Codebook for “Much Ado About Gender: Current Practices and Future Recommendations for Appropriate Gender-Aware Information Access”

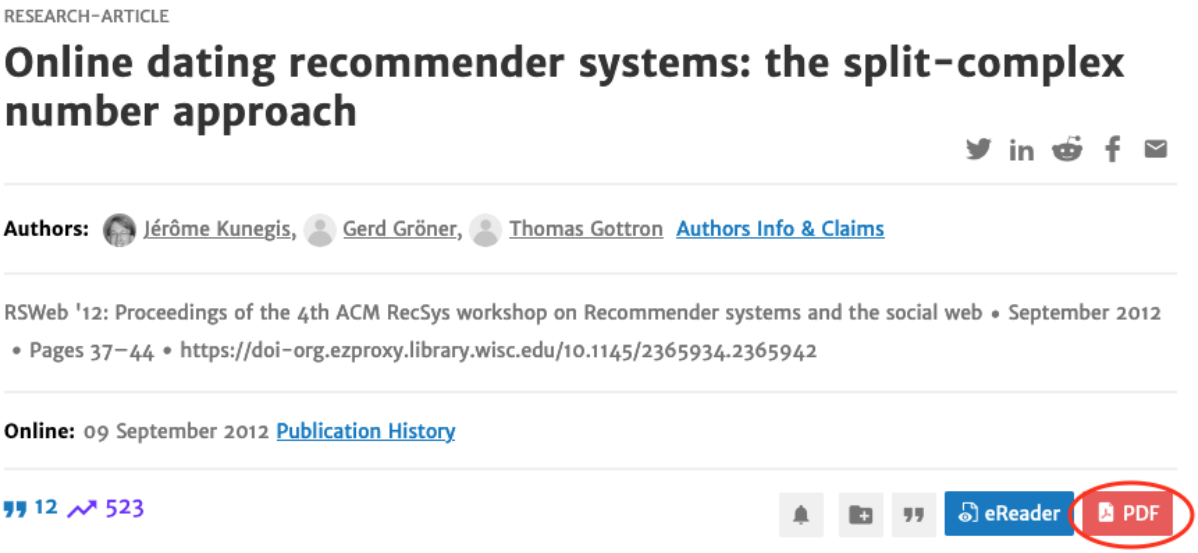
Authors: Christine Pinney, Alex Hanna, Michael K. Ekstrand

Summary

This project looks into how information retrieval and user profiling research treats gender as a variable. Making inferences based on gender is problematic because gender itself is socially constructed and is not always binary. Inferring gender based on user behavior also is problematic because it reduces something as complicated as gender to a few weak signals. In this project, we’re trying to get a field-level view of how IR and UMAP treat gender.

Steps in Accessing an Article

1. Visit the data file. Each coder will have a set of articles beside their name.
2. For the next article assigned to you, click the link in the **link** column.
3. Click the PDF link and download the article for reading.



Search Criteria

All searches with the ACM Digital Library Search.

- Conference / Journal: one of SIGIR, CHIIR, RecSys, UMAP, TOIS
- Date Range: 2017-2021 (some from early 2022 included)
- Keywords: sex OR gender OR male OR female OR men OR women

How to Code

For the article, you are looking for several different variables. These are subject to change as we continue to code articles and discuss them as a group.

You typically will not need to read the whole article to understand how the authors use gender. Sometimes it is sufficient to read the abstract, the introduction section, and the data and/or experiments section. It's also helpful to use the Find functionality in your PDF viewer (this is typically Ctrl+F on Windows and Linux computers, or Command+F on Macs). Look for keywords such as "gender", "sex", "male", "man", "men", "female", "woman", "women", and "non-binary".

For now, we are primarily interested in each of these dimensions of how the authors use gender. These variables will correspond to columns of the spreadsheet.

Paper Type

Regular, Perspective, Short Paper, Doctoral Consortium, Short Abstract (SIRIP - the industry track), RecSys industry track, Workshop Summaries, Late-breaking Results, Journal Article.

Anything other than regular, short papers, or late-breaking results we probably don't care about.

What is the primary referent?

The referent is the group of people who gender is being attributed to.

Example Responses: Users, Providers, Subject of the Text or Image

Are there multiple referents?

Sometimes articles will have multiple referents (e.g. this article, where a group of annotators are gendered, as well as data subjects in an image search). We are expecting this to be a small group of articles.

Responses: Yes | *No* | *Not sure*

Is there a gender variable?

Determine if there is any kind of gender variable in the text at all. This may be implicit or explicit. If there is no gender variable, or you are not sure, then this may make answering the other questions in the spreadsheet impossible.

Applied should be used when the model or experiment in the paper does not use gender, the authors suggest that gender could be used with their method.

N/A should be used in a paper where there is no study or model. This is distinct from a model which does not use a gender variable at all.

Responses: Yes | *No* | *Applied* | *N/A*

If gender variable is "no", "applied", or "N/A", paste the sentence in which sex/gender appears.

What are the gender categories?

Determine what values the gender variable will take. Sometimes these will be explicitly mentioned, but often they will be obscured in a table or implicit in a statistical model, like a regression (e.g. male = 1, female = 0).

Example Responses: Male/Female | *Male/Female/Other* | *Male/Female/Non-binary*

Do they acknowledge that gender is not binary?

Much of the time computer scientists will make this acknowledgement, but then go ahead and keep using binary values for gender.

Responses: *Yes* | *No* | *N/A*

How is gender determined?

Determine how the authors are obtaining the gender label. This may come from self-identification by the user, or is inferred by the authors, third-party annotators, or an automated system.

Example Responses: Self-identification | *Annotator* | *Inferred from [name* | *face* | *<other>]* | *Synthetic*

What is the goal of using gender?

This is a qualitative variable in which you determine what the end goal of using gender is. Is gender used to make an inference about user behavior? Is the system trying to infer gender from a set of signals?

Example Responses: Contextualize Results | *Personalize based on Gender* | *Detect Gender* | *Audit System Behavior* | *Protect Gender Variable* | *Gender Diversity or Inclusion* | *Indexing Clinical Trials* (many more)

Is this paper about bias and/or fairness?

Many papers will be about assessing the bias with a particular system or dataset, or attempt to debias or make fair a dataset or method. If the paper is about bias or fairness, mark this with "Yes."

Responses: Yes | *No* | *N/A*

Notes

Use this field to note anything confusing or complicated which doesn't get captured in the other fields.

Flag for review

Use this field to flag for review at meetings with the rest of the team.