(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
24 April 2008 (24.04.2008)

PCT

(10) International Publication Number
WO 2008/049023 A2

(51) International Patent Classification:
G06F 17/00 (2006.01) *G06F 17/30* (2006.01)

(21) International Application Number:
PCT/US2007/081681

(22) International Filing Date: 17 October 2007 (17.10.2007)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/852,584	17 October 2006 (17.10.2006)	US
11/694,869	30 March 2007 (30.03.2007)	US

(71) Applicant *(for all designated States except US)*: **COMMVAULT SYSTEMS, INC.** [US/US]; 2 Crescent Place, Oceanport, NJ 07757-0090 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **GOKHALE, Parag** [US/US]; 5 Cotswold Circle, Ocean City, NJ 07712 (US). **KOTTOMTHARAYIL, Rajiv** [IN/US]; 7 Skylark Ct., Marlboro, NJ 07746 (US). **ATTARDE, Deepak, Raghunath** [IN/US]; 1 Willow Drive, Apt. #107, Ocean City, NJ 07712 (US). **AHN, Jun, H.** [KR/US]; 7 Waterville Rd, Manalapan, NJ 07726 (US). **PRAHLAD, Anand** [US/US]; 3 Bucknell Drive, East Brunswick, NJ 08816 (US). **SCHWARTZ, Jeremy, A.** [US/US]; 30 Drummond PL, Red Bank, NJ 07701 (US). **NGO, David** [US/US]; 118 Borden St, Shrewsbury, NJ 07702 (US). **BROCKWAY, Brian** [US/US]; 3 Brady Road, Shrewsbury, NJ 07702 (US). **MULLER, Marcus, S.** [US/US]; 4 Fennec Court, Tinton Falls, NJ 07757 (US).

(74) Agents: **BOSWELL, J., Mason** et al.; Perkins Coie LLP, P.O. Box 1247, Seatlle, WA 98111-1247 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

WO 2008/049023 A2

(54) Title: METHOD AND SYSTEM FOR OFFLINE INDEXING OF CONTENT AND CLASSIFYING STORED DATA

(57) Abstract: A method and system for creating an index of content without interfering with the source of the content includes an offline content indexing system that creates an index of content from an offline copy of data. The system may associate additional properties or tags with data that are not part of traditional indexing of content, such as the time the content was last available or user attributes associated with the content. Users can search the created index to locate content that is no longer available or based on the associate attributes.

METHOD AND SYSTEM FOR OFFLINE INDEXING OF CONTENT AND CLASSIFYING STORED DATA

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] The present application claims priority to U.S. Provisional Application No. 60/852,584 (Attorney Docket No. 60692-8047.US00) entitled "METHOD AND SYSTEM FOR COLLABORATIVE SEARCHING," and filed on October 17, 2006, which is hereby incorporated by reference.

BACKGROUND

[0002] Computer systems contain large amounts of data. This data includes personal data, such as financial data, customer/client/patient contact data, audio/visual data, and much more. Corporate computer systems often contain word processing documents, engineering diagrams, spreadsheets, business strategy presentations, and so on. With the proliferation of computer systems and the ease of creating content, the amount of content in an organization has expanded rapidly. Even small offices often have more information stored than any single employee can know about or locate.

[0003] Many organizations have installed content management software that actively searches for files within the organization and creates an index of the information available in each file that can be used to search for and retrieve documents based on a topic. Such content management software generally maintains on index of keywords found within the content, such as words in a document.

[0004] Creating a content index generally requires access to all of the computer systems within an organization and can put an unexpected load on already burdened systems. Some organizations defer content indexing until off hours, such as early in the morning to reduce the impact to the availability of systems. However, other operations may compete for system resources during off hours. For example, system backups are also generally scheduled for off hours. Systems may be placed in an unavailable state during times when backups are being performed, called the

backup window, to prevent data from changing. For organizations with large amounts of data, any interruption, such as that from content indexing, jeopardizes the ability to complete the backup during the backup window.

[0005] Furthermore, traditional content indexing only identifies information that is currently available within the organization , and may be insufficient to find all of the data required by an organization. For example, an organization may be asked to produce files that existed during a past time period in response to a legal discovery request. Emails from five years ago or files that have been deleted or are no longer available except in offsite backup tapes may be required to answer such a request. An organization may be obligated to go through the time consuming task of retrieving all of this content and conducting a manual search for content related to the request.

[0006] There is a need for a system that overcomes the above problems, as well as providing additional benefits.

BRIEF DESCRIPTION OF THE DRAWINGS

[0007] Figure 1 is a block diagram that illustrates components of a system, in one embodiment of the invention.

[0008] Figure 2 is a block diagram that illustrates flow of data through the system, in one embodiment.

[0009] Figure 3 is a flow diagram that illustrates processing of a content indexing component of the system, in one embodiment.

[0010] Figure 4 is a flow diagram that illustrates processing of an index searching component of the system, in one embodiment.

[0011] Figure 5 illustrates a data structure containing entries of a content index, in one embodiment.

[0012] In the drawings, the same reference numbers and acronyms identify elements or acts with the same or similar functionality for ease of understanding and convenience. To easily identify the discussion of any particular element or act, the most significant digit or digits in a reference number refer to the Figure number in

which that element is first introduced (e.g., element 1104 is first introduced and discussed with respect to Figure 11).

[0013] The headings provided herein are for convenience only and do not necessarily affect the scope or meaning of the claimed invention.

DETAILED DESCRIPTION

Overview

[0014] A method and system for creating an index of content without interfering with the source of the content including an offline content indexing system that creates an index of content from an offline copy of data is provided. In general, organizations may have a primary or production copy of source data and one or more offline or secondary copies of data. Secondary copies can be created using various storage operations such as snapshots, backups, replication, migration, and other operations. The offline content indexing system can create an index of an organization's content by examining secondary copies of the organization's data (e.g., backup files generated from routine backups performed by the organization). The offline content indexing system can index content from current secondary copies of the system as well as older offline copies that contain data that may no longer be available on the organization's network. For example, the organization may have secondary copies dating back several years that contain older data that is no longer readily available, but may still be relevant to the organization. The offline content indexing system may associate additional properties with data that are not part of traditional indexing of content, called metadata, such as the time the content was last available or user attributes associated with the content. For example, user attributes such as a project name with which a data file is associated may be stored.

[0015] Members of the organization can search the created index to locate content that is no longer readily available or based on the associated attributes. For example, a user can search for content related to a project that was cancelled a year ago. Thus, users can find additional organization data that is not available in traditional content indexing systems. Moreover, by using secondary copies, content indexing does not impact the availability of the system that is the original source of the content.

[0016] In some embodiments, members of the organization can search for content within the organization through a single, unified user interface. For example, members may search for content that originated on a variety of computer systems within the organization. Thus, users can access information from many systems within the organization and can search for content independent of the content's original source. Members may also search through multiple copies of the content, such as the original copy, a first secondary backup copy, and other secondary or auxiliary copies of the content.

[0017] Various attributes, characteristics, and identifiers (sometimes referred to as tags or data classifications) can be associated with content. The system may define certain built-in tags, such as a document title, author, last modified date, and so on. Users of the system may also define custom tags, or the system may automatically define custom tags. For example, an administrator may add tags related to groups within an enterprise, such as a tag identifying the department (e.g., finance, engineering, or legal) that created a particular content item. Individual users may also add tags relevant to that user. For example, a user might add a descriptive field, such as a programmer adding a check-in description to identify a change made to a version of a source code document. For content that is inherently unstructured or appears random outside of its intended purpose, tags are an especially effective way of ensuring that a user can later find the content. For example, United States Geological Survey (USGS) data is composed of many numbers in a file that have little significance outside of the context of a map or other associated viewer for the data. Tags allow descriptive attributes or other meaningful information to be associated with the data, for example, so that a searching user can know at a glance that particular USGS data refers to a topological map of a nearby lake. Tags may be associated with offline and online data through a metabase or other suitable data structure that stores metadata and references to the content to which the metadata applies. Figure 5, discussed below, describes one exemplary data structure used to store user tags associated with content.

[0018] The invention will now be described with respect to various embodiments. The following description provides specific details for a thorough understanding of, and enabling description for, these embodiments of the invention. However, one skilled in the art will understand that the invention may be practiced

without these details. In other instances, well-known structures and functions have not been shown or described in detail to avoid unnecessarily obscuring the description of the embodiments of the invention.

[0019] The terminology used in the description presented below is intended to be interpreted in its broadest reasonable manner, even though it is being used in conjunction with a detailed description of certain specific embodiments of the invention. Certain terms may even be emphasized below; however, any terminology intended to be interpreted in any restricted manner will be overtly and specifically defined as such in this Detailed Description section.

Creation of an Offline Copy

[0020] As discussed above, the offline content indexing system may create a secondary copy, such as an offline copy, as part of an existing backup schedule performed by an organization. For example, an organization may perform weekly backups that contain a complete copy of the organization's data. It is generally not necessary for the offline content indexing system to consume any further resources of the computer systems within the organization that contain source content, since all of the needed data is typically available in the backup data files. The offline content indexing system may restore the backed up data to an intermediate computer system that is not critical to the operation of the organization, or may operate on the backup data files directly to identify and index content. The offline content indexing system may also create the offline copy using copies of data other than a traditional backup, such as a snapshot, primary copy, secondary copy, auxiliary copy, and so on.

[0021] In some embodiments, the offline content indexing system uses a change journal to create an offline copy of content. Modern operating systems often contain built in change journaling functionality that stores a journal entry whenever data is changed within a computer system. The change journal generally contains a step-by-step, sequential, or ordered log of what data changed and how the data changed that can be processed at a later time to recreate the current state of the data. The change journal may be used in conjunction with a full data backup or other data protection mechanisms. The full backup can be used to establish the

state of the data at a prior point-in-time, and then the change journal entries can be used to update the state with subsequent changes.

[0022] In some embodiments, the offline content indexing system or other system uses a data snapshot to create an offline copy of content. Newer operating systems and several data storage companies offer snapshot software capable of taking a snapshot of the content currently on a computer system with minimal impact to the availability of the system. For example, the snapshot may simply note the current entry in a change journal, and keep track of subsequent change journal entries for updating the snapshot. These snapshots can be transferred from the host system and read on another, less critical system or can be used to replicate the data to a different. The offline content indexing system can then access this intermediate system to identify content and perform content indexing. Other technologies that will be recognized by those skilled in the art, such as disk imaging, mirroring, incremental backups, and so on, may be used in a manner similar to create an offline copy of data for content indexing.

[0023] In some embodiments, the offline content indexing system selects an offline copy of data for indexing among several available offline copies. For example, an organization may have several copies of data available on different types of media. The same data may be available on a tape, on a backup server, through network attached storage, or on fast mounted disk media. The offline content indexing system may take into consideration factors such as the access time of a particular media and the scheduled load on a particular offline copy when selecting a copy to use for indexing. For example, an offline copy stored on a hard drive may be preferred over a copy stored on tape due to the faster access time of the hard drive copy and the ability to randomly seek among the data rather than accessing the data sequentially. Alternatively or additionally, a backup server storing or responsible for otherwise desirable data to index scheduled to perform an intensive operation such as encrypting content may be skipped in favor of using a different server responsible for an offline copy that is not expected to be needed by other systems during the time expected to index the content. Similarly, the offline content indexing system may prefer an unencrypted offline copy over an encrypted one due to the extra effort required to decrypt the content to index it.

Indexing of Content

[0024] In some embodiments, the offline content indexing system may wait to index content until a request related to the content is received. Searches for offline content may not be as time sensitive as searches for currently available content such that the effort of indexing the content can be postponed until the content is required. For example, in a legal discovery request there may be several days or even weeks available to find content responsive to the request, such that indexing before a request is received would unnecessarily burden an organization's systems.

[0025] In some embodiments, the offline content indexing system may postpone content indexing until other storage operations have been performed. For example, one storage operation, called single instancing, may reduce or eliminate redundant files contained in backup data caused by many systems containing the same operating system or application files. By postponing content indexing until after single instancing has occurred, the offline content indexing system does not have to search as much data and may complete the indexing process sooner and with less burden to the organization's systems. A storage policy or other system parameter setting or preference may define how and when content indexing is done, and what other operations are performed before and after content indexing (e.g., indexing content after single instancing). A storage policy is a data structure that stores information about the parameters of a storage operation. For example, the storage policy may define that only some content is to be indexed, or that content indexing should occur late at night when system resources are more readily available.

[0026] In some embodiments, the offline content indexing system may update a content index according to an indexing policy. An indexing policy is a data structure that stores information about the parameters of an indexing operation. For example, an organization may create a full backup once a week, and may create an indexing policy that specifies that the index should be updated following each weekly full backup. Indexing the full backup creates a reference copy that the organization can store according to legal requirements (e.g., ten years) to respond to any compliance requests. The indexing policy may also specify that incremental updates are performed on the index based on incremental backups or other incremental data protection operations such as updates from a change journal or snapshot

application. For example, incremental backups may be created that only specify the data that has changed since the last full backup, and content changes identified within the incremental backup may be used by the offline content indexing system to update the index to reflect the new state of the content. If the backup data indicates that content has been deleted, the indexed content may be retained, but may be flagged or otherwise identified as having been deleted.

Content Tags

[0027] In some embodiments, the offline content indexing system tags or otherwise identifies indexed content with additional information that may help identify the information, for example, in a search for content. For example, indexed content may be tagged with the location of the offline copy in which the information was found, such as a particular backup tape or other offline media. The system may also tag online content, such as tagging a new file with the name of its author. If the content is later deleted, the indexed content may be tagged with the date the content was deleted, the user or process that deleted the content, or the date the content was last available. Deleted content may later be restored, and the indexed content may be identified by a version number to indicate versions of the content that have been available on computing systems throughout the content's history. Other information about the content's availability may also be stored, such as whether the content is stored onsite or is archived offsite, and an estimate of the time required to retrieve the content. For example, if the content is stored offsite with an external archival company, the company may require one week's notice to retrieve the content, whereas if the content is stored on a tape within the organization, the content may be available within an hour. Other factors may also be used to provide a more accurate estimate, such as the size of the content, the offset of the content if it is on tape, and so on. During a search, the search results may indicate whether the time required to retrieve certain content would exceed a retrieval threshold. The system may also prohibit transferring content beyond a given retrieval time to ensure compliance with a policy of the organization.

[0028] In some embodiments, the offline content indexing system tags content with classifications. For example, the offline content indexing system may classify content based on the type of application typically used to process the content, such

as a word processor for documents or an email client for email. Alternatively or additionally, content may be classified based on the department within the organization that generated the content, such as marketing or engineering, or based on a project that the content is associated with such as a particular case within a law firm. Content may also be classified based on access rules associated with the content. For example, some files may be classified as confidential or as only being accessible to a certain group of people within the organization. The system may identify keywords within the content and classify the content automatically based on identified keywords or other aspects of the content.

Searching

[0029] In some embodiments, the offline content indexing system searches for content based on temporal information related to the content. For example, a user may search for content available during a specified time period, such as email received during a particular month. A user may also search specifically for content that is no longer available, such as searching for files deleted from the user's primary computer system. The user may perform a search based on the attributes described above, such as a search based on the time an item was deleted, or based on a project that the item was associated with. A user may also search based on keywords associated with user attributes, such as searching for files that only an executive of the organization would have access to, searching for files accessed by a particular user, or searching for files tagged as confidential.

[0030] In some embodiments, the offline content indexing system provides search results that predict the availability of content. For example, content stored offsite may need to be located, shipped, and then loaded back into the organization's systems before it is accessible. The offline content indexing system may provide a time estimate of how soon the content could be available for searching as well as providing limited information about the content immediately based on data stored in the index. For example, the content indexing system may maintain a database of hardware and libraries of media available with the organization, as well as the current location of each of these items such that an estimate can be generated for retrieving the hardware or libraries of media. For example, certain tape libraries may be stored offsite after a specified period of time,

and content stored within the tape library may take longer to retrieve than content in a tape library stored onsite in the organization. Similarly, the offline content index system may estimate that data stored on tape will take slightly longer to retrieve than data that is available through magnetic storage over the network.

Figures

[0031] Unless described otherwise below, aspects of the invention may be practiced with conventional systems. Thus, the construction and operation of the various blocks shown in Figure 1 may be of conventional design, and need not be described in further detail herein to make and use the invention, because such blocks will be understood by those skilled in the relevant art. One skilled in the relevant art can readily make any modifications necessary to the blocks in Figure 1 (or other embodiments or Figures) based on the detailed description provided herein.

[0032] Figure 1 is a block diagram that illustrates components of the system, in one embodiment. The offline content indexing system 100 contains an offline copy component 110, a content indexing component 120, an index searching component 130, an index policy component 140, a data classification component 150, a single instancing component 160, an encryption component 170, and an archive retrieval component 180. The offline copy component 110 creates and identifies offline or other secondary copies of data, such as backup data, snapshots, and change journal entries. The content indexing component 120 creates and updates a content index based on offline copies of data. The index searching component 130 searches the index based on user requests to identify target content. The index policy component 140 specifies a schedule for updating the content index incrementally or refreshing the content index, such as from a full weekly backup. The data classification component 150 adds data classifications to the content index based on various classifications of the data, such as the department that created the data, and access information associated with the data. The single instancing component 160 eliminates redundant instances of information from offline copies of data to reduce the work involved in creating an index of the offline copy of the data. The encryption component 170 encrypts and decrypts data as required to permit access to the data for content indexing. The archive retrieval component 180

retrieves archived content from offsite storage, tape libraries, and other archival locations based on requests to access the content and may also provide estimates of the time required to access a particular content item.

[0033] Figure 1 and the following discussion provide a brief, general description of a suitable computing environment in which the invention can be implemented. Although not required, aspects of the invention are described in the general context of computer-executable instructions, such as routines executed by a general-purpose computer, e.g., a server computer, wireless device or personal computer. Those skilled in the relevant art will appreciate that the invention can be practiced with other communications, data processing, or computer system configurations, including: Internet appliances, hand-held devices (including personal digital assistants (PDAs)), wearable computers, all manner of cellular or mobile phones, multi-processor systems, microprocessor-based or programmable consumer electronics, set-top boxes, network PCs, mini-computers, mainframe computers, and the like. Indeed, the terms “computer,” "host," and "host computer" are generally used interchangeably herein, and refer to any of the above devices and systems, as well as any data processor.

[0034] Aspects of the invention can be embodied in a special purpose computer or data processor that is specifically programmed, configured, or constructed to perform one or more of the computer-executable instructions explained in detail herein. Aspects of the invention can also be practiced in distributed computing environments where tasks or modules are performed by remote processing devices, which are linked through a communications network, such as a Local Area Network (LAN), Wide Area Network (WAN), Storage Area Network (SAN), Fibre Channel, or the Internet. In a distributed computing environment, program modules may be located in both local and remote memory storage devices.

[0035] Aspects of the invention may be stored or distributed on computer-readable media, including magnetically or optically readable computer discs, hard-wired or preprogrammed chips (*e.g.*, EEPROM semiconductor chips), nanotechnology memory, biological memory, or other data storage media. Indeed, computer implemented instructions, data structures, screen displays, and other data under aspects of the invention may be distributed over the Internet or over other networks (including wireless networks), on a propagated signal on a propagation

medium (e.g., an electromagnetic wave(s), a sound wave, etc.) over a period of time, or they may be provided on any analog or digital network (packet switched, circuit switched, or other scheme). Those skilled in the relevant art will recognize that portions of the invention reside on a server computer, while corresponding portions reside on a client computer such as a mobile or portable device, and thus, while certain hardware platforms are described herein, aspects of the invention are equally applicable to nodes on a network.

[0036] Figure 2 is a block diagram that illustrates the flow of data through the system 100, in one embodiment. Content is initially stored on a data server 210 that may be a user computer, data warehouse server, or other information store accessible via a network. The data is accessed by a backup manager 220 to perform a regular backup of the data. The backup manager 220 may be contained within the data server 210 or may be a separate component as shown. For example, the backup manager 220 may be part of a server dedicated to managing backup or other storage operations. Backup data is stored in a backup data store 230 such as a network attached storage device, backup server, tape library, or data silo. The content indexing system 240 accesses data from the backup data store 230 to perform the functions described above. As illustrated in the diagram, because the content indexing system 240 works with an offline copy of the data, the original data server 210 is not negatively impacted by the operations of the content indexing system 240.

[0037] Figures 3-4 are representative flow diagrams that depict processes used in some embodiments. These flow diagrams do not show all functions or exchanges of data, but instead they provide an understanding of commands and data exchanged under the system. Those skilled in the relevant art will recognize that some functions or exchange of commands and data may be repeated, varied, omitted, or supplemented, and other (less important) aspects not shown may be readily implemented.

[0038] Figure 3 is a flow diagram that illustrates the processing of the content indexing component 120 of the system 100, in one embodiment. The component is invoked when new content is available or additional content is ready to be added to the content index. In step 310, the component selects an offline copy of the data to be indexed. For example, the offline copy may be a backup of the data or a data

snapshot. In step 320, the component identifies content within the offline copy of the data. For example, the component may identify data files such as word processing documents, spreadsheets, and presentation slides within a backup data file. In step 330, the component updates an index of content to make the content available for searching. The component may parse, process, and store the information. For example, the component may add information such as the location of the content, keywords found within the content, and other supplemental information about the content that may be helpful for locating the content during a search. After step 330, these steps conclude.

[0039] Figure 4 is a flow diagram that illustrates the processing of the index searching component 130 of the system 100, in one embodiment. In step 410, the component receives a search request specifying criteria for finding matching target content. For example, the search request may specify one or more keywords that will be found in matching documents. The search request may also specify boolean operators, regular expressions, and other common search specifications to identify relationships and precedence between terms within the search query. In step 420, the component searches the content index to identify matching content items that are added to a set of search results. For example, the component may identify documents containing specified keywords or other criteria and add these to a list of search results. In step 425, the component generates search results based on the content identified in the content index. In step 430, the component selects the first search result. In decision step 440, if the search result indicates that the identified content is offline, then the component continues at step 450, else the component continues at step 455. For example, the content may be offline because it is on a tape that has been sent to an offsite storage location. In step 450, the component retrieves the archived content. Additionally or alternatively, the component may provide an estimate of the time required to retrieve the archived content and add this information to the selected search result. In decision step 455, if there are more search results, then the component loops to step 430 to get the next search results, else the component continues at step 460. In step 460, the component provides the search results in response to the search query. For example, the user may receive the search results through a web page that lists the search results or the search results may be provided to another component for additional processing through an

application programming interface (API). The component may also perform additional processing of the search results before presenting the search results to the user. For example, the component may order the search results, rank them by retrieval time, and so forth. After step 460, these steps conclude.

[0040] Figures 5 illustrates some of the data structures used by the system. While the term "field" and "record" are used herein, any type of data structure can be employed. For example, relevant data can have preceding headers, or other overhead data preceding (or following) the relevant data. Alternatively, relevant data can avoid the use of any overhead data, such as headers, and simply be recognized by a certain byte or series of bytes within a serial data stream. Any number of data structures and types can be employed herein.

[0041] Figure 5 illustrates a data structure containing entries of the content index, in one embodiment. The offline content indexing system uses this and similar data structures to provide more intelligent content indexing. For example, the offline content indexing system may index multiple copies of data and data available from the multiple copies using a secondary copy of data stored on media with a higher availability based on the location or other attributes indicated by the data structure described below. As another example, the offline content indexing system may prefer an unencrypted copy of the data to an encrypted copy to avoid wasting time unnecessarily decrypting the data. The table 500 contains a location column 510, a keywords column 520, a user tags column 530, an application column 540, and an available column 550. The table 500 contains three sample entries. The first entry 560 specifies a location to a file on the corporate intranet using a web universal resource locator (URL). The entry 560 contains keywords "finance," "profit," and "loss" that identify content within the file. The entry 560 contains tags added by a user that specify that the content comes from the accounting department and is confidential. The entry 560 indicates that a spreadsheet program typically consumes the content, and that the entry is immediately available. Another entry 570 specifies data stored on a local tape that is a personal email, and can be available in about an hour. Another entry 580 specifies an offsite tape that is a presentation related to a cancelled project. The entry 580 refers to offsite data that is available within one week due to the delay of retrieving the archived data from the offsite location.

Conclusion

[0042] From the foregoing, it will be appreciated that specific embodiments of the offline content indexing system have been described herein for purposes of illustration, but that various modifications may be made without deviating from the spirit and scope of the invention. For example, web pages are often unavailable and their content may change such that the offline content indexing system could be used to retrieve point in time copies of the content useful for conducting historical analysis. As another example, although files have been described, other types of content such as user settings, application data, emails, and other data objects can all be indexed by the system. Accordingly, the invention is not limited except as by the appended claims.

[0043] Unless the context clearly requires otherwise, throughout the description and the claims, the words “comprise,” “comprising,” and the like are to be construed in an inclusive sense, as opposed to an exclusive or exhaustive sense; that is to say, in the sense of “including, but not limited to.” The word “coupled”, as generally used herein, refers to two or more elements that may be either directly connected, or connected by way of one or more intermediate elements. Additionally, the words “herein,” “above,” "below," and words of similar import, when used in this application, shall refer to this application as a whole and not to any particular portions of this application. Where the context permits, words in the above Detailed Description using the singular or plural number may also include the plural or singular number respectively. The word "or" in reference to a list of two or more items, that word covers all of the following interpretations of the word: any of the items in the list, all of the items in the list, and any combination of the items in the list.

[0044] The above detailed description of embodiments of the invention is not intended to be exhaustive or to limit the invention to the precise form disclosed above. While specific embodiments of, and examples for, the invention are described above for illustrative purposes, various equivalent modifications are possible within the scope of the invention, as those skilled in the relevant art will recognize. For example, while processes or blocks are presented in a given order, alternative embodiments may perform routines having steps, or employ systems having blocks, in a different order, and some processes or blocks may be deleted, moved, added, subdivided, combined, and/or modified. Each of these processes or

blocks may be implemented in a variety of different ways. Also, while processes or blocks are at times shown as being performed in series, these processes or blocks may instead be performed in parallel, or may be performed at different times.

[0045] The teachings of the invention provided herein can be applied to other systems, not necessarily the system described above. The elements and acts of the various embodiments described above can be combined to provide further embodiments.

[0046] These and other changes can be made to the invention in light of the above Detailed Description. While the above description details certain embodiments of the invention and describes the best mode contemplated, no matter how detailed the above appears in text, the invention can be practiced in many ways. Details of the system may vary considerably in implementation details, while still being encompassed by the invention disclosed herein. As noted above, particular terminology used when describing certain features or aspects of the invention should not be taken to imply that the terminology is being redefined herein to be restricted to any specific characteristics, features, or aspects of the invention with which that terminology is associated. In general, the terms used in the following claims should not be construed to limit the invention to the specific embodiments disclosed in the specification, unless the above Detailed Description section explicitly defines such terms. Accordingly, the actual scope of the invention encompasses not only the disclosed embodiments, but also all equivalent ways of practicing or implementing the invention under the claims.

[0047] While certain aspects of the invention are presented below in certain claim forms, the inventors contemplate the various aspects of the invention in any number of claim forms. For example, while only one aspect of the invention is recited as embodied in a computer-readable medium, other aspects may likewise be embodied in a computer-readable medium. Accordingly, the inventors reserve the right to add additional claims after filing the application to pursue such additional claim forms for other aspects of the invention.

CLAIMS

I/We claim:

1. In a data management system residing within a private computer network, a method for indexing content, comprising:

selecting an offline copy of the content from the private computer network, wherein the offline copy of the content is a copy of the content that is not a production copy of the content, and wherein the production copy is available from a live data server within the private computer network;

identifying at least some of the content within the offline copy; and

creating or updating a content index based on the identified content.

2. The method of claim 1 wherein selecting an offline copy comprises examining backup data.

3. The method of claim 1 wherein selecting an offline copy comprises examining a change journal.

4. The method of claim 1 wherein selecting an offline copy comprises examining a data snapshot.

5. The method of claim 1 wherein updating a content index comprises updating the content index in response to receiving a search request.

6. The method of claim 1 wherein updating a content index comprises updating the content index in response to an index update policy

7. The method of claim 1 further comprising before updating the content index, eliminating duplicate content within the selected offline copy.

8. The method of claim 1 wherein updating a content index comprises updating the content index incrementally based on incremental changes to the content.

9. A computer-readable medium containing instructions for controlling a computer system to identify archived content, by a method comprising:

receiving a search request, wherein the search request contains criteria for finding target content;

searching a content index for target content to create search results, wherein the content index contains information identifying at least one content item that is not available from mounted disk media or faster media;

for search results identifying content items that are not available from mounted disk media or faster media, retrieving the target content from an archive location; and

providing the search results in response to the search request.

10. The computer-readable medium of claim 9 wherein searching a content index comprises searching based on user-added attributes.

11. The computer-readable medium of claim 9 wherein searching a content index comprises receiving an availability criteria related to the content, searching based on information about the availability of the content, and generating search results indicating the time required to access the content.

12. The computer-readable medium of claim 9 wherein searching a content index comprises receiving a range of time during which deleted content was last available, searching based on a time the content was deleted, and generating search results for accessing the deleted content.

13. The computer-readable medium of claim 9 wherein searching a content index comprises searching based on a time range.

14. The computer-readable medium of claim 9 wherein searching a content index comprises searching a reference copy of the content.

15. The computer-readable medium of claim 9 further comprising after receiving the search request, creating the content index dynamically based on the content available in the system.

16. A computer system for indexing and searching content, comprising:

an offline copy component configured to select an offline copy of the content;

a content indexing component configured to create and update a content index based on the selected offline copy of the content; and

an index searching component configured to identify indexed content based on a received search query,

wherein the index of the content is created without consuming additional resources of a system that is the source of the content.

17. The system of claim 16 wherein the content indexing component decrypts encrypted content.

18. The system of claim 16 wherein the content indexing component updates the content index based on an indexing policy.

19. The system of claim 16 further comprising a data classification component configured to classify content and add classifications to the content index.

20. The system of claim 16 wherein the content indexing component selects a copy to use for indexing from among multiple offline copies of the data based on the time required to access each of the multiple offline copies.

21. In a data management system residing within a private computer network, a method for indexing content, comprising:

selecting an offline copy of the content from the private computer network, wherein the offline copy of the content is a copy of the content that is not a production copy of the content, and wherein the production copy is available from a live data server within the private computer network;

identifying at least some of the content within the offline copy; and

updating a content index by classifying the identified content based on attributes of the identified content.

22. The method of claim 21 wherein updating the content index comprises determining a state of protection of the identified content.

23. The method of claim 21 wherein updating the content index comprises determining whether the identified content is encrypted.

24. The method of claim 21 wherein updating the content index comprises determining whether the identified content has associated access control information.

25. The method of claim 21 wherein updating the content index comprises determining a topology of a network in which the identified content is stored.

26. The method of claim 21 wherein updating the content index comprises determining whether the identified content contains one or more specified keywords.

27. The method of claim 21 further comprising before updating the content index, eliminating duplicate content within the selected offline copy.

28. The method of claim 21 wherein updating a content index comprises updating the content index incrementally based on incremental changes to the content.

29. A computer-readable medium containing instructions for controlling a computer system to identify archived content, by a method comprising:

receiving a search request, wherein the search request contains classifications associated with target content;

searching a content index for target content to create search results, wherein the content index contains information identifying at least one content item that is not available from mounted disk media or faster media, wherein the faster media has a retrieval time or accessibility that is faster than mounted disk media;

for search results identifying content items that are not available from mounted disk media or faster media, retrieving information about the target content from an archive location; and

providing the search results in response to the search request.

30. The computer-readable medium of claim 29 wherein searching a content index comprises searching based on user-added attributes.

31. The computer-readable medium of claim 29 wherein searching a content index comprises receiving an availability criteria related to the content, searching based on information about the availability of the content, and generating search results indicating the time required to access the content.

32. The computer-readable medium of claim 29 wherein searching a content index comprises receiving a range of time during which deleted content was last available, searching based on a time the content was deleted, and generating search results for accessing the deleted content.

33. The computer-readable medium of claim 29 wherein searching a content index comprises searching based on a time range.

34. The computer-readable medium of claim 29 wherein searching a content index comprises searching a reference copy of the content.

35. The computer-readable medium of claim 29 further comprising after receiving the search request, creating the content index dynamically based on the content available in the system.

36. A computer system for indexing and searching content, comprising:

an offline copy component configured to select an offline copy of the content;

a content indexing component configured to store in a content index attributes of the content within the selected offline copy; and

an index searching component configured to identify indexed content based on a received search query and the attributes stored within the index,

wherein the index of the content is created without consuming additional resources of a system that is the original source of the content.

37. The system of claim 36 wherein the content indexing component decrypts encrypted content.

38. The system of claim 36 wherein the content indexing component updates the content index based on an indexing policy that specifies a schedule on which the content should be indexed.

39. The system of claim 36 further comprising a data classification component configured to classify content and add classifications to the content index.

40. The system of claim 36 wherein the content indexing component selects a copy to use for indexing from among multiple offline copies of the data based on the time required to access each of the multiple offline copies.

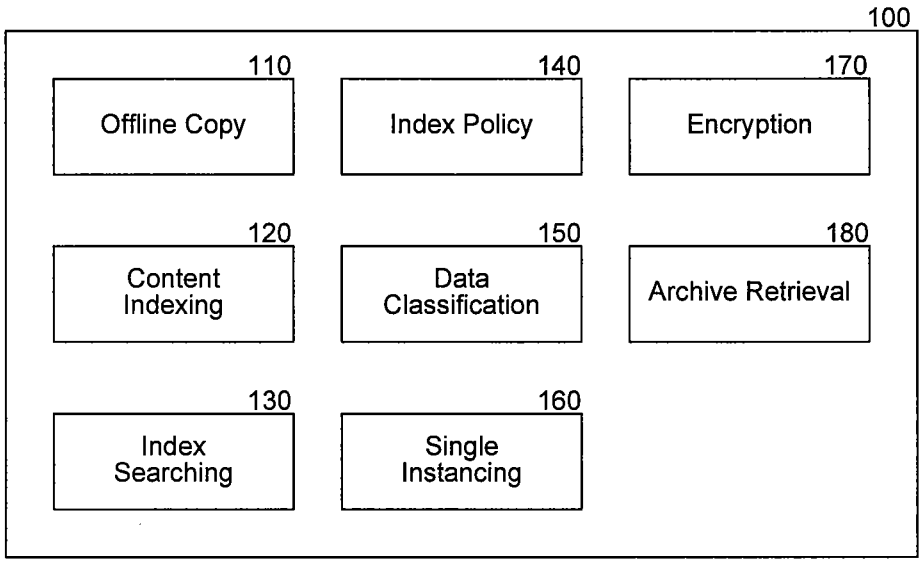
100
110
Offline Copy
140
Index Policy
170
Encryption
120
Content Indexing
150
Data Classification
180
Archive Retrieval
130
Index Searching
160
Single Instancing

FIG. 1

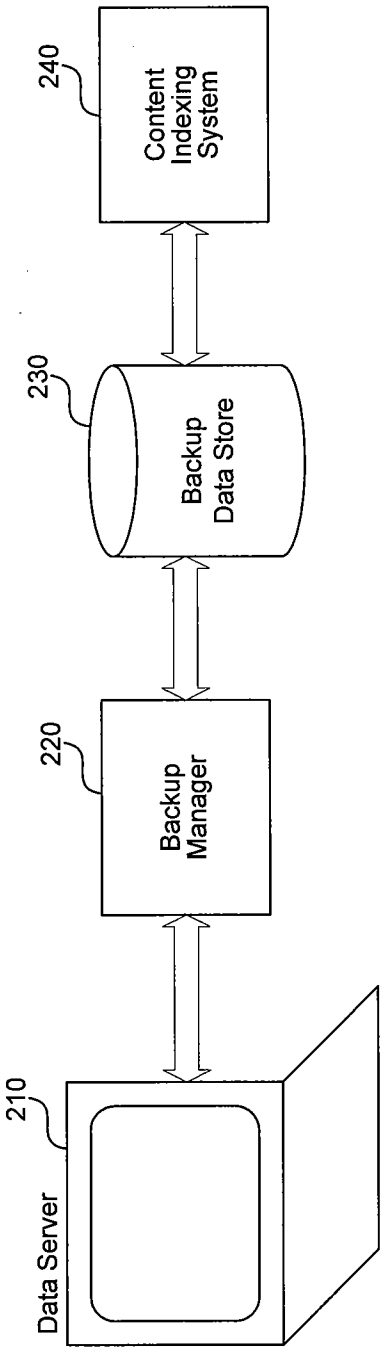
240
Content Indexing System
230
Backup Data Store
220
Backup Manager
210
Data Server

FIG. 2

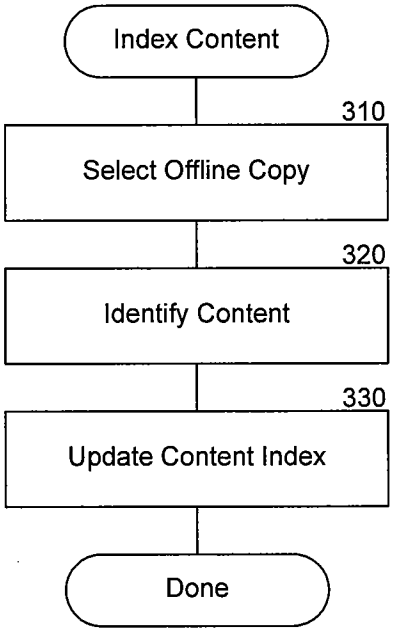
Index Content
310
Select Offline Copy
320
Identify Content
330
Update Content Index
Done

FIG. 3

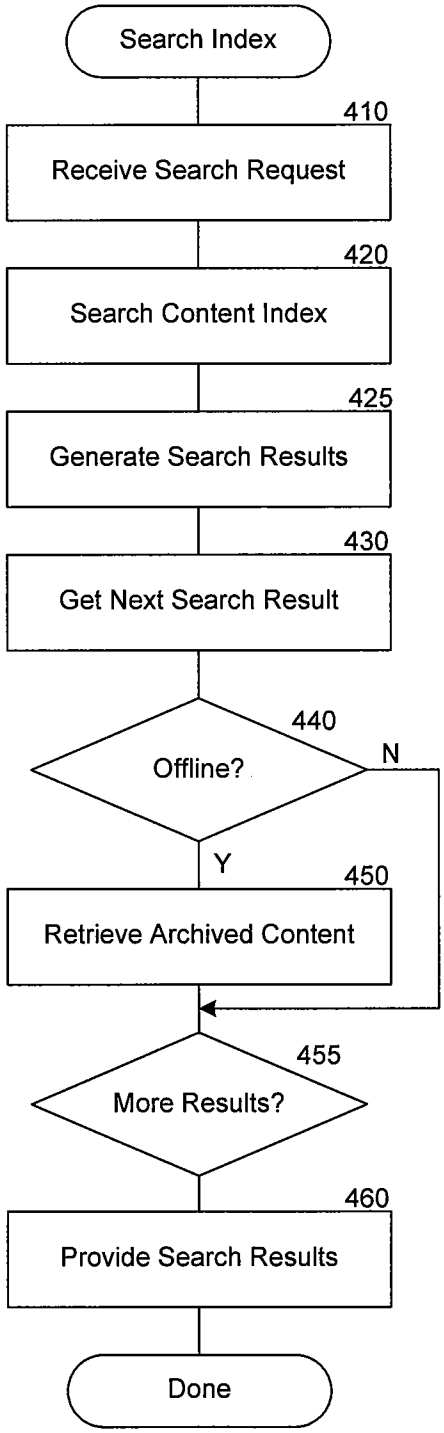
Search Index
410
Receive Search Request
420
Search Content Index
425
Generate Search Results
430
Get Next Search Result
440
Offline?
N
Y
450
Retrieve Archived Content
455
More Results?
460
Provide Search Results
Done

FIG. 4

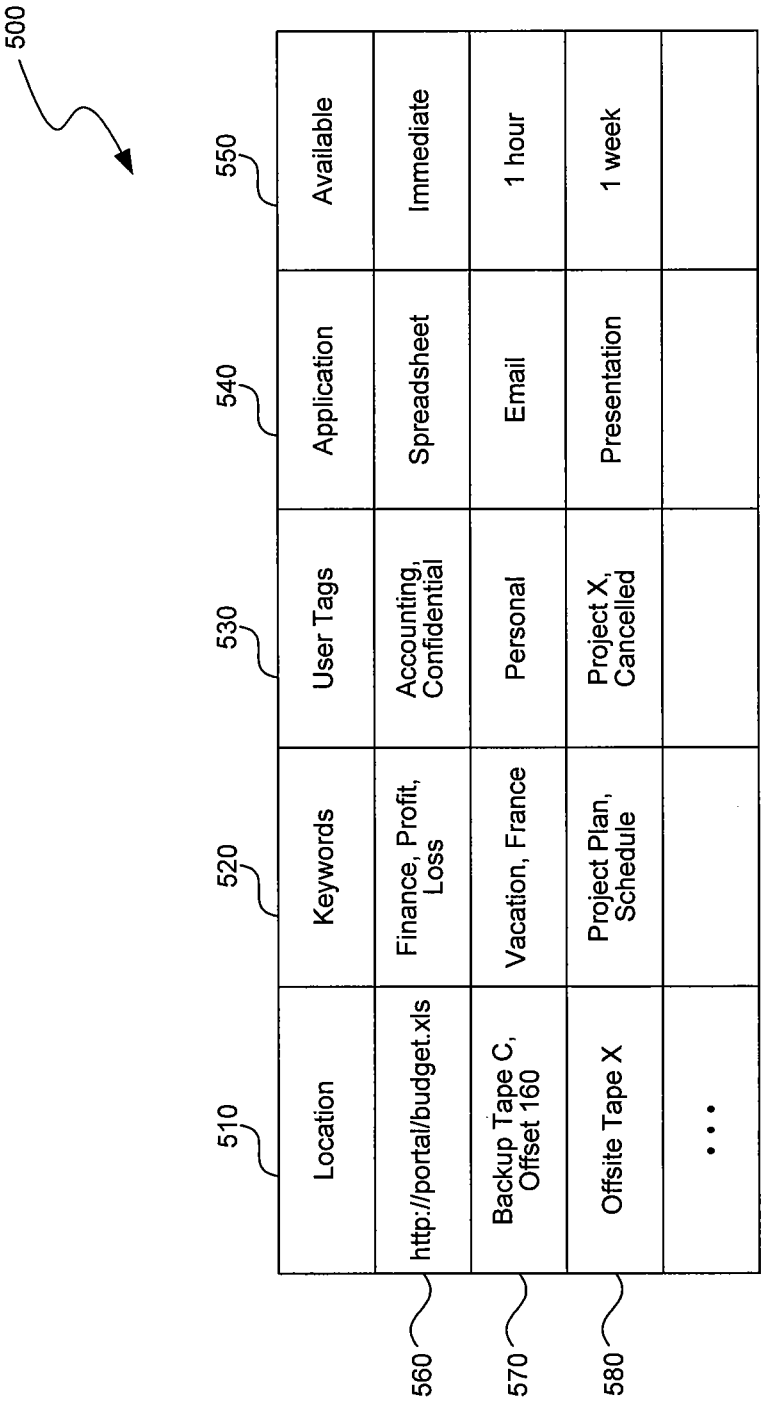
500
510
520
530
540
550
Location
Keywords
User Tags
Application
Available
http://portal/budget.xls
Finance, Profit, Loss
Accounting, Confidential
Spreadsheet
Immediate
Backup Tape C, Offset 160
Vacation, France
Personal
Email
1 hour
Offsite Tape X
Project Plan, Schedule
Project X, Cancelled
Presentation
1 week
560
570
580

FIG. 5