SYSTEM AND PROCESS FOR CREATING A STRUCTURED TAG REPRESENTATION OF A DOCUMENT

Field of the Invention

This invention relates to the field of extracting content from an existing document into a structural representation of the document.

Background of the Invention

The use of content created for a particular purpose for another use has increasingly become a problem. Frequently, it becomes necessary to extract the content from a document created for a particular purpose, such as for print, into a form that can be utilized by other applications. For example, the content for sites on the World Wide Web, hereinafter referred to as the Web, as well as for other sites on the Internet, Intranets or other interconnected electronic information sharing systems, is often already present in existing documents created for print purposes. The content, such as text and graphics, and possibly even audio, video or embedded programs, such as Java or applets, is bound into the documents prepared for print purposes by layout and/or style attributes. In order for this content to be useful, it must be extracted from the constraints created by these attributes. One alternative presently used is to perform a "cut and paste" operation to extract this content. However, this procedure disassociates substantially all of the presentational attributes (which describe or constrain the layout of the content) and the style attributes (which describes or constrains the "look" of the content) from the content. Thus, the content must be restructured in order to provide the presentation and style of the new document, even though the content is to be the same. This can become a tedious and time-consuming task.

The extracted content in order to be usable for other applications, whether for use on the Web, or in other presentation applications, needs to be structured for presentation. For example, presently there are mark up languages for structured documents, such as Standard Generalized Markup Language, (hereinafter referred to as "SGML") and eXtensible Markup Language (hereafter referred to as "XML"). These languages uti-

lize a structured document model approach. One such structured document model is referred to as Document Type Definition or "DTD". These models are typically provided in advance but can be arbitrarily created as needed as well. XML and SGML (and other mark up languages as they develop) use the DTD or other structured document models to associate the content with the appropriate mark up commands to enable the content to be displayed with a desired presentation and style. The mark up language adds identifiers for each of the "elements" or parts of the document for identification purposes. For instance, a DTD may define a document model as having a title, a main paragraph and several secondary paragraphs. The mark up language then adds identifiers, called a "tag", to designate the beginning and the end of a particular element. The presentational attributes and/or the style attributes can also be associated by additional tags or by association with separate style sheets. The use of such structural document models also can be used in converting existing documents into content which can be presented in other applications.

Presently, a structural representation of the existing document must be manually created for that document into which the content extracted from that document can be placed. This step is necessary before the mark up languages or other applications can be utilized. This step is tedious even for documents created with word-processing applications which have relatively few or simple design constraints. The use of more complex documents created with desktop publishing programs and which are tightly bound by many design constraints, such as presentational attributes and style attributes, cause this step to become even more of a burden.

There presently is no effective system for extracting the content from an existing document into a structural representation of that document without extensive intervention from an editor. There is a need for such a system and process for doing so.

Summary of the Invention

The present invention accomplishes these needs and others by providing a system and process for systematically extracting content from a document into a tagged structural representation of the document. In one preferred embodiment of the present invention, the system "interro-

gates" the document in order to systematically extract the content from document based on the defined rules or "hints". The system is able to structure the content in accordance with a defined structural document model (such as a DTD) to create a structural representation of the document. This structural representation can then be used to create a document to enable a meaningful presentation of the extracted content, such as on a browser or other presentation applications. There is no or little need for manual intervention in extracting the content from the document. Thus the system is able to quickly extract content from an existing document into a structured representation of the existing document. This is particularly useful when it is desired to create a plurality of documents of similar type from existing documents, or when there is a need to frequently update documents.

In one preferred embodiment, the present invention utilizes a structural document model, such as a Document Type Definition ("DTD") to define the structure of the content based on the elements of the document to be represented. The DTD is graphically represented by a logical structure tree, which allows the document elements to be easily inserted and/or moved about. Elements can be defined (or nested) within other elements. These elements can be, for example, a Title element, a Headline element, an Author element, a Body element, a Photo element, and so forth. The structure or sequence of particular elements are grouped together to form a particular DTD. Typically, most documents, particularly documents of a similar type, are represented by an existing DTD. For instance, a magazine will typically use a standard DTD for feature articles, another DTD for columns, and the like.

The system also includes a Hint Set to associate particular presentational or style attributes to each of the designated elements of a particular DTD. These presentational and/or style attributes are associated by selecting from a menu or by other means. For instance, the Title element for a particular DTD may be associated with the presentational attribute of being the first text box in the document or by a style attribute of having a particular font size. The Headline element may be associated with the presentational attribute of being the second text box. A Keyword element may be associated with the style attribute of having a particular character style, such as italics. Hint Sets can be applied to different DTD's as well.

The user can select a particular Hint Set from a menu of Hint Sets and associate that Hint Set to a particular DTD.

Once the Hint Set has been defined for or associated with a particular DTD, then the system is able to quickly "interrogate" the document, extract the content and create a structural representation of the extracted content in accordance with the DTD based on the selected Hint Set.

The system parses the document file to search for the attributes assigned to each of the elements in the Hint Set for that DTD. Once it finds a Hint or defined attributes associated with a particular element, it extracts the content associated with those attributes and associates that content with the element in the DTD to which that Hint have been associated. For example, a DTD may have a Title element, a Headline element, a Body element with a Picture element a subset of the Body element. The Hint Set for that DTD would have certain attributes associated with each of those elements. The system analyzes the document to which the DTD was applied. As it finds the presentational and style attributes associated with the Title element, it extracts the content to which those attributes were associated, associates that extracted content to the Title element and represents it in the structure for the Title element for that DTD. The system continues to search for the attributes for the remaining elements. As it finds each of the attributes or style sheets for each of the elements, it extracts that content, associates the extracted content with the element and represents that content with the element in the structure defined in the DTD.

The system of the preferred embodiment also employs heuristic techniques to improve the efficiency of the process. The system may encounter multiple options for various attributes in analyzing the document. The system is capable of intelligently resolving an appropriate path among these multiple options by the use of previous history, by looking ahead to the following sequence of Hints and elements, and by other intelligence. Also, the system of the preferred embodiment will query the user if unrecognized style sheets or attributes are encountered or if there are irreconcilable unresolved options. The decision provided by either the system in resolving the best option or by input from the user will then be used in other documents when those problems are encountered.

These and other features of the present invention are described in greater detail in the ensuing description of a preferred embodiment and in the drawings.

Brief Description of the Drawings

Figure 1 illustrates a screen shot of a document from which content is to be extracted under a preferred embodiment of the present invention.
Figure 2 illustrates a screen shot of a DTD and Hint Set of a preferred embodiment of the present invention.
Figure 3 is a screen shot of a structural representation of a document from which content has been extracted.
Figure 4 is another screen shot of Figure 3.
Figure 5 is an illustration an XML encoded version of the document of Figure 1.

Detailed Description of a Preferred Embodiment

The present invention provides a process and system for extracting content into a structural representation of a defined structural document model from an existing document. In one preferred embodiment of the present invention, the system "interrogates" the document to find the elements of the document based on a set of hints or rules associated with a selected structural document model, extracts the content for each of the document elements and structurally represents the extracted content in accordance with the selected structural document model. It is to be expressly understood that the exemplary description that is discussed herein is for descriptive purposes only and is not meant to limit the scope of the inventive concept. Other implementations of the inventive concept are considered to be within the scope of the appended claims.

There are numerous programs available for the electronic preparation of documents, particularly for print purposes. One such program is QuarkXPress™ distributed by Quark Distribution, Inc. It is to be expressly understood that the present inventive concept is intended for use with

documents created with other programs as well. This program, as well as other word-processing and/or desktop publishing systems, allow the user to input text and graphics into a user-defined layout in electronic digital form. The user is able to utilize presentational attributes such as design objects, including text boxes, picture boxes, lines, color fills, as well as locations, dimensions, spacing and the like. The user may also add style attributes to the content of the document, such as fonts, indentation, spacing, color, image types, and many other attributes. Selective groupings of certain attributes may be assigned designations as "style sheets". These style sheets can then be saved to allow reuse. The style sheets, for example, can be applied to a single element (such as a title, headline, paragraph, etc.) or to a group of elements (such as an article, book, etc.). For example, a title is normally the first text box and is often characterized by a center-justified sentence, in bold letters with a large font. This could be identified as a title style sheet. A headline style sheet may be the second text box while a keyword style sheet may be the third text box and/or having characters in a different style than the other elements, such as in italics. A paragraph is often characterized by a text box, with an indented sentence, followed by one or more other sentences and ending with a "hard return". This could be identified as a paragraph style sheet.

A plurality of presentational and/or style attributes can be grouped together to form a document. For instance, a technical note document may include a title style sheet, a headline style sheet, a keyword style sheet, a body text style sheet in which a series of paragraph style sheets could be included, and so forth. It is to be expressly understood that this description is intended for explanatory purposes only, and is not meant to limit the claimed inventions to this embodiment. The use of other embodiments of document types, and programs for creating them are considered to be within the scope of the claimed inventions.

An example of a document from which the content may be extracted is illustrated in Figure 1. This document (also referred to as an "Article"), prepared under QuarkXPress uses a Style Book which includes a Headline style ("When the Bough Breaks"), a SubHead style ("Mothers tell 20 secrets of keeping children from catching those nasty winter colds"), a Body style, a Photo style, and a PullQuote style ("those of us with newborns know the terror one experiences when children come down with their first case of the sniffles"). The Body style includes a BodySubHead

style and several paragraphs. The Photo style includes a Source style, Dimension style and a Caption style. These styles are all standard for this particular style of article and was creating in accordance with Style Books historically used by the industry.

The present system utilizes these elements to provide an intelligent heuristic and user-definable process for extracting the content of a document into structural representation of the original document. The user selects or defines a Hint Set for the extraction and structural representation of the content from a document. The user first creates or selects a Document Type Definition ("DTD") for the extraction process. It is to be expressly understood that DTD is only one example of a structural document model which could be used under the present invention. A window, such as illustrated in Figure 2, allows a user to define a DTD, or select one already created from a library. In this example, an Article DTD is selected, which is the same DTD used in creating the Article illustrated in Figure 1. It is defined as having a Headline element, a SubHead element, a Byline element, a Body element having a BodySubHead element, a p1 (first paragraph) element and a p (additional paragraphs) nested within the Body element, a Photo element having a Source element, a Width element, a Height element and a Caption element nested within it and a PullQuote element. The DTD is graphically represented by a logical structural tree, as shown in Figure 2. It is to be expressly understood that representations other than the logical structure tree embodiment can be utilized under the present invention.

Next, a "Hint Set" is associated with the selected DTD. The Hint Set associates certain presentational and/or style attributes or style sheets to each of the elements of the DTD. The system will "search" for these attributes in the original document based on the associated Hint Set. An example of a Hint Set is illustrated in Figure 2. The Hint Sets may be selected from a menu of defined Hint Sets, or defined by the user. The user is able to associate the sets of presentational or style attributes to the elements of the DTD as necessary or desired. The elements and attributes can be associated in the DTD and Hint Set by selection from a menu or by other known techniques.

In one preferred embodiment, existing style sheets may be used for the Hint Sets. For example, a style sheet may have already been defined for assigning presentational and style attributes for associating with con-

tent to create a Headline for the existing document. This style sheet can thus also be used in the Hint Set for association with Headline element. Regardless of whether the user utilizes defined style sheets or individually assigns the attributes, each of the elements in the DTD is associated with certain presentational (such as being the first text box, first paragraph, location, etc.) and/or style attributes (font types, character styles, color, etc.). An example is illustrated in Figure 2, where the Headline element for the Article DTD is associated with a Headline style sheet, the BodySubHead element is associated with a Sub sub head style sheet, the p1 element is associated with a Body style sheet, and the p element is associated with a Pull Out Quote style sheet.

The decision as to whether an element is mandatory or optional, that is, if the Hint for a particular element is not encountered, the system determines whether it can resolve the location or type of element, skip that Hint or element or query the user is defined in the DTD. In this example, the Byline element and the Subhead element are designated as optional (not shown). Thus, if the system is unable to find the style sheets in the document associated with the Byline element and/or the Subhead element, it ignores those elements. Also, decision as to whether there may be multiple occurrences of an element, for instance multiple first paragraphs or secondary paragraphs in the Body element or multiple Photos is also defined in the DTD.

The selected DTD and associated Hint Set is then applied to the desired document. In one preferred embodiment, the system of the present invention parses the document by checking the attributes or style sheets of the document. It analyzes those attributes in the document based on the Hint Set for the selected DTD. In this example, the system recognizes the original document as an Article. It then moves to the next Hint, a Headline style sheet, that it expects would contain the attributes for the Headline style sheet. If the system does find the attributes for the Headline style sheet, it extracts the content associated with the Headline style sheet and associates that extracted content with Headline element in the DTD. The system then parses to the next Hint, a Sub-head style sheet. The system continues in this fashion until it has analyzed each of the style sheets or sets of attributes set forth in the Hint Set. If the system is unable to find a Hint, or if it encounters attributes or style sheets which are not listed in the Hint Set, then it employs heuristic techniques to

resolve these issues. For example, if the system is expecting to encounter a particular Hint and does not, it may attempt to resolve the missing Hint by determining whether the Hint is mandatory or optional, whether there is another style sheet that may be used as the style sheet defined in the missing Hint, whether a previous decision based on previous history when this Hint is missing provides instruction on how to proceed, obtain guidance by "looking" ahead to the next sequence of Hints to determine whether to use another style sheet, or by other "intelligent" decisions. The system is also able to employ multiple paths to attempt to resolve this dilemma, such as skipping the Hint to see if the continuing sequence of Hints can be resolved. If the system is able to successfully resolve this issue, then this resolution goes into future decision making. If the system is unable to successfully resolve this issue, then the system may query the user for assistance. If the user provides assistance, or later corrects the structural representation, this assistance or correction can be later used by the system to resolve future dilemmas.

As more documents are interrogated by the system for a particular DTD and associated Hint Set, the more intelligence the system will acquire. Thus, the efficiency of the process will increase as more documents are processed. The Hint Set can be saved and applied to other document types. This is particularly useful when a number of similar documents are processed, or if a particular document is frequently updated. In one preferred embodiment of the present invention, the document from which the content is to be extracted is applied to an existing DTD using the desired Hint Set to create the tagged structural representation of the document.

By way of example, the document illustrated in Figure 1 is applied to the DTD and Hint Set illustrated in Figure 2. The system analyzed the document for the occurrence of the Hints for the applied DTD, as shown in Figure 2. The system recognized the attributes for the Headline style sheet, and extracted the content associated with that style sheet. This extracted content was associated with the DTD element Headline. The system then proceed to analyze the document for the style sheet Subhead. The system was unable to find this style sheet and since the SubHead element was designated as optional, ignored this element and proceeded on. Similarly, the system was unable to find the stylesheet Byline, and thus ignored the Byline element. The system was able to find

multiple style sheets for BodySubHead, p1 and p. Multiple occurrences of these elements had been allowed by the DTD and/or Hint Sheets, thus the system extracted the content associated with those style sheets and associated each of the extracted content to the appropriate element. The system extracted the content associated with each of these Hints and associated the extracted content to the appropriate structural elements of the DTD. Similarly, the system analyzed the document for the style sheets associated with Photo element, Source element, Width element, Height element and Caption element and associated the extracted content with the appropriate elements. A graphical structural representation of the extracted content is illustrated in Figures 4 and 5. In the preferred embodiment, the nested elements may be hidden for conciseness purposes in Figure 4. The nested elements may be viewed in a tree structure by opening the parent element, as shown in Figure 5.

In the preferred embodiment, as shown in Figure 5, as each element is highlighted in the structural representation, the extracted content associated with that element is displayed. This provides an efficient method for verifying the accuracy of the extraction process. Once the content has been systematically extracted and associated with elements in a structural representation of the existing document, that content can then be processed into a format that can be viewed or otherwise utilized. One example is the use of XML to create a graphically viewable presentation. Figure 6 illustrates the content extracted from the document shown in Figure 1 by a preferred embodiment of the present invention with XML tags applied. The XML tags provide the identifiers for each of the elements represented in the structural representation shown in Figures 4 and 5. The entire process, once a DTD and Hint Set has been selected, can extract the content from an existing document prepared for print into a structured representation of that document from which a presentation of that content can be created, such as for use on a Web site.

One feature of the preferred embodiment of the present invention, is the use of the complexities of the document itself to create a more efficient process for extracting the content into a structural relationship. The more "complexities", that is, the more presentational and/or style attributes present in the document, the more "hints" there are for the system to analyze the document and content for structural relationships. Previously, the greater the density of these attributes to create a stylistic document

increased the difficulty in extracting the content in a meaningful manner. The present system is able to efficiently utilize these attributes to extract the content into a structural relationship, and provide greater structural detail with higher density of attributes in the document.

While the descriptive embodiment is particularly useful in processing documents in QuarkXPress, other embodiments may also be used in conjunction with other publishing and/or word-processing systems.

The above embodiments are provided for descriptive purposes only and are not meant to unduly limit the scope of the present inventive concepts as set forth in the claims.

Claims

What is claimed is:

1. A process for extracting content from a document into a structural representation of the document, said process comprising the steps of:

arbitrarily defining a structural document model having elements;

designating attributes to said elements of the structural document model;

searching the document for said designated attributes;

extracting content from the document that is associated with said designated attributes; and

associating said extracted content with the elements to which said designated attributes associated with the extracted content are designated.

2. A system for extracting content into a structural representation of a document, said system comprising:

means for arbitrarily defining a structural document model having elements;

means for designating attributes to each of said elements;

means for searching the document for said designated attributes;

means for extracting content from the document that is associated with said designated attributes; and

means for associating said extracted content with said elements to which said designated attributes associated with said extracted content are designated.