Title (en)

METHOD AND SYSTEM FOR AUTOMATICALLY EXTRACTING DATA FROM WEB SITES

Title (de)

VERFAHREN UND SYSTEM ZUM AUTOMATISCHEN EXTRAHIEREN VON DATEN AUS WEBSITES

Title (fr)

METHODE ET SYSTEME POUR EXTRAIRE AUTOMATIQUEMENT DES DONNEES A PARTIR DE SITES WEB

Publication

EP 1910918 A2 20080416 (EN)

Application

EP 06787271 A 20060714

Priority

- US 2006027335 W 20060714
- US 69951905 P 20050715

Abstract (en)

[origin: WO2007011714A2] In accordance with an embodiment, data may be automatically extracted from semi-structured web sites. Unsupervised learning may be used to analyze web sites and discover their structure. One method utilizes a set of heterogeneous "experts," each expert being capable of identifying certain types of generic structure. Each expert represents its discoveries as "hints." Based on these hints, the system may cluster the pages and text segments and identify semi-structured data that can be extracted. To identify a good clustering, a probabilistic model of the hint-generation process may be used.

IPC 8 full level

G06F 7/00 (2006.01)

CPC (source: EP)

G06F 16/951 (2018.12); **G06F 40/30** (2020.01)

Citation (search report)

See references of WO 2007011714A2

Cited by

CN111625719A

Designated contracting state (EPC)

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

Designated extension state (EPC)

AL BA HR MK RS

DOCDB simple family (publication)

WO 2007011714 A2 20070125; **WO 2007011714 A3 20071004**; **WO 2007011714 A9 20070308**; CA 2614774 A1 20070125; EP 1910918 A2 20080416

DOCDB simple family (application)

US 2006027335 W 20060714; CA 2614774 A 20060714; EP 06787271 A 20060714