

Title (en)

IDENTIFYING SIMILARITIES AND HISTORY OF MODIFICATION WITHIN LARGE COLLECTIONS OF UNSTRUCTURED DATA

Title (de)

IDENTIFIZIEREN VON ÄHNLICHKEITEN UND VORGESCHICHTE DER MODIFIKATION IN GROSSEN SAMMLUNGEN UNSTRUKTURIERTER DATEN

Title (fr)

IDENTIFICATION DE SIMILARITES ET D'HISTORIQUE DE MODIFICATION DANS DE GRANDES COLLECTIONS DE DONNEES NON STRUCTUREES

Publication

**EP 1590748 A2 20051102 (EN)**

Application

**EP 04704049 A 20040121**

Priority

- US 2004001530 W 20040121
- US 44246403 P 20030123
- US 73891903 A 20031217
- US 73892403 A 20031217

Abstract (en)

[origin: WO2004066086A2] A technique for efficient representation of dependencies between electronically-stored documents, such as in an enterprise data processing system. A document distribution path is developed as a directional graph that is a representation of the historic dependencies between documents, which is constructed in real time as documents are created. The system preferably maintains a lossy hierarchical representation of the documents indexed in such a way that allows for fast queries for similar but not necessarily equivalent documents. A distribution path, coupled with a document similarity service, can be used to provide a number of applications, such as a security solution that is capable of finding and restricting access to documents that contain information that is similar to other existing files that are known to contain sensitive information.

IPC 1-7

**G06F 17/30**; **G06F 17/10**

IPC 8 full level

**G06F 17/10** (2006.01); **G06F 17/30** (2006.01); **G06F 21/10** (2013.01); **G06F 21/60** (2013.01); **G06F 21/64** (2013.01)

IPC 8 main group level

**G06F** (2006.01)

CPC (source: EP)

**G06F 16/10** (2018.12); **G06F 16/334** (2018.12)

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 PT RO SE SI SK TR

DOCDB simple family (publication)

**WO 2004066086 A2 20040805**; **WO 2004066086 A3 20050120**; CA 2553654 A1 20040805; CA 2553654 C 20140422; EP 1590748 A2 20051102; EP 1590748 A4 20080730; JP 2006516775 A 20060706; JP 4667362 B2 20110413

DOCDB simple family (application)

**US 2004001530 W 20040121**; CA 2553654 A 20040121; EP 04704049 A 20040121; JP 2006501066 A 20040121