Web Users Interests and Web Content Placement: The Gugubarra Project

Ioanna Constantiou*, Natascha Hoebel, Naveed Mushtaq, Clemens Schefels, Karsten Tolle, Roberto V. Zicari

Goethe University of Frankfurt,
Computer Science and Mathematics Department

* Copenhagen Business School
Department of Informatics
"Self Revealing".

I call the user's "self revealing" process, a process by which a person allows traits of his self or presumed self, to appear, be visible and be traceable through his background and present experiences, both in real situations and in a digital context.
Background experiences & Digital Traces

I call background experiences any learning-oriented situations a person has experienced, which have happened before a given point in time. Background experiences are very relevant when we consider the user behavior and the user decision process. The fact that, now days, a part of the person’s background experiences occur in a digital world, allows their digital tracebility. I call this digital traces.
Foreground
experience

Combined with what I call the foreground experience, that is the current context in which a decision process occurs, such experiences do influence the local choice. When part of the background experiences occur in a digital world, digital traces are left and can be stored permanently in a digitalized format. They can then be seen, analyzed, processed and interpreted.
Real Self or Presumed Self: User Profiles

The digital traces the user allows to leave, characterize a part of his real self or presumed self. Digital traces are characterized by a number of factors related to the experiences the users had in the digital world. For example the actions the user has performed on a given web site, the time spent on specific web pages, the links to other pages or web sites the user has followed. The internet is full of such digital traces. When a person is making a decision, we can look at the digital traces he has left in his digital background experiences, and at the information we might have of his real background experiences and relate them together into a user profile.
User interaction

The user profile can be enhanced by the user interaction with his context, be the context real and or digital. Any interaction the user is willing to have with his context reveals traits of his self or presumed self.
Self Revealing

Sharing photos with selected friends in a web social networks, is one example of such interactions. It is therefore possible to use the digital traces left by a user as a result of background experiences and present interactions, for a number of purposes. The interesting thing is that digital background experiences merge with real background experiences to form a coherent set of information related to a user, for a given period of time. Present user interactions, alters these background information to form a new set of background experiences. The process repeats itself over time, and reveals traits of the user’s self. Therefore the name Self Revealing process.
Self Revealing and the Future of Retail.

This Self Revealing process, when properly handled, can be used as the base for a different shopping experience for the customers.

Opportunities and challenges for both the owner of a retail business (chains, department stores and Internet retail firms), and the customers. A recent report by PricewaterhouseCoopers and TNS Retail Forward examines major changes that could hit the retail industry between now and 2015. The aging of the Baby Boom generation and coming of age of Generation Y will cause a change in consumer demand and spending habits, the report says. Retail stores that are able to adapt to the new consumer demographics will thrive while those unable to adjust will fail.
The report says: "Retailing will become an industry that realizes, more and more, that it must tailor its offerings to select customers, as opposed to the mass appeal approach of the 1980s, in order to win over customers and foster greater customer loyalty".

Better understanding of the consumers therefore, become imperative.
Gugubarra: Main Research Areas

1. **Infer the (supposed) interests of Web users** --> By
   - ---> By looking at the Digital “traces” left & Asking Users Feedback.

2. **Cluster Web Users with similar (supposed) interests**

3. **PlaceWeb**: Definition of a *methodology* for placement of information on a Website
   ---> Applying Framing Strategy to Web content.
Gugubarra: Main Research Areas

1. Gugubarra Engine

2. Framing
   Strategy
   Assumptions

3. PlaceWeb: Definition of a Methodology for Placement of Information on a Website

feedback

validates

2. Framing and user choice Experiments

1. User Profiles
1. The Gugubarra Engine

- A Community of users **registered on a Website**.

  *For each user a profile is built.*

- The profiles go beyond collecting the obvious information the user is willing to give at the time of registration. In Gugubarra, a user profile contains two parts: the obvious profile, given directly by the user and a **non obvious profile (NOP)**, inferred by the user’s behavior (“digital traces”) during his visits on the site & user’s feedback.
Gugubarra: Applicability

- Enable the **owner (manager) of a Website** to **better understand** the user community that is visiting his site.

- **Qualify** better the members of the user community.
Approach

- Build user profiles representing the “traceable supposed users’ interests”:
  - Non-Obvious user Profiles (NOPs)
  - vs. Obvious Profiles (OPs)

- Cluster groups of users by “similar interests”
  
  Detect shifts in interests of single users or user groups.
  
  Detect upcoming trends of the community.
A Website and a supermarket: Some similarities…
There are special “locations” (Zones)...

- for meat...
- for wine...
- for frozen food
You can go from one section to another one and search for goods
Customers perform actions....
Customers perform different actions…
Some customers pay…. But not all of them
Who are the Customers? What are their interests?
Back to users of a Website…

We want to offer to the owner of a Website a tool to create users profile, that can be *parameterized* and *customized* according to the owner's business needs.
Design principle of the Gugubarra Engine

- NOPs are calculated by using different parameters that can be chosen by the owner of a Web portal.

- The combination of the chosen parameters implements a specific *strategy* to deploy and manage NOPs.

- The common element for all strategies is as follows: The creation of NOPs is done by looking at the “traceable behavior” of the Web user and by taking into account a feedback mechanism.
Project history

- The Gugubarra project began in 2004.
- Three research prototype systems implemented between 2004 and 2010 (current version Gugubarra 3.0).
- In use at:

http://www.dbis.informatik.uni-frankfurt.de/
Interpretation of user interest

- The user profiles reflect the “inferred” interests of the users related to a set of pre-defined topics defined by the owner of the Website.

- Topics are related to the content of the Website. (more on this later …)
A Website

We consider a Website (or a domain) as a collection of Web pages that are linked together within the site. Each page has a specific content.
Content of a Web Page
We can define Topics (with different importance…)

1. BROADBAND
2. HOTELS
3. ROUTES
Actions can be performed on a Web Page

We can define some actions as “more important” than others...

1. Download

2. Click
Concepts for the NOP calculation

- the following main concepts are used to create a user’s NOP:
  - **Zones**,  
  - **Topics**,  
  - **Actions**,  
  - **Durations**,  
  - **Weights**.
Zones

A zone defines an area on the Website. It can be a set of pages, a set of parts within a page, a set of parts of several pages, or any combination thereof.

A zone has one of three states.

- The state ON indicates that this zone is being used to calculate the NOPs of the visitors, state OFF indicates the zone is not used. In the third state OFF-ACTION-SENSITIVE, only actions the user does within the zone are taken into account, not the duration.
Topics

- *Topics* are pre-defined by the owner of the Website and are related to the content.

They are defined global by the owner of the Website and then associated to *zones*.

- A *weight* indicates the relative importance of the topic in respect to a scale from 0 (not relevant) to 1 (extremely relevant) in the zone.

**Different semantics**
Predefined Semantics of Topic Weights

1. **DEFINE**
   *Manager’s goal oriented.* The Web manager controls the information and *defines its relevance* according to his needs/goals.

2. **(SELF-DISCOVER)**
   *Manager’s interpretation of the content.* The manager discovers the content of a Website manually on a high granularity level.

3. **(AUTO-DISCOVER)** *Topic weighting by term frequency.*
   Weights are defined as frequency of topics and automatically extracted.
An Example of Topics and Zones

Global topics:
Tp1 = ‘Broadband’,
Tp2 = ‘Finding a Location’,
Tp3 = ‘Calculating Routes’,
Tp4 = ‘Insurance’,
Tp5 = ‘Hotels’.

Page 1:
Z1 = {‘Broadband’, 0.5},
Z2 = {‘Finding a Location’, 0.6},
{‘Calculating Routes’, 0.8}),
Z3 = {‘Finding a Location’, 0.3},
{‘Calculating Routes’, 0.3}).

Page 2:
Z4 = {‘Insurance’, 0.5},
{‘Calculating Routes’, 0.9}),
Z5 = {‘Calculating Routes’, 0.9},
{‘Finding a Location’, 0.9}, {‘Hotels’, 0.5}).
Actions

*Actions* are also global, defined by the owner of the Website, so they are applicable to any zone defined for the Website.

Each action has an associated *weight* which indicates the importance given to such action by the owner of the Website, ranging from 0 as minimum up to $n$ as max.
An Example of Actions

The owner has defined the following global actions with their weights:

- **A0**: PAGE REQUEST, $aw_0 = 1$ (default)
- **A1**: SEND VALUE, $aw_1 = 3$
- **A2**: CLICK EXIT, $aw_2 = 1$
- **A3**: CLICK STAY, $aw_3 = 2$.

Note that the owner has decided to give the highest weight to action $A1$. 
Calculation of a User Profile

The user profile is (re)-calculated dynamically:

- after each session (actions/time) and/or
- an explicit feedback is given by the user and/or
- a set of events occurred which are related to the user’s behavior and to certain “locations” of the Website.
Calculating the NOP

A NOP is a set of values between 0 and 1 for a given set of topics defined for the Website.

A NOP is determined by two parts:

- **Action Profile**: takes into account actions a user does in the zones on the topic $T_{pi}$,
- **Duration Profile**: takes into account the time a user spends on pages associated to the topic $T_{pi}$. 
Setting the Influence of Actions and Time

These two parts can be parameterized by the Website owner in accordance to his needs.

This means the owner can increase or decrease the impact of the time a user spends on a page compared to the actions he did.
**Action Profile**

\[
ActP(i) = \frac{\sum_q \left( \sum_t aw_t \cdot v(Tp_i, Z_q) \right)}{\sum_s aw_s}
\]

- we determine the zone \(q\), where the action occurred and obtain the associated topic weight \(v(Tp_i, Zq)\).
- we multiply this value by the sum of all weights for all occurred actions in this zone.
- we calculate the sum of all zones, where an action occurred and the associated topic lists contain topic \(i\), divided by the sum off all occurred action weights.
Duration Profile

\[
\text{DurP}(i) = \frac{\sum_{j} \left( \text{duration}(P_j) \times v(Tp_i, P_j) \right)}{\sum_{k} \text{duration}(P_k)}
\]

We consider each visited page \(P_j\), that contains the topic \(Tpi\) and multiply the time the visitor spent on this page \(\text{duration}(P_j)\) by its topic weight \(v(Tpi, Pj)\).

We sum these values and divide it by the total time.
Setting the Influence of Actions and Time

\[ x_i = a \times ActP(i) + b \times DurP(i), \]

where \( (a + b) = 1 \)

the parameters \( a \) and \( b \) are used to customize the ratio between \( ActP(i) \) and \( DurP(i) \)
User Feedback

To measure the “accuracy” of a NOP we ask the user for a feedback.

the user is asked directly to enter his preferences for a given set of topics.

We use the Feedback mechanism to “learn and compare” the interests of the users.

At first we do not question the user feedback.
User Feedback Profile

\[ FP_{um, tn} = \begin{pmatrix} 1.0 \\ 0.5 \\ 0.0 \end{pmatrix} \leftarrow \begin{array}{c} T_1 \\ T_2 \\ T_3 \end{array} \]
Four Measurements: Derived Profile

We calculate four different measures:

- NOP difference (ND) at time $t_i$
- Difference (D) at time $t_{i-1}$
- Difference (D) at time $t_i$
- FP difference (FD) at time $t_i$

First we calculate the differences (ND, D, FD) of the different profiles. These differences are then used to calculate a Derived Profile (DP) based on rules, trying to filter specific situations.
Advanced: Managing User Feedback

- Step 1: Definition of a Scope
- Step 2: Definition of a Filter
- Step 3: Obtaining Explicit Users Feedback
- Step 4: Filtering the User Feedbacks.
- Step 5: Clustering
- Step 6: Consistency check
- Step 7: Interpreting the results of the consistency check
Combing NOP and User Feedback

RP is calculated by integrating the two available profiles for the user: the Non Obvious Profile (NOP) which is automatically calculated based upon the behavior of the user (i.e., his actor profile and duration profile), with the explicit feedback given by the user in his Feedback Profile (FP), and filtered to the relevant scope.

\[
RP_{um, tn}(T_i) = \frac{NOP_{um, tn}(T_i) + f_i(S_i) \ast FP_{um, tn}(T_i)}{a + b + f_i(S_i)}
\]
Clustering Web Users

- The use of NOPs opens up several interesting possibilities, for example to cluster together visitors of a Website with “similar” interests and offer them then targeted/personalized e-services.

- Clustering Web users by their behavior can also be useful for measuring “trends” in a Web community, and again it can be a valuable information for creating customized e-services.

- Clustering Web users is also useful for e-CRM, where we build long-term-relationships and increase e-customer loyalty that is the degree to which a Web customer will stay with a specific vendor or a brand.
Architecture of Gugubarra 3.0.

- Clustering User Profiles
- Framing Strategy and User Behaviour Experiments: Preliminary Results
Architecture and Clustering
Gugubarra Architecture 3.0

**Analyzer**

- Strategy
  - Parameter \( a \), \( b \), threshold, ...

- GUI
- NOP-Engine (Web)
  - NOPs, Users

- Clustering-Module (serverside)
  - Clusters

- Interface

**Designer**

- Semantics Strategy

- JavaServerFaces
  - Spring

- Design-Module (Web)
  - Zones, Topics, Actions

- Java

- Hibernate

- DB Persistence

- Interface

- CMS 1 (OpenCMS)
  - Plug-in
  - Web-Service (Axis2)

- CMS 2
  - Plug-in

- GUI Web client

- Java Applet, ...
Clustering Web Users

- In most research, Web users are clustered by their click streams or by their visited pages.

- By using the Non-Obvious Profiles approach we have the possibility to cluster Web users with "supposed similar," interests, by considering: the content of the Web pages, the users activities and their explicit feedback.

- The use of NOPs opens up several interesting possibilities, for example to cluster together users of different Websites.
Steps for Clustering NOP, OP and FP

1. Choose different attributes
   - Topics from the NOP
   - Attributes from the OP e.g. nationality, age
   - Other profile attributes e.g. user activity (Web 2.0)

2. Combine with queries
   - e.g. only users with high interest in Java in FP

3. Cluster all
Scale of Interest

- Ordinal scale -> \{no, little, strong, total\} interest. (g=4)

![Diagram showing the scale of interest with centroids and interpretations]

- Interpretation:
  - Strong interest in Tp2 and total interest in Tp1
Example:
Given the following Profiles…

- NOP, OP, FP and more from different users:

```
<table>
<thead>
<tr>
<th>NOP</th>
<th>User 1</th>
<th>User 2</th>
<th>User 3</th>
<th>User 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big Cars</td>
<td>0.2</td>
<td>0.15</td>
<td>0.9</td>
<td>0.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FP</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Big Cars</td>
<td>0.2</td>
<td>0.5</td>
<td>0.8</td>
<td>0.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OP</th>
<th>User 1</th>
<th>User 2</th>
<th>User 3</th>
<th>User 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-mail</td>
<td>X@Y</td>
<td>Y@Z</td>
<td>C@P</td>
<td>P@B</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RP</th>
<th>User 1</th>
<th>User 2</th>
<th>User 3</th>
<th>User 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big Cars</td>
<td>0.2</td>
<td>0.2</td>
<td>0.8</td>
<td>0.3</td>
</tr>
</tbody>
</table>
```
Customize the Clustering (2)

- Select *scale of interest* for NOP:
  e.g. {little, medium, strong}

- NOP values are aggregated to the *scale of interest*:

<table>
<thead>
<tr>
<th>NOP</th>
<th>User 1</th>
<th>User 2</th>
<th>User 3</th>
<th>User 4</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big Cars</td>
<td>0.2</td>
<td>0.15</td>
<td>0.9</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(little)</td>
<td>(little)</td>
<td>(strong)</td>
<td>(medium)</td>
<td></td>
</tr>
</tbody>
</table>
Customize the Clustering (3)

| User 1  | User 2  | User 3  | User 4  | ...
|--------|--------|--------|--------|--------
| Big Cars | (little) | (little) | (strong) | (medium) |

- Select **Algorithm Parameters:**
  
e.g. No. of Clusters = 3, randomly assigned Centroids
Fuzzy Clustering: Membership Degree

\[ \sum_{i=1}^{k} \mu_{ij} = 1, \quad 1 \leq j \leq n \]
Framing Strategies and Web User Behaviour

Reputation, framing strategies and user’s choice of content on the Web: an empirical study
Ioanna Constantiou, Natascha Hoebel, and Roberto V. Zicari
Several Experiments conducted

- Work in cooperation with the Copenhagen Business School (CBS)

- Methodology based on Behavioural Economics (Framing Effect)


Pre-Test

Asked selected users “What factors are important to the users, when choosing a link on a Website?”

- Source of content
- Reputation of the source
- Date of content development
- Content type (link vs. attachment)
- Time to obtain the content
- I know the given source
- The source of content is mentioned
- The source was recommended
- Place of content (top or bottom of the page)
Pre-Test: Initial Findings

**Reputation of the source** has a significant impact on user attention for Web-related context information and consequently choice compared to other attributes such as date of content.

Experiments: Explore relationship between *background context and local context* respect to reputation of the source.
The Research Framework

H1: Reputation affects the choice of an online information service.

H2: Enriched content affects the choice of an online information service.

H3: The interaction between reputation and enriched content affects the choice of online information service.
Experiments

In order to explore if content providers can affect the individual’s choice of an information service offered online.

Four experiments, based on a factorial independent group 2 x 2 design.

ANOVA Analysis, Pearson Chi-Square.
Type of Questions

- Choice questions, enriched content

  - Known source (BBC)
  - Unknown source (Blog)

Matching Tasks

- **Known source** with percentage of reads

- **Unknown source**: “propose a percentage for that would make both links equally attractive”
Example Choice Question

<table>
<thead>
<tr>
<th>1st of 9 Questions <em>Buddhism</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>If you had to read more about Buddhism from the sources listed below, which one would you choose? Please make your choice without copying the URL to your Browser!</td>
</tr>
<tr>
<td>Source B:</td>
</tr>
</tbody>
</table>
1 Please choose your preferred news item:

<table>
<thead>
<tr>
<th>Obama, Democrats cut health care deal</th>
<th>Obama, Democrats cut health care deal</th>
<th>I would not prefer any of these two news items.</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBC News - 2 hours ago</td>
<td>Seeking Alpha (blog) - 2 hours ago</td>
<td>The White House and congressional leaders returned to the White House this afternoon to negotiating other differences between the bills passed by the House and Senate last year. “We are hopeful that by the end of day they will have worked through -- House, Senate and the president -- worked through any number of issues that remain outstanding,” said White House spokesman Robert Gibbs. Obama told House Democrats this afternoon that the new legislation will cover more Americans and reduce health care costs. He dismissed claims by Republicans it will do the opposite, and urged Democrats to hang tough in the face of political problems over the legislation. “Believe me, I know how big a lift this has been,” Obama said. “I see the polls ... The worst fears will prove groundless.” If Republicans want to campaign against the health care bill, Obama said, “that is a fight I want.”</td>
</tr>
</tbody>
</table>

50% of 90 users recommended this news
70% of 10 users read this news
Initial Results

Framing Strategies affecting user choice

Source (Reputation)

Text (no next, short text, long text)

Nice/attractive picture
Content related picture/video

Percentage of recommendations
(10%, 50%, 70%, 90%)

Percentage of readership
10%, 50%, 70%, 90%

Total Number of Users
(10 users, 90 users, 200 users, all users)
Future Work

3. PlaceWeb: Definition of a **Methodology** for Placement of Information on a Website

2. Framing and user choice
   - Experiments
   - Strategy
   - Assumptions

1. Gugubarra Engine

feedback

Web 2.0 User Profiles
No work on User Profiles can avoid issues on Ethics.

We would like to distinguish between two classes of Ethics problems:

- What we call the *Business Code*, including issues such as Data Protection, Security, and a broader issue we call

- *Social Code*: which relates to the issues on how addictive could be the new Internet technologies which encourage a higher level of “stickness” to a Website.

We would like to raise the awareness in the research community of the danger that improper use of such technology can be very damaging. We believe this issue requires a broad discussion and attention in the research and industrial community.
Gugubarra Resources

http://www.dbis.informatik.uni-frankfurt.de/research/
Thank You!

Contac: gugubarra@dbis.cs.uni-frankfurt.de