An Intelligent Component Retrieval System Using Conversational CBR

Mingyang Gu¹ and Xin Tong²

¹,²Department of Computer and Information Science, Norwegian University of Science and Technology, Sem Sælands vei 7-9, N-7491, Trondheim, Norway
¹mingyang, ²tongxin}@idi.ntnu.no

Abstract

One difficulty in component retrieval comes from users’ incapability to well define their queries. In this paper, we propose a conversational component retrieval model (CCRM) to alleviate this difficulty. In CCRM, a knowledge-intensive conversational case-based reasoning method is adopted to infer potential knowledge from current known knowledge, calculate the context-based semantic similarities between users’ queries and stored components, and prompt users the most discriminative questions to extract more information to refine their queries interactively and incrementally.

1. Introduction

Component retrieval, about how to locate and identify appropriate components for users, is one of the major problems in component reuse. This problem becomes more critical as more reusable components come from on-line component markets instead of from an in-house component library and the number of available components is dramatically increasing [1].

Several retrieval methods have been put forward in the component reuse community [1], such as the free-text-based retrieval method, the pre-enumerated vocabulary method, the signature matching method, the behavior-based retrieval method, and the faceted selection method. Most of them assume users can define their component queries clearly and accurately, and get their desired components based on their well defined queries. However, people often lack clear ideas about what they need while they begin searching for components and usually can not define their query accurately. To address this problem, a component retrieval system should be able to infer as much potential knowledge as possible from users’ provided information, to select the components based on their semantic similarities with the component query instead of only on their syntactic similarities, and to prompt users the discriminative questions to extract more information to refine their queries incrementally.

We propose a conversational component retrieval model which combines the knowledge-intensive case-based reasoning method and the conversational case-based reasoning technology [2] to complete the three tasks identified above: dialogue inferencing, semantic similarity calculating and question ranking [3].

2. The Conversational Component Retrieval Model (CCRM)

In CCRM, both components and general domain knowledge are represented in an integrated knowledge base. The component retrieval process in CCRM is illustrated in Fig. 1. Users’ initial free-text-based component requirements are formalized into a structured component query (this step is not our research focus). A dialogue inferencing module is used to infer the potential knowledge by exploring the component query and general domain knowledge to form a new extended component query. Pre-stored components are sorted according to their semantic similarities with the component query, and the top sorted components are returned. Candidate questions,
which can be used to refine users’ query, are identified from these returned components, and are further ranked to guarantee the most discriminative questions are asked first. Both the returned components and ranked questions are displayed to users. Users either find their desired components which means the component retrieval process is completed, or select a question to answer which is followed by a new round of dialogue inferencing, semantic similarity calculating and question ranking based on the updated component query.

**Knowledge Representation**

In CCRM, a frame-based knowledge representation model, which is a part of the CREEK system, is adopted to represent both component-specific knowledge and general domain knowledge. All the knowledge is represented as concepts, and a concept takes the form of a frame-based structure, which consists of a list of relationships. A relationship is described using an ordered triple \(<C_r, T, C_v>\), in which \(C_r\) is the concept described by this relationship, \(C_v\) is another concept acting as the value of this relationship, and \(T\) designates the relationship type. Viewed as a semantic network, a concept corresponds to a node, and a relationship corresponds to a link between two nodes.

**Dialogue Inferencing**

Dialogue inferencing concerns inferring potential knowledge from the current known knowledge (component query and general domain knowledge), so the knowledge that can be inferred implicitly would not be asked from users. For example, if a user has said that he wants a component running in the Win2K environment, the system should infer that “the desired component runs in the Windows operating system”, instead of still asking “what type of OS does your desired component run on, Windows or Unix?” Otherwise, the retrieval efficiency and users’ confidence to the retrieval system will be dramatically harmed. We identify four types of semantic relations to support dialogue inferencing [3]: concept abstraction, dependency relations, causality relations and correlation relations.

**Semantic Similarity Calculating**

In CCRM, the similarities between the component query and stored components are calculated through two steps [1]. In the first step, similarities are calculated syntactically based on how many relationships they share. In the second step, the similarities calculated in the first step are evaluated using the general domain knowledge. The evaluation task involves explaining away mismatched relationships to be unimportant. According to evaluation results, similarity values are adjusted. For instance, there is one mismatched relationship type “has-color-space” between a component query, which acquires an image processing component to be able to operate images with “XYZ” color space, and a component which can deal with images with “RGB” color space. Based on the general domain knowledge represented in the knowledge base, “RGB convert-to XYZ” and “XYZ convert-to RGB”, we can explain that “since the image using RGB color space can be converted to an image using XYZ color space and vice versa, it is possible to use this component to realize the required task”, and the similarity value generated in the first step can be kept unchanged or increased.

**Question Ranking**

Question ranking is used to guarantee the most discriminative questions are asked first. Question ranking is completed in CCRM through two steps. In the first step, general domain knowledge is explored to identify the questions which are constrained to be asked after other questions, for example, if the answer of question A can be caused by the answer of the question B, so question A is constrained to be asked after question B. Four semantic relations, concept abstraction, dependency relations, causality relations, and practical costs, are identified to support semantic question ranking [3]. In the second step, the questions, which have no semantic constraints to be asked after other questions, will further be ranked according to their statistical metrics (e.g. information gain).

### 3. Future Work

The knowledge-intensive similarity calculating process has been realized in the CREEK system, and the dialogue inferencing and question ranking processes are being added. We are evaluating this model using about 200 different image operating components from the DynamicImager system.

### References:

