20000456-2.

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G06F 17/30 — A1

(11) International Publication Number: **WO 99/27469**

(43) International Publication Date: 3 June 1999 (03.06.99)

(21) International Application Number: PCT/GB98/03468

(22) International Filing Date: 18 November 1998 (18.11.98)

(30) Priority Data:
97309446.9 24 November 1997 (24.11.97) EP

(71) Applicant *(for all designated States except US)*: BRITISH TELECOMMUNICATIONS PUBLIC LIMITED COMPANY [GB/GB]; 81 Newgate Street, London EC1A 7AJ (GB).

(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: WEEKS, Richard [GB/GB]; 44 Glemsford Close, Felixstowe, Suffolk IP11 8UG (GB).

(74) Agent: DUTTON, Erica; BT Group Legal Services, Intellectual Property Dept., Holborn Centre, 8th floor, 120 Holborn, London EC1N 2TE (GB).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, 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: INFORMATION MANAGEMENT AND RETRIEVAL

(57) Abstract

A method and apparatus is provided for extracting key terms from a data set, the method including the steps of identifying a first set of one or more word groups of one or more words that occur more than once in the data set, and removing from this first set a second set of word groups that are sub-strings of longer word groups in the first set. The remaining word groups are key terms. Each word group is weighted according to its frequency of occurrence within the data set. The weighting of any word group may be increased by the frequency of any sub-string of words occurring in the second set and then dividing each weighting by the number of words in the word group. This weighting process operates to determine the order of occurrence of the word groups. Prefixes and suffixes are also removed from each word in the data set. This produces a neutral form of each word so that the weighting values are prefix and suffix independent.

