Extracción automática de fichas de recursos turísticos de la web

Iker Manterola[1], Xabier Saralegi[1], Sonia Bilbao[2]

[1]Elhuyar I+D
[2]Tecnalia

Resumen

Los recursos turísticos básicos son una pieza clave que permite esbozar artefactos más complejos como pueden ser los productos experienciales. Además de la labor de un correcto diseño, un aspecto fundamental para el fácil desarrollo de productos experienciales es la disposición de una amplia paleta de recursos básicos semilla que ofrezcan mayor libertad en el diseño. Este trabajo presenta una metodología que permite ampliar la cobertura de estos recursos semilla. La estrategia propuesta consiste en acudir a la web para ampliar de manera automática la cobertura de un conjunto de recursos o fichas iniciales. La tarea no es sencilla ya que estas fichas pueden estar contenidas en diferentes webs, y además su contenido puede estar estructurado de manera variable. Para hacer frente a estos problemas proponemos una estrategia que incluye dos pasos. Inicialmente se identifican las webs susceptibles a incluir fichas, y posteriormente se inducen wrappers o extractores de fichas sobre el conjunto de webs obtenidas. En ambos pasos el proceso es totalmente automático y se utilizan las fichas semilla a modo de ejemplo de entrada. Los resultados muestran que es posible incrementar automáticamente la cobertura de los recursos iniciales hasta en un 55%.

Palabras clave: eTurismo, Extracción de Información, Crawling, Wrapper

1. Introducción

Dentro del sector turístico las empresas receptivas ofrecen propuestas innovadoras que aglutinan diferentes recursos turísticos básicos como pueden ser eventos, actividades, etc. Estas ofertas o recursos, que en principio tienen una consistencia limitada, pueden fortalecerse e integrarse en productos experienciales de forma que se favorezca la satisfacción de las necesidades de los viajeros y visitantes.

El proyecto Tourexp[1] tiene como objetivo crear los primeros sistemas de distribución en ruta, y las condiciones necesarias para hacer realidad un modelo de B2B2C en Euskadi. El trabajo presentado se enmarca en dicho proyecto, pero se limita únicamente a un problema concreto dentro de esa compleja tarea. Concretamente, se describe una metodología para aumentar la cobertura de la paleta básica de recursos de la que disponen los receptivos a la hora de diseñar productos experienciales. Al fin y al cabo, un aspecto fundamental para el fácil desarrollo de productos experienciales es la disposición de una amplia paleta de recursos básicos que ofrezcan mayor libertad en el diseño.

En este trabajo se propone acudir a la web para ampliar la cobertura de un conjunto de recursos o fichas iniciales. La web se muestra como un repositorio actualizado de la mayor parte de fichas de recursos turísticos. En nuestro caso el recurso inicial lo compone la base de datos de *Open Data Euskadi* (*ODE*), que incluye información sobre recursos turísticos del País Vasco tales como restaurantes, alojamientos, etc. El objetivo consiste en mejorar la cobertura de las fichas correspondientes a ese tipo de recursos acudiendo a la web. La tarea no es trivial ya que la disposición de estas fichas así como la estructuración de las mismas no siguen un patrón regular. Así, estas fichas pueden estar contenidas en diferentes webs y su contenido puede estar estructurado según diferentes formatos. Para hacer frente a estos problemas proponemos dos estrategias que aplicamos conjuntamente. Primero haremos una selección de webs susceptibles a incluir fichas, y en un segundo paso generaremos *wrappers* de manera automática.

2. Descripción del Estado del Arte

La obtención automática de datos desde la WWW implica la necesidad de convertir el contenido web a estructuras más orientadas a los objetos que se desean identificar. En la mayoría de los casos el contenido web suele estar estructurado en HTML, un lenguaje más apropiado para la maquetación o visualización que para la estructuración de datos. Esta

[1] http://www.tourexp.es/

circunstancia dificulta la tarea de extracción de datos estructurados ya que la estructura se funde con los elementos orientados a la presentación. Por ese motivo surge la necesidad de las herramientas conocidas como *wrappers*, capaces de extraer datos concretos y estructurados de los contenidos web. Los *wrappers* suelen basarse en patrones que localizan estructuras de elementos HTML que coincidan con la información predefinida por el usuario. Esta información suele segmentarse en atributos y sus respectivos valores. Crear y adecuar estos patrones manualmente para cada caso y web es una tarea ardua. Por esa razón, muchos autores han propuesto métodos para inducir estos *wrappers* de manera automática o semiautomática (Kushmerick et al., 1997; Chang et al., 2006).

Los *wrappers* pueden clasificarse según el nivel de supervisión con el que se generan. Los *wrappers* supervisados (Freitag, 1998) se inducen a partir de ejemplos etiquetados a mano donde se determina cual es la información que se desea extraer. Otro tipo de *wrapper* puede ser entrenado con muestras sin etiquetar, es decir, partiendo de una muestra cruda donde sí existe el tipo de información que se desea extraer, pero no se concreta cual es. Este tipo de *wrappers* reciben el nombre semi-supervisados (Chang y Lui, 2001). Una vez extraídos los posibles patrones, es el usuario quien decide y determina qué patrones son los correctos y qué tipo de información se va a extraer mediante su uso. De esta forma se reduce el coste del trabajo manual. Por último, existe otra clase de *wrappers* que son capaces de extraer información sin ninguna supervisión (Crescenzi et al. 2001). Para ello, comparan varias páginas de un mismo dominio con el fin de identificar las estructuras de las partes de texto que no son comunes en todas la páginas analizadas. De esta forma, logran identificar las estructuras que contienen texto variable. Como consecuencia de la menor precisión pueden surgir casos ambiguos donde la desambiguación manual resulta necesaria.

Las técnicas de extracción de información desde la Web también son aplicables en el dominio turístico, ya que mediante ellas se pueden crear nuevos recursos y/o aplicaciones. Por ejemplo, el sistema resultante del proyecto TIScover (Pröll y Retschitzegger, 2000) ofrece al usuario información detallada y actualizada sobre productos turísticos, facilitando a su vez la posibilidad de comprar o consumir dichos productos. TIScover se apoya en la herramienta MIRO-Web (Haller et al., 2000), que es capaz de reunir y estructurar información obtenida desde distintas fuentes previamente definidas (desde bases de datos estructuradas hasta páginas extraídas desde dominios previamente seleccionados de la red). Otro sistema capaz de extraer y utilizar información turística desde la red es MOMIS (Beneventano y

Bergamaschi, 2004). Este sistema extrae la información desde distintos dominios previamente identificados, y posteriormente aplica varios recursos semánticos para poder crear una estructura de datos basada en un esquema global. Por último, mencionaremos SESAMO (Walchhofer et al., 2010). De la misma manera que las anteriores, este sistema comienza extrayendo la información necesaria usando *wrappers* desde dominios turísticos ya identificados, para posteriormente unificar los datos extraídos y ofrecer una completa monitorización de la información requerida.

3. Open Data Euskadi

Open Data Euskadi es una base de datos públicos ofrecidos por el Gobierno Vasco, y que pretende facilitar la interoperabilidad. Entre sus objetivos destacan:

- **Generar valor y riqueza**: Obteniendo productos derivados de los datos por parte de terceros.
- **Generar transparencia**: Reutilizando los datos para analizar y evaluar la gestión pública.
- **Facilitar la interoperabilidad entre administraciones:** Creando servicios que utilicen datos de diferentes administraciones.

Algunos de los datos que ofrece pertenecen al área del turismo (ver tabla 1). La motivación para usar ODE como punto de partida se debe a que se trata de una fuente de datos en formato abierto, estándar y estructurado, que utiliza esquemas y vocabularios consensuados, actualizados, libremente accesibles y reutilizables.

Tipo recurso	**#**	**atributos**
restaurante	578	cp, mail, tipo, provincia, teléfono, foto, persona contacto, localidad, municipio, url, fax
alojamiento	777	cp, teléfono, localidad, fotom fax, nombre, categoría, provincia, mail, persona contacto, url, municipio
patrimonio	720	cp, mail, fax, teléfono, url, provincia, municipio, localidad

Tabla 1. Tipos de recursos incluidos en Open Data Euskadi

4. Arquitectura propuesta

Con el objetivo de enriquecer la base de datos ODE, y teniendo en cuenta los distintos procesos que han de llevarse a cabo para ello, la arquitectura del sistema (ver Figura 1) que presentamos abarcas las siguientes etapas:

1.***Detección de webs candidatas**:* El primer paso para enriquecer la base de datos ODE consiste en obtener fuentes que puedan contener información turística en la red mediante técnicas de *crawling*. Los buscadores de internet son un buen recurso para realizar el *crawling* ya que tienen acceso a todo el contenido de la red (Boley et al., 1999). Partiendo de la información de las fichas existentes en la base de datos ODE, es posible crear consultas para estos buscadores con el fin de conseguir páginas con contenido turístico extraíble.

2.***Inducción de wrapper**:* Como se ha expuesto en el estado del arte, la información requerida está incrustada en estructuras HTML más orientadas a los usuarios que al tratamiento automático. Por ello, es necesario crear o inducir un *wrapper* capaz de extraer y estructurar la información deseada de cada sitio web candidato. El objetivo será reducir al mínimo el trabajo manual. Como en el proceso anterior, la utilización de recursos ya existentes de la base de datos ODE resulta de gran utilidad.

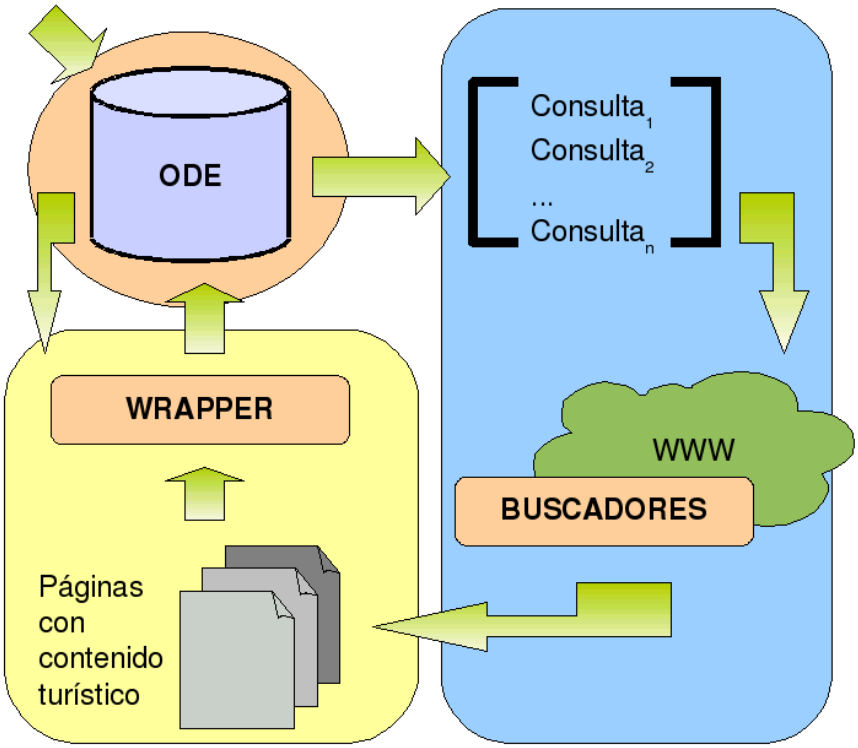

Figura 1. Esquema de la arquitectura propuesta

4.1 Detección de webs candidatas

El método propuesto se basa en el uso de la API proporcionada por el buscador web *Bing*. Mediante consultas construidas siguiendo el procedimiento que explicamos a continuación, localizamos las webs candidatas a contener información adecuada para crear o enriquecer fichas turísticas. Proponemos construir la consulta en base a los resultados que se deseen obtener, es decir con la idea de obtener documentos relevantes a contener fichas turísticas. En nuestro caso, los valores para construir la consulta son los incluidos en las fichas turísticas semilla correspondientes a la base de datos ODE. La consulta $Q=\{F_1(v_1)...F_1(v_i)...F_1(v_n)\ \ldots\ F_j(v_1)...F_j(v_i)...F_j(v_n)\ \ldots\ F_m(v_1)...F_m(v_i)...F_m(v_n)\}$ se construye a partir de valores v correspondientes a n atributos incluidos en m fichas semilla $F_i=\{(a_i,v_i)\}$. Esta consulta no es manejable para los buscadores web comerciales debido a su tamaño y la imposibilidad de utilizar ciertos parámetros lógicos. Por esa razón, dividimos y lanzamos la consulta Q en k consultas Q_i, que contienen m fichas F_i de la consulta Q escogidas aleatoriamente. Posteriormente, agregamos todos los resultados obtenidos por cada consulta. Por ejemplo, si queremos obtener webs que contengan fichas de restaurantes de Guipúzcoa, la consulta Q puede ser creada a partir del nombre y la dirección de fichas de restaurantes situados en Guipúzcoa extraídas de la base de datos ODE:

Q={"Asador Urkiola" "Mayor, 7" "Gran Sol" "San Pedro, 63" "Juanito Kojua" "Puerto, 14" "Altzueta" "Bº Osinaga, 7" "Eula" "Barrio Lategi, 19" ...}

Antes de lanzar la consulta al API del buscador web creamos las subconsultas Q_i:

Q_1={"Asador Urkiola" "Mayor, 7" "Altzueta" "Bº Osinaga, 7"}

...

Q_k={"Juanito Kojua" "Puerto, 14" "Gran Sol" "San Pedro, 63"}

Lanzamos las subconsultas al API y agregamos los diferentes rankings de resultados que corresponderán en su mayor parte a webs que contengan información de más de un recurso turístico a la vez (normalmente páginas de guías o listas de distintos recursos turísticos). De esta forma, el trabajo manual de detección de webs candidatas se reduce al mínimo ya que el proceso es totalmente automático. Las páginas obtenidas en la búsqueda serán descargadas

para tratarlas en la siguiente fase. Cabe destacar que este proceso puede ser aplicado e manera iterativa para mejorar los resultados en términos de precisión y cobertura, ya que cuanto mayor sea la base de datos ODE, mejores serán los resultados obtenidos mediante este proceso.

4.2 Inducción de *wrappers*

Por cada página detectada en la fase anterior se crea un *wrapper* capaz de identificar las estructuras donde aparezcan los atributos y sus respectivos valores de nuevas fichas turísticas. Hay que tener en cuenta que cada página a tratar puede tener una estructura totalmente distinta, por lo que las estructuras a identificar cambiarán para cada página. Por ese motivo cada una debe ser tratada independientemente.

Como en la fase correspondiente a la detección de webs candidatas, proponemos usar información de fichas ya existentes de la base de datos ODE a modo de ejemplo. De esa forma, podemos inducir los patrones de las estructuras donde aparecen los valores de los atributos que buscamos. El algoritmo utilizado se basa en el funcionamiento de los *wrappers* supervisados pero con datos incompletos. Se exige que los patrones detectados se repitan con cierta frecuencia y se utilizan los valores de fichas turísticas extraídas desde la base de datos ODE como restricción. La búsqueda y extracción de las estructuras se lleva a cabo de la siguiente manera:

1) Partiendo de las fichas semilla de la base de datos ODE, se buscan los valores ya conocidos de los atributos que se deseen obtener desde la página candidata. Como resultado, se obtiene un grupo de expresiones que representan la estructura HTML en la que está contenida la información deseada. Por ejemplo, si el atributo que buscamos es el nombre de un restaurante y su valor es *"Bereziartua Sagardotegia"*, según la página HTML mostrada en la Figura 2, la expresión que obtendremos será *"html|body|div|b"*. Si las búsquedas son combinadas, es decir, si el objetivo es extraer por ejemplo el nombre y el teléfono de un restaurante (*"Bereziartua Sagardotegia"* y *"943555798"*), se obtendrán pares de expresiones, que determinarán la completa estructura de la información que buscamos: [*"html|body|div|b"* , *"html|body|div|i"*]. En este caso concreto, la estructura del tipo *"html|body|div"* se extendería hasta las etiquetas *"<b>"* y *"<i>"*

```
<html>
...
<body>
...
<div align="center">
<b>BEREZIARTUA SAGARDOTEGIA</b>
<i>943 55 57 98</i>
</div>
...
</body>
...
</html>
```

Figura 2. Ejemplo de web candidata

2) Partiendo del grupo de expresiones que obtenemos en el paso 1, contamos cuantas veces se repite cada estructura. Si la estructura analizada aparece con suficiente frecuencia, damos por correcta la estructura y pasamos al siguiente paso. Aplicamos esta validación basándonos en el alto porcentaje de guías o listas de recursos que se obtienen en la fase de detección de la webs candidatas.

3) Por cada expresión correcta, extraemos todos los textos que estén contenidos en la misma estructura. De esta forma, construimos nuevas fichas a partir de nuevos pares (valor, atributo) no incluidos en la base de datos ODE. Antes de introducir las nuevas fichas en la base de datos ODE, estas pueden revisarse manualmente con el fin de optimizar la precisión.

5. Evaluación

Con el objeto de evaluar la metodología propuesta se han realizado diversos experimentos para medir su rendimiento. Para ello, se ha propuesto la tarea de enriquecer las fichas pertenecientes a los restaurantes situados en Guipúzcoa. La evaluación está dividida en dos fases. Primero evaluamos la capacidad de la estrategia propuesta para identificar webs candidatas a contener información turística extraíble, y a continuación, evaluamos la funcionalidad de los *wrapper* generados sobre las páginas obtenidas en la primera fase.

Los resultados obtenidos en la fase de identificación de webs candidatas son evaluados en términos de precisión y cobertura. Definimos la cobertura como el nivel de productivad del método cuantificado por el número total de resultados que puede proporcionar, mientras que la precisión nos sirve para estimar en qué porcentaje son correctos dichos resultados.

Fijamos como *baseline* los resultados obtenidos lanzando una consulta creada manualmente. Para la construcción de consultas, nuestro método permite distintas variaciones según el

número de fichas y atributos a utilizar. Por cada variante, creamos 25 subconsultas según diferentes números de fichas *k* y número de atributos *m* por subconsulta. En todos los casos las fichas se seleccionan aleatoriamente desde la base de datos ODE. En el caso de los atributos nos limitamos a el nombre, la dirección, y el teléfono.

Construcción consulta	***#sitios web***	***Precisión***	***Cobertura***
Baseline	100	0,43	43
m=1, k=1	230	0,01	2,3
m=1, k=2	167	0,16	26,72
m=1, k=3	55	0,17	9,36
m=2, k=1	200	0,25	50
m=2, k=2	34	0,4	13,6
m=2, k=3	5	1	5

Tabla 2. Cobertura y precisión para la detección de sitios-web candidatos

Los resultados obtenidos muestran que cuanto más estricta es la consulta (es decir, cuanto más atributos y/o valores se consulten) menor es el número de webs candidatas que se obtienen. El número de atributos y el número de valores también influyen directamente tanto en la precisión como en la cobertura, ya que cuanto más rigurosos son, mayor es la precisión y menor la cobertura. Esta tendencia es más clara en los casos en los que la rigurosidad va marcada por el número de atributos usados. Como se puede observar en la Tabla 2, los resultados obtenidos mediante una consulta creada manualmente pueden ser superados cuando se introducen más de un atributo y más de un valor por cada atributo en la misma consulta.

En el caso del *wrapper* o extractor, evaluamos el método en términos de productividad, es decir, la cantidad de nuevas fichas extraídas. Para ello, partimos de una muestra que contiene todas las webs candidatas que hemos obtenido en la fase de detección. Consideramos que hemos obtenido una nueva ficha si como mínimo obtenemos el nombre y el teléfono o la dirección del recurso turístico. Todas las fichas obtenidas se revisan manualmente con el fin de verificar el cumplimiento de dicha condición. Los resultados (ver Tabla 3) muestran que es posible aumentar la cobertura de la base de datos inicial en un 55%. La revisión manual desvela que el número de fichas incorrectas es mínimo (%5). Las fichas con información repetida tambien son descartadas.

Tipo recurso	*ODE*	*Obtenidas*	*Nuevas*	*% Incremento*
Restaurantes	578	408	320	55%

Tabla 3. Aumento de cobertura de recursos de ODE

6. Conclusiones

Los experimentos realizados en este trabajo demuestran que es posible incrementar la paleta de recursos turísticos básicos mediante su extracción automática de la WWW. El método propuesto se apoya en recursos existentes que utiliza a modo de ejemplo para iteratívamente descubrir nuevos recursos. Debido a que el proceso parte de recursos ya existentes, el rendimiento del sistema aumenta cada vez que el propio sistema extrae nueva información, gracias a que el recurso inicial va incrementando en cantidad y calidad.

En el caso de la detección de webs con contenido turístico extraible, el sistema presentado es capaz de detectar nuevas páginas automáticamente con una precisión aceptable. Este es un factor clave ya que la detección manual de esta clase de recursos web puede resultar una tarea muy costosa, sobre todo si el tipo de recurso es variable (restaurantes, hoteles, vuelos, servicios, etc.). Por otro lado, los *wrappers* generados por el método presentado son capaces de extraer una gran cantidad de información nueva, que a pesar de no ser completa (a causa de que no toda la información requerida de las fichas turísticas está accesible en la red), es de gran utilidad para enriquecer la base de datos ODE.

El sistema presentado en este trabajo supone un primer paso hacia el desarrollo de un sistema totalmente automático capaz de enriquecer recursos turísticos, por eso aún hay aspectos que deben ser mejorados. El aspecto más crítico es la relación entre la precisión y la cobertura de la fase de detección, ya que el aumento de la precisión supone un gran descenso de la cobertura y viceversa. Por esa razón es totalmente necesario revisar y mejorar el sistema con el fin de equilibrar estas dos medidas de forma que el sistema sea más preciso y productivo al mismo tiempo. Por otro lado, es necesario realizar experimentos con otro tipo de recursos turísticos (alojamientos, patrimonio, ...) para estimar cómo varía el rendimiento del sistema dependiendo de distintos entornos de trabajo.

Para terminar, cabe destacar que el método utilizado sobre la base de datos ODE puede ser aplicado sobre otras bases de datos basadas en fichas, ya que la metodología que se usa no impone ninguna restricción sobre el tipo de información.

7. Reconocimientos

Este trabajo se enmarca dentro de Tourexp (ER-2010/00005), proyecto financiado por el Departamento de Industria, Innovación, Comercio y Turismo del Gobierno Vasco (programa ETORGAI 2010).

Bibliografía

Beneventano D. y Bergamaschi S. (2004). The MOMIS Methodology for Integrating Heterogeneous Data Sources. *IFIP International Federation for Information Processing, 2004, Volume 156/2004, 19-24*

Boley, D. Gini, M., Gross, R., Han, E., Hastings, K., Karypis, G., Kumar, V., Mobasher, B. y Moore, J. (1999). Document Categorization and Query Generation on the World Wide WebUsing WebACE. *En Artificial Intelligence Review - Special issue on data mining on the Interne, Volume 13 Issue 5-6, 365 – 391*

Chang, C., Kayed, M., Girgis, M. y Shaalan, K. (2006). A Survey of Web Information Extraction Systems. *En IEEE Transactions on Knowledge and Data Engineering. Volume 18 Issue 10, October 2006, 1411 – 1428*

Chang, C.-H. y Lui, S.-C., (2001). *IEPAD: Information extraction based on pattern discovery.* Proceedings of the Tenth International Conference on World Wide Web (WWW), Hong-Kong

Crescenzi, V., Mecca, G. y Merialdo, P. (2001). RoadRunner: towards-automatic data extraction from large Web sites. En *Proceedings of the 26th International Conference on Very Large Database Systems (VLDB), Rome, Italy, 109-118*

Freitag, D. (1998). *Information extraction from html: Application of a general learning approach.* En Proceedings of the 15th Conference on Artificial Intelligence (AAAI-98)

Haller, M., Pröll, B., Retschitzegger, W., Tjoa, A.M., Wagner, R.R. (2000). Integrating heterogeneous tourism information in TIScover: the MIRO-Web approach. *ENTER 2000: 7th. International Congress on Tourism and Communications Technologies in Tourism, Barcelona, Spain, 26-28 April 2000, 71-80.*

Kushmerick, N., Weld, D. y Doorenbos, R. (1997). *Wrapper induction for information extraction.* En Proceedings of the 15th International Joint Conference on Articial Intelligence (IJCAI) , Japan, 1997

Pröll, B. y Retschitzegger, W. (2000). Discovering Next Generation Tourism Information Systems: A Tour on TIScover. *Journal of Travel Research 2000; vol. 39; 182-191*

Walchhofer, N., Hronský, M., Pöttler, M., Baumgartner, R. y Fröschl, K. (2010). *En Semantic Online Tourism Market Monitoring.* Information and Communication Technologies in Tourism