

(19)



Europäisches Patentamt  
European Patent Office  
Office européen des brevets



(11)

**EP 0 889 417 A3**

(12)

**EUROPEAN PATENT APPLICATION**

(88) Date of publication A3:  
**24.11.1999 Bulletin 1999/47**

(51) Int Cl.6: **G06F 17/27, G06F 17/28**

(43) Date of publication A2:  
**07.01.1999 Bulletin 1999/01**

(21) Application number: **98305231.7**

(22) Date of filing: **01.07.1998**

(84) Designated Contracting States:  
**AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU  
MC NL PT SE**  
Designated Extension States:  
**AL LT LV MK RO SI**

- **Pedersen, Jan O.**  
**Fremont, CA 94555 (US)**
- **Schuetze, Hinrich**  
**Stanford, CA 94305 (US)**
- **Kessler, Brett L.**  
**Menlo Park, CA 94025 (US)**
- **Grefenstette, Gregory**  
**38400 St. Martin d'Herès (FR)**

(30) Priority: **02.07.1997 US 51558 P**

(71) Applicant: **XEROX CORPORATION**  
**Rochester, New York 14644 (US)**

(74) Representative: **Skone James, Robert Edmund**  
**GILL JENNINGS & EVERY**  
**Broadgate House**  
**7 Eldon Street**  
**London EC2M 7LH (GB)**

(72) Inventors:  
• **Nunberg, Geoffrey D.**  
**San Francisco, CA 94114 (US)**

(54) **Text genre identification**

(57) A processor implemented method of identifying the genre of a machine readable, untagged text. The processor implemented method begins by generating a cue vector from the text, which represents occurrences

in the text of a first set of nonstructural, surface cues, which are easily computable. Afterward, the processor determines whether the text is an instance of a first text genre using the cue vector and a weighting vector associated with the first text genre.

**EP 0 889 417 A3**



European Patent  
Office

EUROPEAN SEARCH REPORT

Application Number  
EP 98 30 5231

DOCUMENTS CONSIDERED TO BE RELEVANT			
Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.6)
A	EP 0 672 988 A (INTERNATIONAL BUSINESS MACHINES CORPORATION) 20 September 1995 (1995-09-20) * the whole document * ----	1-10	G06F17/27 G06F17/28
A	US 5 371 807 A (REGISTER ET AL) 6 December 1994 (1994-12-06) * the whole document * ----	1-10	
A	US 5 182 708 A (EJIRI) 26 January 1993 (1993-01-26) * the whole document * ----	1-10	
A	KJELL ET AL: "Visualisation of Literary Style" 1992 IEEE INTERNATIONAL CONFERENCE ON SYSTEMS, MAN, AND CYBERNETICS, vol. 1, 18 - 21 October 1992, pages 656-661, XP000366556 Chicago, IL, US * the whole document * -----	1-10	
			TECHNICAL FIELDS SEARCHED (Int.Cl.6)
			G06F
The present search report has been drawn up for all claims			
Place of search BERLIN		Date of completion of the search 5 October 1999	Examiner Abram, R
CATEGORY OF CITED DOCUMENTS		T : theory or principle underlying the invention E : earlier patent document, but published on, or after the filing date D : document cited in the application L : document cited for other reasons ..... & : member of the same patent family, corresponding document	
X : particularly relevant if taken alone Y : particularly relevant if combined with another document of the same category A : technological background O : non-written disclosure P : intermediate document			

EPO FORM 1503 03/82 (P04C01)

**ANNEX TO THE EUROPEAN SEARCH REPORT  
ON EUROPEAN PATENT APPLICATION NO.**

EP 98 30 5231

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report. The members are as contained in the European Patent Office EDP file on  
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

05-10-1999

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP 672988 A	20-09-1995	US 5548507 A JP 2837364 B JP 7262188 A	20-08-1996 16-12-1998 13-10-1995
US 5371807 A	06-12-1994	NONE	
US 5182708 A	26-01-1993	NONE	

EPO FORM P0459

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82