



## (51) International Patent Classification:

*G06F 21/24* (2006.01)      *G06F 15/16* (2006.01)  
*G06F 21/00* (2006.01)      *G06F 9/06* (2006.01)

## (21) International Application Number:

PCT/US2009/068687

## (22) International Filing Date:

18 December 2009 (18.12.2009)

## (25) Filing Language:

English

## (26) Publication Language:

English

## (30) Priority Data:

61/138,827 18 December 2008 (18.12.2008) US

(71) Applicant (for all designated States except US):  
**SUMOOH INC.** [US/US]; 27 Copley Drive, Northborough, MA 01532 (US).

## (72) Inventors; and

(75) Inventors/Applicants (for US only): **JAIN, Sanjay** [US/US]; 27 Copley Drive, Northborough, MA 01532 (US).  
**CHAUDHRY, Puneesh** [US/US]; 116 Clubhouse Lane, Northbridge, MA 01531 (US).(74) Agents: **BAKER, Hollie, L.** et al.; Wilmer Cutler Pickering Hale And Dorr LLP, 60 State Street, Boston, MA 02109 (US).

(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, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, 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, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, 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))

[Continued on next page]

## (54) Title: METHODS AND APPARATUS FOR CONTENT-AWARE DATA PARTITIONING AND DATA DE-DUPLICATION

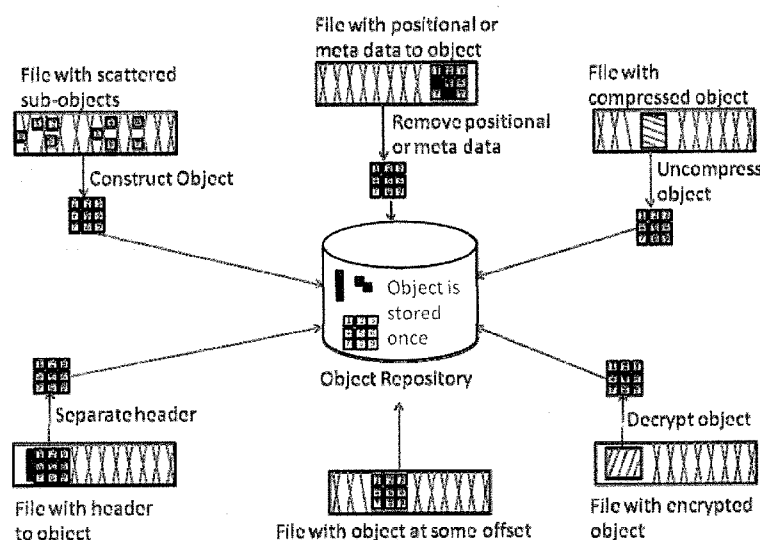


Figure 11

(57) **Abstract:** The systems and methods partition digital data units in a content aware fashion without relying on any ancestry information, which enables one to find duplicate chunks in unrelated units of digital data even across millions of documents spread across thousands of computer systems.



---

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

**(88) Date of publication of the international search report:**  
30 September 2010

## INTERNATIONAL SEARCH REPORT

International application No.  
**PCT/US2009/068687****A. CLASSIFICATION OF SUBJECT MATTER****G06F 21/24(2006.01)i, G06F 21/00(2006.01)i, G06F 15/16(2006.01)i, G06F 9/06(2006.01)i**

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)

IPC8 : G06F

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

Korean Utility models and applications for Utility models since 1975

Japanese Utility models and applications for Utility models since 1975

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS(KIPO internal) "Partition, identification, duplicate, computer, object, logic, storage, determine"

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 6959291 B1 (WILLIAM JOSEPH ARMSTRONG et al.) 25 October 2005 See abstract; column 3 line 65 - column 10 line 60; claims 1-18; figures 1-9.	1-52
A	US 5873104 A (MARC TREMBLAY et al.) 16 February 1999 See abstract; column 7 line 9 - column 25 line 62; claims 1-38; figures 1-17.	1-52
A	US 2008-0082525 A1 (MYUNG HOON CHA et al.) 3 April 2008 See abstract; paragraphs [0027]-[0068]; claims 1-21; figures 1-5.	1-52
A	US 6065046 A (MICHAEL A. FEINBERG et al.) 16 May 2000 See abstract; column 3 line 41 - column 47 line 14; claims 1-36; figures 1-29.	1-52



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 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 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

"&amp;" document member of the same patent family

Date of the actual completion of the international search

02 AUGUST 2010 (02.08.2010)

Date of mailing of the international search report

**03 AUGUST 2010 (03.08.2010)**

Name and mailing address of the ISA/KR

Korean Intellectual Property Office  
Government Complex-Daejeon, 139 Seonsa-ro, Seo-  
gu, Daejeon 302-701, Republic of Korea

Facsimile No. 82-42-472-7140

Authorized officer

Lee Byoung Soo

Telephone No. 82-42-481-5697



**INTERNATIONAL SEARCH REPORT**

Information on patent family members

International application No.

**PCT/US2009/068687**

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 6959291 B1	25.10.2005	JP 03954774 B2	08.08.2007
		JP 2000-357086 A	26.12.2000
		KR 10-2001-0014878 A	26.02.2001
		KR 10-0381503 B1	23.04.2003
US 5873104 A	16.02.1999	EP 0991998 A1	12.04.2000
		EP 0991998 B1	09.02.2005
		JP 03881702 B2	17.11.2006
		JP 2002-506550 A	26.02.2002
		KR 10-2001-0020495 A	15.03.2001
		WO 1999-000732 A1	07.01.1999
US 2008-0082525 A1	03.04.2008	KR 10-0834574 B1	02.06.2008
US 6065046 A	16.05.2000	AU 1998-86671 B2	21.09.2000
		CA 2297069 A1	11.02.1999
		CA 2297069 C	07.05.2002
		EP 1010193 A4	01.09.2004
		EP 1010193 A2	21.06.2000
		JP 2001-512272 A	21.08.2001
		JP 2009-009584 A	15.01.2009
		KR 10-0345460 B1	26.07.2002
		WO 1999-007007 A2	11.02.1999