DataWorkOut 360

Una solución completa para el problema de Identificación de Personas Duplicadas que presentan todas las corporaciones con soporte y desarrollo garantizado en Español e Inglés.

dataworkout 360®

Una solución completa para el problema de Identificación de Personas Duplicadas que presentan todas las corporaciones con soporte y desarrollo garantizado en Español e Inglés.

DataWorkOut 360 ofrece una solución que se adapta a las diferentes corporaciones y mercados, optimizando al máximo los procesos:

ETL Data Cleansing Matching

Todos estos a su vez pueden ser implantados a través de servicios tanto “on-line” como “off-line” en otras palabras, en tiempo real o “back-office” a través de procesos “batch”, garantizando una solución justo a la medida de los clientes, de acuerdo a sus necesidades especificas de operación y mercado en particular. Con este afán DataWorkOut 360 contiene los últimos y mejores algoritmos de comparación e identificación de cadenas, a continuación comentaremos al respecto los más relevantes:

- *Autómatas finitos o su equivalente a través del uso de otros algoritmos*

DataWorkOut 360 entiende el término Autómatas Finitos como algoritmos con código de resolución no exponencial; y hace uso de una gran cantidad de Autómatas Finitos Determinísiticos y Autómatas Finitos no Deterministicos. Usa la estructura de los datos y algoritmos en el matching de expresiones regulares, en el reconocimiento ontológico **(OntologyMatcher)** y de tesauros; así como en la detección de patrones **(RulesMatcher)**.

“El algoritmo de parseo en caso de entrar en una cadena irracional que envíe el parseo a infinito regresa el valor nulo”.

DataWorkOut 360 aprovechando las ventajas de Autómatas Finitos determinísticos provee de:

- *Estructuras de datos eficiente*

Las cuales son usadas para extraer una lista de expresiones con diferente forma (exactitud, minúsculas, normalizadas) de las cadenas de texto. Esta estructura de datos es una mejora del Judy-Array (http://judy.sourceforge.net/application/shop_interm.pdf) con una gran cantidad de mejoras para reducir el uso de RAM (usando prefijos, algoritmos de compresión)

- *Estructuras de datos eficiente basado en B-Tree/DFA para almacenar grandes bases de conocimiento*

Esta estructura de datos es una mejora de String-B-Tree (http://portal.acm.org/citation.cfm?id=301973) con algunas mejoras para siempre reducir el número de lecturas a disco (una lectura a disco por búsqueda en el peor de los escenarios).

- *Comparación de cadenas y/o registros por similitud*

DataWorkOut 360 realiza comparaciones idénticas y métodos de comparación aproximada utilizando técnicas de parseo. La comparación de cadenas la ejecuta basándose en algoritmos de tipo Stemming (Permite reducir una palabra a su raíz), lemmatisation (Agrupa las diferentes formas flexivas de una palabra para que puedan analizarse como un solo elemento) y Damerau-Levenshtein distance.

MÉXICO
Delivery Center
Avenida Lomas de Sotelo 1102
Colonia Lomas Hermosa
Delegación Miguel Hidalgo 11200, Distrito Federal
• www.paramind-usa.com • www.paramind.com.mx Tel. +52.55.4437.7266/69

dataworkout 360®

Algoritmos

Edit distance

DataWorkOut 360 usa Edit Distance (Damerau-Levenshtein distance), el cual se basa en la distancia física entre dos teclas de un teclado. Por ejemplo, g tiene una distancia de 1 de las teclas r, t, y, f, h, v, b, y n. Las diagonales inmediatas como r, y, v y n tienen una distancia de 1 en vez de 1.414 que se tiene para las teclas cercanas de forma horizontal o vertical.

Soundex

DataWorkOut 360 usa el algoritmo Soundex, el cual codifica fonéticamente los nombres, es decir dada una cadena que contiene un nombre, se codifica con este algoritmo y se obtiene una cadena que contiene un código que se supone idéntico para los apellidos fonéticamente cercanos.

Este algoritmo permite identificar si el cambio se produjo a razón de una falta de ortografía, y se apoya del directorio de apellidos para determinarlo.

DataWorkOut 360 usa el algoritmo MetaPhone el cual es una evolución del algoritmo Soundex.

Phonetix

Este algoritmo funciona de forma similar al algoritmo Soundex, sólo que éste último se usa cuando se tienen diferencias tanto de sonido como de ortografía, especialmente entre diferentes culturas o naciones.

Por ejemplo: Thirteen y Thirty

El sonido es similar pero se escriben diferente y su significado es distinto DataWorkOut 360 usa el algoritmo Phonetix, el cual es una implementación open source de Soundex, Metaphone y Double Metaphone. Phonetix contiene una version hardcoded de las reglas fonéticas para el idioma inglés pero no pueden ser cambiadas o editadas. DataWorkOut 360 hizo su propia implementación por razones de performance integrando reglas de idioma.

Hemos desarrollado reglas fonéticas para 19 idiomas (ar ca cs da de en es et fa fi fr he it ja nl no pl pt ro ru sk sl sv) y seguimos mejorando estas reglas.

Karp – Rabin

DataworkOut 360 hace uso del algoritmo Karp-Rabin, el cual hace búsqueda de una cadena dentro de otra. La ventaja es que en vez de buscar el en cada posición del texto, si el patrón ocurre, es más eficiente revisar solamente si el contenido es parecido al patrón.

"Karp – Rabin usa de una función hash para verificar la semejanza entres dos palabras."

Character frecuency

El algoritmo Character Frequency se basa en el conteo del número de ocurrencias de cada carácter en el texto original.

Por ejemplo:

Jesús Ibarra Aragón	Ocurrencias	J 2
Jesús I. Aragón		e 2
		s 4
		u 2
		...

Toma para su corrección los caracteres cuya ocurrencia resulta impar, ya que deben ser pares las ocurrencias para considerarse el mismo registro. Este algoritmo no resulta eficaz en el caso de nombres con las mismas letras por ejemplo: Lina y Alin.

Al calcular el algoritmo Damerau-Levenshtein distance, DataWorkOut 360 permite especificar diferentes pesos a un caracter en específico de acuerdo a su frecuencia.

MÉXICO
Delivery Center
Avenida Lomas de Sotelo 1102
Colonia Lomas Hermosa
Delegación Miguel Hidalgo 11200, Distrito Federal
• www.paramind-usa.com • www.paramind.com.mx Tel. +52.55.4437.7266/69

dataworkout 360®

Algoritmos

- *Fuzzy match*

El algoritmo Fuzzy match hace la comparación del campo con todas las variables posibles que se encuentran en los directorios.

DataWorkOut *360* regresa una identidad basada en Damerau-Levenshtein distance a nivel de palabra, lo que permite tener un fuzzy match a nivel palabra *(olvidar / intercambiar / agregar /borrar algunas palabras),* también tiene fuzzy matching a nivel de caracter para reconocer una palabra basado en Damerau-Levenshtein distance, stemming, lemmatization, phonetization a nivel de caracter.

- *Tecnología de indexación standard (Identity Matching) y compresión de datos*

"DataWorkOut 360 usa algoritmos standard para reducir el tamaño de la RAM de indexación"

(ricecode, offset encoding).

- *Factorización / tree pruning*

Reducen la complejidad del conjunto de reglas.

- *Word level distance (Damerau-Levenshtein)*

DataWorkOut *360* aplica Damerau Levenshtein a nivel de palabras.

- *Character level distance (Damerau-Levenshtein)*

DataWorkOut *360* aplica Damerau-Levenshtein a caracteres Unicode.

- *Hidden Markov Model*

DataWorkOut *360* usa el algoritmo del modelo Hidden-Markov para predicción con algunas modifcaciones para hacer más eficiente el resultado de las observaciones de pesos (basado en tags/vocabularios).

- *Campo aleatorio condicional*

DataWorkOut 360 usa el algoritmo del modelo Hidden-Markov para predicción con algunas modifcaciones para hacer más eficiente el resultado de las observaciones de pesos (basado en tags/vocabularios).

- *Tecnología de indexación standard (Identity Matching) y compresión de datos*

- *Estructuras de datos genéricos para matching de autómatas / regexp open sourced)*

- *Categorización (naives bayes/SVN)*

DataWorkOut *360* implementa la Categorización con estructuras de datos dedicadas basadas en Autómatas Finitos Determinísticos.

- *Adecuación eficiente con la gramática local (transductores deterministas).*

MÉXICO
Delivery Center
Avenida Lomas de Sotelo 1102
Colonia Lomas Hermosa
Delegación Miguel Hidalgo 11200, Distrito Federal
• www.paramind-usa.com • www.paramind.com.mx Tel. +52.55.4437.7266/69

dataworkout 360®

Ventajas

- **1. Estructuración automática de grandes volúmenes de datos sin estructura.**

DataWorkOut 360 se especializa en el procesamiento grandes volúmenes de datos sin estructura. Más allá de simplemente identificar palabras clave en un documento, el procesamiento a fondo significa transformar contenidos sin estructura en una fuente completamente clasificada que puede ser sintetizada con datos estructurados como son las bases de datos corporativas y aplicaciones de negocios.

- **2. Escalamiento efectivo en costo**.

DataWorkOut 360 le permite un fácil escalamiento a un menor costo. El sistema es extremadamente eficiente, soportando la indexación en tiempo real, teniendo la capacidad de procesar grandes cantidades de registros.

El performance puede ser distribuido en múltiples servidores.

- **3. Arquitectura flexible**

DataWorkOut 360 provee de una plataforma flexible y extensible. Su Arquitectura Orientada a Servicios (SOA) y su extensas Application Programming Interfaces (APIs) aseguran:

- Máxima flexibilidad de datos, con la habilidad de conectarse a fuentes internas y externas.
- Ágil y bajo costo de desarrollo, con una capa independiente de datos y soporte para formatos Web standard y protocolos.
- Máxima escalabilidad, performance y disponibilidad, con capacidad de replicación de datos para un escalamiento fácil y de bajo costo.
- Arquitectura asíncrona para correr tareas concurrentes.

MÉXICO
Delivery Center
Avenida Lomas de Sotelo 1102
Colonia Lomas Hermosa
Delegación Miguel Hidalgo 11200, Distrito Federal
• www.paramind-usa.com • www.paramind.com.mx Tel. +52.55.4437.7266/69

DataWorkOut 360

Una solución completa para el problema de Identificación de Personas Duplicadas que presentan todas las corporaciones con soporte y desarrollo garantizado en Español e Inglés.