SWISS CONTRIBUTION

Polish-Swiss Research Programme

On the Subjectivity and Bias of Web Content Credibility Evaluations

Michał Kąkol, Michał Jankowski-Lorek, Katarzyna Abramczuk,
Adam Wierzbicki, Michele Catasta

PROJEKT WSPÓŁFINANSOWANY PRZEZ SZWAJCARIĘ W RAMACH SZWAJCARSKIEGO PROGRAMU WSPÓŁPRACY Z NOWYMI KRAJAMI CZŁONKOWSKIMI UNII EUROPEJSKIEJ

RECONCILE

POLSKO-JAPOŃSKA WYŻSZA SZKOŁA TECHNIK KOMPUTEROWYCH

Agenda

- Introduction
  - What is credibility and why is it important
  - The background - Reconcile
- Related work
- Study overview
- Conclusions
- Future work

Introduction

What is credibility?

- Source credibility = trustworthiness in context of credibility (always subjective, can be dynamic, adapt to changes in information credibility)
- Individual credibility of information: a mental state of evaluating users (consumers of information) = Belief that received information is true
  - Credibility = Believability
- Possibly a third concept: credibility of information as a social opinion

Introduction

Why is it important?

- Internet as information medium
  - Self-explaining
  - Health, finance, relationshps etc.
- Search engine results as tangled relevance and reliability

Introduction

The background – Reconcile project

- Polish-Swiss project
- Series of studies concerning web content credibility
- Aiming at
  - Providing web tool for credibility evaluation
  - Computational trust management: supports evaluation of source credibility
  - Aggregating independent, distributed, diverse opinions
  - Reducing impact of presentation: increasing prominence of chosen features
  - Using data-mining approach and web content classifiers

RECONCILE

Related work

- **Research of Fogg**
  - Prominence-Interpretation theory
- **Augmenting search results**
  - Microsoft experiment
- **(Semi-)Automatic credibility classifiers**
- **Biased ratings**
  - (Amazon, IMBD, BookCrossings)

RECONCILE

THE STUDY: GOALS AND PARTICIPANTS

- Main goals:
  - Verifying hypotheses related to subjectivity of credibility evaluations. Subjectivity can be due to:
    - Socio-economic status
    - Internet efficacy (second level digital divide)
    - Psychological factors
  - User ratings are subjecct to strong bias
- Participants
  - 1503 (2532 invited) IIBR panel participants
  - Control for demographics and social characteristics

RECONCILE

THE STUDY: EXPERIMENTAL TOOLS

- Tools for creating archived copies of Web pages
  - Capable of „freezing” and „unfreezing” of any Web page
- Web-based questionnaire
  - Can be also run in a browser plugin
  - Displays the „unfrozen” Web pages to the users
  - Users are requested to rate Web pages according to selected criteria
  - Users can add textual comments, suggest sources relevant for evaluated page, submit other Web pages for evaluation
- This tool will be the basis of th Reconcile credibility evaluation technology

RECONCILE

THE STUDY: SUMMARY

- Participants and evaluations
  (full, finished evaluations)
  - 1503 participants
  - 4354 evaluations
- Evaluated pages
  - 155
  - 17 categories
  - Average 28 evaluations per page
    - Experiment duration
  - 2012-09-04 to 2012-09-25 (22 days)

RECONCILE

USERS' CATEGORIZATION

We have analysed a number of secondary sources and identified 9 internet activities that are relatively rarely performed. Respondents were asked whether they perform given activity at least once a month.

- creating and publishing own texts (e.g. blog, Wikipedia entry), graphics, music, photos, videos etc.
- creating or modifying WWW site (e.g. code changes, presentation changes)
- gathering materials/information required for learning or work
- gathering information for dealing with administrative matters
- buying products or services via internet
- selling products or services via internet
- commenting on blogs, writing on internet forums/discussion groups
- writing about/reviewing products or services
- using mobile banking

RECONCILE

USERS' CATEGORIZATION

The Web-Use Skill Index is based on a list of 10 internet-related terms. Respondents are asked to rate their level of understanding of these terms on a 1- to 5-point Likert scale. User's score on this scale is given by the sum of points (all the evaluations) and can take on any value between 10 and 50.

- Advanced search
- Tagging
- PDF
- Spyware
- Wiki
- Weblog
- JPG
- Cache
- Malware
- Phishing

RECONCILE

CREDIBILITY EVALUATIONS

- Each user evaluated 3 random pages on the same topic
- Three experimental conditions:
  - TB – Topic Browse: In this condition participants' task was to evaluate 3 presented websites.
  - TS – Topic Search: In this condition participants' task was to evaluate 3 presented websites and answer a question related to their content.
  - TK – Topic Keywords: In this condition participants' task was to evaluate 3 presented websites and propose three keywords to describe their content.

RECONCILE

CREDIBILITY EVALUATIONS

- Pages were categorized into several thematic groups.
  - Healthy lifestyle (e.g. What is Duncan's diet and is it healthy?)
  - Cancer treatment (e.g. Are targeted therapies effective in breast cancer treatment?)
  - Parenting (e.g. How long should breast feeding last?)
  - Personal finance (What are the best ways for you to invest capital?)
  - Etc.

RECONCILE

CREDIBILITY EVALUATIONS

- Each users evaluated websites in seven dimensions:
  - Credibility
  - Presentation
  - Author's expertise
  - Intentions
  - Clarity
  - Completeness
  - Validity

- Clarity and Validity turned out as insignificant and weakly correlated with credibility

RECONCILE

Conclusions

Subjectivity

- Significant
  - Gender, Education, Internet experience
- Just slight effect on the ratings

Bias

- Significant
- Great impact on the ratings
  - Comparison to expert's ratings
  - Acquiescence bias?

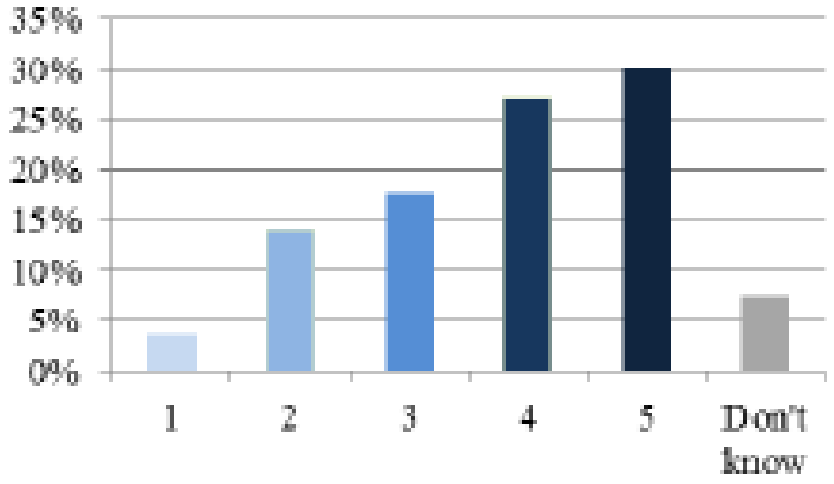

Figure 2. Distribution of all credibility ratings

Conclusions: Subjectivity of gender

- Credibility ratings vs Gender
  - Slightly more females
  - Males seem to give less extremely positive credibility scores

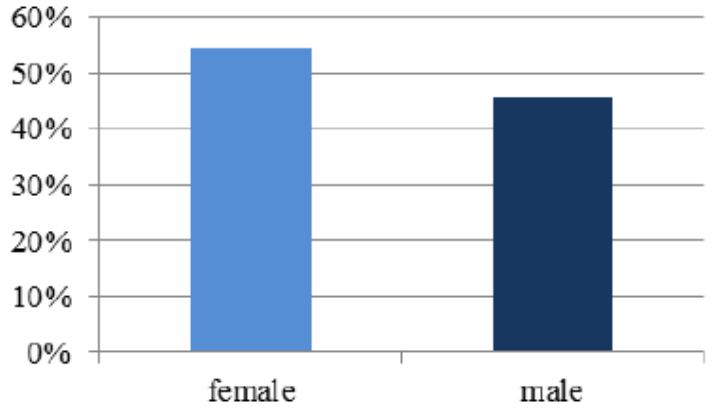

Figure 4. Distribution of genders among the respondents

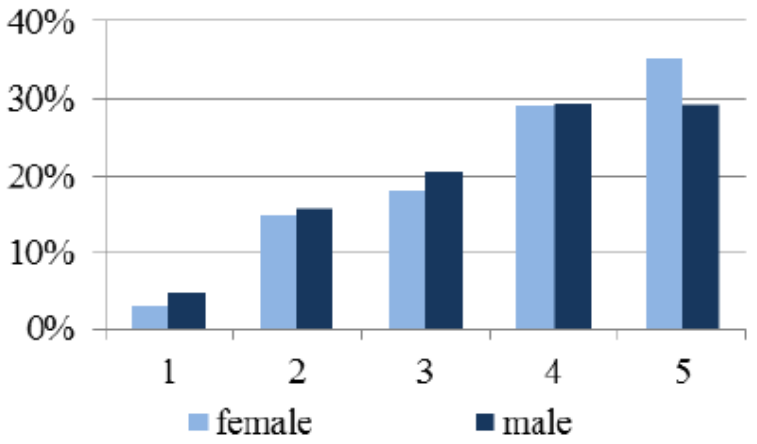

Figure 5. Credibility ratings by genders

Conclusions: Subjectivity of education

- Credibility vs Education
  - 50% respondents with higher education
  - The lower the education level the bigger skew

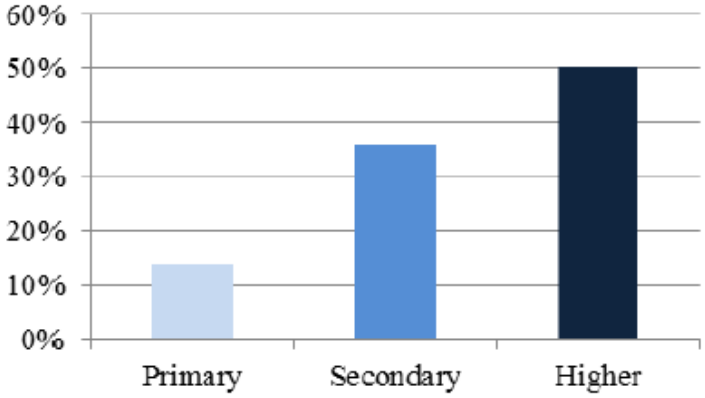

Figure 6. Distribution of education categories among respondents

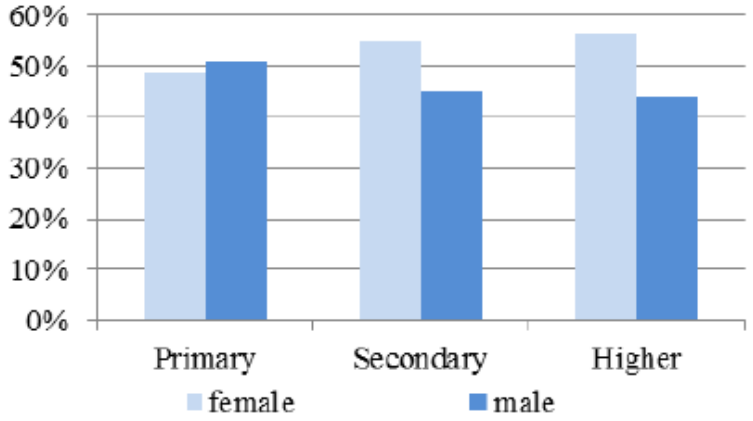

Figure 7. Education categories by genders

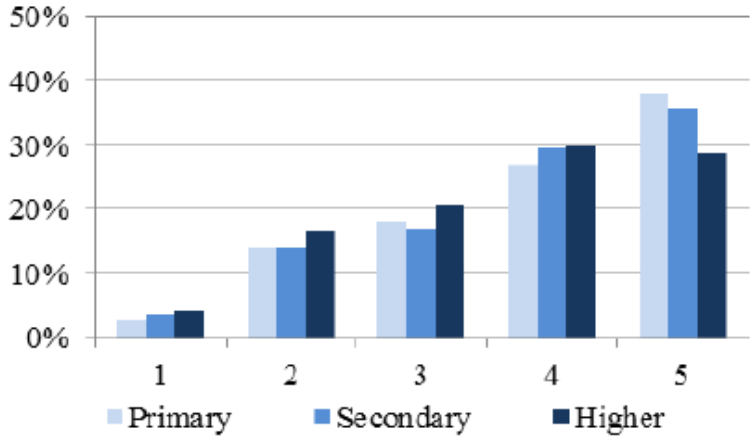

Figure 8. Distribution of credibility ratings by education categories

Conclusions: Subjectivity of Internet experience

- Credibility vs Experience
  - 1/3 of respondents are heavy users
  - Heavy users are mostly male
  - The lighter the user the biger the skew

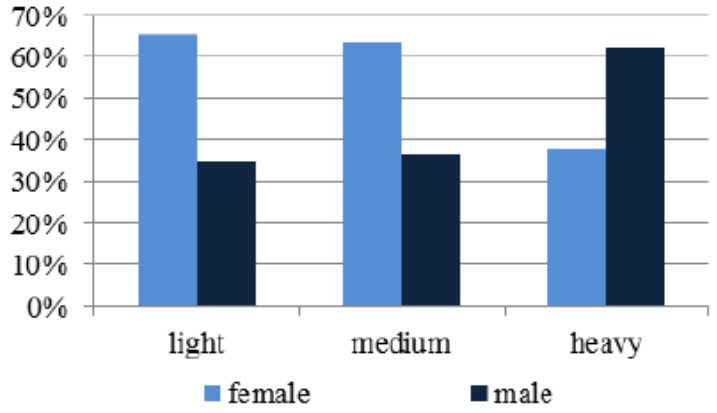

Figure 10. Internet efficacy levels by genders

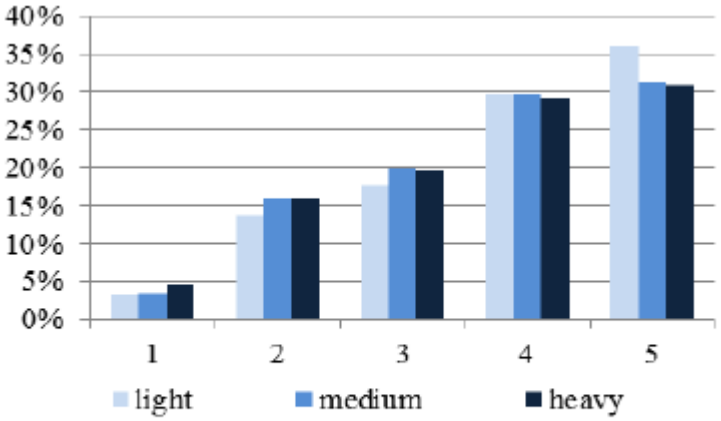

Figure 12. Distribution of credibility ratings by Internet efficacy levels

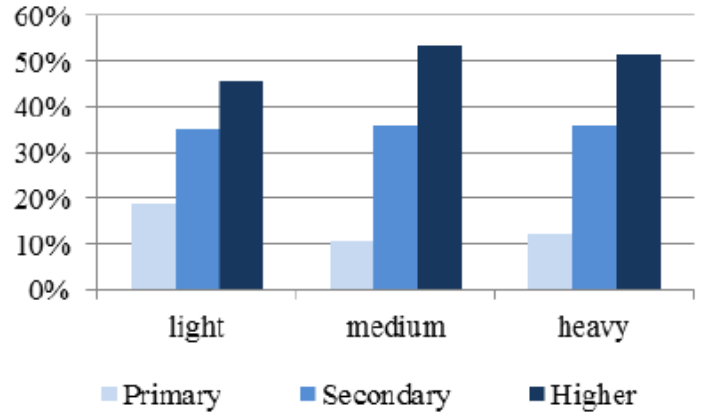

Figure 11. Internet efficacy levels by education categories

RECONCILE

Conclusions: Subjectivity of psychological traits

- Hypothesis: psychological factors have a strong impact on credibility evaluations
  - Two psychological traits: „trust towards others” and „need for cognition”
  - Traits tested by sets of simple questions based on psychological surveys
  - Each respondent replied to these questions before the evaluation of Web pages
  - Main results: not strong impact on credibility evaluations

RECONCILE

Conclusions: Subjectivity of psychological traits

- Credibility vs need for cognition
  - No correlation with credibility ratings
  - High need for cognition and tendency to overrate websites
  - Very low need for cognition and use of low end of credibility scale
- Credibility vs trust towards others
  - Weak but significant correlation with credibility tatings
  - Greater willingness to trust and higher credibility ratings
  - Vice versa

- **Needs further validation.**

RECONCILE

Conclusions: The Bias

- **Ratings distribution is typically negativelly skewed**
  - Controversial categories
- **Domain Experts for comparison**
  - 7 MD, 3 Midwives, Investment Broker
  - Much more regular distribution
- **Interrater agreement**
  - Slight agreement (Experts-Respondents)
  - Almost perfect agreement (Experts)

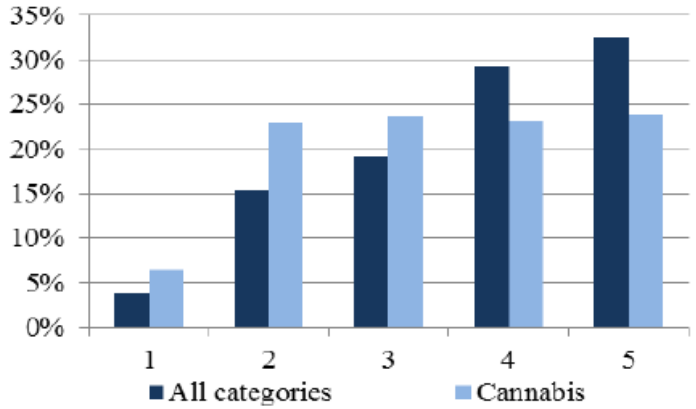

Figure 14. Respondents credibility ratings in all categories versus "Cannabis" category

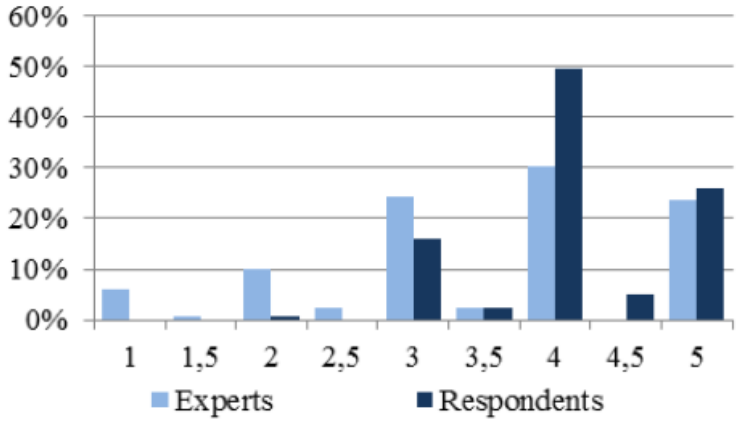

Figure 13. Median ratings of experts and respondents

How to cope with the biast? Future work

- **How to correct the bias?**
  - How to aggregate the ratings?
  - What about the changes in time?
  - How to deal with low number of ratings?
  - Is the 5-point scale enough?
  - Use textual justification left during evaluation

RECONCILE

Future work

- Other extensive studies
  - Bigger online web sites credibility study on Amazon Mechanical Turk (finished on 1st May)
    - 16000 evaluations, 6000 pages, 60 categories, 3500 respondents
    - Extensions are planned
  - Online statements credibility study (in preparation, will start in May)
- Prototype of web credibility evaluation system
  - framework for credibility evaluation in distributed environment

SWISS CONTRIBUTION

Dziękujemy za uwagę.
Thank you for your attention.

PROJEKT WSPÓŁFINANSOWANY PRZEZ SZWAJCARIĘ W RAMACH SZWAJCARSKIEGO PROGRAMU WSPÓŁPRACY Z NOWYMI KRAJAMI CZŁONKOWSKIMI UNII EUROPEJSKIEJ