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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
15 May 2008 (15.05.2008)

PCT

(10) International Publication Number
**WO 2008/057473 A3**

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

(21) **International Application Number:**
PCT/US2007/023233

**(22) International Filing Date:**
5 November 2007 (05.11.2007)

**(25) Filing Language:** English

**(26) Publication Language:** English

**(30) Priority Data:**
11/592,268 3 November 2006 (03.11.2006) US
11/644,009 22 December 2006 (22.12.2006) US

(71) **Applicant** *(for all designated States except US):* **GOOGLE INC.** [US/US]; 1600 Amphitheatre Parkway, Mountain View, CA 94043 (US).

**(72) Inventors; and**
**(75) Inventors/Applicants** *(for US only):* **FURMANIAK, Ralph** [CA/CA]; 43-683 Windermere Road, London, ON N5X-3T9 (CA). **SMITH, Ray** [GB/US]; c/o Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043 (US). **VINCENT, Luc** [US/US]; c/o Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043 (US). **BLOOMBERG, Dan** [US/US]; c/o Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043 (US).

**(74) Agents: MESSINGER, Michael, V.** et al; Sterne, Kessler, Golstein & Fox P.L.L.C, 1100 New York Avenue, N.W., Washington, DC 20005-3934 (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,

*[Continued on next page]*

**(54) Title:** MEDIA MATERIAL ANALYSIS OF CONTINUING ARTICLE PORTIONS

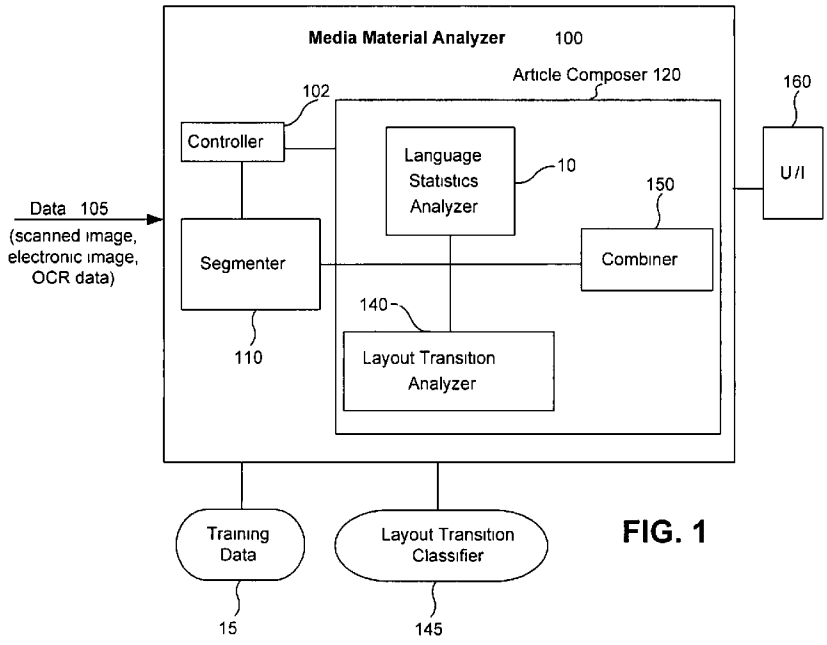


FIG. 1

**(57) Abstract:** The present invention relates to systems and methods for analyzing media material having articles continuing across multiple pages. A media material analyzer includes a segmenter and an article composer. The segmenter identifies block segments associated with columnar body text in the media material. The article composer determines which of the identified block segments belong to a continuing article extending across multiple pages in the media material based on language statistics information and continuation transition information.

WO 2008/057473 A3

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:**
24 July 2008

**INTERNATIONAL SEARCH REPORT**

International application No.
PCT/US 07/23233

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G06F 17/00 (2008.04)
**USPC - 715/255**
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)**
**USPC: 715/255**

**Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched**
**USPC: 715/200, 204, 243, 255, 264, 272**

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Electronic Databases Searched: USPTO WEST (PGPUB, USPAT, USOCR, EPAB, JPAB); GOOGLE SCHOLAR; DIALOG PRO
Search Terms Used: text/printed/media/digital articles/document extraction/metadata/segmentation comparison, multiple/successive column/page spanning, textual keyword/metadata/language/syntax/linguistic analysis, etc.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No |
|---|---|---|
| X | US 2006/0080309 A1 (YACOUB et al.) 13 April 2006 (13.04.2006)<br>Abstract, Para [0017]-[0141] | 1-18 |
| A | US 2006/0184525 A1 (JONES et al.) 17 August 2006 (17.08.2006)<br>Entire Document | 1-18 |
| A | US 2004/012281 1 A1 (PAGE) 24 June 2004 (24.06.2004)<br>Entire Document | 1-18 |
| A | US 2003/0229854 A1 (LEMAY) 11 December 2003 (11.12.2003)<br>Entire Document | **1-18** |

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

| | |
|---|---|
| * **Special categories of cited documents** | "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** |
| "A" **document defining the general state of the art which is not considered to be of particular relevance** | "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** |
| "E" **earlier application or patent but published on or after the international filing date** | "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** |
| "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)** | "&" **document member of the same patent family** |
| "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** | |

| | |
|---|---|
| Date of the actual completion of the international search<br>21 April 2008 (21.04.2008) | Date of mailing of the international search report<br>**14 JUL 2008** |
| Name and mailing address of the ISA/US<br>Mail Stop PCT, Attn: ISA/US, Commissioner for Patents<br>P.O. Box 1450, Alexandria, Virginia 22313-1450<br>Facsimile No. 571-273-3201 | Authorized officer.<br>Lee W. Young<br>PCT Helpdesk. 571-272-4300<br>PCTOSP- 571-272-7774 |

Form PCT/ISA/210 (second sheet) (April 2007)