Publicando listas de correo en la web semántica

Diego Berrueta <diego@berrueta.net>

Junio 2005

Abstract

Los archivos de las listas de correo (es decir, los mensajes antiguos) son frecuentemente publicados en la web e indexados por los buscadores convencionales. La base de conocimientos que introducen en la web es enorme. Sin embargo, una gran cantidad de información se pierde durante la publicación, con el resultado de que los archivos publicados son incómodos de consultar y poco funcionales. Este artículo describe la aplicación de la web semántica para evitar la pérdida de información y habilitar la construcción de nuevas aplicaciones para explotar más convenientemente la información.

Palabras clave

Lista de correo, web semántica, archivos, RDF.

1 Introducción

Las listas de correo son parte fundamental de la comunicación en Internet. Existen listas de correo dedicadas a cualquier tema de interés imaginable. Hoy en día, es común que las listas de correo publiquen sus archivos (los mensajes antiguos) en forma de páginas web, lo que dispara su utilidad, especialmente en combinación con los buscadores actuales. Gracias a esta publicación, es posible consultar los mensajes desde el navegador, sin necesidad de estar suscrito a las listas de correo, y también se puede localizar un mensaje usando Google u otro buscador.

Estos archivos contienen una formidable base de conocimiento, especialmente en temas técnicos. Un uso muy común consiste en introducir un mensaje de error (de una aplicación) en Google y obtener como resultado un mensaje archivado que aborda el problema, probablemente porque alguien se ha encontrado previamente con el mismo error y ha efectuado la consulta en una lista de correo pública. Con suerte, alguna de las respuestas al mensaje localizado contendrá la solución al problema, aportada por un experto suscrito a la lista de correo.

2 Problemas

Por desgracia, consultar los archivos de una lista de correo en la web es más incómodo que hacerlo mediante un cliente de correo electrónico. Por poner sólo algunos ejemplos, el navegador no permite ejecutar ninguna de estas acciones:

- Mostrar el hilo de la conversación en forma de árbol.
- Imprimir el hilo completo.
- Mostrar una lista de los mensajes entre dos fechas arbitrarias.
- Ocultar los mensajes que no tienen respuestas.
- Mostrar sólo los mensajes de una cierta persona.
- Buscar una cadena de texto sólo en los mensajes de un determinado hilo.
- Descargar el hilo como un fichero, o cualquier otra forma de exportar la información para poder acceder a ella desde un cliente de correo electrónico o fuera de línea.
- Responder a un mensaje usando un cliente de correo (o un webmail) y citando el mensaje original.

Al indexar los archivos de las listas de correo, los buscadores se encuentran en ocasiones que los mensajes están replicados en varios servidores (mirrors). Al no tener forma de identificar los mensajes, la desgraciada consecuencia es que los mensajes aparecen varias veces en los resultados de las búsquedas, y sólo el usuario puede darse cuenta de que se trata de una repetición. Naturalmente, el comportamiento ideal sería que los mensajes aparecieran sólo una vez en los resultados del buscador.

3 Origen de los problemas: pérdida de información

En el origen de estos problemas se encuentra una pérdida de información que se produce al convertir los mensajes archivados a HTML para su publicación en la web.

Los gestores más habituales de listas de correo (mailman, majordomo, sympa, etc.) generan un fichero en formato Mailbox (mbox) con los mensajes que han sido enviados a la lista. Otros programas independientes, como hypermail, monharc, pipermail..., se especializan en convertir el fichero Mailbox en un conjunto de páginas web estáticas. Los programas más sofisticados son capaces de generar índices complejos de los archivos (por fecha, por autor, por hilo...), con múltiples referencias cruzadas entre los mensajes en forma de hipervínculos (mensaje anterior, mensaje siguiente, etc.). Pero incluso en el mejor de los casos, esta información sólo es comprensible para el usuario, nunca para la máquina. En consecuencia, es imposible explotarla más allá de las formas previstas por el programa que ha generado los archivos.

Entre la información que se pierde en la publicación, se encuentra:

- El asunto del mensaje.
- El autor del mensaje.
- La fecha del mensaje.
- La referencia a la lista de correo en la que se publicó el mensaje.
- La referencia al mensaje anterior, si existe.
- Las referencias (enlaces) a las posibles respuestas al mensaje.

4 Propuesta para conservar la información

Las tecnologías de la web semántica (y concretamente, RDF) son perfectamente capaces de publicar en la web toda la información señalada en la sección anterior. Dado que la información ya existe en el origen, no es necesario ningún procedimiento manual para enriquecerla. Tan sólo debe considerarse un proceso de conversión que no desprecie la información, sino que la publique junto con los archivos en HTML. De esta forma, las listas de correo se introducirían en la web semántica.

5 Aplicaciones

Enriquecer semánticamente la publicación web de los archivos de las listas de correo abriría la puerta a nuevas aplicaciones:

- Eliminar la aparición repetida de los mismos mensajes en los resultados de los buscadores. Para lograrlo, los buscadores deberían procesar la información semántica para reconocer las copias (mirrors) de los archivos.
- Implementar en los navegadores nuevas funcionalidades, como las apuntadas en la sección 2. Estas capacidades, que mejorarían sensiblemente la comodidad en la consulta de los archivos, podrían añadirse como extensiones o plug-ins de los navegadores actuales.
- Obtener información sobre los suscriptores de una lista de correo. Por ejemplo, conocer en qué otras listas de correo participa una persona. Esta aplicación es especialmente interesante en conexión con FOAF[1]. De este modo, se podría sacar una "orla" con las fotos de los participantes en una lista de correo[2], o situarlos geográficamente en un mapa[3].

[1] http://www.foaf-project.org/

[2] Como hace GNOME, véase http://planet.gnome.org/heads/

[3] Como hace Debian, véase http://www.debian.org/devel/developers.loc

- Facilitar la internacionalización. Al hacer comprensibles las relaciones entre los mensajes por el software, el navegador proporcionaría las opciones de exploración (mensaje siguiente, mensaje anterior, etc.) en el idioma del usuario, independientemente del idioma en el que se encontrasen las páginas HTML.

- Mejorar la accesibilidad de la información. Las tecnologías de accesibilidad podrían informar sobre quién es el autor del mensaje o cuántas respuestas hay, usando la voz u otros medios.

6 Trabajos relacionados

Existen algunos trabajos similares a esta propuesta.

El proyecto DOAML[4] consiste en un vocabulario RDF para describir listas de correo. Como ejemplo, en la web del proyecto se encuentran las descripciones de las listas de correo del W3C. La información de este vocabulario limita sus referencias a los mensajes archivados a un enlace a la versión HTML de éstos.

Por otro lado, EMiR[5]es un esquema RDF para describir mensajes de correo electrónico. En la misma línea se encuentra XMTP[6].

7 Conclusiones

Introducir los archivos de las listas de correo en la web semántica sólo requiere disponer de una aplicación de publicación que utilice la tecnología apropiada (RDF) como complemento al HTML. Con un mínimo esfuerzo, los administradores de todas las listas de correo podrían emplear la aplicación en sus listas, por lo que la implantación sería rápida[7]. Además, al no requerirse la participación de un experto para el enriquecimiento de la información, resultaría posible enriquecer inmediatamente grandes volúmenes de información, incluso listas de correo que lleven muchos años en funcionamiento.

El desarrollo de una aplicación de estas características requeriría, en primer lugar, la creación de un esquema de información, que muy bien podría ser una combinación de los ya existentes (véase la sección 6); y en segundo lugar, el procesamiento de un fichero Mbox para extraer la información que contiene. Dado que existen aplicaciones de software libre que realizan la segunda tarea, lo más razonable parece ser adaptar alguna de ellas.

[4] http://www.doaml.net/

[5] http://xmlns.filsa.org/emir/

[6] http://www.openhealth.org/xmtp/

[7] En realidad, cualquier suscriptor (no necesariamente el administrador) de una lista de correo podría publicar los archivos enriquecidos. Tan sólo debería disponer de todos los mensajes antiguos almacenados en su cliente de correo electrónico, y exportarlos al formato Mbox.