REVISTA COLOMBIANA DE COMPUTACIÓN
Volumen 10, número 2
Págs. 142 - 154

Proceso de Categorización de Consultas Basado en Visibilidad en un Dispositivo Móvil

María Valencia* Santiago Eibe* Ernestina Menasalvas*
Fecha de Recibido: 01/11/2009 Fecha de Aprobación: 14/12/2009

Resumen

La categorización de consultas web es una actividad de creciente interés para las organizaciones debido a que les permite proporcionar a sus usuarios servicios de valor añadido en respuesta a las consultas ellos envían al motor de búsqueda de la compañía. Esto representa un reto en dispositivos móviles no solo por los problemas asociados con la interpretación de una consulta tomando en cuenta el contexto del usuario bajo supuestos de movilidad, sino también por los problemas derivados de las limitaciones de recursos en este tipo de dispositivos. La necesidad de autonomía requerida en una situación donde el minero de datos no está presente, hace que el problema sea aun más desafiante. En este artículo abordamos el problema de categorización de consultas en dispositivos móviles. Para ello presentamos en primer lugar un modelo para la visibilidad de términos. En base a este modelo definimos el proceso para la categorización. La parte innovadora de este proceso comprende la definición de parámetros y tareas, encargadas de controlar las restricciones derivadas por las limitaciones de recursos. Finalmente presentamos un modelo de metadatos, el cual es requerido como base para la automatización.

Palabras clave: *Categorización de Consultas, Motores de Búsqueda, Data Mining Ubicuo.*

Abstract

Web query categorization is an activity of increasing interest due to the fact that it makes possible for organizations to provide users with add-value services in response to queries they submit to the search engine of the company. This is challenging in mobile devices not only because of the problems associated with the interpretation of a query taking into account user contexts under mobility assumptions, but also due to the problems derived from the limitations of resources in this kind of devices. The need of autonomy required in a situation in which the data miner is not present makes the problem even more challenging. In this paper we address the problem of query categorization in mobile devices. To do so, we first define a model for visibility of terms. Based on this model we define the process for categorization. The innovative part of this process realize on the definition of parameters and tasks in charge of controlling the constraints due to resources limitations. We finally present a metadata model required as a basis for automatization.

Keywords*: Query Categorization, Search Engines, Ubiquitous Data Mining.*

* Universidad Politécnica de Madrid, Facultad de Informática,
mvalencia@alumnos.fi.upm.es, {seibe, emenasalvas}@fi.upm.es

1 Introducción

Los avances tecnológicos acontecidos en los últimos años, y en concreto la computación ubicua, permiten que usuarios móviles con pequeños y potentes dispositivos utilicen aplicaciones en casi cualquier lugar. Se genera entonces la necesidad para muchas compañías de ofrecer servicios personalizados de manera local, analizando el comportamiento local de cada usuario, lo cual constituye un nuevo reto y una nueva oportunidad.

El nuevo reto consiste en que para algunas decisiones no se hace necesaria nada más que la información generada localmente. Desde este punto de vista, pasamos de entornos en los que la computación de los modelos se realizaba analizando toda la información disponible acerca de los usuarios y su comportamiento, a un modelo en el que sólo es necesario analizar el comportamiento local de cada usuario para responder a sus necesidades locales. Todo ello plantea la necesidad de componentes locales de mining.

La tecnología plantea un nuevo modelo de computación, el cual presenta diferentes retos entre los que destacan heterogeneidad, escalabilidad, fiabilidad y seguridad, privacidad, movilidad, context-awareness, y resource-awareness entre otros [2].

Actualmente, la tecnología permite contar con información del contexto del usuario: localización, tiempo, características del entorno del usuario y, en general, variables del estado del usuario. Este enfoque de personalización es la motivación de nuestro trabajo, en el que la meta es ofrecer la posibilidad de tomar en cuenta factores que describen mejor al usuario.

No obstante, hacer Data Mining sobre estos datos constituye un reto debido a que se enfrenta a las dimensiones a lo largo de las cuales se presenta la ubicuidad: dispositivos, datos, proceso, usuarios y contexto. En el caso particular de los buscadores web, uno de los retos es el problema de modelar el comportamiento del usuario web basado en su información local y, más en concreto, en el caso de realizar personalización web en base a información local.

El problema de análisis de datos locales ha sido abordado en trabajos anteriores y en diversos campos de aplicación, entre los que se encuentran tecnología asistencial, la web y domótica entre otros. En [7], los autores presentan un sistema de búsqueda móvil que dividen en dos partes lógicas: el motor de búsqueda local y un meta crawling. En concreto, el motor de búsqueda web local proporciona al usuario la

capacidad de adaptar los resultados de la búsqueda a sus propias necesidades, mediante la actualización del índice de búsqueda a partir del análisis conjunto de la información generada localmente por cada usuario y el cambio de contenido. En la misma línea, en [5] se presenta un enfoque de búsqueda personalizada. Esta personalización se propone del lado del cliente y se basa en el mapeo de los intereses a una taxonomía y en la categorización de los resultados, lo que permite desplegar los resultados de acuerdo a los intereses del usuario. Estos intereses son obtenidos a partir de información local que describa los intereses profesionales y habilidades del usuario.

Por otra parte, los autores describen en [4] un enfoque para predecir el contexto del usuario aplicado a entornos asistidos, para lo cual analizan la información que describe la ejecución de un conjunto de tareas por parte de cada usuario. Por otro lado, en [1] se describe el framework SPE (Secure Persona Exchange), que proporciona servicios personalizados a los usuarios en entornos de computación ubicua en base a las preferencias de cada usuario. La información analizada en este enfoque corresponde a las preferencias del usuario que son almacenadas en su dispositivo móvil, y son proporcionadas de manera explícita por el usuario. Sin embargo, este enfoque no tiene en cuenta el contexto dinámico del usuario como localización o la actividad actual y la personalización está restringida a los datos indicados por el usuario.

En este artículo se presenta en primer lugar un modelo que permite clasificar las consultas emitidas por un usuario en base a la frecuencia de aparición de los términos que contiene y la relación que tienen las consultas con el contenido del sitio web en el que se realiza la búsqueda. En este caso, la categorización de consultas permite personalizar el sistema, en tanto que permite ofrecer el servicio adecuado a cada usuario en virtud de la categoría asociada a cada consulta.

En [8], los autores identifican la automatización de los procesos de data mining y el desarrollo de una metodología para data mining como uno de los retos actuales de data mining. Análogamente, en entornos como el que se plantea en este artículo que se basan en el análisis de datos en un dispositivo móvil, es también necesario minimizar la intervención del minero de datos. En este sentido, analizamos la necesidad de autonomía en entornos ubicuos en la sección 3.3.

Además, el aporte presentado en este artículo se complementa con un modelo de metadatos que permite automatizar el proceso de extracción de modelos de personalización de manera semiautónoma. El artículo se ha organizado de la siguiente manera. En la sección 2 definimos el escenario de categorización de consultas basado en visibilidad. En la

sección 3 se describen los retos que se presentan como consecuencia de la ejecución local del proceso de personalización en base a información local. El modelo de metadatos presentado en la sección 4 es una posible solución para afrontar los problemas de autonomía y limitación de recursos. Para concluir la sección 5 presenta la discusión y las líneas futuras de actuación.

2 Definición del Escenario

Para analizar el proceso de minería en dispositivos móviles presentamos el caso particular de las búsquedas en un portal web. En concreto, abordamos el problema de identificar la intención subyacente a las consultas emitidas por un usuario, con el fin de desplegar el contenido apropiado en la página principal del sitio web.

Suponemos un usuario en un dispositivo local (ordenador portátil, PDA, etc.), donde:

- Se registran localmente en el dispositivo las consultas emitidas por el usuario
- Se asume que las consultas anteriores del usuario han sido previamente cargadas, limpiadas y parseadas, actividad que debe ser ejecutada cada vez que se cuente con nuevos registros de consultas
- El contenido desplegado por el sitio web en su página principal es capturado automáticamente en intervalos de tiempo y almacenado en el dispositivo local

Proponemos un modelo de predicción de las consultas del usuario basado en la frecuencia de aparición de las mismas, y la influencia que los contenidos ofertados en la página principal pueden tener en ellas.

De acuerdo a esto, en este enfoque se toman en cuenta los siguientes aspectos:

- Relación de la consulta del usuario con otras consultas del mismo usuario
- Relación de la consulta del usuario con el contenido desplegado en la página principal del sitio web sobre el que el usuario realiza las búsquedas

Para analizar estas relaciones definimos el concepto de visibilidad de los términos que contiene la consulta.

2.1 Visibilidad de los Términos

Definimos la visibilidad de un término como la frecuencia de aparición del término (número de veces que aparece) en un periodo determinado.

En concreto, proponemos medir cómo de visibles son los términos contenidos en una consulta en consultas emitidas previamente por el mismo usuario y cómo de visibles son los términos en la página principal del sitio web.

Para ello definimos medidas que cuantifican el grado en el cual los términos contenidos en la consulta tienen relación con los términos contenidos en consultas recientes y con el contenido de la página principal, lo que permitiría clasificarlos en:

- término estable
- término nuevo
- término frecuente
- término atípico
- término relacionado con el contenido de la página principal
- término que anticipa el contenido de la página principal

2.2 Descripción del Proceso

El modelo propuesto se muestra en la figura 1.

Las fuentes de datos utilizadas como entrada para la aplicación del método son:

1. Datos de Entrada: Log en CLF que contiene las consultas emitidas por el usuario. Cada consulta tiene asociado además el número de resultados que retornó el buscador y el tiempo de respuesta
2. Contenido de la página principal: Porciones de contenido capturados de la página principal del sitio web, para cada uno de los cuales se almacena la fecha y hora en que se capturó, la posición dentro de la página y el texto asociado. En nuestro enfoque diferenciamos el texto de la porción de contenido según corresponda al titular o al cuerpo del texto.

A continuación describimos las fases del proceso que se muestran en la figura 1.

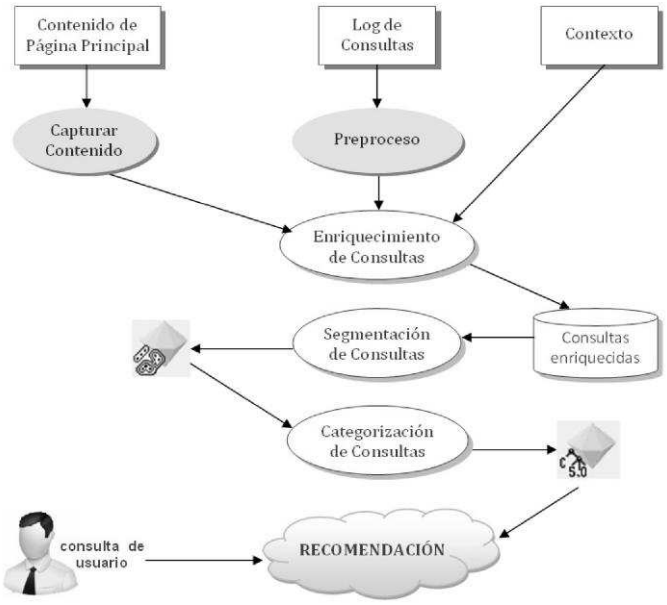

Fig.1. Fases del Proceso

2.2.1 Fase 1: Enriquecimiento de Consultas

La meta de esta fase consiste en enriquecer las consultas de usuario de acuerdo a un conjunto de indicadores que denominamos indicadores de visibilidad. Esta fase consiste en:

- Calcular la visibilidad de cada término contenido en la consulta en diferentes periodos y posteriormente enriquecer la consulta en base a estos coeficientes

- Establecer el mapping entre cada término y el contenido de la página principal que para diferentes periodos indique si el término está o no contenido en página principal.

De acuerdo a esto, las consultas emitidas por el usuario son enriquecidas en función de la visibilidad de los términos que contiene. Se consideran los siguientes casos:

- Visibilidad del término en otras consultas emitidas por el mismo usuario

- Visibilidad del término en el contenido de la página principal del site

Se definen los siguientes indicadores:

- *visible en consultas-término estable*
- *visible en consultas-término nueva*
- *visible en consultas-término frecuente*
- *visible en consultas-término atípica*
- *visible en página principal-término headline*
- *visible en página principal-término body*

Las consultas son enriquecidas con la información relativa a los términos que contienen en periodos de diferente duración: 24 horas, 7 días, día siguiente, semana anterior, etc.

2.2.2 Fase 2: Segmentación de Consultas

La meta de esta fase es identificar tipos de consultas de usuario en base a su relación con otras consultas y con el contenido de la página principal. Para ello se realiza una segmentación de consultas que utiliza los atributos derivados en el paso anterior, los cuales corresponden a los indicadores de visibilidad descritos en la sección 2.1 para diferentes periodos. Ejemplos de estos indicadores son:

- los términos contenidos en la consulta son frecuentes en las consultas emitidas durante la última semana
- los términos de la consulta aparecen por primera vez en el periodo de 24 horas
- los términos de la consulta están contenidos en la página principal del sitio web durante el último mes

2.2.3 Fase 3: Categorización de Consultas

La meta de esta fase es obtener un modelo que permita predecir la intención del usuario a partir de la información local que describe la consulta que realiza. Sin embargo, uno de los principales problemas de la categorización consiste en la necesidad de contar con un conjunto de datos previamente categorizados. Por ello, la entrada de esta fase corresponde al resultado obtenido en la fase anterior, tomando como atributo clasificador el tipo de consulta. El proceso de clasificación se propone en base a los atributos que describen la consulta, entre los que podemos mencionar:

- fecha y hora en que la consulta es emitida
- localización del usuario que realiza la búsqueda
- respuesta del sistema ante la consulta
- tiempo de respuesta del sistema

3 Consecuencias de la Ejecución Local del Modelo en un Dispositivo Móvil

Los enfoques tradicionales de categorización de consultas web ejecutan el proceso de manera centralizada, utilizando para ello la información de las búsquedas realizadas por todos los usuarios. El enfoque descrito en este artículo se diferencia en la naturaleza de la información que utiliza, debido a que el modelo de usuario es calculado en base a información del propio usuario registrada localmente.

La ejecución local del modelo en un dispositivo móvil, tomando como entrada información local del usuario nos enfrenta a los retos impuestos por la ubicuidad. En estos entornos requerimos de componentes de minería que tengan en cuenta la limitación de recursos y que sean adaptables. Pero ello no es suficiente sin incorporar autonomía, puesto que el reto de la ubicuidad plantea además un entorno en el que no podemos asumir que los ingenieros de minería de datos estén disponibles durante el proceso de descubrimiento de conocimiento. Estos aspectos se describen en las secciones siguientes.

3.1 Sistemas ”Resource-Aware”

Debido a que la ejecución del modelo que planteamos se realiza en un dispositivo móvil, la limitación de recursos plantea una problemática que hace necesario que nuestro modelo sea resource-aware. En particular, el modelo debe considerar las restricciones que determinan los recursos de memoria, almacenamiento, conectividad y suministro de energía para el correcto cálculo del modelo.

3.2 Adaptabilidad

Adaptabilidad es definido como la cualidad de ser capaz de acomodarse o ajustarse a nuevas condiciones o usos [6]. En este sentido, las condiciones del entorno en el que ejecutamos el modelo que presentamos hacen que sea necesario que el modelo sea adaptable, por un lado, a los intereses del usuario y, por otro, a las características del entorno.

3.2.1 Adaptabilidad con Respecto a los Intereses del Usuario

Los intereses y el contexto del usuario no son factores estáticos y pueden variar en el tiempo. Por ello, se requiere que el modelo que describe el comportamiento del usuario sea adaptable a estos cambios. Consecuentemente, el modelo debe ser actualizado y validado constantemente, lo cual permite garantizar su validez.

3.2.2 Adaptabilidad con Respecto a las Características del Entorno

Como se vio en la sección anterior, se requiere que el sistema sea consciente de los recursos que están disponibles. Sin embargo, es necesario además que el sistema emplee una estrategia que le permita adaptarse en las situaciones en que varíe la disponibilidad de los recursos.

Por otro lado, debemos considerar el coste computacional que implica la actualización de los modelos en cuanto a almacenamiento de datos, memoria y tiempo de respuesta. Debido a que las restricciones de almacenamiento no permiten en muchas ocasiones almacenar toda la información histórica del usuario, se deben establecer mecanismos que permitan desechar información a medida que se recogen nuevos datos del usuario. Así mismo, también es requerido hacer una actualización incremental de los modelos teniendo en cuenta las acciones recientes del usuario.

3.3 Autonomía

La ubicuidad plantea la necesidad de contar con componentes autónomos capaces de procesar la información recibida y extraer conocimiento sin la intervención del minero. Esto implica que el componente de minería, o un controlador que lo gestione, sea capaz de conocer en cada momento de la ejecución el estado de la misma, las precondiciones de cada proceso, y las restricciones asociadas a los recursos, el modelo y los datos de entrada. En base a esta información, el componente debe ser capaz de decidir si puede continuar la ejecución o no, así como decidir el paso siguiente. Esta situación hace que se minimice, y en el caso ideal que no sea necesaria, la presencia humana que supervise el proceso de mining.

Pero ello requiere sistematizar y automatizar el proceso. En base a esto proponemos un modelo de metadatos que presentamos en la siguiente sección, como soporte al proceso presentado en este artículo.

4 Una Solución Basada en Metadatos

Los retos descritos en la sección anterior plantean la necesidad de disponer de información detallada que describa los datos y el proceso. Ello motiva el desarrollo de una base de datos independiente que dé soporte a la automatización y que almacene esta información que, en adelante, llamamos metadatos. En este modelo hacemos énfasis en los retos asociados a la autonomía y la limitación de recursos.

Un componente autónomo se caracteriza por la capacidad de auto-gestionarse, lo que supone controlar el flujo del proceso y las restricciones asociadas a éste. Para ello es necesario, por un lado, contar con información que permita decidir en cada momento si es posible o no lanzar la ejecución de una tarea y, por otro lado, registrar información a partir de la cual el sistema aprenda y se adapte a nuevas condiciones del entorno.

El modelo de metadatos permite automatizar el proceso en la medida en que proporciona información al sistema que guía y controla la ejecución de las tareas que constituyen el proceso. De esta manera, el sistema consulta los metadatos para determinar si puede o no ejecutar una tarea dada.

En este sentido, identificamos dos planos de información que deben ser integrados en el modelo:

- ***Plano de descripción:***

  En este plano se define la información que da soporte a la automatización y que describe en detalle cada tarea. El soporte a la automatización está dado por la especificación de parámetros y restricciones que permiten la ejecución de la tarea sin supervisión humana. Para definir los elementos que representamos en este plano nos basamos en los estándares existentes, en concreto, en PMML que define el diccionario de datos [3]. También definimos las entradas y salidas asociadas a cada tarea, así como las restricciones sobre cada una de ellas.

  Por otra parte, dado que requerimos la ejecución de las tareas en un entorno resource-aware, incluimos en el modelo información que describa los recursos y las restricciones sobre éstos que permitan garantizar dicha ejecución.

  La figura 2 muestra la parte del modelo de metadatos propuesto correspondiente al plano de descripción, el cual define la información mínima que debe contener. Este modelo permite representar la división de un proceso en tareas, de las cuales se especifican las entradas, salidas, recursos y parámetros asociados.

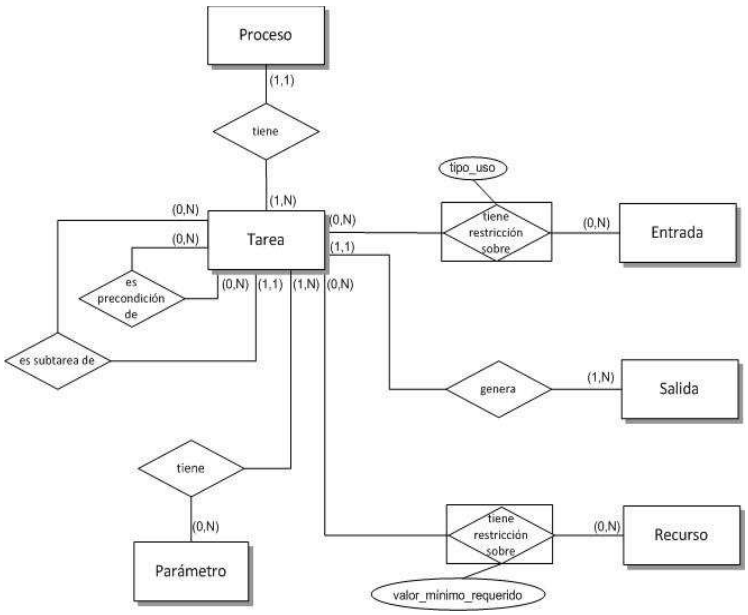

Fig.2 . Definición del plano de descripción

- ***Plano de monitorización:***

  Este plano corresponde al registro de la ejecución de las diferentes tareas. La dinamicidad de esta información permite contar con información actualizada acerca del comportamiento de cada tarea y es utilizada en el proceso de mejora. Un ejemplo de ello es presentado en la figura 3.

  El modelo de metadatos correspondiente a este plano permite describir cada ejecución de una tarea. Esta descripción se presenta mediante la especificación de los parámetros con que se realiza la ejecución y el estado de finalización de la misma. Por otro lado, incluimos en el modelo la marca de ejecución, que permite identificar las ejecuciones que corresponden a un mismo ciclo de ejecución (ejecución de todas las tareas).

  Finalmente y de manera complementaria a la especificación de recursos que incluimos en el plano de descripción, se establece una relación entre cada ejecución y los recursos, lo cual permite registrar la cantidad real de recursos utilizados y, de esta manera, monitorizar las condiciones del dispositivo en las que se ejecutan las tareas.

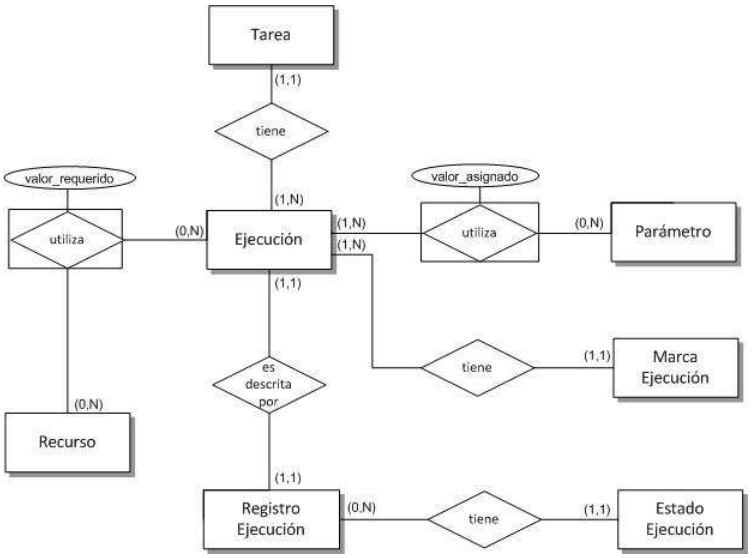

Fig.3 . Definición del plano de monitorización

5 Conclusiones

En este artículo abordamos el problema de identificar la intención del usuario mediante el análisis local de información local en dispositivos móviles. En primer lugar presentamos un modelo de predicción de las consultas del usuario basado en la frecuencia de aparición de los términos que contiene y la influencia que los contenidos del sitio web pueden tener en ellas. En segundo lugar analizamos los retos asociados a la ejecución local del modelo. Estos retos están asociados con la urgente necesidad de desarrollar componentes que reduzcan la intervención manual del experto y que sean capaces de autogestionarse. Se quiere ir entonces hacia la definición de componentes autónomos en dispositivos móviles, lo cual requiere la automatización previa del proceso. Para ello, proponemos un modelo basado en metadatos, que permita al componente conocer en cada momento si puede o no puede ejecutar el proceso. Este trabajo constituye un primer paso hacia la automatización del proceso en dispositivos móviles. En este momento, estamos trabajando en la implementación del modelo de categorización de consultas en un dispositivo móvil, usando para ello los metadatos propuestos. Esta implementación nos ha permitido valorar la generalidad del modelo propuesto. Una de las razones es que facilita la especificación de las tareas que componen el proceso de categorización independientemente de la fase del proyecto a la que corresponda. Por otra parte, las primeras pruebas del sistema conducen a que los tiempos

de ejecución se reducen, debido a que la ejecución de las tareas se realiza de manera secuencial en los momentos en que los recursos están disponibles. Todo ello, en contraposición al esquema tradicional, en el que se requería que un usuario verificara la correcta finalización de una tarea y el estado del dispositivo para lanzar la ejecución de la siguiente tarea.

Referencias

[1] A. Brar y J. Kay. Privacy and security in ubiquitous personalized applications. *In Proceedings of User Modelling Workshop on Privacy-Enhanced Personalization*, Edinburgh pages 47-54, 2005

[2] C. A. da Costa; A. C. Yamin y C. F. R. Geyer. Toward a General Software Infrastructure for Ubiquitous Computing. *IEEE Pervasive Computing*, 7(1):64-73, 2008

[3] DMG. PMML Version 3.2.s http://www.dmg.org/pmml-v3-2.html, May 2009

[4] S. W. Lee et al. Context aware life pattern prediction using fuzzy state q-learning. *In Proceedings of 5th International Conference On Smart homes and health Telematics (ICOST 2007)*, Nara pages 188-195, 2007

[5] Z. Ma; G. Pant y O. R. Sheng. Interest-based personalized search. *ACM Transactions on Information Systems (TOIS)*, 25(1):5, 2007.

[6] Real Academia Española. *Diccionario de la Lengua Española*. Editorial Espasa

[7] P. Tiago et al. Mobile Search - Social network Search Using Mobile Devices. *In Proceedings of 5th IEEE Consumer Communications and Networking Conference*, Las Vegas pages 1201-1205, 2008.

[8] Q. Yang y X. Wu. 10 Challenging Problems in Data Mining Research. *International Journal of Information Technology and Decision Making*, 5(4):597-604, 2006.