

**METHOD AND SYSTEM FOR
COMPRESSING PUBLICATION
DOCUMENTS IN A COMPUTER SYSTEM BY
SELECTIVELY ELIMINATING
REDUNDANCY FROM A HIERARCHY OF
CONSTITUENT DATA STRUCTURES**

TECHNICAL FIELD

The invention relates generally to the field of computer document management, and, more specifically, to the field of computer document compression.

BACKGROUND OF THE INVENTION

Users of computer systems utilize different kinds of application programs ("applications") to perform various kinds of tasks. One such kind of application is a publication application, which may be utilized by users to produce professional-quality printed publications. The users of a typical application can produce publications containing rich text, graphics, extensive formatting, and footnotes.

Publication applications typically store the information relating to a publication in a file or file system object called a publication document. Because of the extensive content that publications may contain described above, publication documents often grow quite large, consuming significant quantities of storage resources and requiring a substantial amount of time and/or significant bandwidth to copy or transmit. This effect is compounded when several publication documents are stored together, e.g., when shipping sample publication documents to the user of a publication application. An effective way to reduce the size of publication documents without significantly reducing the extensiveness of that content would therefore be desirable.

SUMMARY OF THE INVENTION

The present invention provides a method and system for compressing computer documents. In accordance with the invention, a software facility ("the facility") preferably stores a number of computer documents, preferably publication documents, in a compressed format called a "compressed document set." Publication documents are comprised of a multi-level data structure organized as a tree of constituent data structures. For each page in a document, a constituent data structure in the publication document data structure corresponding to the page references a list of constituent data structures each corresponding to one visual element on the page. The visual element data structures each further reference separate constituent data items that contain name, formatting, or content data for the visual element.

A compressed document set, once constructed, is a tree whose general arrangement is similar to the tree data structures that represent the publication documents that it represents, but contains less data: instead of copying constituent data structures from the documents into the compressed document set that would be redundant in the compressed document set, the invention merely inserts in the compressed document set a reference to the copy of the constituent data structure that is already contained in the compressed document set. To construct a compressed document set from documents, therefore, the invention traverses the data structures for these publication documents in depth-first order. For each constituent data structure visited in the traversal of the document, the invention determines whether the compressed document set already contains the constituent data structure. If the compressed document set does not

already contain the constituent data structure, the invention adds the constituent data structure to the compressed document set and inserts a reference in the compressed document set that refers to the newly added constituent data structure.

If, on the other hand, the compressed document set does already contain the constituent data structure, the invention does not add the constituent data structure to the compressed document set, but merely inserts a reference in the compressed document set that refers to the matching constituent data structure already contained in the compressed document set. The result is a compressed document set that contains no duplicate constituent data structures. After compression, any document, and even any page of any document, may be individually extracted from the compressed document set into the normal publication document format.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a high-level block diagram of the general-purpose computer system upon which the facility preferably operates.

FIGS. 2A-2C are publication diagrams showing sample documents whose data structures may be compressed using the facility.

FIGS. 3A and 3B are data structure diagrams showing sample document data structures used by a publication application to represent the sample announcement and product list publication documents, respectively.

FIG. 4 is a data structure diagram showing a compressed document set data structure produced by the facility from the announcement and product list document data structures to represent the sample announcement and product list documents.

FIG. 5 is a data structure diagram showing a concatenated compressed document set data structure.

FIGS. 6A-6B together contain a flow diagram of a Compress Documents routine preferably called by the facility in order to compress a set of documents.

FIG. 7 is a flow diagram showing the Add Data Item routine, which is preferably called by the facility in step 608 (FIG. 6) to add a data item to the data list if necessary.

FIG. 8 is a flow diagram showing the Add Visual Element Descriptor routine, which is preferably called by the facility in step 611 (FIG. 6) to add a visual element descriptor to the object list, if necessary.

DETAILED DESCRIPTION OF THE
INVENTION

A method and system for compressing publication documents is provided. In a preferred embodiment, a software facility ("the facility") stores a number of computer documents, preferably publication documents, in a compressed format called a "compressed document set." Publication documents are comprised of a multi-level data structure organized as a tree of constituent data structures. For each page in a document, a constituent data structure in the publication document data structure corresponding to the page references a list of constituent data structures each corresponding to one visual element on the page. The visual element data structures each further reference separate constituent data items that contain name, formatting, or content data for the visual element.

A compressed document set, once constructed, is a tree whose general arrangement is similar to the tree data structures that represent the publication documents that it repre-