Advanced Engineering Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei

A personalized query expansion approach for engineering document retrieval ☆

Gyeong June Hahm [a], Mun Yong Yi [b], Jae Hyun Lee [c], Hyo Won Suh [a,*]

[a] Department of Industrial & System Engineering, KAIST, Dae-jeon 305-701, Republic of Korea
[b] Department of Knowledge Service Engineering, KAIST, Dae-jeon 305-701, Republic of Korea
[c] Department of Industrial & Management Engineering, DAEGU University, Dae-gu 712-714, Republic of Korea

ARTICLE INFO

Article history:
Received 30 May 2013
Received in revised form 7 April 2014
Accepted 17 April 2014
Available online xxxx

Keywords:
Query expansion
Personalized search
Semantic search
Domain ontology
Engineering document

ABSTRACT

Engineers create engineering documents with their own terminologies, and want to search existing engineering documents quickly and accurately during a product development process. Keyword-based search methods have been widely used due to their ease of use, but their search accuracy has been often problematic because of the semantic ambiguity of terminologies in engineering documents and queries. The semantic ambiguity can be alleviated by using a domain ontology. Also, if queries are expanded to incorporate the engineer's personalized information needs, the accuracy of the search result would be improved. Therefore, we propose a framework to search engineering documents with less semantic ambiguity and more focus on each engineer's personalized information needs. The framework includes four processes: (1) developing a domain ontology, (2) indexing engineering documents, (3) learning user profiles, and (4) performing personalized query expansion and retrieval. A domain ontology is developed based on product structure information and engineering documents. Using the domain ontology, terminologies in documents are disambiguated and indexed. Also, a user profile is generated from the domain ontology. By user profile learning, user's interests are captured from the relevant documents. During a personalized query expansion process, the learned user profile is used to reflect user's interests. Simultaneously, user's searching intent, which is implicitly inferred from the user's task context, is also considered. To retrieve relevant documents, an expanded query in which both user's interests and intents are reflected is then matched against the document collection. The experimental results show that the proposed approach can substantially outperform both the keyword-based approach and the existing query expansion method in retrieving engineering documents. Reflecting a user's information needs precisely has been identified to be the most important factor underlying this notable improvement.

1. Introduction

A product development process is a sequence of activities that are generally intellectual and organizational rather than physical [1]. During the product development process, a large amount of knowledge is organized in the form of documents, drawings, reports, and e-mails. Thus, engineers' productivities are heavily dependent on retrieving and using documents. Several studies have found that engineers spend a large amount of time in searching for information [2,3]. To reduce the searching time, existing retrieval approaches, such as keyword-based searching, should be carefully examined for possible improvement.

However, performance degradation of classic information retrieval (IR) models, such as the Boolean model, vector model, or probabilistic model, is an unavoidable problem for engineering documents retrieval. The main reason for the problem is that engineering documents are different from general documents because of their syntax variations and semantic complexities [4]. Syntax variations mainly occur from the usage of abbreviations, acronyms, and synonyms. Semantic complexities result from the specific relationships among the engineering terms as well as polysemy words. Specific relationships generally originate in the design information, such as product structures or design processes. Meanwhile, the queries of engineers also have ambiguity and complexity. A short query with domain specific terms is a major cause of ambiguity. Also, complexity comes from the specific relationships among the

☆ Handled by W.O. O'Brien.

* Corresponding author. Tel.: +82 423503163; fax: +82 423503110.
E-mail addresses: hamstein@kaist.ac.kr (G.J. Hahm), munyi@kaist.ac.kr (M.Y. Yi), paganini9@gmail.com (J.H. Lee), hwsuh@kaist.ac.kr (H.W. Suh).

http://dx.doi.org/10.1016/j.aei.2014.04.002
1474-0346/

queried terms, e.g., '*part-name* AND *its-functions*' or '*part-name* AND *its-performances*'. Thus, classic IR models offer limited functions for retrieving relevant documents because these models calculate the similarity between a document and a query based on exact term matching. As a result, some documents, which do not contain the query terms, may not be returned to the user even though they are semantically relevant to the given query. The question now arises: how should a retrieval approach be designed in the engineering domain to resolve the problem, namely, of performance degradation?

To tackle the performance degradation problem, we offer a semantic search framework in this paper. This framework uses domain ontology to process the contextual meaning of terms for the indexing and retrieval of documents. Thus, ambiguous terms on engineering documents and keywords can be properly disambiguated and matched. Furthermore, advanced techniques, such as query expansion, document classification, or sophisticated ranking algorithm, can contribute to better retrieval performance. In particular, for engineering document retrieval, an ontology based query expansion approach is a promising direction as queries can be effectively expanded using the domain ontology in which the essence of design knowledge is captured. Thus, through query expansion, meaningful terms can be added to an original query to specify an engineer's information needs. Also, a query expansion approach based on personal preferences, called personalized query expansion, should contribute to enhancing retrieval performance by reflecting the user's interests.

There are several related studies on information retrieval in the engineering domain [2,5–10]. However these studies are mainly focused on a browsing method or structured documents retrieval. While a few studies have been concerned with retrieving unstructured documents, they mainly proposed a semantic search framework without incorporating specific advanced techniques such as query expansion or personalization. Also, they suffer from the lack of empirical validation. For those studies that indeed provide empirical validation of a specific technique, their approach was not as complete as we address in this paper. Although Li et al. [2] proposed a naive ontology-based query expansion approach in their framework, they did not consider user's intent during the expansion process. Also, several ontology-based query expansion studies [11–14] for a general domain lack the consideration of user's intent during the expansion process. An engineer's intent depending on their task context could be an important source to improve the quality of expansion results. Furthermore, few studies for the engineering domain have been conducted to propose a personalized query expansion, which is an attempt to provide more precise expansion results based on user's interests. If both user's interests and intent are considered during the query expansion process, it may produce a synergetic improvement of retrieval performance in the engineering domain.

The purpose of this study is to propose a semantic search framework that includes a personalized query expansion approach. While most of query expansion approaches consider a user's interests only, our approach is designed to consider a user's interests and intent at the same time for a personalized service. An engineer's intent depending on their task context could be an important source to improve the quality of expansion results [15]. Fig. 1 provides an illustrative example of the difference between general approaches and the proposed approach. Existing ontology-based query expansion approaches generally use adjacent concepts/individuals for the expansion. However, user preferences certainly exist among relation types as well as adjacent concepts/individuals. In Fig. 1, we assume that the line thickness for each concept and relation represents the degree of user preference. Between C1 and C6, C6 is selected for expansion because the relation R3 has a higher level of user preference. Between C4 and C5, because C5 has a higher level of user preference, C5 is selected for expansion. Our proposed approach considers user interests (preferences of each individual) and intents (preferences of each relation type) at the same time.

In this paper, we analyze unstructured engineering documents of the vehicle air purification system (VAPS) development process and build a domain ontology for the proposed approach. Using this domain ontology, semantic preprocessing is conducted for document indexing. Also, for a personalized query expansion, a learning method for an ontology based user profile is utilized. At runtime, the learned user profile is used on our query expansion process to reflect the user's interests. During this process, the user's intent is also considered. The meaningful source for capturing user's intent is the distinction of information types by identifying the design stages on which the user is working. When the user's working design stages are considered, terms that are close to preferred

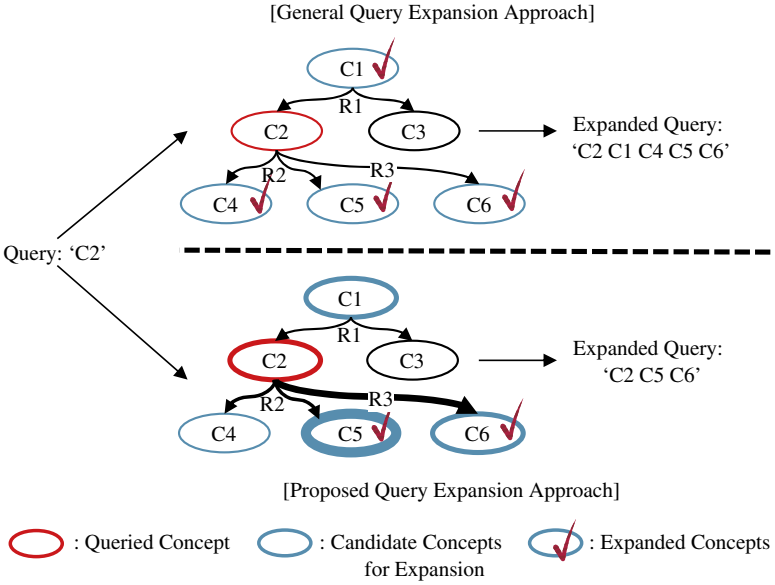

Fig. 1. Illustrative example of the difference among query expansion approaches.

Please cite this article in press as: G.J. Hahm et al., A personalized query expansion approach for engineering document retrieval, Adv. Eng. Informat. (2014), http://dx.doi.org/10.1016/j.aei.2014.04.002