CitEc, un agente para asistir en la recuperación de información basada en citas

José Manuel Barrueco
Universitat de València
Spain

Vicente Julián Inglada
Universidad Politécnica de Valencia
Spain

Resumen El problema del enlace automático de referencias ha sido objeto de numerosos trabajos en los últimos años. El empleo de las citas expresadas en documentos constituye una herramienta ideal para la recuperación de información relacionada a partir de un documento inicial. En este artículo presentamos una solución a la extracción y enlace automático de referencias bibliográficas en documentos electrónicos mediante el empleo de una agente software denominado CitEc, el cual, ha sido evaluado sobre el conjunto de documentos disponibles en la biblioteca digital RePEc.

1 Introducción

Un documento científico se puede caracterizar por los siguientes elementos [6]: Título, Autores, Resumen, Palabras clave, Texto y Referencias bibliográficas. Las referencias contiene la lista de todos los trabajos que han sido citados en el texto, enumerados bajo un formato abreviado establecido, que posibilita la identificación y localización del correspondiente trabajo. Constituyen un indicador del carácter social de la ciencia y una de las características que diferencian la literatura científica de otras representaciones literarias. Igual que no existe el investigador individual, tampoco puede existir un documento aislado. Un documento ocupa así un nodo en la inmensa red de la literatura científica, conectado a través de las referencias bibliográficas con otros nodos relacionados temáticamente con él.

Debido al papel preponderante que juegan las referencias en el sistema de comunicación de la Ciencia, han sido objeto de innumerables estudios desde disciplinas como la Bibliometría, la Sociología de la Ciencia o las Ciencias de la Documentación.

En este último caso, se han utilizado como un instrumento adicional para la recuperación de información. Dado que son representaciones de documentos relacionados temáticamente, constituyen un medio ideal para navegar por la literatura científica partiendo de un documento relevante para una necesidad de información determinada.

Esa navegación por la literatura ha sido tradicionalmente una ardua tarea. La localización de un artículo en la biblioteca o su solicitud por préstamo interbibliotecario si ésta carece de él es un proceso largo y tedioso con el que los investigadores están acostumbrados a convivir. Autores como [4] han ideado sistemas para solucionar esta problemática. Ahora bien, para que esos proyectos tengan éxito son necesarias dos condiciones: (i) que los documentos estén en formato electrónico y (ii) que existan herramientas, como los índices de citas, que faciliten la navegación por la red de citas.

La primera condición, documentos en formato electrónico ya se da. Con ellas ha surgido la posibilidad de tratar los textos automáticamente para extraer las referencias sin intervención humana. De esta forma, si un programa es capaz de determinar dónde se encuentra la bibliografía de un documento, individualizar cada referencia y diferenciar cuál es el título, los autores, la fuente donde ha sido publicado y cuál es su dirección electrónica, cabría la posibilidad de crear un enlace entre la referencia y el documento al que representa que nos permitiera ir de uno a otro a través del hiperespacio electrónico.

La posibilidad de enlazar una referencia con el texto completo del documento al que representa es lo que los anglosajones denominan **reference linking** o enlace de referencias.

Esta idea está siendo objeto de múltiples estudios en la actualidad. Existen iniciativas como ResearchIndex[1] que es capaz de construir verdaderos índices de citas de forma autónoma (autonomous citation indexing), o proyectos como el del Cornell Digital Library Research Group[2] que está trabajando con la extracción de citas desde entornos no homogéneos, etc.

El presente trabajo se inserta en esta línea de investigación que tiende a utilizar las referencias como un instrumento para la recuperación de la información. En esta comunicación describimos la estructura y funcionamiento de un agente para la extracción y enlace automático de referencias bibliográficas en documentos electrónicos. El paradigma de agentes constituye actualmente un área de creciente interés. Numerosas aplicaciones basadas en este nuevo paradigma vienen ya siendo empleadas en infinidad de áreas, dentro de las cuáles se encuentra el área de la Recuperación de Información [5].

El empleo de esta orientación permite automatizar de forma flexible los diferentes procesos en el problema de los enlaces de documentos e incorporar cierto grado de autonomía para poder actuar independientemente del usuario. El agente desarrollado, llamado CitEc, inicialmente se ha diseñado para actuar sobre un universo de datos cerrados, en concreto la biblioteca digital RePEc (Research Papers in Economics)[3], pero podría ser fácilmente ampliado para trabajar con otros tipos documentales.

La biblioteca digital RePEc ha sido descrita en trabajos como [1]. Básicamente se trata de una biblioteca digital distribuida que almacena descripciones bibliográficas de artículos publicados en revistas y documentos de trabajo en Economía. La mayoría de estas descripciones contienen un enlace al texto completo de los documentos que es almacenado fuera de RePEc. Es un sistema distribuido porque a él contribuyen instituciones de todo el mundo. Cada una aporta información sobre los documentos que produce o distribuye. La arquitectura de RePEc divide a las instituciones participantes en dos categorias que no son excluyentes. Los proveedores de datos y los servicios. Estos últimos toman la información de los proveedores para crear un servicio utilizable por los usuarios. Su misión es incluir algún valor añadido a los datos como por ejemplo un sistema de alertas, una base de datos o un sistema de revisión.

El resto del trabajo se estructura de la siguiente forma. La sección dos se dedica a la descripción de la arquitectura del agente CitEc. La sección tres pasa a analizar los resultados obtenidos a través de una serie de pruebas realizadas. Finalmente, en la sección cuatro se presentan una serie de conclusiones.

2 Arquitectura de CitEc

Tal y como se ha comentado en la introducción, CitEc es un agente que trata de operar de forma autónoma en el universo de RePEc para poder extraer y enlazar de forma automática las referencias bibliográficas. De esta forma podríamos decir que básicamente el agente realiza tres funciones: (i) obtiene y analiza documentos, (ii) extrae la información de citas y (iii) enlaza las referencias con el documento si éste está disponible. CitEc por tanto crea de forma automática una red de documentos para facilitar notablemente la labor del usuario.

Basándose en estas ideas, la arquitectura de CitEc consta de tres elementos fundamentales: el sistema de proceso, la base de conocimiento y la interface de comunicación.

2.1 El sistema de proceso

Constituye el núcleo del agente ya que se encarga de las distintas fases de análisis de los documentos. En general, el proceso de enlace de referencias se realiza en las siguientes etapas:

1. Conversión del documento original a un formato que pueda ser analizado. Por ejemplo de PDF a ASCII o de HTML a XML.
2. Análisis del documento para encontrar la lista de referencias. Seguidamente identificar cada referencia y los elementos que la componen. Además es necesario identificar el contexto, es decir, aislar la frase o frases con las cuales el autor se ha referido a esta obra en cuestión.
3. Comprobación de si la obra citada se encuentra disponible en formato electrónico. En caso afirmativo se realizará un enlace.

Cada una de estas tareas la lleva a cabo uno de los siguientes módulos del agente: lectura, análisis y enlace.

[1] http://csindex.org
[2] http://opcit.eprints.org
[3] http://www.repec.org

Lectura. Es el encargado de monitorizar el entorno de datos para detectar cambios. La entrada en el agente está formada por descripciones bibliográficas de documentos. El agente procesa estos datos y como resultado modificará el entorno de RePEc añadiendo un nuevo tipo de descripción llamado "citation template" [2]. En él se incluirá la lista de referencias con los elementos que las componen. Además se incluirá un enlace con el documento al que representa cada referencia si se encuentra disponible en RePEc.

Análisis. El módulo de análisis es el núcleo del sistema. Se encarga de tres tareas. Dado que RePEc sólo contiene información bibliográfica con enlaces al texto completo de los documentos, la primera operación es descargar los ficheros que contienen los documentos. En la versión actual, aquellos que no estén en formato PDF o PostScript son descartados. Esto no supone una gran limitación, dado que en torno al 94% de los documentos se hacen disponibles en alguno de estos formatos. En un segundo lugar, convierte los ficheros del formato original a ASCII para que el texto pueda ser analizado. Finalmente, la última tarea del módulo es analizar el fichero ASCII para buscar si existe una lista de referencias. En caso afirmativo trata de aislar las distintas referencias y por cada una de ellas trata de identificar elementos claves como el autor, título o fecha de publicación.

Enlace. Este módulo lee la información sobre referencias y por cada una, intenta determinar si el documento al que representa se encuentra disponible en RePEc. Para ello compara los elementos de la referencia con la base de conocimiento. Si encuentra una descripción lo suficientemente similar establece un enlace entre ambos. El enlace se materializa en una nueva línea que se añade a la tabla CITA de la base de conocimiento. Este módulo es también el encargado de modificar el entorno RePEc añadiendo nuevas descripciones a la biblioteca.

2.2 La base de conocimiento (BC)

El segundo elemento de CitEc es la base de conocimiento, que se encarga de mantener una representación interna del entorno RePEc. Como se describe en [1], RePEc está basado en una estructura relacional por lo tanto la BC de CitEc se ha implementado como una base de datos relacional utilizando el sistema de gestión MySQL.

Las relaciones más importantes que forman la base son: DOCUMENTO, contiene una ocurrencia por cada documento electrónico disponible en RePEc. Se incluye información básica de cada uno. CITA contiene una ocurrencia por cada referencia que representa a un documento disponible en RePEc. Finalmente, la relación REFERENCIA, contiene una ocurrencia por cada referencia extraida de los documentos procesados.

2.3 La interface de comunicación

Proporciona un canal para la comunicación del agente con el exterior, así como con el resto de servicios RePEc.

Esta parte se ha desarrollado como un módulo **CGI** en **perl** que funciona en el servidor del proyecto en la dirección `http://netec.ier.hit-u.ac.jp/adnetec-cgi-bin/get_data.pl`. Este módulo recibe alguna de las ordenes especificadas más adelante, ejecuta la búsqueda en la base de datos y devuelve una página HTML con los resultados. La posibilidad de devolver el resultado en formato XML queda abierta para una futura ampliación del proyecto.

El módulo admite peticiones tanto utilizando el método POST como GET del protocolo HTTP. Las opciones que soporta son tres:

1. **Handle**: Es el identificador del documento del cual se quiere obtener la información.
2. **Acción**: Es el tipo de información que se quiere obtener del documento. Las posibilidades son las descritas en la arquitectura de enlaces de referencias de [3]:
    (a) **GetMyData**: Obtiene la información bibliográfica del documento especificado. Solamente devolverá el título, autores y año de publicación.
    (b) **GetCurrentCitationList**: Obtiene los identificadores de las obras que han citado al documento.
    (c) **GetReferenceList**: Obtiene una lista de las referencias del documento con enlaces si la obra referenciada se encontrara en RePEc.
    (d) **All**: Incluye todas las opciones anteriores.

3. **Format**: Esta opción sólo se aplica a **GetReferenceList** e indica la cantidad de información que se proporcionará de cada referencia. Ofrece dos posibilidades:
   (a) **Sort**: Solamente se ofrece el texto de la referencia tal cual fue identificado por el módulo de análisis.
   (b) **Long**: Incluye toda la información existente sobre la referencia.

Por el momento CitEc ha sido implementado en tres servicios RePEc diferences: WoPEc [4](Working Papers in Economics) es una biblioteca digital financiada en parte por el Electronic Libraries Programm en el Reino Unido, Socionet[5] en Rusia y la Biblioteca de Ciencias Sociales[6] de la Universitat de Valencia. En la figura 1 se puede ver un ejemplo de implementación del agente por el primero de ellos, WoPEc.

En la imagen tenemos la descripción de un documento de trabajo y en la parte izquierda de la pantalla dos enlaces, uno con el texto "See this paper's referencs" y otro "Cited by". Ambos interrogan CitEc. El primero de ellos es un url de la forma: `http://netec.ier.hit-u.ac.jp/adnetec-cgi-bin/get_data.pl?handle=RePEc:cre:crefwp:34&action=GetReferenceList&format=sort`. Es decir solicita del agente la lista de referencias del documento identificado. Ante esta petición CitEc reponde con una página HTML que se puede ver en la figura 2. La página es una bibliografía en la cual se han incluido enlaces a aquellos documentos que están disponibles en formato electrónico a través de RePEc.

3 Resultados

En la presente sección se hace un análisis de los resultados obtenidos en el procesamiento inicial de todos los documentos disponibles en RePEc. El análisis se ha dividido en función de los distintos módulos de CitEc que intervienen en el proceso.

Para analizar el comportamiento del agente se definieron una serie de estados por los que todo documento debe pasar en el proceso de enlace de referencias. Así, todos parten de un estado *No procesado* que irá cambiando a medida que se superen los distintos módulos hasta llegar al estado final de *Analizado*. Si el proceso falla por algún mótivo, al documento se le asignará un error que identifica el problema detectado. Toda esta información sobre estados es guardada en una tabla específica de la base de conocimiento. Lo que ofrecemos en esta sección es una descripción detallada de los errores con los que nos hemos encontrado en el proceso, el número de documentos analizados correctamente, el número de referencias extraidas, referencias erroneas y sus causas, etc.

El estudio que se presenta corresponde a la situación de RePEc tal y como estaba a mediados de Diciembre de 2001. El procesado inicial de todos los documentos llevó más de un mes.

3.1 El módulo de lectura

En el proceso de lectura se detectaron 58175 descripciones con información sobre documentos electrónicos. Estas se distribuyeron entre 1106 series procedentes de 210 instituciones de todo el mundo. El número de descripciones por institución varía considerablemente desde aquellas que solamente ofrecen 1 o 2 documentos hasta las que ofrecen información procedente de varios centros. Es el caso de **DECOMATE** [7], un proyecto financiado parcialmente por la Unión Europea para crear una biblioteca digital en Economía a nivel europeo. Entre sus actividades está la puesta en formato electrónico de los documentos de trabajo de los departamentos de los Paises Bajos.

3.2 El módulo de análisis

Como vimos en la sección anterior, el módulo de análisis tiene varias funciones. Entre ellas, descargar los ficheros que componen cada documento y convertirlos a formato ASCII para que puedan ser procesados por el siguiente nivel o módulo.

El aspecto más importante de la descarga de documentos es discernir cuáles están realmente accesibles y por lo tanto podrán ser analizados y cuáles no. De los 58175 documentos que ofrecen inicialmente información sobre su disponibilidad electrónica, 28557 (el 50%) están realmente disponibles.

Las causas que indican que un documento no se puede procesar se asocian a los estados

[4] `http://netec.mcc.ac.uk/WoPEc`
[5] `http://socionet.ru`
[6] `http://www.uv.es/bibsoc`
[7] `http://cwis.kub.nl/~dbi/demomate/`

Figura 1: Documento con enlaces al interfaz de CitEc

Error	*Frecuencia*	*Porcen.*
restringido	20878	36%
no procesable	6387	10%
no encontrado	2353	4%
disponible	28557	50%

Tabla 1: Accesibilidad de los documentos

restringido, no encontrado y *no procesable.* En la figura 1 se muestra la proporción de documentos no disponibles por cada uno de estos motivos, junto con la proporción de documentos que sí lo están.

Destacar el elevado número de documentos que no son de acceso gratuito (*restringido*) y que por lo tanto no podemos obtener. La mayor parte de ellos son artículos procedentes de editoriales comerciales o del proyecto JSTOR. JSTOR es un proyecto comercial cuyo objeto es digitalizar y poner en Internet los números antiguos de las revistas más importantes de cada disciplina. Documentos *no encontrados* son aquellos cuya dirección electrónica es incorrecta. Hay por lo tanto un error en los datos suministrados por los centros. Finalmente el apartado de *no procesables* engloba una amplia variedad de errores. Entre ellos, documentos que han sido digitalizados como imágenes, aunque estén en formato PDF, y por lo tanto no pueden ser procesados. Igualmente se incluyen aquellos cuya dirección electrónica no lleva al texto completo sino a su resumen. Esta es una práctica no permitida en RePEc, pero hay algunos centros que la practican.

Una vez descargados los documentos, el agente procede a la conversión de los 28557 documentos disponibles a formato ASCII. En este paso se podrían dar los siguientes errores: *formato incompatible* cuando el fichero no sea de tipo PDF o PostScript, *error de conversión* cuando el programa de conversión falle, *no en inglés* cuando el documento esté en una lengua diferente del inglés, y *sin referencias* cuando se haya convertido a texto pero haya sido imposible encontrar una sección de referencias. En la práctica este último error está relacionado con el segundo, ya que en un elevado porcentaje de casos CitEc no encuentra referencias porque éstas están colocadas al final del documento y el proceso de conversión ha fallado antes de alcanzarlas. El resultado

My References:

1 Armington, P. (1969). A theory of demand for products distinguished by places of production. IMF Staff Papers, 27, 488--526.

2 Backus, D. K., Kehoe, P. J., & Kydland, F. E. (1992). International real business cycles. Journal of Political Economy, 100 (4), 745--775.(RePEc:ucp:jpolec:v:100:y:1992:i:4:p:745-75)

3 Backus, D. K., Kehoe, P. J., & Kydland, F. E. (1994). Dynamics of the trade balance and the terms of trade: The J-curve?. American Economic Review, 84 (1), 84--103.(RePEc:aea:aecrev:v:84:y:1994:i:1:p:84-103)

4 Balassa, B. & Kreinin, M. E. (1967). Trade liberalization under the Kennedy Round: The static effect.

5 Review of Economics and Statistics, 44 (2), 125--137. de Vries, B. A. (1951). Price elasticities of demand for individual commodities imported into the United States. IMF Staff Papers, 1 (3), 397--419.

6 Dellas, H. (1986). A real model of the world business cycle. Journal of International Money and Finance, 5, 381--394.

7 Hansen, G. D. (1989). Technical progress and aggregate fluctuations. Working paper 546, UCLA.(RePEc:cla:uclawp:546)

8 Hodrick, R. & Prescott, E. C. (1980). Post-war U.S. business cycles: An empirical investigation. Graduate School of Industrial Administration, Carnegie Mellon University.

9 Kakwani, N. C. (1972). On the bias in estimates of import demand parameters. International Economic Review, 13 (2), 239--244.(RePEc:ier:iecrev:v:13:y:1972:i:2:p:239-44)

10 Kemp, M. C. (1962). Errors of measurement and bias in estimates of import demand parameters. Economic Record, 38, 369--372.

11 Kreinin, M. E. (1961). The effects of tariff changes on the price and volume of imports. American Economic Review, 51 (3), 310--324.

12 Kydland, F. E. & Prescott, E. C. (1988). The workweek of capital and its cyclical implications. Journal of Monetary Economics, 21, 343--360.

13 Kydland, F. E. & Prescott, E. C. (1991). Hours and employment variation in business cycle theory. Economic Theory, 1 (1), 63--81.

Figura 2: Página devuelta por CitEc con la bibliografía de un documento

Error	*Frecuencia*	*Porcen.*
error de conversión	6931	24%
no en inglés	963	3%
sin referencias	3476	12%
formato incompatible	1665	6%
convertidos	15522	55%

Tabla 2: Errores en la conversión a ASCII

Error	*Frecuencia*	*Porcen.*
error en referencias	2081	13%
analizados	13446	87%

Tabla 3: Documentos excluidos al identificar las referencias

es un fichero de texto correcto pero incompleto.

De los 28557 documentos que han llegado a este punto, 15522 lo han pasado satisfactoriamente. En la tabla 2 puede verse la distribución de errores.

Hay que destacar aquí el elevado número de documentos que no ha sido posible convertir satisfactoriamente. Un punto a tener en cuenta en sucesivas actualizaciones del sistema será probar nuevos programas de conversión de PDF a ASCII. [7] hace una interesante comparativa de herramientas de libre distribución disponibles para realizar esta operación.

15522 documentos fueron analizados para localizar sus referencias bibliográficas e identificar cada uno de los elementos que las componen. En este paso el único problema que puede aparecer para que un documento sea excluido del análisis, en su totalidad, es que el sistema encuentre un número de referencias que no esté entre los límites especificados. Dichos límites son configurables y se han fijado entre 1 y 70. Como se puede apreciar en la tabla 3, casi el 90% de los documentos quedan englobados dentro de estos límites.

A partir de este momento los errores que se pueden producir son en la localización de los elementos que componen cada referencia. Hay tres elementos clave que deben ser encontrados en todas ellas: el año de publicación, el autor o autores y el título del documento. Si el proceso falla en identificar alguno, la referencia es descartada. Los resultados pueden verse en la tabla 4. En total

	Frecuencia	*Porcentaje*
Referencias	293069	
Correctas	283932	97%
Incorrectas	9137	3%
De ellas:		
Error en título	7921	87%
Error en autor	1216	13%

Tabla 4: Causas de exclusión de referencias

se identificaron 293069 referencias en 13446 documentos. Ello nos da una media de 22 referencias por cada uno. De ellas, CitEc fue capaz de identificar correctamente 283932 (97%) y falló solamente en 9137 casos (3%). Entre las causas de error tenemos fallos en identificar el título en 7921 referencias y en identificar el autor en 1216 casos.

3.3 El módulo de enlace

El último módulo analizado es el de enlace que, como se comentó, se encarga de relacionar cada referencia correctamente identificada, con el texto completo del documento al que representa, si se encuentra disponible en RePEc. De las 283932 referencias identificadas, 52352 (el 18.4%) representaban a un documento cuyo texto electrónico está disponible en RePEc y por lo tanto se puede establecer un enlace entre ambos. Este porcentaje podría ser incrementado considerablemente si se añadieran enlaces a los documentos de los que solamente tenemos información referencial. No obstante esto queda fuera del objeto de este trabajo. Igualmente queda fuera del mismo el enlace a documentos disponibles electrónicamente fuera de la biblioteca digital RePEc.

3.4 Resultados globales

Finalmente en la figura 3 se presenta una síntesis de los resultados obtenidos. El 48% de los documentos disponibles en RePEc fueron enlazados de forma satisfactoria. Es decir, el agente fue capaz de extraer y enlazar sus referencias. Por otro lado, esto significa que más de la mitad de los documentos disponibles no pudieron ser enlazados por diferentes razones. La causa de error más importante es el proceso de conversión de PDF a ASCII. En segundo lugar, el agente fue incapaz de encontrar una lista de referencias en el 12% de los documentos. Dado que no es frecuente encontrar un documento científico sin bibliografía, se podría concluir que el algoritmo de análisis necesita ser mejorado. Otros problemas con menor incidencia son documentos donde el agente ha sido incapaz de aislar cada una de las referencias. Este problema está relacionado también con la efectividad de los algoritmos de análisis. Otra causa de errores son los documentos distribuidos en formatos distintos a los aceptados (PDF o PS), ésta es una parte mínima de errores. Finalmente, hay un 3% de documentos que no han sido procesados porque no están en inglés.

4 Conclusiones

A lo largo de este trabajo se ha descrito una aproximación que hace posible la extracción automática de información sobre citas de los documentos existentes en una biblioteca digital. Dicha aproximación difiere de iniciativas similares en el sentido que se ha trabajado en un entorno distribuido, con documentos procedentes de instituciones de todo el mundo y con una gran variedad de formatos.

El agente desarrollado es capaz de determinar cuando ese entorno cambia y actuar en consecuencia. Su actuación se concreta en la recuperación del texto de los nuevos documentos desde cualquier parte de Internet, extracción de sus referencias bibliográficas cuando existen y establecimiento de enlaces con los documentos citados que se encuentran también en la biblioteca. Finalmente, también es capaz de codificar la información obtenida en unos *templates* especiales y ponerla a disposición de aquellos servicios RePEc que deseen utilizarla para mejorar sus aplicaciones.

En la sección anterior se ha demostrado hasta qué punto este sistema se ha mostrado efectivo. No obstante quedan algunos puntos que inciden directamente en la calidad de los resultados, como por ejemplo:

- La colaboración de los administradores de archivos es un punto crucial para el buen funcionamiento del sistema. Lo fundamental es que proporcionen unas buenas descripciones de sus documentos. Entre otras cosas, deben asegurarse que las direcciones de los textos de los documentos son correctas y deben proporcionar información sobre las fechas de creación de los documentos.
- La conversión de PDF/PS a ASCII es la parte del sistema con mayor índice de errores. En este sentido es necesario

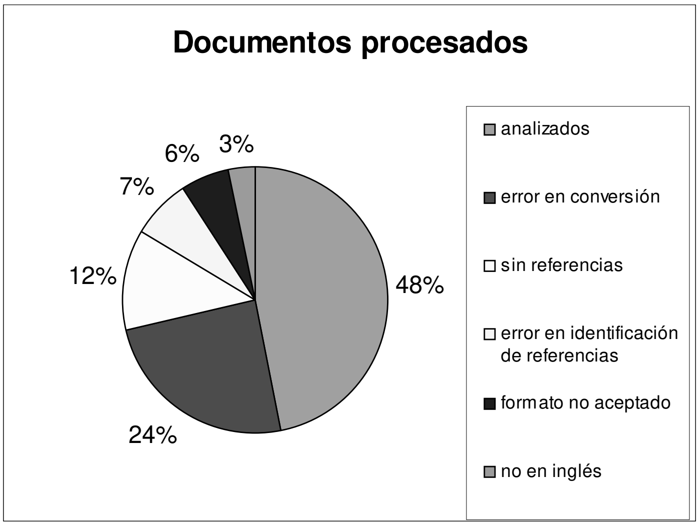

Figura 3: Resumen de resultados

desarrollar mejores herramientas de conversión que las existentes actualmente.

- Los algoritmos de análisis del documento y búsqueda de referencias necesitan ser mejorados para reducir el 12% de casos en los que el agente no es capaz de encontrar la bibliografía.

Referencias

[1] José M. Barrueco and T. Krichel. *Metadata and organizing educational resources on the Internet*, chapter Distributed cataloging on the Internet: the RePEc project, pages 227–241. Haworth Information Press, 2000.

[2] Jose M. Barrueco and Thomas Krichel. Automatic extraction of citation data in a distributed digital library. In *4th International Conference on Enterprise Information Systems. New Developments in Digital Libraries*, 2002.

[3] Donna Bergmark and Carl Lagoze. Reference linking the web's scholarly papers. Technical report, Cornell Digital Library Research Group, 2000.

[4] Robert D. Cameron. A universal citation database as a catalyst for reform in scholarly communication. *First Monday*, 2(4), 1997.

[5] N.R. Jennings and M.J. Wooldridge, editors. *Agent Technology. Foundations, Applications and Markets*. Springer-Verlang, 1998.

[6] Bruno Maltrás. *Los indicadores bibliométricos en el estudio de la ciencia: fundamentos conceptuales y aplicación en política científica*. PhD thesis, Universidad de Salamanca, Facultad de Filosofía, 1996.

[7] Nicholas Robinson. A comparison of utilities for converting from postscript or portable document format to text. CERN-OPEN-2001-065.