Integration of database management system and document retrieval system: An overview

Xin Lu
School of Library and Information Science
University of Western Ontario
London, Ontario N6G 1H1

Starting in 1977, the study of integrating database management systems (DBMS) and information retrieval systems (IRS) has kept attracting the attention of information scientists. The feasibility of integrating is definite from the theoretical point of view. But the picture is still vague in terms of the practical point of view. It is thought, therefore that before the investigation shifts from abstract models proposing to real IRS designing and testing, based on those suggested models, a thorough review and a comprehensive bibliography are indispensible so that the outcomes and problems could be could have the clarified starting pints.

The activities of integrating aim to incorporate those most significant features of DBMS such as simplicity, data independence, multiple views of data, and reduced data redundancy, and the same time to retain those powerful capabilities that the current IRS has (i.e. inverted file system). Another two more important objectives are of increasing the productivity of information processing and of individualising information searching. Still, the research on this topic has practical implication to library automation, especially to some systems like online public access catalog (OPAC).

The approaches can be identified: building an IRS on the top of a DBMS, and designing a hybrid IRS by applying the idea of the data models of DBMS. It is interesting to notice that both approaches tend to favour relational data model since this model, theoretically, not only has a simple and natural view by also has a solid mathematical basis. Along this line, the first approach, constructing an IRS in a DBMS environment does have a remarkable progress. The second approach, on the other hand, seems not very promising. Probably it is because that merely a few people are working on it. Finally, there exists so far only one reported evaluation. It compared the efficiency of a DBMS based IRS and a conventional IRS and concluded that the two systems are approximately equal to each other.

There are two most serious issues currently. The investigators are in a dilemma that normalization into first normal form forces the breaking up of a tuple with nonatomic entries onto a number of tuples with atomic entries and thus losing the direct association between these values. By contrast, when this situation is rectified by employing the non-first normal form relations, several great features of relational data model disappear immediately, which are just intended to be incorporated in integrated IRS.

Closely associated with normalization, the second issue is the end-user interface, or mostly retrieval language. It si out question of user-friendly as long as the language itself requires general users to have knowledge about the logical structure of data, for example, SEQUEL. IF it is the case, the data physical independence is reached but never data logical independence. To sum up, these issues are basic ones in examining and comparing different models.

In addition, difficulties could be expected in the near future when starting designing and testing prototypes of the integrated IRS, such as response time, imperfect relational operators, and combination with other retrieval techniques or algorithms.