Title (en)

METHOD AND SYSTEM FOR CATEGORIZING ARABIC TEXT

Title (de)

METHODE UND SYSTEM ZUR KATEGORISIERUNG VON ARABISCHEM TEXT

Title (fr)

PROCEDE ET SYSTEME DE CATEGORISATION D'UN TEXTE ARABE

Publication

EP 1652107 A1 20060503 (EN)

Application

EP 04740317 A 20040513

Priority

- EP 2004006906 W 20040513
- EP 03368072 A 20030723
- EP 04740317 A 20040513

Abstract (en)

[origin: WO2005015434A1] The present invention is directed to a system, method and computer program for categorizing Arabic documents based on the text content. More particularly, the invention is a frequency based method using a learning approach that exploits, Arabic lexical look-up, Arabic morphological analysis, and a number of interconnected Arabic linguistic filters, to categorize Arabic texts. The present Arabic text categorization method comprises two phases namely: the learning phase, and the automatic categorization phase. During the learning phase, lemma forms (called stems) of specific noun types are extracted from manually categorized Arabic texts and then filtered, using Arabic morphological analysis. Based on these lemma forms and on the normalized frequency of these lemma forms for each predefined category, it is possible to automatically assign new Arabic texts to predefined categories during the automatic text categorization phase. As a result, categorization of Arabic texts is more precise and less sensitive to noise than prior art solutions. The present invention relates to a method for automatically assigning Arabic texts to predefined categories supporting information retrieval. For example, the method can be used to filter out Arabic documents that are unlikely to contain extractable data and can be used to route Arabic texts to processing mechanisms that are category specific.

IPC 1-7

G06F 17/30

IPC 8 full level

G06F 17/30 (2006.01)

CPC (source: EP)

G06F 16/353 (2018.12)

Citation (search report)

See references of WO 2005015434A1

Designated contracting state (EPC)

AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HU IE IT LI LU MC NL PL PT RO SE SI SK TR

DOCDB simple family (publication)

WO 2005015434 A1 20050217; EP 1652107 A1 20060503; IL 173306 A0 20060611

DOCDB simple family (application)

EP 2004006906 W 20040513; EP 04740317 A 20040513; IL 17330606 A 20060123