# METHOD AND SYSTEM FOR SEARCHING AND RETRIEVING ELECTRONIC DOCUMENTS

## FIELD OF INVENTION

The present invention relates to a method and system for searching and retrieving electronic documents.

## BACKGROUND

Due to the vast amount of information stored in a database, a major issue in the retrieval of relevant electronic documents concerns the speed and efficiency in which these electronics documents could be retrieved. Factors that could affect the speed and efficiency in which these electronics documents could be retrieved, includes different representations of the same document, for example, Microsoft ® Word Document can be formatted as Portable Document Format (PDF), as plain text, as HTML, etc, a storage system may have several versions of the same document and/or documents are componentized and the same paragraph, slide or figure may appear in multiple documents.

In order to speed up the retrieval of the relevant electronic documents in a database, a number of search methodologies have been devised and implemented. For instance, an existing method includes implementing systems for storing and querying XML data. However, the existing systems for storing and querying XML data, such as GoXML™ and Xyleme™, do not support the format of metadata.

An example of an existing device for searching and retrieving electronic document is disclosed in US Patent No. 7054859. In particular, it discloses a method for responding to search requests for stored documents. Each of the stored documents includes metadata, content, and a document type identifier for identifying a type of the document. A search request including document type criteria, metadata criteria, and content criteria, is received by the system. A first set of documents based on the metadata criteria is identified. Documents in the first set that have a document type identifier that matches the document type criteria are identified. A search method is identified from a plurality of search methods for searching the content of the identified documents in the first set based on the document type identifier of the documents. The content of the identified documents in the first set is searched using the identified search method and the

content criteria. However, the system is slow and inefficient in searching for the identified documents.

As described in US Patent No. 7054859, metadata are data that describes information about the XML document instance. It provides users an overview of the document like the author, title, etc. However, US Patent No. 7054859 does not inform the users how to navigate around the document. The method depends on a database of cost function associated with a search method each to determine the best way to navigate. This search method requires a prior knowledge of the content of the XML instance, which is difficult to automate. Additionally, the current method is only capable of retrieving documents that satisfy the search conditions. It does not have the capability to pinpoint the search to the exact location within the document. Further, in US Patent No. 7054859, a XML document instance is not componentized and not treated as an aggregation of individual XML nodes. Accordingly, updating and modification of the document instance are not considered as an embodiment of the idea. Further, the current method requires continuous computation and evaluation of the least cost search methods. As such, there is inherently the issue of maintaining the integrity and reliability of the computed results. Evaluation could only be optimal if there are vast collection of search methods and search evaluation metrics. This would require time to build up, constant vigilance of system administrator and a vast amount of user traffics. Failure of any one of these would mean a breakdown in the system's ability to carry out an efficient search.

Moreover, in US Patent No. 7054859, the efficiency of and the type of XML Parser that is available to the user will affect the speed the user could modify the data. The amount of computing memory the user has at his disposal will also affect the speed of retrieval and replacement of the data. With the system, rapid updates/retrieval of a set of data within a large XML document instance or across multiple XML document instance is difficult as the amount of processing required to locate a particular node among the XML document instances is large. Given this limitation and in the context of real-time computing where "hard" deadline can never be compromised the convenience offered by the XML file is severely limited.

Therefore, it would be desirable to have a retrieval system that is fast and efficient in retrieving stored electronic documents, and at the same time, not limited to XML data. Other types of electronic documents may include Delimiter-Separated Value (DSV) File, PDF, ODF

File (Open Document Format), DVI File (Page Description File), HTML / xHTML File (Hyper-Language Web File) and SVG File (which is a special type of XML File).

It is an object of the present invention to overcome, or at least substantially ameliorate, the disadvantages and shortcomings of the prior art.

In particular, one of the object of the present invention is to ensure that the content of a huge XML document instance will not incur high cost in term of memory and computing power since only a small percentage of the entire file need to be parsed to provide the user with an overview of the information housed within the XML document instance. Additionally, since each and every of the XML node is unique and this enables merging of multiple XML document instances with little unexpected errors.

Other objects and advantages of the present invention will become apparent from the following description, taken in connection with the accompanying drawings, wherein, by way of illustration and example, an embodiment of the present invention is disclosed.

**SUMMARY OF THE INVENTION**

According to a first aspect of the present invention, there is provided a method for searching and retrieving electronic documents, the method comprising providing a document file including a plurality of actual data and a plurality of data identifiers, the data identifier being linked to the actual data, entering a query expression, identifying the data identifier according to the query expression, searching and locating the actual data in the document file, and retrieving the actual data from the document file.

Preferably, the data identifier is a metadata tag.

Preferably, the metadata tag comprise of metadata tag criteria.

Preferably, the document file may be an extended markup language (XML) document.

Preferably, the data identifier is linked to the actual data by a universal identifier.

According to a second aspect of the present invention, there is provided a system for searching and retrieving electronic documents, the system comprising a document file including a plurality of actual data and a plurality of data identifiers, each data identifier being linked to the actual data, a navigation controller for receiving a query and for searching and locating the actual data in the document file based on the query, an array for storing the actual data being searched and located by the navigation controller, an update/retrieve section for updating and retrieving the actual data being stored in the array section, a controller for receiving information from the update/retrieve section and displaying a cursor pointing to the actual data, an output section for editing an arrangement of the cursor and the actual data and sending information to a display block, and the display block for displaying the data and the cursor.

**BRIEF DESCRIPTION OF THE DIAGRAMS**

Further objects of this invention, together with additional features contributing thereto and advantages accruing therefrom, will be apparent from the following description of an embodiment of the present invention which is shown in the accompanying drawings with like reference numerals indicating corresponding parts throughout and which is to be read in conjunction with the following drawings, wherein:

Figure 1 shows a high level block diagram illustrating a XML document instance with metadata tag and actual data.

Figure 2 shows a flowchart of a document retrieving process performed by the document retrieval system according to one embodiment of the present invention.

Figure 3 shows a sample XML document illustrating the arrangement of the navigation tags and the actual data tag. It also shows how the metadata and the actual data are linked by the GUID and the full path. When GUID and full path is combined will aid application to navigate to the exact node, retrieve or modify the data in a highly efficient manner.

Figure 4 shows a block diagram illustrating the actions as explained in the flowchart.

These and additional embodiments of the invention may now be better understood by turning to the following detailed description wherein a preferred embodiment is described.

## DETAILED DESCRIPTION OF THE INVENTION

A method for searching and retrieving electronic documents is described. In the following description, for purposes of explanation, numerous specific details are set forth in order to provide a thorough understanding of the invention. It will be apparent, however, to one skilled in the art that the invention can be practiced without these specific details. In other instances, structures and devices are shown in block diagram form in order to avoid obscuring the invention.

Reference in the specification to "one embodiment" or "an embodiment" means that a particular feature, structure, or characteristic described in connection with the embodiment is included in at least one embodiment of the invention. The appearances of the phrase "in one embodiment" in various places in the specification are not necessarily all referring to the same embodiment.

The present invention generally comprises a document file having a plurality of data identifier at a first end and a plurality of actual data at a second end, the data identifier being linked to the actual data. The plurality of data identifier may also be known as a block of “roadmap” navigation data. Each data identifier, which may also be known as “MetaData Tag”, aids in the navigation and location of the actual data during a searching process and thus enables faster retrieval or modification of actual data in the document file, such as extended markup language (XML) document instance. When in use, a query expression identifies the data identifier which in turn searches and locates the actual data in the document file and retrieve the actual data from the document file.

Figure 1 shows a preferred embodiment of the present invention. As shown, the document file 100, such as a XML document instance, contains a plurality of data identifiers or a block of metadata tags 102, located at the beginning of the XML document 100 to facilitate faster searching and retrieving of stored documents or the actual data 122. The primary purpose of the plurality of data identifiers, such as the block metadata tags 102, is to aid in the navigation during the search process so that retrieving of the actual data 120 takes a shorter time.

The block of metadata tags 102 houses a collection of individual metadata tags 104**.** Each metadata tag 104 provides an overview of an actual section of the XML document instance 100. The metadata tag 104 describes the content of the data that includes the names 106 of actual XML node it point to, the data type of the XML node 108, the last modification and time to this node 110, a universal ID that can be use to identify modification across space

and time 112 and a full path node that upon which a program can pinpoint the exact location of the actual data in this XML document instance 114. As shown in Figure 1, there may be a plurality of metadata tags in which each of the metadata tag 104 is linked to a specific actual data 122.

At the other end of the XML document instance 100 is the plurality of actual data 122. The actual data 122 embodies fully the metadata previously described, which is located at the beginning of the XML document. The actual data 122 is linked to the metadata by means of an universal identifier.

Figure 2 shows a flow chart of a document retrieval process according to a preferred embodiment of the present invention. In step 200, the user enters a query expression to determine a range of data that are to be retrieved or updated. This may be done by manual entry or programmatically. In step 202, where updates to data are desired, the user or computer program needs to submit a set of new values. In step 204, the retrieval system first identifies a set of candidate documents that meet the query expression. In step 206, for each of the identified candidate document, an application reads the entire metadata block 102 (as shown in Figure 1) and parse the metadata block into tree nodes. In step 208, based on the tree nodes created in step 206, the application is able to quickly determine if the documents contain the desired data. If desired data tag is not found in the document, the system retrieves the next candidate document on the list. In step 210, if the desired data tag is found in the document, the system retrieves the path of each tag, which meets the condition and stores the metadata of each and every node that satisfies the condition into an array. In step 212, the system checks if there is anymore document on the candidate document list. If there is, the system reads the metadata block of the next file on the list as in step 206. If the document is the last document on the candidate document list, the system proceeds to step 214. In step 214, the system goes to head of array created in step 210, which would have been populated with a collection of metadata nodes. In step 216, iteration begins to the next metadata tag pointing to exact location of the desired data node. In step 218, the application is able to rapidly place the cursor to the desired node to retrieve or update the data with the existence of full path 114 in the metadata of the XML document instance. In step 220, the system iterates to the next metadata until the metadata is the last element in the array. When last element of the array is reached, the application exits the loop. In step 222, if the system performs retrieval operation, the desired data shall be aggregated and displayed to the user.

The system reports on the status of the operation to the user whether partially or completely successful if the updating operation is performed.

Figure 3 shows a sample print screen of the arrangement of the navigation metadata tag and the actual data tag. Additionally, it also illustrates how the metadata and the actual data are linked by means of the GUID 112 and the full path 114 in the XML document instance 100. Both attributes, when combined will help application to navigate to the exact node in a highly efficient manner. The unique ID allows the merging of different XML files housing different versions of the same/different file. Since the ID is unique over time and space, old and new nodes can be compared and merged into single entity efficiently.

Figure 4 shows a block diagram illustrating a high level abstract of the flow process in Figure 2. Data Storage 401 stores block XML documents instances including metadata tag and actual data. The Navigation Controller 402 receives queries from user/system and searches the data storage by using a metadata tag. Data found are stored into array section 403. Update/Retrieve Section 404 searches the array section 403 for pertinent data as well as collects actual data from Data Storage 401 and updates and retrieves the data. Cursor Controller 405 receives information from update/retrieve section and display a cursor pointing to the actual data. Output Section 406 edits arrangement of data and cursor and sends information to the display. Finally, the display block 407 displays data and a cursor.

Although exemplary embodiments of the present invention have been shown and described, it will be apparent to those skilled in the art that a number of changes, modifications, or alternations to the invention as described herein may be made, none of which depart from the spirit of the present invention. All such changes, modifications and alterations should therefore be seen as within the scope of the present invention.

**CLAIMS**:

1. A method for searching and retrieving electronic documents, the method comprising:

providing a document file including a plurality of actual data and a plurality of data identifiers, each data identifier being linked to the actual data,

entering a query expression,

identifying the data identifier according to the query expression,

searching and locating the actual data in the document file, and

retrieving the actual data from the document file.

2. The method according to claim 1, wherein the data identifier is a metadata tag.

3. The method according to any one of the preceding claims, wherein, the metadata tag comprises metadata tag criteria.

4. The method according to any one of the preceding claims, wherein the document file may be an extended markup language (XML) document.

5. The method according to any one of the preceding claims, wherein the data identifier is linked to the actual data by an universal identifier.

6. A system for searching and retrieving electronic documents, the system comprising:

a document file including a plurality of actual data and a plurality of data identifiers, each data identifier being linked to the actual data,

a navigation controller for receiving a query and for searching and locating the actual data in the document file based on the query,

an array for storing the actual data being searched and located by the navigation controller,

an update/retrieve section for updating and retrieving the actual data being stored in the array section,

a controller for receiving information from the update/retrieve section and displaying a cursor pointing to the actual data,

an output section for editing an arrangement of the cursor and the actual data and sending information to a display block, and

the display block for displaying the data and the cursor.