Available online at www.sciencedirect.com

ScienceDirect

IFAC-PapersOnLine 49-12 (2016) 473–478

KTR: an approach that supports Knowledge extraction from design interactions

Rauscher Francois, Matta Nada and Atifi Hassan

Institute ICD/Tech-CICO, University of Technology of Troyes, 12 rue Marie Curie, BP. 2060, 10010 Troyes Cedex, France,

{ francois.rauscher, nada.matta, hassan.atifi}@utt.fr

Abstract: Computer mediated communication is ubiquitous in Software design projects. Email is used for project coordination, but also for design, implementation and test. Especially with currents agile development methods, it is very common to interact through computer meditated communication like email, instant messaging and other collaborative tools in order to express functional needs, notify of issues and take appropriate decisions. In this paper we propose a Knowledge Trace Retrieval (KTR) system. It addresses the problem of retrieving elements of problem solving and design rationale inside business emails from a project. Even if knowledge management tools and practices are well spread in industry, they are rarely used for small projects. Our system aims at helping user retrieve traces of problem solving knowledge in large corpus of email from a past project. The framework and methodology is based on enhanced context (project data, user competencies and profiles), and use machine learning technics and ranking algorithm.

Keywords: **Knowledge Management, Traceability, Problem Solving, e-mails, project memory.**

1 INTRODUCTION

Computer mediated communication is ubiquitous in Software design projects. Email is used for project coordination, but also for design, implementation and test. Especially with currents agile development methods, it is very common to interact through computer meditated communication like email, instant messaging and other collaborative tools in order to express functional needs, notify issues and take appropriate decisions.

The knowledge produced in these types of interactions is often buried inside email boxes, hence being volatile and not easily reused. Project Memory (Matta et al, 2000) aims at describing organizational and cooperative dimensions of knowledge created during the lifecycle of a project. Our objective is to discover if we can structure and extract knowledge from professional emails in order to trace some phases and decisions inside a project memory. The main questions of our research are: Is it possible to extract knowledge from e-mails? Which techniques can be used for that? Are current knowledge engineering techniques adequate for this type of situations?

For instance: a manager asking to an employee: "What was the exact file format adopted in project "a" of the software XYZ and why?" And the only solution for the employee is to explore tens of gigabytes of emails requesting by keywords to try to find useful information. To answer this type of question, we develop a system called KTR that aims at helping users retrieving relevant (according to a query) traces of problem solving knowledge among emails corpus.

Some works have been done on emails related to topics classification or spam detection (Jindal, 2007), it's rarely used in the context of knowledge management. The use of email is often confined to coordination and planning tasks (Wasiak et al, 2011), (Matta et al, 2010) or in case of legal issues. Enron case turned into a rich corpus for research about communication and social networks (Diesner, 2005). We focus in this paper on problem solving analysis e-mails. We use pragmatics analysis and speech act in order to identify corresponding sentence intentions by also linking speech acts to project context and organization.

2 DESIGN PROBLEM SOLVING

Problem solving plays a central role in design projects. For instance, in software design, developer's deal with needs' analysis, specification documents, implementation, debugging and testing. This is especially true if the development follow an agile method, with several round trips from design to delivery.

According to Hardin (Hardin, 2002), "Any problem has at least three components:

- Givens: information and facts presenting context;
- Goal: desired end state;
- Operations: actions to be performed to reach end state;

Software development evolved quickly in the last two decade from classic waterfall model to Extreme Programming and currently to cyclic and iterative methods like Agile (Beck et al, 2001). As a side effect round trips between product-owner (contractor) and product-manager (developer) are more frequent, leading to increased communication and collaborative work. Typically, problem Solving sequences happen on weekly (sometime daily) basis and imply all the actors of the project, not only the development team.

10.1016/j.ifacol.2016.07.658

Otherwise, knowledge can be produced with interaction between actors (Grundstein, 2000). So, we propose in this paper, to analyze mediated interactions through e-mails and to extract problem solving knowledge. We focus first on problem identification.

3 E-MAILS ANALYSIS

Emails is close to written language and have specific characteristics like asynchronous, not face to face, absence of nonverbal signs compared to usual conversation (Baron, 1998), but in a computer mediated communication approach, as stated by Herring (Herring, et al. 2004), it can be considered as a good candidate for using discourse analysis techniques.

A study from Kalia, et al. (2013) on Enron corpus presents a method to identify and track tasks and commitments inside business emails. This was done by pure NLP technics like using n-grams, part of speech tagging and machine learning. Our system enhanced the context with user competencies and roles, project organization and phases. One can also notes the work of Scerri (Scerri, et al. 2010) that propose a system called Semanta to assist users during their daily emails workflow to track actions items like Meeting Request, Task Assignment, and File Request.

Traceability of requirements in software development is usually through the usage of a simple matrix but does not keep the record of the "how, why and who" during the design and implementation process. In (Matta, et al., 2010) linguistic pragmatics was used on discussion forums to identify criteria that help analyzing messages of coordination in design project. Statistical TextMining methods are not well adapted to emails analysis due to its short content. Some work on TextMining are focused on the stylometric features of e-mails in foresnsic investigations (Iqbal et al, 2010), other work identify gender of e-mails authors based on grammatical styles (Corney et al, 2002), or classify e-mails based on lexicon identification (Sakyrai et al, 2005) and pattern recognition (Secker et al, 2003). TextMining approaches need in general an important volume of data (text recognition is based on the calculus of the occurrence of word) and a pretreatment of the text in order to eliminate noises.

Pragmatic analysis is a relevant approach for that but the choice of the speech acts coding scheme is crucial. Identifying knowledge from e-mails can be shown as structured traceability of actors' interactions. Interactions studies show that discourse sense and intentions is related to the context of the interactions. So, we use in our KTR approach a structured traceability based on one side pragmatics analysis and on another side linking to project context and organization in order to identify possible source of knowledge in e-mails corpus.

4 KTR BASIC PRINCIPLES

As we stated in the introduction, the goal of our KTR system is to help the user retrieve "traces of knowledge". We define *Knowledge Traces* (KT), as messages containing meaningful information regarding the team's members having a problem-solving mediated exchange over email.

As a typical use, the user will input a query and the KTR system will present a list of messages from the email corpus matching the query and having a high score of being part of collaborative problem solving between the project's members. The problem can concern different aspect of a project: phase development, product characteristics and quality, coordination, etc. In this paper, retrieving problem about product characteristics is shown.

Retrieval process can be viewed as ranking or classification problem; in this study we consider it as ranking problem. We will compute a score on each message based on the user query and the KT elements. In order to decide if a message contains KT, we will check if:

1. The message is dealing with topics from the project
2. The message thread contains at least a request (problem statement)
3. The messages in the same thread following the initial request contain elements of answer or a decision.
4. The role and competencies of messages senders are related to the development and the use of the product.

4.1 Topics identification

Our approach is to create a keywords dictionary for the main topics of the project. This dictionary can be built from the following sources:

- Project phasing and specifications documents;
- An expert;
- Domain glossary or ontology if available;

As in project memory context, we choose not to rely on statistical NLP clustering like in (Cselle et al, 2007) but to use existing context knowledge. This dictionary is voluntarily simple and has the form of:

Topic1: *keywords1*, *keywords2*... *keywordsn*.

Using this dictionary (Figure 1:) we classify messages into weighted topics vector (same technic is applied to sentences for a fine granularity analysis). In order to do that we use a cosine similarity based algorithm. We compute a Lucene (Gospodnetic, 2004) ranking between our message and each topic. This gives us a topics matrix T where T_{ij} represents the weight of topic j in message i.