Future Directions in Textual Information Systems

Ian A. Macleod
Department of Computing and Information Science
Queen's University
Kingston
Ontario K7L 3N6

The premise of this work is that it is essential to develop a new generation of information retrieval (IR) systems in order to produce any significant practical innovations in document retrieval. Modern (using the term loosely) information retrieval systems are basically tools. They are not complete information systems but rather they provide a specific service. The effectiveness of these types of retrieval systems is measured in terms of recall and precision. Concern with improving these measurements has become almost obsessive. It is not at all obvious that the effectiveness of retrieval systems can be significantly improved using traditional statistical techniques. At the same time it should be noted that enhancements which have been shown to improve effectiveness have not gained widespread exposure through commercial implementations. This may be because these enhancements are thought to be of limited significance or, more likely, because these commercial systems are not particularly flexible.

The other major information system is the database management system (DBMS). In contrast to IR systems, DBMS are much more flexible. They typically have languages in which queries can be composed. Relationships among different data structures can be established dynamically. Equally importantly, the capabilities of a DBMS can be embedded in a host language so that specialised procedures can be built. The integration of DBMS and IR systems has been a topic of research for some time. However, current DBMS are inappropriate as a vehicle for IR implementations. They are record oriented and cannot conveniently represent real world objects such as documents. STAIRS and similar systems may predefine the structure of documents but for many applications this pre-defined structure is much preferable to the alternative of representing a document in a relational database, for example. The limitations of the record based approach are apparent in many DBMS applications and this has led to the development of what are often called *conceptual models*. The power of the conceptual models derives to a large extent from the range of data abstractions provided. These are intended to facilitate the design of structures which closely relate to real world objects.

There have been a number of important document related activities which suggest the need for a new model for text. ISO standards for document description have been recently developed. These standards view documents as hierarchical objects and it is likely that languages such as SGML will become widely used in the near future for document markup. As structured documents become available, so there will be a need to evolve tools to take advantage of structural knowledge. A separate development has been the renewed interest in *hypertext*. A hypertext structure (or *hyperdocument)* basically consists of a set of linkages to parts of other documents. Again tools are needed to get at the structure of documents in order to build hyperdocuments. The goal of the work described here is to develop such tools. A conceptual model for bibliographic data has been designed. The model is known as Quoits (Queen's Object Influenced Text System). It supports structured documents and provides a query language to retrieve and link information contained in these structures. It is eventually hoped to integrate these capabilities into an existing language, possibly Nial, to provide an extremely powerful facility for document related operations.

Orientations futures des systèmes d'information de texte

Ian A. Macleod
Department of Computing and Information Science
Queen's University
Kingston, ON

On soutient dans cet exposé qu'il est devenu indispensable, si l'on veut voir apparaître de véritables innovations dans le domaine de la recherche documentaire, de mettre au point une nouvelle génération de systèmes de recherche. Les systèmes modernes (ce mot étant employé de façon assez libre) d'information documentaire sont avant tout des outils. Ils fournissent un service précis mais ne sont pas des systèmes d'information complets. L'efficacité de ce type de système de recherche se mesure en termes de rappel et de précision, et le souci d'améliorer ces mesures est devenu presque obsessionnel.. Il n'est pas du tout certain cependant que l'efficacité des systèmes de recherche documentaire puisse être améliorée de manière significative par l'utilisation de techniques statistiques traditionnelles. Les améliorations dont l'efficacité a été démontrée n'ont pas non plus connu de développement commercial significatif. Ceci est peut-être dû au fait que ces améliorations ont été jugées de peu d'importance ou, plus probablement, que les systèmes commerciaux qui les ont incorporées ne sont pas très flexibles.

Un second type de système d'information d'importance est le système de gestion de bases de données (SGBD). Les SGBD sont beaucoup plus flexibles que les systèmes d'information documentaire. Ils possèdent habituellement un language d'interrogation de la base de données et permettent la création de relations dynamiques entre plusieurs structures de données. La possibilité d'intégrer les capacités d'un SGBD à un language d'accueil est tout aussi importante, car elle permet le développement de procédures spéciales. L'intégration des SGBD et des systèmes de recherche documentaire fait déjà depuis quelque temps l'objet de recherches. Cependant, les SGBD actuellement disponibles ne sont pas appropriés pour la réalisation de systèmes d'information documentaire. Ils sont conçus en fonction de notices et peuvent difficilement décrire des objets réels tels, par exemple, des documents. Des systèmes tel *STAIRS* définissent à l'avance la structure d'un document. Souvent, ce type de structure défini à l'avance est de beaucoup préférable à l'alternative que constitue, par exemple, la représentation du même document dans une base de données relationnelle. Les limites imposées par la méthode basée sur le concept de notices sont manifestes dans plusieurs applications de SGBD et ont entraîné le développement des modèles dits conceptuels. La puissance des modèles conceptuels provient en grande partie de la variété d'abstractions de données qu'ils permettent. Ces abstractions sont destinées à simplifier la conception de structures étroitement apparentées à des objets réels.

Plusieurs activités d'importances en recherche documentaire ont permis de souligner le besoin d'un nouveau modèle conçu en fonction du texte. On a assisté récemment à l'adoptions de standards ISO pour la description documentaire. Ces standards définissent un document comme objets dotés d'une structure hiérarchique et il est à prévoir que l'utilisation de langages tels SGML pour l'identification des documents se généralisera bientôt. Le besoin d'outils permettant d'exploiter la connaissance structurelle se développera à mesure que des documents structurés deviendront disponibles. L'intérêt accru porté au concept d'*hypertexte* constitue un développement à part. Une structure hypertexte (ou *hyper-document*) consiste essentiellement en une succession de liens établis entre des sections de documents. De nouveau, des outils seront nécessaires pour permettre de décrire la structure des documents et de construire des hyper-documents. Les travaux décrits dans ce rapport ont pour but de développer de tels outils. Un modèle conceptuel de données bibliographiques, connu sous le nom de Quoits (Queen's Object Influenced Text System) a été mis au point. Ce modèle utilise des documents structurés et offre un langage d'interrogation qui permet de retracer et de relier l'information contenue dans ces structures. On espère, à terme, pouvoir intégrer ces capacités à un langage existant, peut-être Nial, ce qui permettrait d'offrir une installation très performante pour la recherche documentaire.