Collaborative Evaluation and Search

Sergej Zerr
zerr@L3S.de
Outline

• Collaborative Advantages
  • Wisdom of crowds
  • Conditions for a successful collaboration

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

• Input Evaluation
  • Inter Rater Agreement

• Output evaluation
  • Precision-recall

• Small experiment
  • Can we collaborate?

• Discussion
Collaboration

Often we need more than one hand

Sometimes more than one brain

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

.. Wisdom of the crowd ..
Direct Collaboration in WWW can be used:

Collaborative tagging, Favorite assignments, Click logs, Data gathering, Recomendations, ect., ect., ect....

Tags: Rainbow, Sea, Island, Green, Palm tree, Maui
Social Computer

Utility

Expert

Masses

Equivalent, or greater, utility under the curve

# of contributors

10

10,000+
Social Computer

Britannica

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

Wikipedia

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

Utility

Expert $$$$$

Masses $

# of contributors

4000

100,000+

Sergej Zerr
Collaboration: Paradox

Collaborative work needs to be managed Efficiently

Using „Wisdom of crowds” is not always a straight-forward task.

Kasparov won against the world in 1999

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

<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>Some mechanism exists for turning private judgments into a collective.</td>
</tr>
</tbody>
</table>
Groupthink Symptoms:
Irving Lester Janis (26 May 1918 - 15 November 1990)

- Illusion of invulnerability
- Collective rationalization
- Belief in inherent morality
- Stereotyped views of out-groups

- Direct pressure on dissenters
- Self-censorship
- Illusion of unanimity
- Self-appointed ‘mind guards’
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**
  - Wisdom of crowds
  - Conditions for a successful collaboration

- **Obtaining collaborative knowledge**
  - Gathering Data from Social Web / Games with purpose / Mechanical Turk
  - From Data to knowledge (Applications)

- **Input Evaluation**
  - Inter Rater Agreement

- **Output evaluation**
  - Precision-recall

- **Small experiment**
  - Can we collaborate?

- **Discussion**
Many Complementary Games

• ESP Game: label images
  → Image retrieval by text
• Squigl: match the labels to areas
  → Object recognition
• Matchin: find the better image
  → Image ranking
• Fliplt: 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
Re-Using “natural” Human Power

reCAPTCHA IS A FREE ANTI-BOT SERVICE THAT HELPS DIGITIZE BOOKS.

→ LEARN HOW reCAPTCHA WORKS

eBird
Report and track the birds online – any place, any time!
Social Computer

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

- Captcha
- Classification (OCR)
- Image tagging
- Speech recognition
- Face recognition
Ask the Social Computer!
“Hmm…, but what about ambiguity?”

„Mr. Burns saw Homer with the binoculars“
Ask the Social Computer!
“Hmm…, but what about a good Interface?”

Throw the coin and tell us the result

- Head
- Tail

Results

- Head 61
- Tail 39

Better: Some preliminary textual answer

- Coin type?
- Head or tail.

People often tend just to select the first option 😞

Matthew Lease and Omar Alonso: http://de.slideshare.net/mattlease/crowdsourcing-for-search-evaluation-and-socialalgorithmic-search
L3S Research – “What about image privacy?”

Private

Public

Work  Sea  Winter  Water

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

Points: 223

"Private are photos which have to do with the private sphere (like self portraits, family, friends, your home) or contain objects that you would not share with the entire world (like a private email). The rest is public. In case no decision can be made, the picture should be marked as undecidable."

Sergej Zerr, Stefan Siersdorfer, Jonathon Hare, Elena Demidova Privacy-Aware Image Classification and Search, SIGIR’12
Sampling Data from Social Web: Tags

Tags Distribution of 2Mio Flickr "Streetart" Images

Exception number of tags may indicate a spammer.
Sampling Data from Social Web: User activity

Simple random sample can result in a set dominated by few power users.
Can collaboration improve the sensemaking process at any step?
Do you use collaborative search?
What are the typical collaborative search tasks?

- Watched over someone’s shoulder as he/she searched the Web, and suggested alternate query terms.
- E-mailed someone links to share the results of a Web search.
- E-mailed someone a textual summary to share the results of a Web search.
- Called someone on the phone to tell them about the results of a Web search.

Around 90% of Microsoft employees are engaged in collaborative search activities.

Collaborative Search:

- Query formulation
- Search process/browsing
- Save/Bookmark
- Annotate/Organize

How to support users in collaborative searching?
- Ideas
- Tools(Web 2.0)
Indirect Collaboration on Web:

Google uses search/click logs. For Page Rank algorithm each link to a page serves as a vote for that page.

Amazon uses search/click logs. For item recommendation similar users are the indirect voters for the product.

ect., ect.
Inter–Rater Reliability: What about ambiguity/subjectivity?

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

Nederland, netherlands, holland, dutch Rotterdam, wielrennen, cycling, duck le grand depart, tour de france, Reklame, caravan, Funny Fotos
<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
<th>Sure?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BRAZIL</strong> →</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RUSSIA</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>BRAZIL</strong> →</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>GERMANY</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>HONDURAS</strong> →</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RUSSIA</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>GERMANY</strong> →</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>HONDURAS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>BRAZIL</strong> →</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>HONDURAS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>BRAZIL</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Outline

• Collaborative Advantages
  • Wisdom of crowds
  • Conditions for a successful collaboration

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

• Input Evaluation
  • Inter Rater Agreement

• Output evaluation
  • Precision-recall

• Small experiment
  • Can we collaborate?

• Discussion
Inter–Rater Reliability “Where is the cat?”

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>V</td>
<td>X</td>
<td>V</td>
<td>V</td>
<td>X</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>V</td>
<td>X</td>
<td>X</td>
<td>V</td>
<td>X</td>
<td>V</td>
</tr>
<tr>
<td>V</td>
<td>X</td>
<td>X</td>
<td>V</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Results
Inter–Rater Reliability “Where is the cat?”

• Ideas
  • Introduce „Honeypot“ – a set of ground truth objects and select only good raters
  • Select only raters who rate close to average rating.
  • Other strategies?

Results
Inter–Rater Reliability:

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

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

Naive approach: 3 cases out of 6 = 0.5 agreement
Statistic Significance:

First experiment:
Throw the dices with the right hand 100 times.
Compute the average

Second experiment:
Throw the dices with the left hand 100 times.
Compute the average

Claim that e.g. left hand is better.....

The set of numbers is what needs to be compared, not the average value
Inter–Rater Reliability: Cohen’s Kappa

Idea: We need to remove agreement achieved just by chance

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

\[ p_a = \frac{n_{11} + n_{22}}{n_{11} + n_{12} + n_{21} + n_{22}} = \frac{35 + 40}{100} = 0.75 \]

\[ p_e = \frac{55 * 40}{100} + \frac{45 * 60}{100} = 0.49 \]

\[ \hat{\kappa} = \frac{0.75 - 0.49}{1 - 0.49} = .51 \]

| A | B | |\n|---|---|---|
| | Yes | No | total |
| Yes | 35 | 20 | 55 |
| No | 5 | 40 | 45 |
| total | 40 | 60 | 100 |

Kilem L. Gwet, Handbook of inter-rater reliability 2010
Experiment: Inter Rater Agreement for Soccer

Result evaluation
Inter–Rater Reliability: Missing Values

Idea: Use partial ratings to estimate marginal probability only

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

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

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

\[ \hat{\kappa} = \frac{0.74 - .431}{1 - .431} = .54 \]

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

- **Reduce Kappa Paradox:**
  - Split judgments into more certain / marginal.
  - Measures: AC₁, AC₂ (ordinal and interval data)

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

Kilem L. Gwet, Handbook of inter-rater reliability 2010
Inter–Rater Reliability: Kappa Interpretations

<table>
<thead>
<tr>
<th>Koch</th>
<th>Strength of Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kappa</td>
<td></td>
</tr>
<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>0.81 - 100</td>
<td>Almost Perfect</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fleiss</th>
<th>Strength of Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kappa</td>
<td></td>
</tr>
<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>

<table>
<thead>
<tr>
<th>Altman</th>
<th>Strength of Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kappa</td>
<td></td>
</tr>
<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)

Kilem L. Gwet, Handbook of inter-rater reliability 2010
Experiment: Inter Rater Agreement - Fleiss kappa
Useful human power for annotating the Web

- 5000 people playing simultaneously can label all images on Google in 30 days!
- Individual games in Yahoo! and MSN average over 5,000 players at a time

Possible contributions: attach labels to images in other languages, categorize web pages into topics
Outline

- **Collaborative Advantages**
  - Wisdom of crowds
  - Conditions for a successful collaboration

- **Obtaining collaborative knowledge**
  - Gathering Data from Social Web / Mechanical Turk
  - From Data to knowledge (Applications)

- **Input evaluation**
  - Inter Rater Agreement

- **Output evaluation**
  - Precision-recall

- **Small experiment**
  - Can we collaborate?

- **Discussion**
Classification Task – Try to Imitate Human Decisions

London

Person_Name

City_Name
Classification Task – Binary Classification Problem

Query: London

Yes
Search in persons

Mr. Oracle

Person (?)

No
Search in cities
Closer Look on the Classifier Values

Search in persons?

| Classifier (confidence) value |  
|-------------------------------|---|
| St. Johns Bridge              | -0.8 |
| John Lennon Statue            | -0.3 |
| John Barrowman                | 1   |
| John the mule                 | 0.5 |
| John Cena                     | 0.3 |
| John Bishop                   | 0.6 |
| John Mayer                    | 1   |
| John Muir woods               | 0.1 |
| St. John Church               | -1  |
| John Mayer rock god           | -0.9 |

... etc. ...
Classification Evaluation: Comparison

1000 Queries

Mr. Oracle1
- TP 300
- TN 200
- FP 200
- FN 300

Mr. Oracle2
- TP 300
- TN 300
- FP 200
- FN 200
Classification Evaluation: Accuracy

\[
Accuracy = \frac{TP + TN}{TP + TN + FP + FN}
\]

Mr. Oracle1
- TP 100
- TN 200
- FP 50
- FN 750

Mr. Oracle2
- TP 500
- TN 50
- FP 400
- FN 50
Classification Evaluation: Recall / Precision

\[ \text{Precision} = \frac{TP}{TP + FP} \]

\[ \text{Recall} = \frac{TP}{TP + FN} \]
Classification Evaluation: F-Measure

\[ F_1 = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \]

Mr. Oracle1
- TP 100
- TN 200
- FP 50
- FN 750

Mr. Oracle2
- TP 500
- TN 50
- FP 400
- FN 50
Closer Look on the Classifier Values: ROC-Curves

![ROC Curves Diagram](image)

- Red: NetChop C-term 3.0
- Black: TAP + ProteaSMM-i
- Green: ProteaSMM-i
## Closer Look on the Classifier Values

Search in persons?

<table>
<thead>
<tr>
<th>Name</th>
<th>Classifier Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>St. Johns Bridge</td>
<td>-0.8</td>
</tr>
<tr>
<td>John Lennon Statue</td>
<td>-0.3</td>
</tr>
<tr>
<td>John Barrowman</td>
<td>1</td>
</tr>
<tr>
<td>John the mule</td>
<td>0.5</td>
</tr>
<tr>
<td>John Cena</td>
<td>0.3</td>
</tr>
<tr>
<td>John Bishop</td>
<td>0.6</td>
</tr>
<tr>
<td>John Mayer</td>
<td>1</td>
</tr>
<tr>
<td>John Muir woods</td>
<td>0.1</td>
</tr>
<tr>
<td>St. John Church</td>
<td>-1</td>
</tr>
<tr>
<td>John Mayer rock god</td>
<td>-0.9</td>
</tr>
</tbody>
</table>

![Mr. Oracle](image)

![Precision vs. Recall Plot](image)
Precision-Recall Curve

Predict Photo Attractiveness using text and visual features

Figure 3: Precision-recall curves for visual and textual dimensions and their combination (8000 training photos per class, numFav≥5)
Outline

• Collaborative Advantages
  • Wisdom of crowds
  • Conditions for a successful collaboration

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

• Input Evaluation
  • Inter Rater Agreement

• Output evaluation
  • Precision-recall

• Small experiment
  • Can we collaborate?

• Discussion
Weight of an elephant baby.
References and further reading

References:
2) Kilem L. Gwet, Handbook of inter-rater reliability 2010

Further reading: