User Story Clustering in Agile Development: a Framework and an Empirical Study

Bo YANG, Xiuyin MA, Chunhui WANG, Haoran GUO, Huai LIU, Zhi JIN

Frontiers of Computer Science, DOI: 10.1007/s11704-022-8262-9

Problems & Ideas

- Problems of user story clustering:
  - It is urgently required to have a way for automatically clustering the user stories, which will both help promote the construction of user story mapping and greatly save the energy.
  - Nevertheless, up to our best knowledge, there does not exist a simple, efficient and automatic (or at least semi-automatic) technique for clustering user stories.
- Ideas: We propose a Natural Language Processing based approach, which can Cluster User Stories and identify duplicate ones, termed as CUSNLP. In the approach, the sentence patterns of each user story content are first analysed and determined such that the information in the representative tasks can be extracted based on the user story meta-model. The similarity of user stories is then calculated. The clustering effect is finally obtained after the connected graphs are generated

Main Contributions

- Contributions:
  - We propose a new comprehensive framework for user story clustering. Different from existing methods, the method proposed in this paper performs feature extraction on the verb-object and noun-object structures of user stories. Then, we analyse each user story and extract keywords. The adjacency matrix is obtained through similarity calculation. Then, an undirected graph is constructed. The corresponding clustering result is obtained by searching for the connected components;
  - We improve the structure of the user story meta-model originally proposed by Wautelet et al [9], which incorporates capabilities into the Task. In particular, Wautelet et al. differentiated Hard goal and Soft goal, which is not necessary in user story clustering. Therefore, we merge them into Goal;
  - The proposed CUSNLP approach is implemented and evaluated on 13 projects. The experimental result show that our approach can obtain higher precision and recall than the baseline techniques.