05/5102 Information extraction and presentation using interactive agent technologies and text mining tools

Type of activity: Medium Study (4 months, 25 KEUR)

Background:

Based on the success of their conversation agent Smarterchild, Conversagent Inc. has recently developed and released a scripting agent design language called Buddyscript. The Buddyscript software development kit enables developers to build conversation agents and launch them in various information exchange / messaging platforms, like AIM, email, http (with a Web UI), MSN, Jabber, Yahoo, SMS. The agents browse pre-determined sources of information on the web, extract relevant pieces of data, store them internally, and present them in a user-friendly format when requested. The source information typically consists of domain related html pages, the structure of which is exactly known. Thus, these sources can be processed relatively easily by the script language. In the present study, our aim is to collect information from less structured, but still raw documents - typically pdf pages containing less unified data tables or product/design descriptions - and provide an interface which could present the extracted and well-organized information either to the final users (by a chatting UI) or pass it to strictly organized databases.

Study Objectives:

- Asses how the interactive chatting agent technologies would work in a more restricted environment (the software and data environment of the Concurrent Design Facility of ESA), on local, but less structured documents.

- Investigate the applicability of the Buddyscript language to extract information from pdf documents based on well identified keywords and/or names of datatable elements.

- Develop a sample agent (to be launched under http) to extract and present information from the test documents provided by ESA. In case the Buddyscript framework alone does not prove enough to extract the required information, pattern matching and/or text mining tools should be applied and be interfaced with the Buddyscript presentation layer.

- Optionally, further develop the small Buddyscript agent (designed internally) for the space domain to support space engineering work. General information related to space science should be collected and implemented in the Buddyscript SDK creating an interactive “assistant” of the engineers.

- Investigate the robustness of the developed agent in terms of extensibility and maintainability.

References:

[1] Buddyscript SDK. https://buddyscript.conversagent.com/

[2] G. Norton, Creating Lotus Instant Messaging interactive agents with the Buddyscript SDK. Lotus
Software. http://www-10.lotus.com/ldd/today.nsf/0/4e6a4aad9b064e7685256dc1006977cd/$FILE/BuddyScript_pt1.pdf

[3] M. Crochemore and W. Rytter, Text Algorithms, Oxford University Press, New York, 1994.