Towards an Empirical Evaluation of CBR Approaches for Product Recommendation - In Electronic Shops

Stephan Weibelzahl\textsuperscript{1}, Ralph Bergmann\textsuperscript{2}, and Gerhard Weber\textsuperscript{3}

\textsuperscript{1} Universität Freiburg, Germany,  
weibelza\texttt{at}uni-freiburg.de
\textsuperscript{2} Universität Kaiserslautern, Germany  
bergmann\texttt{at}informatik.uni-kl.de
\textsuperscript{3} Pädagogische Hochschule Freiburg, Germany  
weber\texttt{at}cogpsy.uni-trier.de

Abstract. A CBR-approach to product retrieval is often said to be better than just listing the products. In this paper we present an empirical evaluation as an approach to give scientific ground to this hypothesis. Six steps of evaluation are proposed, starting with the specification of the problem and ending with a final decision between the alternatives. Such a procedure as a general practice enhances the objectivity and validity of an empirical evaluation.

1 Introduction

A CBR-approach to product-retrieval (Bergmann, Breen, Göker, Manago & Wess, 1999; Wilke, 1999) is often said to be better than just listing the products. It seems to be obvious that recommending the most similar cases is more comfortable for customers. For example, a product recommendation system will come up with a set of five offers (the most similar cases) instead of overwhelming the customer with messages like "1000 matches found" or leaving him alone with "no product found in database" if the query is formulated too broadly or too narrowly.

Nevertheless, evaluations in this field are hard to find. If carried out at all, evaluations are usually limited to technical criteria such as performance (e.g., retrieval time, amount of main memory, or disk space), correctness, or scaling validity.

We feel that these are important criteria, especially when starting to explore a new method. But one of the major advantages of CBR over traditional approaches to problem solving is the cognitive adequateness combined with the ability to offer solutions in a way which is very comfortable for the user. Thus the CBR-approach is often cited as being "better". However, it is unclear what "better" means in this context and whether there is scientific ground for this claim.
In this paper we introduce an empirical approach to evaluation which applies several criteria concerning the quality of user interaction to demonstrate exemplarily the superiority of a case-based product recommendation system over traditional product catalogues.

2 Evaluation Approaches

In general, as was outlined by Althoff (1997a), we have to distinguish between evaluating one software system, evaluating different software systems and tools, and evaluating different software system development methodologies. In this paper we will concentrate on the second view, i.e., we will compare three different software systems, although the empirical approach can be used for all of the three viewpoints.

For comparing different software systems Althoff (1997b) lists many technically and ergonomically oriented criteria, which are presented in table 1.

<table>
<thead>
<tr>
<th>Ergonomic criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>control management</td>
</tr>
<tr>
<td>data acquisition</td>
</tr>
<tr>
<td>explainability</td>
</tr>
<tr>
<td>user acceptance</td>
</tr>
<tr>
<td>time criticality</td>
</tr>
<tr>
<td>rapid prototyping</td>
</tr>
<tr>
<td>validation</td>
</tr>
<tr>
<td>data maintenance</td>
</tr>
<tr>
<td>acustoming</td>
</tr>
<tr>
<td>organisational and technological impact</td>
</tr>
<tr>
<td>interactiveness</td>
</tr>
<tr>
<td>additional tasks</td>
</tr>
</tbody>
</table>

We argue that in addition to these criteria the quality of user interaction should be a component of evaluations. User interaction satisfaction should (and not to say must) be measured empirically using the system either in real life or under more controlled conditions in the laboratory with users interacting with the system instead of just having experts rating the systems. Users often behave quite differently from what developers expect of them.

Before presenting an example of an empirical evaluation we will outline the structure of the evaluated system.

3 CASTLE: Recommending Vacation Homes

CASTLE (Weibelzahl, 1999b) is a case-based system recommending vacation homes in France based on both the customers’ abstract needs and their actual preferred attributes of the product.

The retrieval is performed in two phases. In the first phase of interaction with a customer the system collects data about the customer, such as which activities (e.g., sports, recreation, culture, etc.) he or she prefers and what the most
important motivation is (e.g. meeting people, relaxing, etc.). Figure 1 presents a screenshot of the input mask. This information is used to construct a user model of this customer. Table 2 presents an example of such a user model.

Then CASTLE retrieves a customer, who has the most similar past user model, from the case base of customers. A case in this case base consists of a user model as the problem and certain product attributes as the solution (see figure 2). As long as no other evidence is available, CASTLE assumes that the current user will prefer the same product attributes as the previous customer, because of similar general preferences. Thus, as long as the current customer does not express a preference for, let’s say, vacation homes with microwave ovens, CASTLE implicitly includes "no microwave oven" in its retrieval query if this was true for the previous customer.

Table 2. Example of a user model

<table>
<thead>
<tr>
<th>User model of needs and desires</th>
<th>Activities</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Motives</strong></td>
<td></td>
</tr>
<tr>
<td>Recreation motives:</td>
<td>Relaxation: yes</td>
</tr>
<tr>
<td>Adventure motives:</td>
<td>Sport: no</td>
</tr>
<tr>
<td>Freedom motives:</td>
<td>Hobby, Fun: no</td>
</tr>
<tr>
<td>Social motives:</td>
<td>Culture, education: no</td>
</tr>
<tr>
<td>Other motives:</td>
<td>Active recreation: yes</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In the second phase, CASTLE builds a query from this newly inferred information (i.e., preferred product attributes) and retrieves the most similar product from the case base of products. The customer may refine this query by changing these preferred product attributes and retrieving similar homes again.

If the customer finally chooses a product from the catalogue, a new case will be created by merging the customer’s user model and the attributes of the chosen product. These attributes obviously meet the customer’s needs and preferences in an optimal way. This experience can be useful for creating offers for other customers.

Thus, the major advantage of CASTLE is its ability to learn a lot of most likely preferred product attributes from a small set of needs (Weibelzahl, 1999a). This makes the interaction more comfortable for the user, who does not need to fill in all these data.

4 Evaluating the Quality of User Interaction

Following the steps of an evaluation proposed by Wottawa (1990) we will give an introduction to empirical evaluations by explaining how we realised the evaluation of the CASTLE-System.
4.1 Step 1: Specification of the Problem

In the beginning of an empirical evaluation the problem should be specified precisely. What do you want to know in the end? Often, as mentioned above, a vague question such as "Is it better?" will be the starting point. The question should contain information about specific aspects to be evaluated and specific target groups. For example, when evaluating CASTLE, we wanted to find out at least two aspects:

- **Study I:** Is a case-based system able to map the abstract language of customers to the concrete attributes of the products by learning on its own? More precisely: Does a case-based system learn to adapt to the needs of customers, and thus increase the customer interaction satisfaction?

- **Study II:** Does a case-based system improve customer interaction satisfaction in comparison to existing systems?

That is, we are interested in the impact of the learning ability and the system in general concerning the user interaction. In this case users are defined as customers searching the internet for holiday flats.

4.2 Step 2: Identifying Alternatives

Secondly, the alternatives for the comparison are chosen. When evaluating the total quality of a system other existing approaches to the same problem might be used. But when comparing specific aspects of a system such as the quality of interaction the choice must be made carefully. The alternatives should differ only with respect to the question (e.g., CBR-approach vs. traditional approach), but all remaining aspects should be equivalent. What seems to be commonplace at first glance, is often difficult to realise. For example, different retrieval methods are often confounded with other variables such as response time which might have impact on the user interaction, but are not intended to be evaluated.

In the CASTLE-Evaluation, the alternatives for study I were chosen as follows: In the first phase of the study the system constructs the customer case base by interacting with real customers. People interested in renting a vacation home in France were able to use CASTLE from their home computers while logged into the Internet. The customer case base was initialised with 14 artificial cases judged to be realistic. Note that this was not necessary, as it is theoretically sufficient to feed the model with a single case that does not even need to be realistic. But in order to accelerate the learning phase and to not deter the customers by absurd offers, expert knowledge was used.

Within 6 weeks, 38 new cases were generated in the course of CASTLE's interaction with new 38 customers asking to book a certain home. After this period the system was updated with the new information and run again for another 3 weeks, interacting with 10 new customers.

That is, the first "dumb" system used by 38 customers (group 1) is compared to the updated system used by 10 customers (group 2) to address the question of whether the system is able to learn to adapt to the needs of customers.
To test the second hypothesis, participants were assigned to one of three conditions in a laboratory experiment:

- **Experimental group 1 (EG₁):** In this condition participants interacted with the same version resulting from the first study, that is the complete CASTLE-system.

- **Experimental group 2 (EG₂):** This group used a system without user modelling. Thus, instead of first exploring customer characteristics the system started with an attribute search. This means that only the second phase of the CASTLE retrieval system is used.

- **Control group (EG₃):** The third group used a simple electronic catalogue on a web page, organised by geographical region. No other search criteria were available. This electronic catalogue contains almost the same information as a hard-copy catalogue available at travel agencies. The order in which products are listed within a category is random.

The goal of this categorisation was to vary the degree of user modelling (EG₁ vs. EG₂) and the use of cases (EG₁ and EG₂ vs. EG₃). Both alternatives are used in daily practice, but by reducing the functionality while retaining unchanged interface surface we assured an comparability. The content of the database, the retrieval time, etc. remain the same.

A laboratory task was chosen to assure a more standardised situation under controlled conditions. All subjects were asked to find a suitable vacation home within at least 15 minutes. A sample of 60 persons was explored in this second study.

Note that from a statistical point of view repeated measurement is preferable, but hard to realise in this context. When interacting with the system for the second time the customers are already familiar with the functionality of the system and, even more importantly, with the content of the database. Thus, a between subject design was applied.

### 4.3 Step 3: Collecting Criteria

Third, the abstract question formulated above must be reformulated with respect to specific measurable criteria.

Customer satisfaction is a complex variable that is difficult to measure. In this context it is defined as follows: customers are more satisfied with an interaction if (a) they don’t have to decide about many attributes; (b) if they find a suitable product very quickly; (c) if they don’t have to start many retrievals.

Four different measures were used for defining customer satisfaction:

- total duration of interaction \((T)\): sum of the duration of interaction before \((T_1)\) and after \((T_2)\) the first retrieval (see figure 4)
- total number of answers \((A)\) given by the user before \((A_1)\) and after \((A_2)\) the first retrieval
- number of retrievals \((R)\)
the Questionnaire of User Interaction Satisfaction (QUIS), developed and re-
visited by Chin, Diehl und Norman (1988) is commonly used in HCI research
to compare different types of interfaces with regard to user satisfaction. QUIS
has been shown to yield highly reliable data (e.g., Harper & Norman, 1993;
Slaughter, Harper & Norman, 1994). A subset of 17 items of this question-
naire, which were useful in this context, was translated into German and
presented to the users after the interaction. Two items were added. The first
one explicitly addresses the degree to which the chosen product meets the
customer’s expectations: The vacation home that I chose has what I was
looking for. The second item addresses the user’s certainty that the optimal
product was found: I think there is a more suitable vacation home in the
catalogue.

Figure 4 gives an overview of which measure was applied in which phase of
interaction.

<table>
<thead>
<tr>
<th>Search on the level of customers’ characteristics</th>
<th>Search on the level of attributes</th>
<th>QUIS Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>• duration ((T_1))</td>
<td>• duration ((T_2))</td>
<td></td>
</tr>
<tr>
<td>• number of attributes ((A_1))</td>
<td>• number of attributes ((A_2))</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• number of retrievals ((R))</td>
<td></td>
</tr>
</tbody>
</table>

**Fig. 4.** Phases of a session and related measures

4.4 **Step 4: Target Analysis**

Before the measurement data is collected, the optimal peculiarity of every crite-
ron must be specified. In addition, the relative importance of the criterion may
be specified by a weight.

For CASTLE, we defined that the shorter the duration necessary for finding a
vacation flat, the better the quality of interaction. We expected the time needed
to extract the user model would be compensated by the advantage of the the
better proposal.

The same is true for the criteria number of retrievals and number of attributes:
The less answers and retrievals, the higher the customer satisfaction. The latter
seems to be a less important criterion since many refinements with little change
in the attributes may still be comfortable for the customer. So we will mainly
take duration, QUIS and number of attributes into consideration.

4.5 **Step 5: Measuring and Calculating Profit**

The results of the evaluation can be presented as shown in table 3 and 4.
Table 3. Overview of the results of study I (internet)

<table>
<thead>
<tr>
<th>criterion</th>
<th>group 1</th>
<th>group 2</th>
<th>T</th>
<th>df</th>
<th>α</th>
<th>power(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>total duration (\bar{T}) (s)</td>
<td>628</td>
<td>436</td>
<td>1.79</td>
<td>19</td>
<td>.089</td>
<td>.28</td>
</tr>
<tr>
<td>duration of search (T_2) (s)</td>
<td>390</td>
<td>274</td>
<td>1.24</td>
<td>19</td>
<td>.230</td>
<td>.28</td>
</tr>
<tr>
<td>proportion (\bar{A}_2)</td>
<td>1.80</td>
<td>1.85</td>
<td>-0.10</td>
<td>19</td>
<td>.922</td>
<td>.28</td>
</tr>
<tr>
<td>decisions (A)</td>
<td>82.4</td>
<td>65.1</td>
<td>0.92</td>
<td>19</td>
<td>.369</td>
<td>.28</td>
</tr>
<tr>
<td>refinements (A_2)</td>
<td>56.8</td>
<td>39.0</td>
<td>1.04</td>
<td>19</td>
<td>.313</td>
<td>.28</td>
</tr>
<tr>
<td>retrievals (R)</td>
<td>3.08</td>
<td>1.87</td>
<td>1.16</td>
<td>19</td>
<td>.262</td>
<td>.28</td>
</tr>
</tbody>
</table>

\(^a\) power = 1 − \(\beta\) with \(\alpha = .05\) and an assumed effect-size of \(\omega^2 = .5\)

Table 4. Overview of the results of study II (laboratory)

<table>
<thead>
<tr>
<th>criterion</th>
<th>EG(_1)</th>
<th>EG(_2)</th>
<th>EG(_3)</th>
<th>F</th>
<th>df</th>
<th>α</th>
<th>power(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>total duration (\bar{T}) (s)</td>
<td>1057</td>
<td>1082</td>
<td>1113</td>
<td>0.25</td>
<td>2.59</td>
<td>.782</td>
<td>.39</td>
</tr>
<tr>
<td>duration of search (T_2) (s)</td>
<td>352</td>
<td>1082</td>
<td>1113</td>
<td>81.2</td>
<td>2.59</td>
<td>.000</td>
<td>-</td>
</tr>
<tr>
<td>QUIS</td>
<td>5.75</td>
<td>5.29</td>
<td>6.27</td>
<td>1.93</td>
<td>2.59</td>
<td>.155</td>
<td>.39</td>
</tr>
<tr>
<td>decisions (A)</td>
<td>123</td>
<td>147</td>
<td>-</td>
<td>2.56</td>
<td>1.39</td>
<td>.568</td>
<td>.35</td>
</tr>
<tr>
<td>refinements (A_2)</td>
<td>95</td>
<td>147</td>
<td>-</td>
<td>102</td>
<td>1.40</td>
<td>.027</td>
<td>-</td>
</tr>
<tr>
<td>retrievals (R)</td>
<td>7.48</td>
<td>8.05</td>
<td>-</td>
<td>0.16</td>
<td>1.40</td>
<td>.688</td>
<td>.35</td>
</tr>
</tbody>
</table>

\(^a\) power = 1 − \(\beta\) with \(\alpha = .05\) and an assumed effect-size of \(\omega^2 = .25\)

If available, the total profit in dollars may be calculated. For our purposes a one factor ANOVA\(^1\) was performed for every criterion to test the statistical hypotheses. In both studies there is a trend of all measures in the predicted direction, although only two results reach statistical significance.

Exemplarily, we explain the interpretation of the results for a single criterion: total duration in table 4. Columns \(EG_1\) to \(EG_3\) present the average duration a subject in each group needed to find a suitable vacation home. The next column presents the \(F\)-value which is the proportion \(\frac{\text{variance of treatment}}{\text{variance of error}}\). This value shows the relative amount of variance in the data which is explained by the experimental conditions. By taking into account the number of subjects and the number of treatments, coded as \(\text{degrees of freedom (df)}\), the statistical significance of the difference (\(\alpha\)) is computed. Which column is to be interpreted next depends on the significance of the results. For non-significant results (\(\alpha > .5\)) the test power expresses the probability of finding an existing difference by the applied design. For significant results the effect-size (\(\omega^2\)) shows whether a small, medium, or large (\(\omega^2 > .5\)) effect was found.

In Study I, no criterion reached significance. Nevertheless all results point to the predicted direction. Due to a lack of time the initially planned number of subjects (Study I: 76; Study II: 72) was reduced. This is the reason for the low test power.

In Study II, two criteria reached significance. Especially the number of refinements (\(A_2\)) is of interest. Obviously CASTLE’s user model improved the first

\(^1\) ANalysis Of VAriance: statistical method to compare the variances of two or more data sets in reference to one variable.
proposal clearly, so the customers had to make less refinements. In both studies the reduced number of subjects had a negative impact on the results. More subjects would have provided a better ground for a final interpretation.

4.6 Step 6: Decision

Based on the collected data and its analyses it should be possible to answer the initial question. We draw the following conclusions from the evaluation of CASTLE.

First, case-based sales agents seem to be more popular than choose-out-of-a-list systems ($EG_1$ better than $EG_3$ and $EG_2$ better than $EG_3$ for all criteria). Most systems can thus be improved. The fuzzy search seems to be closer to what customers expect from a sales agent. The quality of the offers further improves when a CBR algorithm considers not only the wording of the query but also the structure of the domain.

Second, user modelling is superior to the traditional CBR approach, i.e., the two-step system with an abstract user model worked better than the one-step system. Customers seem to appreciate being given the opportunity to describe their goals and interests. Most customers were more satisfied when they had this opportunity ($QUIS(EG_1) > QUIS(EG_2)$).

Third, adaptability can be learned from the experience of a system. Customers were more satisfied when interacting with the experienced system (group 2 better than group 1 with respect to all criteria).

5 Conclusion

There are at least four things concerning empirical evaluation to be concluded from the sections above.

Empirical evaluations produce data that is not available using other methods. Assessing the quality of interaction as performed in the example above needs to be researched empirically. The resulting data should be seen in combination with other methods of evaluation such as technical criteria.

The empirical evaluation described above is the first one to give scientific ground to the claim that CBR systems for product recommendation improve the quality of interaction compared to a simple product catalogue.

Sometimes it is possible to observe the system in real life, that means interacting with users that perform a usual task. Directly assessable data such as durations or number of retrievals might help to measure the performance of the evaluated system.

Some situations might require evaluation under more controlled conditions in the laboratory in order to obtain more standardised and less confounded data. There is a trade-off between objectivity on the one hand and validity with respect to daily life on the other hand. A mixture of both strategies might reduce the disadvantages of both views.
Finally, the previous chapter should have pointed out that evaluation is never completely objective, as there are many issues, criteria, and measures to be defined without obvious specifications. An open, extended documentation of all tests and collaboration of researchers favouring different systems can enhance the quality of an evaluation considerably.

References


