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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
8 November 2007 (08.11.2007)

PCT

(10) International Publication Number
WO 2007/127360 A3

(51) International Patent Classification:
H03M 7/00 (2006.01)

(21) International Application Number:
PCT/US2007/010222

(22) International Filing Date: 26 April 2007 (26.04.2007)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
11/414,600 28 April 2006 (28.04.2006) US

(71) Applicant *(for all designated States except US)*: **NETWORK APPLIANCE, INC.** [US/US]; 495 East Java Drive, Sunnyvale, CA 94089 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **ZHENG, Ling** [CN/US]; 495 East Java Drive, Sunnyvale, CA 94089 (US). **STAGER, Roger** [US/US]; 495 East Java Drive, Sunnyvale, CA 94089 (US). **JOHNSTON, Craig** [US/US]; 495 East Java Drive, Sunnyvale, CA 94089 (US). **TRIMMER, Don** [US/US]; 495 East Java Drive, Sunnyvale, CA 94089 (US). **FRANDZEL, Yuval** [US/US]; 495 East Java Drive, Sunnyvale, CA 94089 (US).

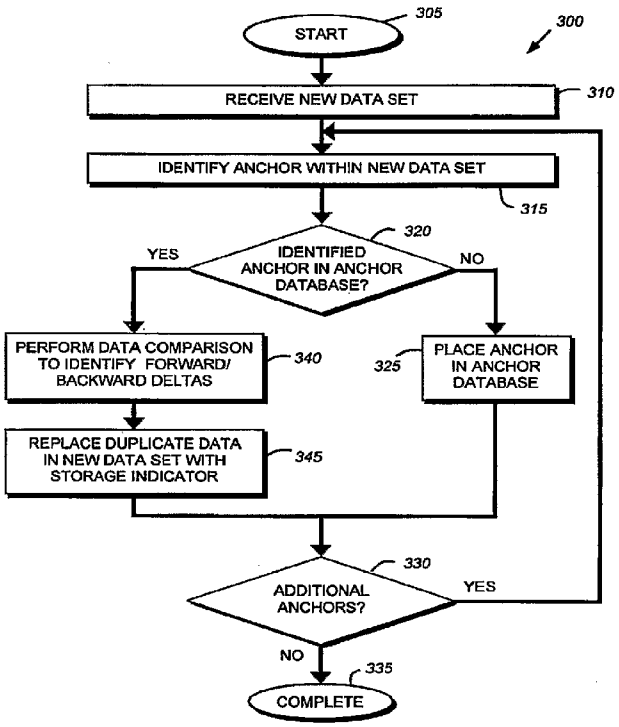
(74) Agent: **BARBAS, Charles, J.**; Cesari And Mckenna, LLP, 88 Black Falcon Avenue, Boston, MA 02210 (US).

(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, 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,

[Continued on next page]

(54) Title: SYSTEM AND METHOD FOR SAMPLING BASED ELIMINATION OF DUPLICATE DATA



(57) Abstract: A technique for eliminating duplicate data is provided. Upon receipt of a new data set, one or more anchor points are identified within the data set. A bit-by-bit data comparison is then performed of the region surrounding the anchor point in the received data set with the region surrounding an anchor point stored within a pattern database to identify forward/backward delta values. The duplicate data identified by the anchor point, forward and backward delta values is then replaced in the received data set with a storage indicator.

WO 2007/127360 A3

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:

— *with international search report*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

(88) Date of publication of the international search report:
5 June 2008

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2007/010222

A. CLASSIFICATION OF SUBJECT MATTER
INV. H03M7/00

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)
H03M

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 practical, search terms used)
EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	M.A. HERNANDEZ AND S.J. STOLFO: "Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem" DATA MINING AND KNOWLEDGE DISCOVERY, vol. 2, 1998, pages 1-31, XP002475116 the whole document	1-28
X	M. BILENKO AND R.J. MOONEY: "Adaptive duplicate detection using learnable string similarity measures" PROCEEDINGS NINTH ACM SIGKDD CONFERENCE, [Online] August 2003 (2003-08), pages 1-10, XP002475117 Washington DC Retrieved from the Internet: URL:http://citeseer.ist.psu.edu/bilenko03adaptive.html> [retrieved on 2008-04-04] the whole document	1-28

-/--

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* 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 document 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
4 April 2008	15/04/2008

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Van Staveren, Martin

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2007/010222

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	MONG LI LEE ET AL: "IntelliClean: a knowledge-based intelligent data cleaner" ACM INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, XX, XX, 2000, pages 290-294, XP002337104 the whole document	1-28
A	M. CROCHEMORE AND T. LECROQ: "Pattern Matching and Text Compression Algorithms"[Online] 8 January 2003 (2003-01-08), pages 1-49, XP002475118 Retrieved from the Internet: URL:http://citeseer.comp.nus.edu.sg/595025.html> [retrieved on 2008-04-04] the whole document	1,13,19, 28
A	REICHENBERGER C: "DELTA STORAGE FOR ARBITRARY NON-TEXT FILES" PROCEEDINGS OF THE 3RD INTERNATIONAL WORKSHOP ON SOFTWARE CONFIGURATION MANAGEMENT, 12-6-91, TRONDHEIM, NEW YORK, NY, US, 14 June 1991 (1991-06-14), pages 144-152, XP008030925 cited in the application the whole document	1-28
A	US 5 990 810 A (WILLIAMS ROSS NEIL [AU]) 23 November 1999 (1999-11-23) cited in the application the whole document	1-28
E	US 2007/255758 A1 (ZHENG LING [US] ET AL) 1 November 2007 (2007-11-01) the whole document	1-28

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2007/010222

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5990810 A	23-11-1999	WO 9625801 A1	22-08-1996
US 2007255758 A1	01-11-2007	WO 2007127360 A2	08-11-2007