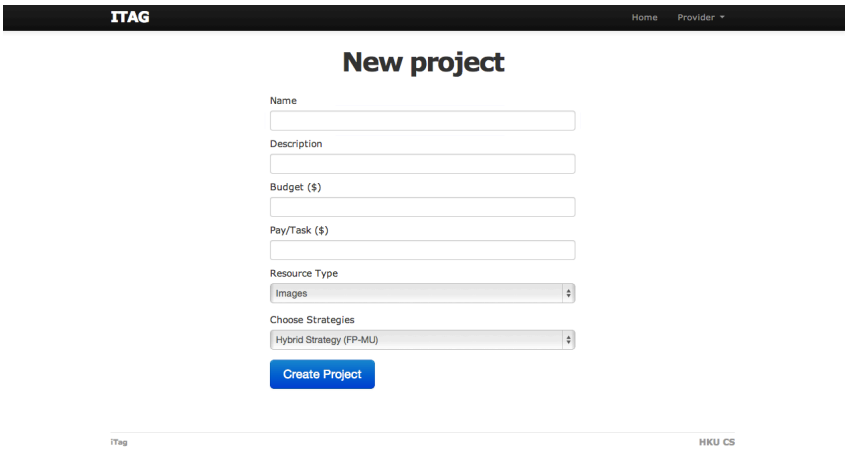
24



- Existing projects created by the Provider

iTag (3)

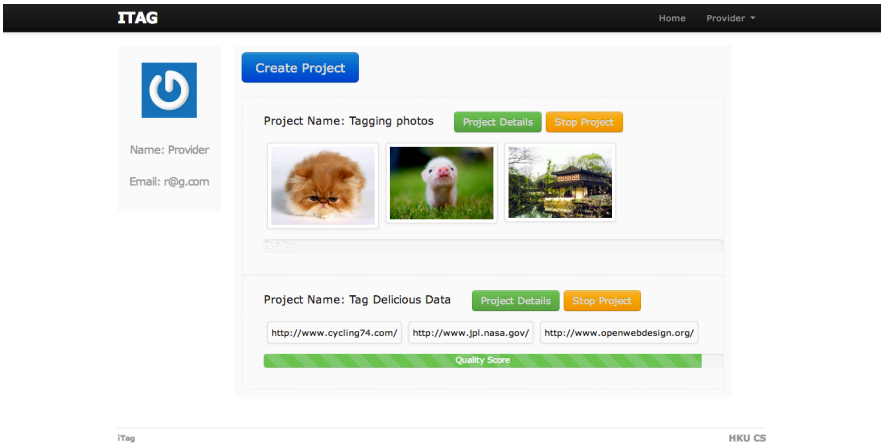
25



- Create a new project

iTag (4)

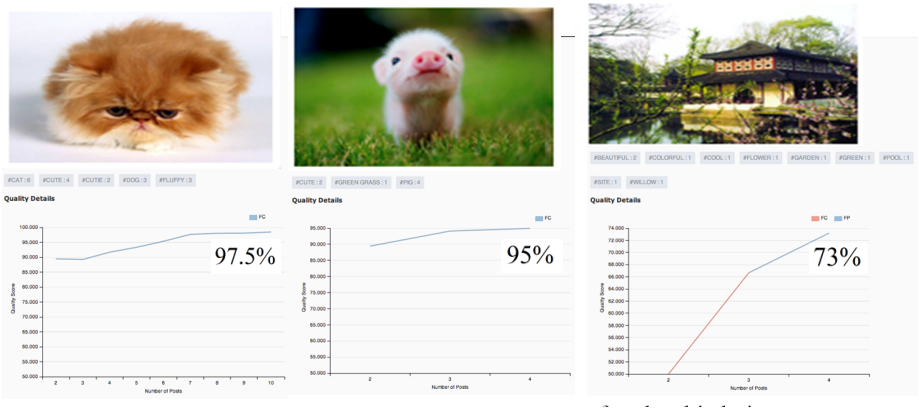
26



- A previously created project

iTag (7)

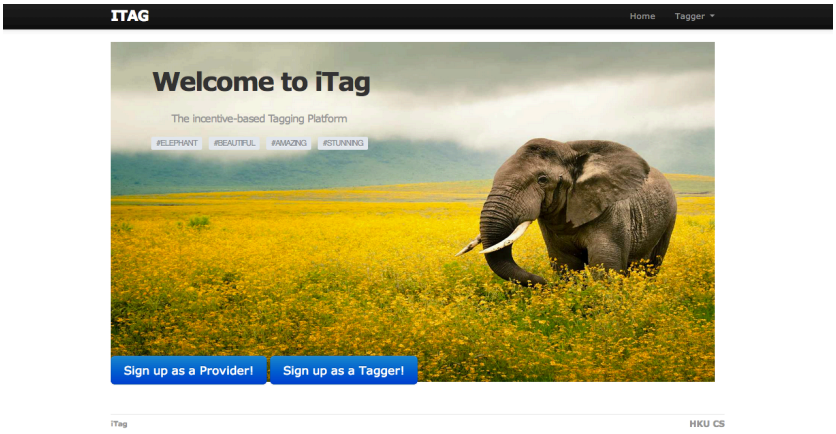
27



- The quality scores for three pictures

iTag (9)

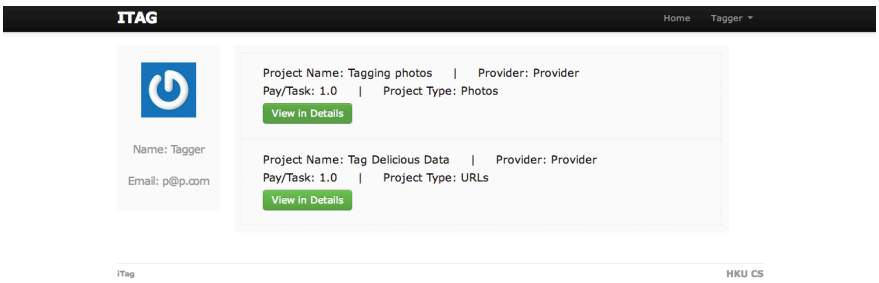
28



- Welcome page for Tagger

iTag (10)

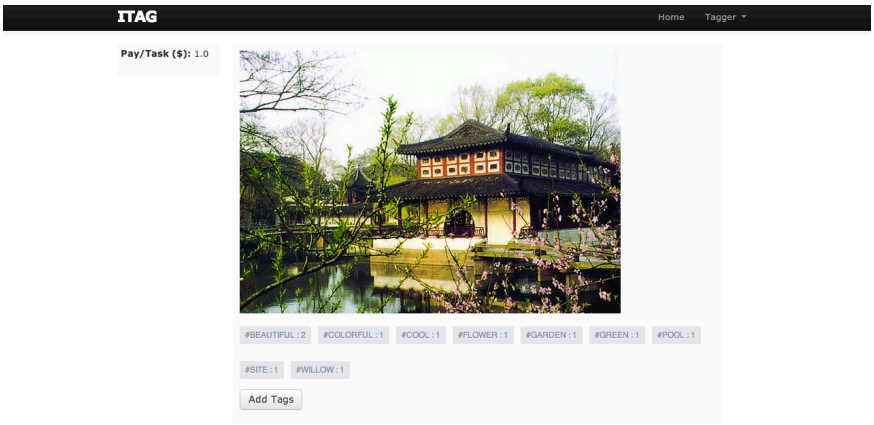
29



- Select a project to tag

iTag (11)

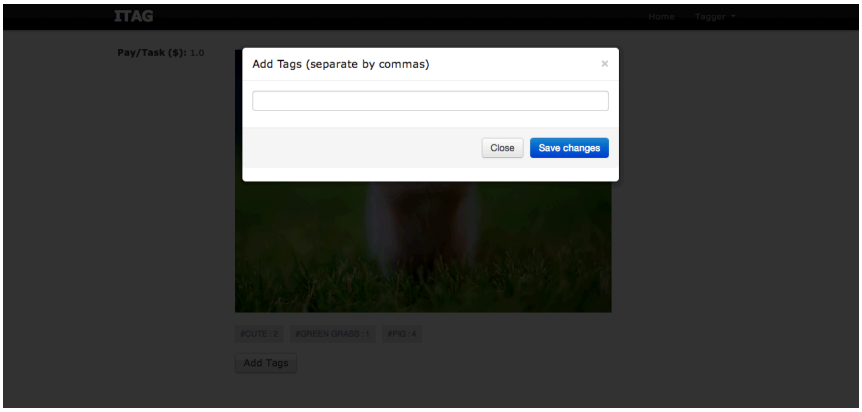
30



- A picture with existing tags

iTag (12)

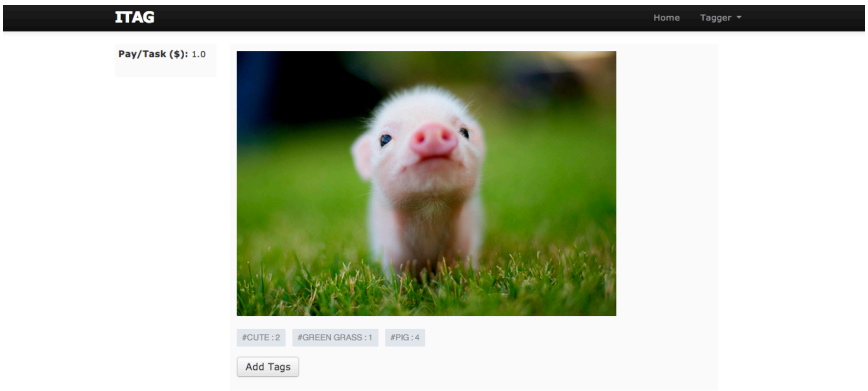
31



- Add your tags for the picture

iTag (13)

32



- Another picture for you to tag

Crowd Sourcing API

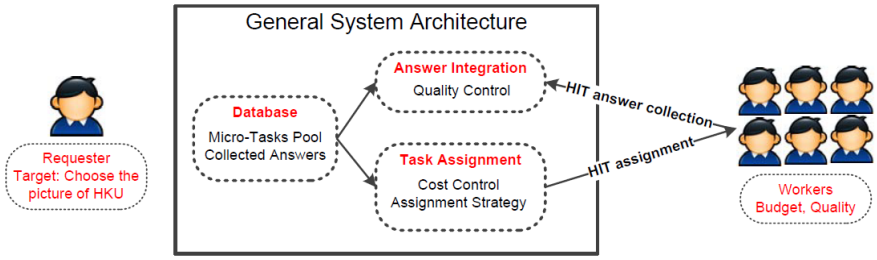
33

- **Consumers**
  - Providing Problems to be solved by Crowd
  - Generating Human Intelligent Tasks
    - Consumer-generated
    - Crowd sourcing service providers assisted (manual, semi-automated, fully automated)
- **Contributors** (Crowd)

- **Crowd Sourcing Service Providers**
  - how many workers does a HIT needs (Plurality)
  - how to integrate crowd contributions to meet the consumers' requirements
  - How to create incentives while lower the cost to make it a profitable business model

Crowdsourcing Framework

34



- 1.Answer Integration:

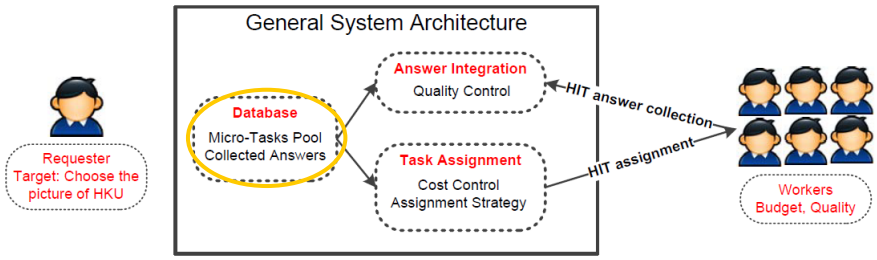
  How to integrate answers from workers ?
- 2.Task Assignment:

  Which tasks are chosen to assign to a worker ?
- 3.Database:

  How to store crowdsourced data?

Crowdsourcing Framework

35



- 1.Answer Integration:

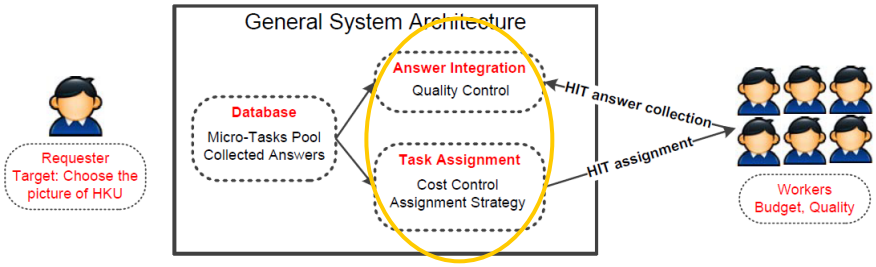
  How to integrate answers from workers ?
- 2.Task Assignment:

  Which tasks are chosen to assign to a worker ?
- 3.Database:

  How to store crowdsourced data?

Crowdsourcing Framework

36



- **1.Answer Integration:**

  How to integrate answers from workers ?

- **2.Task Assignment:**

  Which tasks are chosen to assign to a worker ?

- **3.Database:**

  How to store crowdsourced data?

Classification of HITs

37

- **Format of Questions**

  How are the questions presented to workers?

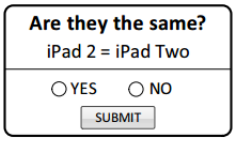
- **Nature of Answers**

  Does the question have a true answer?

Format of Questions (1)

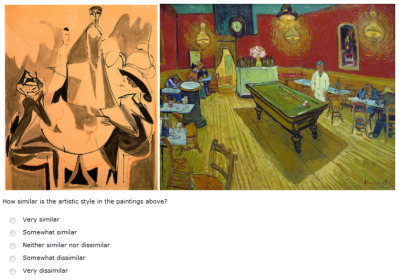
38

- Binary Choice Question (BCQ)



ER (Entity Resolution):
Are two entities equal?

- Multiple Choice Question (MCQ)

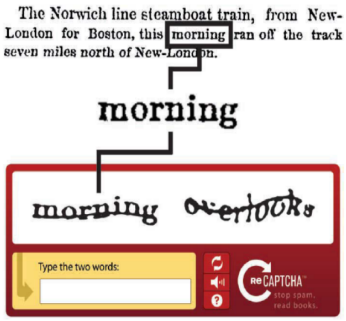


CV (Computer Vision):
similarity between two paintings

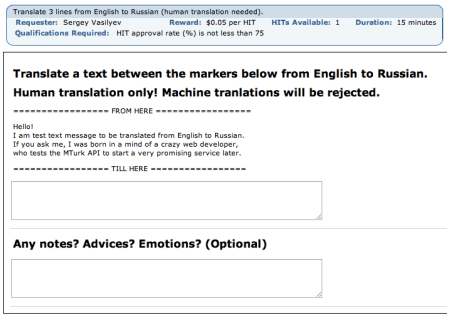
Format of Questions (2)

39

- Open questions with text answers



OCR:
reCAPTCHA



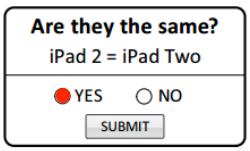
NLP:
language translation

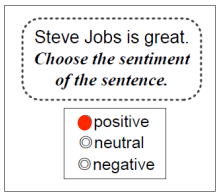
Nature of Answers (1)

40

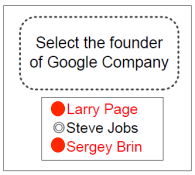
- Ground truth to a question

Single ground truth answer:

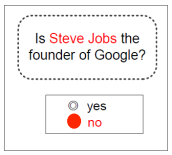




Multiple ground truth answers:



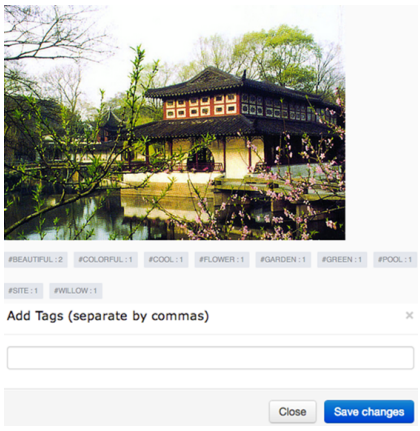
Decomposition



Nature of Answers (2)

41

- No ground truth to a question



A Classification of HITs

42

- TWO dimensions

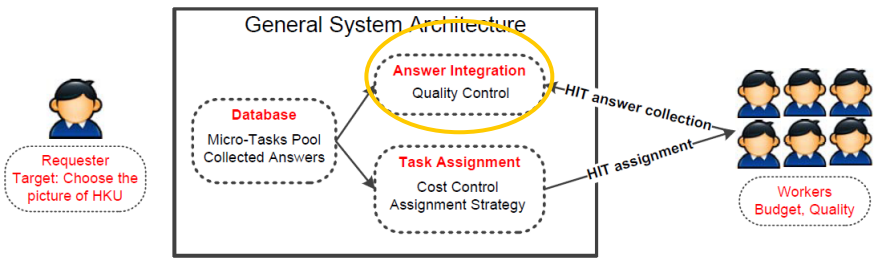
  Questions: Binary Choice, Multiple Choice / Open

  Answers: With / Without Ground Truth

Answer | Question	BCQ	MCQ	Open
With Ground Truth	ER	ER	OCR
Without Ground Truth		CV	Sheep, NLP, Tagging

Crowdsourcing System Framework

43



- 1.Answer Integration:

  How to integrate answers from workers ?
- 2.Task Assignment:

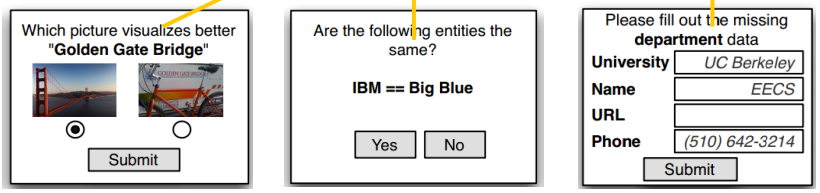
  Which tasks are chosen to assign to a worker ?
- 3.Database:

  How to store crowdsourced data?

Answers with Ground Truth

44

Answer | question	BCQ	MCQ	Open
With Ground Truth			
Without Ground Truth			



M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin. Crowddb: answering queries with crowdsourcing. In SIGMOD Conference, pages 61-72, 2011.

Voting Strategies

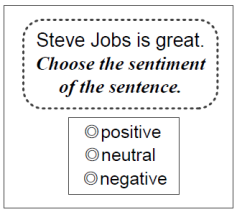
45

2 dimensions (nature | format)	BCQ	MCQ	Open
With Ground Truth			
Without Ground Truth			

- Half Voting
- Majority Voting
- Bayesian Voting

Example

46



quality		positive	neutral	negative
0.7	worker 1	+		
0.5	worker 2		+	
0.6	worker 3		+	
0.4	worker 4			+
0.3	worker 5			+
0.8	worker 6	+		
0.6	worker 7		+	

- Half Voting Output 'NULL
- Majority Voting
  Output 'neutral'

Example

47

quality		positive	neutral	negative
0.7	worker 1	+		
0.5	worker 2		+	
0.6	worker 3		+	
0.4	worker 4			+
0.3	worker 5			+
0.8	worker 6	+		
0.6	worker 7		+	

quality		positive	neutral	negative
0.7	worker 1	0.7	0.15	0.15
0.5	worker 2	0.25	0.5	0.25
0.6	worker 3	0.2	0.6	0.2
0.4	worker 4	0.3	0.3	0.4
0.3	worker 5	0.35	0.35	0.3
0.8	worker 6	0.8	0.1	0.1
0.6	worker 7	0.2	0.6	0.2

- Bayesian Voting

positive: 0.7*0.25*0.2*0.3*0.35*0.8*0.2
neutral: 0.15*0.5*0.6*0.3*0.35*0.1*0.6
negative: 0.15*0.25*0.2*0.4*0.3*0.1*0.2

normalized distribution:
(positive, neutral, negative) = (66%, 32%, 2%)

Observation

48

quality		positive	neutral	negative
0.7	worker 1	+		
0.5	worker 2		+	
0.6	worker 3		+	
0.4	worker 4			+
0.3	worker 5			+
0.8	worker 6	+		
0.6	worker 7		+	

Half Voting

Output 'NULL'

Majority Voting

Output 'neutral'

Bayesian Voting

Output a distribution (66%, 32%, 2%)

★ Bayesian Voting outputs a distribution by considering workers' qualities

Quality is important !

49

quality		positive	neutral	negative
0.7	worker 1	+		
0.5	worker 2		+	
0.6	worker 3		+	
0.4	worker 4			+
0.3	worker 5			+
0.8	worker 6	+		
0.6	worker 7		+	

- How to represent worker's quality ?
- How to derive worker's quality ?

How to represent worker's quality ?

50

- A simple parameter q in [0,1]

  q=0.7 indicates that the person will have 70% to correctly answer a question.

- Confusion Matrix M

	positive	neutral	negative
positive	0.6	0.3	0.1
neutral	0.15	0.8	0.05
negative	0.1	0.2	0.7

	positive	neutral	negative
positive	0.7	0.15	0.15
neutral	0.15	0.7	0.15
negative	0.15	0.15	0.7

X. Liu, M. Lu, B. C. Ooi, Y. Shen, S. Wu, and M. Zhang. Cdas: A crowdsourcing data analytics system.PVLDB, 5(10):1040-1051, 2012.
P. Ipeirotis, F. Provost, and J. Wang. Quality management on amazon mechanical turk. InSIGKDD workshop, pages 64-67,2010.

How to learn worker's quality ?

51

- Golden Questions

  Hire some experts to give the answers for a subset of questions, called "golden questions".

  7 correctly answered questions in 10 golden ones. q= 7/10 =0.7

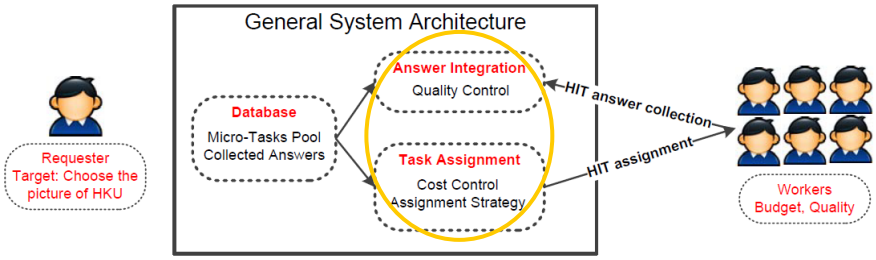
- Learning (Expectation Maximization)

  L() function is not convex/concave, iteratively construct a concave upper bound function(E-step) and update the parameters(M-step).

X. Liu, M. Lu, B. C. Ooi, Y. Shen, S. Wu, and M. Zhang. Cdas: A crowdsourcing data analytics system.PVLDB, 5(10):1040-1051, 2012.
P. Ipeirotis, F. Provost, and J. Wang. Quality management on amazon mechanical turk. InSIGKDD workshop, pages 64-67,2010.

Crowdsourcing Framework

52



- **1.Answer Integration:**
  How to integrate answers from workers ?
- **2.Task Assignment:**
  Which tasks are chosen to assign to a worker ?
- **3.Database:**
  How to store crowdsourced data?

Optimizing Plurality for HITs (CIKM'13)

53

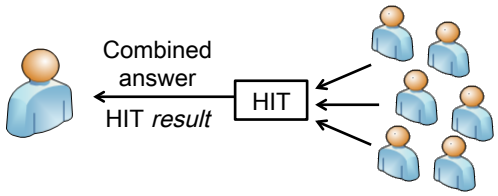
Answer | Question	BCQ	MCQ	Open
With Ground Truth	ER	ER	OCR
Without Ground Truth		CV	Sheep, NLP, Tagging

- How to set plurality for HITs?
- How to develop effective and efficient plurality setting algorithms for HITs

Plurality of HITs

54

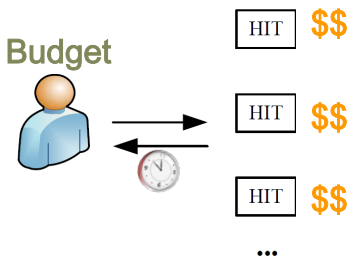
- Imperfect answers from a single worker
  - make careless mistakes
  - misinterpret the HIT requirement
- Specify *sufficient plurality* of a HIT (number of workers required to perform that HIT)



Plurality Assignment Problem (PAP)

55

- Plurality has to be *limited*
  - A HIT is associated with a cost
  - Requester has limited budget
  - Requester requires time to verify HIT results



PAP:
wisely assign the right pluralities to various HITs to achieve overall high-quality results

Interesting Problem

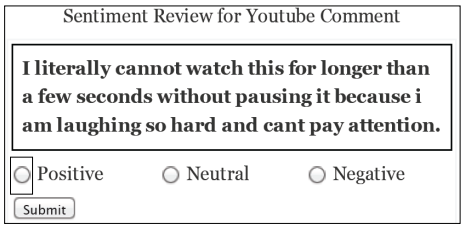
56

- Manually assigning pluralities is tedious if not infeasible
  - AMT on 28th October, 2012
  - 90,000 HITs submitted by Top-10 requesters
- Algorithms for automating the process of plurality assignment are needed!

Multiple Choice Questions (MCQs)

57

- Most popular type
  - AMT on 28th Oct, 2012
  - About three quarters of HITs are MCQs
- Examples
  - Sentiment analysis, categorizing objects, assigning rating scores, etc.



Data Model

58

- Set of HITs $\mathcal{T} = \{t_1, t_2, \ldots, t_n\}$
- For each HIT t_i
  - contains a single MCQ
  - plurality k_i (i.e., k_i workers are needed)
  - cost c_i (i.e., c_i units of reward are given for completing t_i)

Quality Model

59

- Capture the goodness of HIT result's for t_i
- MCQ quality
  - likelihood that the result is correct after it has been performed by *k* workers
- Factors that affect MCQ quality
  - plurality: *k*
  - Worker's accuracy (or accuracy): p_i for HIT t_i
    - probability that a randomly-chosen worker provides a correct answer for t_i
    - estimated from similar HITs whose true answer is known

$$\zeta_i(k) = \sum_{l=\lceil \frac{k}{2} \rceil}^{k} \binom{k}{l} p_i^l (1-p_i)^{(k-l)}$$

Problem Definition

60

- Input
  - budget B
  - Set of HITs $\mathcal{T}$, and $\zeta_i(k), c_i$ for $t_i \in \mathcal{T}$
- Output
  - plurality for every HIT $\vec{k} = \{k_1, k_2, \ldots, k_n\}$
- Objective
  - maximize overall average quality

$$\mathcal{Q}(\mathcal{T}, \vec{k}) = \frac{1}{n}\sum_{i=1}^{n}\zeta_i(k_i) \qquad \zeta_i(k) = \sum_{l=\lceil \frac{k}{2} \rceil}^{k}\binom{k}{l}p_i^l(1-p_i)^{(k-l)}$$

Solutions

61

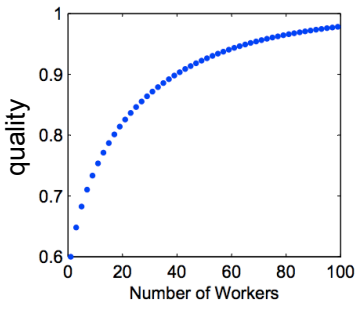
$$\begin{aligned} &\textbf{given} && B, \mathcal{T} \text{ and } p_i, c_i \text{ for } t_i \in \mathcal{T} \\ &\textbf{maximize} && \mathcal{Q}_t(\mathcal{T}, \vec{k}) \\ &\textbf{subject to} && \sum_{i=1}^{n} k_i c_i \leq B \ \textbf{and} \\ & && k_i = 0 \text{ or a positive odd number} \end{aligned}$$

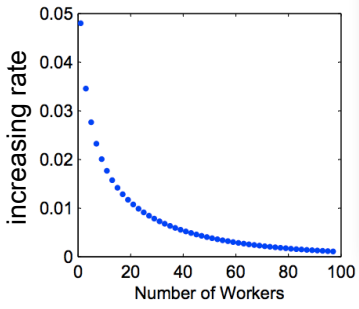
- Optimal Solution
  - Dynamic Programming
  - Not efficient for HIT sets that contain thousands of HITs
    - 60,000 HITs extracted from AMT
    - Execution time: 10 hours

Properties

62

- Monotonicity and Diminishing Return
  - the quality function increases with plurality;
  - the rate of quality improvement drops with plurality.





Greedy: 2-approximate algorithm

63

- Properties of MCQ quality function
  - Monotonicity
  - Diminishing return
  - Plurality Assignment Problem (PAP) is approximable for HITs with these two properties
- Greedy
  - Select the "best" HIT and increase its plurality until budget is exhausted
  - Selection criteria: the one with largest *marginal gain*
  - Theoretical approximation ratio = 2

Grouping techniques

64

- Observations
  - Many HITs submitted by the same requesters are given the same cost and of very similar nature
- Intuition
  - Group HITs of the same cost and quality function
  - More or less the same plurality for HITs in one group
- Main idea
  - Select a "representative HIT" from each group
  - Evaluate its plurality by DP or Greedy
  - Deduce each HIT's plurality from the representative HIT

Experiments

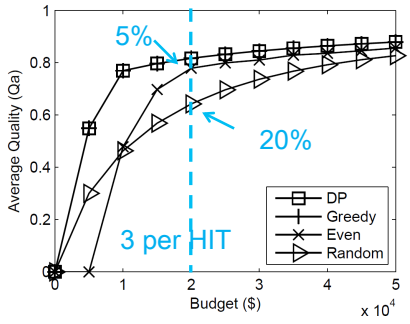
65

- Synthetic
  - Generated based on the extraction of an AMT requester's HITs information on Oct 28th, 2012
  - Statistics
    - 67,075 HITs
    - 12 groups (same cost and accuracy)
    - Costs vary from $0.08 to $0.24
    - Accuracy of each group is randomly selected from [0.5, 1]

Effectiveness

66

- Competitors
  - *Random*: arbitrarily pick a HIT to increase its plurality until budget is exhausted
  - *Even*: divide the budget evenly across all HITs

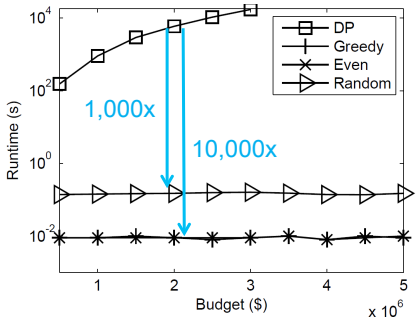


Greedy is close-to-optimal in practice

Performance (1)

67

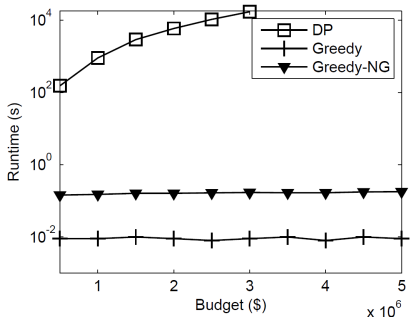
- DP and Greedy are implemented using grouping techniques
- Greedy is efficient!



Performance (2)

68

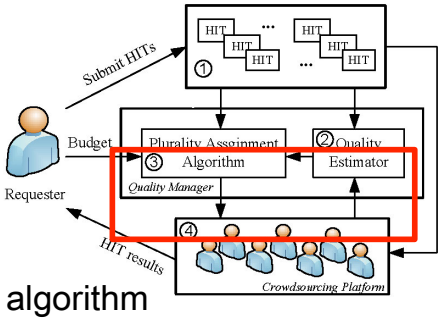
- Grouping techniques
  - 20 times faster than non-group solutions
    - 12 groups vs. 60,000 HITs



Other HIT Types

69

- Examples
  - Enumeration Query
  - Tagging Query
- Solution framework
  - Quality estimator
    - Derive $\zeta_i(k)$
    - accuracy p_i in MCQ
  - Plurality Assignment algorithm
    - Greedy for HITs demonstrate monotonicity and diminishing return



Summary

70

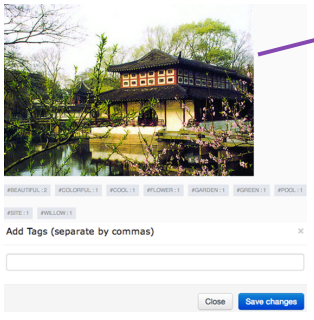
Answer/Question	BCQ	MCQ	Open
With Ground Truth	ER	ER	OCR
Without Ground Truth		CV	Sheep, NLP, Tagging

- Problem of setting plurality for HITs
- Develop effective and efficient plurality algorithms for HITs

Tagging

71

2 dimensions (nature | format)	BCQ	MCQ	Open
With Ground Truth	ER	ER	OCR
Without Ground Truth		CV	Sheep, NLP, Tagging



Worker inputs tags for the picture

Open question without ground truth

X. S. Yang, D. W. Cheung, L. Mo, R. Cheng, and B. Kao. On incentive-based tagging. In Proc. of ICDE, pages 685-696. 2013

Siyu Lei, Xuan S. Yang, Luyi Mo, Silviu Maniu, Reynold Cheng iTag: Incentive-Based Tagging. ICDE 2014 demo.

On incentive-based tagging (ICDE'13)

72

2 dimensions (nature | format)	BCQ	MCQ	Open
With Ground Truth	ER	ER	OCR
Without Ground Truth		CV	Sheep, NLP, Tagging

- How to define Tag Data Quality
- Incentive-Based Tagging
- How to select pictures for workers to tag

Collaborative Tagging Systems

73

- Example:
  - Delicious, Flickr
- Users / Taggers
- Resources
  - Webpages
  - Photos
- Tags
  - Descriptive keywords
- **Post**
  - **Non-empty set of tags**

Applications with Tag Data

74

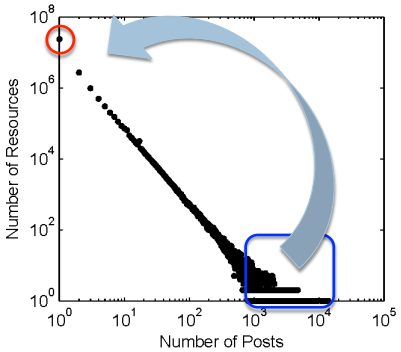
- Search
- Recommendation
- Clustering
- Concept Space Learning

Optimizing web search using social annotations. S. Bao et al. WWW'07
Can social bookmarking improve web search? P. Heymann et al. WSDM'08
Structured approach to query recommendation with social annotation data. J. Guo CIKM'10
Clustering the tagged web. D. Ramage et al. WSDM'09
Exploring the value of folksonomies for creating semantic metadata. H. S. Al-Khalifa IJWSIS'07

Incentive-Based Tagging

75

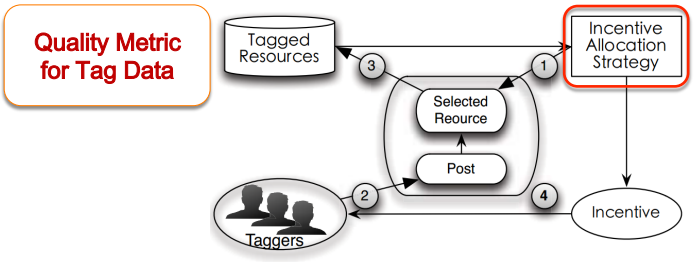
- Guide users' tagging effort
  - **Reward** users for annotating under-tagged resources
- Reduce the number of under-tagged resources
- Save the tagging efforts wasted in over-tagged resources



Incentive-Based Tagging (cont'd)

76

- Limited Budget
- Incentive Allocation
- Objective: Maximize Quality Improvement



Summary

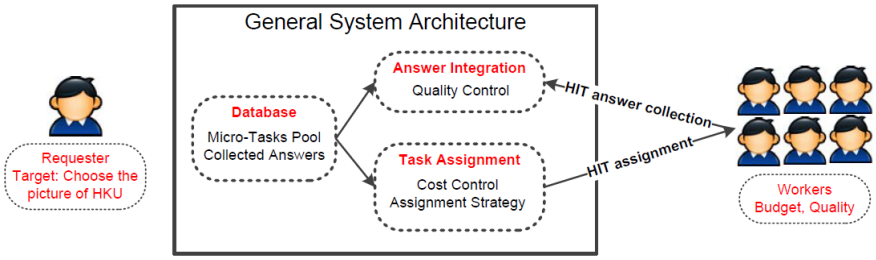
77

2 dimensions (nature | format)	BCQ	MCQ	Open
With Ground Truth	ER	ER	OCR
Without Ground Truth		CV	Sheep, NLP, Tagging

- Define Tag Data Quality
- Incentive-Based Tagging
- Effective Assignment Strategies

Summary : Framework

78



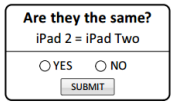
- **1.Answer Integration:**
  How to integrate answers from workers ?
- **2.Task Assignment:**
  Which tasks are chosen to assign to a worker ?
- **3.Database:**
  How to store crowdsourced data?

Summary:Answer Integration

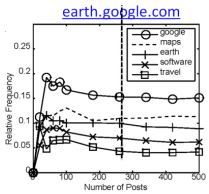
79

2 dimensions (nature | format)	BCQ	MCQ	Open
With Ground Truth	ER	ER	OCR
Without Ground Truth		CV	Sheep, NLP, Tagging

- Half Voting [CIKM'13]



- Relative Frequency
  [ICDE'13, ICDE'14]



Summary:Task Assignment

80

2 dimensions (nature | format)	BCQ	MCQ	Open
With Ground Truth	ER	ER	OCR
Without Ground Truth		CV	Sheep, NLP Tagging

- Assignment Strategy[ICDE'13, ICDE'14]

  Free Choice, Round Robin, FPF,…

- Cost Control [CIKM'13]

Challenge 1: Manage answer integration & task assignment

81

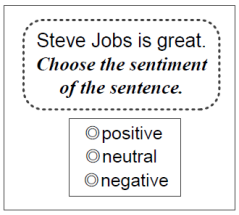
2 dimensions (nature | format)	BCQ	MCQ	Open
With Ground Truth	ER	ER	OCR
Without Ground Truth		CV	Sheep, NLP Tagging

How about majority voting and Bayesian voting?

Which kind of answer integration techniques are better?

Challenge 2: Uncertainty of Crowdsourced Data

82



Bayesian voting
('positive':66%,
'neutral':32%,
'negative':2%)

How can database handle this uncertain data (or data with distribution) ?

Traditional and Crowdsourced database does not provide support for handling uncertain data

Types of Uncertainty

83

- Tuple-level uncertainty

A tuple is a random variable that has a Boolean domain: it is true when the tuple is present and false if it is absent

entity	entity	prob.
ipad2	Ipad two	0.8
ipad2	Iphone 2	0.3

- Attribute-level uncertainty

An attribute represents a random variable, whose domain is the set of values that may take

No.	sentiment		
1	positive	Neutral	Negative
	0.66	0.32	0.02

D. Suciu, D. Olteanu, C. R´e, and C. Koch. Probabilistic Databases . Morgan-Claypool, 2011.

Challenge 3: Machine Learning and Crowdsourcing

84

- **Can we develop a better method which gets the best from the two communities?**

- For example,
  Apply machine learning to handle some clear cases
  Use human effort to work on the difficult questions