Ingeniería y Competitividad, Volumen 13, No. 1, p. 61 - 75 (2011)

Be4SeD: Benchmarking para evaluación de técnicas de descubrimiento de servicios

Luís J. Suárez-Meza[1§], Luis A. Rojas-Potosí[2], Juan C. Corrales[3], Oscar M. Caicedo[4]

[1§]*Programa de Ingeniería Electrónica y Telecomunicaciones, Universidad del Cauca*
ljsuarez@unicauca.edu.co

[2]*Programa de Ingeniería Electrónica y Telecomunicaciones, Universidad del Cauca*
luisrojas@unicauca.edu.co

[3]*Programa de Ingeniería Electrónica y Telecomunicaciones, Universidad del Cauca*
jcorral@unicauca.edu.co

[4]*Programa de Ingeniería Electrónica y Telecomunicaciones, Universidad del Cauca*
omcaicedo@unicauca.edu.co

(Recibido: Abril 30 de 2010 - Aceptado: Abril 25 de 2011)

Resumen

Actualmente, el creciente número de procesos de negocio y servicios ofrecidos, es fuente de innumerables proyectos de investigación, orientados a generar mecanismos de descubrimiento; teniendo como resultado un sinnúmero de algoritmos para recuperar servicios. Sin embargo, dichos proyectos no utilizan una base común para evaluar sus técnicas de búsqueda, impidiendo que las evaluaciones sean objetivas. Por lo tanto, se hace necesaria una herramienta pública, que proporcione una referencia común, que permita comparar y valorar los resultados de los diferentes algoritmos utilizados en el emparejamiento de servicios, con el fin de mejorar la calidad de las técnicas de descubrimiento propuestas. Este artículo presenta una aplicación pública, que implementa una metodología de benchmarking para evaluar la calidad de recuperación de las técnicas de emparejamiento de servicios. Este benchmarking está compuesto de un mecanismo de evaluación intuitivo, de un módulo de ingreso de los datos correspondientes al algoritmo a evaluar y un componente que entrega resultados estadísticos: recall, precision, overall, k-precision y p-precision. Sus funcionalidades se ofrecen como servicio web para facilitar la integración con las implementaciones de algoritmos a evaluar. Finalmente se evalúa un algoritmo de emparejamiento, el cual evidencia el uso de la plataforma Be4SeD en este contexto.

Palabras Claves: Benchmarking, Procesos de negocio, Algoritmos, Descubrimiento de servicios.

Be4SeD: benchmarking for evaluation of service discovery techniques

Abstract

The growing number of business processes and services, has resulted countless develop research projects, aimed at generating discovery mechanisms. While there are a number of algorithms to retrieve services, such work does not use a common basis to evaluate the used search techniques, preventing objective evaluations. Therefore, it is necessary a public tool that provides a common reference for comparing and evaluating the results of the algorithms used in the matching of services, in order to improve the quality of the proposed discovery techniques. This article presents a public application, which implementing a benchmarking methodology to evaluate the quality of recovery of service matching techniques. This benchmarking is comprised of an intuitive assessment mechanism, an input module data for the algorithm to evaluate and a component that provides statistical results: recall, precision, overall, k-precision and p-precision. Its features are offered as a web service to facilitate integration with the implementations of algorithms to evaluate. Finally a matching algorithm is evaluated for evidence the use of Be4SeD platform in this context.

Keywords: Benchmarking, Business processes, Algorithms, Service discovery.

1. Introducción

En los últimos años los modelos de negocio han respondido a la creciente evolución de la infraestructura TI. En este sentido, el concepto de Arquitectura Orientada a Servicios, enfoca sus esfuerzos en la reducción de costos, con el fin de generar nuevas funcionalidades de una manera más rápida y efectiva, haciendo uso de módulos o servicios existentes. Es así como el concepto de composición de servicios, relacionado con la reutilización y adaptación de funcionalidades, ha sido ampliamente desarrollado y utilizado.

La composición a nivel empresarial, tiene la ventaja que al planearse con detenimiento, puede garantizar la creación de funcionalidades complejas y efectivas. Pero en un ambiente, en el cual el avance de las tecnologías de la información ha llevado a un aumento en el acceso a internet, y sobre todo a la oferta de servicios a los usuarios; la capacidad de componer manualmente nuevos servicios se ve diezmada debido al tiempo que toma la búsqueda individual de los mismos. Es por esto que además de la composición, el concepto de descubrimiento de servicios, como proceso previo a la misma, ha llamado la atención de diversos grupos de investigación. El descubrimiento es el proceso encargado de encontrar los servicios más pertinentes con respecto a un servicio solicitado.

Los trabajos relacionados con el descubrimiento de servicios desarrollan una referencia propia para evaluar los resultados de sus algoritmos, lo cual es válido. Sin embargo, el principal problema de esta situación es que al carecer de una referencia común para hacer evaluaciones, la objetividad de las mismas se afecta por no tener el mismo patrón de medida. Para subsanar este inconveniente, se presenta una herramienta pública, basada en una metodología de *Benchmarking*, que evalúa la calidad de recuperación de las técnicas de emparejamiento de servicios utilizando una referencia, ante la cual los algoritmos de emparejamiento comparan sus resultados de selección de servicios, para determinar la eficacia de las técnicas de recuperación empleadas, que definimos como *Benchmark de Referencia*, que es el resultado de comparaciones manuales realizadas por evaluadores expertos en el tema. La herramienta se compone de: un mecanismo de evaluación intuitivo, un módulo de ingreso de los datos correspondientes al algoritmo a evaluar y un componente que entrega resultados estadísticos: *recall, precision, overall, k-precision y p-precision*, Zhang et al. (2004.).

El presente documento describe una aplicación de *benchmarking* que centra su atención en la construcción de una referencia que represente el criterio de un usuario, y se estructura como sigue: la sección 2 presenta el estado del arte relacionado con las diferentes técnicas para la evaluación de mecanismos de recuperación de servicios. La sección 3 aborda una descripción de la solución propuesta. Las secciones 4 y 5 respectivamente, describen el prototipo implementado y evidencian su aplicación por medio de la evaluación de un algoritmo de emparejamiento de servicios. Los resultados de la técnica de descubrimiento evaluada están consignados en la sección 6. Finalmente, se presentan algunas conclusiones.

2. Estado del arte

Esta sección presenta un estudio de trabajos relacionados con la evaluación de técnicas de descubrimiento de servicios, y también investigaciones sobre técnicas de Recuperación de Información (IR), las cuales pueden aplicarse en el análisis de resultados.

En los últimos años se ha observado un creciente desarrollo en el campo de IR (Martínez, 2004, Egghe, 2008, ECIR, 2008). Sin embargo, uno de los principales inconvenientes en este dominio gira en torno a la evaluación de la calidad de las diferentes técnicas de recuperación propuestas por diversos autores. Es por ello que metodologías fiables y herramientas web de evaluación son fundamentales para el progreso científico de este campo.

De acuerdo con Voorhees (2001), la evaluación de IR ha sido dominada en cuatro décadas por el *paradigma de Cranfield,* el cual se caracteriza por el uso de los criterios *Recall y Precision*. Este paradigma considera una referencia común para la evaluación de las técnicas de recuperación la cual es construida a partir de juicios de expertos en el dominio de aplicación, y soportada en una colección de prueba compuesta de: un conjunto de

documentos (datos de prueba), un conjunto de necesidades de información (temas o consultas) y los documentos que deben recuperarse. Si bien es cierto que éste paradigma presenta una contribución importante para la IR, Voorhes (2001) resalta que los experimentos basados en suposiciones propuestos por el *paradigma Cranfield* son procesos que generan ruido, pero permiten obtener resultados útiles a la hora de valorar el rendimiento de diferentes sistemas evaluados por el mismo experimento. Según Küster et al. (2007) y Küster et al. (2009) no es adecuado mirar la recuperación de servicios como un simple problema de IR, ya que las principales diferencias entre estos enfoques son la expresividad del formalismo y el razonamiento empleado. Sin embargo, al momento de valorar los resultados del descubrimiento de servicios se puede hacer uso de medidas de evaluación de IR como lo dice *Cranfield*.

Los autores del *S3 Matchmaker Contest* (S3, 2008) aplican el *paradigma de Cranfield* a la evaluación de técnicas de descubrimiento de servicios. Esta aproximación define una colección de servicios OWL-S y la evaluación de las técnicas de emparejamiento de Servicios Web Semánticos (SWS) se basa en las clásicas medidas de *Precision, Recall, F1*,y considera también un promedio de los tiempos de respuesta de las consulta. Este tipo de aplicaciones del *paradigma de Cranfield* para el dominio del emparejamiento SWS tiene un alcance limitado, ya que no permite una evaluación comparativa de diferentes estándares para descripción de servicios.

Por otro lado, un problema común para diferentes enfoques y evaluaciones en el dominio del emparejamiento de servicios es el uso de bancos de prueba adecuados. Los servicios no necesariamente tienen que ser reales o extremadamente complejos para poner a prueba las características de un sistema de emparejamiento. Lo importante, es la descripción formal de los servicios utilizada (Grafos, Redes de Petri, Autómatas de Estado Finito, etc.), la cual facilita la tarea del emparejamiento y por ende la de evaluación. Además, los servicios también se deben describir con suficiente detalle para permitir un significativo descubrimiento. Después de todo debe haber una ventaja de usar las anotaciones semánticas en comparación con el simple uso de las técnicas tradicionales de IR (Küster et al., 2009).

En WSBEn (Seog-Chan & Lee, 2009) se construye un *benchmark* a partir de las descripciones de servicios encontradas en un banco de WSDLs denominado PUB06. Con base en las relaciones entre servicios, operaciones y parámetros encontrados, se generan redes de nodos tomando tres modelos de Redes de servicios Web, como son *random, small-world*, and *scale-free*, que según Albert & Barabasi (2002), son suficientes para modelar redes de servicios en el mundo real. Este trabajo proporciona un *benchmark* para ejecutar pruebas, archivos auxiliares para realizar análisis estadísticos y utiliza también una representación formal de grafos, pero no implementa una herramienta que permita llegar más allá de las evaluaciones y obtener un *Benchmarking*, considerando a este último como una comparación entre dos o más evaluaciones o *Benchmarks*.

En Toma et al. (2007) se presenta un *framework* que permite evaluar diferentes enfoques de descubrimiento de servicios Web y entornos Grid a nivel semántico, de acuerdo con aspectos como: leguaje de consulta y publicación, escalabilidad, soporte de razonamiento, emparejamiento versus intermediación, y soporte de mediación. El principal aporte de este trabajo está en el estudio realizado, más que en el *framework* de comparación implementado. Mientras que el *framework* da directrices para una comparación estructurada, este no ofrece una prueba concreta, medidas, estadísticas, *benchmarks* o procedimientos para una evaluación comparativa y objetiva.

En el estado del arte presentado se expuso diferentes estudios relacionados con la evaluación de técnicas de recuperación de servicios, y se observa un marcado interés en su desarrollo. Se aprecia además, que aún después de los esfuerzos realizados, no se ha logrado tener una base común para la evaluación de técnicas de recuperación de servicios, lo cual, como se mencionó, disminuye la objetividad en el momento de seleccionar el mejor algoritmo de descubrimiento. A partir de estos argumentos, el objetivo que persigue el presente

artículo gira en torno a proporcionar una herramienta pública, que permita evaluar la eficacia de las técnicas de recuperación de servicios, utilizando un *Benchmark de Referencia*, relacionando el desempeño de los algoritmos y la observación consignada por los evaluadores. La calidad del método de descubrimiento es determinada por un conjunto de medidas de desempeño: *recall, precision, overall, k-precision y p-precision.*

3. Arquitectura genérica de Be4SeD

La Figura 1 presenta los subsistemas de Be4SeD. En esta herramienta, evaluadores expertos en el tema de descubrimiento de servicios comparan manualmente por parejas los servicios contenidos en el repositorio, con el fin de generar su propio benchmark. Una vez todos los evaluadores registrados en la plataforma concluyen la evaluación de los servicios, el administrador de Be4SeD ordena la creación del Benchmark de Referencia, ejecutando las políticas que permiten generalizar los resultados de la evaluación de cada experto. Este Benchmark de Referencia es la característica más relevante del trabajo expuesto en el presente artículo, ya que para evaluar y determinar la calidad de un algoritmo es necesaria una base confiable que pueda compararse con los resultados arrojados por la técnica de emparejamiento y así inferir sobre la calidad de los mismos.

Por otro lado, Be4SeD permite a los autores de las diferentes técnicas de recuperación de servicios (usuarios) crear su propio *Benchmark del Algoritmo*, con el fin de comparar los resultados con el *Benchmark de Referencia*, generado por los expertos evaluadores. Por último, los usuarios también pueden acceder al sistema de *Análisis Estadístico*, para obtener información sobre la evaluación de su algoritmo y generar su propio análisis. A continuación se describen los subsistemas de Be4SeD.

3.1 Banco de Servicios

Es una colección común de servicios utilizada para evaluar los algoritmos de descubrimiento. Estos servicios son clasificados como *Query y Target*. La evaluación se realiza entre un número definido de *servicios Query* y todos los *servicios Target* (1:N). Vale la pena aclarar que los servicios *Query* están incluidos como servicios *Target*. Finalmente, se resalta que no es posible generar evaluaciones entre parejas de servicios *Target* y mucho menos generar evaluaciones de parejas repetidas.

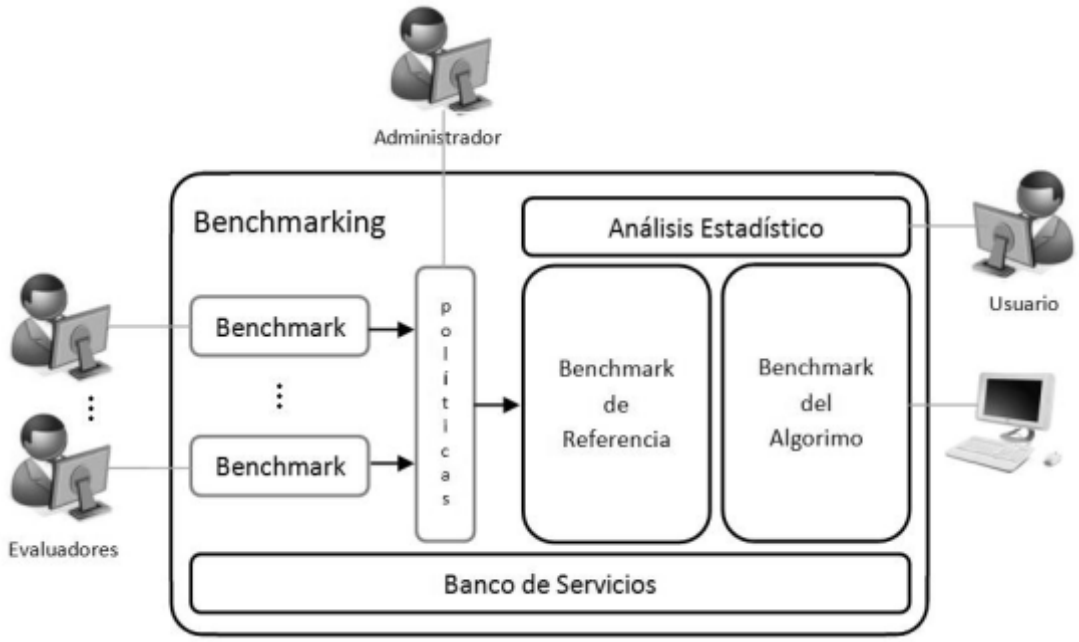

Figura 1. Arquitectura genérica de Be4SeD.

3.2 Benchmark de Referencia

Para garantizar que la evaluación de algoritmos de emparejamiento de servicios sea objetiva y confiable, se necesita una referencia común que proporcione una "verdad absoluta", producto del criterio de expertos en el tema de Descubrimiento de Servicios, es decir: una evaluación ejemplo, resultado de la generalización de las evaluaciones realizadas por los expertos, que permita tener una referencia común para contrastar con los resultados de los algoritmos. Por esta razón el *Benchmark de Referencia* constituye un aspecto clave en esta investigación. Este subsistema representa los datos de referencia, ante los cuales los algoritmos de emparejamiento comparan sus resultados de selección de servicios, con el fin de determinar la eficacia de las técnicas de recuperación empleadas. El *Benchmark de Referencia* es creado considerando las siguientes políticas:

3.2.1 Políticas de Creación del Benchmark de Referencia

El valor de similitud para cada comparación se calcula de la siguiente manera, Zhang et al. (2004.):

a. Evaluación del Emparejamiento por usuario:

$$EMu = \sum_{i=1}^{n} (Wi * Sui) \quad (1)$$

Donde: ***Wi*** (es el peso asignado por el evaluador según el grado de relevancia del atributo del servicio evaluado), ***Sui*** (Calificación) y ***n*** (cantidad de atributos a considerar), el valor *EMu* es la similitud estimada por un usuario para una pareja de servicios.

La media de similitud para cada uno de los servicios evaluados es la siguiente:

b. Evaluación Total:

$$TE(EM) = \frac{\sum_{u=1}^{n} EMu}{n} \quad (2)$$

Donde ***EMu*** es la similitud de una comparación y ***n*** es el número de evaluadores.

3.3 Benchmark del Algoritmo

Es la colección de los resultados de recuperación de servicios obtenidos de la ejecución de los algoritmos de emparejamiento a evaluar. Dichos algoritmos son ejecutados sobre las parejas de *servicios Query* y todos los servicios *Target* (1:N) contenidas en el *Banco de Servicios*. Se debe resaltar que los resultados de recuperación varían según las técnicas empleadas para determinar la similitud entre servicios.

3.4 Análisis estadístico

El desempeño general del sistema se establece utilizando las medidas: *recall* (*r*), *precision* (*p*), *overall* (*o*), *top-k precision* (P_k) *y p-precision* (P_p). Para evaluar la calidad del algoritmo de recuperación, se comparan los servicios (*P*) retornados por el *Algoritmo* con los servicios (R) obtenidos en el *Banco de Servicios*. De esta forma se puede determinar un conjunto de verdaderos positivos (*I*), servicios correctamente identificados; igualmente se determina un conjunto de falsos positivos, servicios falsos recuperados (*F = P/I*), y falsos negativos, es decir servicios relevantes no recuperados (*M = R/I*) (Corrales et al., 2008). $Retrel_k$ es el conjunto de servicios relevantes para un *top k* de servicios recuperados, mientras *Rel-p* determina cuantos de los servicios de $Retrel_k$ están en la misma posición del ranking de referencia del *Banco de Servicios*, Zhang et al. (2004.). Con base en la cardinalidad de estos conjuntos se tiene:

$$p = \frac{|I|}{|P|} \quad (3)$$

$$r = \frac{|I|}{|R|} \quad (4)$$

$$o = r * \left(2 - \frac{1}{p}\right) \quad (5)$$

$$p_k = \frac{|Retrel_k|}{k} \quad (6)$$

$$p_p = \frac{[Retrel_k] Rel - p}{k} \quad (7)$$

La medida *precision* estima la fiabilidad de los servicios relevantes recuperados por el algoritmo, en tanto *recall* especifica el porcentaje de los servicios relevantes recuperados. Por su parte la medida *overall* valora la calidad del emparejamiento, teniendo en cuenta el esfuerzo necesario para la eliminación de falsos positivos y los servicios no recuperados, Zhang et al. (2004.).

Las medidas establecidas anteriormente se calculan para cada una de los servicios empleados en el *Banco de Servicios*. Para estimar la *precision* y el *recall* de todo el sistema, se emplean los métodos *macro-promedio* y *micro-promedio* (Lewis, 1992), así:

Macro-promedio: es la media de la *precision* y *recall* de los emparejamientos individuales.

$$P = \frac{\sum_{i=1}^{n} P_i}{n} \quad (8)$$

$$R = \frac{\sum_{i=1}^{n} r_i}{n} \quad (9)$$

Donde: ***n*** es el número de emparejamientos realizados**.**

Micro-promedio: tiene en cuenta los verdaderos positivos y los falsos positivos. La *precision* y el *recall* se calculan utilizando los valores globales.

$$P = \frac{\sum_{i-1}^{n} TP_i}{\sum_{i=1}^{n} (TP_i + FP_i)} \quad (10)$$

$$R = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} (TP_i + FN_i)} \quad (11)$$

Donde: ***TPi***: Son los verdaderos positivos, ***FPi***: Son los Falsos positivos, ***FNi***: Son los Falsos Negativos.

A partir de los resultados entregados por este módulo se realiza un completo análisis estadístico que permite determinar la calidad de un algoritmo de emparejamiento de servicios. Este análisis se fundamenta en la evaluación de las siguientes relaciones: *Precision vs. Recall, Overall, K-Precision vs. K, P-Precision vs. K,* aplicadas en tres escenarios: i) evaluación de las medidas de desempeño comparando servicios de entrada contra los de un mismo dominio contenidos en el repositorio, ii) comparación de los servicios de entrada contra aquellos almacenados que pertenecen a un dominio diferente, y iii) comparación de los servicios de consulta contra todos los servicios contenidos en el repositorio.

4. Prototipo

La Figura 2, expone el diagrama de despliegue de la plataforma Be4SeD. Su implementación fue realizada sobre Glassfish V2.1, con J2EE (versión 1.4), utilizando el *Contenedor de EJB* (*Entrerprise Java Beans, versión 2.1*) para desplegar la lógica

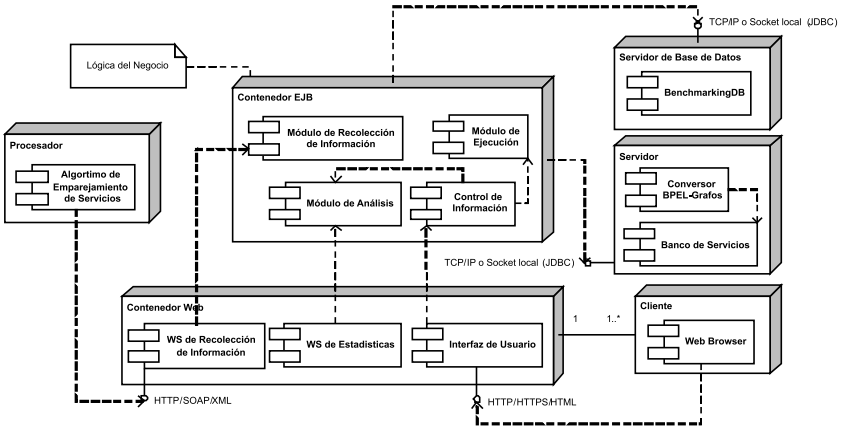

Figura 2. Diagrama de despliegue de la plataforma Be4SeD.

de negocio. La interfaz de usuario y los servicios web se encuentran en el *Contenedor Web*. Adicionalmente, se utilizó PostgresSQL (versión 8.3) como motor de base de datos. El *Banco de Servicios* utiliza el repositorio de procesos de negocio presentado en Vanhatalo et al. (2006).

4.1 Banco de Servicios

Está conformado por actividades básicas descritas en una colección de documentos BPEL, encontrados en el repositorio de procesos de negocio presentado en Vanhatalo et al. (2006.), el cual soporta el almacenamiento y consulta de documentos BPEL (y otros documentos XML). Éste provee un API Java para la manipulación de sus archivos como objetos, y guarda las descripciones de 53 actividades básicas BPEL, agrupadas en 5 dominios: Vacaciones, Compras, Pagos, Disponibilidad de productos e Información de productos, clasificadas como *Actividades Query* (28 actividades) y *Actividades Target* y *Actividades Target* (25). Obteniendo de ellas 1106 parejas, número considerable a evaluar, conformando un gran banco de datos. Finalmente es importante resaltar que a las actividades básicas abstraídas de BPEL, se les adicionó un parámetro de contexto, con el fin de mostrar la capacidad de adaptación del *Banco de Servicios*, permitiendo así la adopción de nuevos atributos y diferentes tipos de representación de servicios, dependiendo de su dominio de aplicación. Esto se realizó considerando la propuesta presentada en Hermida et al. (2009).

4.2 Conversor BPEL-Grafos

Transforma las descripciones de comportamiento (BPEL) en su equivalente en grafos, implementando la estrategia presentada por Mendling & Ziemann (2005). El algoritmo emplea un proceso de transformación recursivo para cada tipo de actividad estructurada, tomando una aproximación de arriba-abajo (top-down). Las actividades básicas BPEL son transformadas en nodos y las secuencias son obtenidas conectando los nodos requeridos por medio de aristas. Las actividades estructuradas son representadas por medio de operadores lógicos XOR y AND (Corrales, 2008).

4.3 Interfaz de Usuario

Facilita la interacción de los expertos con la plataforma. Su lógica de presentación es implementada en el *Módulo de Interfaz de Usuario*. En la Figura 3, se muestra la vista que permite al experto seleccionar el dominio y la

Figura 3. Interfaz de selección de actividades a evaluar.

actividad a evaluar para posteriormente construir su *benchmark*. En la Figura 4, se observa la evaluación hecha entre dos nodos *Query* y *Target*. El evaluador realiza la comparación de las actividades, asignando una calificación a cada uno de los atributos según el nivel de similitud. El valor de la calificación a asignar está entre 0 y 5, donde 0 es la mínima y 5 la máxima similitud. Además, el evaluador, al momento de registrarse en Be4SeD, fija un peso de acuerdo a la importancia de cada uno de los atributos. La suma de todos los pesos debe ser igual a 100%.

4.4 Control de Información

Es el encargado de procesar las peticiones de la interfaz de usuario, y encontrar los datos que ésta necesita. Por lo tanto, este módulo toma información tanto del banco de servicios como de los módulos de análisis y ejecución.

4.5 BenchMarkingDB

Almacena las valoraciones realizadas por los expertos en un formato relacional.

4.6 Módulo de Ejecución

Implementa las políticas para la generación del *Benchmark de Referencia*. El administrador lo ejecuta una vez los expertos completen la valoración del *Banco de Servicios*. Sin embargo, es importante aclarar que este módulo no es el encargado de realizar análisis sobre los datos obtenidos en este proceso (ver sección 4.7).

4.7 Módulo de Recolección de Información

Se encarga de obtener los resultados de algoritmos de emparejamiento de servicios, generando el *Benchmark del Algoritmo*. Para esto, el módulo de recolección de información implementa la lógica que soporta las siguientes operaciones: *Autenticación* - valida el ingreso de datos a la plataforma. Como parámetros de entrada recibe un *login* y un *password*. Retorna una cadena de caracteres (serial) utilizada para ingresar nuevos datos al sistema. *Obtener atributos de las parejas de actividades* - retorna el valor de un parámetro específico de los servicios que conforman la pareja consultada. Como parámetros de entrada recibe el

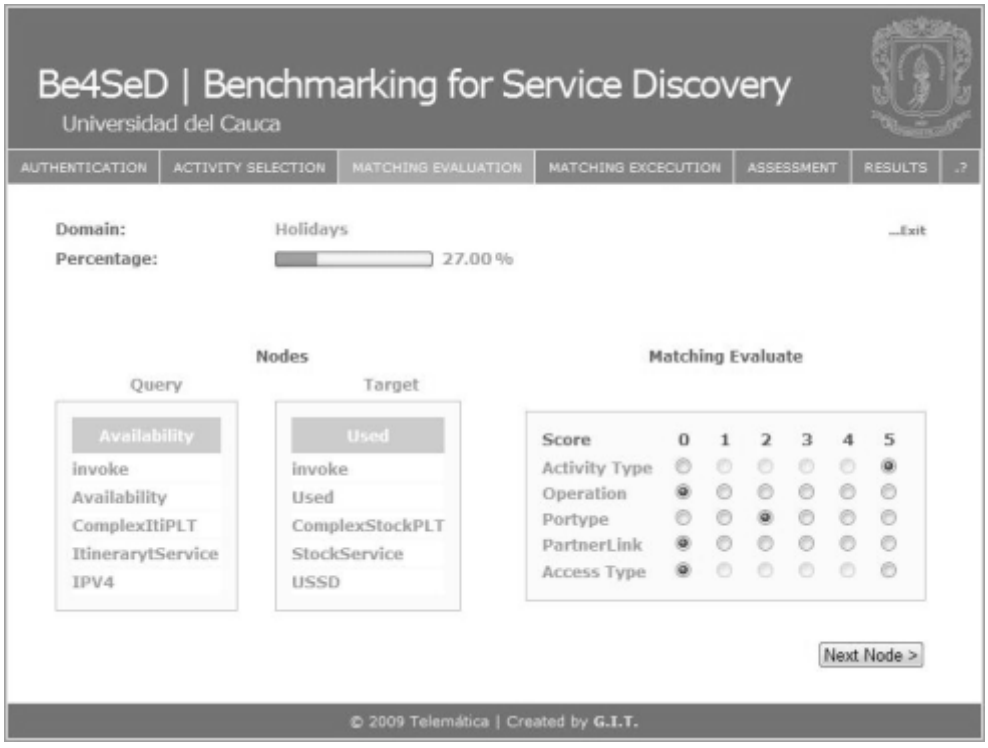

Figura 4. Interfaz de evaluación.

serial, el identificador de la pareja (0-1105), nodo ("a" o "b") y atributo (0-(n-1)). Para este caso las actividades cuentan con 5 atributos, n=5. *Activity Type*, *Operation*, *Portype*, *PartnerLink* y *Access Type*. *Evaluar similitud* - almacena el resultado de la comparación de la pareja evaluada por el algoritmo.

Los parámetros de entrada son: serial, identificador de la pareja y la similitud (*Score*). Las anteriores operaciones son expuestas por medio del *WS de Recolección de Información*, facilitando su consumo por parte de las implementaciones de distintos algoritmos.

4.8 Módulo de Análisis

Es el encargado de entregar los datos estadísticos, generados como resultado de la comparación entre el *Benchmark del Algoritmo* y el *Benchmark de Referencia*. Estos datos son una adaptación de medidas propias del campo de descubrimiento de servicios a una metodología de *benchmarking*, demostrando su flexibilidad y por ende la gran utilidad de ésta en diversos entornos. Este componente implementa la lógica de operaciones que retornan los valores de *Precision*, *Recall*, *Overall*, *P-presision y k-presicion* utilizando las técnicas de *Macro-promedio* y *Micro-promedio*. También contiene una operación de autenticación similar al expuesto en el *Módulo de Ejecución*. El usuario (Figura 1) puede acceder a esta información consumiendo el *WS de Estadísticas*, que expone las operaciones descritas. Lo anterior permite adaptar la plataforma Be4SeD a una aplicación externa que utilice estos datos estadísticos para generar un análisis sobre la calidad de su algoritmo. Con la información generada por el módulo de *Análisis*, la respectiva *Interfaz de Usuario* de Be4SeD permite visualizar una lista ordenada de los resultados arrojados por el algoritmo evaluado (*Benchmarking del Algoritmo*) y la evaluación hecha por los expertos (*Benchmarking de Referencia*), para todas las actividades de consulta contenidas en el *Banco de Servicios*, ver Figura 5. Esta es una característica

Figura 5. Ranking de servicios.

muy importante de la plataforma Be4SeD. La columna a la izquierda expone el ranking de servicios del *Benchmark de Referencia,* el nodo *Query* es *Payment,* y el *K es* igual a 5. Como se presentó, los *servicios Query* están incluidos como *servicios Target,* es por ello que el primer nodo encontrado en la lista es el mismo nodo *Payment,* el cual obviamente posee la máxima similitud; mostrando que el *Benchmarking de Referencia* es confiable para determinar la calidad de algoritmos de emparejamiento de servicios.

5. Metodología

La plataforma aplica una metodología de *benchmarking,* que consiste en inferir análisis a partir de la comparación de evaluaciones realizadas a dos o más técnicas o sistemas, que tengan una base común de entradas. A estas evaluaciones se las denominó *Benchmark del Algoritmo* y *Benchmark de Referencia,* siendo *Benchmarking* el proceso encargado de comparar los dos *Benchmarks.* Después de la construcción del *Benchmark de Referencia,* se da paso a la recolección de datos para generar un *Benchmark del Algoritmo.* Para evidenciar la utilidad de Be4SeD, se evaluó el algoritmo de descubrimiento propuesto en Hermida et al. (2009). El cual expone una plataforma de descubrimiento de servicios en ambientes ubicuos. La fase de descubrimiento presentada, es soportada por una técnica que realiza un emparejamiento a nivel atómico, comparando actividades básicas BPEL.

La función principal a tener en cuenta para esta evaluación es *basicActivityMatch* que compara las actividades básicas de entrada con las contenidas en un repositorio (Hermida et al., 2009). La función *BasicActivityMatch* (ver Algoritmo 1) toma como entradas dos nodos, que representan actividades básicas de BPEL (*receive, invoke, reply*), y calcula la distancia semántica entre los dos. Cada nodo posee dos atributos *Operación* (Op) y el *PortType* (PT). La función de emparejamiento prioriza la comparación de la *Operación*, si las dos operaciones son similares (SimOperation > 0) se calcula la similitud del *PortType* y se estima la distancia entre las dos actividades (DistanceNode). Los pesos *Wop* y *Wpt* indican la contribución de la similitud de *Operación* y *PortType* a la similitud de las actividades (0 = Wop = 1 y 0 = Wpt = 1). Para calcular la similitud de los atributos se emplea la función *LS*.

Algoritmo 1. Algoritmo de BasicActivityMatch

INPUTS: ($Node_i$, $Node_j$)
$Node_i$: Struct (Op_i, PT_i), $Node_j$: Struct (Op_j, Pt_j)
OUTPUT: *DistanceNode*
BEGIN
Calculate Operation Similarity $SimOperation = LS(Op_i, Op_j)$
if SimOperation = 0 (different Operations) **then**
Return *DistanceNode* = 1
else
Calculate Port Type Similarity $SimPortType = LS(PT_i, PT_j)$
Calculate *DistanceNode*

$$DistanceNode = 1 - \frac{w_{op} * SimOperation + w_{pt} * SimPortType}{w_{op} + w_{pt}}$$

end if
END

La implementación de Hermida et al. (2009) consultó los parámetros correspondientes a los nodos de las parejas del banco de servicios, evaluó su similitud y utilizó el módulo de recolección de información para almacenar ese resultado. Después de realizar esto para todas las parejas, el módulo de análisis de Be4SeD, a través de la interfaz de usuario, entregó la información de calidad del algoritmo evaluado. Este último proceso se encargó de comparar los resultados del algoritmo de emparejamiento contra el *Benchmark de Referencia*, "Verdad Absoluta", y mostró al usuario estadísticas sobre la evaluación de su algoritmo, utilizando medidas como *recall, precision, overall, k-precision y p-precision*.

6. Resultados

Esta sección presenta los resultados de la evaluación hecha por la plataforma Be4SeD al algoritmo de Hermida et al. (2009), considerando las medidas presentadas en la sección 3.4 las cuales son actualmente empleadas en múltiples investigaciones, cuyo principal objetivo es evaluar la calidad de las estrategias de búsqueda y recuperación de información. Estos resultados permiten estimar la fiabilidad de los servicios recuperados por el algoritmo evaluado, especificar el porcentaje de los servicios relevantes entregados y por ende determinar la calidad de la técnica empleada en Hermida et al. (2009). A continuación se expone la forma como se podría interpretar y analizar los datos arrojados por la plataforma Be4SeD.

La gráfica de *P-Precision* determina cuantos de los servicios recuperados están en la misma posición del ranking de referencia. Para esta medida las curvas de la Figura 6 son decrecientes a medida que se incrementa el número de actividades *k*, comportamiento presentado en los tres escenarios de evaluación (igual dominio, dominio diferente y todos los dominios). La medida de precision con respecto al número de actividades recuperadas por el algoritmo, es presentada en la Figura 7 y determina el conjunto de servicios relevantes para un *top k* de servicios recuperados. El desempeño total del sistema se puede apreciar en la Figura 8, donde se identifica el umbral de similitud óptimo para el algoritmo de emparejamiento. En la Figura 9, se observa la relación entre las medidas de *precision y recall* para los tres escenarios planteados.

De la información arrojada por Be4SeD se concluye que: el umbral de similitud óptimo del algoritmo evaluado equivale a 4,41, valor en el cual se alcanza el máximo desempeño del sistema de recuperación, se evidenció además que el desempeño es mejor cuando se emplean valores de similitud superiores al umbral, ya que en valores bajos de similitud la medida de *recall*, es muy pobre al descartar demasiadas actividades consideradas como relevantes. A partir de este umbral se concluye que el número k de actividades debe estar entre 1 y 7, rango en el que los valores de *precision* son adecuados para un desempeño óptimo. Este rango de valores para el *k* y la similitud, permiten establecer los parámetros para obtener las actividades más relevantes para el usuario, teniendo en cuenta el documento BPEL de consulta.

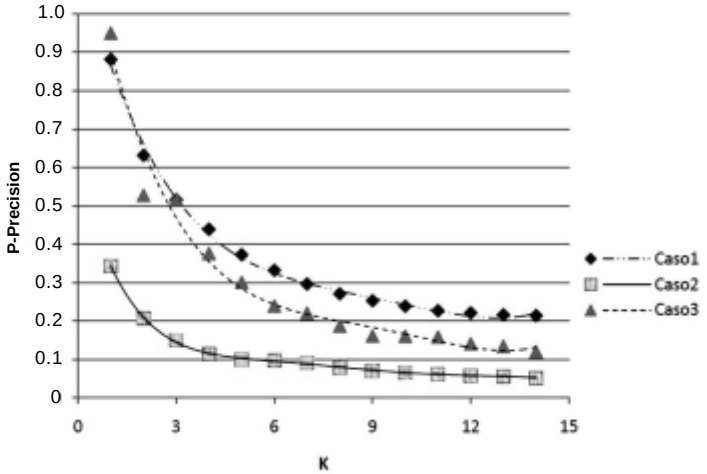

Figura 6. Análisis estadístico de la relación P-Presicion vs. K

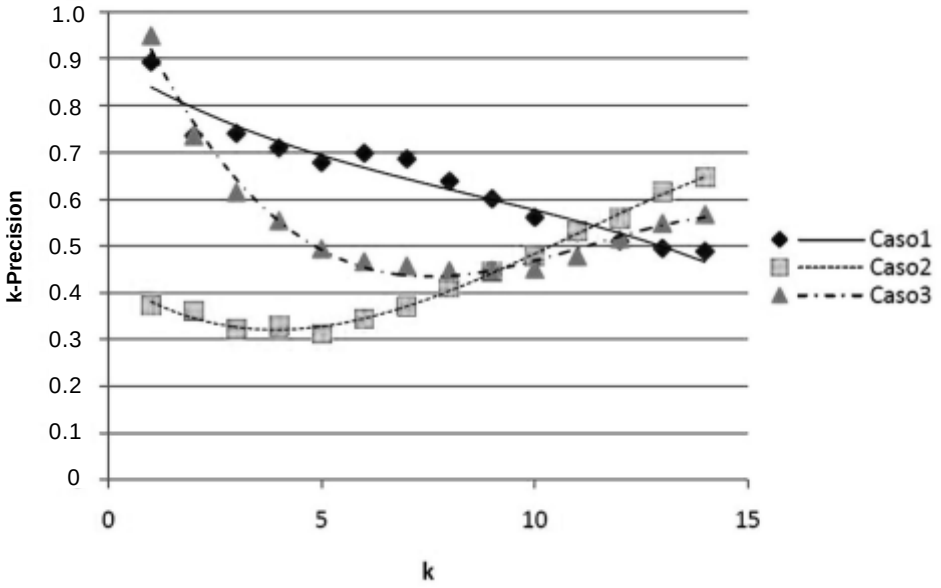

Figura 7. Análisis estadístico de la relación K-Presicion vs. K

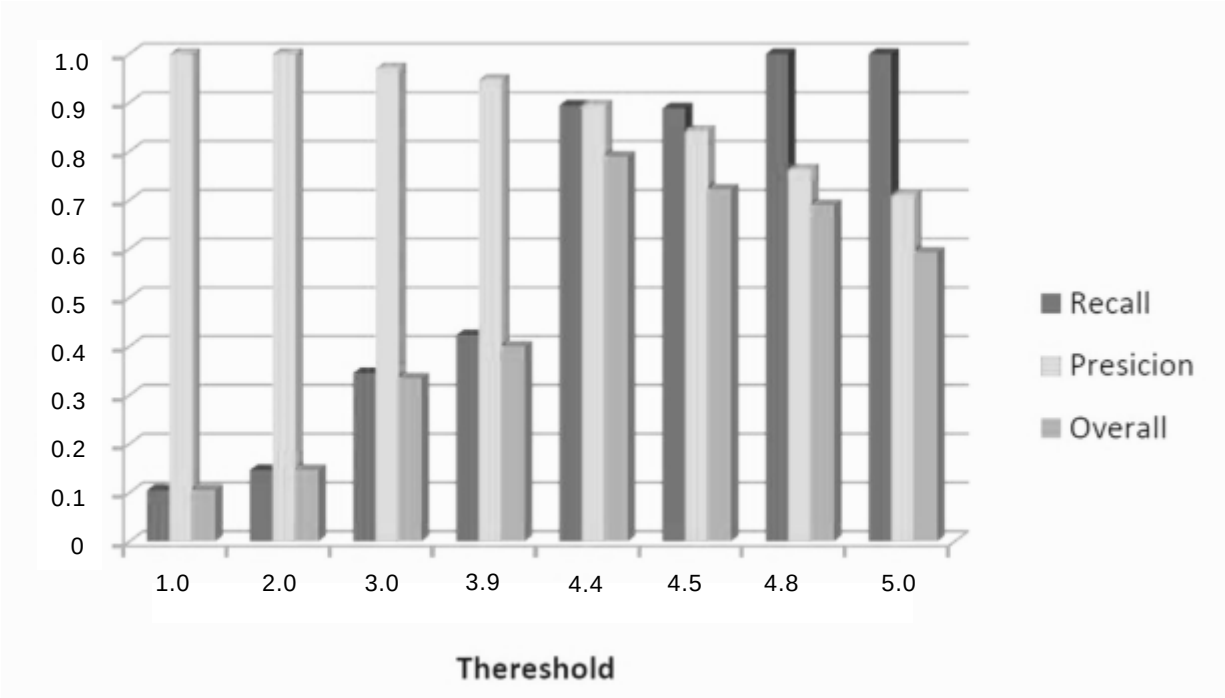

Figura 8. Análisis estadístico para determinar el Umbral de Similitud Óptimo

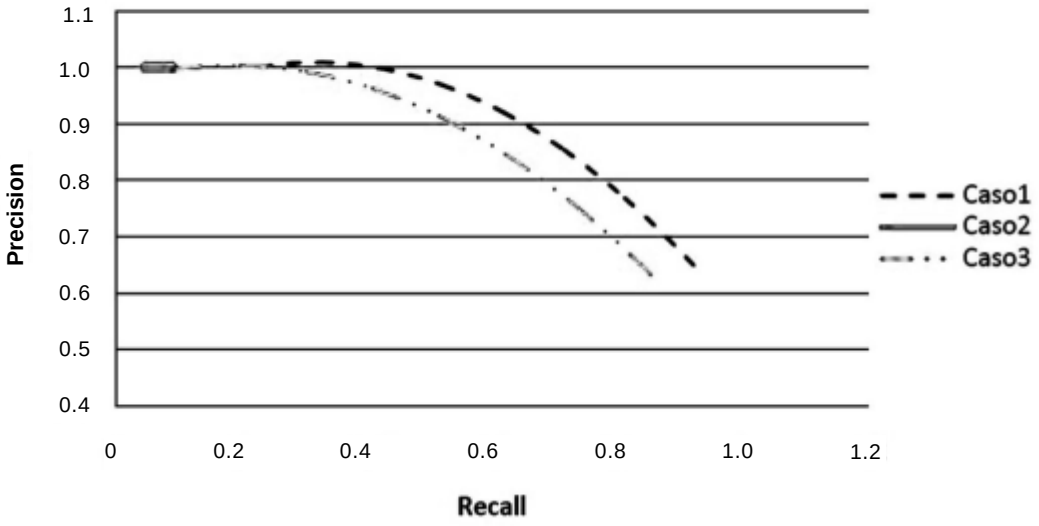

Figura 9. Análisis estadístico de la Relación entre las medidas de precision y recall para los tres escenarios planteados.

Las estadísticas presentadas por Be4SeD son de gran utilidad para determinar si la solución propuesta en Hermida et al. (2009) obtiene el comportamiento esperado, o si necesita ajustar su algoritmo de emparejamiento para lograr mejores resultados.

7. Conclusiones

En este artículo se presenta una plataforma que permite evaluar la eficacia de las técnicas de recuperación de servicios, contrastando sus resultados con un *Benchamark de Referencia*, producto de comparaciones manuales realizadas por evaluadores expertos en el tema. Este *Benchamark de Referencia* sirve como punto de comparación en el proceso de *Benchmarking*, y depende directamente del criterio de evaluadores expertos en el tema de descubrimiento de servicios.

La aplicación de la metodología de benchmarking en la evaluación de técnicas de recuperación de servicios, llevó a adaptar nuevas medidas, propias de éste campo de estudio como: *recall, precision, overall, top-k precision* y *p-precision*, demostrando la flexibilidad y por ende, gran utilidad de esta metodología en diversos entornos. Así, los resultados del presente trabajo, han permitido su adopción al interior del Grupo de Investigación en Ingeniería Telemática de la Universidad del Cauca, para evaluar algoritmos de recuperación de servicios; proporcionando medidas, útiles al momento de tomar decisiones como: ajustar sus pesos o reestructuración del algoritmo, para mejorar su calidad; facilitando su optimización, por medio de la comparación objetiva de versiones del mismo. Como en el caso de estudio presentado.

Este trabajo presenta un *Benchmarking* público con fines académicos, orientado a fortalecer un campo del conocimiento, brindando la posibilidad de que nuevas investigaciones puedan enriquecer este proceso, y lograr la construcción de una guía que permita avanzar de manera efectiva, en pro de perfeccionar los algoritmos de emparejamiento de servicios, y alcanzar metodologías de evaluación estándar.

8. Referencias

Albert, R., & Barabasi, A. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics* 74(1), 47-95.

Corrales, J.C. (2008). *Behavioral matchmaking for service retrieval.* Doctoral Thesis, Department of Computer Science, University of Versailles Saint-Quentin-en-Yvelines, Versailles, France.

Corrales, J.C., Grigori, D., Bouzeghoub, M. & Burbano, J.E. (2008). *Bematch: A platform for matchmaking service behavior models*. In Proceedings of EDBT, 695-699.

ECIR (30th European Conference on Information Retrieval).(2008). *Workshop on Novel Methodologies for Evaluation in Information Retrieval*. Glasgow, United Kingdom.

Egghe, L. (2008). The measures precision, recall, fallout and miss as a function of the number of retrieved documents and their mutual interrelations. *Information Processing and Management* 44(2), Diepenbeek, Bélgica, 856-876.

Hermida, V., Caicedo, O., Corrales, J.C., Grigori, D. & Bouzeghoub, M. (2009**).** *Service Composition Platform for Ubiquitous Environments Based on Service and Context Matchmaking*. En el Cuarto Congreso Colombiano de Computación 4CCC, Bucaramanga, Colombia.

Lewis, D. (1992). *Representation and learning in information retrieval. Doctoral Thesis,* Department of Computer and Information Science, University of Massachusetts, USA.

Martínez, F. J. (2004). *Recuperación de Información: Modelos, Sistemas y Evaluación.* Murcia: JMC Kiosko Ediciones.

Mendling, J. & Ziemann, J. (2005). *Transformation of bpel processes to epcs*, In Proceedings of PK2005, Hamburg, Germany, 41-53.

S3 (Annual International Contest S3 on Semantic Service Selection Retrieval Performance). (2008). *Retrieved from Evaluation of Matchmakers for Semantic Web Services.* Http://www-ags.dfki.uni-sb.de/~klusch/s3/html/2008.html

Seog-Chan, Oh., & Lee, D. (2009). WSBen: A Web Services Discovery and Composition Benchmark Toolkit. *Journal of Web Services Research (JWSR)* 6(1),1-19.

Toma, I., Iqbal, K., Roman, D., Strang, T., Fensel, D., Sapkota, B., Moran, M., Gomez, J. (2007). Discovery in grid and web services environments: A survey and evaluation. *Journal on Multiagent and Grid Systems* 3(3), 341-352.

Küster, U., König-Ries, B. (2009). *Relevance Judgments for Web Services Retrieval - A Methodology and Test Collection for SWS Discovery Evaluation*. In Proceedings of the 7th IEEE European Conference on Web Services. Einhoven, The Netherlands,17-26.

Küster, U., Lausen, H., König-Ries, B. (2007). *Evaluation of Semantic Service Discovery - A Survey and Directions for Future Research*. In Proceedings of the 2nd Workshop on Emerging Web Services Technology (WEWST07) in conjunction with the 5th IEEE European Conference on Web Services (ECOWS07). Halle (Saale), Germany.

Vanhatalo, J., Koehler, J. & Leymann, F. (2006). *Repository for business processes and arbitrary associated metadata*. In Proceedings of the BPM Demo Session at the Fourth International Conference on Business Process Management. Vienna, Austria, 25-31

Voorhees, E. (2001). *The philosophy of information retrieval evaluation*. In Evaluation of Cross-Language Information Retrieval Systems Second Workshop of the Cross-Language Evaluation Forum. Darmstadt, Germany, 355-370.

Zhang, Y. , Dong, X. , Halevy, A., Madhavan, J., Nemes, E. (2004.) *Similarity Search for Web Services*. In Proceedings of the 30th VLDB conference. Vol. 30. Toronto. 372-383.