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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
12 September 2013 (12.09.2013)

WIPO | PCT

(10) International Publication Number
WO 2013/132476 A1

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

(21) International Application Number:
PCT/IL2012/000240

(22) International Filing Date:
19 June 2012 (19.06.2012)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
13/413,748 7 March 2012 (07.03.2012) US

(71) Applicant *(for all designated States except US)*: **VARONIS SYSTEMS, INC.** [US/US]; 1250 Broadway, 31st Floor, New York, New York 10001 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **FAITELSON, Yakov** [IL/IL]; 3 Mishol Hasapir Street, 44814 Herzeliya (IL). **KORKUS, Ohad** [IL/IL]; 11 Galgaley Haplada Street, 46733 Herzeliya (IL). **BASS, David** [IL/IL]; 26 Hatamar Street, 99797 Carmei Yoseph (IL). **KRETZER-KATZIR, Ophir** [IL/IL]; 23 Tomer Street, 71799 Reut (IL).

(74) Agents: **SANFORD T. COLB & CO.** et al.; P.O. Box 2273, 76122 Rehovot (IL).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, 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, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, 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, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

(54) **Title**: ENTERPRISE LEVEL DATA MANAGEMENT

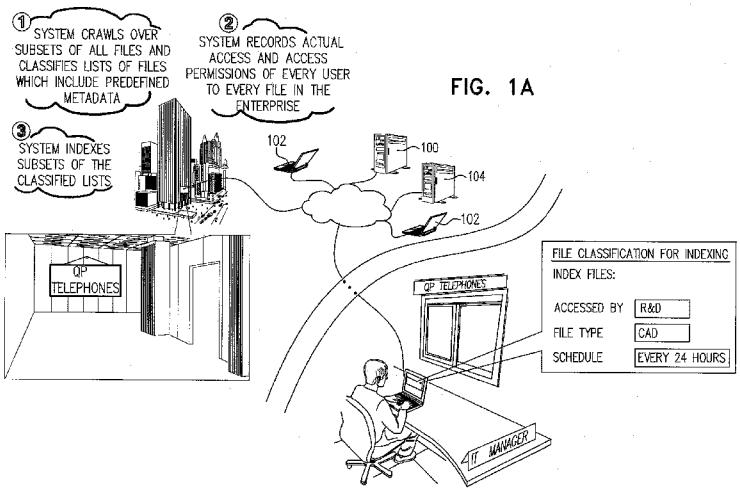


FIG. 1A

(57) **Abstract**: A system for indexing data of interest within a multiplicity of data elements residing on multiple platforms in an enterprise, the system including background data characterization functionality operable for characterizing the multiplicity of data elements to provide a background data characterization output, the characterizing being based on at least one of at least one access metric thereof, the at least one access metric being selected from data access permissions and actual data access history thereof and metadata thereof, background data classification functionality operative to classify the data of interest based at least partially on the background data characterization output, and providing a background data of interest classification output, and indexing functionality operative to index the data of interest based at least partially on the background data of interest classification output.

WO 2013/132476 A1

ENTERPRISE LEVEL DATA MANAGEMENT

REFERENCE TO RELATED APPLICATIONS

Reference is made to the following patents and patent applications, owned by assignee, the disclosures of which are hereby incorporated by reference, which are believed to relate to subject matter similar to the subject matter of the present application:

U.S. Patent Nos. 7,555,482 and 7,606,801;

U.S. Published Patent Application Nos. 2007/0244899, 2008/0271157, 2009/0100058, 2009/0265780 and 2009/0119298; and

U.S. Patent Application Nos. 12/498,675; 12/673,691 and 13/413,748.

FIELD OF THE INVENTION

The present invention relates to data management generally and more particularly enterprise level data management.

BACKGROUND OF THE INVENTION

The following patent publications and articles are believed to represent the current state of the art:

U.S. Patent Nos.: 7,031,984; 6,338,082; 6,928,439; 7,555,482; 7,606,801; 6,393,468; 5,899,991; 7,068,592 and 5,465,387.

U.S. Published Patent Application Nos.: 2003/0051026; 2004/0249847; 2004/0186809; 2005/0108206; 2005/0278334; 2005/0203881; 2005/0120054; 2005/0086529; 2006/0064313; 2006/0184530; 2006/0277184; 2006/0184459 and 2007/0203872.

SUMMARY OF THE INVENTION

The present invention provides improved systems and methodologies for data management.

There is thus provided in accordance with a preferred embodiment of the present invention a system for indexing data of interest within a multiplicity of data elements residing on multiple platforms in an enterprise, the system including background data characterization functionality operable for characterizing the multiplicity of data elements to provide a background data characterization output, the characterizing being based on at least one of at least one access metric thereof, the at least one access metric being selected from data access permissions and actual data access history thereof and metadata thereof, background data classification functionality operative to classify the data of interest based at least partially on the background data characterization output, and providing a background data of interest classification output, and indexing functionality operative to index the data of interest based at least partially on the background data of interest classification output.

Preferably, the characterizing the multiplicity of data elements is based on both at least one access metric thereof, the at least one access metric being selected from data access permissions and actual data access history thereof, and metadata thereof.

Preferably, the system also includes near real time data matching functionality operable for selecting data of interest by considering only data elements which have the at least one access metric thereof from among the classification output.

Preferably, the indexing functionality is operative to index the data of interest also based on the background data characterization output. Preferably, the at least one access metric is a dynamic metric which changes over time during operation of the enterprise.

Preferably, the near real time data matching functionality includes searching functionality operable to employ an output of the indexing functionality for searching for data elements which have the at least one content characteristic thereof, identification functionality operable for identifying data elements from among the

multiplicity of data elements in accordance with the at least one access metric, and combining functionality operable for combining results of the searching and the identifying.

Preferably, the searching functionality and the identifying functionality are provided by separate entities. Preferably, the metadata includes at least one of file size, file type and keywords.

There is also provided in accordance with another preferred embodiment of the present invention a method for indexing data of interest within a multiplicity of data elements residing on multiple platforms in an enterprise, the method including characterizing the multiplicity of data elements to provide a background data characterization output, the characterizing being based on at least one of at least one access metric thereof, the at least one access metric being selected from data access permissions and actual data access history and metadata thereof, classifying the data of interest based at least partially on the background data characterization output and providing a background data of interest classification output, and indexing the data of interest based at least partially on the background data of interest classification output.

Preferably, the characterizing the multiplicity of data elements is based on both at least one access metric thereof, the at least one access metric being selected from data access permissions and actual data access history, and metadata thereof.

Preferably, the method also includes selecting, in near real time, data of interest by considering only data elements which have the at least one access metric thereof from among the classification output.

Preferably, the indexing also includes indexing the data of interest based on the background data characterization output. Preferably, the at least one access metric is a dynamic metric which changes over time during operation of the enterprise.

Preferably, the selecting includes employing an output of the indexing for searching for data elements which have the at least one content characteristic thereof, identifying data elements from among the multiplicity of data elements in accordance with the at least one access metric, and combining results of the searching and the identifying.

Preferably, the searching and the identifying are performed by separate entities. Preferably, the metadata comprises at least one of file size, file type and keywords.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention will be understood and appreciated more fully from the following detailed description, taken in conjunction with the drawings in which:

Figs. 1A and 1B are simplified pictorial illustrations of an example of the operation of the system and methodology of the present invention;

Fig. 2 is a simplified block diagram illustration of the system and methodology of the embodiment of Figs. 1A & 1B; and

Fig. 3 is a simplified block diagram illustration of the use of the system and methodology of Figs. 1A – 2 for selecting data of interest from among a multiplicity of data elements by considering only data elements which are characterized by a given content classification, a given characteristic and a given access metric thereof.

DETAILED DESCRIPTION OF A PREFERRED EMBODIMENT

Reference is now made to Figs. 1A and 1B, which are simplified pictorial illustrations of an example of the operation of the system and methodology of the present invention. In the example of Figs. 1A and 1B, there is provided a system for indexing data of interest within a multiplicity of data elements residing on multiple platforms in an enterprise, the system preferably including:

background data characterization functionality operable for characterizing the multiplicity of data elements to provide a background data characterization output, the characterizing being based on at least one of:

at least one access metric thereof, the at least one access metric being selected from data access permissions and actual data access history thereof; and

metadata thereof;

background data classification functionality operative to classify the data of interest based at least partially on the background data characterization output, and providing a background data of interest classification output; and

indexing functionality operative to index the data of interest based at least partially on the background data of interest classification output.

As shown in Fig. 1A, the system preferably resides on a server 100 which is connected to an enterprise level network 102. Network 102 also preferably comprises a plurality of computers 102 and computer servers 104 connected thereto. Computers 102 and computer servers 104 may be located at disparate locations and are preferably operable for storing data elements, which are typically numbered in the thousands, hundreds of thousands or even millions.

Preferably, the system also utilizes a database 110 for storing information recorded thereby.

As illustrated pictorially in Fig. 1A, the system preferably operates in the background to record actual access and access permissions of every user to every data element in enterprise network 102. It is appreciated that the scope of activities of the system may be restricted to exclude certain users and certain data elements.

The system also preferably operates in the background to continuously crawl through subsets of all data elements in the enterprise and to classify lists of data elements of the subsets which data elements include predefined metadata. The metadata may include, for example, one or more specified keywords, file type and file size. A file type may be, for example, CAD files, text files and spreadsheets.

Preferably, the subsets of files are selected in accordance with access permission metrics. For example, a subset of files to which research and development personal have access permissions may be selected for classification as potentially being related to research and development.

The predefined metadata may be selected by an authorized manager as being appropriate for each subset. Thus, for example, for research and development files, keywords such as "design", "architecture" and "algorithm" may be appropriate for classifying the data elements as being related to research and development. Additionally, files which are CAD files may also be classified as being related to research and development.

It is a particular feature of this embodiment of the present invention that the system also preferably operates in the background to regularly index the classified lists of data elements. It is appreciated that indexing of the classified data elements is operative to allow rapid searching of the classified data elements for any particular string.

As shown in Fig. 1A, an IT administrator of the system residing on server 100 utilizes the system to classify a list of files which are to be regularly indexed. The IT administrator selects files to which research and development personal have access permissions, and which are of a CAD file type. The IT Administrator also preferably schedules the indexing to be performed every twenty-four hours.

Turning now to Fig. 1B, it is seen that at stage A, the CEO of a company notices a headline in a newspaper announcing the launch of the company's new confidential product. As shown at stage B, the CEO then contacts the IT manager of enterprise network 102 and demands to know how information regarding the confidential product reached the press. As shown at stage C, the IT Manager responds that all research and development related CAD files are regularly indexed and therefore

the file containing the information which reached the press can be rapidly located, and users who have recently accessed the file can be rapidly identified.

The IT Manager preferably proceeds to utilize the system to rapidly search all indexed files which were classified as research and development related files for references to the new confidential product. As shown at stage D, the IT Manager receives, in near real time, a list of relevant files. As thereafter shown at stage E, the IT manager can then utilize the actual access and access permissions information which is continuously collected by the system to determine which users have recently accesses the relevant confidential file, and to thereby ascertain which employee was responsible for providing the confidential information to the press.

It is a particular feature of the present invention that due to the background operation of the system whereby the history of actual access of every user to every file in the enterprise is recorded, classified lists of files which include predefined metadata are maintained and the classified lists are regularly indexed, the system enables the IT Manager to receive the results of his search for relevant files in near real time. The system achieves this near real time response by combining available actual access and access permissions information of classified lists of data elements with indexing information relating to the data elements.

Reference is now made to Fig. 2, which is a simplified block diagram illustration of the system and methodology of the embodiment of Figs. 1A & 1B. As seen in Fig. 2 and described hereinabove in Figs. 1A & 1B, the system and methodology of the present invention includes the following functionality which takes place in the background:

Actual access and access permissions of every user to every file in the enterprise is preferably continuously monitored and stored in a database. This functionality is embodied in a system, commercially available under the trademark DatAdvantage by an affiliate of the assignee of the present invention, Varonis Systems Inc. of New York, NY and is described in U.S. Patent 7,606,801 and in U.S. Published Patent Application 2009/0265780 of the assignee, the disclosures of which are hereby incorporated by reference. Access permissions and/or actual access are together designated as access metrics and may be used to designate subsets of all of the files in the enterprise.

Additionally, the system preferably continuously crawls through subsets of all files in the enterprise which are selected in accordance with the access metrics and classifies lists of files which include predefined metadata.

Additionally, the system preferably continuously operates in the background to regularly index at least part of the classified lists of data elements.

Reference is now made to Fig. 3, which is a simplified block diagram illustration of the use of the system and methodology of Figs. 1A - 2 for selecting data of interest from among a multiplicity of data elements by considering only data elements which are characterized by a given content classification, a given characteristic and a given access metric thereof.

As shown in Fig. 3, upon receipt of a query, which could, for example, include a request for a list of files of a particular classification which contain particular keywords and which have certain access metrics associated therewith, the system preferably combines indexing information relating to files of the particular classification with access metrics information such as that provided by the crawling functionality described hereinabove, to provide a response which indicates which files of the particular classification contain the particular keywords and have associated access metrics as specified in the query.

It will be appreciated by persons skilled in the art that the present invention is not limited by what has been particularly shown and described hereinabove. Rather the scope of the present invention includes both combinations and subcombinations of the various features described hereinabove as well as modifications thereof which would occur to persons skilled in the art upon reading the foregoing description and which are not in the prior art.

CLAIMS

1. A system for indexing data of interest within a multiplicity of data elements residing on multiple platforms in an enterprise, the system comprising:

background data characterization functionality operable for characterizing said multiplicity of data elements to provide a background data characterization output, said characterizing being based on at least one of:

at least one access metric thereof, said at least one access metric being selected from data access permissions and actual data access history thereof; and

metadata thereof;

background data classification functionality operative to classify said data of interest based at least partially on said background data characterization output, and providing a background data of interest classification output; and

indexing functionality operative to index said data of interest based at least partially on said background data of interest classification output.

2. A system for indexing data of interest within a multiplicity of data elements residing on multiple platforms in an enterprise according to claim 1 and wherein said characterizing said multiplicity of data elements is based on both:

at least one access metric thereof, said at least one access metric being selected from data access permissions and actual data access history thereof; and

metadata thereof.

3. A system for indexing data of interest within a multiplicity of data elements residing on multiple platforms in an enterprise according to claim 1 and also comprising near real time data matching functionality operable for selecting data of interest by considering only data elements which have said at least one access metric thereof from among said classification output.

4. A system for indexing data of interest within a multiplicity of data elements residing on multiple platforms in an enterprise according to claim 1 and

wherein said indexing functionality is operative to index said data of interest also based on said background data characterization output.

5. A system for indexing data of interest within a multiplicity of data elements according to claim 1 and wherein said at least one access metric is a dynamic metric which changes over time during operation of the enterprise.

6. A system for indexing data of interest within a multiplicity of data elements residing on multiple platforms in an enterprise according to claim 3 and wherein said near real time data matching functionality comprises:

searching functionality operable to employ an output of said indexing functionality for searching for data elements which have said at least one content characteristic thereof;

identification functionality operable for identifying data elements from among said multiplicity of data elements in accordance with said at least one access metric; and

combining functionality operable for combining results of said searching and said identifying.

7. A system for indexing data of interest within a multiplicity of data elements residing on multiple platforms in an enterprise according to claim 6 and wherein said searching functionality and said identifying functionality are provided by separate entities.

8. A system for indexing data of interest within a multiplicity of data elements residing on multiple platforms in an enterprise according to claim 1 and wherein said metadata comprises at least one of file size, file type and keywords.

9. A method for indexing data of interest within a multiplicity of data elements residing on multiple platforms in an enterprise, the method comprising:

characterizing said multiplicity of data elements to provide a background data characterization output, said characterizing being based on at least one of:

at least one access metric thereof, said at least one access metric being selected from data access permissions and actual data access history; and

metadata thereof;

classifying said data of interest based at least partially on said background data characterization output and providing a background data of interest classification output; and

indexing said data of interest based at least partially on said background data of interest classification output.

10. A method for indexing data of interest within a multiplicity of data elements residing on multiple platforms in an enterprise according to claim 9 and wherein said characterizing said multiplicity of data elements is based on both:

at least one access metric thereof, said at least one access metric being selected from data access permissions and actual data access history; and

metadata thereof.

11. A method for indexing data of interest within a multiplicity of data elements residing on multiple platforms in an enterprise according to claim 9 and also comprising selecting, in near real time, data of interest by considering only data elements which have said at least one access metric thereof from among said classification output.

12. A method for indexing data of interest within a multiplicity of data elements residing on multiple platforms in an enterprise according to claim 9 and wherein said indexing also comprises indexing said data of interest based on said background data characterization output.

13. A method for indexing data of interest within a multiplicity of data elements according to claim 9 and wherein said at least one access metric is a dynamic metric which changes over time during operation of the enterprise.

14. A method for indexing data of interest within a multiplicity of data elements residing on multiple platforms in an enterprise according to claim 11 and wherein said selecting comprises:

employing an output of said indexing for searching for data elements which have said at least one content characteristic thereof;

identifying data elements from among said multiplicity of data elements in accordance with said at least one access metric; and

combining results of said searching and said identifying.

15. A method for indexing data of interest within a multiplicity of data elements residing on multiple platforms in an enterprise according to claim 14 and wherein said searching and said identifying are performed by separate entities.

16. A method for indexing data of interest within a multiplicity of data elements residing on multiple platforms in an enterprise according to claim 9 and wherein said metadata comprises at least one of file size, file type and keywords.

FIG. 1A

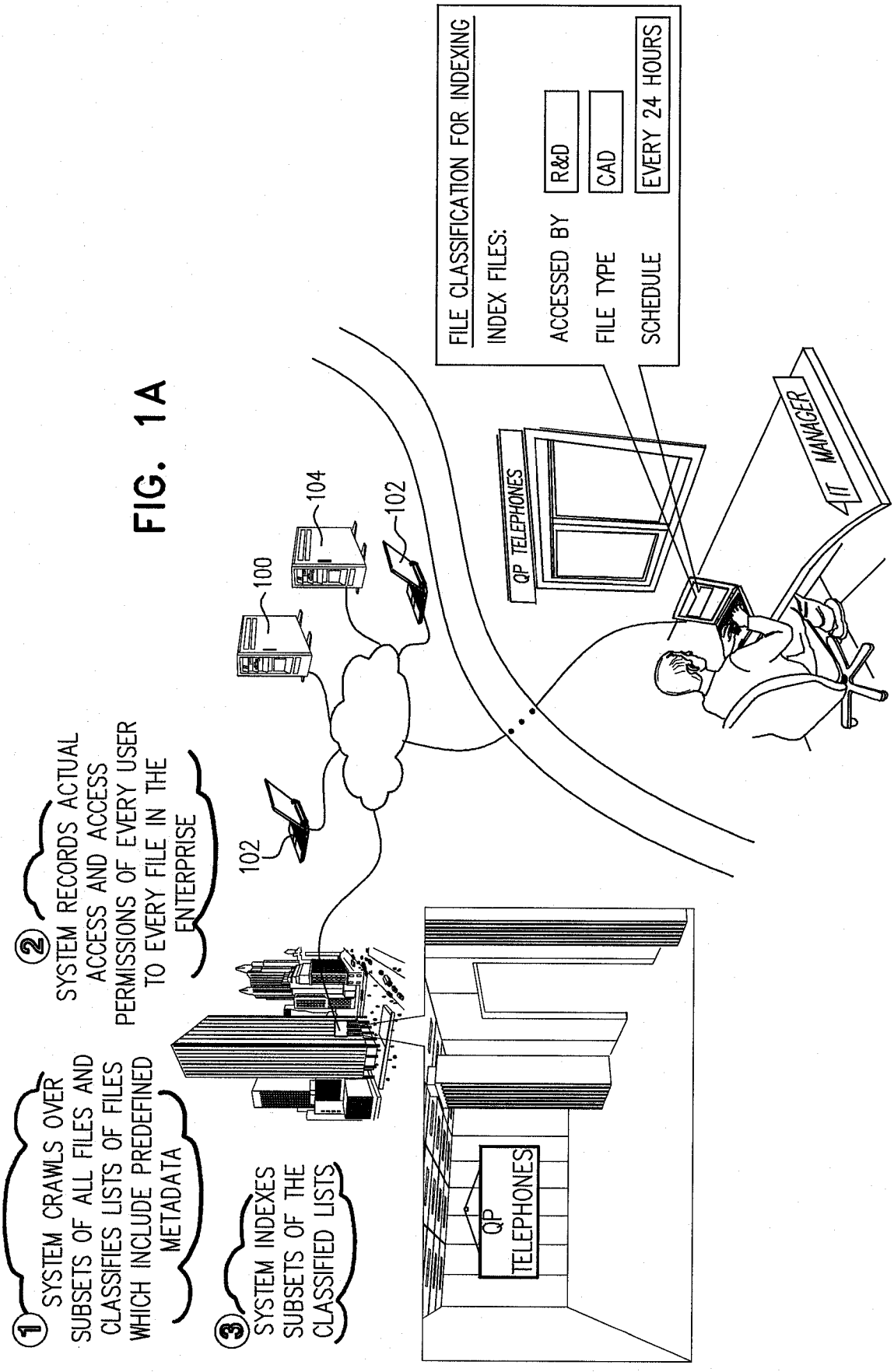
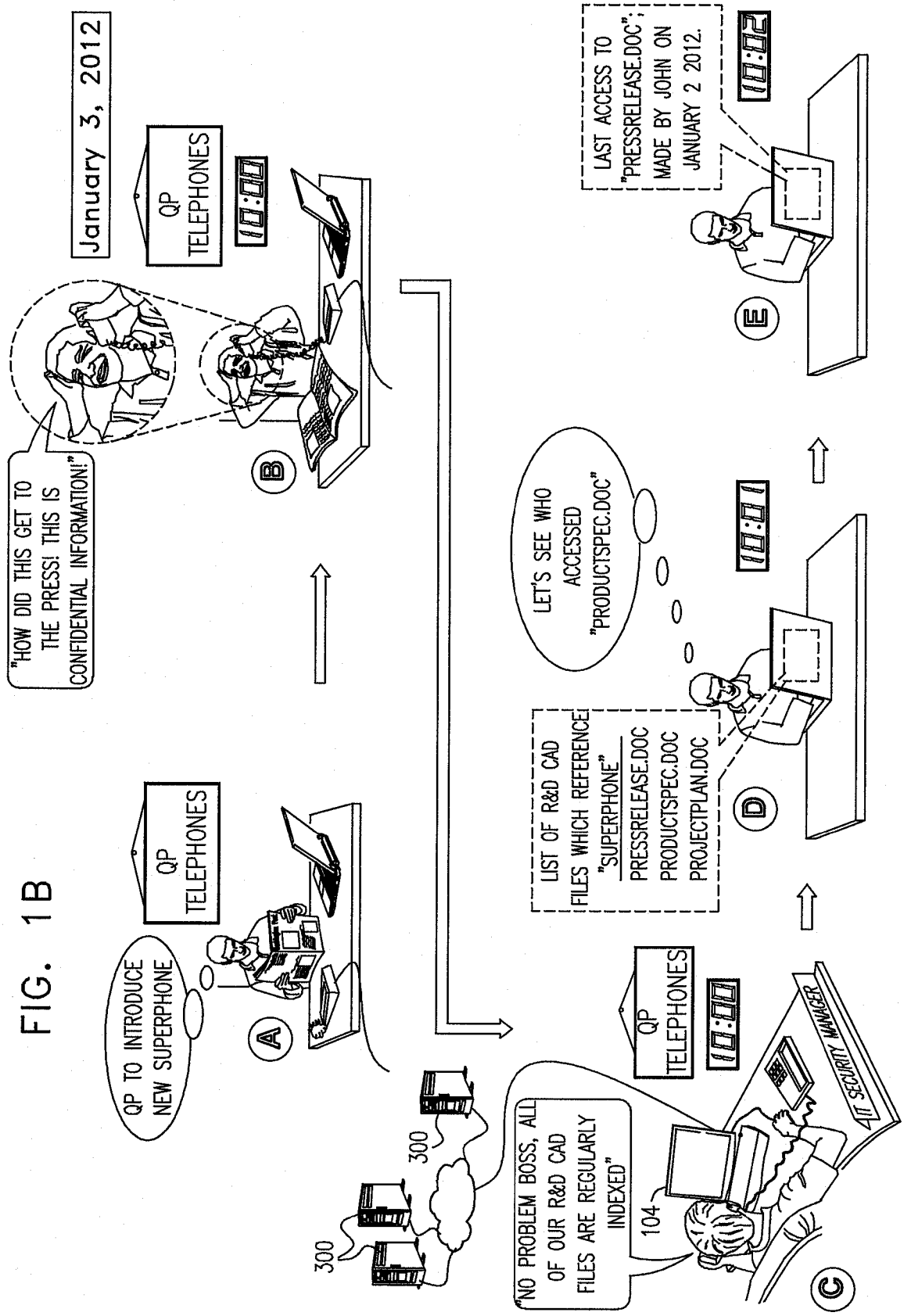

1 SYSTEM CRAWLS OVER SUBSETS OF ALL FILES AND CLASSIFIES LISTS OF FILES WHICH INCLUDE PREDEFINED METADATA
2 SYSTEM RECORDS ACTUAL ACCESS AND ACCESS PERMISSIONS OF EVERY USER TO EVERY FILE IN THE ENTERPRISE
3 SYSTEM INDEXES SUBSETS OF THE CLASSIFIED LISTS
QP TELEPHONES
100
102
102
104
QP TELEPHONES
IT MANAGER
FILE CLASSIFICATION FOR INDEXING
INDEX FILES:
ACCESSED BY R&D
FILE TYPE CAD
SCHEDULE EVERY 24 HOURS

FIG. 1B


January 3, 2012
QP TELEPHONES
10:00
"HOW DID THIS GET TO THE PRESS! THIS IS CONFIDENTIAL INFORMATION!"
B
QP TELEPHONES
QP TO INTRODUCE NEW SUPERPHONE
A
LAST ACCESS TO "PRESSRELEASE.DOC"; MADE BY JOHN ON JANUARY 2 2012.
10:02
E
LET'S SEE WHO ACCESSED "PRODUCTSPEC.DOC"
10:01
LIST OF R&D CAD FILES WHICH REFERENCE "SUPERPHONE"
PRESSRELEASE.DOC
PRODUCTSPEC.DOC
PROJECTPLAN.DOC
D
QP TELEPHONES
10:00
"NO PROBLEM BOSS, ALL OF OUR R&D CAD FILES ARE REGULARLY INDEXED"
104
IT SECURITY MANAGER
C
300
300

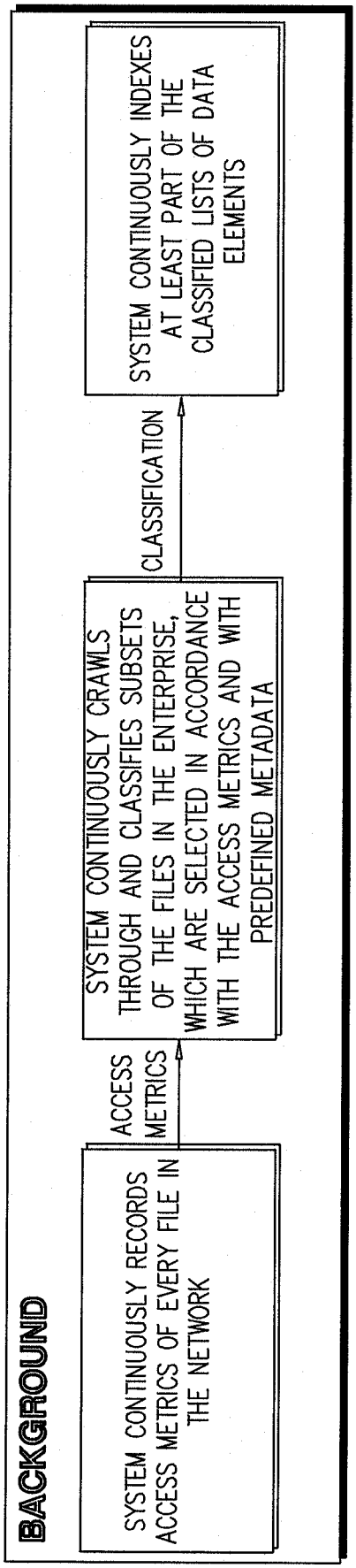

BACKGROUND
SYSTEM CONTINUOUSLY RECORDS ACCESS METRICS OF EVERY FILE IN THE NETWORK
ACCESS METRICS
SYSTEM CONTINUOUSLY CRAWLS THROUGH AND CLASSIFIES SUBSETS OF THE FILES IN THE ENTERPRISE, WHICH ARE SELECTED IN ACCORDANCE WITH THE ACCESS METRICS AND WITH PREDEFINED METADATA
CLASSIFICATION
SYSTEM CONTINUOUSLY INDEXES AT LEAST PART OF THE CLASSIFIED LISTS OF DATA ELEMENTS

FIG. 2

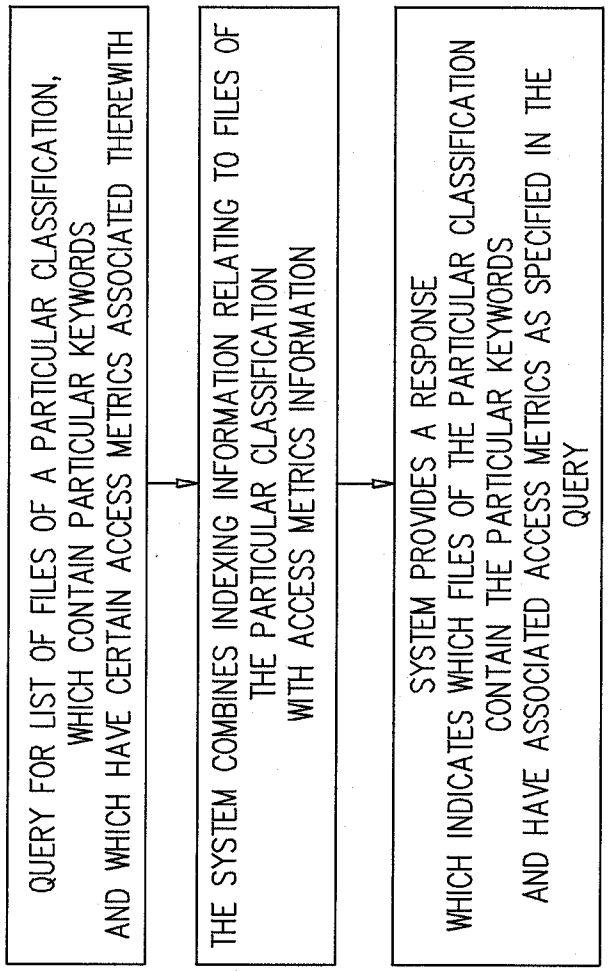

QUERY FOR LIST OF FILES OF A PARTICULAR CLASSIFICATION, WHICH CONTAIN PARTICULAR KEYWORDS AND WHICH HAVE CERTAIN ACCESS METRICS ASSOCIATED THEREWITH
THE SYSTEM COMBINES INDEXING INFORMATION RELATING TO FILES OF THE PARTICULAR CLASSIFICATION WITH ACCESS METRICS INFORMATION
SYSTEM PROVIDES A RESPONSE WHICH INDICATES WHICH FILES OF THE PARTICULAR CLASSIFICATION CONTAIN THE PARTICULAR KEYWORDS AND HAVE ASSOCIATED ACCESS METRICS AS SPECIFIED IN THE QUERY

FIG. 3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/IL2012/000240

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G06F 17/00 (2012.01)
USPC - 707/999.001
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC(8) - G06F 17/00, 17/30, 19/00 (2012.01)
USPC - 707/999.001, 999.003, 999.007

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
PatBase, Elsevier Inc Engineering Village: Compendex, Inspec, NTIS

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2009/0265780 A1 (KORKUS et al) 22 October 2009 (22.10.2009) entire document	1-16
Y	US 7,124,272 B1 (KENNEDY et al) 17 October 2006 (17.10.2006) entire document	1-16
A	US 2009/0182715 A1 (FALKENBERG) 16 July 2009 (16.07.2009) entire document	1-16
A	US 7,529,748 B2 (WEN et al) 05 May 2009 (05.05.2009) entire document	1-16

☐ Further documents are listed in the continuation of Box C. ☐

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
20 September 2012	01 OCT 2012
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Authorized officer: Blaine R. Copenheaver PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (July 2009)