La información de los proveedores de servicios de Internet en la estadística sobre usos de TIC: Avance de resultados del proyecto europeo Diastasis

Vicenç Meléndez, Idescat

Resumen

El desarrollo de Internet es tan rápido que en solamente tres años desde la aparición del WWW alcanzó los 50 millones de usuarios, superando la implantación de otros sistemas y tecnologías como el ordenador, la televisión, el teléfono, etc. La estadística oficial ha respondido desarrollando un sistema de estadísticas para dar cuenta del fenómeno con el retraso previsible teniendo en cuenta la necesidad de estandarizar conceptos y definiciones para poder comparar. Sin embargo, las demandas de los usuarios siempre van por delante y no se satisfacen completamente; por eso, una vez existentes ya los datos oficiales de disponibilidad de Internet y algunas medidas sobre su uso, los investigadores especulan ahora sobre el impacto que tiene, exigiendo datos más detallados que los recogidos actualmente por las oficinas de estadística. El Idescat participa desde diciembre de 2002 en un proyecto de UE llamado Diastasis que tiene como objetivo obtener más detalle que la estadística actual de Internet aprovechando los datos digitales brutos disponibles en los ficheros log de los proveedores de servicios Internet (ISPs). Usando técnicas estadísticas apropiadas, es posible combinar estos datos con los microdatos actuales que provienen de las encuestas estándar sobre uso de las TIC en la población en general. Un sistema reducido de variables comunes a ambos conjuntos de datos, caracterizando a grupos de usuarios, permite producir algunas respuestas virtuales. Esas respuestas pueden ser agregadas como si fueran preguntadas a los individuos en una encuesta sobre la misma base de población. Se podría de esta manera verificar y ampliar las respuestas de los usuarios y, adicionalmente, obtener datos técnicos que estos desconocen y no depender exclusivamente de la memoria de los entrevistados. Además, la metodología usada permitiría, también, la combinación de conjuntos independientes de datos sobre una misma población, para los que no se ha previsto a priori que fueran a relacionarse, como por ejemplo determinados hechos sociológicos y el uso de Internet.

La información de los proveedores de servicios de Internet en la estadística sobre usos de TIC: Avance de resultados del proyecto europeo Diastasis

<u>Índice</u>

1.- Objetivo e integrantes del proyecto

El proyecto Diastasis tiene por objetivo analizar el aprovechamiento de la información en soporte digital disponible en ficheros log - o ficheros registro - de conexiones de los usuarios de los Proveedores de Servicios de Internet (ISPs) para complementar las encuestas que se llevan a cabo a la población de usuarios de Internet, facilitando un mayor detalle y más precisión en los indicadores y resultados estadísticos oficiales actualmente vigentes sobre usos de TIC. Con este fin, se aplican técnicas de minería de datos para indizar y sistematizar la información digital disponible y técnicas estadísticas de fusión de datos que permiten combinar ambas fuentes d'información.
El proyecto contempla el desarrollo de nuevas herramientas informáticas que incorporen las metodologías y procedimientos estadísticos, las cuales se evaluarán mediante una prueba piloto basada en datos de la comunidad académica y profesional de la Universitat Politècnica de Catalunya y de su proveedor de servicios de Internet.
La duración del proyecto es de 26 meses, de diciembre de 2002 a enero 2005, con un presupuesto de 1.973.965 euros y recibe financiación del Information Society Technologies programme, dentro del V Programa Marco de Investigación de la UE.
Los participantes en el consorcio que lleva a término el proyecto son: European Dynamics – Grecia -, Fraunhofer-Institut für Autonome Intelligence Systeme – Alemania -, junto con la Universitat Politècnica de Catalunya (UPC), el Institut d'Estadística de Catalunya (IDESCAT) y UPCnet.
La contribución del Idescat se centra principalmente en la fase de Análisis y Diseño del proyecto (workpackage 2), en especial con respecto a: Requerimientos de los Usuarios, Estudio de las Fuentes de Información existentes sobre Internet y Aspectos Legales aplicables a la utilización de los datos en soporte digital referente a los accesos a Internet por parte de los usuarios.
Adicionalmente, el Idescat aporta su experiencia para la realización de una encuesta al personal de la UPC incluida en la prueba piloto y efectúa una valoración de conjunto de los resultados del proyecto.

2.- Justificación

Internet se considera la tecnología que más rápidamente se ha desarrollado, alcanzando más de 50 millones de usuarios en apenas tres años desde la adopción del estándar de WWW. Para la estadística oficial, esto supone la necesidad de nuevas metodologías para obtener potenciales indicadores del uso del Internet y poder hacer frente a estos nuevos fenómenos.

Durante los primeros años de Internet, se utilizó un modelo de recogida de datos estadísticos que primero identificó la presencia de las tecnologías de información y de la comunicación (ICT) y que posteriormente se amplió para medir su grado de uso. Hoy, el modelo ha llegado a ser obsoleto puesto que la necesidad real es determinar los impactos que causa la tecnología. Ello requiere, en primer lugar, medir el uso actual del Internet con un detalle considerable. La utilización de la cantidad enorme de información de los usuarios del Internet, disponible en los ficheros log, o de registro, de los ISPs, que propone Diastasis, podría ser la estrategia más útil para alcanzar el grado necesario de detalle, aunque otras estrategias están también disponibles. Los paneles de usuarios del sector privado persiguen, igualmente, conocer mejor qué está sucediendo, lo mismo que la encuestas con cuestionarios muy extensos para relacionar Internet con muy diversas actividades humanas.

La metodología de Diastasis al ser capaz de unir conjuntos de datos oficiales y de otras fuentes que tienen mucho detalle y mucha extensión, ofrece la estrategia más flexible puesto que carece de las limitaciones en cuanto a volumen de usuarios, de los paneles y por otra parte permite relacionar, a posteriori, Internet con otras realidades y hechos sociales no previstos a priori. Un reto de Diastasis es verificar la eficacia de las metodologías estadísticas en que se basa, referentes al desarrollo de los mecanismos para relacionar la información de las encuestas oficiales actuales con otras, mediante un grupo común de variables.

Otra de las preocupaciones importantes dentro del proyecto DIASTASIS se refiere a la protección de datos personales utilizados, procedentes de la recogida y almacenaje subsiguiente en los ficheros log. Se analiza el marco jurídico existente con respecto a la protección de individuos por lo que respecta al proceso de sus datos personales, obtenidos con los procedimientos legales habituales en la estadística oficial y se investigan las implicaciones de Diastasis con respecto a esta cuestión.

DIASTASIS tiene como objetivo, además, la mejora de las encuestas actuales relativas al uso de las TIC en los hogares, definiendo indicadores nuevos, o mejorando los existentes mediante las metodologías desarrolladas. Se sitúa, por tanto, dentro del contexto de los indicadores e-Europa 2005. Dadas las características de la población que se analizará en la prueba piloto, los relacionados con la educación serían el objeto predominante del estudio.

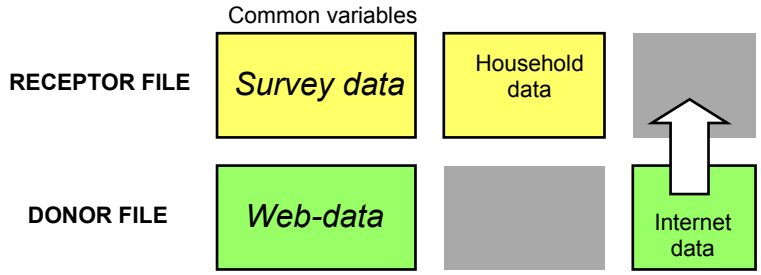
3.- Descripción técnica del proyecto

3.1.-Técnicas estadísticas usadas

Los aspectos metodológicos del proyecto corren a cargo del departamento de Estadística de la UPC, por lo que respecta a la metodología estadística y de Fraunhofer Institute für Autonome Intelligente Systeme (FhG/AIS), para los aspectos de minería de datos en páginas web (ver referencias metodológicas citadas).

File grafting in Diastasis

Fusion of households surveys with Internet data



Objective: To produce a representative file simulating complete real data: survey data (actual) with web data (imputed).

Gráfico 1 Esquema de asignación de variables a los microdatos de la encuesta oficial (gráfico producido por Tomàs Aluja)

Como puede observarse en el gráfico 1, anterior, los datos del conjunto de registros donante – el de los usuarios de Internet reflejados en los ficheros log – aportará campos o completará los registros de la encuesta oficial de uso de las TIC en hogares por lo que respecta al uso de Internet.

El nexo entre ambos conjuntos de datos se establece a partir de variables comunes que hay que identificar, tratando de que sea un conjunto mínimo y con alto valor predictivo para las variables específicas que hay que producir en el conjunto receptor de datos. Dicha identificación se lleva a cabo analizando los datos de la encuesta de uso de TIC en los hogares, receptora de datos. La metodología estadística que se emplea para efectuar la transferencia de datos de uno a otro conjunto es la denominada ***hot deck***, método muy conocido que se basa en la idea simple de que un individuo se comporta como sus similares. En concreto el método usado es el k NN o k nearest neighbors, en el cual los registros donantes y receptores se sitúan en el mismo subespacio de referencia definido por las variables comunes y para los cuales se han creado unos

grupos o clusters que agrupan a los más comunes entre sí. De este modo para cada registro receptor se determina un conjunto de donantes parecidos. La forma en que se imputan los valores, inexistentes en el conjunto de datos receptor, es mediante selección estocástica. Los resultados así obtenidos no presentan incoherencias, y mantienen la variabilidad y la correlación entre variables comunes y específicas, por el contrario, ofrecen menor precisión que otros.
No es imprescindible que los datos que se obtienen de la fuente que aporta información provengan de una muestra aleatoria, es decir, no se necesita que el fichero de donantes con la información registrada en los logfiles tenga carácter aleatorio.

Posible variables comunes para la población general
Tipo de conexión o amplitud de banda.*
Frecuencia de uso semanal.*
Promedio de tiempo de uso por día (en horas).*
Parte principal del día en cuanto a conexión (noche, fin semana,…).*
Conexión doméstica o en trabajo.
Promedio de e-mails por día.*
Años de uso de Internet.
Principal uso del ordenador (uso avanzado, de oficina o para ocio,…).
Edad
Ocupación /profesión.
Nivel de educación.
Sexo
Frecuencia de uso del móvil (por día) .
Presencia de hijos (entre 10 y 14 años)

Posibles variables comunes para el personal de la UPC
Ancho de banda
Frecuencia de conexión
Edad
Tipología de profesor/trabajo administrativo
Departamento o centro
Años de uso de Internet
Años en el puesto de trabajo de profesor o en trabajo administrativo
Número promedio de e-mails por día
Tiempo promedio de conexión por día Uso para trabajo científico Uso para transacciones …

Gráfico 2

(*) Preguntas que pueden obtenerse directamente del ISP

3.2.- Proceso de ficheros log de registro de peticiones http de los ISP

La actividad de los usuarios de Internet genera una cantidad de datos enorme que puede ser analizada con métodos estadísticos. Las fuentes de información usadas son normalmente los ficheros log o registros de acceso a las páginas web que contienen secuencias ordenadas cronológicamente de peticiones http que emite un navegador.

```
Sat Feb 1 00:51:47 2003 010.010.010.043 http://arc6.msn.com/ADSAdClient31.dll?
Sat Feb 1 00:51:47 2003 010.010.010.043 http://svcs.microsoft.com/svcs/mms/ads.asp?
Sat Feb 1 00:53:06 2003 010.010.010.043 http://www.msn.es/40231/ULI-e0a57f63.gif
Mon Feb 3 14:22:37 2003 010.010.010.016 http://www.bcn.es/imatges/flecha3azul.gif
Mon Feb 3 14:22:37 2003 010.010.010.016 http://www.bcn.es/imatges/punto.gif
```

Gráfico 3, Ejemplos de peticiones http (la descarga de un solo documento puede suponer varias peticiones http)

La minería de contenidos web es el proceso de extraer conocimiento de los documentos. Pude producirse un análisis numérico de diferentes aspectos contenidos o bien clasificar el texto según categorías existentes o agrupar las páginas que resulten similares en clusters. Con carácter previo y dado que en los contenidos de las páginas web están en distintas lenguas, hay que identificarlas puesto que los clasificadores se desarrollan para cada lengua independientemente.

El objetivo de la minería de texto de Diastasis es asociar las URL de los ficheros log con categorías textuales significativas. Clasificar texto requiere un conjunto de páginas de “entrenamiento” clasificadas con las categorías deseadas, las cuales pueden sacarse las páginas indizadas manualmente como DMOZ o Yahoo. Para cada categoría se crea un clasificador que una vez entrenado puede clasificar, con una rapidez de 50 documentos por segundo, documentos nuevos. Es necesario un preproceso de las páginas a clasificar que se inicia ya con la descarga de la parte documental a que hace referencia la URL.

3.3.-Operación piloto

Por razones de disponibilidad de datos y de posibilidad de contrastar los resultados obtenidos mediante las técnicas en estudio con los reales, la operación piloto se llevará a cabo sobre un a población específica: el personal de la UPC (ver gráfico 4). UPCnet es el responsable de la captura de los datos de uso de Internet: URL pedido y fecha y hora de petición. Posteriormente se recupera la página descargada para su análisis, previo proceso de adecuación; como resultado de dicho análisis FhG asignará a cada petición de documento una probabilidad de asignación a una de las 200 categorías de Yahoo que se estudian. Al mismo tiempo se procederá, por parte de Idescat a efectuar una encuesta al personal de la UPC al cual, se solicitará permiso para su seguimiento por Internet, en cumplimiento de la Ley de protección de datos. Dicha encuesta estará implantada en un formulario en una página web que realizará UPCnet. UPC llevará a cabo el estudio de la fusión de datos mediante el cual se generará un fichero de microdatos que tendrá una parte real – los datos de la encuesta - y otra virtual - proveniente de los usos de Internet registrados en los ficheros log -. ED aportará la informatización de las técnicas estadísticas usadas.

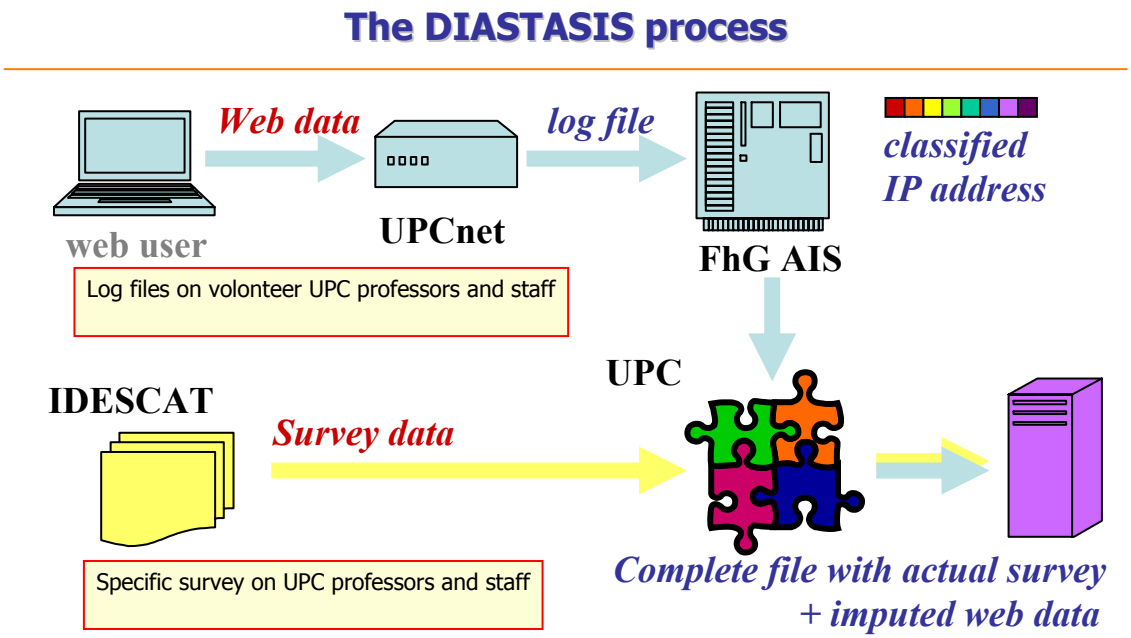


Gráfico 4 (producido por Tomàs Aluja)

4.- Posibles aportaciones de Diastasis a la estadística oficial, factibilidad

Mediante las aportaciones de Diastasis - y en función de la logística de recogida de datos de los usuarios de ISPs, aplicada - la encuesta de uso de tecnologías de la información y las comunicaciones en los hogares podría mejorarse, en general y especialmente en la parte correspondiente al uso individual. Podrían verificarse y ampliarse los resultados obtenidos de uso de Internet por edad, nivel de educación, ocupación y sexo y en general las variables comunes.
La naturaleza de las actividades en Internet que se preguntan en la encuesta, mejorarían en detalle.

- Conocimiento de la ocurrencia real de recepción o envío de e-mails/telefonía sobre Internet/Videoconferencia
- Navegación en la red en relación a escuchar la radio/Ver al televisión/Jugar/Descargar música/Leer/Descargar periódicos
- Interactuar con entidades bancarias/Comprar bienes/Vender bienes (subastas)
- Obtener información de la administración/Descargar formularios oficiales/Rellenar formularios oficiales
- Tener acceso a servicios educativos o de formación y aprendizaje
- Buscar información de temas de salud
- Tipo de bienes o servicios que se busca: compra de productos alimenticios y para la casa, películas, música, libros, revistas, material para formación, ropa, artículos deportivos, compra de valores mobiliarios, adquisición de servicios financieros, de seguros, de turismo y viajes, etc.

El detalle por actividades y tipo de usuarios resultaría muy mejorado, puesto que se tendrían muchos más datos de ellos y su número sería mayor que el de las encuestas, en donde hay usuarios y no usuarios de Internet. El detalle geográfico referente a dichos usuarios, también mejoraría. El intervalo temporal abarcado, sería, sin embargo, más limitado que el de las encuestas.
La periodicidad general de acceso puede obtenerse directamente del ISP, aunque a nivel de password, es decir, sin relación con las variables comunes. El número de horas en conexión con el ISP, también puede proporcionarlo este. Las horas de uso y la periodicidad en diferentes estratos de usuarios puede deducirse por su representación dentro de los datos recogidos por el ISP, en relación con el mismo estrato entre los usuarios de Internet en la encuesta.
El destino de las peticiones, número de páginas y el volumen de información transferido puede obtenerse en general y en relación con los usuarios monitorizados. Las lenguas de las páginas y de la información transferida pude asimismo comprobarse igualmente en ambos casos.
Se dispondría de mayor confidencialidad para obtener los mismos datos, puesto que los encuestados no deberán responder determinadas preguntas y habría una menor carga de trabajo para la población en respuesta. Aunque por

otro lado el acceso a los ficheros log debería ser reglamentado para proteger la intimidad de los usuarios.

La realización de la operación para el público en general, requiere resolver aspectos logísticos que dan pie a la selección de individuos para inquirir las variables comunes y que deben resolverse actuando conjuntamente con los ISPs. La participación o ganas de colaborar tanto de estos como de las personas seleccionadas, según las opiniones vertidas en los informes de revisión efectuados por expertos seleccionados por parte de los miembros del consorcio Diastasis, constituye un punto clave y a la vez un punto débil del proyecto. La participación está relacionada con la credibilidad que el ciudadano confiere a la estadística oficial y subsiste aunque se trate de una estadística declarada oficial y por tanto de respuesta obligatoria. La exhibición de unas normas de protección de los datos personales especialmente adecuadas podría ser un instrumento para vencer reticencias por parte de la opinión pública.

Referencias

Tomàs Aluja, UPC. **Data Fusion Methodologies**. 8 pag.

Gerhard Paass y Ingo Schwab, FhG/AIS. **Web Mining and Data Fusion Methodologies**. 11 pag.