PLATAFORMA PARA LA EVALUACIÓN DE SISTEMAS DE RECUPERACIÓN DE SERVICIOS BASADOS EN COMPORTAMIENTO

LAURA SANDINO PERDOMO

UNIVERSIDAD DEL CAUCA
FACULTAD DE INGENIERÍA ELECTRÓNICA Y TELECOMUNICACIONES
DEPARTAMENTO DE TELEMÁTICA
GRUPO DE INGENIERÍA TELEMÁTICA
POPAYÁN, SEPTIEMBRE DE 2010

PLATAFORMA PARA LA EVALUACIÓN DE SISTEMAS DE RECUPERACIÓN DE SERVICIOS BASADOS EN COMPORTAMIENTO

LAURA SANDINO PERDOMO

Documento final de trabajo de grado presentado como requisito para obtener el título de Ingeniera en Electrónica y telecomunicaciones

Director:
Juan Carlos Corrales Muñoz

UNIVERSIDAD DEL CAUCA
FACULTAD DE INGENIERÍA ELECTRÓNICA Y TELECOMUNICACIONES
DEPARTAMENTO DE TELEMÁTICA
GRUPO DE INGENIERÍA TELEMÁTICA
POPAYÁN, SEPTIEMBRE DE 2010

TABLA DE CONTENIDOS

LISTA DE TABLAS

LISTA DE FIGURAS

CAPÍTULO 1

INTRODUCCIÓN

1.1 CONTEXTO

Los Sistemas para la Recuperación de Información (SRI), desde su inicios hacia los años setenta con la aparición de los proyectos de Cranfield [CLE72] han experimentado grandes transformaciones, y en gran medida esta evolución se debe a la necesidad de satisfacer las exigencias informativas de sus usuarios, de forma que su aplicación provea respuestas acertadas en tiempos cada vez más cortos. La importante acogida de este tipo de herramientas cada vez más difundidas en la Web, fue la inspiración para el desarrollo de algoritmos para descubrir automáticamente en fuentes de datos, algo más que sólo documentos. Este es el caso de los servicios web, y más tarde los procesos de negocio.

Los servicios Web usados recientemente para implementar SOA (Arquitectura Orientada a Servicios) [RAG05], han hecho posible construir y ejecutar procesos de negocio invocando servicios dinámicamente sobre Internet. Para ello, se han invertido esfuerzos en la investigación y desarrollo de sistemas para el descubrimiento de este tipo de componentes, con el fin de facilitar y agilizar el desarrollo de nuevos procesos.

En sus comienzos, los desarrolladores de herramientas para el descubrimiento de flujos de trabajo (workflows) [BKB05, GLG06, WOR06], investigaron los criterios utilizados por los humanos para comparar la similitud entre secuencias de procesos, mediante experimentos en los cuales realizaran esta tarea y expresaran las razones que motivaron sus juicios. De esta manera, se obtuvieron diversos conjuntos de características importantes para evaluar semejanzas, y una idea de los procesos cognitivos aplicados en esta actividad. Estos experimentos, tenían el objetivo de crear modelos automáticos que homologaran las comparaciones humanas con el fin de optimizar el tiempo empleado en ellas. A partir de estas experiencias, se han implementado nuevos sistemas de recuperación de procesos de negocio [COR08, FIC10] que aplican los conocimientos previos en este dominio, para la generación de sus técnicas de búsqueda y comparación.

Al igual que los SRI se convirtieron en el impulsor de nuevas técnicas de recuperación, sus modelos de evaluación son la guía para la valoración de la efectividad de ellas. Como consecuencia, se observa que los criterios y las medidas empleadas en la estimación del funcionamiento de los sistemas de recuperación de

cualquier tipo, son en general las mismas. Por ejemplo, a lo largo de toda la investigación que recopila este documento, podrá apreciarse que el cálculo de parámetros que describen la efectividad de la recuperación de los sistemas en términos de la certitud de los elementos recuperados, es el método más difundido para conocer la calidad y el nivel de satisfacción provista por una herramienta de este tipo, y encuentra sus cimientos en la evaluación de los SRI.

Como todo proyecto que desee evolucionar, las herramientas para la recuperación de procesos de negocio deben ser evaluadas para encontrar sus inconsistencias, corregirlas y garantizar el correcto cumplimiento de su objetivo fundamental, que es proveer a las empresas un punto de partida en la automatización de sus procesos mediante la reutilización de aquellos que sean útiles en sus propósitos.

En conclusión, la aparición de los sistemas de recuperación desde los SRI hasta los de servicios web y procesos de negocio, han generado la necesidad de crear modelos de evaluación que permitan calcular su utilidad y ayuden a identificar los aspectos que deben mejorarse.

1.2 ESCENARIOS DE MOTIVACIÓN

Como puede intuirse, la generación de modelos de evaluación que faciliten la valoración del desempeño de las herramientas automáticas para el descubrimiento de procesos de negocio, es la base fundamental para garantizar su utilidad y por ende contribuir en la competitividad de las empresas que se benefician de ellas en la implementación de sus procesos, y adicionalmente en beneficios económicos y reconocimiento a los desarrolladores de los sistemas.

La evaluación de las herramientas para la recuperación de procesos de negocio es un campo de investigación estrechamente ligado al desarrollo de estos sistemas. Gracias a este mecanismo puede garantizarse su mejoramiento y posterior difusión, ya que los beneficios percibidos por un usuario, motivarán su utilización por parte de muchos más, quienes darán pie a nuevos requerimientos que repercuten en su evolución.

El grupo de Ingeniería Telemática (GIT) de la facultad de Ingeniería Electrónica y Telecomunicaciones de la Universidad del Cauca, ha enfocado gran parte de sus esfuerzos en la investigación de técnicas de recuperación de procesos de negocio, y en desarrollo de sistemas que las apliquen [FIC10]. Desafortunadamente, no se cuenta con una herramienta que permita aplicar modelos de evaluación de su efectividad y motiven su progreso. Por esta razón, el trabajo presentado en este documento, suple esta necesidad mediante la creación de una plataforma para la

evaluación manual de procesos, y la generación de un modelo de evaluación que beneficie a los trabajos relacionados con la implementación de sistemas de recuperación de este tipo.

1.3 PLANTEAMIENTO DEL PROBLEMA

Como se mencionó con anterioridad, la evaluación del desempeño de los sistemas para la recuperación de procesos de negocio, facilita el hallazgo de falencias funcionales, garantiza una constante evolución y su mejora continua, y asegura la satisfacción de las necesidades de sus usuarios. Por estas razones, cualquier herramienta para el descubrimiento de procesos que desee mantenerse vigente, deberá actualizarse constantemente y someterse a valoraciones de su rendimiento en términos de la eficacia de ejecución, el efectivo almacenamiento de los datos y la efectividad de la recuperación [BYF99]. Como puede intuirse, los dos primeros parámetros son fácilmente apreciables, no obstante, el tercero requiere del previo conocimiento del conjunto de elementos que deberían ser recuperados a partir de una consulta, para así determinar la relevancia de los resultados obtenidos del sistema.

De esta manera, la medición de calidad de la recuperación de un sistema es un punto importante en la evaluación de su desempeño, y para llevarla a cabo es necesario testear la herramienta sobre una colección de prueba de la cual se conozcan los resultados deseables. Desafortunadamente, las colecciones de procesos de negocio para los fines mencionados, no son asequibles o están en desarrollo [KUK08].

En la actualidad existen grupos y conferencias como el *S3 Contest*[1], el *SWS Challenge*[2] y FUSION[3] de la universidad de Jena en Alemania, que trabajan en la generación y aplicación de modelos de evaluación comparativa entre sistemas de recuperación de servicios web, y en la construcción de colecciones de servicios de prueba descritos en diversos lenguajes, sobre las cuales aplicar los modelos. No obstante, para participar en las actividades programadas por dichos grupos, es necesario tener un nivel de competitividad que sólo se logra mediante pruebas preliminares.

En consecuencia, para evaluar la efectividad de la recuperación de un sistema para el descubrimiento de procesos de negocio, se requiere un conjunto de procesos de prueba sobre los cuales ejecutar el algoritmo a evaluar. Adicionalmente, deben

[1] http://www-ags.dfki.uni-sb.de/~klusch/s3/index.html
[2] http://sws-challenge.org/wiki/index.php/Main_Page
[3] http://fusion.cs.uni-jena.de/professur

establecerse unos elementos dentro de la colección que servirán como consultas; y finalmente, es necesario obtener el subconjunto de procesos similares para cada una de ellas. Esta última tarea se realiza con la ayuda de unos expertos en este dominio, quienes determinarán el nivel de semejanza existente entre las consultas y cada elemento de la colección; y de esta manera adquirir los resultados que un algoritmo debe retornar para ser considerado efectivo. No obstante, no existe una herramienta que permita hacer las comparaciones manuales de procesos, que almacene la información recopilada, y que provea la clasificación de elementos relevantes para cada consulta según la evaluación intuitiva.

1.4 PROPUESTA

El trabajo presentado en esta monografía, propone generar y aplicar un modelo de evaluación para sistemas de recuperación de procesos de negocio, que comprenda la recopilación de conjuntos de procesos de prueba, su evaluación manual de similitud, el cotejo de dichos resultados con los obtenidos automáticamente por la herramienta de Figueroa [FIC10], y la aplicación de medidas que permitan valorar la calidad de la recuperación de esta última.

1.5 CONTRIBUCIONES

Las contribuciones de este proyecto son las siguientes:

- **Colección de procesos de negocio de prueba:** Para la evaluación de la herramienta automática [FIC10] era necesario poseer un conjunto de procesos sobre los cuales aplicar la herramienta, por esta razón, se creó una colección de prueba compuesta por 40 procesos del dominio del geoprocesamiento, y 60 del dominio de las telecomunicaciones.

- **Criterios para la evaluación manual de similitud entre procesos:** Debido a que la efectividad de la recuperación de un sistema para el descubrimiento de procesos se mide de acuerdo con la relación existente entre sus resultados y los esperados por los usuarios, fue necesario pedirle a un grupo de personas que extrajeran de la colección de procesos de negocio de prueba, aquellos que presentaban algún nivel de similitud con unos procesos de consulta establecidos. Para facilitar esta tarea, se investigaron y proporcionaron unos criterios de comparación teniendo en cuenta las características fundamentales en la evaluación de semejanzas entre flujos de trabajo.

- **Plataforma para la evaluación intuitiva de procesos de negocio:** Se creó una aplicación web para la comparación intuitiva de procesos de negocio, que permite

la recopilación de los elementos de la colección de prueba, la selección de los procesos que servirán de consulta, la evaluación manual de similitud entre procesos mediante la aplicación de los criterios mencionados anteriormente, y la obtención de clasificaciones jerarquizadas de los elementos similares para cada consulta de acuerdo con los juicios de semejanza proveídos por los evaluadores.

- **Conjunto de medidas para la evaluación de la calidad de la recuperación de procesos de negocio:** Se investigaron y seleccionaron las medidas para evaluar sistemas de recuperación de procesos de negocio de forma que se compararan los resultados obtenidos automáticamente, con aquellos provistos por las evaluaciones manuales. De esta manera, se conoce por medio de la valoración de estos parámetros el nivel de acierto en la recuperación de un sistema.

- **Modelo de evaluación:** Se generó un modelo de evaluación que incluyó la generación de los procesos para la colección de prueba, su adición a la plataforma para la evaluación manual, la selección de 7 procesos establecidos como consultas, la comparación intuitiva de procesos para hallar los elementos de la colección con algún nivel de similitud para cada consulta, la jerarquización de los resultados de acuerdo con los valores de semejanza obtenidos, la aplicación de la herramienta automática [FIC10] sobre la misma colección de prueba, y finalmente la aplicación de las medidas para determinar la calidad de su recuperación.

- **Enriquecimiento de la base de conocimiento existente correspondiente a la evaluación de sistemas para la recuperación de procesos de negocio:** Se provee toda la investigación referente a la evaluación de los SRI, y su influencia en los mecanismos para este mismo propósito utilizados en la recuperación de procesos de negocio. Adicionalmente, se proporciona la base documental que sustenta el modelo de evaluación propuesto y aplicado en el desarrollo de este trabajo.

1.6 CONTENIDO DEL DOCUMENTO

El documento que se presenta a continuación consta de 6 capítulos adicionales:

- **Capítulo 2:** Presenta un resumen acerca de la evolución y el estado en el que se encuentran los mecanismos de evaluación para los SIR y los sistemas para la recuperación de servicios web. Adicionalmente, hace un recuento de algunos trabajos relacionados, sus ventajas, desventajas y aportes a este trabajo.

- **Capítulo 3:** Hace un recuento de los criterios utilizados para evaluar la recuperación de los SRI y las herramientas para el descubrimiento de servicios web; y finaliza con la selección y explicación de los criterios empleados en la

evaluación manual de similitud de procesos aplicada en la plataforma desarrollada en este proyecto.

- **Capítulo 4:** Explica las medidas utilizadas comúnmente en la evaluación de la calidad de la recuperación de los SRI y los sistemas de recuperación de servicios web. Analiza la conveniencia de la utilización de estas medidas, y presenta la selección de las que se utilizaron en la evaluación de la herramienta automática [FIC10].

- **Capítulo 5:** Explica la arquitectura y el funcionamiento de la plataforma para la evaluación manual de similitud entre procesos de negocio desarrollada en este trabajo de grado.

- **Capítulo 6:** Presenta los resultados obtenidos de las evaluaciones manuales y de la aplicación de la herramienta automática, muestra los valores obtenidos de las medidas de calidad de la recuperación, y analiza la efectividad del sistema de acuerdo a sus resultados.

- **Capítulo 7:** Expone las conclusiones de la aplicación del modelo de evaluación, las recomendaciones y trabajos futuros.

CAPÍTULO 2

ESTADO DEL ARTE

2.1 CONTEXTO GENERAL

Una revolución tecnológica de gran envergadura ha surgido a partir de la aparición de las tecnologías de la información y de las comunicaciones (TIC), adquiriendo gran importancia y convirtiéndose en pieza clave de un proceso de transformación económica, ya que de la rápida aparición de innovaciones de procesos y productos en materia digital, han surgido nuevas actividades productivas.

La confluencia y las interrelaciones entre el proceso de digitalización y sus usos productivos, generan un flujo interactivo entre la demanda y la oferta, a través de mecanismos innovadores. Esto, a su vez, produce importantes aumentos de productividad y competitividad de economías, sectores o empresas. Es por ello que las empresas buscan crear, adaptar, modificar o integrar componentes existentes de manera rápida y fiable con el fin de ofrecer nuevos servicios que satisfagan las exigencias de la demanda [FIC10]. Con este fin, diversos estudios han sido dirigidos hacia la creación de herramientas para el descubrimiento de componentes reutilizables, que permitan reducir el tiempo de despliegue de nuevos servicios.

Los sistemas de recuperación de servicios deben ser, como cualquier sistema, sometidos a procesos de verificación para que los usuarios puedan evaluar su efectividad, logren detectar ineficiencias y consigan plantear mejoras encaminadas a su evolución. De ahí la importancia de aplicar mecanismos y medidas que evalúen el desempeño de estas herramientas en todos los campos que sea posible. Ya que los sistemas de recuperación de información (SRI) y su evaluación tienen una trayectoria amplia y conocida, y han sido punto de partida para las herramientas de recuperación en general, la primera parte de este capítulo recopilará información acerca de la evaluación de los SRI. A continuación, se mostrará el estado en el que se encuentra la evaluación de los mecanismos para el descubrimiento de servicios web y, finalmente, se referenciarán trabajos afines a éste.

2.1.1 Evaluación de sistemas de recuperación de Información

La Recuperación de Información (RI) se puede definir como el problema de la selección de información, depositada en un medio de almacenamiento, en

respuesta a consultas realizadas por un usuario [LOP06, BYF99]. Los sistemas de recuperación de información, son herramientas que acceden a bases de datos de documentos mediante consultas que reflejan la necesidad de información de un usuario, y que como resultado obtienen elementos relevantes que satisfacen la consulta. La evaluación de estos sistemas se ha desarrollado paralelamente a su evolución encontrándose estrechamente relacionados con la investigación y el desarrollo de la recuperación de información.

De esta forma, varios estudios se han dirigido a plantear la mejor forma de evaluar dichos sistemas; por ejemplo, Baeza-Yates [BYF99] manifiesta que los sistemas de recuperación de información pueden ser evaluados por diversos criterios como su *eficacia de ejecución*, el *efectivo almacenamiento de los datos*, y *la efectividad de la recuperación de la información*.

La eficacia de ejecución se mide por el tiempo que le toma al sistema o una parte del mismo llevar a cabo una operación. Este es un parámetro importante porque un tiempo excesivo de recuperación puede provocar que un usuario deje de utilizar un sistema. Por otra parte, la eficiencia del sistema se evalúa de acuerdo al número de bytes que se precisan para almacenar los datos. Finalmente, la efectividad de la recuperación se basa en la satisfacción de la necesidad real de información de un usuario, haciendo de esta una medida subjetiva [LOP06].

Borlund [BOR00] por su parte, diferencia entre evaluar el *acceso físico* que es el que concierne a cómo la información es recuperada y representada de forma física al usuario, y el *acceso lógico* a los datos que está relacionado con la localización de la información deseada. El primer caso está muy vinculado con las técnicas de recuperación y de presentación de la información, mientras que el segundo tiene que ver con la *relevancia* del objeto localizado con una determinada petición de información. De forma parecida, Baeza-Yates [BYF99] afirma que existen dos tipos de evaluaciones: la del funcionamiento del sistema y la de calidad de la recuperación, siendo la segunda modalidad la que analiza cómo los documentos recuperados se clasifican de acuerdo a su *relevancia* con la pregunta efectuada.

Teniendo en cuenta las características que un SRI debe ofrecer a los usuarios, Rijsbergen [RIJ99] analiza seis medidas principales que permiten cuantificar su desempeño: “la cobertura de una colección; el tiempo de respuesta del sistema a una petición; la forma de presentación de los resultados; el esfuerzo realizado por el usuario; la *exhaustividad* del sistema y la *precisión* del sistema”. Rijsbergen opina que la proporción de material relevante recuperado como respuesta a una petición de búsqueda (*exhaustividad ó recall*), y la proporción de material recuperado que es realmente relevante (*precisión*), son los parámetros que verdaderamente

pretenden medir la *efectividad* de los sistemas, siendo esta una medida de la capacidad del sistema para satisfacer al usuario en términos de la *relevancia* de los documentos recuperados.

En la actualidad existen dos corrientes de investigación enfocadas a la evaluación de los SRI, la primera se centra en la evaluación de los algoritmos y en las estructuras de datos necesarias para optimizar la eficacia y la eficiencia de las búsquedas en bases de datos documentales; y la segunda considera el papel del usuario y de las fuentes de conocimiento implicadas en la RI [INW95].

En la corriente algorítmica, las conferencias TREC[4] (Text REtrieval Conference), se han convertido en el foro de intercambio científico más prestigioso del campo de la recuperación de información y en consecuencia en el eje central donde gravita la evolución de los SRI. TREC reúne a creadores de diferentes sistemas y compara los resultados que éstos obtienen en diferentes pruebas, previamente estandarizadas y acordadas por todos. La primera conferencia, TREC-1 (1992), presentó como resultado principal la existencia de una amplia similitud entre los SRI que hacen uso de técnicas basadas en lenguaje natural y los basados en el modelo probabilístico y los basados en el modelo del vector. En la conferencia TREC-2 (1993), se detectó una significativa mejoría de la recuperación de información, con respecto a la anterior. Las siguientes conferencias aportaron nuevas prestaciones a los experimentos: localización de información en varias bases de datos de manera simultánea, presencia de errores ortográficos con el fin de valorar el comportamiento de los SRI ante ellos y recuperación de información en idiomas distintos del Inglés -se eligieron el Español y el Chino- para valorar los posibles cambios de comportamiento de los SRI [MMR04].

Otros foros de evaluación con relevancia internacional en el estudio y evaluación de los sistemas RI son: MUC[5] (Message Understanding Conferences), SUMMAC[6] (Summarization Conference), CLEF[7] (Cross Language Evaluation Forum), entre otros.

En la segunda corriente, la evaluación se centra en la representación de los problemas de información, el comportamiento en las búsquedas y los componentes humanos de los SRI en situaciones reales, y se fundamenta en la psicología cognitiva y en las ciencias sociales. La búsqueda de información y la formulación de la necesidad de información se contemplan como procesos cognitivos del usuario individual, siendo el SRI y los intermediarios funcionales

[4] http://trec.nist.gov/
[5] http://evaluacion-recuperacion.iespana.es/
[6] http://www-nlpir.nist.gov/related_projects/tipster_summac/index.html
[7] http://www.clef-campaign.org/

(como la interfaz del sistema) componentes fundamentales de este proceso de contextualización [OLV99].

En el marco del modelo cognitivo, se han propuesto distintas medidas de evaluación del SRI relacionadas con el concepto de relevancia basada en el usuario. Entre ellas se encuentran la proporción de cobertura o alcance ("coverage ratio"), definida como la fracción de documentos relevantes conocidos por el usuario que han sido recuperados, y la proporción de novedad ("novelty ratio"), que se define como la fracción de documentos relevantes recuperados que son desconocidos por el usuario [BYF99, KOR97]. También se ha considerado la satisfacción del usuario como medida de la eficacia del SRI en este marco de trabajo [KOR97], aun cuando no se ha propuesto una forma adecuada para medirla por tratarse de un criterio muy subjetivo. Precisamente, esta última es la mayor crítica al modelo cognitivo, que también utiliza otras medidas como beneficios y frustraciones, utilidad, etc. que no son objetivas y que no evalúan directamente el sistema sino el efecto que provoca en el usuario [LOP06].

2.1.2 Evaluación de sistemas de recuperación de Servicios Web

La reutilización de servicios Web dispersos en la Internet, pretende facilitar a las empresas el despliegue de nuevos servicios en forma rápida y confiable mediante la creación, modificación, adaptación e integración de dichos componentes para desarrollar tareas más complejas. Con el propósito de agilizar el hallazgo de los elementos necesarios para implementar los procesos de negocio de las empresas, diversos estudios [FIC10, DDG09] se han dedicado a encontrar las mejores formas para hacer búsquedas y comparaciones de similitud entre servicios web, con el fin de encontrar a partir de una petición, aquel o aquellos servicios que la satisfacen. Así, se han ido planteando diferentes técnicas de emparejamiento y, con ellas, herramientas automáticas para el descubrimiento de servicios Web.

Los sistemas tradicionales de RI normalmente crean el modelo sobre el cual operan de manera autónoma, por lo tanto, desde el punto de vista de la evaluación ellos trabajan sobre los datos originales y, en consecuencia, diferentes sistemas de RI pueden ser evaluados sobre un mismo banco de datos de prueba (como una colección de documentos) [KUK09]. Sin embargo, en los sistemas de recuperación de servicios, el modelo se crea a partir de las anotaciones semánticas que son escritas por personas expertas para obtener una recuperación eficiente y precisa. De hecho, no existe un acuerdo acerca de qué formalismo y modelo semántico (WSMO[8], OWL-S[9], SAWSDL[10]) ofrece la mejor expresividad, usabilidad y complejidad computacional.

[8] http://www.w3.org/Submission/WSMO/

En los últimos años, se han invertido grandes esfuerzos y dinero en la investigación de técnicas para el descubrimiento de servicios web semánticos obteniendo herramientas cada vez más sofisticadas y maduras. No obstante, deviene sorprendente el poco esfuerzo puesto en la evaluación de estas propuestas. Las razones que argumentan en [KUK08] es que no se han establecido metodologías de evaluación teóricamente bien fundadas, y que los bancos de pruebas para las evaluaciones comparativas dependen en gran medida del formalismo utilizado para describir los servicios (WSMO, OWL-S, SAWSDL) lo cual dificulta la aplicación del paradigma de Cranfield [CLE72].

S3 CONTEST[11] *(Sematic Services Selection Contest)*

Proporciona los medios y un foro para la evaluación comparativa del desempeño de sistemas para el descubrimiento de SWS mediante la aplicación del paradigma de Cranfield, y promueve el desarrollo de colecciones de prueba en los formatos más representativos para la descripción de servicios semánticos como OWL-S, WSML y el estándar SA-WSDL.

Una de las principales críticas que ha recibido el S3 se relaciona con la aplicación del paradigma de Cranfield en evaluaciones que impiden la participación de herramientas de emparejamiento de diferentes lenguajes de descripción de servicios [KUK08]. Pero, en los últimos años, el S3 ha unido esfuerzos con otros grupos de trabajo como el SWS Challenge[12] y el proyecto FUSION[13] (FUnctionality Sharing In Open eNvironments) de la Universidad de Jena en Alemania, para la generación de repositorios de servicios para los diferentes lenguajes de descripción, y colecciones de servicios de diferentes campos de aplicación descritos en todos los lenguajes, facilitando la evaluación comparativa de herramientas heterogéneas.

La iniciativa para la formación del S3 (Semantic Service Selection), se presentó en la Quinta Conferencia Internacional de Web Semántica (ISWC, por sus siglas en inglés) realizada en Athens, Estados Unidos en 2006, con el propósito de fomentar el desarrollo rápido e innovador de herramientas para el descubrimiento de servicios semánticos. Desde entonces, proponen anualmente unos procesos de evaluación de diferentes herramientas y presentan sus resultados en la ISWC (International Semantic Web Conference).

[9] http://www.w3.org/Submission/2004/SUBM-OWL-S-20041122/
[10] http://www.w3.org/TR/sawsdl/
[11] http://www-ags.dfki.uni-sb.de/~klusch/s3/index.html
[12] http://sws-challenge.org/wiki/index.php/Main_Page
[13] http://fusion.cs.uni-jena.de/professur

Anualmente, el S3 propone unas actividades a realizar que son evaluadas y socializadas al final, en el Taller Internacional SMR2 (Service Matchmaking and Resource Retrieval in the Semantic Web) del ISWC[14] (International Semantic Web Conference). En 2009 por ejemplo, se propuso evaluar herramientas de emparejamiento OWL-S y SAWSDL sobre repositorios de servicios descritos en su correspondiente lenguaje (OWLS-TC3 y SAWSDL-TC2 respectivamente). Además, sobre la colección Jena Geography Dataset JGD[15] cuyos servicios están descritos cada uno en los lenguajes de prueba, se examinaron herramientas con diferentes técnicas de descubrimiento y distintos lenguajes.

Para 2010, se proponen las mismas actividades incluyendo evaluaciones sobre un nuevo repositorio de servicios descritos en WSML[16]. La presentación final de esta edición se realizará en el ISWC, con la colaboración de la campaña de evaluación de SWS del proyecto SEALS[17] (Semantic Evaluation At Large Scale).

SWS CHALLENGE

El SWS Challenge es otra iniciativa de la comunidad dedicada a evaluar tecnologías de SWS. Sin embargo, no está dirigida a hacer competencias entre las diferentes herramientas, sino que busca constatar sus virtudes y criticar sus deficiencias en la búsqueda de mejoras continuas de cada propuesta. Sin embargo, este grupo se identifica con el S3 Contest en cuanto que ambos trabajan en el desarrollo de colecciones de servicios de prueba que permitan la evaluación de los diferentes formalismos utilizados en las herramientas.

Esta iniciativa organiza una serie de talleres en los que los participantes tratan de modelar y solucionar problemas descritos en las colecciones de prueba que se encuentran disponibles al público pero, el conjunto de servicios descritos hasta el momento es reducido.

2.2 TRABAJOS RELACIONADOS

Pese a los grandes esfuerzos invertidos en el desarrollo de herramientas para el descubrimiento de servicios web, las investigaciones encaminadas al desarrollo de metodologías formales para la evaluación de dichas herramientas son incipientes y aún poco difundidas. A continuación se describen algunos trabajos que propusieron metodologías de evaluación y montajes experimentales para evaluación de herramientas para el descubrimiento de SWS.

[14] http://iswc.semanticweb.org/
[15] http://fusion.cs.uni-jena.de/professur/jgd/
[16] http://www.w3.org/Submission/WSML/
[17] http://www.seals-project.eu/

En [TAH06] se abordan algunos mecanismos sobre la evaluación de la eficacia de la recuperación de los sistemas para el descubrimiento de SWS. Partiendo de que los esquemas de evaluación tradicionales no captan plenamente el valor agregado de la semántica del servicio, y tampoco tienen en cuenta su clasificación (expresado a través del grado de correspondencia DoM), la cual es adoptada por la mayoría de los motores de descubrimiento SWS, el principal objetivo de este trabajo es proponer un sistema de evaluación para el descubrimiento de SWS basado en las teorías de IR.

Para empezar, el concepto de "Grado de Correspondencia" ó DoM (por su nombre en inglés Degree of Match) se puede definir como el valor obtenido de una escala ordenada de valores, que expresa la semejanza entre dos entidades con respecto a una métrica de similitud o relevancia. Esta medida se introdujo para que las herramientas de descubrimiento de servicios, al calcular el DoM, entregaran resultados ordenados de acuerdo al grado de similitud que presenta cada elemento recuperado con la consulta realizada. Para los sistemas de recuperación de servicios este parámetro se define como RSV (Retrieval Status Value) que calcula el grado en el que las representaciones de los documentos satisfacen los requisitos expresados en la consulta y recupera aquellos documentos que son relevantes a la misma [LIH04].

La evaluación de cualquier sistema de recuperación, se sustenta en la medición del *rendimiento* y la *eficacia de recuperación*. El *rendimiento* implica la complejidad computacional, los tiempos de respuesta del sistema, etc.; y la *eficacia de recuperación* evalúa la forma correcta o adecuada en que la herramienta está descubriendo los servicios relevantes que han sido especificados por un experto en el dominio. Este trabajo se enfoca en proponer unas métricas y una metodología para evaluar la eficacia de recuperación de las herramientas para el descubrimiento de SWS.

Un esquema básico de recuperación de servicios incluye un conjunto de servicios ofrecidos S_i, un servicio de consulta proporcionado por el usuario *R*, y una herramienta de emparejamiento encargada de hacer el descubrimiento mediante la asignación de un DoM ***e****(R, S_i)* a cada servicio publicado S_i.

Para evaluar la eficacia de recuperación los autores representan la similitud o relevancia entre un conjunto de servicios de consulta *Q*, y el conjunto de servicios de comparación *S*, como un arreglo de vectores $\boldsymbol{r}: Q * S \rightarrow W \;\; y \;\; \boldsymbol{e}: Q * S \rightarrow W$, donde ***W*** es el conjunto de valores que representan el grado de relevancia para ***r*** (determinado por un experto), ó el grado de correspondencia DoM para ***e*** (dado por la herramienta de descubrimiento). ***W*** puede tomar diferentes tipos de valores: Booleano (W={0,1}), Números reales (W=[0,1]), Términos difusos (W={"relevante", "irrelevante",…}), etc. De esta manera, la evaluación de una

herramienta para el descubrimiento de servicios estará determinada por el nivel de aproximación del vector ***e*** al ***r***.

Teniendo en cuenta lo anterior, los autores presentan dos esquemas de evaluación: El Booleano donde se establecen dos valores, 0 ó 1 para los grados de similitud[18] y de correspondencia[19] entre un servicio de consulta y uno de comparación, siendo "1" cuando los dos servicios presentan algún nivel de afinidad, y "0" cuando no. De acuerdo con esto, la eficacia de una herramienta de descubrimiento estará dada por el número de elementos recuperados (proporcionado por la herramienta) que son relevantes (proporcionados por el experto), y estos valores darán paso a la aplicación de medidas de *exhaustividad y precisión* que se analizarán más adelante. El otro esquema de evaluación propone una escala de valores de similitud (valores numéricos [0,1], términos difusos {"relevante", "irrelevante", etc.}), que permite ordenar los resultados de acuerdo al nivel de semejanza que presentan los servicios de consulta y los de comparación. En este caso, la evaluación se hace de acuerdo a la equivalencia que existe entre el ordenamiento de los servicios proporcionado por el experto, y el obtenido por la herramienta.

En un esquema de evaluación booleano, una herramienta calcula para cada consulta su correspondiente valor de relevancia ***e***: Q × S → {0,1}, donde ***e***(R, S_i) = 1 si el servicio de comparación S_i ∈ S es "relevante" para un servicio de consulta R ∈ Q; y ***e***(R, S_i) = 0 si S_i es "irrelevante" para R. De forma análoga, los expertos asignarán un grado de similitud 1 ó 0 a los servicios de comparación de acuerdo con la consulta especificada.

En este caso, se utilizan las medidas estándar de *Precisión* P_B y *Exhaustividad* R_B [BYF99] para calcular la efectividad de la herramienta. Así pues, la *Precisión* P_B para una consulta dada, se define como el porcentaje del número de servicios relevantes recuperados, sobre el número de servicios recuperados; y la *Exhaustividad* R_B por su parte, se define como el porcentaje del número de servicios relevantes recuperados, sobre el número de servicios relevantes en la colección[20].

$$R_B = \frac{|RT \cap RL|}{|RL|} \; ; \; P_B = \frac{|RT \cap RL|}{|RT|} \qquad \begin{matrix} RT \rightarrow Servicios\ Recuperados \\ RL \rightarrow Servicios\ Relevantes \end{matrix}$$

No obstante, la mayoría de las herramientas para el descubrimiento de servicios SWS soportan DoM múltiples, lo cual obliga a mapear los grados de similitud obtenidos a un esquema binario y, ello implica el establecimiento de un umbral

[18] El grado de similitud entre dos servicios lo proporciona la evaluación manual realizada por un dominio experto.
[19] El grado de correspondencia es el que asigna la herramienta automática a cada servicio recuperado según su similitud con la consulta.
[20] El subíndice B indica que son las medidas para el esquema de evaluación Booleano.

desde el cual, todos los servicios recuperados catalogados por encima de éste son "relevantes" (tendrán un valor de 1), y los que se encuentren por debajo son "irrelevantes" (un valor de 0). El principal problema de esta nueva clasificación, es que al perder los grados de correspondencia, se desperdicia la facultad de la herramienta para explotar la semántica de los servicios, y se corre el riesgo de considerar como "irrelevantes" servicios que podrían tener algún grado de similitud con la consulta.

De otra parte, las evaluaciones booleanas de similitud por parte de los expertos, asignan una relevancia total a todos los servicios que tengan algún nivel de semejanza. Esto, contrasta con uno de los objetivos principales del descubrimiento de SWS que busca una recuperación más eficaz y precisa.

Lo anterior llevó a concluir que, al transformar los resultados de la herramienta automática de una escala de valores múltiples a una booleana, se ignoran el emparejamiento y la semántica de los servicios. Además, que la transformación de estos valores implica el establecimiento de umbral y esta no es una tarea trivial. Por último, que la evaluación booleana de la relevancia no es muy específica y no siempre refleja las intenciones de los expertos.

Para superar estos inconvenientes, este trabajo propuso que los grados de similitud y DoM no fueran medidos a la manera booleana. Así, para los expertos se propuso adoptar un método lingüístico para discriminar los servicios de acuerdo a su relevancia con la consulta, de modo que, un servicio de comparación será "muy relevante", "algo relevante", etc., para la consulta.

En este caso una variable lingüística se caracteriza por una tupla (L, H(L)) [BOP93], donde L es el nombre de la variable ("relevancia"), y H(L) el conjunto valores lingüísticos de L ("relevante", "irrelevante", "muy relevante", etc.). Cada uno de estos valores puede ser caracterizado por una variable difusa *u* cuyo rango está especificado dentro del universo U. El grado de pertenencia de un elemento $u \in U$ se define por una función de pertenencia μ_u, de manera que μ_u: U → [0,1], donde 1 significa que existe una relación y 0 que no la hay. Para la evaluación de los SWS el nombre de la variable lingüística utilizada por el dominio de los expertos es la "relevancia" y el conjunto H("relevancia")={"irrelevante", "poco relevante", "algo relevante", "relevante", "muy relevante"}.

Para medir la eficacia de la recuperación utilizaron una generalización de las medidas de *precisión* y *exhaustividad* presentada en [BUK81], calculadas a partir de los dos rankings de la evaluación de relevancia ***fe*** entregada por la herramienta, y ***fr*** entregada por el experto donde: [21]

[21] El subíndice G se refiere a que son las medidas generalizadas para la precisión y la exhaustividad.

$$R_G = \frac{\sum_{S_i \epsilon S} \min\{f_r(R,S_i), f_e(R,S_i)\}}{\sum_{S_i \epsilon S} f_r(R,S_i)},$$

$$P_G = \frac{\sum_{S_i \epsilon S} \min\{f_r(R,S_i), f_e(R,S_i)\}}{\sum_{S_i \epsilon S} f_e(R,S_i)}$$

$$\boldsymbol{fe}: \mathrm{Q} \times \mathrm{S} \rightarrow [0,1]$$

$$\boldsymbol{fr}: \mathrm{Q} \times \mathrm{S} \rightarrow [0,1]$$

En este caso se toman en cuenta todos los valores proporcionados para las relaciones entre los servicios de consulta y los servicios comparados. Teniendo en cuenta las ecuaciones anteriores, la *precisión* será máxima cuando las estimaciones de relevancia de la herramienta sean más rigurosas haciendo que $min\{f_r f_e\} = f_e$, observándose el caso contrario para la *exhaustividad*.

Para las pruebas, utilizaron la herramienta de código abierto OWLS-MX Matcher [KFK05] en la comparación automática de los servicios. Este sistema establece un DoM entre dos servicios (el de consulta *Q* y el de comparación *S*) como se describe a continuación.

DoM	**Definición (Informal)**
Exacto	Si las entradas y las salidas de ambos procesos (*Q* y *S*) son conceptos equivalentes
Plugin	Si las salidas de *S* son subclases directas de las salidas de *Q* y las entradas de *Q* son subsumidas por las entradas de *S* en la ontología de dominio.
Subsumes (Inclusión)	Si las salidas de *S* son subsumidas por las salidas de *Q* y las entradas de *Q* son subsumidas por las entradas de *S* en la ontología de dominio.
Subsumed- by	Si las salidas de *Q* son subclases directas de las salidas de *S* y las entradas de *Q* son subsumidas por las entradas de *S* en la ontología de dominio.
Fallo	Si ninguno de los criterios anteriores aplica

Tabla 1 Definición de DoM

La herramienta se configuró para aplicar solamente algoritmos de emparejamiento basados en la lógica, y el umbral se estableció en Fallo (Fail), para que recuperara todos los servicios que presentaran algún nivel de similitud sin importar el DoM.

La aplicación de los esquemas de evaluación planteados, se realizó sobre un conjunto de servicios del dominio de la educación pertenecientes a la colección

TC2[22]. Dicho conjunto contenía 135 servicios de comparación, 6 de consulta (Q15-Q20), y el conjunto de servicios relevantes para cada una de ellas. Sobre este subconjunto, realizaron comparaciones manuales de similitud mediante una escala de valores y se calculó su desviación con respecto a las evaluaciones booleanas existentes para el TC2.

Al finalizar el experimento, se comparó la eficacia de la recuperación para el modelo de evaluación booleano (EVS1) y el propuesto en este trabajo (EVS2), mediante la aplicación de las medidas de calidad de recuperación *Recall* y *Precision*, y el paralelo entre sus resultados.

Query ID	**EVS 1**		**EVS 2**	
	R_B	P_B	R_G	P_G
Q15	77%	77%	77%	77%
Q16	60%	92%	87%	96%
Q17	57%	92%	77%	89%
Q18	73%	92%	90%	88%
Q19	100%	65%	100%	71%
Q20	80%	71%	95%	72%

Tabla 2 Precisión y Exhaustividad para EVS1 y EVS2

En este estudio comprobaron la existencia de una gran sensibilidad de las medidas generalizadas para calcular la eficacia de la recuperación de la herramienta suponiendo cambios en los resultados del DoM ó de los resultados proporcionados por los expertos. Por ejemplo, para Q16 obtuvieron que la herramienta no recuperó el servicio S_1 que había sido descrito por los expertos como *"algo relevante"*; de igual manera el S_2 catalogado como *"irrelevante"* la herramienta lo recuperó con un DoM *"subsumes"*. Si se calcula la *precisión* asumiendo que a S_2 se le asignó el DoM *"exacto"* pudo observarse que mientras P_B permaneció constante en 92%, el P_G decreció de un 96% a 93%. Igualmente si S_1 hubiera sido *"muy relevante"*, R_G decrecería de 87% a 84%.

Así mismo, hallaron diferencias significativas entre los resultados de R_G y R_B obtenidos en los experimentos Q16, Q17, Q18 y Q20 (entre un 15% y un 27%). Esta diferencia la atribuyeron a que para estas consultas los expertos caracterizaron algunos servicios como "relevantes" en la evaluación booleana,

[22] http://projects.semwebcentral.org/projects/owls-tc/

mientras que en el otro esquema de evaluación estos mismos servicios pudieron ser catalogados como *"poco relevantes"*.

Como conclusiones plantearon que el esquema de evaluación propuesto mejora en gran medida los inconvenientes presentados en el esquema booleano, sin embargo, presenta otra clase de problemas. La mayoría de los expertos no están dispuestos a especificar la similitud entre dos servicios mediante la escala propuesta por considerarla una tarea dispendiosa; en vez de eso prefieren simplemente valorar la relevancia de forma booleana (es decir, "si ó no"). De aquí, se plantean la incógnita de si es posible inferir un valor de similitud a partir de una evaluación booleana. Es decir, si existe la forma de medir el grado de relevancia entre dos servicios tomando un valor binario proporcionado por un experto. Debido a que la "relevancia" es un concepto subjetivo, es difícil si no imposible saber la interpretación otorgada por cada persona a este parámetro.

No obstante, estos autores proponen que si se logran identificar los componentes de la relevancia, se podría inferir alguna información acerca de ella a partir de un valor booleano. Algunos componentes de este parámetro son la interpretación lógica del dominio del discurso, reglas basadas en la experiencia del usuario, una métrica de similitud heurística, entre otros componentes que afectan la relevancia entre dos conceptos.

Asumieron que una representación apropiada de los servicios podría ser una ontología como la propuesta en [LIH04] donde sus conceptos son expresiones complejas de la Lógica Descriptiva (DL). Si se clasifican los servicios a comparar dentro de la ontología, se pueden obtener la relación que tienen dichos servicios dentro de la estructura.

De lo anterior, asumieron una matriz de inferencias para mapear la relevancia difusa a partir de valores booleanos proporcionados por los expertos, teniendo en cuenta una clasificación de las relaciones de los servicios dentro de una ontología.

Estas relaciones las describieron de la siguiente forma:

- *Eq*: El servicio de comparación S_i es equivalente al de consulta R.
- *DSup*: S_i es un súper concepto directo de *R.*
- *DSub*: S_i es un subconcepto directo de *R.*
- *Sib*: S_i y *R* son *hermanos* dentro de la estructura.
- *No*: no existe una relación directa entre S_i y *R*.

Relación Lógica	Eq	DSup	DSub	Sib	No
Valor Booleano	1	1	1	1	1
Valor difuso Inferido	V	R	R	R	SW
Relación Lógica	**Eq**	**DSup**	**DSub**	**Sib**	**No**
Valor Booleano	0	0	0	0	0
Valor difuso Inferido	SW	S	S	I	I

Valores difusos de Relevancia

- *V*: Muy Relevante
- *R*: Relevante
- *SW*: Algo relevante
- *S*: Poco relevante
- *I*: Irrelevante

Tabla 3 Matriz de inferencias

Para evaluar la efectividad de esta propuesta, la aplicaron sobre la misma colección de prueba sobre la que habían hecho las evaluaciones anteriores, y compararon los resultados. De allí, concluyeron que es posible obtener valores de relevancia difusa a partir de valores booleanos de este parámetro proporcionados por expertos sin que las medidas de calidad de recuperación varíen sustancialmente; y que aplicando algoritmos de inferencia más sofisticados podrían calcular la relevancia difusa de forma más realista.

Otro aspecto a tener en cuenta, es el problema que se presenta cuando los valores de DoM proporcionados por una herramienta, no concuerdan en número con los valores para la relevancia difusa, es decir, cuando $|H(\text{“relevance”})| \neq |H(\text{“DoM”})|$. Para esto proponen la utilización de modificadores difusos, por ejemplo *diluciones* o *concentradores*, con el fin de alinear las escalas, omitiendo pruebas al respecto.

El gran aporte del estudio expuesto anteriormente al trabajo que se presenta en este documento, es la adopción de una escala múltiple para evaluar la relevancia, ya que se pudo comprobar que tiene un mejor desempeño que un método de evaluación booleano.

La propuesta descrita por Tsetsos es quizá, el punto de partida para todos los trabajos que pretenden practicar una evaluación experimental tendiente a establecer juicios de relevancia en aras de evaluar una herramienta específica, y para nuevos planteamientos que busquen ofrecer nuevos métodos de evaluación en este ámbito. Sin embargo, las comparaciones entre servicios tienen varios tópicos a considerar, por ejemplo, las herramientas de emparejamiento de servicios evalúan la concordancia entre las interfaces de los servicios, la secuencia o el flujo de tareas del servicio, entre otros aspectos para al final, entregar un valor de similitud entre los dos servicios. Observando, tal vez si se les da la oportunidad a los expertos de evaluar diferentes características, es probable que puedan emitir juicios de relevancia más acertados; no obstante ese trabajo no se ocupa de evaluar esta posibilidad.

Otro aspecto, es que en [TAH06] no se especifica la forma de exhibir a los expertos los servicios a comparar. Desde la perspectiva de este proyecto, se plantea su importancia como que un buen entendimiento de los servicios puede contribuir eficazmente en el momento de calificar una similitud.

Resumiendo, el trabajo que se desarrolla en este proyecto tiene en cuenta la gradación de la relevancia para hacer los juicios de similitud, pero busca la forma de mejorar este tipo de evaluación adicionando unos criterios que le permitan a quien realiza las comparaciones tener en cuenta diversas características de los servicios web, tal como lo hacen las herramientas automáticas. Igualmente, se propone proveer la mayor cantidad de información relacionada con cada servicio y descripciones gráficas, que garanticen la abstracción correcta de ellos por parte de los evaluadores.

Un segundo trabajo relacionado con la evaluación de herramientas para el descubrimiento de SWS es el planteado por Küster [KUK09] quien propone una metodología y un montaje experimental para la evaluación de herramientas de descubrimiento de SWS partiendo de las diferencias entre estas y los sistemas de IR tradicionales. Para esto, Küster presentó una colección de servicios en un entorno real e investigó la consistencia de tres escalas diferentes para catalogar la relevancia y sus consecuencias sobre la metodología de evaluación propuesta.

El acercamiento más obvio a la evaluación de la recuperación de Servicios Web Semánticos (SWS) es adoptar el enfoque utilizado para IR, es decir, el paradigma de Cranfield representado por TREC. De acuerdo con este paradigma la efectividad de la recuperación se mide principalmente por medio de la *precisión* y la *exhaustividad*, es decir las proporciones de elementos recuperados y elementos relevantes. La relevancia se basa en la similitud temática como la obtenida por los juicios de expertos en el dominio. Una colección de prueba, por lo tanto, tiene tres componentes: un conjunto de documentos (Datos de prueba), un conjunto de

necesidades de información (temáticas o consultas), y un conjunto de juicios de relevancia (listas de documentos que deberían ser recuperados para cada consulta).

Los sistemas de IR tradicionales crean los modelos sobre los cuales trabajan de manera autónoma, por lo tanto desde el punto de vista de la evaluación, ellos operan sobre los datos originales, facilitando que sobre una misma colección de datos se puedan evaluar diferentes sistemas de recuperación. Para la recuperación de SWS, el modelo se forma a partir de anotaciones semánticas que no son creadas automáticamente por el sistema de recuperación, sino escritas manualmente por expertos para obtener la recuperación precisa y eficiente de los servicios.

En términos de la evaluación, el uso de anotaciones semánticas escritas manualmente tiene tres implicaciones importantes. La primera, valorar la acertada recuperación de los servicios teniendo en cuenta la semántica formal y midiendo los efectos de la expresividad del formalismo utilizado, la calidad de las anotaciones, las capacidades del algoritmo que opera sobre las anotaciones; y la alineación de las anotaciones y el algoritmo que las procesa.

Segunda, los datos de entrada a los sistemas de recuperación de SWS, es decir, descripciones de servicios semánticamente anotados, no son de fácil acceso. Por ejemplo, Seekda[23] ofrece cerca de 27.000 descripciones de servicios WSDL, sin embargo en la web sólo unos cientos están semánticamente anotados [KLZ08], y se encuentran descritos en formalismos diferentes y siguen mecanismos de modelado distintos. Por la naturaleza manual del enriquecimiento semántico de los servicios, la generación de colecciones de prueba para los diferentes formalismos tardará un poco hasta alcanzar las magnitudes de las colecciones existentes para IR.

La tercera se deriva de la precedente: las suposiciones del paradigma de Cranfield se validan mediante la comparación de los resultados de la aplicación de diferentes sistemas de recuperación sobre una misma colección de datos. Pero, las colecciones de servicios descritos en diferentes formalismos (WSMO, OWL-S, SAWSDL) son difíciles de construir y su aplicabilidad para herramientas de emparejamiento con formalismos diferentes son difíciles de construir y aún no hay alguna disponible. Además, se sabe que los juicios humanos que proporcionan juicios de relevancia de referencia no concuerdan, comprometiendo la confiabilidad de las evaluaciones, a menos que las colecciones de prueba sean del orden de los miles de elementos y los resultados sean promediados sobre varias docenas de consultas. Como, a futuro cercano no se podrán tener colecciones de servicios de esta naturaleza, no hay un modo de aplicar de manera segura este tipo de evaluación.

[23] http://seekda.com/

Dentro de la propuesta de evaluación que se presenta, está la creación de una colección de servicios descritos en lenguaje natural en vez de una descripción semántica formal, facilitando así los juicios humanos de relevancia y aminorando los efectos adversos provocados por los desacuerdos en los resultados de estos. Una ventaja adicional, es que los juicios de relevancia no tienen que venir necesariamente de expertos en el dominio con conocimiento en web semántica. No obstante, el enriquecimiento semántico de los servicios debe mantenerse en la colección, por ser el sostén de trabajo de los algoritmos.

Debido a que las colecciones de servicios disponibles como OWLS-TC[24], SAWSDL-TC[25] ó SWS-TC[26], no proveen información acerca de los sitios web que proveen el servicio, y los archivos WSDL originales no contienen información acerca de la descripción lógica del servicio original, se reunieron 201 operaciones de servicios reales de fuentes como Seekda, Xmethods[27], Webservicelist[28], Programmableweb[29] y Geonames[30]. La reducción de los servicios a ese número tuvo el fin de facilitar su evaluación manual.

Los servicios fueron inspeccionados y almacenados en una base de datos[31], guardando el nombre y la fuente de cada servicio, los enlaces a su implementación y, si se disponía, su descripción WSDL. Además se almacenó en lenguaje natural, documentación y una lista de las entradas y salidas que fueron tomados de los sitios web que proveyeron el servicio. Adicionalmente, se vincularon de forma manual los tipos de dato de las entradas y salidas al WordNet synsets[32] para evitar ambigüedades semánticas sin recurrir a la lógica de un formalismo específico.

El conjunto de servicios resultante contenía 119 basados en WSDL y los restantes en REST (Representational State Transfer)[33], en su gran mayoría comerciales tomados del dominio de la geografía y la geocodificación. Este dominio permite obtener el número deseado de servicios sin necesidad de crearlos o hacerles modificaciones, ya que existen un gran número de ellos similares disponibles al público.

Para la elaboración de los juicios de relevancia se realizó una comparación entre tres formas diferentes de catalogarla. La primera es la binaria que ha sido comúnmente utilizada, pero presenta diversas críticas por cuanto no concuerda con las escalas de relevancia proporcionadas por las herramientas. La segunda

[24] http://www.semwebcentral.org/projects/owls-tc/
[25] http://www.semwebcentral.org/projects/sawsdl-tc/
[26] http://projects.semwebcentral.org/projects/sws-tc/
[27] http://xmethods.com/ve2/index.po/
[28] http://webservicelist.com/
[29] http://www.programmableweb.com/
[30] http://www.geonames.org/
[31] http://fusion.cs.uni-jena.de/professur/jgd/
[32] http://wordnet.princeton.edu/
[33] http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm

presenta una escala de múltiples valores para medir la similitud entre dos servicios como en [TAH06, KKO08], y la llamaron relevancia unidimensional graduada. Finalmente, la tercera, proporciona una escala múltiple para evaluar diferentes aspectos a comparar entre dos servicios, a este tipo de evaluación de relevancia la llamaron Relevancia multidimensional graduada.

- *Relevancia multidimensional graduada:* Considera tres aspectos a evaluar:
  - La *equivalencia* que determina la semejanza funcional entre el servicio ofrecido y el de consulta. Responde a la pregunta: ¿El servicio ofrecido provee cualitativamente de forma exacta la funcionalidad deseada o sólo algo similar? Y los valores posibles son: *Igual (Equal)*, *Posiblemente Igual (PossEqual)*, *Aproximado (Approximate)*, *Posiblemente aproximado (PossApproximate)* y *No relacionado (Not related)*.
  - El *alcance (scope)* representa la integridad funcional de la oferta con respecto a la petición. Responde a la pregunta: ¿La oferta provee cuantitativamente toda la funcionalidad requerida o sólo en parte? Y los valores posibles son: *Coincide (Match)*, *Posiblemente coincide (PossMatch)*, *Parcial (Partial)*, *Posiblemente parcial (PossPartial)* y *No coincide (NoMatch)*.
  - La *interfaz* determina si la interfaz de la oferta coincide con la requerida. Responde a la pregunta: ¿Las entradas ofrecidas se encuentran todas en el formato esperado y la oferta provee todas las salidas requeridas en los formatos esperados? Y los valores posibles son: *Compatible (Compatible)*, *Posiblemente compatibles (PossCompatible)* e *Incompatible (Incompatible)*.

  Los niveles de relevancia catalogados como *posiblemente (Poss)*, se relacionan con la falta de información para dar un juicio exacto, esto debido a que tanto las descripciones en lenguaje natural como los formalismos pueden proporcionar datos incompletos y dar lugar a interpretaciones o a suposiciones.

- *Relevancia unidimensional graduada*

  Se presentan los siguientes niveles de relevancia:
  - *Concuerda (Match)* cuando el servicio ofrecido coincide exactamente con el requerido.
  - *Posiblemente concuerda (PossMatch)* cando la oferta podría coincidir perfectamente con la consulta, pero la documentación es insuficiente para determinarlo con certeza.
  - *Parcialmente concuerda (ParMatch)* cuando la oferta provee partes de la funcionalidad requerida.
  - *Posiblemente concuerda parcialmente (PossParMatch)* cuando la oferta podría proveer partes de la funcionalidad requerida.

- *RelationMatch* cuando la oferta proporciona una funcionalidad que es cualitativamente similar a la requerida ó la provee pero las interfaces no concuerdan.
- *No concuerdan (NoMatch)* cuando ninguno de los valores anteriores es posible y la oferta es irrelevante para la consulta.

- *Relevancia Binaria*

  Para este tipo de relevancia la escala la definieron como *Coincide (Relation-Match)* y *No Coincide (NoMatch)*. De la escala de relevancia unidimensional graduada como *Irrelevante* y los otros cuatro niveles como *Relevante*.

  Para el experimento formularon tres servicios de consulta, el primero pedía un servicio que convirtiera una dirección en los Estados Unidos a su ubicación geográfica (Geocodificación US); el segundo, un servicio que proveyera la distancia entre dos ciudades alrededor del mundo (Distancia); y el tercero, requería un servicio que proporcionara toda la información disponible de una ciudad en los Estados Unidos (Información de ciudad de US).

  Los juicios de relevancia fueron dados por cuatro jueces entre los que se encontraban los autores de este estudio y dos estudiantes de ciencias de la computación con gran experiencia en programación. Las evaluaciones se realizaron a través de un portal web HTML que permitía escoger el servicio de consulta y cada una de las ofertas, mostrando para cada par de servicios toda la información disponible y permitiendo ingresar los valores de relevancia considerados.

  Los 201 servicios fueron evaluados por los 4 jueces con respecto a las 3 consultas planteadas, mediante las diferentes escalas de relevancia. De allí obtuvieron 12060 evaluaciones, de las cuales 2412 fueron binarias al igual que las de la escala unidimensional, y 7236 del método multidimensional.

  En los resultados observaron que existe una gran diferencia en la forma en que los jueces evalúan los servicios. Esto se hizo evidente al sumar el número de juicios proporcionados por cada evaluador para cada grado de relevancia unidimensional teniendo en cuenta las tres consultas. Por ejemplo, mientras el primer juez encontró 24 servicios que coincidían perfectamente el segundo sólo encontró 3, el tercero encontró 7 y el cuarto 19. Esto mismo sucede para los otros grados de relevancia.

  Los resultados de los juicios de relevancia multidimensional se convirtieron a la escala unidimensional (o binaria), esperando encontrar una consistencia entre los resultados de la escala multidimensional reducida y la unidimensional. Sin embargo, al observar los resultados para cada juez encontraron que dependiendo

del servicio de consulta se presentaban diferentes niveles de inconsistencia. Por ejemplo, para la consulta de la distancia el cuarto juez tuvo una consistencia del 96%, sin embargo al evaluar el servicio de los datos de una ciudad de EU su consistencia fue tan sólo del 67%. Algo similar sucedió para los otros jueces, exceptuando el primero que tuvo una consistencia entre el 90% y el 96% para los diferentes servicios de consulta.

Por otra parte, evaluaron el desacuerdo en los juicios para las diferentes escalas de relevancia entre 2, 3 y los cuatro jueces, teniendo en cuenta la escala binaria, la reducción de la escala multidimensional (MD) a binaria, la reducción de la unidimensional a binaria, la escala unidimensional, la reducción de MD a unidimensional, una escala MD relajada y finalmente una MD estricta. La escala MD estricta analiza si los jueces concuerdan en los juicios para los tres aspectos evaluados, es decir, que si el juicio para alguno de los aspectos no coincide, el desacuerdo entre los jueces será del 100%. La escala MD relajada por su parte, considera cada uno de los aspectos de manera independiente, esto es, que si los jueces coinciden en su apreciación para dos de ellos, el grado de desacuerdo será sólo del 33%.

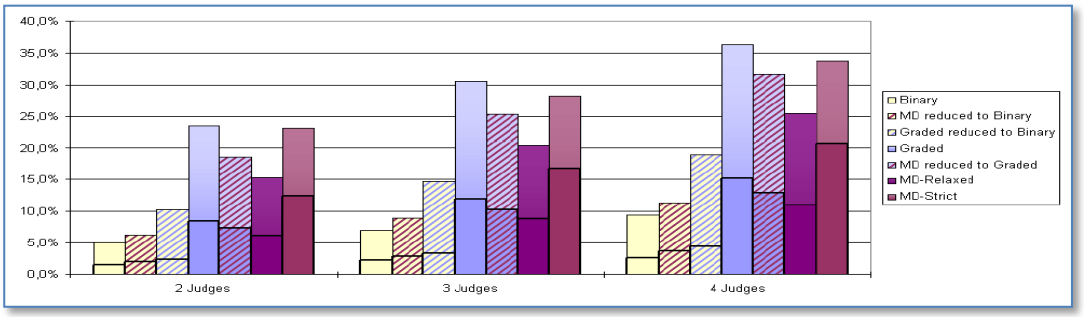

Figura 1 Desacuerdo entre los juicios de relevancia antes (barras completas) y después de la resolución de problemas (parte baja de las barras)

Como es obvio, el grado de desacuerdo aumente conforme se tienen en cuenta más jueces, es decir, que los porcentajes de este parámetro son menores cuando se tienen en cuenta los resultados de dos jueces que cuando se consideran los 4. Sin embargo, al comprar el grado de desacuerdo para la escala unidimensional con respecto a la multidimensional, observaron que esta última reduce el número de errores. Este resultado lo atribuyen a que el uso de una relevancia multidimensional obliga a los jueces a construir sus decisiones en una forma más cuidadosa y estructurada, dando lugar a menores errores pero a una inconsistencia ligeramente superior. Esta interpretación la soportaron además en el hecho de que los juicios unidimensionales obtenidos mediante la reducción de MD, mostraron un nivel de desacuerdo menor al obtenido de la evaluación unidimensional. De forma similar, observaron que el grado de desacuerdo

obtenido de la reducción de MD a escala binaria, es menor que el mostrado por la reducción de la escala unidimensional a binaria. No obstante, ambas reducciones presentaron mayor valor de desacuerdo que la escala binaria. La explicación dada a este fenómeno es que al realizar la evaluación binaria después de haber terminado las otras dos evaluaciones, el conocimiento ganado mediante estos procesos pudo haber conducido a juicios más consistentes.

Debido a que los resultados fueron un tanto decepcionantes, adicionaron al experimento una fase de resolución de conflictos donde los jueces discutieron sus evaluaciones y las compararon con las de los demás. De esta manera, pudieron cambiar o modificar sus juicios y el resultado del grado de desacuerdo se redujo sustancialmente como se muestra en la figura 1.

Ahora bien, en lo que atañe al presente trabajo, los aportes importantes son la utilización de la escala multidimensional, que permite hacer una evaluación más completa de los servicios; y la presentación, en lenguaje natural, de información relacionada con los servicios para facilitar las evaluaciones manuales.

La relevancia multidimensional graduada utilizada por Küster [KUK09] posibilita evaluaciones de diferentes aspectos del servicio, mejora la calidad de las apreciaciones de un evaluador acerca del grado de relevancia dado a un servicio frente a una consulta cuando estos no son iguales ni completamente diferentes. La consideración de caracteríasticas como la funcionalidad, el flujo de tareas, las inerfaces, obliga a los jueces a construir sus decisiones en una forma más cuidadosa y estructurada. Esto pudo comprobarse con los diferentes resultados obtenidos de este experimento mediante la observación de los cálculos de desacuerdos, donde se vio que las escalas reducidas de MD (unidimensional y binaria) presentaron un mejor comportamiento que la relevancia unidimensional y la reducción de esta (binaria).

Por otra parte, la inclusión de información de los servicios en lenguaje natural, mejora el entendimiento de estos y por lo tanto optimiza la calidad de los juicios de relevancia.

Sin embargo, como metodología de evaluación, el trabajo de Küster carece de medidas para analizar la efectividad de una herramienta automática de recuperación de servicios de acuerdo con las escalas graduadas de relevancia, por lo tanto para el desarrollo del proyecto descrito en esta monografía, se decidió adoptar el enfoque de Tsetsos para este aspecto.

RESUMEN: Se puede concluir que las metodologías para la evaluación de sistemas para el descubrimiento de servicios, están aún incipientes y muy ceñidas a los mecanimos propuestos para la evaluación de los SRI, por lo que aún no se establece con certeza la mejor manera de definir la relevancia de un servicio con respecto a una consulta. Además, hasta hace muy poco, todos los esfuerzos estaban centrados en el desarrollo de herramientas automáticas de descubrimiento, pero no muchos estudios en generar formas de evaluarlas. Adicionalmente, se observó que la mayoría de los autores referenciados, culpan de la falta de modelos de evalaución a la inexistencia de colecciones de servicios de prueba que faciliten la aplicación de las herramientas sobre entornos reales, y que permitan la comparación del desempeño de diferentes sistemas.

A pesar de lo mencionado anteriormente, autores como Tsetsos [TAH06] y Küster [KUK09], proponen metodologías de evaluación de la eficiencia de la recuperación, basados en juicios de relevancia realizados por expertos en el dominio. Los trabajos referenciados en este capítulo, realizaron tres contribuciones concretas a este proyecto:

- La utilización de una escala graduada para establecer la relevancia.
- Hacer una evaluación de la relevancia multidimensional, en la que se tengan en cuenta características de los servicios para realizar las comparaciones.
- Adicionar descripciones en lenguaje natural a los servicios del repositorio para proporcionar información de fácil entendimiento a los usuarios, y así facilitar en parte la comprensión de los servicios a comparar.

CAPÍTULO 3

SELECCIÓN DE CRITERIOS BASADOS EN LA EXPERIENCIA PARA LA COMPARACIÓN DE PROCESOS DE MANERA INTUITIVA

La selección de los criterios para la comparación manual de similitud entre procesos, es un paso muy importante, ya que la valoración de ellos dará como resultado la percepción humana de semejanza, y el punto de partida para la evaluación de la calidad de recuperación de las herramientas automáticas para el descubrimiento de servicios web, que es parte del propósito central de este trabajo.

Como se vio en el capítulo anterior, la evaluación de sistemas de recuperación de información, constituye la base para el planteamiento de metodologías y métricas de evaluación para los sistemas de recuperación en general. A continuación, se hará un resumen de los criterios utilizados en la evaluación de los SRI, la influencia de estos criterios en la elaboración de aquellos a utilizar para valorar las herramientas de descubrimiento de servicios web, y finalmente, se numerarán y justificarán los criterios establecidos para esta evaluación.

3.1 CRITERIOS UTILIZADOS PARA LA EVALUACIÓN DE LA RECUPERACIÓN DE INFORMACIÓN

"La propia naturaleza de los SRI propicia su necesidad crítica de evaluación, justo como cualquier campo de trabajo que aspire a ser clasificado como campo científico" [BLA90]. De allí que muchos autores, desde la creación de los SRI, hayan enfocado sus esfuerzos en encontrar los criterios que permitan evaluarlos integralmente.

Baeza-Yates [BYF92] por ejemplo, manifiesta que los sistemas de recuperación de información pueden ser evaluados por diversos criterios como:

- *Eficacia de ejecución*: Es la medida del tiempo que tarda el sistema o a parte de este, para realizar una operación. La importancia de este parámetro radica en que un excesivo tiempo de recuperación, conduzca al desuso del sistema. Los requerimientos no funcionales de un SRI normalmente especifican el tiempo máximo aceptable para una búsqueda y para las operaciones de mantenimiento de una base documental, tales como añadir y borrar documentos [LOP06].

- *Efectivo almacenamiento de los datos*: Es el espacio necesario para el almacenamiento de los datos. Una forma común de medir en bytes la eficacia del almacenamiento, es la relación de la dimensión de los ficheros índice, más la del documento sobre la dimensión de los archivos del documento, denominado espacio general.
- *Efectividad en la recuperación de la información*: Generalmente esta medida se basa en la *relevancia* de los documentos recuperados, es decir en la satisfacción de la necesidad real de información del usuario.

Borlund [BOR00] por su parte, diferencia entre evaluar el *acceso físico* y el *acceso lógico*:

- *Acceso físico* es el que concierne a cómo la información es recuperada y representada de forma física al usuario, y está muy vinculado con las técnicas de recuperación y de presentación de la información.
- *Acceso lógico* a los datos que está relacionado con la localización de la información deseada. Tiene que ver con la *relevancia* del objeto localizado con una determinada petición de información

De forma parecida, Baeza-Yates [BYR99] afirma que existen dos tipos de evaluaciones: la del funcionamiento del sistema y la del funcionamiento de la recuperación, siendo la segunda modalidad la que analiza cómo los documentos recuperados se clasifican de acuerdo a su *relevancia* con la pregunta efectuada.

Debido a la coincidencia de la necesidad de evaluar la relevancia como un criterio primordial en la evaluación de estos sistemas, se hace necesaria la definición de dicho término. *Relevancia* según el Diccionario de la Lengua Española, significa "cualidad o condición de relevante, importancia, significación", por lo tanto un documento será relevante en tanto su contenido posea significación o importancia de acuerdo a la necesidad informativa. Sin embargo, se presentan algunos inconvenientes para determinar si un documento es relevante o no debido a que esta apreciación es subjetiva, la valoración será distinta dependiendo de quién la haga, el momento en que la haga o la necesidad específica que tenga, entre otros obstáculos.

Estos impedimentos obligan a distintos planteamientos; Saracevic [SAR97] por ejemplo, habla de dos tipos de relevancia, la *objetiva* que hace hincapié en los sistemas, y normalmente define cómo la materia de los elementos recuperados coincide con la de la petición; y la *subjetiva* que es la relevancia mirada desde el punto de vista del usuario. Por otra parte, [GRE00] aporta la idea de "utilidad de un documento" o *pertinencia*, considerando que es mejor definir a la *relevancia* en términos de la percepción que un usuario posee sobre la utilidad de un documento

recuperado, es decir, *si el mismo le va a ser útil o no*. Así pues, la *relevancia* queda asociada con el concepto de la relación existente entre los contenidos de un documento con una temática determinada y *pertinencia* se restringe a la *relación de utilidad* existente entre un documento recuperado y una necesidad de información individual.

Por todo lo anterior, es evidente que un campo importante a tener en cuenta en la evaluación de los SRI es la satisfacción del usuario, y por ende la medición de criterios que muestren el desempeño de las herramientas en este ámbito. Los criterios definidos a continuación pretenden describir las características principales que un SRI debe ofrecer al usuario [LOP06]:

- *Exhaustividad* ó habilidad del sistema para presentar todos los ítems relevantes. De manera estándar, representa la proporción de material relevante que es recuperado como respuesta a una petición de búsqueda.
- *Precisión* ó la habilidad del sistema para recuperar sólo ítems relevantes. Como medida, representa la proporción de material recuperado que es relevante.
- *Esfuerzo* Intelectual ó físico, requerido por el usuario en la formulación de las consultas, en el manejo de la búsqueda y en el escrutinio de los resultados.
- *Tiempo* transcurrido entre que el sistema recibe la consulta por parte del usuario y entrega los elementos recuperados.
- *La presentación* de los resultados de la búsqueda, que influye en habilidad del usuario para hacer uso de la información recuperada.
- *El alcance o cobertura* de la colección documental, ó la proporción de ítems relevantes conocidos por el usuario dentro del material recuperado.

Rijsbergen [RIJ99] opina que la proporción de material relevante recuperado como respuesta a una petición de búsqueda (*exhaustividad ó recall*), y la proporción de material recuperado que es realmente relevante (*precisión*), son los parámetros que verdaderamente pretenden medir la *efectividad* de los sistemas, siendo esta una medida de la capacidad del sistema para satisfacer al usuario en términos de la *relevancia* de los documentos recuperados. Finalmente, por regla general se consideran de mayor importancia las medidas basadas en la *relevancia* que aquellas basadas en el *proceso*, principalmente porque estas últimas sirven para diferenciar unos sistemas de otros con base en las prestaciones de la aplicación informática subyacente, pero no valoran adecuadamente los aspectos relacionados con el contenido de los documentos [MMR04].

3.2 CRITERIOS UTILIZADOS PARA LA EVALUACIÓN DE LA RECUPERACIÓN DE SERVICIOS WEB

El acercamiento más obvio para la evaluación de la recuperación de SWS es adoptar un enfoque de evaluación más común como el de los SRI, es decir, el paradigma de Cranfield [CLE72], representado por el TREC, donde la eficacia de la recuperación de un SRI es medida principalmente por la *exhaustividad* y la *precisión*. Estas medidas están basadas en la *relevancia* de los documentos recuperados de acuerdo con una similitud temática, como la proporcionada por los juicios de expertos en el dominio. Las colecciones de prueba asociadas a este tipo de sistemas contienen un conjunto de documentos de prueba, un conjunto de necesidad de información (temáticas ó consultas), y una lista de documentos que deberían ser recuperados a partir de una consulta [KUK09].

Al igual que otros sistemas de recuperación, como los motores de búsqueda web, los sistemas para el descubrimiento de Servicios Web Semánticos (SWS) deberían ser evaluados en términos del *rendimiento* y *la eficacia de recuperación*, donde el primero se mide por la complejidad computacional, la escalabilidad, los tiempos de respuesta, etc.; y la eficacia ilustra el acierto del sistema en el descubrimiento de servicios relevantes [TAH06], enfocándose particularmente en la evaluación de las herramientas en términos de la eficacia de la recuperación.

Las evaluaciones de los SRI se han centrado en la aplicación del paradigma de Cranfield, en el cual se desarrollan evaluaciones de los diferentes sistemas mediante la comparación de sus desempeños (rendimiento y eficacia), sobre colecciones extensas de documentos. Sin embargo, la aplicación de este paradigma en la evaluación de los sistemas de recuperación de SWS (Semantic Web Services) no es tan sencilla, y uno de los motivos fundamentales es que, generalmente, los conjuntos de servicios sobre los cuales se pretende hacer las pruebas, no son lo suficientemente grandes para poder asegurar que los resultados obtenidos van a ser aceptables. Las razones para esto son variadas, una de ellas, y tal vez la más importante, es la existencia de los diferentes lenguajes para describir los servicios y las distintas ontologías disponibles para anotarlos semánticamente; esto indudablemente da pie para la creación de herramientas diversas para el descubrimiento de servicios, y la generación de colecciones para evaluar cada una de ellas, repercutiendo en un avance lento en el crecimiento de dichos conjuntos de servicios para hacer pruebas. Además, mientras la evolución de los SRI viene desde su aparición a principios de los 40, el avance de los sistemas de recuperación de servicios web tiene apenas alrededor de una década.

Conforme a lo expuesto anteriormente, si bien las metodologías de evaluación de la recuperación de SWS toman su punto de partida en la evaluación de los SRI, por el momento los esfuerzos más grandes, se están haciendo en la generación de colecciones de servicios de prueba para cada uno de los lenguajes de descripción y

así poder evaluar herramientas afines; y en la creación de conjuntos de servicios heterogéneos (descritos cada uno en los diferentes lenguajes), para poder hacer comparaciones híbridas de herramientas al nivel que se hacen para los SRI.

Pese a lo anterior, estudios recientes se han enfocado en formular metodologías, métricas y montajes experimentales, encaminados a evaluar la eficacia de la recuperación de herramientas para el descubrimiento de SWS [TAH06, MDG06, SAK04, KUK09]. Como se ha referido en precedencia, esta clase de evaluación está basada en la relevancia de los elementos recuperados de acuerdo con una consulta dada, y dos puntos importantes a definir para medir este parámetro, son los valores que puede tomar, y los aspectos a evaluar.

Por ejemplo en [KUK09] proponen una relevancia multidimensional graduada, esto es, que cada dimensión corresponde a un aspecto de la relevancia a evaluar, y cada dimensión se mide de acuerdo a una escala. Los aspectos a evaluar fueron:

- *La equivalencia* que mide cualitativamente las semejanza funcional entre un servicio consultado y uno ofrecido.
- *El alcance* que representa cuantitativamente si el servicio ofrecido satisface toda la funcionalidad requerida por la consulta.
- *La interfaz* determina si las entradas y salidas del servicio ofrecido coinciden con las requeridas en la consulta, y si todas ellas se encuentran en el formato deseado.

Teniendo en cuenta lo anterior podría intuirse que una de las formas para determinar los aspectos que evalúan la relevancia, podría ser extrayendo los métodos utilizados por las herramientas automáticas para el descubrimiento de servicios. De acuerdo con [FIC10], estos métodos están centrados en cuatro niveles:

- *Las interfaces* que para comparar similitud entre servicios, toma en cuenta las entradas y las salidas de estos.
- *La estructura* que se centra en la organización de las tareas dentro de los servicios.
- *El comportamiento* enfocado en el flujo de ejecución.
- *La semántica* que realiza un análisis de los conceptos propios de las tareas de acuerdo con un dominio específico de aplicación.

Habiendo establecido que la eficacia de recuperación de un sistema, se mide de acuerdo a la relevancia de un elemento recuperado con respecto a la consulta, y que la relevancia es aportada por unos expertos en el domino, es decir, que son las personas las que estipulan si un servicio recuperado es relevante o no, habría que analizar en qué se basan estos juicios. Algunos estudios preliminares al desarrollo

de herramientas de emparejamiento, realizaron estudios para conocer de qué manera las personas hacían juicios de similitud, con el fin de aplicar estos parámetros en sus algoritmos [BKB05, GLG06, WOM06, WOR06]. A continuación se describen algunos de estos trabajos, y su aporte en la selección de los criterios para la evaluación intuitiva propuesta en este proyecto.

Bernstein [BKB05], recopilan un catálogo de medidas de similitud basadas en ontologías, que son empíricamente comparadas con un "patrón" obtenido al encuestar a 50 personas. Los métodos que usaron para medir la similitud semántica entre objetos de una ontología fueron la medición de la distancia ontológica, espacio vectorial, edición de distancia o distancia de Levenshtein, entre otras. De acuerdo con el estudio de Budanitsky and Hirst [BUH01], decidieron que los juicios humanos de similitud proporcionan la mejor evaluación de las bondades de una medida, y realizaron el cotejo de los resultados obtenidos de la aplicación de las medidas planteadas, frente a los resultados de la evaluación experimental.

Para la ejecución de su estudio, escogieron como ontología subyacente el MIT Process Handbook Ontology [MCL99, MCH03], y de esta seleccionaron 40 procesos que fueran entendibles para una audiencia general. Adicionalmente, los procesos fueron descritos en lenguaje natural, y listadas las partes de cada uno. Los procesos fueron presentados a 50 participantes escogidos de tres rangos universitarios involucrados en las ciencias de la computación y lingüística computacional, para que compararan un par de procesos, los calificaran de acuerdo a su similitud en una escala de 1 (diferentes) a 5 (idénticos) y, explicaran el método utilizado optando de entre una lista:

- 1. Nombre de proceso
- 2. Descripción del proceso
- 3. Partes/relaciones del proceso
- 4. Una combinación de 1 y 3
- 5. Usando otro método de evaluación

*L*os resultados obtenidos tanto de las encuestas como de la aplicación de las medidas fueron comparados usando la correlación de Spearman [BKB05], de allí pudieron concluir que las evaluaciones humanas y de los algoritmos son notoriamente variables; que los sujetos y los algoritmos pueden ser agrupados de modo que, la aplicación de métodos para evaluar similitud debe ser personalizada para optimizar los resultados.

Teniendo en cuenta los métodos utilizados por las personas para evaluar la similitud entre los objetos presentados, formaron dos grupos de medidas. En el

primero pusieron las medidas enfocadas en las partes del objeto (el modelo vectorial y los dos de edición de distancia), es decir sus atributos y relaciones. Y en el segundo, las medidas basadas en la ubicación del proceso dentro de la ontología (information-theoretic y la distancia ontológica).

Por otra parte, Carole Goble durante sus estudios en la Escuela de Ciencias de la Computación de la Universidad de Manchester, presenta un estudio empírico [GLG06] de los criterios utilizados por un personal del área de la bioinformática para descubrir workflows, y de los parámetros que debe evaluar una herramienta automática de descubrimiento para ser eficaz. Para esto realizaron dos experimentos.

En el primero, evaluaron los criterios considerados importantes por los científicos en la búsqueda de workflows para ser reutilizados. En él, participaron 21 personas, 15 del área de la bioinformática, y 6 desarrolladores de software quienes fueron interrogados sobre la importancia de aspectos como las entradas, las salidas producidas, la descripción de las tareas, el proveedor del servicio, entre otros, a la hora de hacer descubrimiento de workflows. Para ello, se les pidió que valoraran la importancia de dichos criterios por medio de una escala de 1 a 5 donde 1 denotaba el menor valor de importancia y 5 el mayor.

Como resultado obtuvieron una jerarquía de los criterios: la descripción de las tareas está en el lugar de mayor importancia seguido por, las entradas, las salidas producidas, la documentación en línea, el proveedor del servicio, y por último los recursos subyacentes (Ej. Una BD particular). Al preguntarles a los participantes qué otros criterios tienen en cuenta, propusieron *parámetros de calidad del servicio,* en particular medidas de *rendimiento y fiabilidad.* Otras personas expresaron que ellos no sólo se basan en las características de funcionamiento sino que esperan usar información estructural como los servicios contenidos en el workflow y las subtareas dirigidas o iniciadas por este. Ello sugiere un descubrimiento basado en la estructura de los workflows usando información de su comportamiento.

En el segundo experimento, la encuesta evaluó la similitud entre workflows mediante apreciaciones humanas que aplicaron los criterios arrojados como relevantes en el experimento anterior. En este nuevo proceso participaron 9 personas del área de la bioinformática y 4 desarrolladores de software. La encuesta se difundió a través del servicio de evaluación de Keysurvey[34] y el repositorio de workflows fue publicado y accedido en línea gracias a *my*Grid/Taverna

[34] www.keysurvey.com

workbench[35]. El repositorio contenía 89 Workflows de los cuales 66 fueron creados por el mismo autor como soporte para la investigación de enfermedades graves. El desarrollo de la actividad se dividió en tres secciones principales. En la primera, se les pidió a los usuarios alguna información personal y se les presentó un modelo de un workflow de referencia del que debían describir su funcionalidad con el fin de garantizar un correcto entendimiento.

En la segunda parte, cinco workflows diferentes en tamaño, organización de los nodos, etc., fueron seleccionados del repositorio. Cada uno de estos modelos fue enfrentado al workflow de referencia para que los participantes juzgaran su similitud dando valores entre 1 (Idénticos) y 9 (completamente diferentes), teniendo en cuenta la funcionalidad y la forma general de los modelos. Además, se les presentaron los criterios de comparación para ser calificados de acuerdo a la utilidad que tienen durante la comparación siendo 1 muy útil y 4 nada útil. Así mismo los participantes dieron una medida de confiabilidad de su evaluación entre alta-media-baja, siendo 1 alta y 5 baja.

Los criterios considerados fueron los siguientes:

- Tiene sentido en el contexto de la biología, tener este workflow como parte del workflow de ejemplo
- Tiene sentido en el contexto biológico tener este workflow como una superposición del workflow de ejemplo
- Forma del workflow: número de entradas y salidas compartidas
- Forma del workflow: correspondencia del tipo de servicio
- Forma del workflow: composiciones de servicios compartidos
- Forma del workflow: flujos compartidos entre la entrada y la salida

En la tercera y última parte, los participantes valoraron la importancia de adicionar información como descripciones textuales de funcionamiento y clasificación de los servicios, y finalmente, se preguntó el nivel de dificultad de la actividad completa.

Al analizar los resultados de las encuestas, encontraron que los encuestados reportaron una confianza “media” sobre todos sus juicios. El 66,7% de los participantes, hallaron la estimación de la funcionalidad entre “muy difícil” y “difícil”; mientras que la dificultad del análisis de similitud de la forma de los worflows, fue sólo de un 25%, de lo que se concluye que la representación gráfica de flujos es un aspecto de fácil comprensión por parte de los evaluadores, lo que convierte a este criterio en fiable a la hora de emitir juicios. Por otra parte, al analizar la importancia de la aplicación de los criterios, los resultados arrojados mostraron que no hay un consenso ya que estos varían de acuerdo con las

[35] www.mygrid.org.uk

personas y con los workflows que estén comparando. No obstante, los criterios que tienen en cuenta si un workflow hace parte del otro ó es una superposición, los encuestados los catalogaron como “importantes” ó “muy importantes”, cuando los esquemas evaluados presentaban esta característica.

Cuando le preguntaron a los evaluadores por la importancia de añadir descripciones textuales de las entradas, las salidas y los servicios, un 69,2% estuvo de acuerdo es que es “muy importante” y el 30,8% restante consideró que era “importante”. Por otra parte, a la pregunta de si es importante añadir información acerca de la clasificación del servicio, en términos de su rol en el proceso de la bio-informática, el 46,2% lo consideró “muy importante”, el 30,8% “importante”, y el 23% restante como “neutral”. Acerca de agregar información de clasificación de entradas y salidas, el 23,1% lo encontró “muy importante”, el 53,8% “importante” y el 23,1% “neutral”. Finalmente, a la pregunta de si es importante agregar información acerca de la clasificación de combinaciones de servicios dentro de múltiples workflows, el 38,5% lo calificó como “muy importante”, el 15,4% como “importante” y el 46,2% como “neutral”.

A la complejidad encontrada por los evaluadores al comparar diagramas de workflows, el 16,7% lo encontró “muy difícil”, el 25% “difícil” y el 58,3 “Neutral”. Finalmente, concluyeron que no se pueden determinar unos criterios estándar que reflejen los aplicados por los humanos para evaluar similitud, ya que cada individuo utiliza formas distintas para establecer diferencias; por lo tanto es difícil crear una herramienta automática de descubrimiento de workflows que recree exactamente la habilidad de comparación.

Wombacher presenta otro estudio empírico cuyo objetivo principal era obtener una noción de los criterios aplicados por los humanos para medir la similitud en las diferentes descripciones de workflows [WOR06, WOM06].

Para este experimento se desarrolló un cuestionario de 23 ejercicios, cada uno presentaba un Workflow de referencia y 3 o 4 más para ser comparados con el primero mediante la organización descendente de acuerdo a su similitud y, la descripción de la razón por la cual se hizo el ordenamiento. El cuestionario se diseñó de modo que fuera entendible para que no requiriese más de 60 minutos para ser desarrollado, es por esto que los workflows fueron representados como máquinas de estado finito de procesos muy sencillos de entender y evaluar.

Para descubrir los criterios que los humanos utilizan para medir similitud entre procesos, fueron planteadas unas hipótesis con respecto a la importancia de tres aspectos fundamentales en la descripción de un workflow. Los aspectos a evaluar

eran el *lenguaje* o las posibles secuencias de ejecución de un autómata; la *semántica* referida a las etiquetas de transición que determinan la semántica del Workflow, y la *estructura* o representación gráfica de los procesos. Para probar las hipótesis planteadas con respecto a esos tres ítems importantes, el cuestionario se diseñó para evaluar con los ejercicios la veracidad de cada una. Esto quiere decir que, mediante la evaluación de las respuestas obtenidas del cuestionario, podría probarse la certitud de cada hipótesis.

Al examinar los resultados, tuvieron en cuenta el ordenamiento que los participantes dieron a los workflows en cada ejercicio y las razones que sustentaron su clasificación, de acuerdo a esto, se escogieron las respuestas que apoyaban la hipótesis, las que no y, las que cuyo aporte era nulo para la evaluación.

De esta manera, los autores analizaron los resultados obtenidos por la encuestas y concluyeron que para determinar similitud de workflows son más efectivas las medidas de lenguaje que las medidas de estructura. Incluso si el lenguaje no es exactamente igual, los autómatas con estructura análoga son considerados menos similares. Además, los súper autómatas se consideran más afines que aquellos que presentan transiciones antes o entre los caminos del workflow de referencia. Sin embargo, cuando transiciones extra son adheridas al autómata como agregación final, estos son considerados más parecidos. Por otra parte, un alfabeto similar es importante, pero cuando la estructura y las secuencias cambian mucho se prefieren los autómatas con diferente alfabeto pero secuencias semejantes. Asimismo, el reemplazo de una transición por un bucle, por otra transición ó la eliminación de esta tienen el mismo impacto en las medidas de similitud. Finalmente la influencia de la semántica en dichas medidas es baja cuando se presentan súper autómatas con lenguaje afín, pero si el lenguaje es distinto se prefieren los autómatas con semántica análoga.

3.3 SELECCIÓN Y DESCRIPCIÓN DE LOS CRITERIOS A EVALUAR INTUITIVAMENTE

El objetivo a cumplir con esta selección es establecer los criterios que servirán a los evaluadores para hacer las comparaciones de similitud entre procesos. Teniendo en cuenta el numeral anterior, las propuestas recientes para la evaluación de herramientas para el descubrimiento de SWS [TAH06, KUK09], constituyen el punto de partida para establecer estos criterios.

Los criterios escogidos deben permitir la evaluación de la relevancia de un servicio de acuerdo a una consulta dada. Debido a que la mayoría de los sistemas para la recuperación de servicios, mediante la asignación de grados de correspondencia (DoM), entregan sus resultados como un conjunto jerarquizado de elementos

similares para cada consulta [TAH06], se hace necesario que los evaluadores puedan asignar niveles de relevancia a cada comparación. De igual manera, en [KUK09] se corroboró que al evaluar diferentes características de un servicio se obtienen medidas de la relevancia más confiables ya que, esto obliga a las personas a construir sus decisiones en una forma más cuidadosa y estructurada. Por lo anterior, se tomó la decisión de hacer una evaluación de la relevancia mediante un esquema multidimensional graduado, partiendo de la propuesta expuesta en [KUK09].

Para la selección de las dimensiones (características de los servicios), se tuvieron en cuenta los estudios experimentales de Wombacher [WOM06, WOR06], Bernstein [BKB05] y Goble [GLG06], que proveen los diferentes aspectos que los humanos utilizan a la hora de emitir juicios de similitud.

Del experimento de Goble, se destaca la obtención de una lista de criterios jerarquizados encabezada por las descripciones de las tareas, seguidas por sus entradas y salidas, la documentación en línea, el proveedor del servicio y el uso de recursos subyacentes como una BD particular. Además, se evidenció que la estructura gráfica de los flujos es un criterio fácil de evaluar, y que en aquellos esquemas que presentan algún tipo de relación de inclusión, considerarla se convierte en un aspecto importante. De igual manera, los encuestados catalogaron como importante la inclusión de descripciones textuales de los procesos, lo que confirma la necesidad de proveer este tipo de información.

Bernstein por su parte, concluyó que la intuición humana para comparar no puede ser descrita por un criterio particular, cada persona utiliza razonamientos distintos para sus procesos cognitivos y por lo tanto, aplica sus propios juicios a la hora de determinar similitudes. Sin embargo, cabe resaltar los criterios que ofreció para realizar las comparaciones; estos fueron: el nombre del proceso, la descripción del proceso, las partes del proceso y sus relaciones, y una combinación del nombre y las partes del proceso. Se destaca de este estudio, que utilizaron descripciones en lenguaje natural de la funcionalidad de los procesos a comparar, así como un listado de los elementos que los componían.

Las conclusiones del estudio de Wombacher, proporcionan una idea de los razonamientos humanos a la hora de encontrar las semejanzas entre procesos, mostrando la preferencia de uno o varios criterios sobre otros de acuerdo a los casos presentados.

Teniendo en cuenta las evaluaciones experimentales, y los diferentes métodos utilizados comúnmente para hacer descubrimientos automáticos de servicios, se concluye que:

- Los criterios servirán para evaluar la similitud entre las diferentes características de los servicios, y a su vez dar un juicio de relevancia global.
- Los criterios deben ser medidos mediante escalas que permitan asignar diferentes grados de similitud.
- Los criterios deben estar soportados en los resultados de las evaluaciones experimentales para garantizar que puedan ser medidos por las personas, y que reflejen su proceso cognitivo de evaluación de similitud.
- Los criterios deben tener en cuenta al menos los cuatro métodos utilizados comúnmente para el descubrimiento automático de servicios, es decir, las interfaces, la estructura, el comportamiento y la semántica; y así obtener resultados coherentes con las distintas herramientas que podrá evaluar la plataforma.

De acuerdo con lo anterior se establecieron los criterios de la siguiente manera:

- Para la comparación de los procesos completos se tienen en cuenta la estructura y el comportamiento, y para cada uno de ellos, se establecen los criterios que los evalúan:

    - *Estructura:* Se refiere a la representación formal y total de un proceso que permite la visualización de la secuencia de las tareas y las relaciones entre ellas. Los resultados de Wombacher mostraron que en casos donde la estructura y las secuencias cambian mucho se prefieren los autómatas con diferente alfabeto pero secuencias similares, haciendo de la estructura un factor para determinar la diferencia entre procesos. De igual manera, los resultados de las encuestas de Goble, reflejaron que la estructura es un criterio fácil y fiable para evaluar. Para su evaluación se establecieron los siguientes criterios:
        - Dependencia Causal: Evalúa qué tan parecidos son dos procesos teniendo en cuenta las relaciones entre los eventos y las tareas que los componen. En muchos casos estas relaciones pueden cambiar por completo la finalidad total de un proceso, y por lo tanto son un factor importante para establecer similitud. Este criterio se evaluó mediante la asignación de valores numéricos de 0 (Completamente diferentes) a 4 (Idénticos).

        - Estructura gráfica: Provee al evaluador una idea visual de la secuencia de las tareas, que complementada con una descripción verbal del proceso permitirá descartar rápidamente elementos disímiles. Del estudio de Goble se concluyó que el entendimiento de los procesos depende en gran medida de la forma como ellos sean descritos

gráficamente, de ahí la importancia de escoger un modelo estandarizado que sea comprensible.

Para la evaluación de este criterio se tuvo en cuenta la relación estructural entre el proceso de consulta y el de comparación, siendo esta relación *fallida* cuando no existe ni una relación entre tareas que sea común para los dos procesos; *inexacta* cuando el proceso de comparación presenta flujos del proceso que no son exactamente iguales al de consulta, pero que guardan alguna relación; *inclusión* cuando el proceso de consulta está contenido, es parte o hereda del proceso de comparación; *complemento* cuando el proceso de comparación está contenido, es parte o hereda del proceso de consulta; y *exacto* cuando los procesos de consulta y comparación representan el mismo proceso.

- *Comportamiento*: Tiene en cuenta el flujo de ejecución, o de control, de las tareas dentro del proceso de negocio, es decir, los constructores específicos que permiten observar la manera en que un proceso de negocio se comporta [FIC10].
  - Flujo de control: Muestra el orden en que son realizadas las tareas para cumplir con el objetivo del proceso. Como se vio en Wombacher las secuencias pueden determinar la diferencia entre dos procesos en casos donde los nombres de las tareas, no son claros, son diferentes o simplemente insuficientes. Este criterio se evaluó mediante la asignación de valores numéricos de 0 (Completamente diferentes) a 4 (Idénticos).

- Para la comparación entre las tareas que componen los procesos se tuvieron en cuenta la semántica y las interfaces, y se determinaron los criterios para evaluarlas.
  - *Semántica:* En una herramienta para el descubrimiento de procesos, es el análisis de conceptos propios de las tareas de acuerdo a un dominio específico de aplicación. Para el caso de las personas, un concepto propio de una tarea puede ser su nombre que por sí sólo puede dar una idea de lo que hace, pero que al agregar una descripción verbal de su funcionalidad, teniendo en cuenta el estudio de Bernstein, proporciona una idea clara de la participación de la tarea en el proceso.
    - Nombres: Para evaluar este criterio, se comparan los nombres de las actividades de cada uno de los procesos, encontrado para cada tarea del proceso de consulta, si existe una y sólo una tarea del proceso de

comparación que se le asemeje. Al relacionar un par de tareas se asigna un valor numérico de similitud a esta relación, que va desde 1 cuando la similitud es baja, hasta 4 cuando los nombres son exactamente iguales.

- Descripciones: De Goble se concluyó que una clara comprensión de lo que se compara, reduce las respuestas aleatorias producidas por las dudas y por las dificultades a la hora de evaluar. Además, la asignación de nombres a las tareas se hace en general de forma subjetiva, por lo que dos tareas con nombres disímiles, pueden tener una funcionalidad parecida o hasta igual. De acuerdo con las descripciones de las tareas, se busca si para cada tarea del proceso de consulta existe una y sólo una tarea del proceso de comparación cuya funcionalidad sea semejante. Al relacionar un par de tareas se asigna un valor numérico de similitud a esta relación, que va desde 1 cuando la similitud es baja, hasta 4 cuando los nombres son exactamente iguales.

- *Interfaces:* Son las entradas y las salidas de una tarea. De Goble se obtuvo que las interfaces eran un elemento importante de comparación, ya que dos tareas pueden tener una funcionalidad similar, pero pueden ser disparadas por parámetros distintos o incluso arrojar como resultado algo diferente a lo deseado.

  - Entradas: Se listan las entradas de las tareas del proceso de consulta y se comparan con la lista de las mismas para el proceso de comparación. Para valorar este criterio, se busca si para cada tarea del proceso de consulta existe una y sólo una tarea del proceso de comparación, cuya lista de entradas posea elementos semejantes. Al relacionar un par de tareas se asigna un valor numérico de similitud a esta relación, que va desde 1 cuando la similitud es baja, hasta 4 cuando los nombres son exactamente iguales.

  - Salidas: Se listan las salidas de las tareas del proceso de consulta y se comparan con la lista de las mismas para el proceso de comparación. Para valorar este criterio, se busca si para cada tarea del proceso de consulta existe una y sólo una tarea del proceso de comparación, cuya lista de salidas posea elementos semejantes. Al relacionar un par de tareas se asigna un valor numérico de similitud a esta relación, que va desde 1 cuando la similitud es baja, hasta 4 cuando los nombres son exactamente iguales.

Mediante la aplicación de estos criterios en la comparación intuitiva de procesos, se busca ayudarle a los evalaudores de la relevancia, a emitir juicios estructurados y por lo tanto acertados acerca de la similitud. Además, la inclusión de las descripciones en lenguaje natural tanto de la funcionalidad del procesos, como de las actividades que los componen, facilitarán el correcto entendimiento de ellos y que se observará en la valoración de cada uno de los criterios. Por lo anterior, se espera que las discrepancias entre las evalauciones emitidas por diferentes personas, no sean tan marcadas y por lo tanto se obtenga una clasificación confiable de los procesos.

RESUMEN: La elección de los criterios a evalauar para las comparaciones manuales de procesos de negocio, es el paso más importante en el desarrollo de este proyecto ya que de ellos depende la extracción correcta de los elementos relevantes de una colección para las cosultas especificadas. A lo largo de este capítulo se hizo un recuento de los aspectos que tienen en cuenta las personas para evaluar similitud tanto en el campo de la recuperación de información, como en la de servicios; y finalmente se excogió un conjunto de ellos, teniendo en cuenta los niveles que consideran las herramientas automáticas para sus comparaciones, y las características más importantes de los procesos que los comparadores analizan cuando evalaúan semejanzas.

CAPÍTULO 4

DEFINICIÓN DE LAS MEDIDAS DE CALIDAD DE LA RECUPERACIÓN DE PROCESOS DE NEGOCIO

Evaluar los sistemas de recuperación de procesos de negocio ha sido una tarea dispendiosa por los muchos inconvenientes que presenta la adopción de metodologías utilizadas para evaluar SRI (Sistemas de recuperación de Información). Entre los problemas comúnmente expuestos están, las diferencias entre los lenguajes y las técnicas de recuperación utilizadas por las herramientas para el descubrimiento de servicios web y procesos de negocio, ya que esto impide hacer evaluaciones comparativas entre los diferentes sistemas disponibles. Por otra parte, generar colecciones de prueba no es tan sencillo como crear bases de datos documentales para los SRI. Los diversos lenguajes, ontologías y técnicas de emparejamiento, hacen demorada su elaboración.

Sin embargo, evaluar es una necesidad y hace parte del proceso evolutivo de toda investigación. Por esta razón, a pesar de los inconvenientes que se presentan, se siguen adoptando paradigmas relacionados con la evaluación de SRI para aplicar en el campo expuesto en este trabajo.

En este capítulo, se discutirán las medidas que se han venido utilizando para valorar la calidad de la recuperación de los sistemas de recuperación de información y su adopción en la evaluación de las herramientas para el descubrimiento de servicios web y procesos de negocio.

4.1 MEDIDAS DE CALIDAD DE RECUPERACIÓN DE INFORMACIÓN

La evaluación de los sistemas de recuperación de información, se ha desarrollado paralelamente a su evolución por razones científicas y por la necesidad de someterlos, como cualquier otro sistema, a una valoración de su efectividad por parte de los usuarios. Borlund [BOR00] manifiesta que “la tradición de la evaluación de los SRI se estableció desde la realización de los experimentos de *Cranfield* [CLE72], seguido de los resultados y experiencias que Lancaster desarrolló en la evaluación de *MEDLARS* [LAN68], los diversos proyectos *SMART* de Salton [SAL83], y actualmente en las conferencias TREC (Text REtrieval Conference)”.

Como se ha venido expresando a lo largo de esta monografía, los SRI se evalúan mediante dos ejes fundamentales: La evaluación del funcionamiento de la herramienta, y la evaluación de la recuperación, donde el segundo aspecto se valora mediante la comparación de los documentos recuperados automáticamente

frente a aquellos considerados como relevantes para un grupo de expertos en el dominio.

Teniendo en cuenta lo anterior, habría que definir entonces cuando un documento es relevante, sin embargo, dada la naturaleza subjetiva de este parámetro, habría que tener en cuenta diversas circunstancias:

- La relevancia o no de un mismo documento, puede ser determinada de forma diferente por dos personas dependiendo de los motivos que producen la necesidad de información, o el grado de conocimiento de cada una sobre la materia en cuestión. Es más, en un caso extremo, un mismo documento puede parecer relevante o no a una misma persona en instantes diferentes.

- El concepto de relevancia no puede ser definido con exactitud ya que es poco objetivo, es decir, que puede ser explicado de múltiples maneras por personas distintas. Además, un usuario puede afirmar que un resultado es relevante a sus necesidades, pero sin poder determinar los criterios que lo llevaron a esta decisión. Esto, no obstante, no le resta importancia al concepto como tal, sino que hace parte del amplio conjunto de procesos cognitivos cotidianos que llevan a cabo los humanos pero que generalmente, no se describen de forma clara [BLA90].

- Finalmente, un resultado no puede catalogarse categóricamente como relevante con un tema, debido a que, en ocasiones un apartado del documento puede tener información relacionada con el tema consultado, pero no en el resto de sus contenidos. Por esta razón, algunos autores advierten que la relevancia no puede medirse en términos binarios (sí/no), sino en términos de una función continua (muy relevante, relevante, escasamente relevante, mínimamente relevante, etc) [MAR02].

Por lo anterior, se infiere que el concepto de *relevancia* podría ser inadecuado para evaluar los resultados de un SRI debido a su alto grado de subjetividad. Acerca de esto, Cooper [COO73] opina que la forma correcta de afrontar el problema es definir la *relevancia* en términos de la "utilidad de un documento", de esta manera, es más fácil para una persona explicar cuándo una información es útil. De igual manera, Blair [BLA90] considera que aunque la evaluación de la "utilidad" es subjetiva, es un criterio más fácil de medir, y refleja la calidad de un sistema en tanto que proporciona la satisfacción de un usuario con los resultados.

Así mismo, Frants [FRA97] propone otra acepción de *relevancia* como "eficiencia funcional", donde un SRI alcanzará altos niveles de este valor cuando la mayoría de

los documentos recuperados satisfagan la demanda de información de quien hace una consulta, es decir, cuando los resultados sean útiles.

Lancaster [LAN93], por su lado, opina que la *relevancia* de un documento recuperado no debe coincidir estrictamente con los juicios que en este aspecto hacen los expertos, sino que depende de la satisfacción del usuario y la utilidad provista por los contenidos, sugiriendo que es mejor relacionar este caso con el término *pertinencia*. De esta forma, la *relevancia* se asocia con la relación entre la temática consultada y el contenido de los documentos recuperados, mientras que la *pertinencia* se restringe a la relación de utilidad de un resultado y una necesidad informativa individual.

Foskett [FOS72] tiene una ponencia similar, define como *documento relevante* a aquel "documento perteneciente al campo/materia/universo del discurso delimitado por los términos de la pregunta, establecido por el consenso de los trabajadores en ese campo", igualmente define como documento pertinente a "aquel documento que añade nueva información a la previamente almacenada en la mente del usuario, y que le resulta útil en el trabajo que ha propiciado la pregunta".

Estas designaciones acerca de la *relevancia* se encuentran aún vigentes y por lo tanto, Martínez [MAR02] propone que un documento se considere relevante cuando aporte algún contenido relacionado con la petición expuesta, y así, se puede hablar de *pertinencia* siempre que se haga referencia al punto de vista del usuario final que realiza una operación de recuperación de información.

Como se ha dicho anteriormente la evaluación de los SRI se divide en lo que llamaron "acceso físico" y "acceso lógico", cabe ahora introducir las medidas que reflejan estos conceptos.

En 1966 Cleverdon [CLE72] presentó seis medidas para determinar qué evaluar de un SRI. Estas son la cobertura de la colección, el tiempo de respuesta del sistema a una petición, la forma de presentación de los resultados, el esfuerzo realizado por un usuario, la exhaustividad y la precisión del sistema. Rijsbergen [RIJ, 1999], opina que las cuatro primeras medidas son intuitivas y fácilmente estimables, pero que "la efectividad es una medida de la capacidad del sistema para satisfacer al usuario en términos de la *relevancia* de los documentos recuperados", y por lo tanto son la *precisión* y la *exhaustividad* las medidas indicadas para este fin.

Chowdhury [CHO99], propone también seis medidas divididas en dos grupos:

- El primero lo conforman *la cobertura* (proporción de las referencias que potencialmente podrían haberse recuperado), la *exhaustividad* y el *tiempo de respuesta del sistema*.

- El segundo lo forman la *precisión*, la usabilidad (el valor de las referencias considerado en términos de fiabilidad, comprensión, actualidad, etc.) y la *presentación* (la forma en la que los resultados de la búsqueda son presentados al usuario)".

Junto a las medidas anteriormente mencionadas, diversos autores han empleado una amplia serie de estas, basadas en otros criterios. Meadow las sintetiza todas en tres grupos: medidas basadas en la *relevancia*, medidas del proceso y medidas del resultado [MEA93], donde las primeras, son las de mayor importancia y se han venido discutiendo a lo largo de este trabajo. Por otra parte, las medidas del segundo grupo, las basadas en el proceso, son utilizadas para diferenciar unos sistemas de otros, a partir de las prestaciones de la aplicación informática subyacente, y no permiten la evaluación de aspectos relacionados con el contenido de los documentos. Finalmente, el tercer grupo, se encuentra muy relacionado con las medidas basadas en la relevancia, aunque introducen algunos aspectos diferenciadores.

- *Medidas basadas en la Relevancia*
  - *Precisión*: Determina el porcentaje de acierto de una operación de recuperación de información. Como se ha dicho en apartados anteriores, corresponde a los documentos relevantes recuperados divididos entre el total de documentos recuperados.
  - *Exhaustividad*: Determina la profundidad de una operación de recuperación de información. Corresponde a los documentos relevantes recuperados dividido entre el total de documentos relevantes.
  - *Promedio de la efectividad*: Promedios de la efectividad en pares de valores de exhaustividad y precisión.

- *Medidas basadas en el Proceso*
  - *Selección*: Mide cuántos documentos hay en la base de datos, el grado de solapamiento con otras relacionadas y lo que se espera de la base de datos antes de las búsquedas.
  - *Contenido*: Tipo y temática de los documentos de la base de datos.
  - *Traducción de una consulta*: Verifica si el usuario puede plantear la consulta directamente o precisa de intermediación.
  - *Errores en establecimiento de la consulta*: Media de los errores sintácticos en la escritura de la búsqueda que propician la recuperación de conjuntos vacíos y erróneos.
  - *Tiempo medio de realización de la búsqueda*: Tiempo medio de realización de una estrategia de búsqueda.

- *Dificultad en la realización de la búsqueda*: Se refiere al tiempo mencionado anteriormente, adicionando los problemas que los usuarios inexpertos pueden encontrar.
- *Número de comandos precisos para una búsqueda*: Promedio de instrucciones necesarias para realizar una búsqueda.
- *Coste de la búsqueda*: Costes directos e indirectos en su realización.
- *Número de documentos recuperados*: extensión del resultado de una búsqueda.
- *Número de documentos revisados por el usuario*: Promedio del número de documentos que los usuarios están dispuestos a revisar.

- *Medidas de resultado*
  - *Precisión, exhaustividad y promedio de efectividad:* Son las mismas que se definieron en las medidas basadas en la relevancia.
  - *Medidas promedio de la satisfacción del usuario*: Pretende medir la reacción de los usuarios ante el resultado de una búsqueda.

Como se ha venido diciendo a lo largo del documento, las medidas basadas en la relevancia son las que proporcionan una evaluación de la eficacia de la recuperación de un sistema, por esta razón se estudiarán con más detalle.

La efectividad de la recuperación es una medida puramente de la habilidad de un sistema para satisfacer al usuario en términos de los documentos recuperados. Inicialmente, Baeza-yates [BAY99] se concentró en medirla mediante la precisión y la exhaustividad, y para esto, introdujo una tabla de contingencia como la que se muestra a continuación.

	RELEVANTE	**NO RELEVANTE**	
RECUPERADO	$A \cap B$	$\bar{A} \cap B$	B
NO RECUPERADO	$A \cap \bar{B}$	$\bar{A} \cap \bar{B}$	$\bar{B}$
	A	$\bar{A}$	N

(N = Número de documentos de la colección)

Tabla 3 Tabla de contingencia para la relevancia

Un gran número de medidas para determinar la eficacia de un sistema de recuperación se pueden derivar de esta tabla. Como por ejemplo:

- *La precisión:* Generalmente se define como la habilidad del sistema para recuperar solamente elementos relevantes. Representa la fracción del resultado que es relevante para una consulta específica [SAL83]. Esta medida está relacionada con los conceptos de ruido y silencio informativo, presentándose mayor ruido en tanto más se aproxime a cero este parámetro.

$$Precision = \frac{|A \cap B|}{|B|}$$

- *La exhaustividad*: En Inglés "*recall*", es la proporción de material relevante recuperado, del total de los documentos que son relevantes en la base de datos, independientemente de que éstos, se recuperen o no. Esta medida fue llamada más adelante como *probabilidad condicional de un ítem* por Swet [SWE63] y en 1964, como *sensibilidad* por Goffman y Newil [GOF64]. Salton [SAL83] propuso la siguiente ecuación para calcularla:

$$Exhaustividad = \frac{|A \cap B|}{|A|}$$

Una recuperación perfecta se presenta cuando el resultado de la exhaustividad es 1, es decir, el sistema evaluado recuperó todos los documentos relevantes en la base de datos, y no se exhibe ruido ni silencio informativo.

De la precisión y la exhaustividad se puede obtener un par ordenado de resultados, y cuando estos puntos se unen geométricamente se obtiene una curva, que generalmente describe el desempeño del sistema para cada consulta. Para medir el desempeño total de una herramienta, el conjunto de curvas antes descritas se combinan para producir una curva promedio [BAY99].

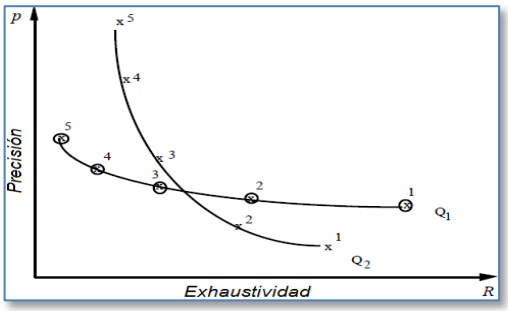

Figura 2 Curva de Precisión-Exhaustividad para dos consultas. Los números indican los valores de un parámetro de control

- Fallout, es una medida que relaciona los resultados obtenidos que no son relevantes, con respecto al total de ellos en la colección. Es decir que refleja la porción de elementos erróneos recuperados.

$$Fallout = \frac{|\bar{A} \cap B|}{\bar{A}}$$

Existe una relación funcional entre los tres parámetros anteriores llamado *Generalidad* (*G*) la cual mide la densidad de documentos relevantes en la colección.

$G = \frac{\bar{A}}{N}$ y $P = \frac{R*G}{(R*G)+F(1-G)}$

Pasando a otro punto, para medir adecuadamente la efectividad de la recuperación de información de manera estandarizada, se necesita una colección de prueba que contenga:

- Una colección de documentos
- Un conjunto de necesidades informativas de prueba, expresadas como consultas
- Un conjunto de juicios de relevancia para cada una de las consultas planteadas, respecto de cada uno de los documentos de la colección.

La forma estandarizada de evaluar un SRI se basa en la forma binaria de establecer la relevancia, es decir, si cada elemento recuperado es o no, relevante a la consulta formulada. Sin embargo, como se ha venido diciendo la relevancia también se puede medir mediante escalas graduadas que permiten establecer niveles para este parámetro.

Baeza-Yates [BAY99] hace un estudio acerca de las medidas estandarizadas para evaluar la eficacia de la recuperación, y la generalización de ellas para los resultados graduados de relevancia. Por ejemplo, cuando la salida de de una estrategia de recuperación depende de un parámetro como la posición en una clasificación (ranking), ó el número de términos que una consulta tiene en común con un documento (nivel de coordinación); las medidas varían para cada caso, y en consecuencia también lo hacen las ecuaciones planteadas para la precisión y la exhaustividad.

Cuando la recuperación se hace mediante el nivel de coordinación (λ), *S* es el conjunto de consultas, y:

$\left|\tilde{A}\right| = \sum_{s \in S} |A_s|$, donde A_s es el conjunto de documentos relevantes para una consulta *s*.

$\left|\tilde{B}_\lambda\right| = \sum_{s\in S}|B_{\lambda s}|$, donde $B_{\lambda s}$ es el conjunto de documentos recuperados bajo el nivel de coordinación λ.

De lo anterior se obtienen las ecuaciones para la Precisión P_λ y la Exhaustividad R_λ.

$P_\lambda = \sum_{s\in S}\frac{|A_s \cap B_{\lambda s}|}{|\tilde{A}|}$ y $R_\lambda = \sum_{s\in S}\frac{|A_s \cap B_{\lambda s}|}{|\widetilde{B_\lambda}|}$

La insatisfacción con los métodos para medir la efectividad mediante un par de números (precisión y exhaustividad) que pueden co-variar de forma imprecisa, han llevado a plantear medidas compuestas, las cuales están en gran parte basadas en la tabla de contingencia (tabla 4). El inconveniente de estas medidas es que muchas de ellas son planteadas para un fin específico, y no son justificadas en una forma racional. Algunos ejemplos son:

- La suma de la precisión y la exhaustividad: $S = P + R$
- Una medida propuesta por Borko [BAY99]: $BK = P + R - 1$
- Otras:
    - $Q = \frac{R-F}{R+F-2RF}$ (F=Fallout)
    - $V = 1 - \frac{1}{2\left(\frac{1}{P}\right)+2\left(\frac{1}{R}\right)-3}$, Medida de Vickery [MRS09].

Baeza-Yates [BYA99], hace un estudio amplio de las medidas que se aplican para evaluar la efectividad de la recuperación de los SRI, presentando diferentes escenarios y referenciando su aplicación en otros trabajos. No obstante, cabe resaltar que en su gran mayoría, las medidas presentadas allí son redefiniciones de la precisión y la exhaustividad y combinaciones de ellas, de acuerdo con los casos de estudio.

4.2 MEDIDAS DE CALIDAD DE RECUPERACIÓN DE SERVICIOS WEB

Como se ha mencionado con anterioridad, las metodologías para la evaluación de los sistemas para el descubrimiento de servicios web, toman como base los estudios que al respecto se han hecho para los SRI. El caso de las medidas para evaluar la calidad de la recuperación automática de procesos de negocio, no es la excepción. Al igual que para los SRI, la efectividad de la recuperación se mide en términos de la *relevancia* de los elementos recuperados de acuerdo con una consulta especificada.

En el capítulo 2 de este documento, se citaron trabajos como el de Tsetsos [TAH06] y el de Küster [KUK09] donde se compararon distintas formas de establecer la *relevancia*. Ambos autores coinciden en que este parámetro puede medirse de

forma binaria, asignando un valor de "1" cuando el servicio ofrecido presenta algún grado de similitud con la consulta, y "0" cuando no existe ninguna relación entre ellos. No obstante, los dos trabajos critican esta manera de evaluar por considerar que dos servicios pueden presentar diferentes niveles de semejanza entre sí, y por lo tanto no es lo mismo establecer como completamente relevante un resultado que satisface completamente la consulta, a aquel que sólo lo logra parcialmente.

Para la evaluación de la efectividad de la recuperación a partir de unos valores de relevancia binarios, Tsetsos cita a Baeza-Yates [BAY99] para la utilización de las medidas estándar de *precisión* y *exhaustividad* que se muestran a continuación:

$$Precision = \frac{Porcentaje\ de\ los\ elementos\ relevantes\ recuperados}{porcentaje\ de\ los\ elementos\ recuperados}$$

$$Exhaustividad = \frac{Porcentaje\ de\ los\ elementos\ relevantes\ recuperados}{porcentaje\ de\ los\ elementos\ relevantes}$$

No obstante, Tsetsos hace notar que la mayoría de los sistemas para el descubrimiento de servicios, presentan los resultados jerarquizando los elementos recuperados de acuerdo con un grado de concordancia con la consulta, lo que indica que para poder aplicar las medidas, habría que convertir los resultados del sistema a un formato binario. De nuevo, cabe recordar que los problemas que produce esta conversión de escalas se explicaron en el capítulo 2.

Por otra parte, para superar la inconveniencia de las evaluaciones booleanas, Tsetsos propone crear una escala de *relevancia* en la que se puedan asignar diferentes niveles de este parámetro en la comparación de servicios. En este caso, un elemento ofrecido para una consulta, puede ser *muy relevante, relevante, algo relevante, poco relevante* e *irrelevante*. Al igual que para la evaluación binaria, este autor referencia a Buell [BUK81] para escoger las ecuaciones de *precisión* y *exhaustividad* para el caso en cuestión. Las medidas mostradas a continuación son una generalización de las estándar, y se calculan a partir de la clasificación de los servicios recuperados por el sistema (f_e) y el ranking obtenido de los juicios de relevancia (f_r).

$$Precisión = \frac{\sum_{S_i \epsilon S} \min\{f_r(R,S_i), f_e(R,S_i)\}}{\sum_{S_i \epsilon S} f_e(R,S_i)} \qquad Exhhaustividad = \frac{\sum_{S_i \epsilon S} \min\{f_r(R,S_i), f_e(R,S_i)\}}{\sum_{S_i \epsilon S} f_r(R,S_i)}$$

De estas ecuaciones se deduce que la *precisión* será máxima cuando los resultados del sistema son más rigurosos que los juicios de relevancia, es decir, cuando el grado de correspondencia asignado a un elemento recuperado, es menor que el

proporcionado por los expertos. El comportamiento de la *exhaustividad* es, como siempre, inversamente proporcional a la *precisión.*

Por otra parte, una medida que relaciona a la precisión y a la exhaustividad, es "overall" ó "match accuracy" que representa la calidad de la recuperación, teniendo en cuenta el esfuerzo adicional que debe hacer un usuario para borrar los elementos irrelevantes y agregar los elementos no obtenidos.

$$Overall = exhaustividad\left(2 - \frac{1}{Precisión}\right)$$

De esta medida se debe tener en cuenta que sólo tiene sentido cuando la precisión no es menor a 0.5, es decir que al menos la mitad de los elementos recuperados sean correctos, de otro modo esta medida se hace negativa [COR08].

Otra forma de medir la efectividad de la recuperación teniendo en cuenta el ranking de los servicios recuperados por el sistema, y la clasificación obtenida de los juicios de relevancia, es calcular la *precisión* en términos de la cantidad de elementos recuperados que son relevantes y que concuerdan con el ordenamiento de este último. Para esto, se establece el número de resultados a evaluar, es decir, que para el ranking de la plataforma y los resultados de los juicios de relevancia se toman los primeros *k* de cada uno y se comparan entre sí. De esta forma, la *precisión* puede calcularse obteniendo el número de servicios dentro de la clasificación provista por herramienta automática, que se encuentran en la misma posición que el ordenamiento obtenido de los expertos.

$$Precisión_k = \frac{[RecRel_k]Rel_p}{k}$$

Donde $[RecRel_k]Rel_p$, hace referencia al número de servicios recuperados como relevantes que se encuentran en la misma posición que el ranking proporcionado por una evaluación manual de similitud, y *k* es la cantidad de resultados a tener en cuenta [COR08].

Adicionalmente, se presenta la *precisión media* que es una medida ampliamente difundida para capturar el desempeño del sistema, y permite medir este parámetro en términos de las exhaustividad. Si se designa *L* como el conjunto de elementos recuperados como respuesta a una consulta, y a *r* como $1 \leq r \leq |L|$, podría decirse que $isrel(r) = 1$ cuando el elemento recuperado en la posición *r* es relevante, y será "0" en el caso opuesto. Finalmente, el número de elementos relevantes entre el top *r* de los elementos recuperados está dado por la siguiente

suma: $count(r) = \sum_{i=1}^{r} isrel(r)$. De acuerdo con esto, se ofrece la ecuación para esta medida.

$$AveP = \frac{1}{|R|} \sum_{r=1}^{|L|} isrel(r) \frac{count(r)}{r}$$

Alrededor del año 2000, se incrementó el interés en las medidas basadas en relevancia continua o graduada y, varias propuestas se han hecho para generalizar la medida anteriormente descrita para evaluar la relevancia en estos términos, en donde la mayoría de dichas propuestas, se basan o pueden ser expresadas en términos de una *ganancia acumulativa*. Intuitivamente, esta medida en una posición del ranking *r*, mide la ganancia que un usuario recibe al escanear los elementos del top *r* en una lista ordenada de resultados. De manera más formal, un $g(r) \geq 0$ denota el valor de ganancia (ó nivel de relevancia) de un ítem en una posición *r.* Para este caso $isrel(r) = 1$ si $g(r) \geq 0$, de otro modo será "0". De esta forma la *ganancia acumulativa* se define como: $cg(r) = \sum_{i=0}^{r} g(r)$. De manera similar, se puede definir una ganancia acumulativa ideal $icg(r)$ cuando $\forall(r > 1, r \leq |R|): isreal(r) = 1$, y $\forall(r > 1): g(r) \leq g(r-1)$ [KKO08].

Como $cg(r)$ puede tomar valores arbitrarios grandes para consultas con varios ítems de relevancia, esta ganancia debe ser normalizada para calcular la media o para comparar los resultados a través de las consultas. En consecuencia, la *ganancia acumulativa normalizada* se define como el desempeño relativo de la relevancia para un óptimo comportamiento de la recuperación.

$$ncg(r) = \frac{cg(r)}{icg(r)}$$

Lo anterior permite una extensión directa de *Precisión media* que en ocasiones es llamada *Precisión media ponderada* (AWP por sus siglas en inglés Average Weighted Precisión) donde:

$$AWP = \frac{1}{|R|} \sum_{r=1}^{|L|} isrel(r) \frac{cg(r)}{icg(r)}$$

Desafortunadamente, $ncg(r)$ tiene un defecto que hereda AWP, y es que como $icg(r)$ tiene un límite superior fijo ($icg(r) \leq icg(|R|)$), $ncg(r)$ y AWP no pueden penalizar la recuperación tardía de elementos relevantes de manera adecuada porque $ncg(r)$ no puede distinguir a partir de qué rango los elementos relevantes son recuperados para rangos mayores que R. Esto se puede ilustrar comparando

$ncg(r)$ y la $precisión_r$ para el último en la fila en una salida completa ($R \subseteq L$). En este caso, $ncg(|L|) = 1$ pero la $Precisión(|L|) = \frac{|R|}{|L|}$ es usualmente mucho menor que "1".

Para resolver el problema anterior, Järvelin and Kekäläinen [JAK02] proponen el uso de una *Ganancia acumulada reducida* $dcg(r) = \sum_{i=0}^{r} \frac{g(r)}{disc\,(r)}$, y *Ganancia acumulada reducida ideal* (*idcg(r)*). De estas nuevas medidas se desprende una nueva adopción de la AWP que se llama *Precisión media ponderada reducida* AWDP:

$$AWDP = \frac{1}{|R|} \sum_{r=1}^{|L|} isrel(r) \frac{dcg(r)}{idcg(r)}$$

De forma similar, Kishida [KIS05] propone una generalización para AveP que además evita el defecto del AWP:

$$genAveP = \frac{\sum_{r=1}^{|L|} isrel(r) \frac{cg(r)}{r}}{\sum_{r=1}^{|R|} \frac{icg(r)}{r}}$$

Adicionalmente, Sakai [SAK04] propone una integración del AWP y el AveP llamada medida-Q que hereda propiedades de ambas medidas y propone un parámetro β para controlar si esta medida se comporta más como AWP o como AveP.

$$Qmesure = \frac{1}{|R|} \sum_{r=1}^{|L|} isrel(r) \frac{\beta cg(r) + count(r)}{\beta icg(r) + r}$$

Como se observó a lo largo de este capítulo, la calidad de la recuperación provista por un algoritmo, se mide en general por la relación existente entre los elementos (documentos, servicios, procesos, etc) considerados como relevantes por un grupo de personas, y el conjunto retornado por una herramienta automática a partir de una consulta. Esta relación se mide comúnmente aplicando ecuaciones para determinar la precisión y la exhaustividad en las diferentes formas estudiadas con anterioridad, y adicionalmente, por medio de la combinación de ellas para determinar desde diversas perspectivas el acierto de una recuperación.

A continuación, se exponen las medidas escogidas para medir la efectividad del algoritmo a evaluar en este trabajo, y lo que se espera obtener de cada una de ellas.

4.3 SELECCIÓN, JUSTIFICACIÓN Y DESCRIPCIÓN DE LAS ECUACIONES PARA MEDIR LA EFECTIVIDAD DE LA RECUPERACIÓN (EXHAUSTIVIDAD, PRECISIÓN Y OVERALL)

La finalidad de este capítulo es obtener un conjunto de medidas que describan el desempeño de una herramienta automática para el descubrimiento de procesos de negocio, en términos de su acierto en la recuperación de material relevante. Por este motivo, las ecuaciones expuestas aquí van encaminadas a dar un juicio de efectividad, teniendo en cuenta los resultados obtenidos de las evaluaciones manuales realizadas sobre la plataforma desarrollada en este trabajo.

En general, la evaluación de la efectividad de las herramientas automáticas para el descubrimiento de procesos de negocio, depende de su capacidad para satisfacer una consulta de acuerdo con lo que considera relevante un grupo de personas. Sin embargo, al principio de este capítulo se percibieron opiniones contradictorias acerca de la conveniencia de usar este parámetro debido a su naturaleza subjetiva, y se dieron diferentes definiciones para sortear este inconveniente. No obstante, debido a que en este trabajo, los juicios de relevancia emitidos por los participantes en la evaluación de procesos, son emitidos en razón de la similitud entre ellos y no por una necesidad personal, se decidió que en adelante, el término *relevancia* hará referencia a los niveles de semejanza entre los elementos comparados.

Como se puede notar a lo largo de este documento, las medidas más difundidas y aceptadas para la evaluación de la efectividad de la recuperación en términos de la relevancia son las medidas de Precisión y Exhaustividad, calculadas a partir de valores binarios o graduados de este parámetro. Para el caso que aquí se estudia, tanto la plataforma para la comparación manual de procesos como el algoritmo, proporcionan valores de similitud en una escala de cero a uno, haciendo obvia la necesidad de aplicar ecuaciones que permitan tener en cuenta esta graduación. En el apartado anterior de este capítulo se trataron algunas medidas utilizadas en trabajos relacionados. Específicamente Tsetsos, propone la aplicación de una generalización de las medidas convencionales para la exhaustividad y la precisión, que tienen en cuenta resultados escalados, y que serán adoptados en este trabajo.

A pesar de las discrepancias acerca de medir la calidad de recuperación mediante una relevancia binaria, se considera interesante analizar la variación de las medidas de efectividad cuando se tienen en cuenta solo el número de elementos recuperados, y cuando se tiene en cuenta la graduación. Por esta razón, se aplicarán dos formas de evaluar la efectividad que se llamarán en adelante, evaluación binaria y evaluación graduada.

Con la evaluación binaria se busca medir la habilidad de la herramienta automática para extraer procesos que tienen algún nivel de similitud, sin importar el valor que le asigne. Con este fin, se convertirán los datos obtenidos de las dos fuentes (algoritmo y manualmente), de una serie de valores a una escala binaria, asignando un "1" a todas las comparaciones cuyo valor sea diferente de cero; y "0" en el caso contrario. Así, se mide la precisión calculando la porción de los elementos recuperados que son relevantes, y se obtiene una idea de la capacidad de acierto del algoritmo. Además, la exhaustividad medida en estas condiciones, revela la fracción de los elementos relevantes que la herramienta automática encontró algún grado de similitud, y así conocer la sensibilidad del algoritmo. Las ecuaciones para este caso son las que se muestran a continuación, donde *RelRec* representa el número de elementos recuperados que son relevantes; *Rec* es la cantidad de procesos recuperados por el algoritmo; y *Rel* conjunto de elementos considerados relevantes en las evaluaciones manuales.

$$Precisión = \frac{RelRec}{Rec} \qquad Exhhaustividad = \frac{RelRec}{Rel}$$

Por otra parte, en la evaluación teniendo en cuenta la graduación (evaluación graduada) [TAH06], se pretende además de corroborar los resultados obtenido de forma binaria, analizar la rigurosidad del algoritmo. De esta manera se observará que la precisión será mayor en tanto el algoritmo asigne valores menores que los provistos por las personas, y por lo tanto se considere más riguroso; o cuando los valores de ambas fuentes no sean muy disímiles. En conclusión, entre más severo sea el algoritmo al asignar valores de similitud a los elementos que encuentra relevantes, y entre menos procesos erróneos retorne, la precisión será más alta. Contrario a lo anterior, la exhaustividad crecerá en tanto la herramienta automática sea menos severa, y que a su vez asigne algún valor (que no sea cero), a un gran número de procesos relevantes. Un total alto de esta medida, si bien podría interpretarse como una falta de rigurosidad del algoritmo, muestra su cercanía a los resultados de las personas, y por lo tanto se intuye que proveerá una mayor satisfacción. Las ecuaciones al final de este párrafo, ya han sido explicadas con anterioridad en este capítulo y representan las medidas para esta evaluación.

$$Precisión = \frac{\sum_{S_i \epsilon S} \min\{f_r(R,S_i), f_e(R,S_i)\}}{\sum_{S_i \epsilon S} f_e(R,S_i)} \qquad Exhhaustividad = \frac{\sum_{S_i \epsilon S} \min\{f_r(R,S_i), f_e(R,S_i)\}}{\sum_{S_i \epsilon S} f_r(R,S_i)}$$

Adicionalmente, se calcularán los valores para overall en ambos casos (evaluación binaria y gradada) mediante la ecuación mostrada a continuación. Como se expuso en el apartado anterior de este capítulo, esta medida representa la calidad de la

recuperación, teniendo en cuenta el esfuerzo adicional que debe hacer un usuario para borrar los elementos irrelevantes y agregar los elementos no obtenidos. De estos resultados se analizará el esfuerzo disminuye o aumenta cuando se tienen en cuenta la escala de valores. No obstante, hay que recordar que esta medida tiene sentido cuando la precisión es mayor que 0.5, es decir, cuando más de la mitad de los procesos recuperados son relevantes, de otro modo este valor sería negativo y significaría que el trabajo para conseguir el conjunto no recuperado, sería apenas comparable con tener que hacer todas las comparaciones manualmente.

$$Overall = exhaustividad\left(2 - \frac{1}{Precisión}\right)$$

Finalmente, se analizará la precisión de la herramienta automática al contabilizar la porción de los elementos que se recuperaron en la misma posición que las personas estipularon. Para ello, se calculará la ecuación precisión top k, para k=20, 15, 10 y 5. Pero además, se aplicará esta medida sin tener en cuenta el orden, y así analizar la capacidad de recuperación del algoritmo dentro de estos rangos. A continuación se recuerda la ecuación.

$$Precisión_k = \frac{[RecRel_k]Rel_p}{k}$$

RESUMEN: Este capítulo expuso un breve recorrido por las medidas más importantes para evaluar la efectividad de la recuperación de las herramientas automáticas para descubrimiento tanto de información, como servicios web. De allí se pudo observar que la forma más difundida para esta evaluación, es la aplicación de ecuaciones que miden la efectividad en términos de la relevancia de los elementos recuperados, y que los parámetros que más se ajustan son la precisión y la exhaustividad. A lo largo de esta sección de la monografía se presentaron diversas interpretaciones y formas de calcular estas medidas, y finalmente se escogieron las que más se ajustan al análisis que se desea hacer en este trabajo.

CAPÍTULO 5

PLATAFORMA PARA LA EVALAUCIÓN MANUAL DE PROCESOS DE NEGOCIO

La plataforma para la evaluación manual de procesos de negocio desarrollada en este trabajo, está diseñada para permitir la comparación intuitiva de ellos a partir de juicios de relevancia emitidos por los usuarios comparadores, mediante la evaluación de los criterios establecidos en el capítulo 3. Adicionalmente, la herramienta provee un módulo de gestión de procesos, donde se pueden generar y editar el conjunto de procesos de negocio de prueba. Y finalmente, proporciona una visualización de resultados que despliega el ranking de procesos relevantes para una consulta determinada, a partir de los resultados de las evaluaciones manuales.

A lo largo de este capítulo se describirán las funcionalidades de la herramienta, se expondrá y explicará su arquitectura, y finalmente se analizarán algunas de las interfaces de usuario más importantes.

5.1 DESCRIPCIÓN DE LA PLATAFORMA

En esta sección, se presentan los tipos de usuarios que soporta la plataforma, y se describen y explican las funcionalidades principales de la herramienta, teniendo en cuenta quienes tienen acceso a ellas.

La plataforma soporta dos tipos de usuario, el *Administrador* y un usuario *Comparador*, que se describen a continuación:

- *Administrador:* Es el usuario que tienen acceso a la información prioritaria de la aplicación. Dentro de sus funciones se encuentra la de gestionar los procesos de la colección que se carga en la plataforma. Además puede hacer gestión de todos los usuarios, es decir, que puede crear un usuario de cualquier tipo, editar su información y borrarlos. Adicionalmente, puede revisar los resultados de las comparaciones para cada proceso de consulta. Sin embargo, este tipo de usuario no tiene acceso a la realización de las comparaciones.

- *Comparador*: Es el encargado de realizar las evaluaciones manuales de los procesos. Es decir, es quien interactúa directamente con el módulo de comparación. Una persona puede registrarse como comparador, pero no tendrá acceso a los derechos del administrador. O sea, que un usuario de este tipo no podrá gestionar usuarios, ni gestionar procesos, y tampoco podrá consultar los resultados.

Cuando un usuario de este tipo se registra (Ver anexo A - Agregar Usuario), desarrolla una encuesta para medir su nivel de conocimiento acerca de servicios y procesos web, workflows, BPMO (Business Process Model Ontology); y su dominio del idioma Inglés. De acuerdo con los resultados, el *Comparador* se catalogará dentro de uno de tres tipos:

- Experto: Es quien alcanza más de un 60% en su nivel de conocimiento.
- Intermedio: Será el usuario cuyo conocimiento se encuentre dentro del 30% y el 60%.
- Novato: Será quien su conocimiento no alcance el 30%.

Una vez registrados, los usuarios tendrán acceso a las diferentes funcionalidades que ofrece la plataforma, las cuales se encuentran distribuidas dentro de los módulos que se describen en la arquitectura.

5.2 ARQUITECTURA

La herramienta está dividida en tres capas: La de aplicación que provee las interfaces de usuario y las funcionalidades básicas de la plataforma; la capa de mediación donde se encuentran todas las interfaces de programación de aplicaciones (API's) utilizadas en el desarrollo de la herramienta; y finalmente la capa de cimiento que incluye el software básico que se requiere para su funcionamiento.

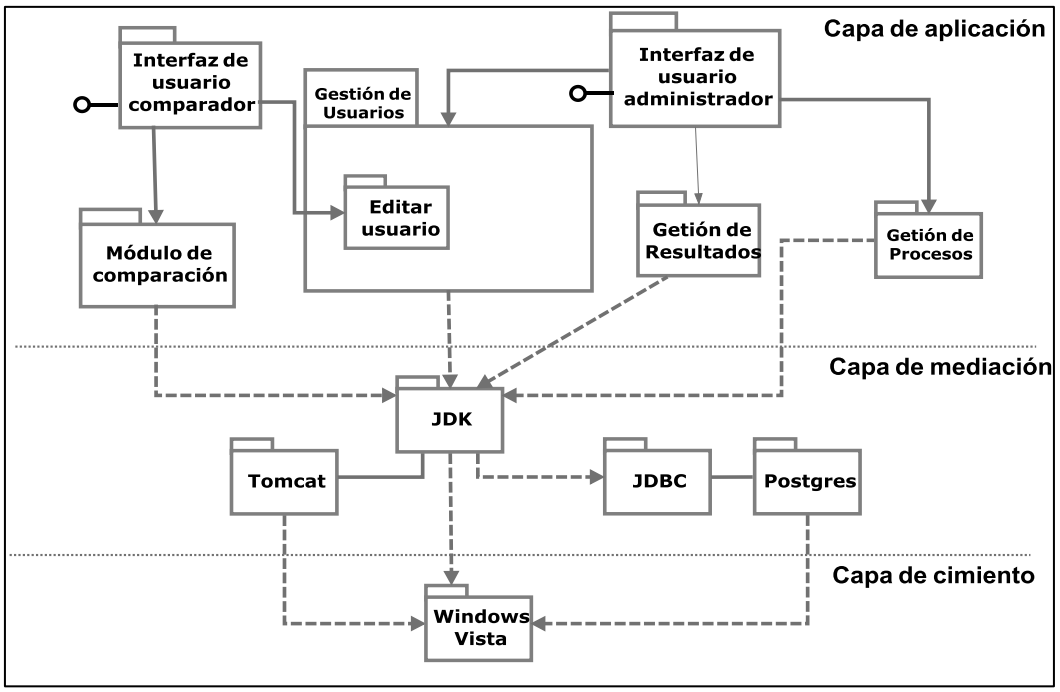

Figura 3 Arquitectura

- *Capa de aplicación*

  - *Interfaz de usuario administrador*: Proporciona al administrador el acceso a todas las funcionalidades de las que es responsable. Da acceso a las gestiones de usuarios, de resultados y de procesos.

  - *Interfaz de usuario comparador*: Provee a los usuarios las interfaces y el acceso a la funcionalidad, para realizar las comparaciones manuales entre procesos.

  - *Módulo de gestión de usuarios:* Este módulo permite agregar, editar y borrar usuarios en la plataforma y se comunica con la base de datos relacional del módulo de almacenamiento, para editar las tablas relacionadas con los usuarios. Sus principales funciones son:

    - *Agregar Usuario:* Un *Administrador* de la plataforma puede agregar cualquier tipo de usuario, sin embargo es el único que posee los permisos para agregar usuarios con su mismo rol. No obstante, en la página principal de la plataforma existe el link "*Register*" que permite a cualquier persona interesada en participar de las comparaciones, hacer un registro como *Comparador*.

    - *Editar Usuario*: El administrador tiene acceso a toda la información de los usuarios desde su menú de gestión de usuarios (Ver Anexo A Figura 3), y puede modificar la información relacionada con cualquiera de ellos. Por otra parte, una vez un *Comparador* ha iniciado su sesión, puede modificar la información que suministró en su registro inicial.

    - *Eliminar Usuario*: El derecho para borrar un usuario de la plataforma es exclusivo del *Administrador*. Desde la página de gestión de usuarios, puede eliminar usuarios de cualquier tipo y su información relacionada.

  - *Módulo de gestión de procesos:* La gestión de los procesos permite administrar los procesos de la colección WSML del Módulo de Almacenamiento, es decir, permite crear los procesos, especificar los que servirán de consulta, visualizarlos y eliminarlos. A las funciones que provee este módulo, sólo tienen acceso los Administradores de la plataforma. La gestión de los procesos incluye:

    - Agregar Procesos: Desde la interfaz para la gestión de los procesos (ver Anexo A figura 6), se puede acceder a la adición de Procesos a la colección. Para esto, el *Administrador* debe suministrar los archivos que describen el proceso a ingresar (.wsml y .wsml.layout).

Para la representación de los procesos de negocio se escogió BPMO (Business Process Modeling Ontology), que proporciona una representación gráfica clara y entendible para todo tipo de usuario, ya que fue pensada para crear una comunicación eficiente entre los analistas de negocio y los desarrolladores de software. Además, el modelador WSMOStudio[36] versión 0.7.3, permite adicionarles enriquecimiento ontológico y descripciones en lenguaje natural a todas los elementos de los procesos, y así facilitar su emparejamiento en herramientas automáticas que utilizan ontologías para las comparaciones; y adicionalmente, proveer la información textual a los comparadores como ayuda en la comprensión de los procesos. Asimismo, el lenguaje WSML describe todo el flujo de actividades representado por BPMO, agilizando la creación de colecciones de prueba. Un ejemplo de un proceso BPMO es el que se muestra en la siguiente figura 4:

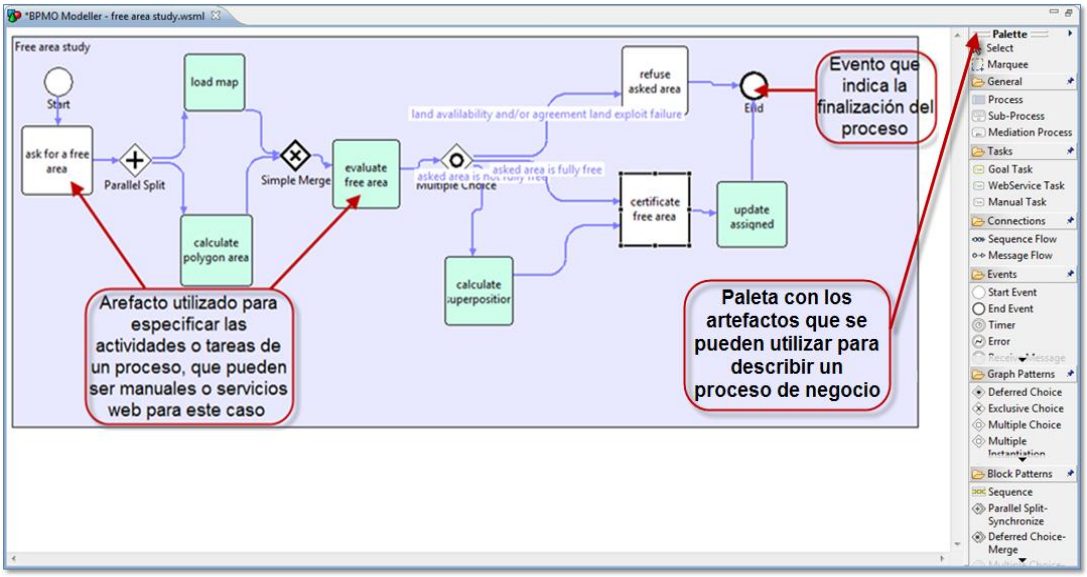

Figura 4 Ejemplo de un proceso BPMO

"Los servicios web semánticos SWS y las ontologías necesitan lenguajes formales para su especificación y así permitir su procesamiento automático. En cuanto a las descripciones de la ontología, la recomendación del W3C para este caso OWL[37] (Ontology Web Language) tiene sus limitaciones, tanto a nivel conceptual y en lo que respecta a algunas de sus propiedades formales [BPL05]. Una propuesta para la descripción de servicios web semánticos es OWL-S[38] (Semantic Markup for Web Services). Sin embargo, tiene serias limitaciones en el plano

[36] http://www.wsmo.org/
[37] http://www.w3.org/TR/owl-features/
[38] http://www.w3.org/Submission/OWL-S/

conceptual y además, las propiedades formales del lenguaje no están claras [LLR05]. Por ejemplo, OWL-S, ofrece la posibilidad de elegir entre diferentes lenguajes para la especificación de requisitos y efectos. No obstante, la descripción de entradas y salidas se hace con OWL pero no se aclara su interacción con los demás lenguajes. Estas cuestiones no resueltas fueron la principal motivación para ofrecer una alternativa, el lenguaje unificado para WSMO"[39].

En los objetivos a cumplir en el desarrollo de este proyecto, se planteaba la evaluación de la herramienta automática referenciada en [COR06], sin embargo durante el transcurso del trabajo, el grupo de investigación GIT (Grupo de Ingeniería Telemática) presentó una evolución del algoritmo encaminada al emparejamiento de procesos de negocios descritos en BPMO, por lo anterior, se decidió analizar el sistema con las mejoras [FIC10] y no el propuesto inicialmente.

Como se mencionó anteriormente, para adicionar un nuevo elemento a la colección en la plataforma, se requieren los archivos wsml y wsml.layout que describen el proceso BPMO. Una vez se han definido sus rutas, la aplicación extrae y guarda la información relacionada con el nuevo proceso, crea un gráfico que describe su flujo y lo almacena en una carpeta como un archivo de imagen.

- Visualizar Procesos: Mediante esta función el *Administrador* tiene acceso a toda la información relacionada con un proceso, es decir, puede ver el nombre, la descripción en lenguaje natural del funcionamiento del proceso, la lista de las actividades que lo componen con sus descripciones, parámetros de entrada y salida, la estructura gráfica (BPMO), y si el proceso es de consulta o no (Ver anexo A figura 8).

- Eliminar Procesos: Desde la página de gestión de procesos, el *Administrador* puede eliminar un proceso, borrando de la base de datos toda su información relacionada. No obstante, los archivos (wsml y wsml.layout) se conservan en la carpeta donde se almacenan los procesos de la colección.

- *Módulo de comparación:* A éste módulo sólo tienen acceso los *Comparadores*, y es el que permite hacer las evaluaciones manuales de los procesos. Su funcionamiento se encuentra explicado en el siguiente numeral (Uso de la herramienta) o en el Anexo A de este trabajo (Manual de usuario).

[39] http://www.wsmo.org/TR/d16/d16.1/v1.0/#sec:wsml-web-service-specification

- *Módulo de Resultados:* Es el módulo que permite obtener y visualizar la clasificación de los procesos relevantes para una consulta según las evaluaciones proporcionadas por los *Comparadores*. A las funciones que provee este módulo sólo tienen acceso los *Administradores*, y se describen en el siguiente numeral (Uso de la herramienta), y en el anexo A de este trabajo (Manual de usuario).

- *Capa de mediación*

  - *Tomcat:* Implementa los JSF (Java Server Faces) proporcionando un ambiente de ejecución en cooperación con el servidor web, para el código Java .

  - *JDK (Java Development Kit):* Es un ambiente de desarrollo integrado (IDE) para escribir aplicaciones Java. Se compone por un ambiente de ejecución que se encuentra en la capa superior del sistema operativo, y permite compilar, depurar y ejecutar la herramienta descrita en este capítulo.

  - *JDBC:* Es un API para Java que define la forma como un usuario debe acceder a la base de datos, y provee métodos para consultarla y editarla. Esta interfaz es utilizada por los paquetes de la capa de aplicación que implementan la lógica de la aplicación, para gestionar los datos dentro de la base de datos Postgres.

  - *Postgres:* Es el sistema para la administración de bases de datos utilizado para almacenar la información de los usuarios, y los datos recopilados de las evaluaciones manuales.

- *Capa de cimiento:*

  - *Windows Vista Home Premium*: Es el sistema operativo sobre el cual se ejecutan todas las aplicaciones necesarias para el funcionamiento de la plataforma.

5.3 EJEMPLOS DE APLICACIÓN

El manual de usuario de la herramienta se puede consultar en el anexo A de este trabajo, sin embargo, vale la pena explicar la forma detallada cómo funcionan dos de los módulos más importantes: El módulo comparación y el de gestión de resultados.

5.3.1 Módulo de comparación

El módulo de comparación es el que permite visualizar y comparar una lista de procesos de consulta con el conjunto de procesos del repositorio.

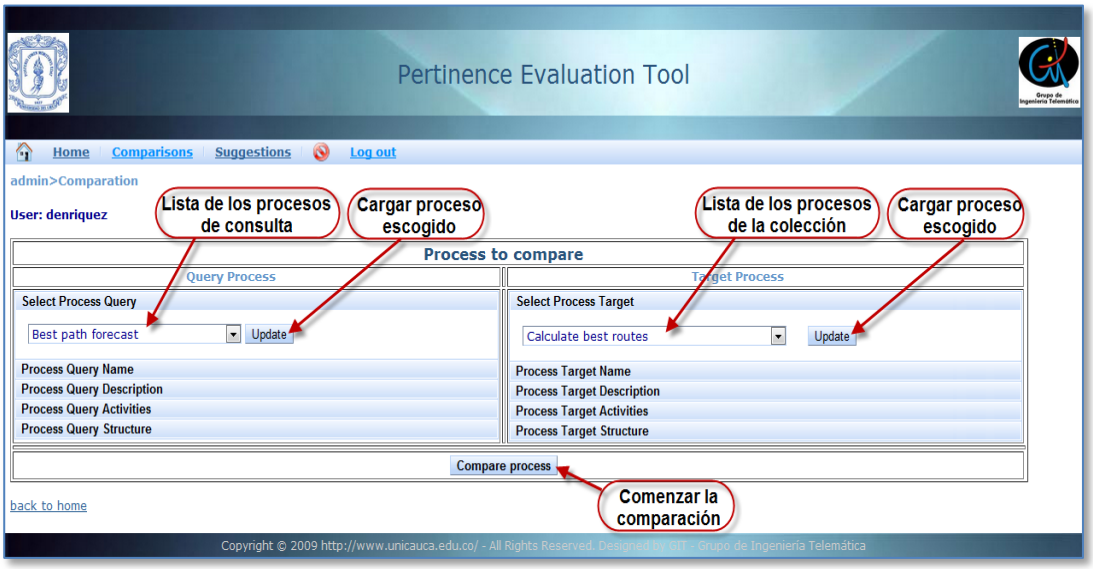

Figura 5 Interfaz principal para la comparación de procesos

La interfaz principal de comparación, muestra dos bloques paralelos en los que se expone la información de los procesos a comparar. En el bloque de la izquierda, se presenta la lista de los procesos de consulta, es decir, aquellos que servirán de modelo para buscar sus *relevantes* dentro de la colección de procesos que se muestra en el bloque de la derecha. Para ambos bloques es posible seleccionar los procesos que se desean evaluar y cargar su información. Cada proceso tiene disponible para mostrar, su nombre, una descripción en lenguaje natural de su objetivo principal, una representación gráfica que muestra el flujo de ejecución, y una lista de las actividades que lo componen, con una breve descripción y la lista de las entradas y las salidas.(figura 5).

Cada vez que un usuario *Comparador* entra a la interfaz de comparación, la aplicación consulta en la base de datos las evaluaciones finalizadas, y en consecuencia, para cada proceso de consulta carga sólo los procesos del repositorio que aún no se han evaluado, y aquellos cuyas comparaciones están incompletas.

Cuando ya se han escogido los procesos a comparar, el usuario empieza la evaluación dando clic en el botón "*Compare Process*" que se encuentra en la parte inferior de la pantalla. Inmediatamente, la aplicación muestra el primer formulario con los criterios para comparar los procesos completos, y una

ventana emergente con la información de cada uno de ellos (figuras 6 y 7 respectivamente).

La evaluación consta de cinco partes. La comparación de los procesos completos y las comparaciones de las actividades por nombre, descripción, entradas y salidas.

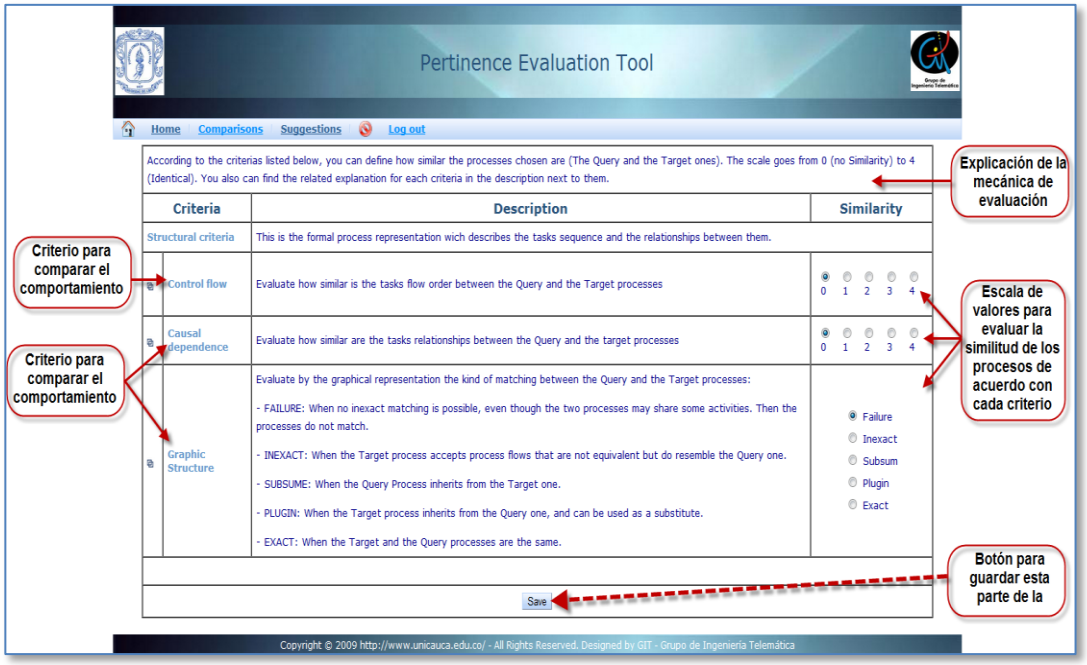

Figura 6 Interfaz de comparación para los criterios relacionados con la estructura y el comportamiento

En la primera parte, se exponen los criterios relacionados con la estructura y el comportamiento de los procesos. Como se estableció en el capítulo 3 de este trabajo, los criterios relacionados con la estructura son *la dependencia causal,* y *la estructura gráfica*; mientras que para el comportamiento se comparan los procesos de acuerdo con el *flujo de control*.

En la interfaz (figura 6), se muestra una breve explicación de la forma como debe llevarse a cabo la evaluación, se definen cada uno de los tópicos antes mencionados, y finalmente se presenta una escala para establecer el grado de similitud de los procesos a comparar según cada criterio. Una vez terminada esta porción de la evaluación, el usuario guarda los datos suministrados y estos son almacenados en la base de datos.

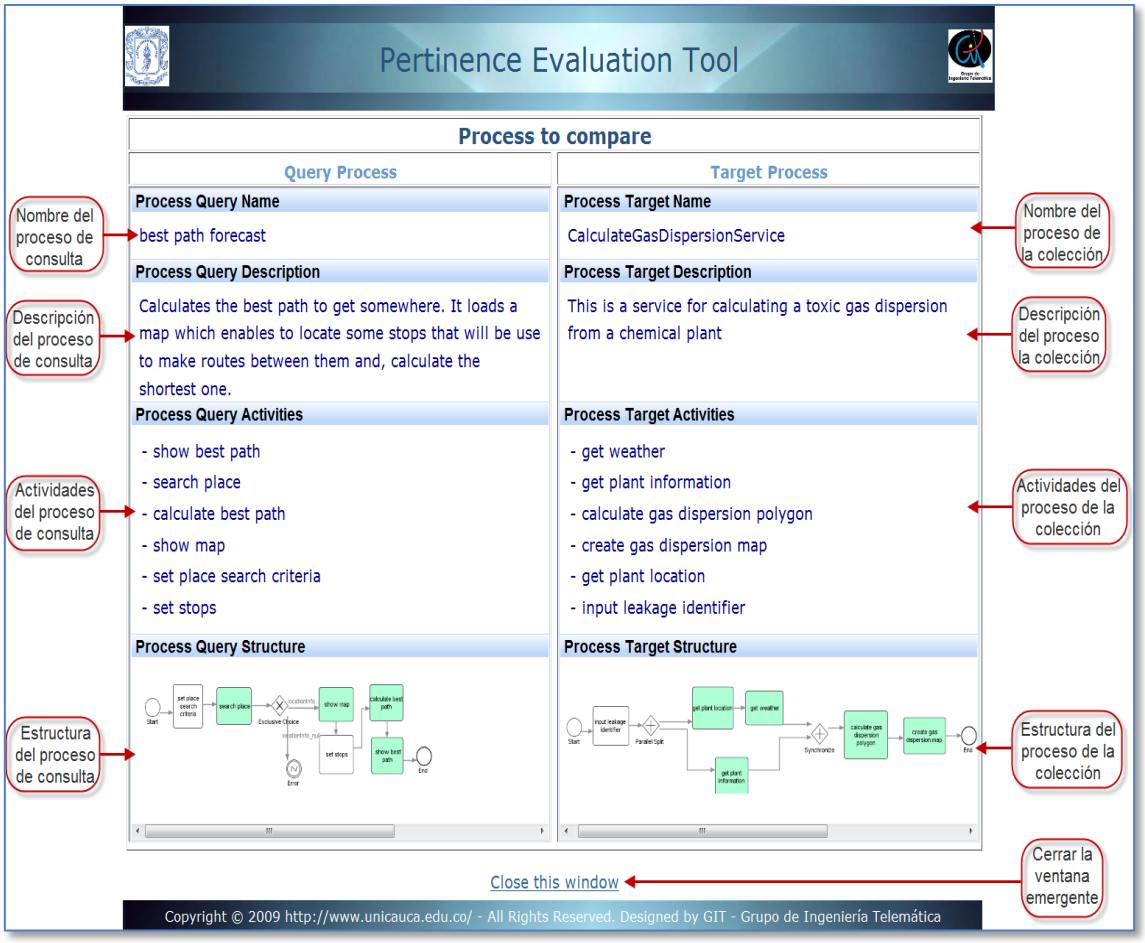

Figura 7 Ventana emergente con la información de los procesos a comparar

Cuando se establecen los niveles de similitud para cada criterio, el usuario hace clic en el botón "*Save*" que se encuentra en la parte inferior de la pantalla. Para almacenar los resultados, la base de datos cuenta con una tabla donde se relacionan los identificadores de los procesos comparados, el identificador del usuario, un número de comparación y un campo de chequeo para cada parte de la evaluación. Al finalizar esta primera parte, se guardan los datos anteriormente mencionados y se establece como "verdadero" el campo relacionado con este segmento. Adicionalmente, existe en la base de datos una tabla para almacenar los resultados de esta parte de la evaluación, donde se recolectan los valores obtenidos para cada criterio y se relacionan con el número de la comparación.

La idea de asignar valores verdaderos a cada segmento de evaluación finalizado, tiene como propósito dejar que los usuarios realicen las comparaciones al ritmo que deseen. De esta manera no se pierde información ya recopilada y se retoma el proceso desde el punto donde quedó. Es decir, que si una persona abandona la evaluación una vez ha guardado la información del primer

segmento, cuando regrese a comparar, la aplicación cargará el formato para el segundo.

A continuación, comienza la evaluación de la segunda parte, y en este punto se carga la primera interfaz para la comparación de las actividades de los procesos (figura 8). Para empezar, se evalúa si existe similitud entre los nombres. La interfaz muestra una lista de los nombres de las actividades del proceso de consulta, y frente a cada una de ellas una lista desplegable con los nombres de las actividades del proceso del repositorio. Así, para cada actividad del primer proceso, se le puede relacionar una y sólo una actividad del segundo, estableciendo un valor de similitud para esta relación. No obstante, si el *Comparador* no encuentra actividades afines puede simplemente continuar con la evaluación y no guardar ningún dato para esta parte. De forma similar, se realizan las comparaciones de las descripciones, de las entradas y las salidas de las actividades.

Al igual que en la fase inicial de evaluación, cada vez que se termina una sección se establece como verdadero el campo correspondiente a la parte en cuestión en la tabla de comparación. Para los datos obtenidos en la comparación de las actividades, existe otra tabla en la base de datos que recolecta los valores de similitud relacionados con los identificadores las actividades comparadas y el número de comparación de la tabla principal.

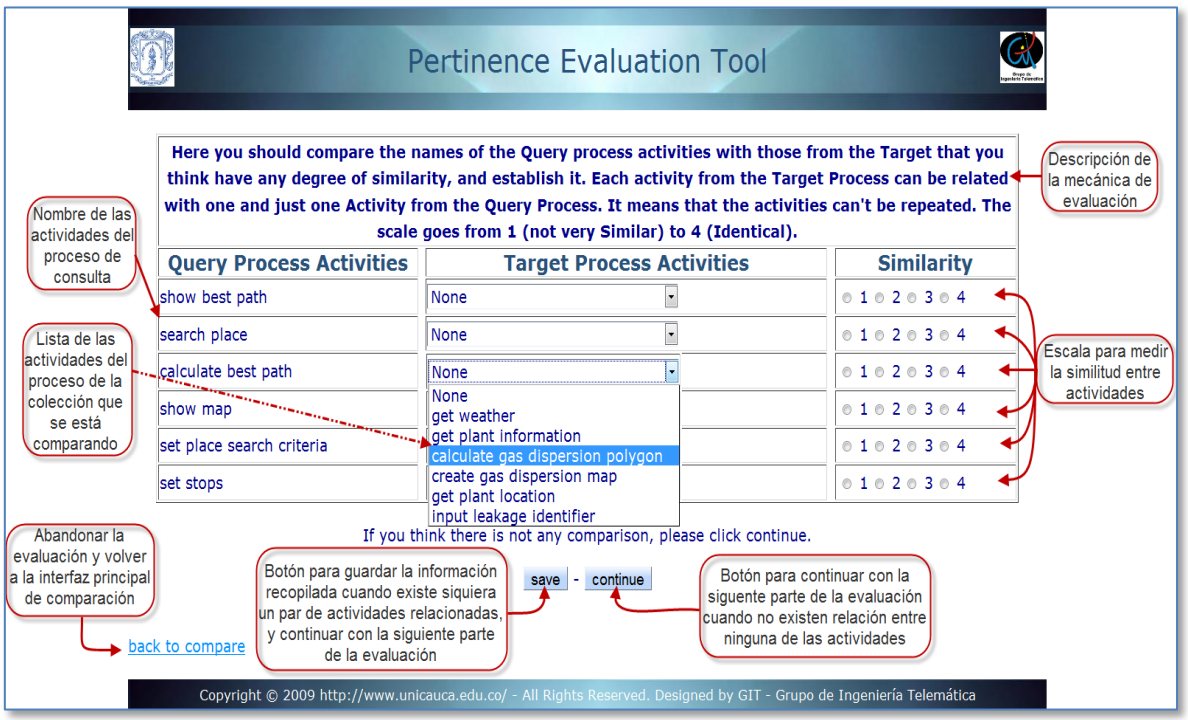

Figura 8 Interfaz de comparación actividades por nombre

La metodología para la comparación de actividades por descripción, entradas y salidas, es igual a la formulada para la evaluación de las actividades por nombre, no obstante, las interfaces muestran una tabla adicional en la que se proporciona la información pertinente para cada etapa. La figura 9 expone un ejemplo de la comparación de las actividades por sus entradas, como se puede observar, la tabla adicional provee una lista de las entradas a las actividades y la tabla para la asignación de valores de similitud es exactamente igual a la de la figura 9.

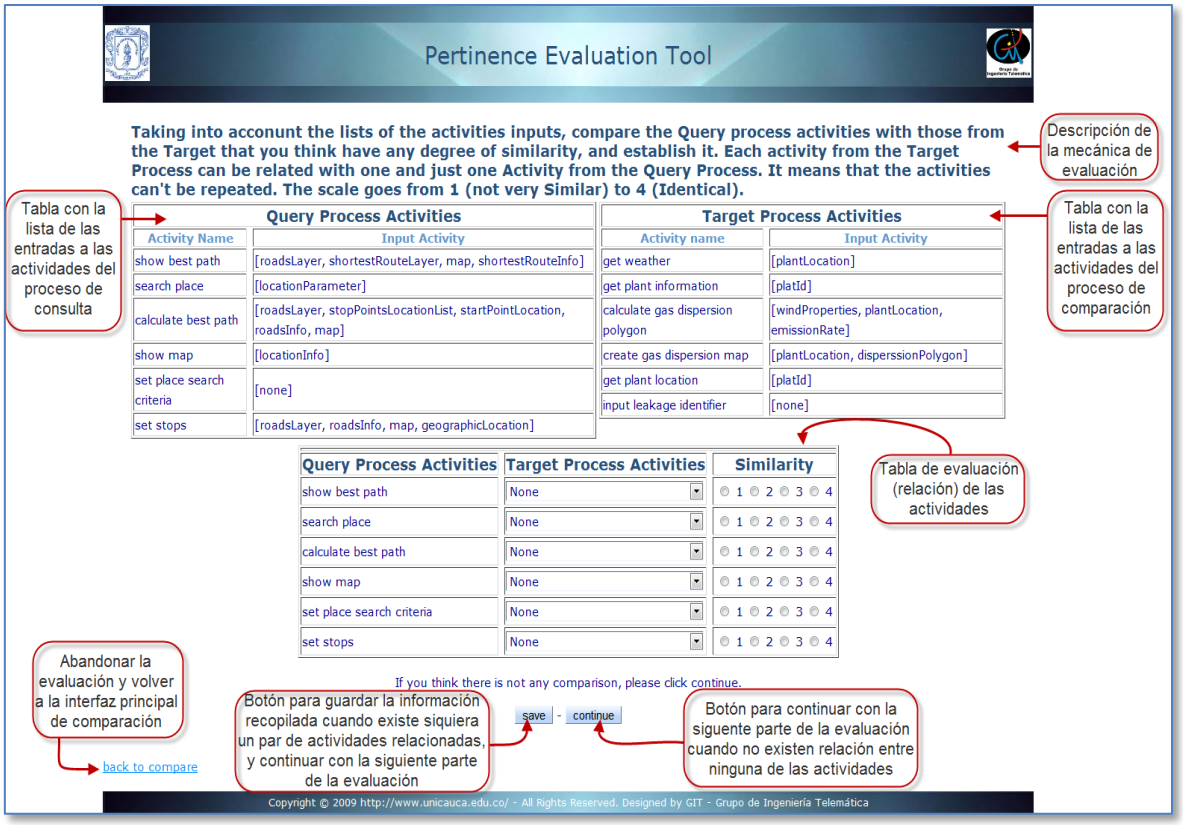

Figura 9 Interfaz para la comparación de actividades por sus entradas

5.3.2 Módulo de gestión de resultados

Para la consulta de los resultados, el *Administrador* de la plataforma accede a la página principal de estos, haciendo clic en el vínculo "*Results*" que se encuentra en la barra de links asociada a este tipo de usuario.

Al acceder a la página principal de los resultados, la interfaz muestra una tabla con los procesos de consulta para los cuales ya hay resultados. Como puede verse en la figura 10, la lista de los procesos muestra además de su nombre, el identificador y un link llamado "*Top*". Cuando el usuario pasa el cursor sobre cada identificador, se muestra un cuadro con información adicional del proceso (descripción y estructura gráfica).

Para encontrar el ranking de los procesos relevante para una consulta, debe darse clic en link "*Top*" asociado al proceso que describe la consulta deseada. Una vez escogido, la aplicación carga todos los resultados obtenidos para este proceso, normaliza los resultados para cada criterio y guarda los datos en una tabla en la base de datos. La normalización de los datos se utiliza para unificar en una misma escala los valores resultantes de la evaluación de los criterios.

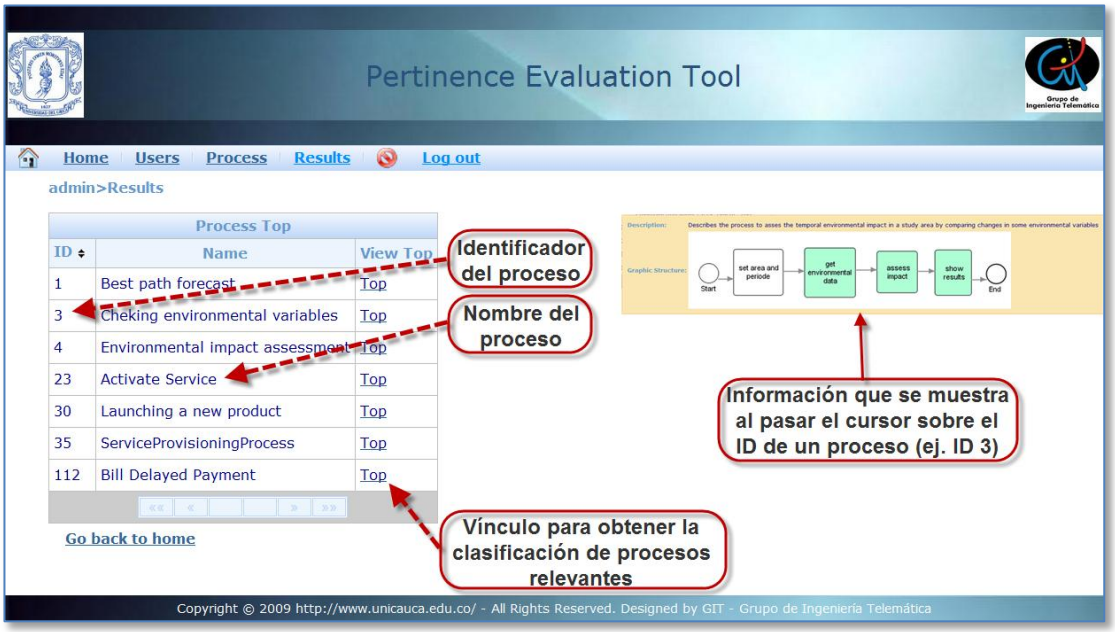

Figura 10 Interfaz principal de los resultados

Paralelo a la normalización de los datos, la plataforma carga una interfaz que permite al *Administrador* escoger los criterios de los cuales desea obtener información y el nivel de importancia que quiere darle a cada uno. De esta manera, para cada criterio se muestra una lista desplegable con un porcentaje que va de cero a cien, y así, al final de la elección, el porcentaje total entre los criterios seleccionados debe ser cien (figura 11). Por ejemplo, si se desea encontrar el ranking de procesos relevantes para una consulta teniendo en cuenta el "*control de flujo*" y "*dependencia causal*", cada uno con el mismo nivel de importancia, se asigna el 50% para ambos criterios, y el resto de ellos quedan con un 0%.

Finalmente, cuando los criterios son escogidos y se les han asignado sus porcentajes de importancia, se debe dar clic en el botón "*Get the top*" que se encuentra al final de la tabla, para calcular los porcentajes de similitud entre procesos y armar el ranking para el proceso de consulta solicitado.
Para calcular el porcentaje de similitud para una comparación, se aplican los criterios anteriormente escogidos sobre los resultados normalizados y se multiplica por cien el resultado obtenido. Cuando este procedimiento se ha

hecho para todos los resultados del proceso de consulta solicitado, se procede a calcular una moda para los valores arrojados por todos usuarios para cada par de procesos comparados. Para entender mejor, si se tienen los dos conjuntos siguientes:

- Procesos Consulta={A,B}
- Procesos Colección={a,b,c}

Y se cuenta con los siguientes *Comparadores*:

- Conjunto de Usuarios={1,2,3}

El conjunto de las evaluaciones posibles por usuario será:

- Conjunto de Comparaciones={(A,a),(A,b),(A,c),(B,a),(B,b),(B,c)}

Para cada par ordenado existe un valor de similitud por usuario, es decir, que en este caso hay 3 valores (usuarios 1, 2 y 3), para cada una de las comparaciones posibles ((A,a), (A,b), (A,c), (B,a), etc). Entonces, para obtener sólo un valor de similitud para una comparación, debe calcularse un valor que refleje la estimación de semejanza de la mayoría de los *Comparadores*. En estadística, *la Moda* es el valor con una mayor frecuencia (repetición) en una distribución de datos [CAN88], se considera el cálculo de esta medida como la mejor solución al planteamiento anterior.

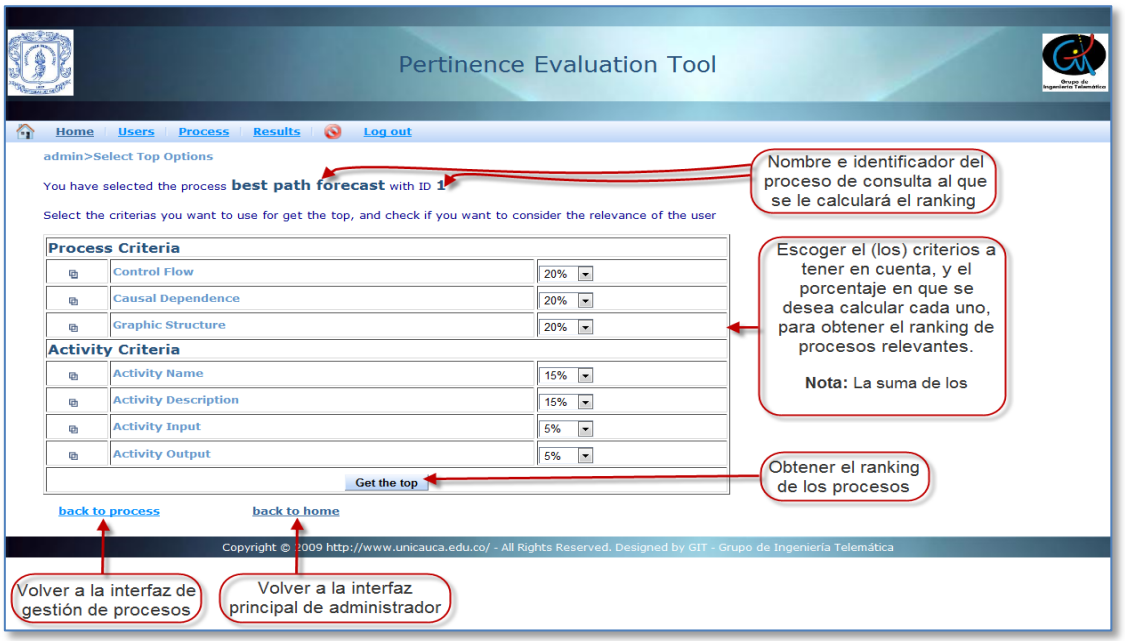

Figura 11 Asignación de criterios y sus porcentajes para calcular el ranking de procesos relevantes para un proceso de consulta (en este caso el proceso Best path forecast)

Una vez se calcula un valor único de similitud para cada comparación, se organizan de mayor a menor los procesos relevantes para una consulta. Este cálculo se hace para cada tipo de comparador (experto, novato e intermedio) y los resultados se visualizan en la interfaz que se muestra en la figura 11 del Anexo A. Cada tabla muestra los primeros 15 procesos relevantes, y se pueden ver los siguientes dando clic en las barras de desplazamiento de cada una.

RESUMEN: Este capítulo describe la plataforma para la evaluación manual de procesos, mediante la presentación de los tipos de usuarios soportados, la arquitectura de la aplicación y la definición de todos sus componentes. Finalmente, se exponen figuras de algunos pantallazos de la plataforma para explicar sus funcionalidades básicas. Este último apartado está soportado por el anexo A donde se ofrece el manual para los usuarios de la herramienta.

CAPÍTULO 6

EVALUACIÓN Y ANÁLISIS DE RESULTADOS

En este capítulo se exponen los resultados obtenidos de las comparaciones manuales a partir del uso de la plataforma desarrollada en este proyecto, y se utilizan como modelo para analizar la efectividad de la recuperación del algoritmo [FIC10] mediante la aplicación de las medidas de calidad presentadas en el capítulo 4.

A continuación se explicará la metodología de evaluación empleada para la recopilación de los datos y el análisis de estos; posteriormente, se mostrarán algunos ejemplos de aplicación del análisis de resultados; y finalmente se evaluará el nivel de concordancia entre los jueces a partir de los datos obtenidos de las comparaciones manuales de procesos.

6.1 METODOLOGÍA DE EVALUACIÓN

Esta sección describe la forma como se evaluó la efectividad de la recuperación del algoritmo desarrollado por Figueroa [FIC10] en su tesis de maestría. En seguida se enumeran los pasos seguidos para obtener los datos y analizarlos.

Primero, por medio del uso de la plataforma para la evaluación manual de procesos *Pertinence Evaluation Tool*, un grupo de personas calificaron la relevancia de los procesos de negocio de la colección de prueba (Anexo B) de acuerdo con unas consultas. Segundo, se aplicó el algoritmo sobre el mismo conjunto de procesos para obtener los similares para cada consulta. Tercero, mediante las medidas de calidad descritas en el capítulo 4, se analizó si los resultados de la herramienta automática satisfacen lo considerado relevante por las personas. Y finalmente, a partir de los valores obtenidos del punto anterior se emitió un juicio acerca de la efectividad de la herramienta evaluada.

La plataforma para la evaluación de sistemas de recuperación de procesos de negocio *Pertinece Evaluation Tool*, se publicó en http://pertinence.blogdns.com:8081/pertinence/ durante los meses de Febrero a Junio de 2010, con todas las funcionalidades descritas en el capítulo 5 y cuyo manual de usuario puede consultarse en el anexo A. Para las comparaciones manuales, se ofreció una colección de procesos compuesta por 40 del dominio del geoprocesamiento y 60 del dominio de las telecomunicaciones (Anexo B - Tabla 1); y de estos se escogieron 7, 3 del primero y 4 del segundo como elementos de consulta. Adicionalmente, 13 personas se registraron para desarrollar las evaluaciones pero sólo 5 las concluyeron en su totalidad, por consiguiente, el

análisis se efectuó con dichos resultados. En adelante, se llamarán Jueces a los 5 evaluadores (Anexo C - Tabla 2), y este grupo estará dividido en los tres rangos de comparadores: Experto (3 jueces), Intermedio (1 juez) y Novato (1 juez). De acuerdo con lo anterior, cada juez suministró 700 comparaciones, 100 por cada elemento de consulta, para un total de 3500 evaluaciones.

Por otra parte, la plataforma para la evaluación manual de procesos, permite calcular y visualizar los resultados de las comparaciones intuitivas, teniendo en cuenta la asignación de porcentajes a los criterios analizados (capítulo 3), y así estudiar la efectividad de la recuperación del algoritmo de Figueroa desde diferentes perspectivas.

Para la aplicación de la herramienta automática se utilizó la misma colección de procesos de negocio y se recopilaron datos de similitud para los 7 procesos de consulta. No obstante, el algoritmo emplea una gran porción de memoria virtual para las comparaciones, retornando errores por memoria insuficiente en varios casos. Para superar este inconveniente, se probó reservar un espacio exclusivo para la ejecución de la aplicación, pero el IDE NetBeans 6.9 sólo permitió hacerlo hasta 1536MB y desafortunadamente no fue suficiente para obtener la totalidad de los resultados.

Además, el algoritmo presentó una falencia al cotejar procesos que tienen realimentación en su flujo, y debido a que el conjunto de procesos contenía 3 de este tipo (*Broadband service activation, DSL provisión, Test of the environmental impact*), se presentó una pérdida de 3 resultados para cada consulta. Adicionalmente, la herramienta automática no arrojó resultados para el servicio *Web Service Approval* al ser cotejado con las consultas Q1 y Q2. En resumen, puede consultarse la tabla 1 del anexo C que muestra los nombres e identificadores para cada proceso de consulta y el número de comparaciones obtenidas manual y automáticamente; y en el anexo B se relacionan los procesos para los cuales no hubo resultados. Dado que las comparaciones para las cuales no se obtuvo un valor de similitud se presentaron por la carencia de memoria virtual y no por un error del algoritmo, se decidió hacer el análisis con los procesos para los cuales hubo resultados del sistema automático y no con la totalidad de los elementos de la colección. De otro modo, si se les asigna un valor de similitud de cero (0) a todas las comparaciones sin respuesta, se considerarían como irrelevantes procesos que el algoritmo hubiera podido recuperar si las condiciones del IDE hubieran sido propicias.

La herramienta automática provee 3 valores de semejanza para cada comparación, analizando una similitud básica que cuantifica la distancia estructural entre dos

procesos; una similitud de nodo que relaciona el valor antes mencionado con el número de nodos comparados; y finalmente una similitud de secuencia que vincula la primera medida con el número de secuencias analizadas de acuerdo con de nodos coincidentes en los procesos comparados. La tabla 4 del anexo C muestra los parámetros antes mencionados y para cada uno de ellos diferentes formas de analizarlos teniendo en cuenta diversos resultados de las evaluaciones manuales. En general, se obtuvieron valores para cada uno de los nueve casos expuestos en la tabla, no obstante en este capítulo se analizarán tres de ellos, pero las gráficas y los resultados de las medidas de calidad para todos los análisis, pueden examinarse en el anexo C.

Para el análisis de la efectividad de la recuperación de la herramienta automática estudiada, se utilizaron las ecuaciones descritas en la última sección del capítulo 4. Como se expuso allí, se harán dos tipos de evaluación, la binaria y la gradada, y se aplicarán las medidas de calidad de la recuperación para cada una de ellas.
En resumen las medidas a evaluar son:

- R y P (evaluación binaria) relacionan el número de procesos relevantes recuperados con los elementos considerados como relevantes por los usuarios y los recuperados por el algoritmo. Para medir estos parámetros, se tomaron como relevantes los procesos para los cuales las evaluaciones manuales arrojaron algún valor de similitud.

- Rg y Pg (evaluación gradada) relacionan el valor de similitud asignado por los expertos y el algoritmo para cada comparación ofreciendo una noción de la rigurosidad de la recuperación de la herramienta automática. Es decir que mientras R y P proporcionan una idea del acierto en el número de elementos relevantes recuperados, Rg y Pg muestran la relación entre los valores de similitud obtenidos por cada medio (manual y automático).

- P top k que evalúa el número de procesos recuperados por la herramienta automática que se encuentran en igual posición que la esperadas por las comparaciones manuales. K hace referencia a la cantidad de elementos a tener en cuenta, que en este caso serán 20, 15, 10 y 5. Adicionalmente, se hará el mismo análisis descrito aquí, pero sin tener en cuenta el ordenamiento, sino sólo el número de procesos recuperados que son relevantes dentro de los intervalos mencionados.

6.2 EJEMPLOS DE APLICACIÓN

Como se dijo en el apartado anterior, la plataforma para la evaluación manual de procesos permite obtener múltiples conjuntos de resultados relevantes para una consulta, dependiendo de los criterios que se deseen tener en cuenta. Para el análisis presentado en seguida, se evaluaron los tres resultados proveídos por el algoritmo por cada consulta, con respecto a los datos obtenidos de las evaluaciones intuitivas mediante las combinaciones de criterios presentadas en la tabla 4 del anexo C.

- *Similitud básica*

  Este valor retorna la similitud como una función de la distancia estructural encontrada por el algoritmo de comparación, y se evaluó frente a los resultados de relevancia obtenidos considerando dos combinaciones de todos los criterios y una con los criterios estructurales, relacionados en los análisis A, B y C de la tabla 4 del anexo C.

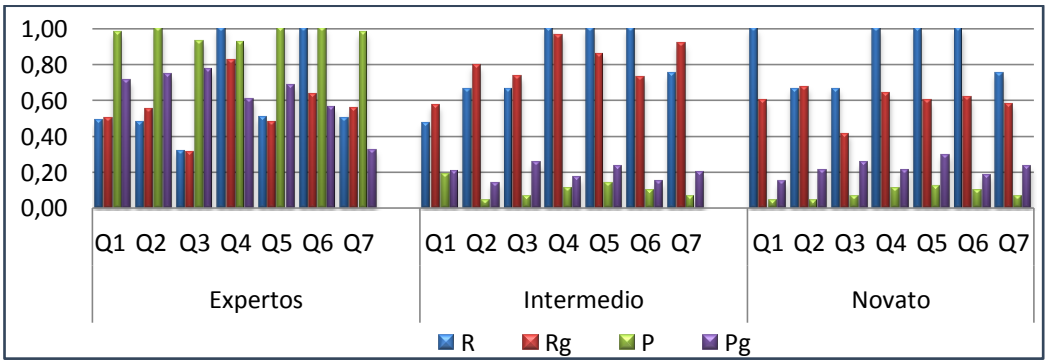

Figura 12 Resultados comparativos entre jueces para la precisión y la exhaustividad – Análisis A

La figura 12 expone los valores para la precisión y la exhaustividad obtenidos a partir de los resultados de similitud básica provistos por el algoritmo, y la clasificación de procesos relevantes de acuerdo con la distribución de criterios para el análisis A (tablas 6 y 7 Anexo C).

Experto: Los resultados de la exhaustividad binaria muestran que para los usuarios de este tipo, el algoritmo recuperó alrededor de la mitad de los elementos relevantes para las consultas Q1, Q2, Q5 y Q7, sin embargo, al observar su comportamiento para las mismas teniendo en cuenta la graduación de la similitud, se evidencia que los valores para este parámetro aumentan en mínimas proporciones, haciendo indiscutible que la rigurosidad del algoritmo es pobre. En cuanto a la precisión se observan valores de más de un 90% para cada consulta, comprobando que los

procesos recuperados por la herramienta automática son en su mayoría relevantes en cada caso. No obstante, los porcentajes para esta medida disminuyen al tener en cuenta los valores reales de similitud, esto puede deberse a que el algoritmo asigna valores mayores a los establecidos manualmente haciéndolo poco riguroso al evaluar.

De otra parte, se observa que para la consulta Q3 el valor de la exhaustividad es bajo, esto se debe a que el algoritmo recuperó una porción pequeña de elementos considerados relevantes por los expertos. Además, el nivel de este parámetro al tener en cuenta la gradación no sufrió una variación significativa, pero la precisión disminuyó, por lo tanto podría decirse que los valores de relevancia (jueces) fueron menores o cercanos a los de similitud (algoritmo), corroborando una vez más que el algoritmo no es muy riguroso. A pesar de lo anterior, la herramienta automática está siendo acertada con respecto al material recuperado, ya que por el resultado de la precisión puede concluirse que al menos los elementos extraídos en su mayoría son relevantes.

Las consultas Q4 y Q6 tienen un tratamiento especial. Dado que para este análisis sólo se tuvieron en cuenta los procesos para los cuales el algoritmo proporcionó algún nivel de similitud, es poco confiable el resultado de la exhaustividad ya que no se puede analizar si todos los elementos relevantes se recuperaron. En este caso los valores para este parámetro son altos (1), pero eso lo único que garantiza es que al menos lo que recuperó es relevante. Por su parte, la precisión confirma lo anterior, aunque para Q4 se evidencia la presencia de unos pocos *falsos positivos,* para Q6 mostró un valor perfecto.

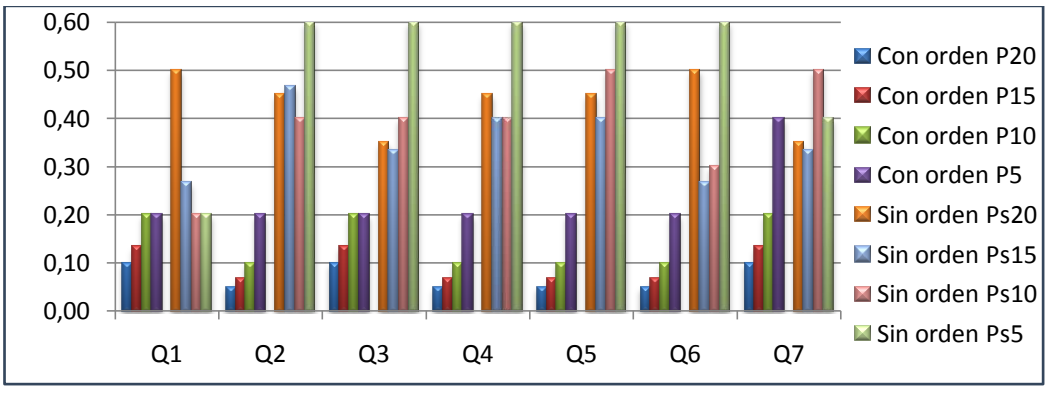

Figura 13 Resultado de la precisión top k para expertos – Análisis A

Como se mencionó en el apartado anterior de este capítulo, la precisión top k sólo fue posible medirla para los jueces expertos, ya que en la mayoría de

los casos, los otros dos tipos de evaluadores consideraron relevantes menos de cinco procesos para cada consulta.

En la figura 13 se muestran los valores obtenidos para la precisión top k. Como es evidente para todas las consultas, los niveles de esta medida aumentan cuando no se toma en consideración el orden de las clasificaciones, y esto sucede porque el algoritmo está recuperando, en muchos casos, más de la mitad de los procesos para cada intervalo, pero no en el mismo orden.

En los niveles de precisión, puede observarse que entre los veinte primeros resultados, la herramienta automática no alcanza más de dos resultados en el mismo orden y en todos los casos, excepto el de Q7, sólo uno de ellos hace parte del top 5. Pese a eso, cuando el orden no es importante, el algoritmo alcanza en ocasiones a coincidir hasta en la mitad de los resultados aunque no en la misma posición, es decir, que de los 20 primeros elementos recuperados por el algoritmo, 10 de ellos se encontraban dentro del top 20 de procesos relevantes; y como puede verse de Q2 a Q6, para el top 5 la herramienta alcanza a recuperar 3 de ellos.

Intermedio y Novato: Desde el punto de vista de estos dos tipos de usuario, el funcionamiento del algoritmo cambia drásticamente, ya que como se advierte en la figura 12 los valores para la exhaustividad aumentan con respecto a los jueces expertos, pero así mismo la precisión cae a niveles muy bajos. Lo anterior se atribuye a que estos tipos de comparadores (novatos e intermedios), sólo consideran relevante un conjunto muy reducido de procesos, y en casos como el de Q4, Q5 y Q6 la herramienta automática recupera todo el conjunto. No obstante, el número de elementos proveídos por el algoritmo es mucho mayor a lo que consideraron relevantes los intermedios y los novatos, y por lo tanto, para ellos la recuperación está llena de ruido y la herramienta no es efectiva.

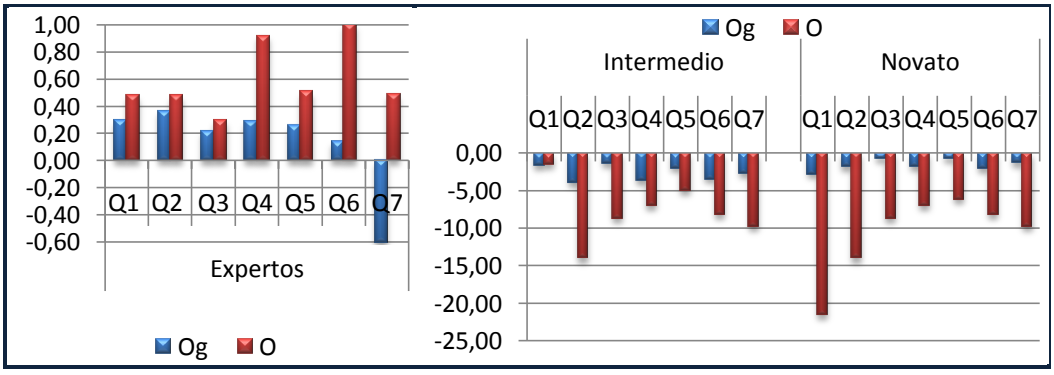

Figura 14 Resultados comparativos entre jueces para Overall – Análisis A

Finalmente, la figura 14 muestra los resultados que describen el esfuerzo que los potenciales usuarios del algoritmo tendrían que hacer para eliminar los elementos recuperados que no son relevantes, y encontrar los que hicieron falta. Esta medida, como se explicó en el capítulo 4, sólo tiene sentido cuando la precisión es al menos 0.5, y como para los intermedios, novatos y los expertos en la consulta Q7 este requisito no se cumple, se evidencia que todos los valores para este parámetro son negativos. Como se mencionó anteriormente, el problema en este caso no sería encontrar los elementos que hicieron falta, sino eliminar todo lo que se consideraría como irrelevante.

Para interpretar los valores de esta medida, es necesario saber que será máxima (1) cuando la precisión y la exhaustividad también lo sean, y por lo tanto el usuario de la herramienta automática no tendrá que manipular los resultados. Analizando los valores obtenidos, se ve que para la mayoría de las consultas el nivel no llega ni a 0.5 lo que quiere decir que el esfuerzo del usuario sería importante. Además, cuando se toman en cuenta las gradaciones, los resultados disminuyen haciendo que la efectividad del algoritmo no sea suficiente para satisfacer medianamente a ninguno de los tipos de jueces. Como se mencionó con anterioridad, las Q4 y Q6 tienen un tratamiento especial. Teniendo en cuenta los niveles de overall para estas consultas, se puede decir que si el número de resultados obtenidos hubiera sido el que se debía tener y la medida aún conservara su valor, entonces la herramienta hubiera tenido un buen desempeño para estas consultas, sin embargo, es sólo una hipótesis que podrán resolver en trabajos futuros.

Pasando a los análisis B y C (anexo C), se puede concluir que a pesar de variar la importancia de los criterios a tener en cuenta, la efectividad de la herramienta automática no mejora en ningún caso, por lo tanto, la medida de similitud básica es un resultado que denota una baja calidad del algoritmo en la recuperación.

- *Similitud de nodo:* Este valor es calculado a partir de la similitud básica y el número de nodos en el grafo resultante y hace referencia a la dependencia causal entre los procesos comparados. Al igual que la similitud básica, esta medida fue analizada de acuerdo con las distribuciones de los criterios mencionadas en la tabla 4 del anexo C, y son específicamente los análisis D, E, F y G.

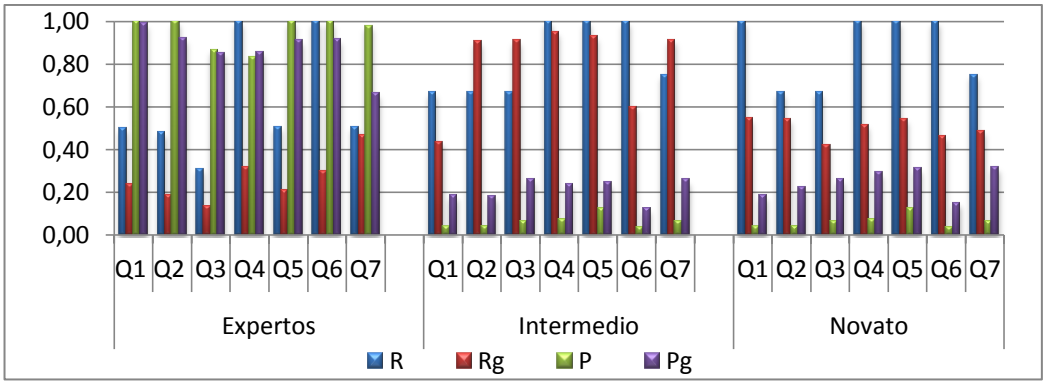

Figura 15 Resultados comparativos entre jueces para la precisión y la exhaustividad – Análisis G

La figura 15 muestra los resultados de la efectividad de la recuperación del algoritmo al evaluar sus valores de similitud de nodo, frente a los procesos relevantes considerando sólo el criterio de dependencia causal.

Experto: Al igual que en el análisis anterior la herramienta automática recupera al rededor de la mitad de los elementos relevantes para cada consulta, arrojando valores cercanos a 0.5 para la exhaustividad. En cuanto a la precisión se aprecian resultados altos, mayores de 0.8, para todas las consultas. Esto evidencia que si bien el algoritmo no recupera la totalidad de los elementos relevantes, al menos el conjunto de procesos ofrecidos como respuesta, es casi en su totalidad satisfactorio para el usuario. Adicionalmente, se observa que los valores para la precisión gradada se mantienen constantes con respecto al valor de la evaluación binaria. Esto se debe a que en este caso la herramienta es mucho más rigurosa, y califica la similitud más estrictamente que los jueces.

Si se comparan los resultados aquí descritos con los mostrados en el análisis anterior, podría decirse que la medida de similitud de nodo mejora en gran medida los valores para la precisión binaria y gradada, dando una noción de rigurosidad a la recuperación del algoritmo. Pese a lo anterior, se nota una disminución en los valores de la exhaustividad, lo que indica que el peligro que se corre al hacer más precisa y más rigurosa la herramienta automática, es que deje de recuperar elementos relevantes.

La figura 16 expone los resultados para la precisión top k para los usuarios expertos. Si se comparan estos resultados con los presentados anteriormente, se puede observar una mejora evidente en los resultados para este parámetro al considerar el orden. Mientras que el otro análisis mostraba máximo 2 elementos coincidentes por cada 20 recuperados, en este caso se ve que en casos como Q5 llega hasta 5 procesos correctamente

recuperados, y además que en consultas como Q4 llega a ordenar correctamente 3 de los 5 primeros. Igualmente, mejoran los valores para la precisión top k sin tener en cuenta el orden, donde en los resultados para Q6 se ve que 14 de los primeros 20 se recuperaron, y que 8 de los 10 primeros elementos recuperados se encontraban dentro del top 10 de los procesos relevantes.

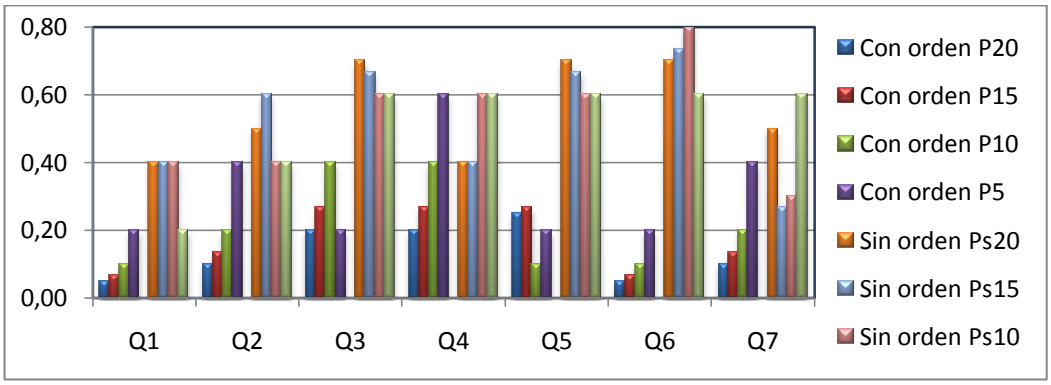

Figura 16 Resultado de la precisión top k para expertos – Análisis G

Si bien estos resultados no mejoraron para todas la consultas, es evidente que el valor de la similitud de nodo mejora un poco la calidad de la recuperación de la herramienta automática, incluso al analizarlo teniendo en cuenta todos los criterios (análisis D anexo C). No obstante, si se revisan los resultados del cotejo de esta medida frente a los valores de relevancia que se obtienen al analizar sólo los criterios concernientes con los nodos (análisis E anexo C), se observa que al disminuirse el número de procesos considerados relevantes, la precisión cae a niveles menores que 0.5 así se mantengan los valores de la exhaustividad, haciendo intuir que además de no recuperar un porcentaje importante del material esperado, el conjunto de procesos reportados está compuesto en gran medida por falsos positivos. Por todo lo anterior, podría concluirse que al encontrar un desempeño aceptable del algoritmo para los análisis D, F y G (anexo C), el análisis E no es un buen parámetro para analizar la herramienta automática, y por lo tanto los criterios allí tenidos en cuenta no se satisfacen con el valor de la similitud de nodo.

Intermedio: La figura 15 muestra que los resultados esperados por este tipo de usuario son satisfechos por la herramienta automática si no en su totalidad, en gran medida. Lo anterior puede deducirse al observar los valores casi perfectos para la exhaustividad, donde se nota que el algoritmo está recuperando el material considerado relevante por el nivel intermedio. Además, los valores para este parámetro teniendo en cuenta la gradación, mantienen la medida antes mencionada y en casos como la consulta Q7 la

superan. Esto quiere decir a pesar de que la herramienta automática se presentó rigurosa para los jueces expertos, para los intermedios no lo es tanto, produciendo la recuperación de muchos más procesos pero en los que se encuentran más de los deseados. Lo anterior puede verse con más claridad si se analizan los valores evidentemente bajos de las precisiones. Es decir, que al tener una exhaustividad alta y una precisión tan baja, el algoritmo está retornando al usuario un conjunto de resultados dentro de los que se encuentran los esperados, pero acompañados de una gran cantidad de falsos positivos.

Novato: Los resultados para este tipo de usuario son semejantes a los del nivel intermedio. Se recupera en gran medida el conjunto de procesos deseados, pero sumados a una gran cantidad de elementos irrelevantes. Pese a esto, vale la pena observar que a diferencia de los jueces del tipo anterior, la herramienta automática se presenta algo más rigurosa frente a los novatos, y puede verse en la notoria disminución de la exhaustividad frente a los valores teniendo en cuenta la gradación.

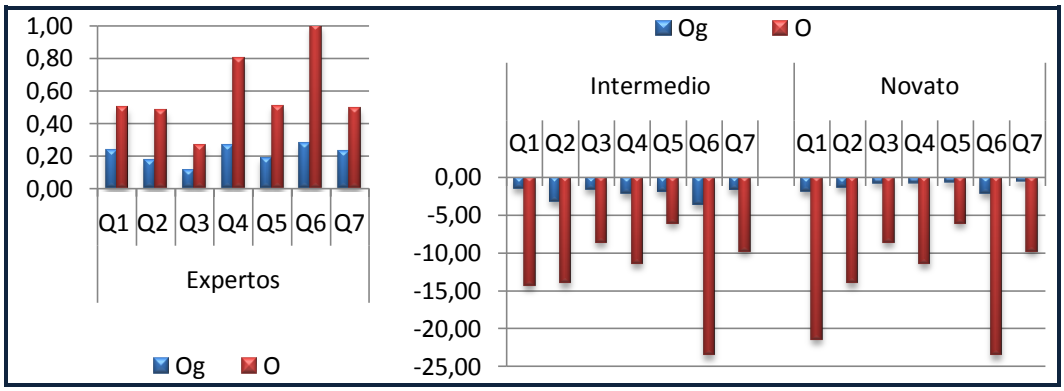

Figura 17 Resultados comparativos entre jueces para Overall – Análisis G

La figura 17 expresa los resultados para el overall. Como en el análisis para la similitud básica, los resultados para los tipos intermedio y novato continúan en niveles negativos, y la razón innegable siguen siendo los bajísimos niveles de precisión provocados por la cantidad de material irrelevante recuperado. Esto quiere decir, que si dentro de las primeras posiciones del ranking ofrecido por la herramienta automática no se encuentran los resultados certeros, el proceso para descartar elementos innecesarios va a ser extenuante, y por lo tanto el algoritmo será considerado para estas personas, como poco efectivo.

Por otra parte, los resultados para los expertos se mantienen en los niveles anteriores, donde sigue siendo evidente el hecho de que para este tipo de usuarios la herramienta automática no recupera la totalidad del material

esperado. Sin embargo, no se puede dejar de lado que los valores para la precisión subieron, la rigurosidad del algoritmo fue mucho mayor, y por consiguiente, la cantidad de elementos coincidentes en las clasificaciones mejoró. En conclusión el algoritmo presenta una efectividad aceptable para los usuarios de tipo experto, recupera los elementos esperados por los jueces intermedio y novato, pero la precisión para estos últimos continúa siendo muy baja.

- *Similitud de secuencia:* Esta medida retorna un valor como una función de la similitud básica y el número de secuencias encontradas para la relación de nodos de la medida anterior. Este parámetro podría considerarse como una evaluación del control de flujo de los procesos comparados, y por esta razón se coteja con el cálculo de los resultados manuales para este criterio (Análisis I anexo C).

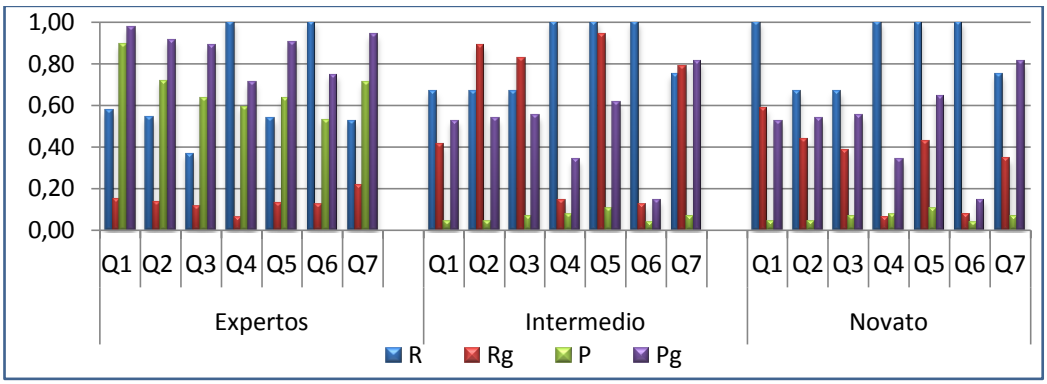

Figura 18 Comparativos entre jueces para la precisión y la exhaustividad – Análisis I

En la figura 18, se describen los resultados para las medidas de calidad de recuperación para los valores de la similitud de secuencia, frente a los calculados mediante el análisis I del anexo C para las evaluaciones manuales.

Experto: En la gráfica se observa que a excepción de Q3, el algoritmo recupera al menos la mitad de los elementos relevantes para cada consulta, es decir que la exhaustividad es en cada caso mayor a 0.5. Como sucedió anteriormente con Q3, la herramienta automática recupera un número de procesos mucho menor que para el resto de las consultas, y además, el número de elementos relevantes es pequeño respecto a lo esperado por este tipo de usuarios.

Por otra parte, la precisión aunque no es tan alta como en el análisis precedente exhibe resultados mayores a 0.5, y estos valores mejoran hasta

en un 0.3 al tener en cuenta la gradación. De acuerdo con esto, los elementos retornados por el algoritmo son en su mayoría relevantes, sin embargo, sólo alcanza a recuperar alrededor de la mitad del conjunto esperado por los expertos. Además, se puede decir que por el aumento percibido en los valores de la precisión cuando se toman en consideración los valores gradados, la herramienta es muy rigurosa siendo esta posiblemente la causa de los valores de la exhaustividad.

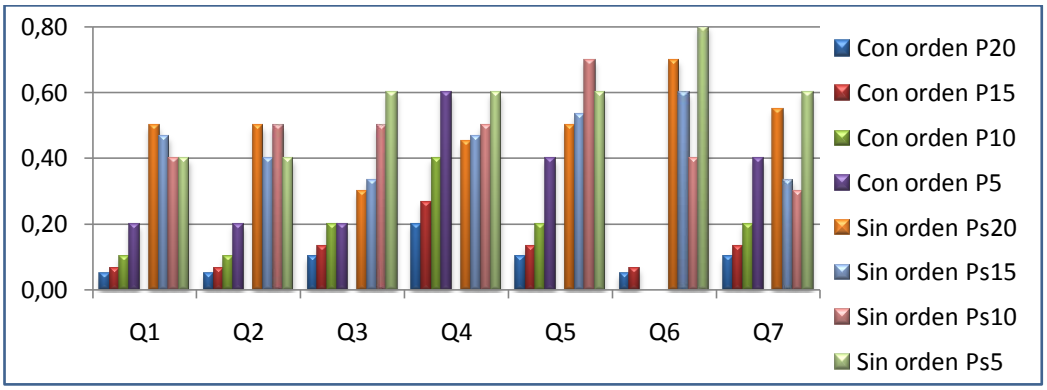

Figura 19 Resultado de la precisión top k para expertos – Análisis I

En la figura 19 se describe la capacidad del algoritmo para recuperar elementos relevantes en la misma posición dentro de unos intervalos. Como se aprecia en la gráfica para las consultas Q1, y Q2, la herramienta automática coincidió sólo en un elemento para todos los rangos estudiados, mientras que Q3, Q5 y Q7 presentaron dos procesos ordenados en la misma posición hasta el top 10, sin embargo para la consulta Q3 sólo uno de ellos entró en el top 5. Entretanto, la consulta Q4 ostenta el mayor número de elementos coincidentes al presentar 3 procesos recuperados en la misma colocación esperada por los expertos en el top 5. A pesar de lo anterior, el algoritmo sólo acertó en un proceso dentro del top 15 para la Q6.

Como sucedió en los análisis previos para la precisión top k, este valor mejora sustancialmente cuando no se tiene en cuenta el orden, y el ejemplo más diciente es el caso de Q6 que a pesar de sólo obtener un proceso coincidente en posición dentro del top 15, para los resultados sin ordenamiento presenta 14 elementos dentro del top 20, 9 en el top 15, y cuatro en los top 10 y 5. Lo anterior evidencia que aunque la herramienta no logra coincidir en el ordenamiento de los resultados, al menos en una gran proporción, los usuarios podrán encontrar lo que buscaban dentro de los rangos evaluados.

Intermedio y novato: Como en los análisis previos el comportamiento de la herramienta frente a las expectativas de estos usuarios tiene un comportamiento bastante regular. Aún para este caso, las exhaustividades continúan presentando resultados bastante aceptables, sin embargo, la precisión no mejora. En conclusión, los jueces de estos rangos buscan elementos con similitudes obvias, y por consiguiente no están interesados en los procesos que tengan semejanzas a niveles muy profundos, así que la única forma de que el algoritmo llene sus expectativas, es que dentro de los resultados presentados a este tipo de usuarios, se encuentren los que consideraron relevantes dentro de los primeros propuestos en la clasificación.

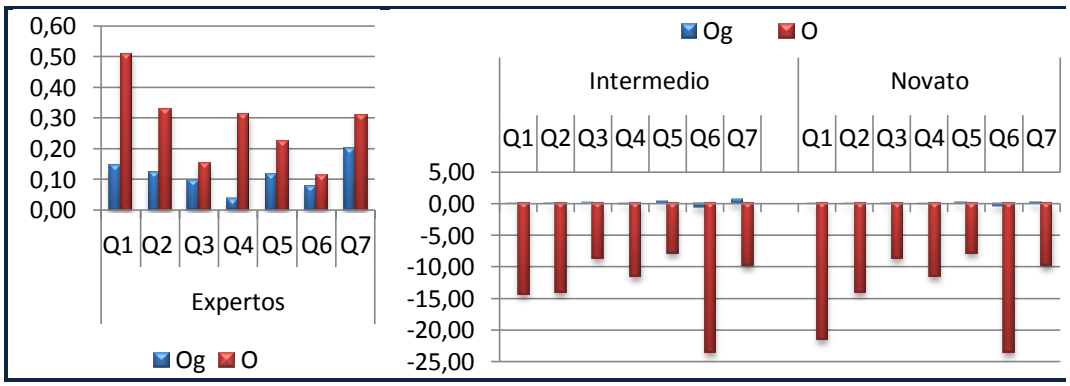

Figura 20 Resultados comparativos entre jueces para Overall – Análisis I

La figura 20 muestra la efectividad de la recuperación de la herramienta en términos de overall. Se puede notar en la gráfica que el resultado para esta medida es recurrente en todos los análisis. Los valores negativos de esta medida para los jueces intermedio y novato, son producto del número reducido de elementos relevantes esperados y la cantidad de procesos recuperados por la herramienta automática. Además, para los expertos, sigue haciéndose notoria la falencia del algoritmo para retornar más de la mitad de los procesos relevantes de la colección.

En términos generales, se puede concluir que:

- Los expertos tienen en cuenta muchos detalles para determinar la similitud entre dos procesos, por lo tanto el número de elementos considerados relevantes es mucho mayor a la cantidad recuperada por la herramienta automática.

- Los jueces intermedio y novato no explotaron al máximo el análisis que se esperaba de los criterios utilizados para efectuar las comparaciones entre procesos, y esto se hace evidente en el número reducido de elementos considerados relevantes por estos grupos.

- La herramienta automática no muestra una rigurosidad continua para todos los análisis, exhibida en los valores variables para la precisión y la exhaustividad al tomar en cuenta la gradación de los resultados.

- En los casos donde la herramienta se mostró más rigurosa (análisis D y H – Anexo C), se notó un crecimiento en los valores de la precisión para todos los tipos de usuario, llevando a la conclusión que la rigurosidad del algoritmo asegura que el número de elementos irrelevantes recuperados no sea significativo, pero disminuye su capacidad para obtener más elementos relevantes de la colección.

- Ninguno de los resultados provistos por el algoritmo proporciona un ordenamiento aceptable para los expertos, aún cuando generalmente el panorama mejora al no tener en cuenta la posición de los elementos recuperados, los usuarios de esta herramienta automática no podrán estar seguros de encontrar siquiera dentro de los primeros 5, el conjunto de elementos que llenen por completo sus expectativas.

- De los valores de similitud provistos por el algoritmo, podría concluirse por los resultados de los análisis A, D y H (anexo C), donde se analizan estas medidas con la misma distribución de los criterios, que la similitud básica es la que mejor se ajusta a los intereses de los usuarios, ya que a pesar de no exhibir niveles de precisión tan altos como la similitud de secuencia, es el que ostenta una mejor exhaustividad sin un detrimento importante de la precisión.

6.3 RELACIÓN COMPARATIVA ENTRE JUECES DE TIPO EXPERTO

Los participantes de la evaluación intuitiva de procesos fueron el eje fundamental del desarrollo de este proyecto, ya que gracias al desarrollo de sus comparaciones fue posible determinar la calidad de la recuperación de la herramienta automática. Esta es la razón para dedicar este apartado del capítulo, a la observación de los resultados provistos por los jueces expertos con el fin de analizar el comportamiento entre ellos y frente a la herramienta automática.

Esta sección se divide en tres partes, la primera evalúa la concordancia de los resultados de los expertos teniendo en cuenta la cantidad de procesos que consideraron relevantes para cada consulta; la segunda parte hace un recuento de la coincidencia existente en el ranking top 5 para cada evaluador; y finalmente se hará un análisis de rigurosidad entre los jueces.

- *Análisis de concordancia:*

  Siguiendo la línea del trabajo realizado por Tsetsos [TAH06], se hará un análisis de la concordancia existente entre los jueces de acuerdo al número de procesos considerados relevantes para cada consulta, y como componente adicional, se incluyó el número de elementos recuperados por la herramienta automática en cada caso.

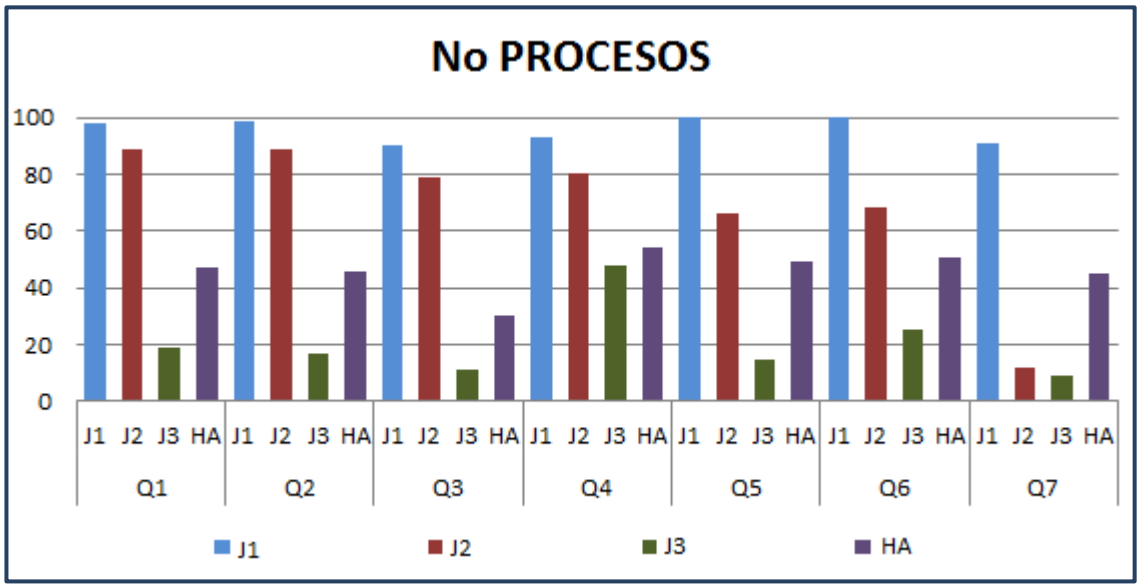

Figura 21 Análisis comparativo del número de procesos considerados relevantes por los jueces expertos y la herramienta automática

En la figura 21 se observa que existe un nivel de concordancia en el número de procesos considerados relevantes para los jueces J1 y J2, sin embargo, el juez J3 parece establecerse por la línea de los intermedios y novatos, siendo poco minucioso en la evaluación de similitud, pero recuperando un poco más de procesos que ellos. Además, se observa que el evaluador J1 es bastante detallista en las comparaciones y encuentra relación a cada consulta con casi todos los elementos de la colección. Este juez es por lo tanto quien mejor aplicó los criterios para la evaluación intuitiva, sin embargo, al establecer niveles de similitud para casi todos los procesos evaluados, ocasiona una precisión muy alta de la herramienta automática, pero a costa de niveles de exhaustividad bastante bajos, como efectivamente pudo observarse en el análisis del apartado anterior. Al hacer el conjunto de elementos relevantes cercano al número de procesos de la colección de prueba, ocasiona que todos los procesos recuperados por el algoritmo sean relevantes, pero a menos que el algoritmo sea muy poco riguroso, le será en extremo difícil alcanzar a recuperar la totalidad de elementos esperados.

En conclusión, la gráfica 21 muestra que los resultados utilizados para la evaluación de la efectividad de la herramienta automática, son proveídos en su gran mayoría por el juez J1 quien comparó a fondo la similitud entre los procesos y encontró niveles de relevancia para casi todos los elementos de la colección.

- *Análisis top 5*

Este análisis es similar a la precisión top k medida para evaluar la herramienta automática, pero en este caso tiene en cuenta el número de procesos coincidentes entre los 5 primeros retornados por cada juez. De esta manera se verá hasta qué punto los jueces fueron concordantes en este aspecto.

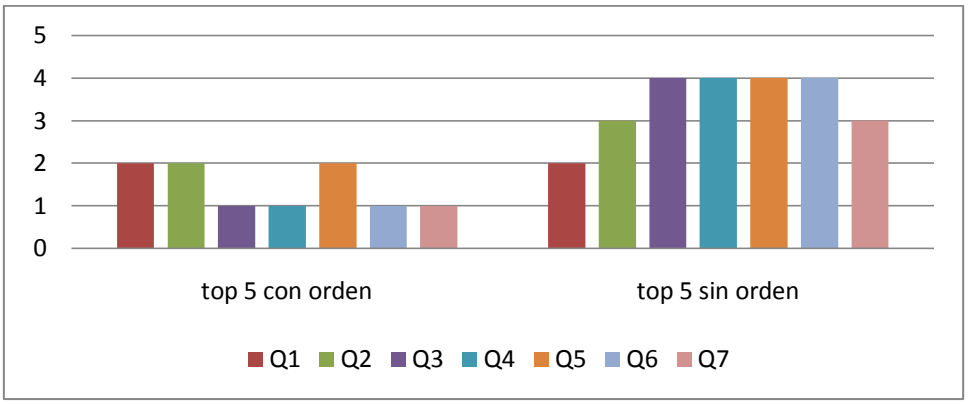

Figura 22 Número de procesos coincidentes entre los jueces para el top 5

La figura 22 expone el análisis explicado anteriormente, y en ella puede observarse que los jueces sólo coincidieron en máximo 2 procesos en el mismo nivel de relevancia para las consultas Q1, Q2 y Q5. A pesar de eso, la cantidad de elementos relevantes dentro de los primeros 5, aumenta cuando no se toma en cuenta el ordenamiento, ostentando hasta 4 procesos coincidentes para los 3 jueces en 4 de las 7 consultas.

En este apartado puede concluirse que la concordancia en los primeros elementos relevantes es alta si no se considera el ordenamiento, pero que las diferencias en los puntos de vista de los 3 jueces son notorias al observar que no hubo unanimidad en el posicionamiento de dichos procesos. Como recomendación para un trabajo futuro, sería evaluar la importancia y el desempeño de los criterios para la comparación manual de procesos, para replantearlos de una forma que ayude a disminuir las ambigüedades en estos resultados, y por lo tanto provean una jerarquización de procesos relevantes más confiable.

- *Análisis de rigurosidad entre jueces*

La rigurosidad en la apreciación de la similitud entre procesos, es un parámetro importante para determinar el número de elementos que conformarán el conjunto de componentes relevantes para una consulta dentro de una colección de prueba. Por lo anterior se decidió hacer este análisis y determinar los factores por los cuales cada juez consideró relevantes un número determinado de procesos.

A continuación se estudiará la rigurosidad para los tres jueces, y para los dos que obtuvieron el mayor número de elementos relevantes. Para esto se utilizó el mismo mecanismo empleado para calcular la precisión en la evaluación graduada, descrito en la siguiente ecuación.

$$Rigurosidad_{J_x} = \frac{\sum_{s=1}^{100} \min\{(r_{J1})_s, (r_{J2})_s, (r_{J3})_s\}}{\sum_{s=1}^{100} (J_x)_s}$$

En la ecuación, J_x es el juez (J1, J2 ó J3) a quien se evalúa la rigurosidad, r_j hace referencia al resultado de relevancia de cada evaluador en el proceso "*s*" evaluado, donde "*s*" representa cada uno de los procesos de la colección.

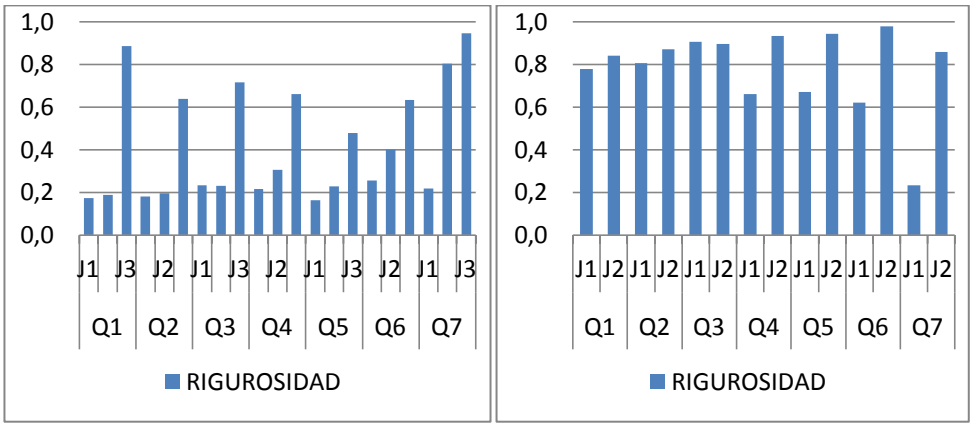

Figura 23 Análisis de rigurosidad de los jueces

Como se observa en la figura 23, al analizar la rigurosidad entre los tres jueces (figura de la izquierda), se deduce que el juez J3, para todas las consultas, es el más riguroso de los 3. Esta tendencia puede observarse además en el número reducido de elementos relevantes establecidos por este comparador. Contrario a lo anterior, el juez J1 es el menos riguroso, y como se mencionó en análisis precedentes, encontró semejanzas entre casi todos los procesos de la colección para cada consulta.

En la figura de la derecha, se describe la rigurosidad teniendo en cuenta sólo los jueces J1 y J2, para los cuales se percibió una concordancia en el número de elementos considerados relevantes. En este caso se ve un nivel de rigurosidad alto y parejo para ambos jueces, que se debe a la poca

diferencia existente en el número de procesos con algún nivel de similitud. No obstante, la figura muestra que el juez J2 presenta un aumento en el parámetro evaluado a medida que avanza en las consultas. Al estudiar los resultados arrojados por cada juez, se notó que el J2 iba disminuyendo el número de procesos considerados relevantes conforme avanzaba en el análisis de cada consulta. Esto puede deberse a que la persona que realizó estas evaluaciones fue obviando detalles mientras hacía las comparaciones, tal vez producto de la fatiga, por falta de concentración, quizás el afán de terminar, ó por la monotonía que produce hacer procesos repetitivos.

Concluyendo esta sección, se puede afirmar que:

- El nivel de concordancia para los jueces J1 y J2 es relativamente parejo, aunque fue disminuyendo a medida que avanzaban las consultas. No obstante, hubieran sido un muy buen patrón para evaluar la herramienta automática.

- El juez J3 no explotó al máximo los criterios para la evaluación intuitiva de procesos y por esta razón, el número de elementos relevantes para este comparador fueron mucho menores que los proveídos por J1 y J2. Sin embargo, al observar la cantidad de procesos recuperados por la herramienta automática, se podría deducir que la exhaustividad y la precisión en la recuperación del algoritmo para este juez, deben ser en gran medida mucho más estables.

- De lo anterior, se puede deducir además, que el algoritmo tienen más posibilidades de satisfacer a un usuario que no requiera niveles de similitud muy profundos, o que busque funcionalidades excesivamente específicas en los procesos.

- Los resultados proveídos por los jueces J1 y J2 son muy útiles si se desea calibrar una herramienta automática ya que ofrece un nivel de detalle profundo acerca de la similitud de los procesos.

RESUMEN: En este capítulo se hace un recuento acerca de la metodología utilizada en la evaluación de una herramienta automática para el descubrimiento de procesos de negocio, mediante la aplicación y análisis los resultados de las medidas de calidad de recuperación. Aquí se explica desde la forma como se obtuvieron los resultados del algoritmo [FIC10] y de las evaluaciones manuales, su análisis y las conclusiones que describen la efectividad de la herramienta automática. Adicionalmente se analizó el

nivel de concordancia existente entre los resultados de las evaluaciones manuales provistas por los jueces de tipo experto.

Este capítulo puede considerarse como un modelo de evaluación de sistemas de recuperación de procesos de negocio, ya que describe todas las etapas y análisis realizados para determinar la efectividad de la recuperación de un algoritmo de esta naturaleza.

CAPÍTULO 7

CONCLUSIONES Y TRABAJOS FUTUROS

En este trabajo se propuso crear una plataforma para la comparación manual de procesos negocio, que proporcionara el conjunto de elementos relevantes presente en una colección de prueba, para permitir la evaluación de la efectividad de la recuperación de una herramienta automática de descubrimiento de procesos de negocio.

Durante el desarrollo de esta propuesta, se generó una base documental acerca de la evolución de las metodologías existentes para la evaluación de sistemas de recuperación de información y las herramientas para el descubrimiento de servicios web, para finalmente generar el modelo aplicado en la valoración del algoritmo implementado por Figueroa [FIC10].

Adicionalmente, se generó una colección de prueba compuesta por procesos de negocio del dominio del geoprocesamiento y las telecomunicaciones, para aplicar sobre ella el algoritmo y las comparaciones manuales, que sirvieron de sustento para la evaluación de la herramienta automática.

Finalmente, se emitió un concepto acerca de la efectividad del algoritmo evaluado sustentado en el análisis de los juicios de relevancia provistos por las comparaciones manuales, y los resultados de la aplicación de la herramienta automática sobre la colección de prueba descrita anteriormente.

A continuación, se presentan las conclusiones del trabajo realizado, algunas recomendaciones y la descripción de posibles trabajos futuros.

7.1 CONCLUSIONES

Durante la aplicación de la herramienta desarrollada para la comparación manual de procesos, se realizó la investigación de los modelos empleados para la evaluación del desempeño de los sistemas de recuperación que sirvieran como ejemplo para la ejecución de este trabajo. De esta recopilación de información puede concluirse que:

- Existen dos ejes fundamentales sobre los cuales se evalúa la calidad de un algoritmo de recuperación. Estos son el desempeño en ejecución, y la relevancia de los elementos recuperados de acuerdo con la utilidad percibida por el usuario, siendo este el punto más importante a considerar para analizar la efectividad de una herramienta de este tipo.

- Además, se estableció que los parámetros que mejor evalúan la efectividad son las medidas basadas en la relevancia de los elementos recuperados, siendo este último un concepto emitido por los usuarios de una herramienta de acuerdo con la utilidad de las respuestas y su relación con la consulta planteada.

- Las medidas más empleadas para valorar la efectividad de una herramienta automática son la *precisión* y la *exhaustividad*, que evalúan la calidad de la recuperación en términos del material retornado por el algoritmo, y los elementos considerados relevantes en una colección por un dominio experto.

- La recopilación manual de los elementos relevantes para una consulta dentro de una colección de prueba puede realizarse asignando valores binarios de similitud, o mediante la aplicación de una escala que describa diversos niveles de semejanza entre dos elementos. Dependiendo de la técnica utilizada, existen reinterpretaciones de las ecuaciones para las medidas de calidad de recuperación que permiten la emisión de conceptos de efectividad. No obstante, se observó en los trabajos relacionados, que el uso de escalas de relevancia proporcionan resultados más certeros de las comparaciones manuales de servicios web y procesos de negocio.

- La generación de conjuntos de procesos de negocio de prueba, es un trabajo dispendioso y demorado debido a la cantidad de información que cada proceso debe contener para ser útil en la comparación intuitiva y en la recuperación automática. Por ejemplo, durante la investigación se encontró, que la inclusión de descripciones en lenguaje natural acerca del funcionamiento de los procesos, ayuda a las personas en el entendimiento de estos, viéndose su utilidad reflejada en los juicios de relevancia. Adicionalmente, en los últimos años se ha incrementado el interés por el enriquecimiento ontológico de los procesos, para facilitar su publicación y posterior localización. Por lo anterior puede evidenciarse que la creación de una colección de procesos con estas características requiere de mucho trabajo y dedicación.

Por otra parte, durante la ejecución del modelo de evaluación utilizado para la valoración de la herramienta automática se encontraron varios puntos a destacar.

- La realización de las comparaciones manuales por parte de los usuarios arrojaron resultados con poca concordancia entre los diferentes tipos de comparadores, e incluso dentro de un mismo grupo. Esto se presenta porque el nivel de rigurosidad de cada persona al realizar las comparaciones es diferente. Adicionalmente, en los resultados provistos

por un juez, se evidenció el decrecimiento en el número de elementos relevantes conforme avanzaba en el análisis de las consultas. Lo anterior puede deberse a que la fatiga producida por las evaluaciones tan largas, causa desconcentración y provoca respuestas aleatorias. Se recomienda para posteriores aplicaciones de la herramienta sintetizar la cantidad de criterios a evaluar, procurando no bajar la calidad de las comparaciones, pero buscando una agilidad en el proceso.

- La efectividad de la herramienta para los usuarios intermedio y experto, no son muy buenas, pero en gran parte esto se debe a la pobreza de los juicios de relevancia emitidos por este tipo de jueces. Por esta razón, se recomienda enfocar los análisis a un dominio experto, y procurar la evolución de los algoritmos hacia la satisfacción de este tipo de usuarios.

7.2 RECOMENDACIONES Y TRABAJOS FUTUROS

La gran amplitud del campo de la evaluación de sistemas de recuperación de procesos de negocio y su constante evolución, abren el camino para una amplia variedad de trabajos afines a este, ó continuaciones y mejoras que provean mecanismos más eficientes para este proceso.

Para trabajos futuros, podrían emplearse los resultados obtenidos en los juicios de relevancia, para evaluar una evolución de la herramienta automática de Figueroa [FIC10] donde se utilice enriquecimiento ontológico para la comparación de procesos, y observar si la efectividad de esta mejora.

Por otra parte, al analizar los resultados de los comparadores expertos, se observó que tienen en cuenta muchos detalles al emitir juicios de similitud, y por lo tanto al obtener el conjunto de los elementos relevantes para una consulta, la cantidad de material relevante se acerca mucho al número de procesos de la colección de prueba. Para sortear este inconveniente, podría incluirse un criterio de evaluación adicional, que permita al usuario determinar si un elemento es relevante para una consulta, tomando en cuenta los niveles de semejanza encontrados durante la valoración de los demás criterios. De esta manera, se obtendrían unos valores de similitud útiles para la calibración de los algoritmos de descubrimiento, y adicionalmente, para analizar su efectividad, sólo se tendrían en cuenta los niveles de semejanza de aquellos procesos considerados relevantes.

Adicionalmente, podría desarrollarse un módulo de evaluación de efectividad que reciba directamente los resultados de una herramienta automáticas, y los compare con los valores de relevancia obtenidos, aplique las ecuaciones para medir la calidad de la recuperación, y despliegue las gráficas que describan esta medición.

Este nuevo módulo ahorraría tiempo en análisis de datos y facilitaría la valoración de la efectividad de un algoritmo.

Finalmente, es importante reevaluar la forma como se realizan las comparaciones manuales, ya que debido a la falta de concordancia entre los resultados de los comparadores, se sugiere sintetizar los criterios utilizados aquí, en unos que mantengan los aspectos evaluados, pero que acorten el tiempo de comparación.

8. REFERENCIAS

[BKB05] A. Bernstein, E. Kaufmann, C. Burki, y M. Klein: *How similar is it? Towards personalized similarity measures in ontologies*. In 7. Tag. Wirt. Informatik, 2005.

[BLA90] Blair, D.C. Language and representation in information retrieval. *Elsevier Science Publish-ers*, Amsterdam, 1990.

[BOP93] G. Bordogna, G. Pasi. "A fuzzy linguistic approach generalizing Boolean information retrieval: A model and its evaluation", *Journal of the American Society for Information Science*, 44, 70-82, 1993

[BOR00] P. Borlund, "Information retrieval, experimental models and statistical analysis". *Journal of Documentation*, vol 56, nº 1, pp. 71-90, January 2000.

[BPL05] J. de Bruijn, A. Polleres, R. Lara and D. Fensel. *OWL DL vs. OWL Flight: Conceptual Modeling and Reasoning for the Semantic Web*. Fourteenth International World Wide Web Conference (WWW2005), 2005.

[BUH01] A. Budanitsky, G. Hirst: *Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures*. Second meeting of the North American Chapter of the Association for Computational Linguistics *(NAACL2001)*, Pittsburgh, PA, 2001.

[BUK81] D. A. Buell, D. H. Kraft. "Performance measurement in a fuzzy retrieval environment". 4th Annual international ACM SIGIR Conference on information Storage and Retrieval, Oakland, California, May, 1981.

[BYF92] R. Baeza-Yates, y W.B. Frakes, "Information retrieval: data structures & algorithms". Englewood Cliffs, New Yersey, Prentice Hall, 1992.

[BYF99] R. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval*, Addison-Wesley, 1999.

[CAN88] G. Canavos. *Probabilidad y estadística – Aplicaciones y Métodos*. 1988.

[CHO99] G. G. Chowdhury. Introduction to modern information retrieval. London: Library Asso-ciation, 1999.

[CLE72] C.W. Cleverdon, *On the inverse relationship of recall and precision*, Journal of Documentation 28 (1972), 195-201.

[COO73] W.S. Cooper. 'On selecting a Measure of Retrieval Effectiveness'. Journal of the American Society for Information Science, v. 24, March-April 1973. p.87-92

[COR08] C. Corrales, “Behavioral matchmaking for service retrieval”, Tesis Doctoral presentada a la *Unicersidad de Versailles Saint-Quentin-en-Yvelines*, Enero 2008.

[DDG09] R. Dijkman, M. Dumas, and L. García-Bañuelos: *Graph Matching algorithms for Business Process Model Similarity Search.* 7th Int. Conference on *Business Process* Management (BPM'09). Ulm, Germany, 2009.

[FIC10] C. Figueroa Martínez, J. Corrales. *Descubrimiento Automático de Procesos de Negocio Basado en Semántica del Comportamiento.* Febrero 2010.

[FOS72] D.J. Foskett. ‘A Note on the Concept of Relevance’. Information Storage and Retrieval, 8 (2):77-78, April 1972

[FRA97] V.I. Frants, et al. *Automated information retrieval : theory and methods*. San Diego [etc.] : Academic Press, cop.1997. XIV, 365 p.

[GLG06] A. Goderis, P. Li, C. Goble: *Workflow discovery: the problem, a case study from e-Science and a graph-based solution.* IEEE International Conference on Web Service, 2006.

[GOF64] W. Goffman. 'On relevance as a measure', *Information Storage and Retrieval,* 2, 201-203 (1964).

[GRE00] H. Greisdorf, “Relevance: An interdisciplinary and Informacion Science perspective”. *Informing Science: Special Issue on Information Science Research*, Vol 3 No 2, 2000.

[INW95] P. Ingwersen and P. Willet, *An introduction to algorithmic and cognitive approaches for information retrieval*, Libri 45 (1995), no. 3-4, 169-177.

[JAK02] K. Järvelin y J. Kekäläinen: Cumulated gain-based evaluation of ir techniques. ACM Transactions on Information Systems 20(4) (2002) 422-446

[KFK05] M. Klusch, B. Fries, M. Khalid, K. Sycara. OWLS-MX: Hybrid Semantic Web Service Retrieval. In *1st Intl. AAAI Fall Symposium on Agents and the Semantic Web*, AAAI Press, Arlington VA, 2005.

[KIS05] K. Kishida: Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments. Technical Report NII-2005-014E, National Institute of Informatics, Tokyo, Japan (2005)

[KKO08] U. Küster and B. König-Ries, “Evaluating semantic web service matchmaking effectiveness based on graded relevance,” in *Proceedings of the 2nd International Workshop SMR*2 *on Service Matchmaking and Resource Retrieval in the Semantic Web at the 7th International Semantic Web Conference (ISWC08)*, Karlsruhe, Germany, October 2008.

[KLZ08] M. Klusch and X. Zhing, “Deployed semantic services for the common user of the web: A reality check,” in *Proceedings of the 2nd IEEE International Conference on Semantic Computing (ICSC2008)*, Santa Clara, CA, USA, August 2008.

[KOR97] R. Korfhage, *Information storage and retrieval*, Wiley, New York, 1997.

[KUK08] U. Küster, B. König-Ries. Towards Standard Test Collections for the Empirical Evaluation of Semantic Web Service Approaches. *International Journal of Semantic Computing 2(3)*. 2008.

[KUK09] U. Küster, B. König-Ries. Relevance Judgments for Web Services Retrieval – A Methodology and Test Collection for SWS Discovery Evaluation. *Proceedings of the 7th IEEE European Conference on Web Services*, 2009.

[LAN68] F.W. Lancaster. *Evaluation of the MEDLARS Demand Search Service*. Library of Medicine, Betsheda, Md, 1968.

[LAN93] Lancaster, F. W. and Warner, A.J. *Information Retrieval Today*. Arlington, Virginia : Information Resources, 1993.

[LLR05] R. Lara, A. Polleres, H. Lausen, D. Roman, J. de Bruijn, and D. Fensel. *A Conceptual Comparison between WSMO and OWL-S*. WSMO Deliverable D4.1v0.1, 2005. http://www.wsmo.org/2004/d4/d4.1/v0.1/

[LIH04] L. Li, I. Horrocks, “A Software Framework for Matchmaking Based on Semantic Web Technology”,*International Journal of Electronic Commerce*, 6(4), 39- 60, 2004.

[LOP06] A. López-Herrera: *Modelos de Sistemas de Recuperación de Información Documental Basados en Información Lingüística Difusa*. Memoria de Tesis al grado de Doctor en Informática. Granada, 2006.

[MAR02] F.J. Martínez-Méndez. *Propuesta para el desarrollo de un modelo para la evaluación de la recuperación de información en Internet*. Tesis Doctoral, Facultad de ciencias de la documentación, Universidad de Murcia, 2002.

[MCH03] T. W. Malone, K. Crowston, y G. A. Herman, “Organizing Business Knowledge: The MIT Process Handbook”, Cambridge, MA: MIT Press. 2003.

[MCL99] T. W. Malone, K. Crowston, J. Lee, B. Pentland, C. Dellarocaset et al., “Tools for inventing organizations: Toward a handbook of organizational processes”, *Management Science 45*, 1999: pp. 425-443.

[MDG06] V.D. Mea, G. Demartini , L.D. Gaspero, S. Mizzaro: Measuring retrieval effectiveness with average distance measure (ADM). Information Wissenschaft und Praxis 57(8) (2006) 405–416.

[MEA93] C. T.Meadow. *Text Information retrieval Systems*. San Diego: Academic Press, 1993.

[MMR04] F. J. Martínez Méndez, y J. V. Rodríguez Muñoz, "Reflexiones sobre la evaluación de los sistemas de recuperación de información: Necesidad, utilidad y viabilidad". *Anales de Documentación* 7:pp. 153-170, 2004.

[MRS09] C. Manning, P. Raghavan y H. Schütze. *Introduction to* Information Retrieval. Cambridge University Press. 2008

[OLV99] M.D. Olvera, *Evaluación de la recuperación de información en internet: Un modelo experimental*, Tesis doctoral, Universidad de Granada, 1999.

[POR00] Pors, N. O. 'Information retrieval, experimental models and statistical analysis'. *Journal of Documentation*, vol 56, nº 1 January 2000. p. 55-70, 2000.

[RIJ99] C.J. Rijsbergen, "Information Retrieval", Glasgow, [En línea] Glasgow, University, 1999, http://www.dcs.gla.ac.uk/~iain/keith.

[SAK04] T. Sakai: New performance metrics based on multigrade relevance: Their application to question answering. In: Fourth NTCIR Workshop on Research in Information Access Technologies, Information Retrieval, Question Answering and Summarization (NTCIR04), Tokyo, Japan (2004).

[SAL83] G. Salton and M.J. Mc Gill. *Introduction to Modern Information Retrieval*. New York: Mc Graw-Hill Computer Series, 1983.

[SAR97] T. Saracevic, "Relevance: A review of and a framework for the thinking on the notion in information science". *Readings in Information Science*, San Francisco, Morgan Kaufmann Publisher, 1997.

[SWE63] J.A. Swets. 'Information retrieval systems', *Science,* 141, 245-250 (1963).

[TAH06] V. Tsetsos, C. Anagnostopoulos, S. Hadjiefthymiades. On the Evaluation of Semantic Web Service Matchmaking Systems. *Fourth IEEE European Conference on Web Services (ECOWS'06)*. 2006.

[WOM06] A. Wombacher: *Evaluation of Technical Measures for Workflow Similarity Based on a Pilot Study*. Springer-Verlag Berlin Heidelberg, 2006.

[WOR06] A. Wombacher, M. Rozie: *Piloting an Empirical Study on Measures for Workflow Similarity*. IEEE International Conference on Services Computing, 2006.