BDD-Based Combinatorial Keyword Query Processing under a Taxonomy Model

Shin-ichi Minato, Hokkaido University, Sapporo, Japan
Nicolas Spyratos, Université de Paris-Sud, Orsay, France

ABSTRACT

Digital libraries are one of the key systems for an IT society, and supporting easy access to them is an important technical issue between a human and an intelligent system. Here the authors consider a publish/subscribe system for digital libraries which continuously evaluates queries over a large repository containing document descriptions. The subscriptions, the query expressions and the document descriptions, all rely on a taxonomy that is a hierarchically organized set of keywords, or terms. The digital library supports insertion, update and removal of a document. Each of these operations is seen as an event that must be notified only to those users whose subscriptions match the document’s description. In this chapter, the authors present a novel method of processing such keyword queries. Their method is based on Binary Decision Diagram (BDD), an efficient data structure for manipulating large-scale Boolean functions. The authors compile the given keyword queries into a BDD under a taxonomy model. The number of possible keyword sets can be exponentially large, but the compiled BDD gives a compact representation, and enabling a highly efficient matching process. In addition, the authors’ method can deal with any Boolean combination of keywords from the taxonomy, while the previous result considered only a conjunctive keyword set. In this chapter, they describe the basic idea of their new method, and then the authors show their preliminary experimental result applying to a document set with large-scale keyword domain under a real-life taxonomy structure.

Keywords: Binary Decision Diagram (BDD), Digital Library, Filtering, Keyword Query, Taxonomy Model

INTRODUCTION

Digital libraries are one of the key systems toward an IT society, and supporting easy access to them is an important technical issue between a human and an intelligent system. The publish/subscribe interaction paradigm provides subscribers with the ability to express their interest in classes of events generated by publishers. A system that supports this paradigm must be able to find, for each incoming event e, the subscriptions that match e, in order to determine which subscribers should be notified. Many typical web applications can be seen as variants of this general framework, including auction sites, on-line rental offices, virtual bookshops, etc. They act as brokers which store only descriptions of the published items.

We now consider a publish/subscribe system for digital libraries which continuously evaluates queries over a large repository.
containing document descriptions (Frej, 2006). The subscriptions, the query expressions and the document descriptions, all rely on a taxonomy that is a hierarchically organized set of keywords, or terms. The digital library supports insertion, update and removal of a document. Each of these operations is seen as an event that must be notified only to those users whose subscriptions match the document’s description.

In this chapter, we present a novel method of processing such keyword queries. Our method is based on Binary Decision Diagram (BDD) (Bryant, 1986), an efficient data structure for manipulating large-scale Boolean functions. We compile the given keyword queries into a BDD under a taxonomy model. The number of possible keyword sets can be exponentially large, but the compiled BDD gives a compact representation, and enabling a highly efficient matching process than using a naïve representation. In addition, our method can deal with any Boolean combination of keywords from the taxonomy, while the previous result (Frej, 2006) considered only a conjunctive keyword set. In this chapter, we will show the results of an experiment in which we applied our method to document-tagging keywords from the ACM Computing Classification, a taxonomy list with over 1,000 entries. We will also mention our future work to consider more advanced keyword expansion.

**PRELIMINARIES**

Here we describe the basic framework of subscription-notification service which we consider in this paper.

**Subscription-Notification Service in Digital Library**

Figure 1 shows the basic processing flow of our system. A document is represented in the digital library repository by a description of its content together with an identifier (say, the document’s URI) allowing to access the document’s content. A number of documents are incoming to the digital library every day, and the identifier list of recent new documents is notified to each user periodically (daily, weekly, or monthly). Each user can specify a subscription to show the composed keywords that are of interest to that user. Every document has a description with a set of keywords, and if one of the keywords matches to some users’ subscriptions, the identifier of the document is added to the notification list. Here, we assume that a user’s subscription is not so frequently updated as the notification periods.

If we consider a nation-wide digital library system, the number of documents and the number of users become quite huge, and it is an important problem to reduce the computation cost for matching (or filtering) of the documents to each user.

*Figure 1. Basic processing flow of subscription-notification service*
BHA2: Bio-inspired Algorithm and Automatic Summarisation for Detecting Different Types of Plagiarism
[www.igi-global.com/article/bha2/169887?camid=4v1a](www.igi-global.com/article/bha2/169887?camid=4v1a)

Weighted Average-Based Multi-Objective Optimization of Tube Spinning Process using Non-Traditional Optimization Techniques