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
Gugubarra Main Research Areas

1. Inferring the “traceable“ interests of Web users.

2. Framing strategies and user choice of content on the Web.

3. PlaceWeb: Definition of a methodology for placement of information on a Website.
Gugubarra Main Research Areas

1. User Profiles

2. Framing and user choice Experiments

validates

feedback

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

2. Framing Strategy Assumptions

1. Gugubarra Engine
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 during his visits on the site.
Goals

- Enable the owner (manager) of a Website to understand better the 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”…

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 wanted 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.
- Two research prototype systems implemented between 2004 and 2007.
- Currently designing a new prototype called Gugubarra 3.0.
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 Page ==> Topics (Not all with same importance…)

BROADBAND

ROUTES

HOTELS
Actions can be performed on a Web Page
Some actions are “more important” than others...

[ All Experts ] [ Paul Harmon articles ] [ Paul Harmon Bio ]

Paul Harmon
An OMG Update For Managers

1. The Continuing Relevance of the OMG
posted: 2005.05.04 - Size: 24 KB - Count: 1

In the months ahead I will explore a number of distributed computing issues facing IT managers. Most IT managers are caught in the Internet-e-business whirlwind. How should your company respond to the Internet? How should your company organize your Web site initiatives? What kind of B2B and B2C applications should your company build? The list goes on and on and usually ends with; How soon can it be done?
Is there a role that the Object Management Group (OMG) and its CORBA architecture can plan in any of this?

2. Java, Enterprise JavaBeans and CORBA
posted: 2005.05.04 - Size: 27 KB - Count: 1

It's no secret that most companies are developing new e-business applications in Java and many are relying on Enterprise JavaBeans. Indeed, a recent Cutter Consortium survey reported that 75% of the companies surveyed were using EJB, application server products for new enterprise development, while only 25% were using CORBA. This is an important point for those responsible...
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 a Page with Zones and Topics

Page *P1* with three zones, *Z1*, *Z2* and *Z3*.

- In zone *Z1*: banners of the topic ‘**Broadband**’ are displayed,

- In zone *Z2* and zone *Z3* information on **routes** and maps are displayed, BUT the content of the zone *Z2* and *Z3* differs:
  - Zone *Z2*: Contains the *Search* field to find a place and the *Calculate Route* field to get a route calculation.
  - Zone *Z3*: Contains Marketing information’s etc.
An Example of Topics and Zones

- **Global topics:**
  - $\text{Tp1} = \text{Broadband}$,
  - $\text{Tp2} = \text{Finding a Location}$,
  - $\text{Tp3} = \text{Calculating Routes}$,
  - $\text{Tp4} = \text{Insurance}$,
  - $\text{Tp5} = \text{Hotels}$.

- **Page 1:**
  - $\text{Z1} = \{\text{Broadband}, 0.5\}$,
  - $\text{Z2} = \{\text{Finding a Location}, 0.6\}$, \{Calculating Routes, 0.8\},
  - $\text{Z3} = \{\text{Finding a Location}, 0.3\}$, \{Calculating Routes, 0.3\}.

- **Page 2:**
  - $\text{Z4} = \{\text{Insurance}, 0.5\}$, \{Calculating Routes, 0.9\},
  - $\text{Z5} = \{\text{Calculating Routes}, 0.9\}$, \{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

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

- we determine the zone \( q \), where the action occurred and obtain the associated topic weight \( v(Tp_i, Z_q) \).
- 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**

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

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

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

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

the parameters \(a\) and \(b\) are used to customize the ratio between \(\text{ActP}(i)\) and \(\text{DurP}(i)\)
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.
Four Measurements: Derived Profile

We calculate four different measures:

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.
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
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)

**INTEREST Tp₂**

- total 1.0
- strong
- little
- no

**“centroids“**

interpretation: strong interest in Tp₂ and total interest in Tp₁
Clustering Algorithm for Categorical Data

We want to cluster the NOP, OP, FP with a variation of the clustering algorithm:

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>
<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>Color</td>
<td>light blue</td>
<td>red</td>
<td>blue</td>
<td>purple</td>
<td></td>
</tr>
<tr>
<td>Nationality</td>
<td>Italian</td>
<td>Italian</td>
<td>German</td>
<td>Swiss</td>
<td></td>
</tr>
<tr>
<td>Activity</td>
<td>little</td>
<td>little</td>
<td>strong</td>
<td>strong</td>
<td></td>
</tr>
</tbody>
</table>
Customize the Clustering (1)

Select attributes:

<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>Color</td>
<td>light blue</td>
<td>red</td>
<td>blue</td>
<td>purple</td>
<td></td>
</tr>
<tr>
<td>Nationality</td>
<td>Italian</td>
<td>Italian</td>
<td>German</td>
<td>Swiss</td>
<td></td>
</tr>
<tr>
<td>Activity</td>
<td>little</td>
<td>little</td>
<td>strong</td>
<td>strong</td>
<td></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>
</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>
<tr>
<td>(little)</td>
<td>(little)</td>
<td>(strong)</td>
<td>(medium)</td>
<td></td>
</tr>
<tr>
<td>Color</td>
<td>light blue</td>
<td>red</td>
<td>blue</td>
<td>purple</td>
</tr>
<tr>
<td>FP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Customize the Clustering (3)

<table>
<thead>
<tr>
<th></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>(little)</td>
<td>(little)</td>
<td>(strong)</td>
<td>(medium)</td>
<td></td>
</tr>
<tr>
<td>Color</td>
<td>light blue</td>
<td>red</td>
<td>blue</td>
<td>purple</td>
<td></td>
</tr>
</tbody>
</table>

Select **Algorithm Parameters:**

e.g. No. of Clusters = 3, randomly assigned Centroids
Why Cluster?
Suppose you want to sell a blue Big Car

<table>
<thead>
<tr>
<th></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>(little)</td>
<td>(little)</td>
<td>(strong)</td>
<td>(medium)</td>
<td></td>
</tr>
<tr>
<td>Color</td>
<td>light blue</td>
<td>red</td>
<td>blue</td>
<td>purple</td>
<td></td>
</tr>
</tbody>
</table>

$\uparrow$

$\text{Big Car}$
Fuzzy Clustering: Membership Degree

\[ \sum_{i=1}^{k} \mu_{ij} = 1, \quad 1 \leq j \leq n \]
Fuzzy Clustering: Distance Measurement for Categorical Data

Comparing attribute values

\[ d_c(V_i, X_j) = \sum_{l=1}^{p} \delta \left( v_{i,l}, x_{j,l} \right) \]

\[ \delta \left( v_{i,l}, x_{j,l} \right) = \begin{cases} 0, & v_{i,l} = x_{j,l} \\ 1, & v_{i,l} \neq x_{j,l} \end{cases} \]
Assign Users to Cluster…

Example of two users in one cluster

Cluster A
Big Cars = strong
Color = blue

User Sara
Big Cars = medium
Color = blue

User Leo
Big Cars = medium
Color = blue

Then update the cluster Centroids …
Minimize Distance

Comparing attribute values

\[ d_c(V_i, X_j) = \sum_{l=1}^{p} \delta (v_{i,l}, x_{j,l}) \]

\[ \delta (v_{i,l}, x_{j,l}) = \begin{cases} 
0, & v_{i,l} = x_{j,l} \\
1, & v_{i,l} \neq x_{j,l} 
\end{cases} \]

\[ J_m (U, V : X) = \sum_{i=1}^{k} \sum_{j=1}^{n} (u_{ij})^m d_c(V_i, X_j) \] → Minimize

- membership degree
- distance
Example: Updating the Centroids

\[ X_i = [\text{color, size}] \]

\[
\sum_{x_{j,i}=a_l^{(r)}} \mu_{ij}^m \geq \sum_{x_{j,i}=a_l^{(r)}} \mu_{ij}^m, \quad 1 \leq t \leq n_l
\]

\[
\sum \mu_{1j} = 1 + 0.75 = 1.75 \quad \text{size = big}
\]

\[
\sum \mu_{1j} = 0.7 + 0.7 + 0.75 = 2.15 \quad \text{size = small}
\]

\[ 2.15 > 1.75 \quad \rightarrow \quad V_1 = [\text{blue, small}] \]
Framing Strategies and Web User Behaviour: initial results of our experiments...
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 in 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”
Preliminary Results

Which of the following Framing Strategies affects user choice?

- nice/attractive picture
- content related picture
- coloured font and font-background
- content related video
- percentage of recommendations
- percentage of readership
**Future Work**

1. **Gugubarra Engine**
   - Feedback

2. **Framing and user choice**
   - Experiments

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

**Web 2.0 User Profiles**
Business and Social Ethics

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 “stickiness” 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/?mode=art&l=e&aid=260&tmid=5&smid=16](http://www.dbis.informatik.uni-frankfurt.de/research/?mode=art&l=e&aid=260&tmid=5&smid=16)

- Talk at Google, HQ, Mountain View, California, 25th of July, Roberto V. Zicari: [http://video.google.com](http://video.google.com)

- Creating User Profiles of Web Visitors using Zones, Weights and Actions (EEE 08)

- On Clustering Visitors of a Web Site by Behavior and Interests (AWIC 2007)

- The Design of Gugubarra 2.0: A Tool for Building and Managing Profiles of Web Users (WI 2006)


- Building and Evaluating Non-Obvious User Profiles for Visitors of Web Sites (CEC 04)
Thank You!

- **Take Home Message**

  Gugubarra is a tool to support Website owners in understanding the interests of the users. It takes defined Zones, Actions and Durations into account to create user NOPs.

- **Contact:** gugubarra@dbis.cs.uni-frankfurt.de