Text Retrieval and the Relational Model

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

In this article, Macleod examines the suitability of the relational model in the context of storing and retrieving documents. The relational model is compared with other traditional models and their strengths and weaknesses are compared. Through out the article, Macleod largely compares the relational model against the text model. He states that ‘the terminology text model is used to refer to inverted file base systems.’ (Macleod)

Using the text model, documents are retrieved by specifying the combination of terms that are to be found within them and retrieval techniques can easily be employed. Documents are indexed by their attributes (author, published date, etc) or by the words within the content which are listed. The text model is better suited to the straightforward storage and retrieval of stand alone textual databases, for example, a Library Catalogue.

The major characteristic of the relational model is that all data must be stored in tables and each entry in the table must represent a single atomic piece of data. The process of transforming data into a tabular form is called normalization. The relational model provides the capability to retrieve information from a number of sources and also allows relationships to be established dynamically between arbitrary pieces of data. Retrieval operations are performed using SQL.

Macleod wraps up the article by reiterating the fact that both models were developed in response to quite different needs and that neither is better than the other. He also goes on to discuss other applications that have been developed due to the insufficiencies that both the textual model and the relational model have involving text storage and retrieval. ‘This has led to the development of a new class of database models, variously known as the conceptual models, semantic models and object oriented models.’ (Macleod)

In concurrence with Macleod, with the development of new applications, it is apparent that both the text model and the relational model have their strengths and weaknesses, and are used for different purpose and neither the relational model or the text model is necessarily better than the other. There were also certain issues that Macleod did not talk about that effect both the text and retrieval model, or storage and retrieval in general.

Macleod largely talks about the storage and retrieval of documents however, does not touch on the topic of storage and retrieval of images. In most cases, documents are fully converted to electronic representation which can be indexed automatically. However, there are many factors such as high costs, low document quality, and the problem with non-text components being unable to be adequately presented in converted form that may prohibit its complete conversion.
Word processing applications allow us to create and edit electronic documents with ease, compress documents’ electronic representation which allows efficient storage and transmission, and the text model focus’ on the processing, indexing and retrieval of text from large databases.

For documents that have always been in electronic form, these tasks are relatively straightforward and therefore, are easily searched entities. Unfortunately, for those documents which were created manually, systems must deal with scanned images of the hard copy.

In attempts to achieve automatic indexing, the documents must be fully converted to electronic form. Although text can be easily transformed into electronic form, low document image quality can inhibit accurate symbol recognition (OCR) and make it difficult to preserve stylistic features such as font and layout. Furthermore, it may not be possible to provide a suitable representation for certain scanned graphic and picture regions, other than as a digital image.

The main issue then becomes the need for robust ways to access or ‘index’ the information these images contain. With electronic text documents, the index information can be extracted directly from the text, but we do not have that luxury with document images. The ‘content’ of the document images is not directly available since the internal representation is simply a set of pixels, so it is difficult to perform automatic indexing. For some archival applications a simple document ID, such as a case number, may be sufficient for indexing, and can be provided manually.

‘Indexing images may be done using, texture, key word spotting, characterisation, logo recognition and indexing image captions.’ (Doermann)

A major issue concerning both the text model and retrieval is the issue of lexical ambiguity. An issue that Macleod does not touch on, that affects nearly all methods of information retrieval. Lexical ambiguity is an enveloping problem in natural language processing.

There are two types of lexical ambiguity: syntactic and semantic. Syntactic ambiguity refers to differences in syntactic category (e.g., love can occur as either a noun or verb). Semantic ambiguity refers to differences in meaning, and is further broken down into homonymy or polysemy, depending on whether or not the meanings are related. ‘The bark of a dog versus the bark of a tree’, is an example of homonymy. An example of polysemy where the words can have related meanings, ‘he will review the review this weekend.’

‘The goal of an information retrieval system is to locate relevant documents in response to a user’s query. Documents are typically retrieved in a ranked list where the ranking is based on estimations of relevance.’ (Croft and Krovetz)

Many retrieval systems represent documents and queries by the words they contain, and base the comparison on the number of words they have in common. The more words the query and document have in common, the higher the document is ranked; this is referred to as a ‘coordination match.'
There are two problems with using words to represent the content of documents. The first problem is that words are ambiguous and this ambiguity can cause documents to be retrieved that are not relevant. One response to this problem is to use phrases to reduce ambiguity by specifying more accurate search terms. For example ‘computer chair instead of chair’. However, it is not always possible for the user to provide phrases in which the word occurs only with the desired sense. This would also place burden on the user to think of relevant search terms.

The second problem is that a document can be relevant even though it does not use the same words as those that are provided in the query. The user is generally not interested in retrieving documents with exactly the same words, but with the concepts that those words represent. Retrieval systems address this problem by expanding the query words using related words from a thesaurus. Macleod talks about how the text model useful in this sense. For example, the word “term” could be synonymous with “word” (as in a vocabulary term).

In conclusion, I agree with Macleod in the sense that both the text model and the relation model were developed for different purposes and therefore, did not provide a stance on which he though was better. He only states that both have their weaknesses and their strengths. However, he did not talk about a lot of other issues in relation to data storage and retrieval, which I have talked about briefly.
Bibliography:

