

Title (en)

SYSTEM AND METHOD FOR DETECTING TEXT SIMILARITY OVER SHORT PASSAGES

Title (de)

SYSTEM UND VERFAHREN UM TEKSTÄHNLICHKEITEN ÜBER KURZE PASSAGEN ZU ENTDECKEN

Title (fr)

SYSTEME ET PROCEDE DE DETECTION DE SIMILARITE DE TEXTE SUR DE COURTS PASSAGES

Publication

**EP 1203309 A1 20020508 (EN)**

Application

**EP 00951059 A 20000619**

Priority

- US 0040238 W 20000619
- US 13993099 P 19990618

Abstract (en)

[origin: WO0079426A1] A system and method are provided for determining similarity in short text segments. The method provides a definition of similarity which is appropriate for the small text setting (100). Small text segments are compared to determine if there exist common primitive features, such as words, noun phrases, synonyms, verbs with a common semantic class, proper nouns and the like (105). From the primitive features identified, the small text segments are evaluated to determine whether composite features are present (110). Composite features are defined as predetermined relationships between primitive features. The common primitive features and composite features are applied as inputs to an appropriate machine learning algorithm which is trained to ascertain a similarity measure based on the primitive and composite features common to the text segments (115).

IPC 1-7

**G06F 17/21**

IPC 8 full level

**G06F 17/21** (2006.01)

CPC (source: EP)

**G06F 40/10** (2020.01)

Designated contracting state (EPC)

AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE

DOCDB simple family (publication)

**WO 0079426 A1 20001228**; EP 1203309 A1 20020508; EP 1203309 A4 20060621

DOCDB simple family (application)

**US 0040238 W 20000619**; EP 00951059 A 20000619