Chapter I

Potential Cases, Database Types, and Selection Methodologies for Searching Distributed Text Databases

Hui Yang, University of Wollongong, Australia
Minjie Zhang, University of Wollongong, Australia

ABSTRACT

The rapid proliferation of online textual databases on the Internet has made it difficult to effectively and efficiently search desired information for the users. Often, the task of locating the most relevant databases with respect to a given user query is hindered by the heterogeneities among the underlying local textual databases. In this chapter, we first identify various potential selection cases in distributed textual databases (DTDs) and classify the types of DTDs. Based on these results, the relationships between selection cases and types of DTDs are recognized and necessary constraints of database selection methods in different cases are given which can be used to develop a more effective and suitable selection algorithm.

This chapter appears in the book, Intelligent Agents for Data Mining and Information Retrieval, edited by Masoud Mohammadian. Copyright © 2004, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.
INTRODUCTION

As online databases on the Internet have rapidly proliferated in recent years, the problem of helping ordinary users find desired information in such an environment also continues to escalate. In particular, it is likely that the information needed by a user is scattered in a vast number of databases. Considering search effectiveness and the cost of searching, a convenient and efficient approach is to optimally select a subset of databases which are most likely to provide the useful results with respect to the user query.

A substantial body of research work has looked at database selection by using mainly quantitative statistics information (e.g., the number of documents containing the query term) to compute a ranking score which reflects the relative usefulness of each database (see Callan, Lu, & Croft, 1995; Gravano & Garcia-Molina, 1995; Yuwono & Lee, 1997), or by using detail qualitative statistics information, which attempts to characterize the usefulness of the databases (see Lam & Yu, 1982; Yu, Luk & Siu, 1978).

Obviously, database selection algorithms do not interact directly with the databases that they rank. Instead, the algorithms interact with a representative which indicates approximately the content of the database. In order for appropriate databases to be identified, each database maintains its own representative. The representative supports the efficient evaluation of user queries against large-scale text databases.

Since different databases have different ways of representing their documents, computing their term weights and frequency, and implementing their keyword indexes, the database representatives that can be provided by them could be very different. The diversity of the database representatives is often the primary source of difficulty in developing an effective database selection algorithm.

Because database representation is perhaps the most essential element of database selection, understanding various aspects of databases is necessary to developing a reasonable selection algorithm. In this chapter, we identify the potential cases of database selection in a distributed text database environment; we also classify the types of distributed text databases (DTDs). Necessary constraints of selection algorithms in different database selection cases are also given in the chapter, based on the analysis of database content, which can be used as the useful criteria for constructing an effective selection algorithm (Zhang & Zhang, 1999).
Related Content

Extensions of Web Browsers useful to Knowledge Workers
www.irma-international.org/chapter/extensions-web-browsers-useful-knowledge/64428/

A New Method for Building Probabilistic Ontology (Prob-Ont)
www.irma-international.org/chapter/a-new-method-for-building-probabilistic-ontology-prob-ont/198606/

www.irma-international.org/chapter/bates-berrypicking-model-1989-2002-2005/127126/

On the Design and Implementation of Interactive XML Applications
www.irma-international.org/chapter/design-implementation-interactive-xml-applications/75898/

Fragmentation in Distributed Database Design Based on Ant Colony Optimization Technique
www.irma-international.org/article/fragmentation-in-distributed-database-design-based-on-ant-colony-optimization-technique/222766/