

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
22 December 2005 (22.12.2005)

PCT

(10) International Publication Number
WO 2005/122004 A3

- (51) International Patent Classification:
G06F 15/00 (2006.01)
- (21) International Application Number:
PCT/US2005/017919
- (22) International Filing Date: 23 May 2005 (23.05.2005)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
10/859,937 3 June 2004 (03.06.2004) US
- (71) Applicant (for all designated States except US): **VERITY, INC.** [US/US]; 894 Ross Drive, Sunnyvale, CA 94089 (US).
- (72) Inventors: **MOTYKA, Daniel, Richard**; 5884 Bow Crescent NW, Calgary, AB T3B 2B7 (CA). **WALKER, Robert, Norman**; 676 Woodsworth RD S.E., Calgary, AB T2 1M8 (CA). **MAH, Marvin**; 608 Trafford Drive N.W., Calgary, AB T2K-2T3 (CA).
- (74) Agent: **BOTJER, William, L.**; P.O. Box 478, Center Moriches, NY 11934 (US).
- (81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM,

AT, AU, AZ, BA, BB, BG, 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, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

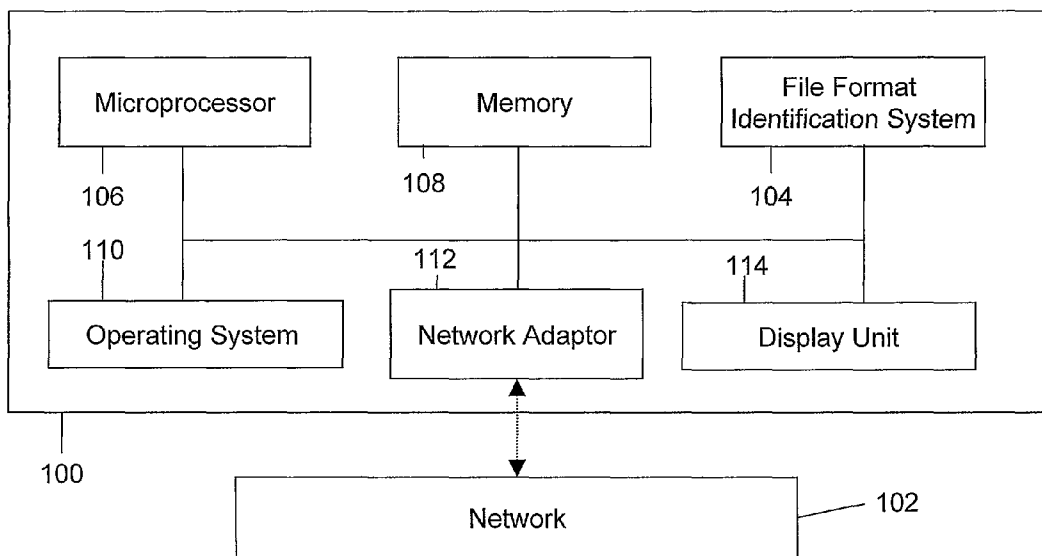
(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, MC, 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:
11 October 2007

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: CONTENT-BASED AUTOMATIC FILE FORMAT IDENTIFICATION



(57) Abstract: A method and system for content-based, automatic file format identification. The invention also relates to a method and system for dynamically selecting a set of bytes for byte-pattern recognition. The invention matches the pre-selected number of bytes of a file with the data signature of selected file formats. The file format information provided by the meta-data linked to the file acts as a filter that selects the file formats, which match the file information. If the attempt for file format identification, mentioned above, is unsuccessful, the invention computes the data type of the file, and subsequently identifies the corresponding text or binary file type. If a compound data type is computed, the invention identifies the file formats present in the compound file format.

WO 2005/122004 A3

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US05/17919

A. CLASSIFICATION OF SUBJECT MATTER

IPC: **G06F 15/00** (2006.01)

USPC: 715/513,515

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)

U.S. : 715/513,515

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)

Google, EAST, ACM, PLUS

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X, E --- Y	US PGPub 2006/0015630 (Stolfo et al) 19 January 2006 (19.1.2006), Paragraph 0007, 0066-0072, 0117, 0119-0122, 0125, 0126	1,5,14,16,17 ----- 2,3,6,7,15
Y	Fileext, "File Extension Details for .Zip", 17 March 2004 (17.03.2004), Retrieved from the Internet, http://web.archive.org/web/20040317233409/http://www.filext.com/detailist.php?extdetail=ZIP [retrieved on 5 May 2006], pg 1-3	2,6
Y	US 6,460,044 (Wang) 1 October 2002 (01.10.2002), Column 3, lines 1-18, 50-67	2,6
Y, P	US 6,785,867 (Shaffer et al) 31 August 2004 (31.08.2004), Column 5, lines 47-58; FIG 4	3, 7

Further documents are listed in the continuation of Box C.

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 application or patent 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

01 June 2007 (01.06.2007)

Date of mailing of the international search report

07 AUG 2007

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

Stephen Hong

Telephone No. 571-272-2100

**Jacqueline A. Whitfield
Special Project Asst.**