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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
12 May 2005 (12.05.2005)

PCT

(10) International Publication Number
WO 2005/043417 A3

(51) International Patent Classification[7]: G06F 17/30

(21) International Application Number:
PCT/US2004/036759

(22) International Filing Date:
3 November 2004 (03.11.2004)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/517,010 3 November 2003 (03.11.2003) US
10/877,735 24 June 2004 (24.06.2004) US

(71) Applicant *(for all designated States except US)*: **CLOUDMARK, INC.** [US/US]; 500 3rd Street, Suite 265, San Francisco, CA 94107 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **PRAKASH, Vipul, Ved** [IN/US]; 767 Bryant Street, #405, San Francisco, CA 94107 (US). **STEMM, Mark** [US/US]; 2301 Harrison, #301, San Francisco, CA 94110 (US).

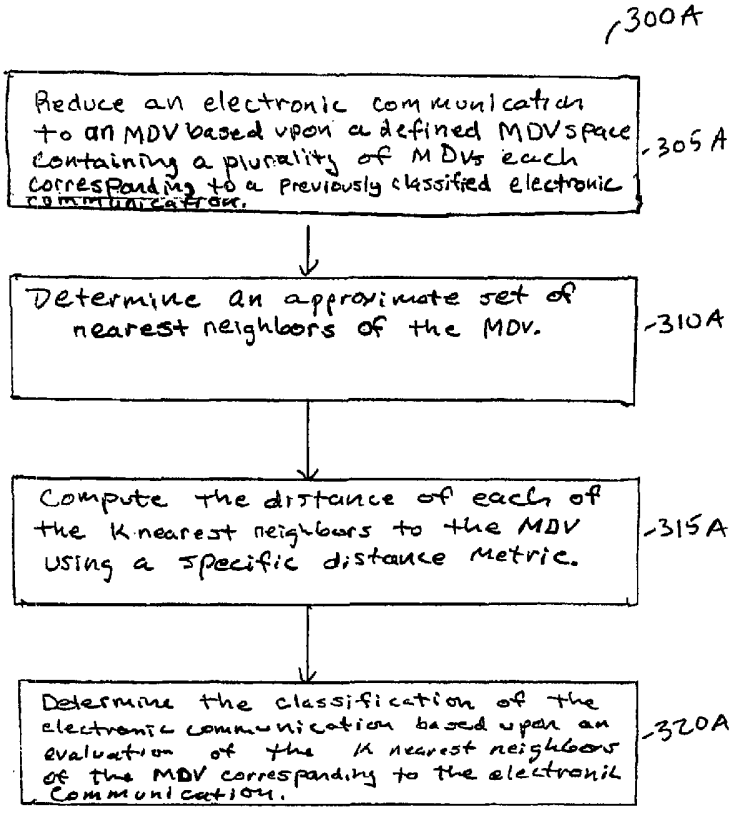
(74) Agents: **WARD, John, P.** et al.; Blakely, Sokoloff, Taylor & Zafman LLP, 12400 Wilshire Boulevard, 7th Floor, Los Angeles, CA 90025 (US).

(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, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, 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,

*[Continued on next page]*

(54) Title: METHODS AND APPARATUSES FOR CLASSIFYING ELECTRONIC DOCUMENTS




(57) Abstract: Embodiments of the invention provide methods and apparatuses for classifying electronic documents (e.g., electronic communications) as either spam electronic documents or legitimate electronic documents. In accordance with one embodiment of the invention, each of a plurality of electronic communications is reduced to a corresponding multidimensional vector based on a mufti-dimensional vector space. The mufti-dimensional vectors represent corresponding electronic documents that have been classified as at least one type of electronic documents. Subsequent electronic documents to be classified are reduced to a corresponding mufti-dimensional vector inserted into the mufti-dimensional vector space. The electronic documents corresponding to an inserted mufti-dimensional vector are classified based upon the proximity of the inserted mufti-dimensional vector to at least one previously classified mufti-dimensional vectors of the mufti-dimensional vector space.

WO 2005/043417 A3

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, 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*

**(88) Date of publication of the international search report:**
18 August 2005

*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.*

# INTERNATIONAL SEARCH REPORT

| | |
|---|---|
| | al Application No<br>PCT/US2004/036759 |

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06F17/30

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)
IPC 7 G06F

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 practical, search terms used)
EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category ° | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| | -/-- | |

[X] Further documents are listed in the continuation of box C.

[ ] Patent family members are listed in 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 document 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 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 | Date of mailing of the international search report |
|---|---|
| 3 March 2005 | 03.06.05 |

| Name and mailing address of the ISA | Authorized officer |
|---|---|
| European Patent Office, P.B. 5818 Patentlaan 2<br>NL - 2280 HV Rijswijk<br>Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,<br>Fax: (+31-70) 340-3016 | Hauck, R |

# INTERNATIONAL SEARCH REPORT

al Application No
PCT/US2004/036759

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

| Category ° | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| X | YONGHONG LI ET AL: "Classification of text documents" PATTERN RECOGNITION, 1998. PROCEEDINGS. FOURTEENTH INTERNATIONAL CONFERENCE ON BRISBANE, QLD., AUSTRALIA 16-20 AUG. 1998, LOS ALAMITOS, CA, USA,IEEE COMPUT. SOC, US, vol. 2, 16 August 1998 (1998-08-16), pages 1295-1297, XP010297856 ISBN: 0-8186-8512-3 | 1,31,61 |
| Y | page 1295, column 1, line 1 - page 1296, column 1, line 15 | 2-6, 17-20, 29,30, 32-36, 47-50, 59,60, 62-66, 77-80, 89,90 |
| | --- | |
| Y | DUDANI S A: "The Distance-Weighted k-Nearest-Neighbor Rule" IEEE TRANSACTIONS ON SYSTEMS, MAN AND CYBERNETICS, vol. SMC-6, no. 4, April 1976 (1976-04), pages 325-327, XP009044290 LOS ALAMITOS, CA, US<br>page 325, column 1, line 1 - page 326, column 2, line 17 | 2-6, 17-20, 29,30, 32-36, 47-50, 59,60, 62-66, 77-80, 89,90 |
| | --- | |
| A | MACLEOD J E S ET AL: "A Re-Examination of the Distance-Weighted k-Nearest Neighbor Classification Rule" IEEE TRANSACTIONS ON SYSTEM, MAN AND CYBERNETICS, vol. SMC-17, no. 4, August 1987 (1987-08), pages 689-696, XP009044291 the whole document | |
| | --- | |

# INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2004/036759

## Box II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This International Search Report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the International Application that do not comply with the prescribed requirements to such an extent that no meaningful International Search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

## Box III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

see additional sheet

1. ☐ As all required additional search fees were timely paid by the applicant, this International Search Report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this International Search Report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this International Search Report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

1-6, 17-20, 29-36, 47-50, 59-66, 77-80, 89, 90

**Remark on Protest**

☐ The additional search fees were accompanied by the applicant's protest.

☐ No protest accompanied the payment of additional search fees.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

This International Searching Authority found multiple (groups of) inventions in this international application, as follows:

1. claims: 1-6,17-20,29-36,47-50,59-66,77-80,89,90

A method for classifying an electronic document in a multi-dimensional vector space based on a proximity of the corresponding multi-dimensional vector to each of a set of nearest neighbor vectors,
details of the disctance metric and the voting mechanism used.

---

2. claims: 11,12,41,42,71,72

A method for classifying an electronic document in a multi-dimensional vector space based on a proximity of the corresponding multi-dimensional vector to each of a set of nearest neighbor vectors,
termination conditions for the nearest neighbor search.

---

3. claims: 13,43,73

A method for classifying an electronic document in a multi-dimensional vector space based on a proximity of the corresponding multi-dimensional vector to each of a set of nearest neighbor vectors,
selecting a number of leader vectors, determining a most proximate leader vector to a given vector, designating the vectors associated to the leader vector as nearest neighbors.

---

4. claims: 14-16,44-46,74-76

A method for classifying an electronic document in a multi-dimensional vector space based on a proximity of the corresponding multi-dimensional vector to each of a set of nearest neighbor vectors,
wherein the vector space grows dynamically, either by adding vectors or by adding dimensions.

---

5. claims: 21-28,51-58,81-88

A method for classifying an electronic document in a multi-dimensional vector space based on a proximity of the corresponding multi-dimensional vector to each of a set of nearest neighbor vectors,
details of selecting and weighting features of the multi-dimensional vectors.

---

6. claims: 7-10,37-40,67-70

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

A method for classifying an electronic document in a multi-dimensional vector space based on a proximity of the corresponding multi-dimensional vector to each of a set of nearest neighbor vectors,
use and details of a fallback mechanism.

---