A clustering approach to assess real user profiles in spoken dialogue systems

Zoraida Callejas [1], David Griol [2], Klaus-Peter Engelbrecht [3], Ramón López-Cózar [1]

[1] Dept. Languages and Computer Systems, University of Granada, Spain
[2] Dept. Computer Science, University Carlos III of Madrid, Spain
[3] Quality and Usability Lab, Deutsche Telekom Laboratories, TU Berlin, Germany

Introduction

- In order to provide a positive user experience, spoken dialogue systems should adapt to their users.

- Despite of the systems designed for specific population groups, the decision of **which user groups must be considered** is not trivial, and it is not clear how it can be evaluated.

Proposal

We present an approach based on **clustering** to assess whether the user groups considered to implement a system establish meaningful differences in their interaction behaviour.

1. Clustering of a real user corpus:
   - Interaction parameters.
   - Subjective judgements.
2. Are the groups balanced between clusters?

Experimental set-up

- Corpus of 62 dialogues of real users interacting with the **INSPIRE system** to control domestic devices via speech.
- Experiments: Use our proposal to assess the appropriateness of considering 4 user groups which are the combinations of age (senior or young) and self-perceived technical affinity (low or high).
    - 32 dialogues by young, 30 by senior users.
    - 26 dialogues by low, 36 by high technical affinity users.
- Clustering:
    - x-means algorithm (estimates the number of clusters to be used).
    - 1,000 interactions.
    - Euclidean distance between centroids using different metrics.

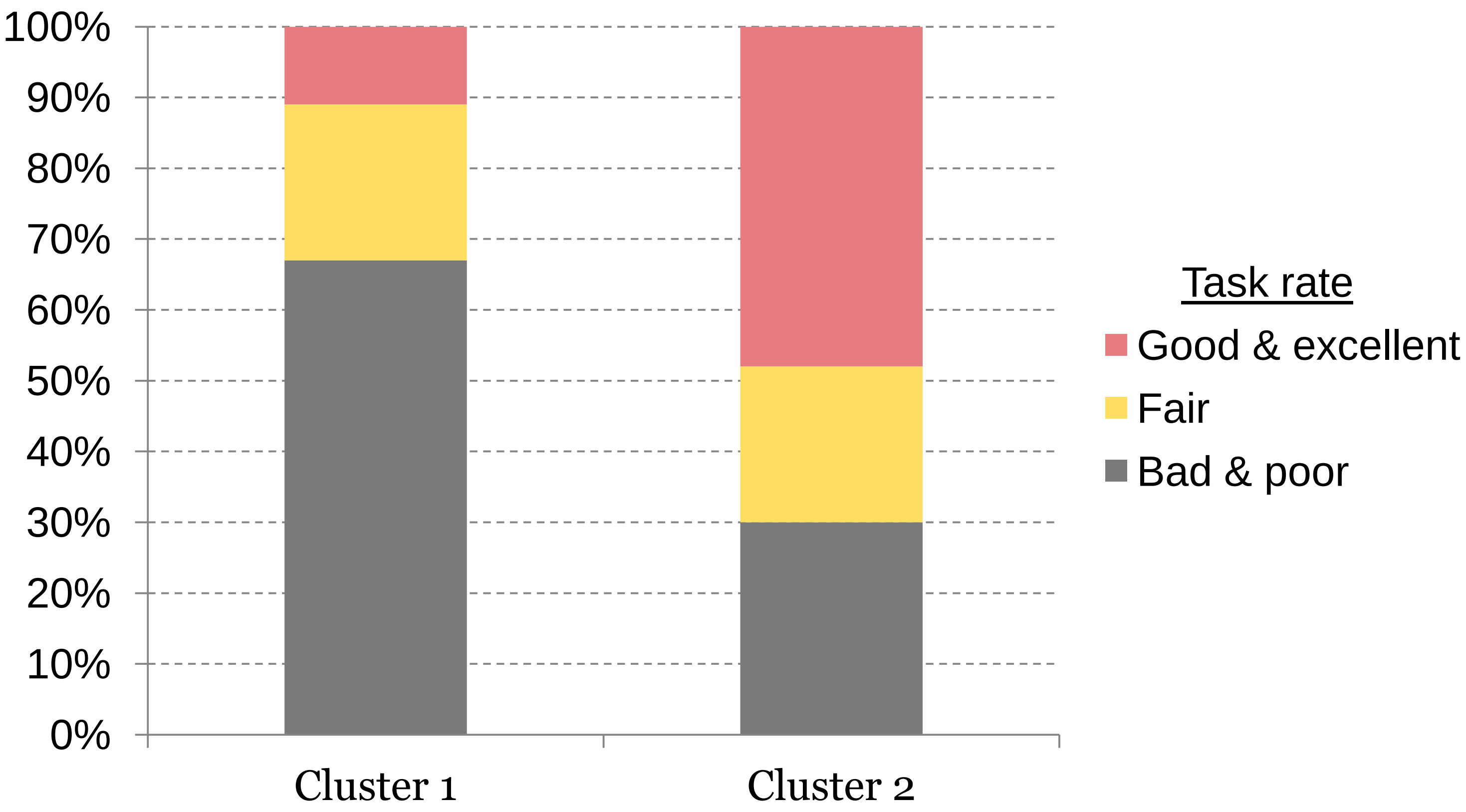
Experimental set-up

Parameters used	
Interaction parameters	User turn duration, system turn duration, number of turns, number of words per user’s utterance, number of words per system’s utterance, number of help requests in the dialogue, task success, concept error rate, number of no matches per dialogue, number of repetitions per dialogue, number of barge-in per dialogue.
User judgements	Task rate, overall impression with the interaction, overall impression of the presented system.
User profile	Technical affinity, age.

Discussion of results

Experiment 1: *Clustering parameters:* interaction parameters.
Parameter studied: overall impression.

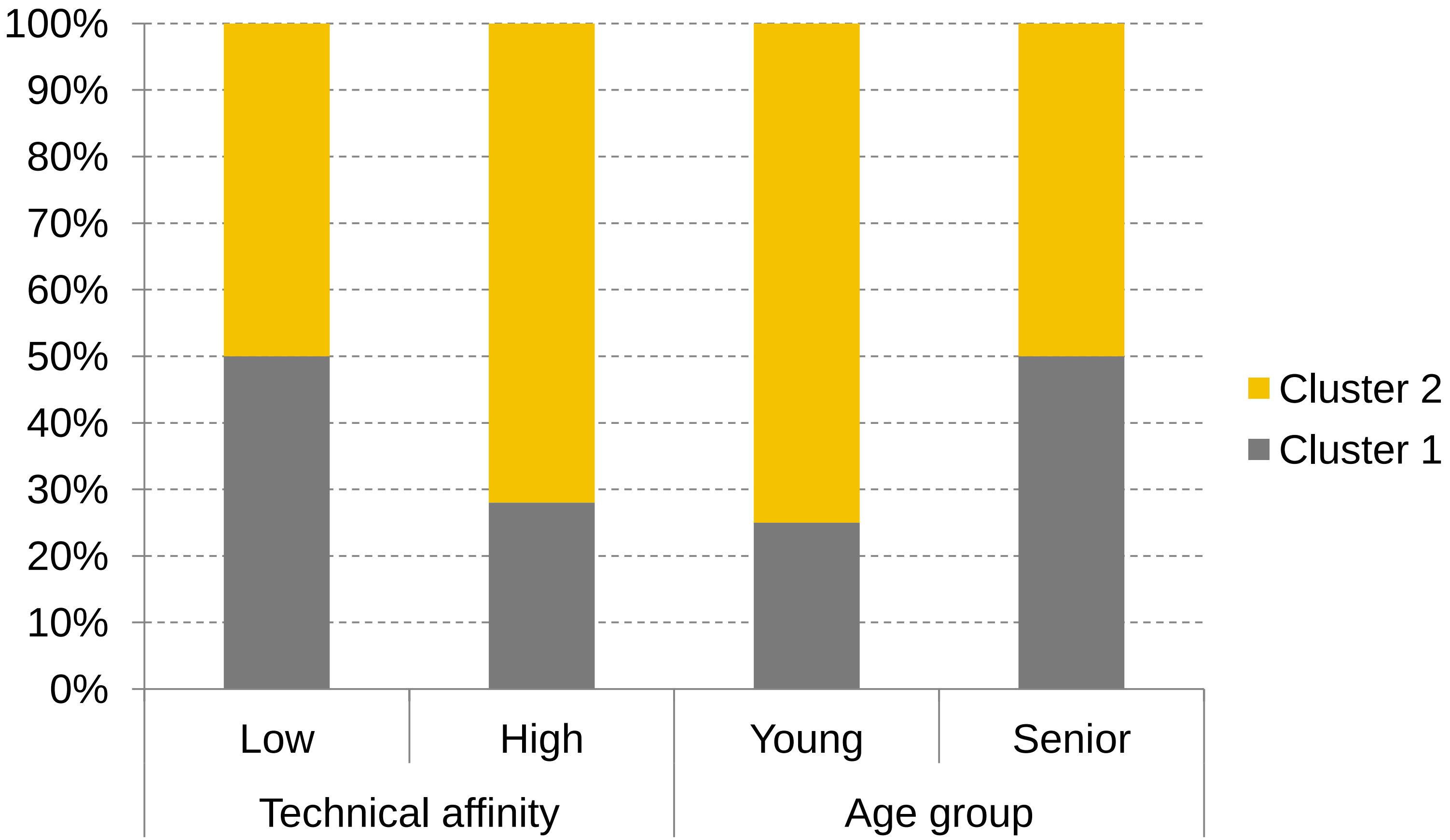
- Interaction parameters did not lead to clusters with distinct overall subjective impressions, with the exception of the judgement of task rate:



Discussion of results

Experiment 2: *Clustering parameters:* interaction parameters.
Parameter studied: user profiles.

- Interaction parameters did not lead to clusters with a clear distinction of user profiles:

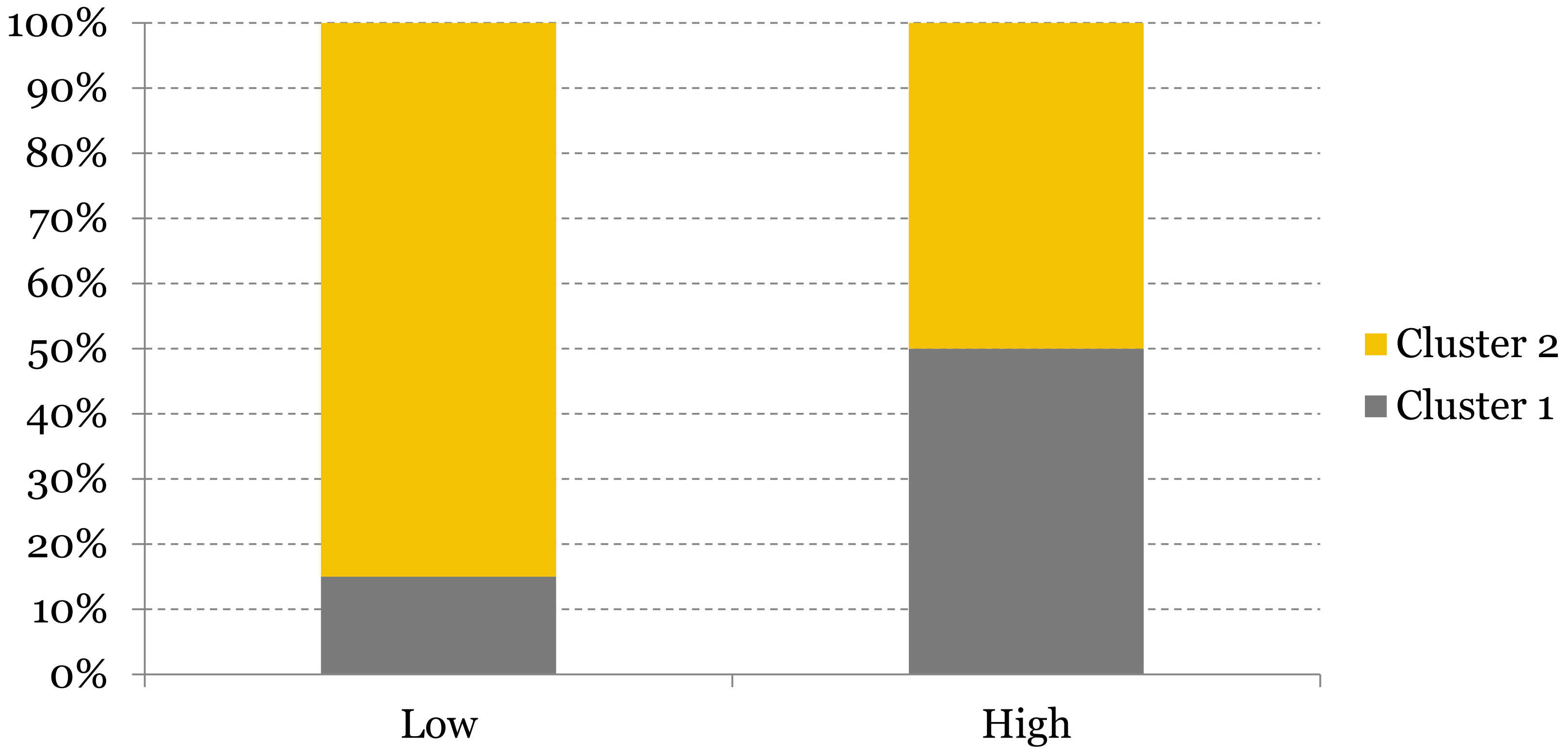


Discussion of results

Experiment 3: *Clustering parameters:* user judgements.
Parameter studied: user profiles.

- The majority of low technical affinity dialogues were classified into the same cluster.

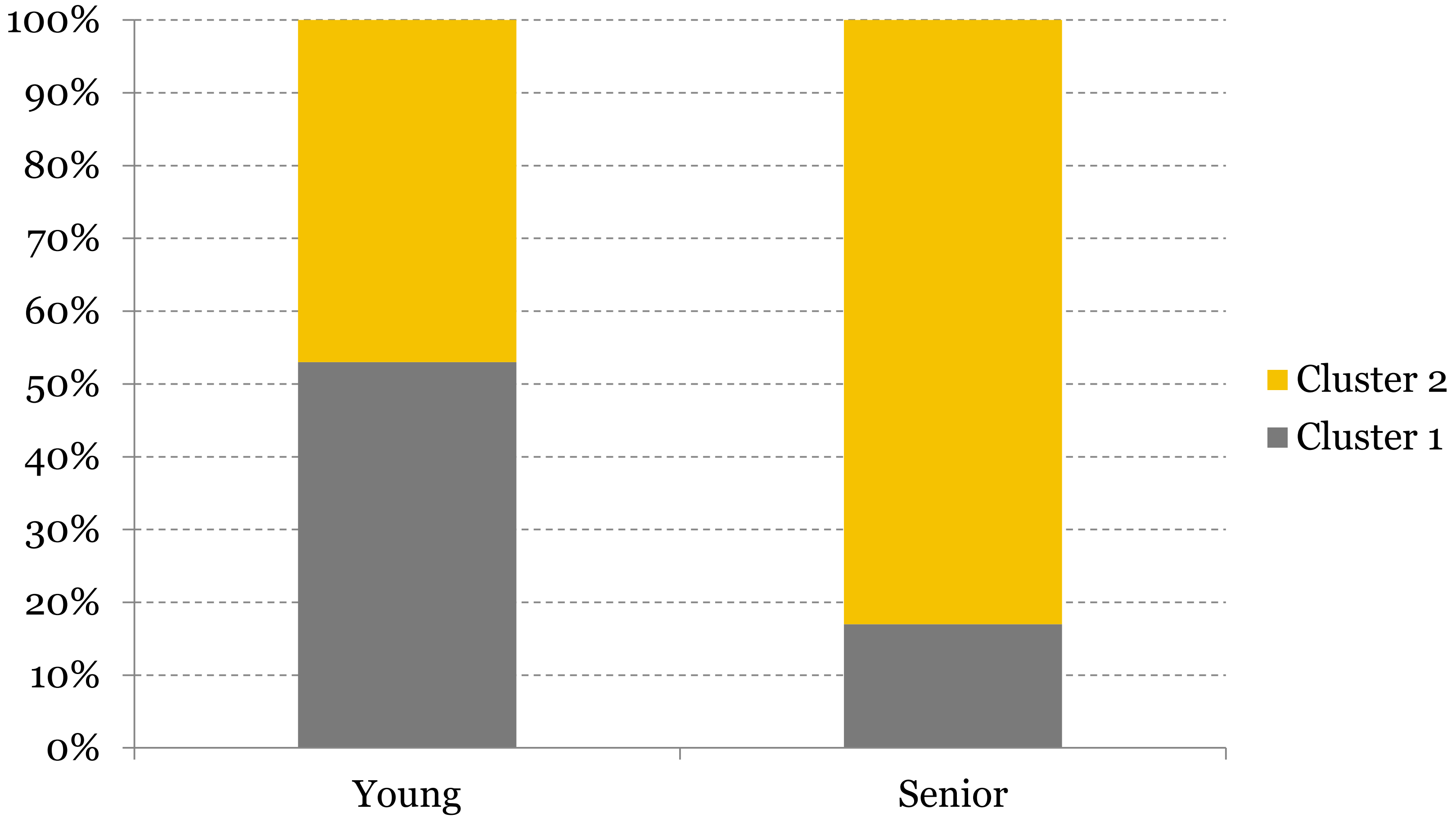
  Users with low affinity systematicallly evaluate the system with worse rates whereas high affinity users provide more varied judgements.



Discussion of results

Experiment 3: *Clustering parameters:* user judgements.
Parameter studied: user profiles.

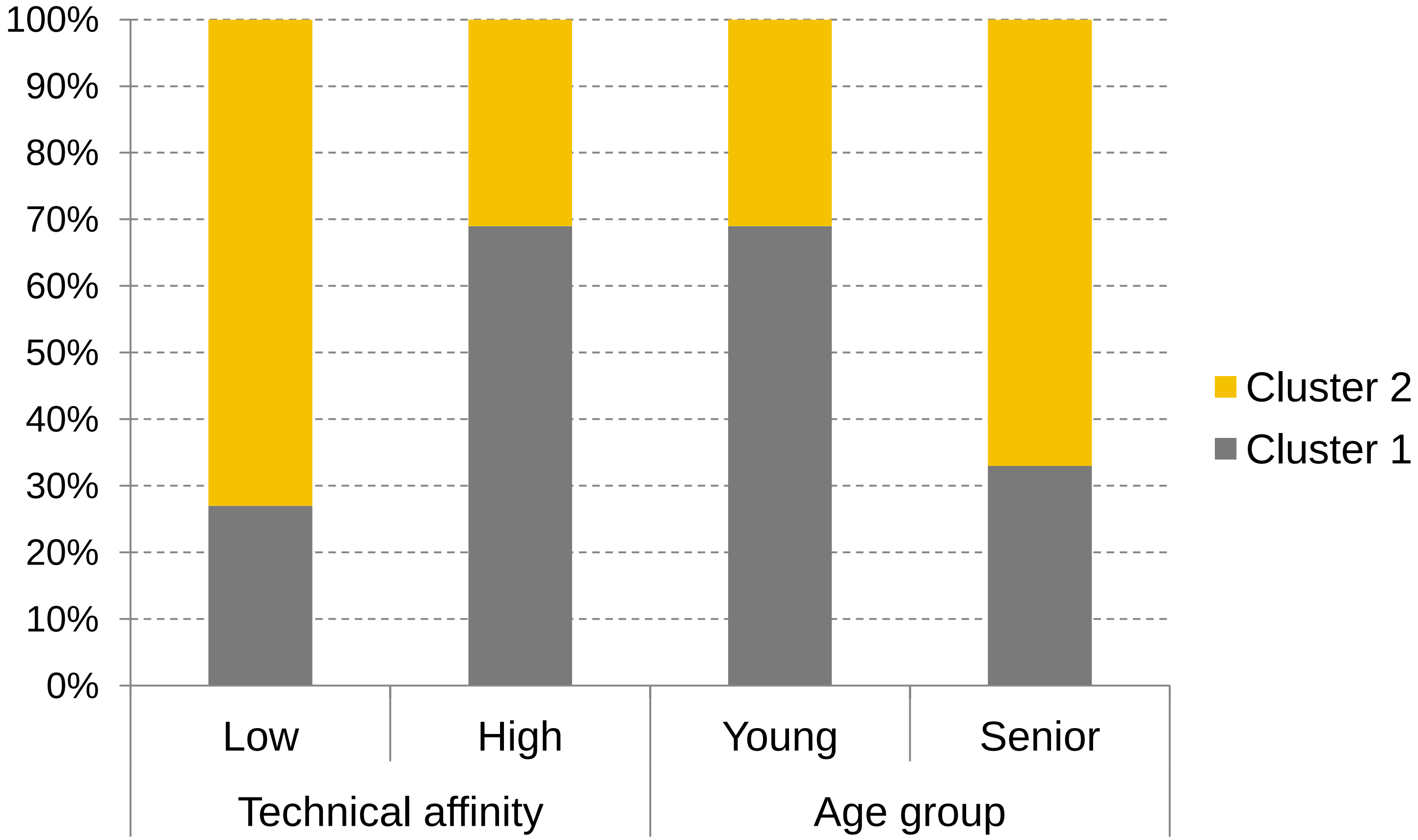
- Subjective features lead to clusters that distinguish senior users in a similar way:



Discussion of results

Experiment 4: *Clustering parameters:* interaction parameters and user judgements.
Parameter studied: user profiles.

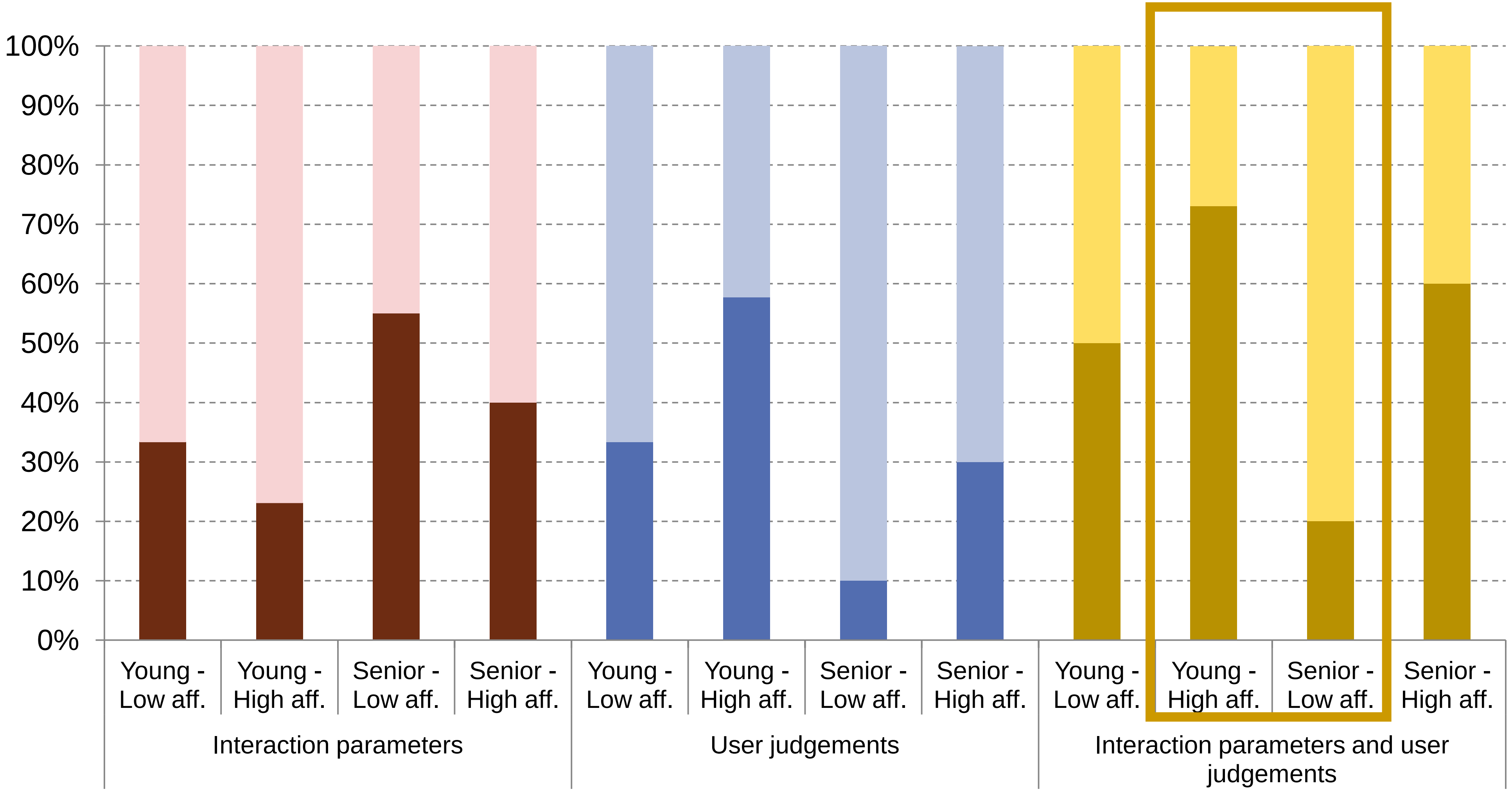
- When both interaction parameters and user judgements were used, the profiles corresponding to high technical affinity and young users were separated better:



Discussion of results

Summary

The real difference strives between young+high technical and senior+low technical profiles:



Conclusions

- The clustering approach provides an efficient way of easily assessing user groupings, which helps to optimize data collection.
- In the case of the INSPIRE system:
  - The profiles of the users elicitated different behaviours when considering 3 groups (young+high affinity, senior + low affinity, remaining), instead of 4 (young+high, young+low, senior+high, senior+low).

Future work

- To assess whether the system adapted to the new groups outperforms:
  - A non-adaptive baseline.
  - A system adaptive to the initial 4 groups.
- To replicate the experiments in other application domains.