Collective Intelligence: Crowdsourcing groundtruth data for large scale evaluation in Information Retrieval

Sergej Zerr
S.Zerr@soton.ac.uk
Challenges in IR Evaluation

- **BigData**
  - Heterogeneity (larger annotation demand)
  - Dynamicity (updates required)
  - Novel tasks (no test collections)

- Relevance ranking
- Search result diversification
- Temporal retrieval
- etc.

### Document Set

#### Test Collections

<table>
<thead>
<tr>
<th>query</th>
<th>documents</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D_1</td>
</tr>
<tr>
<td></td>
<td>D_2</td>
</tr>
<tr>
<td></td>
<td>:</td>
</tr>
<tr>
<td></td>
<td>D_k</td>
</tr>
</tbody>
</table>
Challenges in IR Evaluation

- Better human accessibility
  - WiFi, Mobile Networks, Portable gadgets (larger crowd)

- Challenges:
  - How to motivate the crowd to work?
  - How to obtain meaningful results from the individuals?
  - How to aggregate the crowdsourced results?
  - How to evaluate the output?
Outline

• Collaborative Advantages
  • The wisdom of crowds
  • Conditions for a successful collaboration

• Obtaining collaborative knowledge
  • Crowd motivation
  • Scalability/Efficiency
  • Own work

• Input/Output Evaluation
  • Users and Data
  • Quality assurance

• Discussion
Collaboration

Often we need more than one hand

Also more than one brain

“Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations”
James Suroewicki

.. The Wisdom of Crowds ..
In 1906, the statistician Francis Galton observed a competition at a country fair. The crowd accurately guessed the weight of an ox when their individual guesses were averaged (the average was closer to the ox's true butchered weight than the estimates of most experienced crowd members)
Crowd IQ: aggregating opinions to boost performance

United Brains

Equivalent, or greater, utility under the curve

Utility

# of contributors

Expert

Masses

10

10,000+

10,000+
United Brains

4,000 experts, 80,000 articles, 200 years to develop, Annual Updates

>100,000 amateurs, 1.6 Million articles, 5 years to develop, Real-Time Updates

Utility

# of contributors

4,000

100,000+
(In,-) Direct Collaboration in IR can be used:

- Collaborative tagging,
- Favorite assignments,
- Click logs,
- Data partitioning,
- Recommendations,
- etc., etc., etc....
Collaboration: Paradox

- Using „Wisdom of Crowds” is not always straight-forward to achieve.

- Collaborative work needs to be managed efficiently

- Kasparov won against the world in 1999

http://en.wikipedia.org/wiki/Kasparov_versus_the_World
Collaboration: Success Criteria

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diversity of opinion</td>
<td>Each person should have “private” information.</td>
</tr>
<tr>
<td>Independence</td>
<td>People's opinions aren't determined by the opinions of those around them.</td>
</tr>
<tr>
<td>Decentralization</td>
<td>People are able to specialize and draw on local knowledge.</td>
</tr>
<tr>
<td>Aggregation</td>
<td>Effective mechanism exists for turning private judgments into a collective</td>
</tr>
</tbody>
</table>
Groupthink Symptoms:
Irving Lester Janis (26 May 1918 - 15 November 1990)

- Collective rationalization
- Self-censorship
- Direct pressure on dissenters
- Self-appointed 'mindguards'

https://www.youtube.com/watch?v=fuIXiXqv978
Collaboration

“The best collective decisions are the product of disagreement and contest, not consensus or compromise.”

“The best way for a group to be smart is for each person in it to think and act as independently as possible.”
Outline

- Collaborative Advantages
  - The Wisdom of Crowds
  - Conditions for a successful collaboration

- **Obtaining collaborative knowledge**
  - Crowd motivation
  - Scalability/Efficiency
  - Own work

- Input/Output Evaluation
  - Users and Data
  - Quality assurance

- Discussion
Humans can (yet) solve some tasks more efficient and/or accurate as a machine would do.

- Captcha (OCR)
- Classification
- Image tagging
- Speech recognition
- Face/emotion recognition
Declarative Crowdsourcing Systems

Michael J. Franklin et al: CrowdDB: answering queries with crowdsourcing, SIGMOD 2011
A. Marcus et al: Human-powered Sorts and Joins, VLDB 2011

SELECT c.name
FROM celeb c JOIN photos p
ON samePerson(c.img,p.img)
TASK samePerson(f1, f2) TYPE EquiJoin:
SingluarName: "celebrity"
LeftPreview: "<img src='%s'>",tuple1[f1]
RightPreview: "<img src='%s'>",tuple2[f2]
Combiner: MajorityVote

SELECT image i FROM serengety
ORDER BY CROWDORDER (i, "Which image contains more baby animals");
Gathering Input, Reusing “Natural” Human Power

- Google Flu Trends
- eBird

Report and track the birds online – any place, any time!
Human Computation Platforms and Motivation for Participation

Citizen Science

- Helping/Contribute to something important
- Social pressure
- Virtual goods
- Competitions
- Gaming
- Money

Paid Crowdsourcing
Monetary based Motivation

24.07.2016

2nd Keystone Training School, Sergej Zerr
Mturk: IR Example – Snippet Evaluation

- Study on summary lengths
- Determine preferred result length
- Asked workers to evaluate snippet quality

- Payment between $0.01 and $0.05 per HIT
- 12,790 queries - 40K judgments 400$-2000$ (300h of work)

IR Example – Relevance Assessment

- Replace TREC-like relevance assessors with MTurk?
- Selected topic “space program” (011)
- Modified original 4-page instructions from TREC
- Workers more accurate than original assessors!
- 40% provided justification for each answer

- Payment between $0.02 per HIT
- 1 topic, 29 documents - 290 judgments (6$)


July 24, 2011 Crowdsourcing for Information Retrieval: Principles, Methods, and Applications 51
Games

- ESP Game: label images
  - Image retrieval by text
- Squigl: match the labels to areas
  - Object recognition
- Matchin: find the better image
  - Image ranking
- FlipIt: memory with similar images
  - Near duplicate detection

- Other areas covered as well: label songs, find synonyms, describe videos
- See: www.gwap.com by Luis von Ahn
Useful human power for annotating the Web

- 5000 people playing simultaneously could label all images on Google in 30 days!
- Individual games in Yahoo! and MSN average over 5,000 players at a time
Competition based Motivation (Image Privacy)
Gathering average community notion of privacy

- We crawled “most recently uploaded” Flickr photos (2 Months)
- Started a social annotation game (over the course of 2 weeks)
- 81 users (colleagues, social networks, forum users), 6 teams
- Collected around 30K annotated photos

Sergej Zerr, Stefan Siersdorfer, Jonathon Hare, Elena Demidova  Privacy-Aware Image Classification and Search, SIGIR’12
The GUI for Privacy Aware Image IR

(a) Web service GUI for privacy-oriented image classification.

(b) Search results for the query “cristiano ronaldo” (06/06/12).
Motivation: Add Social Pressure

Combine Gamification, Competition and Money

**Problem:** improve time aware cost effectiveness of crowdsourcing

- **Individual reward mechanisms**
  - Competitive game designs for improving the cost effectiveness of crowdsourcing
  - CIKM'14

- **Team-based reward mechanisms**
  - Groupsourcing: Team competition designs for crowdsourcing
  - WWW'15

- **Temporal-based crowdsourcing performance**
  - Just in Time: Controlling Temporal Performance in Crowdsourcing Competitions
  - WWW'16
Reward Distribution 1: “Pay-per-Task” (Baseline)

- Fixed reward rate $c$ ($ per task) for each worker
- Reward of workers proportional to value produced by worker (e.g. no. of annotations, ratings, etc.)

M. Rokicki, S. Chelaru, S. Zerr, and S. Siersdorfer. Competitive game designs for improving the cost effectiveness of crowdsourcing. CIKM'14
Reward Distribution 2: Competitions

- Workers compete during limited time period
- Workers obtain scores based on their performance (e.g. no. of tasks fulfilled)
  - Ranking of the workers based on their performance
  - Distributing of the rewards according to the rank

![Diagram of Reward Distribution Methods]

- "Winner-takes-All"
- "Exp-decreasing"
Workers’ View: Tasks
Information Policies

- How much information about fellow workers to provide during competition?
- Information: scores, rank

<table>
<thead>
<tr>
<th>Rank</th>
<th>Worker</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>abc</td>
<td>1128379</td>
</tr>
<tr>
<td>2</td>
<td>xcvasl</td>
<td>231977</td>
</tr>
<tr>
<td>3</td>
<td>wpoe</td>
<td>123232</td>
</tr>
<tr>
<td>4</td>
<td>afaolik</td>
<td>99934</td>
</tr>
<tr>
<td>5</td>
<td>qwope</td>
<td>65312</td>
</tr>
<tr>
<td>6</td>
<td>Xcmy</td>
<td>42387</td>
</tr>
<tr>
<td>7</td>
<td>GKafskdö</td>
<td>3628</td>
</tr>
</tbody>
</table>

Tradeoff

Lots of information

Low amount of information

Sweet spot

? Myself 567

124 Myself 567
# Performance of Strategies (Captcha Task)

<table>
<thead>
<tr>
<th>Experiment</th>
<th>No. Captchas</th>
<th>USD/ Hour</th>
<th>Cent/Captcha</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Performance-based payment</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>exp-open</td>
<td>67,951</td>
<td>0.300</td>
<td>0.074</td>
</tr>
<tr>
<td><strong>exp-med</strong></td>
<td>154,188</td>
<td>0.138</td>
<td>0.032</td>
</tr>
<tr>
<td>exp-res</td>
<td>25,853</td>
<td>0.605</td>
<td>0.193</td>
</tr>
</tbody>
</table>

| pay-per-task | 58,635 | 0.352 | 0.077 |
Competition among Top-10 Workers

(a) Baseline

(b) Exp-med

R=1
R=2
R=3

R=4
R=5
R=6

R=7
R=8
R=9

R=10
Team-based reward mechanisms
Can we use work groups to further improve the performance?

monetary incentives

individuals

teams

competitions

M. Rokicki, S. Zerr, and S. Siersdorfer. Groupsourcing: Team competition designs for crowdsourcing. WWW’15
Rewards:
- Non-linear distribution among teams
- Individual share proportional to contribution

Communication:
- Team chats with notifications

Combinations with individual reward
- balanceTS
- ind-balanceTS
# Performance of the Strategies

<table>
<thead>
<tr>
<th>Experiment</th>
<th>No. Images</th>
<th>Cent / Hour</th>
<th>Cent / 100 Images</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ind</td>
<td>298,332</td>
<td>9.895</td>
<td>3.352</td>
</tr>
<tr>
<td><strong>Balanced Teams</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>balanceTS</td>
<td>327,073</td>
<td>8.967</td>
<td>3.057</td>
</tr>
<tr>
<td>ind-balanceTS</td>
<td>391,620</td>
<td>8.059</td>
<td>2.553</td>
</tr>
</tbody>
</table>
Results: Team Contributions

(a) balanceTS

(b) ind-balanceTS
Workers Interaction

- Communication in team chats
  - 2,500 messages by over 200 participants
- Encouragement
- Help and clarification of rules
- Discussing strategy
- Democratic team administration
- Discussing our strategies

**Messages:**

- **User 1:** Lets go team!!! we are 5, team A are 3. We can reach them!!!
- **User 2:**
  - What if I answer wrong?
  - We will lose 20 points :)

**User 1:**
- Im trying to get to number 5 spot because he/she stopped clicking.
- User 2:
  - Yeah but u need 2000 thousand more buddy, and you know that he/she will be careful now :/
  - she will check again to see if you will attack and then he/she will start doing more [...]
- **User 1:**
  - good point

**Message:**

- [...] this system.. its stable and perfect.. all in our hands(public) but not of system automatically selecting arranging them in teams..
Temporal-based crowdsourcing performance
Can we control the crowd to annotate at right times?

Peak vs NonPeak

(a) short notice, low bonus

(f) long notice, high bonus
Number of correct annotations per minute around a typical bonus hour.
Annotation dynamics

(a) Baseline individual scores
(b) Long notice, medium bonus individual scores
Workers Interaction

- Communication in team chats
  - 3,400 messages by over 200 participants

Bonus coming 14 minutes from now. Prepare everyone, we must try that first place :D

Our rank fall from rank 4-5 to 12 bcoz the other team work on time of bonus

As soon as they announce the time of the next bonus I will email you
Contributions

**Individual reward mechanisms**
- Competitive game designs for improving the cost effectiveness of crowdsourcing
  - CIKM’14

**Team-based reward mechanisms**
- Group sourcing: Team competition designs for crowdsourcing
  - WWW’15

**Temporal-based crowdsourcing performance**
- Just in Time: Controlling Temporal Performance in Crowdsourcing Competitions
  - WWW’16

---

Medium Information policy and exponential rewards significantly increase crowd performance by 300%

Balanced teams + individual rewards further increase performance and make the work more attractive (+30%)

Framework using our strategies additionally increased output in peak times by more than 300%
Motivation: Contribute to Science ("Zooniverse")

https://www.zooniverse.org/
Motivation: Contribute to Science ("Cities at Night")

- Classification of night photos from ISS to estimate artificial light pollution in cities
- Observe temporal development, measure impact on citizens and biosphere

http://stars4all.eu
http://www.citiesatnight.org/
Motivation: Contribute to Science (“Cities at Night”)

- Task 1: “Dark skies” – Find night cities in a photo stream (over 100K annotated)

- Task 2: “Lost at night” – identify the city on the photograph (around 500 identified)

- Task 3: “Night cities” – position, rotate and scale the image to the map.
Combine Human and Machine Input

- Assign DBPedia class to entities
  - Baby food, Petroleum industry in Nigeria, Light infantry
  - "Region", "Locality", "Settlement"

Using microtasks to crowdsource Dbpedia entity classification: A study in workflow Design. Qiong Bu, Elena Simperl, Sergej Zerr and Yunjia Li
Combine Human and Machine Input
Output Aggregation

- Statistical models
  - Majority voting
- Graphical models
- Optimization models
Outline

• Collaborative Advantages
  • The Wisdom of Crowds
  • Conditions for a successful collaboration

• Obtaining collaborative knowledge
  • Gathering Data from Social Web / Mechanical Turk
  • From Data to knowledge (Applications)
  • Own work

• Input/Output Evaluation
  • Users and Data
  • Quality assurance

• Discussion
Asking questions

- Ask / formulate the right questions
- Part art, part science
- Instructions are key
- Workers may not be IR experts (don’t assume the same understanding in terms of terminology)
- Show examples

Quality: Ambiguity and Subjectivity

What is relevant?
“Snow. Snow is relevant.”

„Alice saw Bob with the binoculars“
Quality: Data from Social Web

- Simple random sample can result in a set dominated by few power user
Demographic Bias (Zooniverse)

<table>
<thead>
<tr>
<th>Region</th>
<th>Classifications</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe (UK, Germany, France)</td>
<td>3688453</td>
<td>48.2%</td>
</tr>
<tr>
<td>North America (USA, Canada, Mexico)</td>
<td>3071134</td>
<td>40.2%</td>
</tr>
<tr>
<td>Oceania (Australia, New Zealand, Tanzania)</td>
<td>347818</td>
<td>4.6%</td>
</tr>
<tr>
<td>Asia (Singapore, India, Japan)</td>
<td>277536</td>
<td>3.6%</td>
</tr>
<tr>
<td>Far East</td>
<td>37278</td>
<td>0.5%</td>
</tr>
<tr>
<td>Middle East</td>
<td>15318</td>
<td>0.2%</td>
</tr>
<tr>
<td>South America (Brazil, Argentina, Chile)</td>
<td>154807</td>
<td>2.0%</td>
</tr>
<tr>
<td>Africa, Egypt, Kenya</td>
<td>50045</td>
<td>0.7%</td>
</tr>
</tbody>
</table>

Rater Reliability “Where is the cat?”

Results
Quality Assurance

- Qualification Tests
- Test questions
- „Static“ Honeypots
- „Dynamic“ honeypots
- Workers‘ reputation mechanisms
- Inter-rater agreement
Test Questions

Throw the coin and tell us the result

• Head
• Tail

Results

• **Head 61**
• **Tail 39**

People often tend just to select the first option 😊

Better: Some preliminary textual answer

• Coin type?
• Head or tail.

Matthew Lease and Omar Alonso: http://de.slideshare.net/mattlease/crowdsourcing-for-search-evaluation-and-socialalgorithmic-search
Honeypots

- **Static honeypots**
  - Let the workers perform the task. Reject the results with honeypot errors

- **Dynamic batches with injected honeypots**
  - Only reject the low quality batches

```
| b1 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |

| b2 |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
```
Measure the Inter–Rater Reliability

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>No</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Naive approach: 3 cases out of 6 = 0.5 agreement

Kilem L. Gwet, Handbook of inter-rater reliability 2010

<table>
<thead>
<tr>
<th></th>
<th>Attractive?</th>
</tr>
</thead>
<tbody>
<tr>
<td>MR_GOOD</td>
<td>1 1 0 1</td>
</tr>
<tr>
<td>MR_GOOD</td>
<td>0 1 0 1</td>
</tr>
</tbody>
</table>

24.07.2016
2nd Keystone Training School, Sergej Zerr
First experiment:
Throw the dices with the right hand 10 times.

\begin{itemize}
  \item \includegraphics[width=0.3\textwidth]{left_hand}
  \item Compute the average 3.7
\end{itemize}

Second experiment:
Throw the dices with the left hand 10 times.

\begin{itemize}
  \item \includegraphics[width=0.3\textwidth]{right_hand}
  \item Compute the average 4.4
\end{itemize}

Claim that the left hand is better.....
Inter–Rater Reliability: Cohen’s Kappa (1960)

- Idea: We need to remove agreement achieved just by chance

\[ \hat{\kappa} = \frac{p_a - p_e}{1 - p_e} \]

\[ p_a = \frac{n_{11} + n_{22}}{n_{11} + n_{12} + n_{21} + n_{22}} \]

\[ p_e = \frac{n_{11} * n_{22}}{100} + \frac{n_{12} * n_{21}}{100} + \frac{n_{21} * n_{22}}{100} \]

\[ \hat{\kappa} = \frac{.75 - .49}{1 - .49} = .51 \]
Inter–Rater Reliability: Missing Values

- Idea: Use partial ratings to estimate the marginal probability only

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>X</td>
</tr>
<tr>
<td>---</td>
<td>-----</td>
<td>----</td>
<td>---</td>
</tr>
<tr>
<td>B</td>
<td>Yes</td>
<td>30</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>5</td>
<td>34</td>
</tr>
<tr>
<td>X</td>
<td>5</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>total</td>
<td>40</td>
<td>55</td>
<td>5</td>
</tr>
</tbody>
</table>

\[ p_a = \frac{n_{11} + n_{22}}{n - (n_{x1} + n_{x2} + n_{1x} + n_{2x})} = \frac{30 + 34}{100 - (5 + 8)} = .74 \]

\[ p_e = \frac{50}{100} \ast \frac{40}{100} + \frac{42}{100} \ast \frac{55}{100} = 0.431 \]

\[ \hat{\gamma}_k = \frac{0.74 - .431}{1 - .431} = .54 \]
Inter–Rater Reliability: Extensions

- **Multiple Raters/Categories:**
  - Fleiss 1971 – Average over random pairs of raters for random objects

- **Adjustment for Ordinal and Interval Data, Weighting:**
  - weight judgments using distances between categories.
  - Measures: $AC_1$, $AC_2$ (ordinal and interval data)

- **Check for statistical significance:**
  - The number of categories and/or raters matters.

Inter–Rater Reliability: Kappa Interpretations

<table>
<thead>
<tr>
<th>Kappa</th>
<th>Strength of Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;0.0</td>
<td>Poor</td>
</tr>
<tr>
<td>0.0 – 0.20</td>
<td>Slight</td>
</tr>
<tr>
<td>0.21 - 0.40</td>
<td>Fair</td>
</tr>
<tr>
<td>0.41 - 0.60</td>
<td>Moderate</td>
</tr>
<tr>
<td>0.61 - 0.80</td>
<td>Substantial</td>
</tr>
<tr>
<td>&gt;0.75</td>
<td>Almost Perfect</td>
</tr>
</tbody>
</table>

**Fleiss**

<table>
<thead>
<tr>
<th>Kappa</th>
<th>Strength of Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0-0.40</td>
<td>Poor</td>
</tr>
<tr>
<td>0.41 – 0.75</td>
<td>Intermediate to Good</td>
</tr>
<tr>
<td>&gt;0.75</td>
<td>Excellent</td>
</tr>
</tbody>
</table>

**Altman**

<table>
<thead>
<tr>
<th>Kappa</th>
<th>Strength of Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;0.20</td>
<td>Poor</td>
</tr>
<tr>
<td>0.21 - 0.40</td>
<td>Fair</td>
</tr>
<tr>
<td>0.41 - 0.60</td>
<td>Moderate</td>
</tr>
<tr>
<td>0.61 - 0.80</td>
<td>Good</td>
</tr>
<tr>
<td>0.81 - 100</td>
<td>Very Good</td>
</tr>
</tbody>
</table>

**Please note:** These interpretations were proven to be useful mostly in medical domain (diagnosis)
Summary

- Wisdom of the Crowd: Collective Intelligence and Groupthinking
- Obtaining Collaborative Knowledge: Motivation in Paid Crowdsourcing and Citizen Science
- Result Aggregation and Quality Assurance
Outline

• Collaborative Advantages
  • The Wisdom of Crowds
  • Conditions for a successful collaboration

• Small experiment
  • Can we collaborate?

• Obtaining collaborative knowledge
  • Crowd motivation
  • Scalability/Efficiency
  • Own work

• Input/Output Evaluation
  • Users and Data
  • Quality assurance

• Discussion
The real weight of the Babyphant: 112 KG
Average of the 38 estimates: 113.32 KG
Max/Min guesses: 300/2

The graph shows the single estimations as blue points, the average after each estimate as the grey dotted line and the real value as the orange line.

[Image of graph showing estimations and averages]

Baby Elephant: [Link to Zimbio picture]

Results of the Experiment:
References

Lei Chen (HKUST), Dongwon, and Meihui Zhang: Crowdsourcing in Information and Knowledge Management, 2014 CIKM Tutorial
Matthew Lease, Omar Alfonso: Crowdsourcing for Search Evaluation and Social-Algorithmic Search, 2012 SIGIR Tutorial
Michael J. Franklin et al: CrowdDB: answering queries with crowdsourcing, SIGMOD 2011
A. Marcus et al: Human-powered Sorts and Joins, VLDB 2011
Crowdsourcing for Information Retrieval: Principles, Methods, and Applications 50
Crowdsourcing for Information Retrieval: Principles, Methods, and Applications 51
James Surowiecki : The Wisdom of Crowds, 2004

Sergej Zerr , Stefan Siersdorfer , Jonathon Hare , Elena Demidova Privacy-Aware Image Classification and Search , SIGIR’12
M. Rokicki, S. Chelaru, S. Zerr, and S. Siersdorfer. Competitive game designs for improving the cost effectiveness of crowdsourcing. CIKM’14
M. Rokicki, S. Zerr, and S. Siersdorfer. Groupsourcing: Team competition designs for crowdsourcing. WWW’15