Personalización y Evaluación XML mediante la Simulación de Perfiles de Usuario y Juicios de Relevancia

Luis M. de Campos, Juan M. Fernández-Luna, Juan F. Huete, and Eduardo Vicente-López

Departamento de Ciencias de la Computación e Inteligencia Artificial, E.T.S.I. Informática y de Telecomunicación, CITIC-UGR, Universidad de Granada, 18071-Granada
{lci,jmfluna,jhg,evicente}@decsai.ugr.es

Resumen En este trabajo se presenta una técnica de personalización mediante expansión de consultas, aplicada a un Sistema de Recuperación de Información (SRI) XML. Esta expansión con nuevos términos hace que se distorsione la consulta original del usuario, que es la que representa sus necesidades de información, obteniéndose resultados no deseados. Se proponen dos soluciones que evitan este problema. Normalmente, para evaluar la ganancia de rendimiento del SRI tras aplicarle una técnica de personalización se realiza un estudio de usuario, pero éstos son muy costosos tanto en tiempo como en recursos. Además, en dicho estudio hay que elegir una combinación determinada de los parámetros de configuración, para la técnica de personalización y el modelo de Recuperación de Información (RI). Para facilitar la evaluación de los SRI personalizados se propone una metodología de evaluación, que agiliza el proceso mediante la simulación de perfiles de usuario y juicios de relevancia para un conjunto de consultas. Mediante esta metodología se han optimizado los parámetros de configuración para un estudio de usuario futuro, que pretendemos valide que la metodología propuesta es una alternativa fiable a estos estudios, especialmente indicada en las etapas iniciales de desarrollo del SRI personalizado, o en los casos en los que un estudio de usuario no sea factible.

Keywords: Recuperación de Información, Personalización, Evaluación, XML.

1. Introducción

Es un hecho que la cantidad de información digital se está incrementando de forma exponencial en los últimos años [7], por lo que el acceso a la información relevante se hace más difícil cada día. Los investigadores tratan de mitigar este problema con el desarrollo de SRI, y especialmente con la aplicación de técnicas de personalización [17], que proporcionan resultados más cercanos al usuario mediante el uso de su perfil, que representa sus intereses y preferencias.

Este artículo se centra en los SRI XML [6,11], y entre las posibles técnicas de personalización [15], en la expansión de consultas [5]. La principal ventaja de los SRI XML es que aprovechan la estructura interna del documento, permitiéndoles devolver una parte específica del mismo ó Unidad Estructural (UE), p.e. un párrafo, que incluso el usuario puede establecer como preferencia de recuperación, y no el documento completo como hacen los sistemas tradicionales. De esta forma, el esfuerzo requerido por el usuario para cubrir sus necesidades de información disminuye considerablemente.

A su vez, el uso de la RI XML y de la expansión de consultas presentan algunas dificultades adicionales, entre las que se encuentra determinar el número y valor de ponderación adecuado de los términos de expansión. Esta elección determinará la influencia de estos términos sobre los resultados obtenidos. Por ejemplo una consulta con muchos términos de expansión puede hacer que las UE recuperadas sean de mayor tamaño, puesto que deben contener el mayor número de términos posible de la consulta, ó hacer que aparezcan UE que sólo hablen de los términos expandidos y no de la consulta original del usuario. Por otro lado, la evaluación de documentos XML presenta dos dificultades principales, debido a la dependencia entre las UE del documento: (1) los fallos cercanos en la recuperación, que son UE que están estructuralmente relacionadas con UE realmente relevantes, como un párrafo vecino o una sección contenedora; y (2) los solapamientos, que se dan cuando el mismo texto está contenido en varias UE, como cuando se recuperan tanto un párrafo como su sección contenedora. En la sección 3 se verá cómo evitar estos problemas.

Los anteriores problemas hacen que configurar los parámetros, del SRI XML y de la técnica de personalización usados, sea muy complejo, ya que existe un gran número de posibles configuraciones. El método más usado para evaluar un sistema personalizado es el estudio de usuario [13,12], pero su realización requiere mucho tiempo y recursos. Si ya realizar uno es muy costoso, es totalmente imposible realizar un estudio con cada configuración de parámetros posible. Por ello, hemos desarrollado una metodología automática de evaluación de SRI personalizados, con la ventaja de no necesitar ningún dato previo de los usuarios ni de sus juicios de relevancia (ambos simulados), y es aplicable a documentos planos y estructurados. Esta metodología mide la ganancia de rendimiento de un SRI, tras aplicarle cualquier técnica de personalización en la que el perfil de usuario se corresponda con un conjunto de términos ponderados. Debe quedar claro que esta metodología no pretende sustituir a los estudios de usuario, ya que éstos son la forma más adecuada de evaluar un SRI personalizado. Lo que pretende es ser una alternativa en la evaluación de dichos sistemas, sobre todo en sus etapas iniciales de desarrollo, o cuando su realización no sea factible debido a cualquier circunstancia, como la falta de tiempo o recursos [13,16].

Un trabajo futuro es hacer un estudio de usuario para validar la metodología propuesta. El tener una metodología automática de evaluación permite ejecutar gran cantidad de tests de evaluación, con diferentes conjuntos de consultas y perfiles de usuario. Por ello, hemos usado la metodología para realizar múltiples simulaciones y decidir los valores apropiados de configuración del SRI, para

este futuro estudio de usuario. También permite el rápido desarrollo de nuevas técnicas de personalización, posibilitando la comparación de sus resultados con los que se tuvieran previamente de otras técnicas.

La sección 2 presenta una breve descripción del estado del arte de la personalización de SRI y su evaluación. La sección 3 describe la forma de personalización y evaluación XML usada en este artículo, así como las soluciones propuestas a los problemas expuestos en esta primera sección. La sección 4 explica los componentes, características necesarias y forma de uso de la metodología automática de evaluación propuesta. La sección 5 muestra los resultados obtenidos mediante la aplicación de dicha metología de evaluación. Y por último, en la sección 6 se exponen las conclusiones generales del artículo y los posibles trabajos futuros.

2. Fundamentos de Personalización y Evaluación en SRI

La RI personalizada, considerada como un subcampo de la RI contextual [14], se define en [1] como *"la combinación de tecnologías de búsqueda y conocimiento sobre la consulta y el contexto del usuario, para proporcionar las mejores respuestas a las necesidades de información del usuario"*.

La RI personalizada trata de optimizar la exactitud de la recuperación mediante las dos siguientes etapas [14]:

- *Modelado del contexto del usuario*: tiene dos principales componentes: (1) las técnicas de adquisición de información para construir el perfil de usuario y (2) el modelado y actualización de dicho perfil. En el primer componente se distinguen las técnicas explícitas, donde el usuario rellena su propio perfil, o las técnicas implícitas, donde a partir de diferentes fuentes de información sobre el usuario, como su historial de clicks, tiempo de lectura de cada documento, sus marcadores del navegador, etc, se infiere su perfil. En el segundo componente, el perfil se suele representar como un vector de palabras clave ó como estructuras semánticas de conceptos con la ayuda de ontologías. Este perfil se actualiza mediante la continua extracción de información del usuario a través de las técnicas vistas en el punto uno, y que algoritmos de minería de datos usan para aprender y actualizar el perfil.
- *Recuperación personalizada*: usando el perfil de usuario y la consulta, se intenta dar al usuario la información más acorde para satisfacer sus necesidades de información. Existen varias técnicas como el refinamiento de la consulta mediante relevance feedback o la desambiguación de la consulta. En todos los casos se integra la evidencia extraida del perfil de usuario mediante reformulación de la consulta, re-ranking de los resultados o en el propio modelo de recuperación, siendo las dos primeras técnicas las más utilizadas.

El principal objetivo de la evaluación de los SRI es estimar su rendimiento global. El marco de evaluación de la RI consiste en una colección de documentos de test, un conjunto de consultas y sus juicios de relevancia. La evaluación tradicional se basa en el modelo de laboratorio iniciado por Cleverdon [13] en el proyecto Cranfield II, que es muy usado en concursos como TREC, INEX

o CLEF. En estos concursos, los usuarios reales y los contextos de búsqueda están generalmente fuera del objetivo de evaluación. Por ello, es necesaria una aproximación alternativa de evaluación en RI personalizada o contextual, cuyos principales foros son: IRIX e IIiX. Las mayores contribuciones de estos congresos son el diseño de nuevas estrategias y métricas de evaluación, adaptadas a la RI personalizada. Hay 3 principales métodos de evaluación en RI personalizada o contextual [14]:

- *Extensiones al modelo de laboratorio*: modelan una mínima interacción entre el usuario y el sistema, incluyendo algunos factores contextuales. Su principal objetivo es tener un marco de evaluación más realista, a través de un entorno de evaluación relativamente controlado. Algunos ejemplos son 'TREC Interactive Track' o 'HARD Track'. Aún usando todas estas características del usuario, la evaluación sigue estando controlada por el sistema, y las principales métricas de evaluación están basadas en la exhaustividad y la precisión.
- *Simulaciones contextuales*: simulan usuarios e interacciones con el sistema por medio de escenarios de recuperación bien definidos (hipótesis). Tienen la ventaja de consumir menos tiempo y coste que los métodos con usuarios reales. Se usan mucho para la mejora de las interfaces de búsqueda de los SRI.
- *Estudios de usuario*: su principal objetivo es medir la efectividad del sistema en una situación de recuperación real, y tener un buen balance entre control y realismo. Si se usan distintos usuarios para distintas tareas de búsqueda, los experimentos no son repetibles, dadas las diferencias individuales entre dichos usuarios, y es difícil determinar qué factores influyen en la efectividad de la recuperación.

Como se observa existen muchas metodologías de evaluación en RI personalizada o contextual, pero aún no hay acuerdo en la definición de un marco de evaluación estándar ni en las métricas a usar, debido a que todas estas metodologías tienen limitaciones. Por esta razón, además de las argumentadas en la sección 1, hemos decidido desarrollar una nueva metodología automática de evaluación de SRI personalizados (véase la sección 4).

3. Personalización y Evaluación bajo un SRI XML

Uno de los aspectos fundamentales en personalización es el perfil de usuario. Este trabajo propone que este perfil sea generado automáticamente, basándose en el contenido de los documentos del área de interés de la colección documental en la que el usuario está interesado. Como modelo de representación se ha usado un vector de palabras clave ponderadas. El perfil está formado por aquellos términos en las primeras n posiciones, ordenados por *tf*idf* y ponderados por *idf*. Se ha escogido idf como valor de ponderación, porque cada término está mejor representado por este valor que por el valor tf*idf, considerando la colección documental completa. Un ejemplo de los primeros términos de un perfil

de usuario, aprendido a partir de un área de interés de educación, sería el siguiente: *1.45663*educ, 1.36756*centr, 1.96815*alumn, 2.06333*profesor, 2.11426*enseñ, 2.07116*curs, etc.* Como se observa, el perfil de usuario está formado por raíces de términos ponderadas por su idf, que representa el grado de influencia de cada término en la recuperación de las UE. En este contexto, los valores tf*idf se calculan como sigue:

$$tf(t, X) = \frac{f(t, X)}{terms(X)}, \tag{1}$$

donde *tf (frecuencia del término)* del término t para el área de interés X es el número de apariciones del término t, denotado por *f(t,X)*, en cada documento perteneciente al área de interés X. Y *terms(X)* es el número total de términos que contienen los documentos del área de interés X. Por otro lado,

$$idf(t) = \log \frac{N}{N(t)}, \tag{2}$$

donde el valor *idf (frecuencia documental inversa)* del término t es el resultado de dividir el número total de UE básicas N, de todos los documentos que componen todas las áreas de interés, por el número de UE básicas de todas las áreas de interés que contienen el término t, denotado por *N(t)*. Una UE es básica cuando no está contenida en ninguna otra UE (p.e. párrafos).

En cuanto a la recuperación personalizada hemos utilizado la reformulación de consultas mediante expansión de términos [5], que es una técnica simple, muy utilizada y adecuada para nuestro problema. En líneas generales, esta técnica consiste en añadir n términos, procedentes del perfil de usuario, a la consulta original realizada por dicho usuario. La intención es obtener resultados más cercanos a los intereses y preferencias del usuario. Pero hay que prestar mucha atención, ya que si el SRI usado sólo añade estos nuevos términos a los de la consulta original y ejecuta la consulta expandida, es más que probable que el usuario considere que los resultados no satisfacen sus necesidades de información. El motivo es el gran cambio que se ha producido sobre la consulta original, mayor cuantos más términos de expansión se utilicen. Esta consulta expandida recupera resultados más cercanos a su perfil de usuario que a la consulta original, siendo ésta la que representa sus necesidades de información y no su perfil. Para mitigar este efecto, sabiendo que cada término de la consulta original está ponderado por un peso de uno, cada factor de ponderación asociado a cada término del perfil es normalizado, con el objetivo de que estos pesos no sean superiores a los originales. Por ejemplo, los términos del anterior perfil de educación, tras aplicarles un factor de normalización de 0.66, quedarían como sigue: *0.45471*educ, 0.42691*centr, 0.61439*alumn, 0.6441*profesor, 0.66*enseñ, 0.64655*curs, etc.*

El anterior factor de normalización, propuesto como una primera mejora y denotada por *normRest*, trata de modular la influencia de los términos expandidos sobre la consulta original, pero dichos términos, aunque ponderados en menor medida, aún son utilizados de la misma forma que los términos originales por el sistema para la recuperación de UE (tienen los mismos niveles de exigencia de aparición que los términos originales).

Se ha utilizado el SRI estructurado Garnata [2], basado en modelos gráficos probabilísticos [9]. Garnata ordena los resultados por su RSV (Relevance Status Value). En un paso del cálculo de este valor, Garnata utiliza una función que controla el grado de exigencia de aparición de los términos de la consulta en la UE a recuperar, o dicho de otra forma, sería como modular la intensidad de una puerta AND sobre los términos de la consulta. Esta función tiene un componente paramétrico exponencial n. Cuanto mayor sea n, mayor será la exigencia de aparición de los términos de la consulta en la UE a recuperar, y como efecto colateral, el sistema tenderá a devolver UE de mayor tamaño. Este último hecho se debe a que la probabilidad de que aparezcan todos o casi todos los términos de la consulta es mayor en UE grandes que en UE pequeñas.

Debido a este comportamiento exponencial de Garnata se ha propuesto una segunda mejora, denotada por *normExpRest*, que consiste en evitar que los términos de expansión se vean afectados por dicha función exponencial. Así, sólo se requiere la aparición de los términos de la consulta original en la UE a recuperar, pero no de los términos expandidos, su aparición es un extra no un requisito. Para conocer con más detalle cómo Garnata hace la recuperación de UE, véase [3].

En cuanto a la evaluación, hemos eliminado el problema de los solapamientos, usando la estrategia de presentación de resultados XML *'Focused'* propuesta en INEX [10], donde se eliminan los solapamientos dando preferencia a las unidades con mayor RSV, y a las unidades mayores en caso de igualdad de RSV. En la sección 5.1 veremos cómo hemos evitado el problema de los fallos cercanos.

4. Metodología de Evaluación

La metodología de evaluación propuesta intenta aunar dos objetivos principales: ser automática, y por tanto muy fácil de usar, y que sus resultados sean fácilmente reproducibles y generalizables. Esta metodología se podría encuadrar dentro del grupo de simulaciones contextuales, aunque está más orientada a la efectividad de recuperación que a la interacción usuario-sistema. Además, no necesita definiciones precisas de escenarios de recuperación, ni el uso de colecciones con juicios de relevancia previos. Sólo necesita tener un conjunto de consultas características sobre la colección documental a usar, y las áreas de interés sobre las que se realizará la evaluación. Su principal ventaja, frente a las extensiones del modelo de laboratorio y a los estudios de usuario, es que es automática y no necesita de usuarios reales, con lo que esto conlleva en facilidad de uso y ahorro de recursos. Además, presenta gran parte de las ventajas de la evaluación basada en el modelo de laboratorio, tales como que sus resultados son reproducibles, se pueden hacer comparaciones entre distintas técnicas de personalización, y se pueden hacer evaluaciones iterativas, con lo que es posible identificar problemas y presentar soluciones. La precisión y fiabilidad de la metodología se verificará mediante la comparación entre sus resultados y los del estudio de usuario que realizaremos.

Esta metodología necesita simular el perfil de usuario y los juicios de relevancia sobre los resultados de las consultas enviadas al sistema. A continuación se enumeran los componentes que la forman:

Colección Documental: es el único componente que debe cumplir un requisito clave: sus documentos deben poder ser clasificados en distintas áreas de interés. Esta clasificación la puede realizar el propio sistema, o bien la colección puede estar ya clasificada. Esta característica es un prerrequisito, ya que se va a obtener el perfil de usuario basándose en el contenido de los documentos que forman cualquier área de interés. Este prerrequisito puede ser implícito o explícito a la colección. Un ejemplo explícito sería que los documentos tuvieran varias etiquetas asociadas de un vocabulario controlado, que ayudarían a un clasificador automático a clasificarlos. Un ejemplo implícito sería que los documentos ya estuvieran clasificados por su propia naturaleza, como un periódico con las noticias clasificadas por sus secciones. Las secciones determinarían las áreas de interés, y por tanto, la generación de perfiles de usuario.

Perfiles de usuario: el perfil de usuario se obtiene de un área de interés dada de la colección documental, en la cual el usuario está interesado. Se asume que un usuario con afinidad por un área de interés de la colección documental tendrá un perfil muy similar al perfil obtenido a partir del contenido de los documentos de esa área de interés. En este trabajo no pretendemos determinar el mejor perfil de usuario, está más enfocado a la evaluación, por lo que utilizamos una aproximación simple de éste, expuesta en la sección 3.

Conjunto de consultas y sus juicios de relevancia: se ejecuta cada consulta en el SRI, obteniéndose una lista de resultados ordenada por relevancia. Debido a que se utilizan usuarios simulados, también se necesita un procedimiento para simular sus juicios de relevancia. Se asume la siguiente suposición: las UE relevantes para una consulta bajo un perfil dado, son aquellas UE que pertenecen a documentos de ese perfil, y han sido recuperadas por el sistema entre los x primeros resultados, donde x es un valor relativamente bajo (100 en nuestra experimentación), ya que no tiene sentido considerar relevante una UE que, por ejemplo, se encuentre en la posición 1000 del ranking de resultados. La intuición tras esta suposición es que si una UE está entre las primeras recuperadas por el sistema para la consulta, y además corresponde al área de interés que nos ocupa, entonces dicha UE trata simultáneamente sobre la temática de la consulta original y sobre el área de interés, por lo que será relevante para el perfil correspondiente.

5. Experimentación

La técnica de personalización utilizada es la expansión de la consulta original con los n primeros términos del perfil de usuario. El número de términos de expansión junto con su peso normalizado serán los principales parámetros a ajustar de la técnica de personalización, junto con otros propios del SRI utilizado.

El objetivo de la metodología de evaluación propuesta es evaluar de una forma simple, efectiva y con bajo coste, cómo se ve afectado el rendimiento de

un SRI con el uso de cualquier técnica de personalización (basada en términos). En esta sección se muestra cómo hemos utilizado dicha metodología en nuestro estudio experimental.

5.1. Métrica de Evaluación

Para medir la ganancia de rendimiento entre la aproximación original y personalizada la métrica elegida es la NDCG (Normalized Discounted Cumulative Gain), basada en la métrica DCG [8], pero normalizada por el ranking ideal de los resultados relevantes. Está diseñada para la estimación de la ganancia de relevancia global obtenida por un usuario, al juzgar los documentos en las primeras posiciones del ranking de resultados. Para ello se usa un factor de descuento o penalización que reduce el efecto del documento sobre el valor de la métrica, conforme la posición de éste aumenta en el ranking. El valor de la métrica para una lista de resultados dada, se calcula como sigue:

$$NDCG@x = \frac{1}{N} \sum_{i=1}^{y} \frac{2^{rel(d_i)} - 1}{\log(i+1)}, \tag{3}$$

donde y es la posición hasta la que se evalua (50 en nuestros experimentos), N es el valor DCG ideal para los resultados relevantes (suponiendo todos los resultados relevantes consecutivos desde la primera posición del ranking), i es la posición del ranking de la UE que se está evaluando, d_i es la UE en la posición i, y $rel(d_i)$ es el valor de relevancia de d_i, siendo igual a 0 si el resultado es no relevante e igual a 1 si sí lo es.

Para la consulta original, una UE se considera relevante si pertenece al área de interés bajo la que se está evaluando, y su posición es anterior al umbral x. Al usar recuperación XML, no hay una correspondencia uno a uno entre las UE devueltas para la consulta original y la expandida. Una UE de la consulta original puede dividirse en varias UE más pequeñas en la consulta expandida o viceversa. Esto es una desventaja de la evaluación de los SRI estructurados y no pasa con los SRI no estructurados, ya que se devuelve el documento completo. Por este motivo, y con el fin de calcular el valor de la métrica para la consulta expandida, se considera una UE como relevante bajo un área de interés dada, si dicha UE es un ascendiente o descendiente de cualquier UE relevante perteneciente a la consulta original correspondiente a la actual consulta expandida. De este modo evitamos el problema de los fallos cercanos expuesto en la sección 2.

5.2. Resultados Experimentales

En nuestros experimentos hemos utilizado la metodología de evaluación propuesta en la sección 4. Para ello hemos usado una colección basada en los diarios de sesión de comisiones (comisiones) del Parlamento de Andalucía [4], en la que cada comisión se centra en un área de interés específica. Se han usado algunas de éstas para la simulación de los siguientes 8 perfiles de usuario: *Agricultura, ganadería y pesca, Coordinación, régimen de las administraciones públicas y*

justicia, Cultura, turismo y deporte, Economía, hacienda y presupuestos, Educación, Empleo, Medio ambiente y Salud. Se ha usado un conjunto de 16 consultas[1], propuestas por expertos en la colección documental con la que trabajamos, y sus correspondientes juicios de relevancia, simulados tal y como se ha especificado en el tercer punto de la sección 4.

Existen dos valores de la métrica de evaluación, uno para los resultados de la consulta original o no personalizada $NDCG\_NP$ y otro para los resultados de la consulta personalizada $NDCG\_P$. Si $NDCG\_P > NDCG\_NP$, la inclusión de la técnica de personalización en el proceso de recuperación es ventajosa.

En la figura 1 se muestran los valores promedios NDCG obtenidos mediante la aplicación de todas las consultas sobre todos los perfiles, aplicando personalización para ambas aproximaciones ($normRest = nR$ y $normExpRest = nER$; leyenda), para varios valores de la función exponencial (parámetro n en las gráficas), usando distinto número de términos de expansión (número en leyenda) y para distintos valores del factor de normalización (eje de abcisas). Los valores NDCG obtenidos sin usar personalización (NP) son planos ya que no hay términos de expansión. Su valor sólo cambia muy ligeramente con la variación de la exponencial.

Observando la figura 1 se pueden sacar las siguientes conclusiones:

- El uso de las técnicas de personalización resulta claramente beneficioso, puesto que los resultados de los experimentos con el sistema personalizado, con cualquier combinación de parámetros, son sistemáticamente mejores que los del sistema no personalizado.
- Se observa que la preocupación por la posibilidad de distorsionar la consulta original con la inclusión de los términos expandidos, era real. Esta conclusión se extrae observando la pérdida de rendimiento que conlleva el aumento del número de términos de expansión para la aproximación nR.
- En cambio, se puede observar que cuando se usa la aproximación nER, cuantos más términos de expansión se usen mejores resultados se obtienen. Esto es debido a que no se requiere de forma tan estricta que los términos expandidos de la consulta estén en la UE a recuperar, al contrario de lo que sucede con la aproximación nR.
- El factor de normalización también se comporta de forma opuesta en ambas aproximaciones. En la aproximación nR, cuanto mayor es el peso de los términos de expansión (debido a un factor de normalización mayor) mayor es su influencia, obteniéndose peores resultados. Por el contrario, en la aproximación nER, cuanto mayor es el factor de normalización de los términos expandidos mejores resultados se consiguen. La razón es la misma que en el caso del número de términos de expansión, la mayor o menor exigencia de que los términos de expansión formen parte de las UE recuperadas.
- El parámetro n, que modula la intensidad de la puerta AND sobre los términos de la consulta, parece no tener una influencia determinante, pues los resultados son muy parecidos para los tres valores (3, 7 y 20) probados de

[1] Ejemplos de tales consultas son “informática”, “gastronomía andaluza” o “depuración de aguas”.

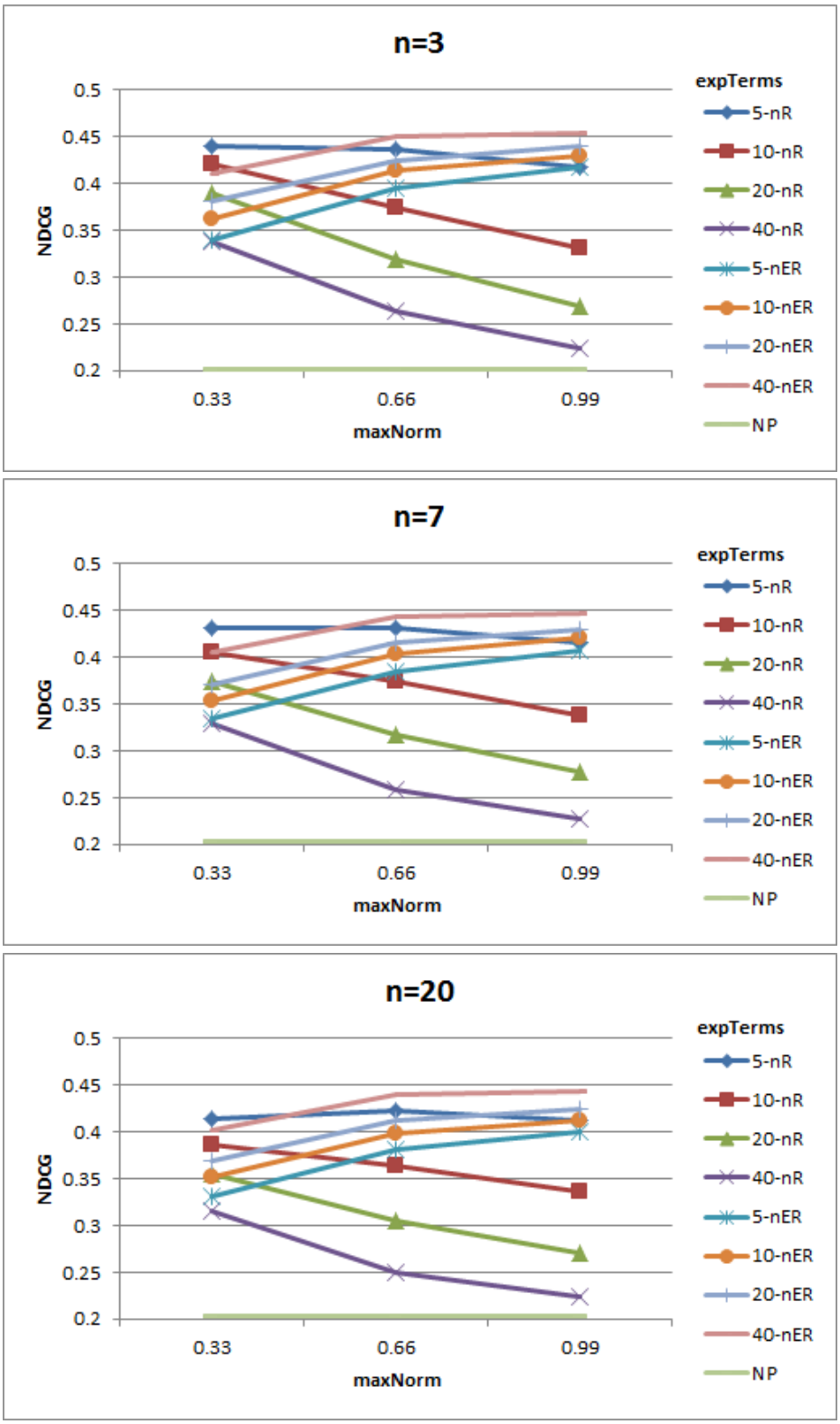

Figura 1. Resultados para todas las posibles configuraciones de los principales parámetros del SRI y de la técnica de personalización aplicada.

este parámetro. Solamente se aprecia un muy leve empeoramiento en la aproximación *nER* al aumentar n.

- En general, la aproximación *nER* se comporta mejor que la *nR*. De hecho, el mejor resultado se ha obtenido usando dicha aproximación, empleando 40 términos y un factor de normalización de 0.99 (los valores máximos probados de los parámetros). Dicho resultado, un valor de NDCG de 0.45355, representa una ganancia respecto al sistema no personalizado (con un valor NDCG de 0.20225) del 124 %.

Como conclusiones generales podemos decir que (1) la aplicación de la metodología propuesta posibilita el desarrollo y mejora de los SRI personalizados, debido a que rápidamente se pueden ver los efectos que tiene cualquier modificación sobre las técnicas de personalización usadas o sobre el propio modelo de RI; y (2) usando las técnicas de personalización propuestas se consiguen porcentajes de mejora bastante importantes en la ganancia de relevancia experimentada por el usuario. No obstante, estos resultados deben ser validados comparándolos con los resultados obtenidos del futuro estudio de usuario. En ese momento, podremos afirmar con seguridad nuestras previsiones de que la metodología propuesta, que ofrece bastantes ventajas sobre el resto de métodos de evaluación existentes, es una alternativa fiable a dichos métodos.

6. Conclusiones y Trabajos Futuros

La personalización sobre un SRI estructurado XML presenta varios problemas, unos por tratar con información estructurada y otros por la técnica de personalización utilizada, expansión de consultas en nuestro caso. Para solventarlos se ha propuesto un método de personalización en la recuperación XML que modula la influencia de los términos de expansión en la recuperación, y que modifica el modelo de RI usado para hacer que dichos términos no distorsionen la consulta original del usuario, y por tanto, sus necesidades de información.

La evaluación de los SRI personalizados, comúnmente realizada a través de estudios de usuario, requiere de un tremendo esfuerzo tanto en tiempo como en recursos, que no siempre son factibles. La metodología propuesta en este trabajo, que pretende mejorar las ya existentes, es capaz de simular perfiles de usuario y sus posibles juicios de relevancia, haciendo posible la evaluación de la inclusión de una técnica de personalización en el SRI con suma facilidad. Esta metodología es especialmente útil en las primeras etapas de desarrollo de un nuevo SRI, para equipos de investigación con bajos recursos, e incluso para la elección de los parámetros de configuración óptimos del costoso estudio de usuario. Y se ha utilizado para el desarrollo y evaluación de las dos anteriores mejoras, obteniéndose unas ganancias de relevancia importantes.

Como trabajo futuro destaca la realización del estudio de usuario con la mejor configuración posible de los parámetros de evaluación. Los resultados de este estudio nos servirán para validar la metodología de evaluación propuesta en este artículo. Otro trabajo futuro será desarrollar otras técnicas de personalización, como por ejemplo diferentes estrategias de re-ranking de resultados.

Agradecimientos. Este trabajo ha sido financiado por la Consejería de Innovación, Ciencia y Empresa de la Junta de Andalucía (P09-TIC-4526), el Ministerio de Educación y Ciencia (TIN2011-28538-CO2-02) y el Programa Consolider Ingenio 2010 (MIPRCV CSD2007-00018).

Referencias

1. J. Allan. Challenges in information retrieval and language modelling. *In Report of a workshop held at the Center for Intelligent Information Retrieval*, University of Massachusetts, Amherst, 2002.
2. L.M. de Campos, J.M. Fernández-Luna, J.F. Huete, and A.E. Romero. Garnata: An information retrieval system for structured documents based on probabilistic graphical models. *In Proc. of the 11th Int. Conf. of Inform. Proc. and Manag. of Uncert. in Knowl. Syst. (IPMU)*, 1024-1031, 2006.
3. L.M. de Campos, J.M. Fernández-Luna, J.F. Huete, C. Martín-Dancausa, and A.E. Romero. New utility models for the Garnata information retrieval system at INEX'08. *Lecture Notes in Computer Science*, 5631:39-45, 2009.
4. L.M. de Campos, J.M. Fernández-Luna, J.F. Huete, C. Martín-Dancausa, A. Tagua-Jiménez, and C. Tur-Vigil An integrated system for managing the Andalusian Parliament's digital library. *Program–Electronic Library and Information Systems*, 43(2):156-174, 2009.
5. P.A Chirita, C.S. Firan, and W. Nejdl. Personalized query expansion for the web. *In Proc. of the 30th annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR '07)*, 7-14, 2007.
6. N. Furh, M. Lalmas, S. Malik, and G. Kazai (Eds.). Advances in XML Information Retrieval and Evaluation. *Lecture Notes in Computer Science*, 3977, 2006.
7. J.F. Gantz. The Diverse and Exploding Digital Universe. *IDC white paper*, 2, pp.1-16, 2008.
8. K. Jarvelin, and J. Kekalainen. Cumulative gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422-446, 2002.
9. F.V. Jensen. Bayesian Networks and Decision Graphs. *Springer Verlag*, 2001.
10. J. Kamps, J. Pehcevski, G. Kazai, M. Lalmas, and S. Robertson. INEX 2007 Evaluation Measures. *Lecture Notes in Computer Science*, 4862:24-33, 2008.
11. M. Lalmas. XML Retrieval. *Morgan & Claypool Publishers*, 2009.
12. F. Qiu, and J. Cho. Automatic identification of user interest for personalized search. *In Proceedings of WWW '06*, 727–736, 2006.
13. E. Santos Jr, Q. Zhao, H. Nguyen, and H. Wang. Impacts of user modeling on personalization of information retrieval: an evaluation with human intelligence analysts. *In 4th Workshop on the Evaluation of Adaptive Systems*, 27-36, 2005.
14. L. Tamine-Lechani, M. Boughanem, and M. Daoud. Evaluation of contextual information retrieval effectiveness: overview of issues and research. *Know. Inf. Syst.*, 24:1-34, 2009.
15. G. Uchyigit, and M.Y. Ma. Personalization Techniques and Recommender Systems. *World Scientific*, 2008.
16. R. Wilkinson, and M. Wu. Evaluation experiments and experience from perspective of interactive information retrieval. *In 3rd Workshop on the Evaluation of Adaptive Systems*, 221-230, 2004.
17. Y. Yang, and B. Padmanabhan. Evaluation of online personalization systems: A survey of evaluation schemes and a knowledge-based approach. *Journal of Electronic Commerce Research*, 6(2):112-122, 2005.