The problem statement

In my master thesis I am going to investigate questionnaire system models for classification. These are questionnaires that depending on the answers a subject gives to the question classify some property of the subject. Questionnaires are used in a lot of different fields, from medical diagnosis to questionnaires about political interest (stemwijzer.nl) and multiple choice examination. The design of these questionnaires is mostly done by hand with as result a static list of questions with which some kind of classification algorithm can identify the class.

Some questionnaires do not result in a static list of questions but have the possibility to use ‘skip-patterns’. The idea is that based on the answer to a certain question in the questionnaire another question becomes irrelevant. For example, after answering the question ‘do you have any children’ it is no longer relevant to ask the question ‘what age are your children’.

A disadvantage of fixed order questionnaires and ‘skip-patterns’ is found in question relevance. Some questions are more relevant for some subjects than other questions. For example, if we answer the question ‘Do you own a car’ with ‘no’, the question ‘Do you travel by public transport’ becomes less relevant because the probability the person uses public transport increases, but could very well still be relevant. Also the question ‘How do you travel to work’ could become more relevant because if the subject would have owned a car, the probability he uses a car to go to work is higher. Another disadvantage any form of questionnaires contains is that the number of questions that can be asked is limited. If we could ask more (and more specific) questions, the classification could become more precise or in other words, if we know more about a subject we can perform better classification.

The solution:

In my thesis I will present a new model, in which question ordering is done automatically with a system that chooses questions from a large pool of candidate questions by looking at the questions relevance with respect to the current answers the subject has given. The questions and actual classification will probably not change.

Research questions

The working title is ‘Optimal probabilistic question ordering in Questionnaires’ and raises the main question:
Is there a way to determine ‘relevance’ of a question and can it be used to create a more optimal questionnaire?”

This again raises a lot of smaller questions:

- How is a classic classification questionnaire system modelled
• How to model questions
• How to measure an answers influence on a classification
• Can we define an optimal way of ordering questions
• What is the impact of changing the order in questions (literature)
• How do we measure if our model performs better than the classic approach

Extra
Because I will write my master thesis at a company ‘Hexon b.v.’, i will also look into designing a way to apply the model to a car search system. The idea behind this is to design a system in which people that want to buy a car can answer a series of questions in which the best car is presented for them.

Literature
Although there is not a lot of literature available on this subject, there are a lot of areas related to it.

About classic and skip-pattern questionnaire models:

Non-adaptability measures in the pseudo-questionnaires context C. Bertoluzza a, V. Doldi b, J. Jiménez c, S. Montes d,

Using Graph Theory to analyze Skip Patterns in Questionnaires
Jim Fagan Brian V. Greenberg Statistical Research Division Bureau of the Census Washington, D.C. 20233
...

About questionnaire design:

Improving Survey Questions: Design and Evaluation
By Floyd J Fowler Edition: 6 Published by SAGE, 1999

Imperfect Answers in Multiple Choice Questionnaires
Javier Diaz1, Maria Rifqi1, Bernadette Bouchon-Meunier1, Sandra Jhean-Larose2, and Guy Denhi`ere2
....

About classification and decision trees:

Hidden Markov Models for Longitudinal Comparisons
S L. Scott, GM. James, and CA. Sugar

A Survey of Decision Tree Classifier Methodology
S. Rasoul Safavian and David Landgrebe
.....

About question order:
Carryover Effects in Attitude surveys
R Tourangeau K, A Rasinski, N Bradburn and R d’Andrade

Measuring new type of question-order effects, additive and substractive
D W Moore

Planning:
In the period of 19-01-2009 to 30-07-2009 I will try to follow this planning:

<table>
<thead>
<tr>
<th>Date</th>
<th>Activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>19 jan - 23 feb</td>
<td>Investigate the problem, read literature, get an idea of the background</td>
</tr>
<tr>
<td>23 feb - 16 mar</td>
<td>Write introduction and read about models and backgrounds</td>
</tr>
<tr>
<td>16 mar - 13 apr</td>
<td>Think out, and write the basics of the model</td>
</tr>
<tr>
<td>13 apr - 18 may</td>
<td>research and answer questions raised by the model</td>
</tr>
<tr>
<td>18 may - 30 jul</td>
<td>Combine the research with a design for Hexon</td>
</tr>
<tr>
<td>18 may - 15 jul</td>
<td>Improve the model, and write it down, research how to test the model</td>
</tr>
<tr>
<td>15 jun - 15 jul</td>
<td>Run tests, finish writing, and revise the thesis</td>
</tr>
<tr>
<td>15 jul - 30 jul</td>
<td>Reserved for unexpected delay</td>
</tr>
</tbody>
</table>

Project specifics:

I. the frequency of the meetings with your supervisor
   Every week (where possible) there will be a meeting between me and my supervisor.

II. the way in which your supervisor gives feedback (meetings or e-mail)
    In the weekly meeting the progress is discussed and feedback given

III. the periods during which you will be absent
    No planned absence