Conceptual Summarization using Ontologies and Nearest Neighborhood Clustering

Elahe Gavagsaz, Mahmoud Naghibzadeh, Mehrdad Jalali
Department of Software Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran
egavagsaz@ms.hdiau.ac.ir, naghibzadeh@um.ac.ir, jalali@ms.hdiau.ac.ir

Abstract - Conceptual summarization aims to provide a database which comprises an abstraction of the entire document content. To effectively provide conceptual summarization, we have presented an approach that is used for conceptual querying. The approach is based on utilizing an ontology for similarity measure between concepts and the nearest neighborhood clustering algorithm for concepts clustering. The results show an improvement in the runtime and tolerant as regards noise.

Keywords-conceptual summarization; ontology; nearest neighborhood clustering

I. INTRODUCTION

We can use ontologies for the organization of concepts, structure and relations within a knowledge domain. Use of ontologies as tools for information access provides a foundation for enhanced, knowledge-based approaches to surveying, indexing and querying of document collections. In this paper we address an approach to conceptual summarization based on instantiated ontology. The main goal is preparing a tool for conceptual summarization of documents when they are used in conceptual querying. We use an ontology that includes the set of concepts, for investigation of the concepts. Conceptual investigation of set of documents can be performed by extracting a set of essential concepts that are the local points of the documents.

Summarization is a process of transforming sets of similar low level objects into more abstract conceptual representations [6], and more specifically, a summary for a set of concepts in the form of a smaller set of concepts. For instance {program, conductor} as summary for {virus, chip, compiler, bus} are or {device} as summary for {printer, monitor, mouse}.

We introduce an approach to conceptual summarization in which ontology plays a key role as reference for the conceptualization. We use nearest neighborhood clustering for concepts clustering and use ontology for similarity measurement between concepts. The semantic grouping that results from the clustering process can then lead to a summary by, for instance, taking a least upper bound of each of the clusters.

The purpose of the ontology in the context and conceptual summarization is to define and relate concepts that may appear in a document collection which may then be used in the summary.

Sources for knowledge base ontologies may have various forms and a taxonomy can be complemented with, for instance, word and term lists as well as dictionaries for definition of vocabularies and for handling of morphology. The well-known resource WordNet [5] is among the more interesting and useful resources for general ontologies.

To establish a general ontology we need an atomic concepts A, semantic relations R, and the set of well-formed terms L that is defined the in following:

\[ L = \{ A \} \cup \{ x [r_1 : y_1], \ldots, r_n : y_n] | x \in A, r_i \in R, y_i \in L \} \]

and taxonomy T over the set of atomic concepts A, we can use an inclusion relation “\( \leq \)” over all well-formed terms of the language L [3],[4], [8]. Therefore, the general ontology \( O = (L, \leq, R) \) encompasses a set of well-formed expressions L derived in the concept language from a set of atomic concepts A, an inclusion relation generalized from the taxonomy relation in T, and a set of semantic relations R.

To form an instantiated ontology, we assume a general ontology \( O = (L, \leq, R) \) and a set of concepts C. The instantiated ontology \( O_C = (L_C, \leq_C, R) \) is a restriction of O to cover only the concepts in C and corresponds to “upper expansion” \( L_C \) of C in O

\[ L_C = C \cup \{ x | y \in C, x \in L, y \leq x \} \]

Conceptual querying concerns retrieval of concepts appearing in an instantiated ontology - thus at a conceptual level to investigate the concepts appearing, or the content of the documents holding these concepts. We will use the knowledge captured in instantiated ontologies for deriving conceptual summarizations, in principle, any collection of text such as a single document, as set of documents. Thus, given a set of concepts C, we want to move towards a smaller set of representative concepts covering C, that is, towards an appropriate summary that includes what’s most characteristic about C. Obviously, this is solely dependent on the structure of the instantiated ontology in use, but since these normally are structured as hierarchies, the use of more general concepts as a cover of subsumed concepts in C will lead to a summary with fewer elements.