Bachelor project 2018 / 2019

Ask Your Repository! An infrastructure to categorize and retrieve project knowledge by combining voice conversational interfaces over project knowledge-bases enhanced by hybrid crowd-machine learning classifiers

System Analysis and Modeling Research Group of Prof. Dr. Giese

Scenario

When faced with a design problem, engineers make quick decisions about which tasks to perform. Many of these decisions are ad hoc or even look arbitrary. However, what engineers are actually doing is to apply creative thinking to relate the current situation to their tacit knowledge. In this sense, engineers use their experience on similar problems to come up with creative solutions. This might involve looking **at the data from the current project as well as data from previous projects.** These data can consist both of artifacts (e.g., specifications, sketches, emails) and tasks (e.g., changes on artifacts, applied methods).

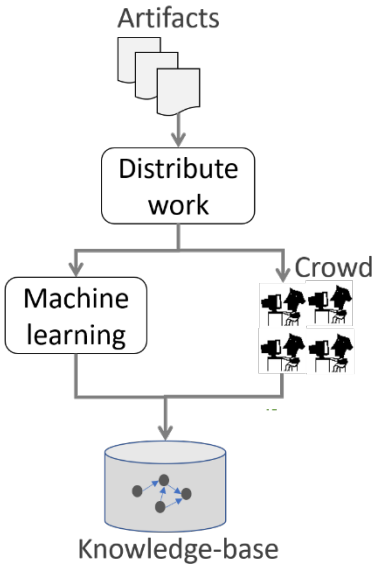

Figure 1 Knowledge-base construction

Challenges

However, retrieving previous experiences and insights depends on one's ability to recall data timely and in enough level of detail. Except for certain types of data that are kept on versioning systems (source code and specifications), most project data are unstructured, e.g., lack categories or dependencies. Hence, this makes it difficult for engineers to accurately and unobtrusively retrieve most data.

Solution

We suggest (1) to **build** a project knowledge-base that can be easily and unobtrusively queried using (2) simple voice commands. For (1) it will be often necessary to make the unstructured data better recognizable. One way is to use **machine learning** to identify categories. Another option is to rely on the intuition of a large group of people (**crowd**) to come up with data categories. To make the knowledge-base available via voice (2), we propose to use an interface that accepts voice commands and translate them into database queries by utilizing a voice recognition APIs.

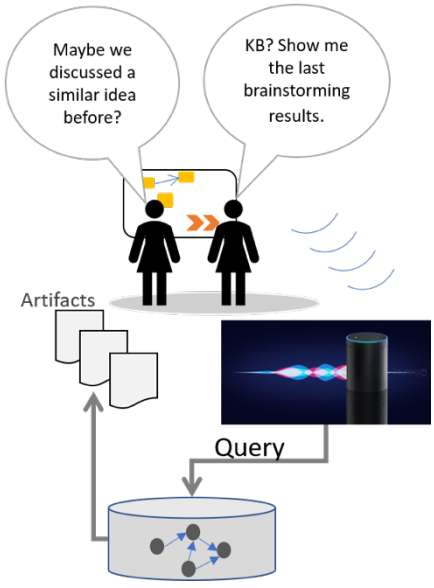

Figure 2 Unobtrusive retrieval of knowledge

Project Objectives

To develop the envisioned solution, we will design and build the following components:

Data enricher consists of two subcomponents responsible to add content to the existing data. One component is a GUI interface to allow people (crowd) to add labels to existing data. The second component scans the knowledge-base to identify datasets that need labeling.

Automatic classifier is a machine learning component that will scale up the categorization work. The crowd will generate a ground truth that will could be used to train a supervised learning model. As an intermediate step, we could use unsupervised learning to suggest categories to the crowd. For audio data, we could also use standard voice-to-text API's to extract categories.

Knowledge-base stores artifacts and tasks from previous projects. We can work both with Open Source Software projects and an existing Design Thinking project repository. The repositories of choice will be transformed into a knowledge-base by means of data enrichment and classification.

Querying by voice will convert simple voice commands into queries in SQL or Cypher format (Graph Database), e.g., "show me the architecture sketch that we talked about last week".

We plan to employ the following **technologies and methods**:

- **Voice recognition** (Google Assistant or Amazon Alexa)
- Voice-to-text translation (Otter AI)
- Graph database or NoSQL database
- **Crowdsourcing services** (Amazon Mechanical Turk)
- **Machine learning** classification methods (e.g., Neural Nets, Decision Trees, SVM)

Main takeaways for students

The project team will apply agile methods which will involve short delivery cycles guided by the concept of MVP (minimum viable product). Students will acquire insights and experiences on:

> New technology (graph and document database, voice recognition)
> New methods (agile, crowdsourcing, machine learning)

Project contacts

This bachelor project is offered by the System Analysis and Modeling chair of Professor Giese. Christian Adriano and Christian Zöllner (room: A-2.27, e-mail, christian.adriano@hpi.de and christian.zollner@hpi.de) will be the main contacts. Christian Adriano will coach the students on the use of machine learning and crowdsourcing methods, which are his areas of expertise. Our external project partner is the D-School represented by Dr. Claudia Nicolai (room: D-1.10, email: claudia.nicolai@hpi.de). The D-School team will be available for project scoping, interviews and feedback-sessions, will provide data-sets and is interested in the use of the outcome.