PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : G06F 17/28, 17/21, 17/30 — A1

(11) International Publication Number: **WO 00/46694**

(43) International Publication Date: 10 August 2000 (10.08.00)

(21) International Application Number: PCT/US00/02747

(22) International Filing Date: 2 February 2000 (02.02.00)

(30) Priority Data:
09/243,744 3 February 1999 (03.02.99) US

(71) Applicant *(for BB BR CA CR CU DM GD LC MX only)*: QUARK, INC. [US/US]; 1800 Grant Street, Denver, CO 80203 (US).

(71) Applicant *(for all designated States except BB BR CA CR CU DM GD LC MX)*: QUARK MEDIA HOUSE SARL [CH/CH]; Puets–Godeet 6a, CH–2000 Neuchatel (CH).

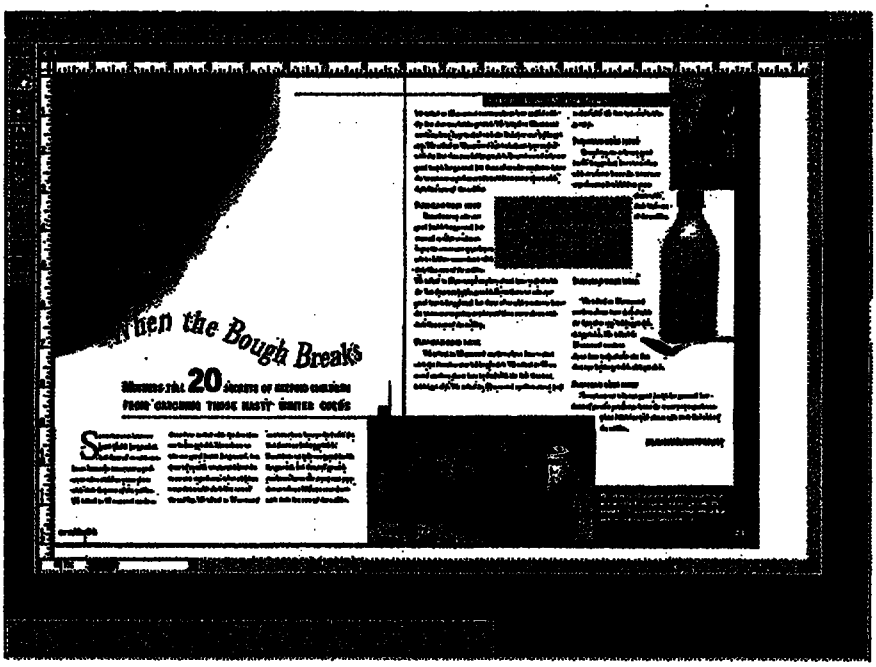
(72) Inventors: GILL, Timothy; 1800 Grant Street, Denver, CO 80203 (US). KNOSHAUG, David; 17105 E. Harvard Avenue, Aurora, CO 80013 (US). LIN, William; 1330 Miramonte Court, Broomfield, CO 80020 (US). NIES, Zachary; 1400 E. Cornell Avenue, Englewood, CO 80110 (US).

(74) Agent: WEBB, Glenn; P.O. Box 951, Conifer, CO 80433 (US).

(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
With international search report.

(54) Title: SYSTEM AND PROCESS FOR CREATING A STRUCTURED TAG REPRESENTATION OF A DOCUMENT



(57) Abstract

A system and process for extracting content from a document into a structural representation of the document. The system utilizes a heuristic process for analyzing a document based on user supplied hints or rules, extracting the content from the document and associating the extracted content with elements of structural document model based on the user supplied hints.