Who told you that?
Transparency and control for privacy-aware personalization

Eelco Herder

27 September 2017
Who is Eelco Herder?
My research in a nutshell

My core research field is User Modeling and Personalization.

The application areas of my research have been extended to various related fields, including social media, technology-enhanced learning and mobile, ubiquitous computing.

In particular, I am interested in how users interact with programs, how they search for information, and how they return to places where they have been before.
Looking for a nice restaurant in Tilburg

You would probably recommend me the most popular or well-known restaurants. Or the ones that you like best.

But did you know that I am a vegetarian, that I prefer mid-price restaurants, love Indian and Italian cuisine, find atmosphere more important than quality of food, would like to take the dog with me, and do not have a car at hand?
So I might better ask Google

Hmm...

- Which aspect (or keyword) is more important?
- The average query length is 4.29 words
We expect interaction to be personalized

In a library, a person looks for some books on China. What will the librarian recommend?¹

- Is the person a *small child* who saw a TV show about China and wants to learn about this exotic country?
- Or a *high school student* working on a paper?
- Perhaps a *prospective tourist*?
- A scholar interested in *Eastern philosophy*?
- Someone who can *read Chinese*?

---

Most likely the librarian will make an educated guess, based on the person’s appearance:

- age, style of clothing, accent, choice of words, ...
This initial guess might be confirmed or refuted by observations.

- It is assumed that a European cannot read Chinese, unless said otherwise
- Children are generally not (yet) interested in Eastern philosophy, but there are exceptions
- ....

The educated guess, a stereotype can be refined with follow-up questions.

Persons expect a personalized advice, even though the librarian does not know them.
And the same seems to yield for Web stores.

Jeff Bezos, amazon.com

If I have 3 million customers on the Web, I should have 3 million stores on the Web
When is personalization useful?

My supermarket is not personalized. Still, I can find all products that I need. Probably just because my needs are similar to everyone else’s needs.

Personalization is deemed useful when:

- there are so many things to choose from that there is a need for guidance or recommendations
- the system is used by people with different goals and backgrounds
The ideal recommender

Your partner, your best friend or your mother probably knows a lot about you:

- the food you like, the books you read, the movies you watch
- things that interest you or that upset you
- your current needs, aspirations and goals
- dates of your birthday, your kids’ birthdays, and holidays
- secret desires and phantasies
Still, this does not guarantee that your mother will buy you a present that you like.

It can be something that
- you already have
- you hate for some reason only known to you
- she bought to surprrize you (sometimes this works out perfectly fine, though)
Personalization techniques

Most information-oriented websites are not personalized. They provide the same kind of content to anyone.

In many cases, this is fine. Users find the information they need - as long as the navigation structure is understandable.
Semi-personalized systems

Semi-personalized systems do not make use of a user profile. They try to adapt the content to the (estimated) needs of everyone. For example:

- trending topics on Twitter
- most read items on a newssite
A personal touch and personal functionality

Particularly transaction sites (online stores) require users to log in or to create an account. It is customary that these sites ‘welcome’ the user, which creates some basic ‘mutual recognition’.
Why bother to log in

A personal greeting is not a very convincing reason for asking users to register.

More convincing - and practical:

- Easier, quicker checkout (less information to fill out when buying an item)
- Access to previous orders, perhaps with the probability to re-order
- Saving of default location, default language, default values
I personally really appreciate this sort of functionality - most users expect (ecommerce) sites to offer such features.

Personalization goes beyond this ‘basic’ functionality.
It’s good that the government has some data about me. I wouldn’t want to fill out the same personal data over and over again. Would you?
Understanding the user

Tools such as Google Analytics show general trends:

- number of visits and users
- where do users come from, which systems do they use
- popular pages and keywords

User modeling is about getting to know the *individual* user.
The adaptation process: a process-oriented view

Explicitly provided user data on the Web

Personal data has always been published and shared in the World Wide Web.

The Web 2.0 has promoted this even more by offering a variety of services where users can publish data without requiring specific technical skills on:

▶ social networking sites
▶ photo and document sharing sites
▶ collaborative work environments
▶ blogs
▶ and many other sites.
My Facebook word cloud
Implicitly collected user data

In addition, many of our activities on our computers and on the Web are logged in some way.

- **Ecommerce sites** register which items we browsed for and which we bought.
- **Social networking sites** keep track of the messages, pictures and likes that we broadcast and send to our friends.
- **Browsers** maintain a history of sites that we visited and store cookies.
- **Desktop search engines** index all programs and files that we used.
Who or what can access the data?

Much personal data is only accessible by ourselves and is used for **private purposes**, such as reflection, archival and refinding.

Many other traces are **shared with others** or broadcast to the world - voluntarily or involuntarily\(^3\).

These public or semi-public traces define our online presence: the way we are seen by the outside world, based on which others judge who we are.

And even personal data that is not disclosed is often **used for inferring** our interests or for providing recommendations.

Information asymmetry

A fundamental privacy concern is the *information asymmetry* between

- website providers as data collectors and
- users as data providers

due to the absence of adequate control mechanisms of how user data are collected and whether they are disseminated.

---

Limited user awareness

In most environments it is hard to figure out what exactly is being logged or to inspect and regulate which traces are stored, used by other applications or published to the outside world.

This raises serious privacy issues, of which the average user is often not aware.
The consequences can be real\textsuperscript{5}

In 2008, 8\% of U.S. companies with 1000 workers or more reported firing an employee based on information released on social networks.

The end of privacy via obscurity

Through the powerful reach of search engines, obscure pieces of personal information - such as email messages sent a decade ago to niche forums or newsgroups - are increasingly retrievable by a simple keyword search.

As a result, any ‘privacy via obscurity’ that generally kept such information from public view has been diminished.

---

The issue with linked data and mash-ups

Mash-ups, clever combinations of existing applications, provide interesting new applications and user-friendly one-stop services.

Further, many social networking applications have methods for synchronizing content and contacts, as well as for integration in other sites.

As a result, users receive better, more integrated services, but they are also exposed to serious threats to informational privacy.

---

Completeness of user profiles

Users often do not fill out their profiles completely. For example, Twitter only asks 6 attributes, but these profiles are only completed up to 49%\(^8\).

Aggregating data from different sources leads to more complete user profiles.

---

The issue is often not the data itself, but the interpretation. Mr. Iwanyk suspected that his TiVo thought he was gay, since it kept recording gay programs. He tried to tame TiVo by recording war movies and other “guy stuff.”

“I overcompensated. It started giving me documentaries on Goebbels and Eichmann. It stopped thinking I was gay and decided I was a crazy guy.”

---

Some interpretations may seem quite far-fetched at first... ¹¹

Apart from the fact that people who like curly fries are more intelligent, it is also possible to find out whether a teen girl is pregnant ¹⁰

¹⁰ https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#77184eaf6668

Ultimately the question is what is being done with your user profile.

Being shown an advertisement for washing machines after having searched (and bought one) at Mediamarkt might be a bit creepy. But it doesn’t have real consequences. Apart from that, I can even check why this advertisement has been shown to me.
Personalized search results

Useful when the personalization is done right. Annoying when factors are taken into account that are wrong or not relevant - and one cannot turn it off.
Google’s personalization is like my mother buying the wrong present

We cannot predict or analyze 100% accurately your current needs and context. Even your mother cannot.

This is also the reason why a filter bubble feels uncomfortable: it is not self-chosen or self-made.

Plus there are other factors at work, like commercial interests. And that’s why transparency and control is needed.
Judy Kay: Scrutable Adaptation

Capable of being understood through study and observation; comprehensible and understandable upon close examination

- What information does the system collect about me to drive this personalisation?
- What does this system do with that information?
- With whom does it share the information?
- What information is in this part of my environment?
- How is it combined with other information about me?
- What is the meaning of "prefers low cost travel"?
- How did the system conclude that I prefer to travel at low cost?
- How can I get a big picture of the models related to my travel?
- How did the system choose to put this information here and now?
- What did it present a week ago?
- What would it present if I were a rich man?
- How do I change any of the above processes?

---

Users were able to scrutinise simple forms of personalisation, but mixed results for more complex forms.

The profile tool was effective and users found it easy to use.

"Surprised and upset that the system made them a member of the Special Interest Group for religious TV programs."

---

The usability of privacy controls is improving quickly.

I wanted to show the arcane privacy controls for Google Ads. Apparently, Google (like Facebook) recently put a lot of effort in the usability.

Still, how much time would users want to spend on settings for advertisements?
Amazon offers methods for controlling my recommendations

- On an item-by-item basis (not scalable)
- It is not explained what else is in my user model
Why not explanations for any type of personalization? This is an active research field in recommender systems\(^{15}\).

<table>
<thead>
<tr>
<th>Aim</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transparency (Tra.)</td>
<td>Explain how the system works</td>
</tr>
<tr>
<td>Scrutability (Scr.)</td>
<td>Allow users to tell the system it is wrong</td>
</tr>
<tr>
<td>Trust</td>
<td>Increase users’ confidence in the system</td>
</tr>
<tr>
<td>Effectiveness (Ef.)</td>
<td>Help users make good decisions</td>
</tr>
<tr>
<td>Persuasiveness (Pers.)</td>
<td>Convince users to try or buy</td>
</tr>
<tr>
<td>Efficiency (Efc.)</td>
<td>Help users make decisions faster</td>
</tr>
<tr>
<td>Satisfaction (Sat.)</td>
<td>Increase the ease of use or enjoyment</td>
</tr>
</tbody>
</table>

Data security and legislation are only part of the solution. Sometimes scrutability is deemed unnecessary or cumbersome, sometimes simply as not wanted.

Why should a company explain why you have to pay a higher price?

Regulations and fines might be a solution. But ultimately you want the user to be in control, or at least to be able to be in control.
Requiring companies to do so without telling them which rules exactly to adhere to doesn’t work - they will find a workaround.

Our job is to provide metaphors, workflows and templates that companies have to adhere to, to ensure that:

- it is transparent how search results are adjusted
- why your Facebook feed contains which items (and which not)
- why you are given a loan (or not)
- ...
Counter argument: our models are too hard to explain

Clustering methods, support vector machines, regression methods, ... they all can provide insight in which factors are most important (i.e. have the highest loadings). Often, a small number of small features provides sufficient insight\textsuperscript{16}.

\textsuperscript{16}Figure taken from: Markus Rokicki, Eelco Herder, Tomasz Kusmierczyk and Christoph Trattner. Plate and Prejudice: Gender Differences in Online Recipes. Proc. UMAP 2016.
Counter argument: how can you tell whether a price or feed is personalised or not?

Some companies might not be willing to provide the full details on how they personalize. This is why more and more studies focus on reverse-engineering the inner workings of personalized search engines and the like.\(^\text{17}\)

---

High-level research agenda

- Understand the impact of personal data on information and services served to the user
- Understand user choices and user understanding regarding online privacy
- Develop generic metaphors and approaches for scrutability (transparency and control)
- Evaluate user understanding, acceptance and adoption of these tools
Questions?

Privacy Engineering, User Modeling, Personalization, Recommendation, Web Usage Mining, Data Analysis and Visualization, Usability, Evaluation