# Co-uso de documentos en una biblioteca digital

**José Manuel Barrueco**
Universitat de València
Spain

**Thomas Krichel**
Long Island University
New York, USA

**Resumen** La biblioteca digital RePEc ofrece la mayor fuente distribuida de documentos científicos disponible actualmente en el mundo. Su arquitectura se basa en la distinción entre proveedores de datos y proveedores de servicios. WoPEc es uno de estos proveedores de servicios al usuario final que está funcionando desde 1993. En esta comunicación intentamos determinar qué documentos en la colección tienen un contenido similar a través del estudio de los accesos de los usuarios. La idea es que si diferentes usuarios acceden a una pareja de documentos repetidas veces, entonces dichos documentos deberían responder a la misma necesidad de información. Presentamos una discusión teórica de este tipo de relación así como una demostración práctica. Para ello introducimos una medida que denominamos de *co-uso* y la aplicamos a los usuarios que han accedido el servicio WoPEc.

## 1. Introducción

La operación de establecer relaciones entre documentos de contenido similar ha sido objeto de múltiples estudios en el área de recuperación de la información. En su manual sobre el tema [4] escribe:

> Probably the single key concept behind information storage and retrieval is document similarity.

La principal motivación para la búsqueda de la similitud entre documentos es mejorar la calidad en los sistemas de recuperación de la información. Cuando un usuario ha encontrado un documento relevante para su necesidad de información, a través de las relaciones entre documentos puede ser reconducido o guiado hacia otros documentos similares que posiblemente merezcan también su atención. Una segunda motivación más teórica sería el descubrimiento de modelos de pensamiento dentro de un corpus de conocimiento que está disperso entre una amplia colección de documentos.

Sin lugar a dudas hay muchas formas a través de las cuales se puede establecer la similitud entre documentos. Una podría ser el juicio de un experto, por ejemplo. Esto es lo que se hace habitualmente cuando se clasifica un documento en función de algún sistema como la CDU (Clasificación Decimal Universal). Dividimos un corpus documental en clases en función de su contenido temático. Este es un método ampliamente usado que tiene el inconveniente de que necesita, al menos por ahora, el trabajo de un experto. De esta forma, son bastantes las colecciones, sobretodo en Internet, en las que no se incluye una clasificación de sus fondos. Por ejemplo, en la colección que vamos a utilizar como campo de pruebas en este trabajo, tres de cada cuatro documentos carecen de una clasificación temática.

Otras medidas clásicas para determinar la similitud de los documentos científicos son las co-citas. Su inconveniente es que las citas requieren la existencia de un Índice de Citas cuya producción es altamente costosa. De hecho en nuestro país no existe ninguno y se ha de recurrir a los elaborados por el Institute for Scientific Information de USA. No obstante con la aparición de los documentos electrónicos el problema del coste está en vías de solución. Se está trabajando en herramientas de software para crear índices de citas que puedan trabajar de forma autónoma [1]. Se puede ver además [2] para una descripción del trabajo de análisis de citas llevado a cabo sobre la misma biblioteca digital que utilizamos en este trabajo.

En esta comunicación introducimos un concepto completamente nuevo de similitud que toma como punto de partida el comportamiento de los usuarios de una biblioteca digital.

Si partimos de la idea de que dos documentos son similares si responden a una misma necesidad de información, entonces podemos establecer que dos documentos son similares si varios usuarios consultan o retiran dichos documentos mientras están trabajando en el sistema. Siempre que tengamos una biblioteca digital con un amplio registro de su utilización por parte de los usuarios, podremos deducir qué documentos se han utilizado juntos frecuentemente. A esto es a lo que nos referimos con el término *co-uso* en el resto del trabajo.

Por el momento no estamos al corriente de otro estudio similar que haya implementado este concepto dentro del ámbito de las bibliotecas digitales. No obstante existen trabajos como el de [3] que han utilizado un concepto similar para aplicarlo a la evaluación de bibliotecas digitales. Por otro lado, si que existen sistemas comerciales que usan un concepto parecido. Así Amazon.com, por ejemplo, sugiere a un consumidor que está considerando comprar un item $x$ que otras personas que compraron $x$, también compraron $y$. La diferencia entre ambas aproximaciones estriba en que en nuestra biblioteca los usuarios son anónimos lo que hace que sea más difícil obtener pruebas del comportamiento de los mismos.

El resto del trabajo se estructura como sigue: en la sección dos se formula una definición precisa del concepto de *co-uso*. En la sección tres presentamos algunos resultados analíticos. La sección cuatro muestra su implementación en nuestra biblioteca digital y la sección cinco concluye este trabajo.

## 2. Midiendo co-uso

Para probar en la práctica la efectividad del concepto de *co-uso* hemos de estudiar el comportamiento de los usuarios de una biblioteca digital. Hemos seleccionado para nuestro trabajo la biblioteca WoPEc (Working Papers in Economics) accesible en la dirección `http://netec.mcc.ac.uk/WoPEc` y en la que los autores venimos trabajando desde hace casi diez años. WoPEc contiene en la actualidad información bibliográfica sobre más de 38.600 artículos publicados en revistas y más de 44.700 documentos de trabajo distribuidos por departamentos y centros de investigación de todo el mundo. Todos ellos están disponibles a texto completo en la red, si bien no necesariamente de forma gratuita. WoPEc tiene tiene tres mirrors: en USA, Reino Unido y Japón. En conjunto los servidores reciben una media de 250.000 accesos al mes con más de 40.000 documentos descargados.

Es habitual que toda biblioteca digital guarde un registro de quien ha accedido al sistema, qué búsquedas ha realizado, cuántos documentos ha recuperado, cuántos ha descargado, etc. En el caso de WoPEc este tipo de archivos de registro de uso del sistema se remontan hasta 1998 lo que constituye un excelente campo de pruebas para nuestro trabajo.

Si la biblioteca mantiene una base de datos de usuarios registrados, como sucede con muchos distribuidores comerciales, es sumamente fácil detectar el comportamiento de los mismos ya que siempre están identificados cuando acceden al sistema. No obstante WoPEc no dispone de una base de datos similar. Pensamos que el exigir un registro es un requisito tedioso que frenaría la llegada de nuevos usuarios o la entrada de aquellos que no tienen una idea clara de qué van a encontrar. En general el usuario se siente reticente a entregar sus datos personales, por mínimos que sean, o su dirección de correo electrónico, si no es con una justificación clara que en este caso no existe al ser la información gratuita.

Además WoPEc es un servicio no comercial. WoPEc no entra en una relación contractual con los usuarios o clientes. Por un lado, este tipo de contratos con clientes son muy costosos de establecer ya que implican complicadas negociaciones, cuestiones de privacidad, etc. Por otro lado, la parte positiva es que los contratos, tanto si implican o no una retribución económica, generan datos sobre los consumidores o visitantes de nuestro sitio web. Los datos sobre consumidores en si mismos contienen un considerable valor. Se puede consultar [5] para ver una interesante discusión sobre el tema.

Al no disponer de estos datos, ni haber utilizado herramientas como por ejemplo las *cookies*, durante los años que vamos a analizar, la única solución para obtener datos sobre nuestros usuarios es leer el fichero de *logs* del servidor HTTP para intentar encon-

trar rastros de utilización del sistema por una misma persona.

Cada vez que un usuario hace una petición al servidor web se generan una o varias lineas en un fichero de *logs*. Este tiene un formato ASCII que tienen la estructura siguiente: 66.196.73.18 - - [15/Sep/2002:06:45:18 -0500] "GET /WoPEc/data/Papers/nbrnberwo5852.html HTTP/1.0" 200 9813 "-" "Mozilla/5.0" Aquí vemos la dirección IP del cliente. La fecha y hora de la petición. El documento que se ha recuperado, el código HTTP devuelto por el servidor, el tamaño del fichero y el tipo de navegador que está utilizando el cliente.

Ahora bien, lo que este fichero ofrece es meramente un rastro de una dirección IP. Durante los últimos años a Internet se accede cada vez más a través de ordenadores personales que de sistemas multiusuario. Sin embargo no podemos estar completamente seguros que un número IP corresponde a la máquina de un usuario individual ya que habitualmente en instituciones como las universidades se crean cachés de páginas, o proxies de forma que todo el tráfico de salida de la institución se canaliza a través de una sola máquina. Por lo tanto debemos tener un cuidado adicional con las direcciones IP que aparecen de forma particularmente frecuente y que al referirse no a usuarios finales sino a máquinas pueden desvirtuar los resultados. Por ello se hace necesario un tratamiento previo de los datos almacenados en los logs.

Si comenzamos asumiendo que los datos de uso que proceden de una misma máquina proceden a su vez de un único usuario, estaremos en disposición de identificar al mismo. Sin embargo para deducir que los documentos que han sido accedidos juntos, lo han sido porque responden a la misma necesidad de información, tendríamos que asumir que la necesidad de información permanece la misma durante todo el periodo de observación. Esto plantea problemas. Normalmente los usuarios ven a WoPEc como una fuente de referencia a la que acuden con diferentes necesidades. Por lo tanto es seguro que el fichero contendrá material que aun referido al mismo usuario contendrá diferentes necesidades de información.

Para solucionar este problema hemos establecido que los usuarios acceden al servicio dentro de "sesiones". Una sesión la podríamos definir técnicamente como:

un subconjunto del total de entradas en el fichero log que proceden de la misma máquina IP dentro de un determinado periodo de tiempo.

Si $L$ es el fichero log, tendríamos que: $L = \{S_1, S_2, ...S_n\}$

El problema está en cómo determinar dicho periodo. Es decir qué tiempo $t$ establecemos como medida de corte para diferenciar dos sesiones procedentes de la misma dirección IP. Para determinar $t$ hemos tomado una muestra representativa de datos y hemos estudiado el tiempo transcurrido entre dos conexiones sucesivas desde la misma dirección IP. El resultado lo podemos ver en el gráfico de la figura 1.

Como podemos comprobar el lapso de tiempo entre dos peticiones es de unos pocos segundos. Concretamente en torno al 85 % de peticiones se realizan con un intervalo inferior a los tres minutos. Si siguiéramos la distribución comprobaríamos que son menos del 4 % por ciento aquellos casos en los que se ha superado la hora de diferencia.

Siguiendo estas conclusiones estableceremos un criterio amplio para definir el punto de corte de las sesiones y asumiremos que no debe haber más de una hora entre dos entradas consecutivas en el fichero de logs para ser consideradas como pertenecientes a la misma sesión.

Además, una restricción adicional es que no consideraremos a efectos de estudio los accesos repetidos al mismo documento dentro de la misma sesión. Es decir, cada documento sólo puede aparecer una vez en cada sesión. Con esto evitamos la desviación que se puede producir por accesos repetidos a un documento por problemas técnicos tales como usuarios que pulsan repetidas veces sobre un enlace al no obtener una respuesta rápida del servidor, o circunstancias como que un usuario no tiene tiempo de leer un artículo la primera vez que lo ve, etc.

Bajo estas restricciones parece conveniente modelar las sesiones como conjuntos de documentos. Podemos ahora proceder a formular una definición formal de co-uso.

Sea $D$ el conjunto de documentos disponibles en la biblioteca digital y $S$ el número de sesiones. Cada sesión $S_i$ contiene un subconjunto del conjunto de documentos, es decir, $S_i \subset D$. Introduzcamos $|\cdot|$ como una notación para el número de elementos

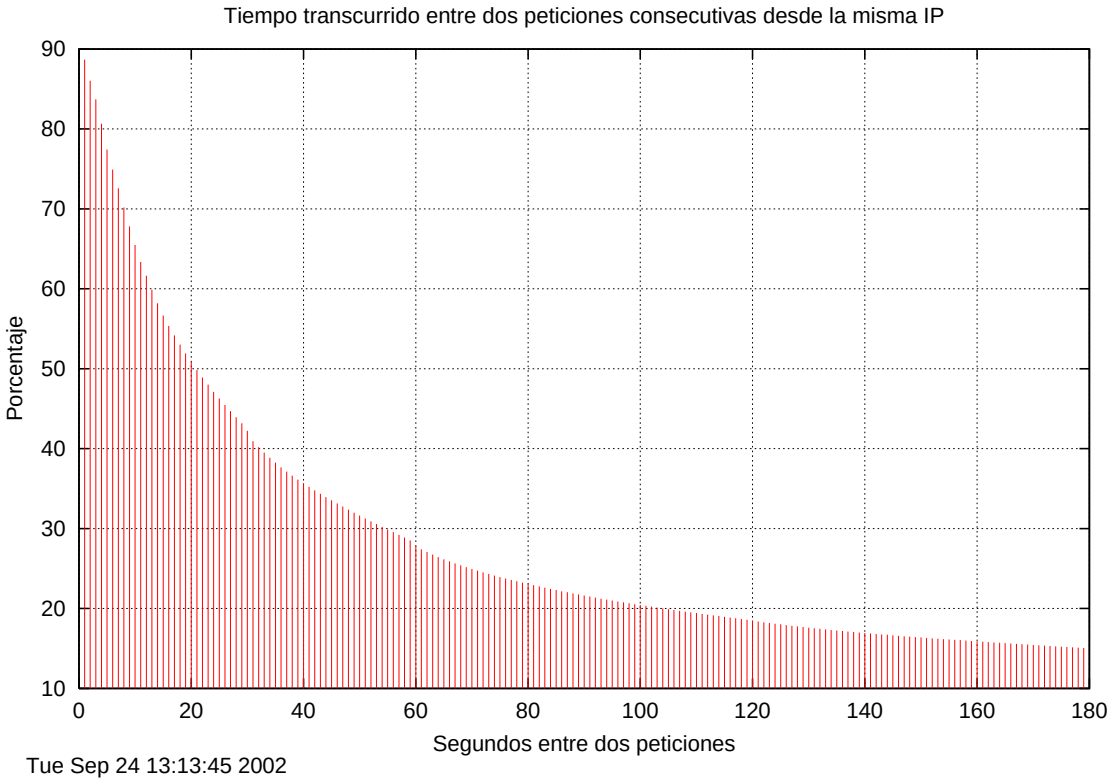


Figura 1: Lapso de tiempo entre dos peticiones procedentes de la misma máquina

en un conjunto. Escojamos dos documentos $d_i$ y $d_j$. Sea $u(d_i, d_j)$ el conjunto de todas las sesiones que contienen ambos $d_i$ y $d_j$. Para expresarlo de forma más sencilla establezcamos $u(d_i) = u(d_i, d_i)$. Entonces proponemos llamar coeficiente de *co-uso* $c(d_i, d_j)$ a lo siguiente:

$$c(d_i, d_j) = \frac{|u(d_i, d_j)|^2}{|u(d_i)||u(d_j)|}$$

Es decir, que el *co-uso* es igual al cuadrado de las veces que han aparecido los dos documentos juntos, partido por el producto del número de veces que han aparecido cada uno.

El motivo para introducir el producto en el denominador es tener la seguridad que no nos encontraremos con situaciones donde un documento aparece frecuentemente junto con otro simplemente por el tamaño de su propio uso. Por otro lado también nos lleva a que $c(d_i, d_j)$ quede en el rango $0 \leq c(d_i, d_j) \leq 1$ y que $c(d_i, d_i) = 1$. Es decir el coeficiente de *co-uso* de dos documentos varía entre el 0 cuando son totalmente diferentes o no han aparecido nunca juntos hasta el 1 cuando tienen un mayor grado de similitud y siempre han aparecido juntos.

Autores como [3] han definido un coeficiente similar pero simplemente añadiendo un incremento fijo cada vez que dos documentos aparecen juntos. Esta medida no nos parece apropiada para nuestros propósitos de recuperación de información ya que no es suficiente para determinar el grado de similitud entre ambos documentos.

Esta claro que la medida que proponemos aquí es de carácter intuitivo. No hay una medida desarrollada científicamente de la probabilidad que, cualquiera que sea el tamaño de la sesión, veamos los dos documentos apareciendo juntos en la misma. Es posible evaluar esta probabilidad, pero estos cálculos no son abordados aquí sino que se plantea como un posible trabajo futuro.

## 3. Resultados

Los autores diseñamos un programa en perl para investigar el *co-uso* de documentos en un periodo de tres años 1999 a 2001. Este programa toma como entrada un fichero de logs que originalmente estaba compuesto por 102.551.867 líneas. Antes de proceder al estudio, el programa ordena el fichero por direcciones IP y dentro de éstas por fechas y hora.

El primer paso como hemos mencionado anteriormente es la limpieza para eliminar los

datos que no nos interesan o que podrían inducirnos a error. En el proceso de limpieza inicial se descartaron 24.836.221 peticiones de imágenes y 401.614 peticiones de ficheros CSS (Cascade Style Seets). También se eliminaron 3.164.598 peticiones que contenían un error HTTP 404, es decir, se referían a un objeto que no existía en el servidor. 9.315 peticiones devolvieron algún otro error del protocolo HTTP.

Del resto de peticiones que fueron analizadas se extrae que 3.966.837 máquinas diferentes han accedido al sistema. Excluimos además aquellas peticiones que vengan de máquinas que hayan accedido el fichero *robots.txt* según el protocolo de exclusión de robots. Además hay direcciones caché y proxies que deben ser eliminados también. Este es un terreno mucho más complicado. La regla que hemos seguido es analizar el log de cada día y detectar todas las máquinas que han emitido más de 100 peticiones. En total hemos detectado 23.068 hosts y descartado 17.484.961 lineas que fueron pedidas por éstos.

Después de esta limpieza nos quedamos con 23.413.849 lineas para ser analizadas. En ellas detectamos 5.251.337 sesiones. Para evitar las interferencias que causarían aquellos usuarios esporádicos que llegan a WoPEc por casualidad, sin una necesidad de información concreta relacionada con nuestro campo de interés, hemos establecido algunas restricciones sobre las sesiones identificadas. Nos limitamos a sesiones que estén en el rango de 3 a 100 accesos y que incluyan al menos una búsqueda. Quedaron para ser analizadas 69.179 sesiones. Lo que nos indica que muchos usuarios llegan a WoPEc mientras están navegando por la red pero no están interesados por su contenido de forma que ven lo que hay y se van rápidamente.

Por razones obvias de espacio no podemos hacer los resultados disponibles en esta comunicación. Por lo tanto los hemos colocado en dirección: `http://openlib.org/home/krichel/kumegawa/report.htm`. En esta página hemos impreso los pares de documentos con un mayor índice de co-uso, hasta un redondeo de 0,5. Cuando miramos a los datos podemos comprobar los siguientes resultados.

Primero, los documentos dentro de la misma colección, son más dados a aparecer usados conjuntamente que los documentos de diferentes colecciones. A menudo son documentos del mismo autor, en diferentes periodos de tiempo disponibles en la misma serie de documentos de trabajo o revista los que aparecen con un mayor índice de co-uso. Esta observación sugiere que la división original de los documentos en series en la que está estructurada WoPEc, no solo proporciona un nivel organizacional para los datos, sino también una semántica.

En segundo lugar, de la lectura a través de las diferentes parejas aparece que un concepto clave que aparece en el título inmediatamente explica el co-uso. De esta forma sería posible profundizar en conceptos clave que son de interés de los usuarios mediante el análisis de los documentos que tienen un elevado índice de *co-uso* y encontrar las palabras comunes. Este sería un tema de investigación interesante en el futuro.

## 4. Implementación

Los resultados del estudio que se ha descrito en esta comunicación han sido implementados con éxito en la base de datos WoPEc. Así a la descripción de cada documento del que disponemos información se ha añadido la lista de aquellos otros documentos con los que está relacionado y el coeficiente de *co-uso* entre ambos. Un ejemplo de esta implementación puede verse en la figura 2.

## 5. Conclusiones

En este documento hemos introducido el *co-uso* como un medio para establecer relaciones entre documentos. El concepto de *co-uso* nos ha llevado a obtener excelentes resultados empíricos. Nuestros resultados sugieren que el concepto de *co-uso* es válido e interesante en el ámbito de la recuperación de información. Los problemas que hemos encontrado para hacerlo funcionar en la práctica son consecuencia de una pobre medición del uso más que fallos en el nivel teórico subyacente. Igualmente que en el caso de las co-citas sobre cómo de útil puede llegar a ser el co-uso. Esperamos haber abierto ese debate con este documento.

Hay otras dos consecuencias adicionales que se desprenden de este documento. A largo plazo pensamos que nuestro trabajo se puede ver como simple pero un paso pionero hacia sistemas de recuperación de la información que serán capaces de aprender. En lugar de tener un sistema que tiene un procedimiento fijo que puede ser aplicable a cualquier

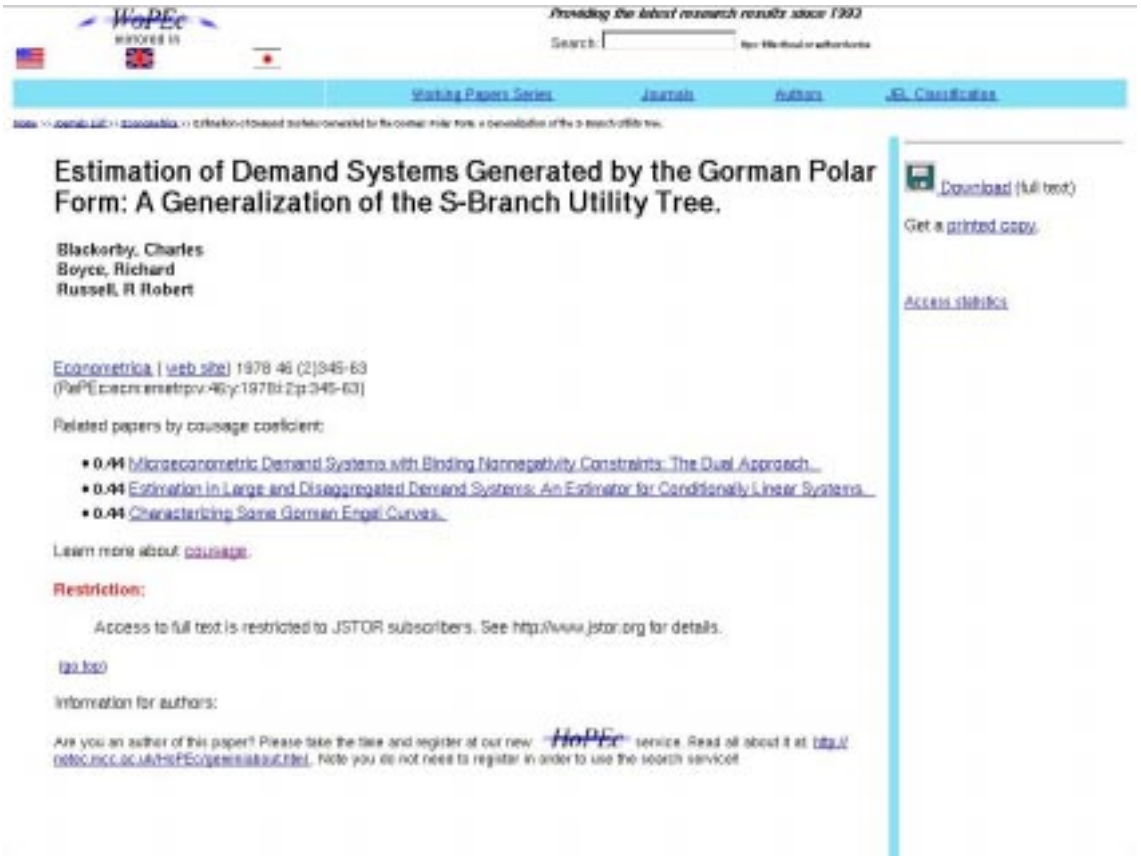


Figura 2: Implementación del co-uso de documentos en WoPEc

pregunta en la colección podemos imaginar que los sistemas futuros aprenderán del comportamiento de usuarios pasados para servir mejor al usuario presente. El *co-uso* aparece como un concepto clave para avanzar en esta dirección.

## 6. Anexo: Ejemplos de documentos relacionados por co-uso

- Marvel, Howard P & Ray, Edward John. Intraindustry Trade:
  - 1 Krugman, Paul R. Intraindustry Specialization and the Gains from Trade.
- Arabmazar, Abbas & Schmidt, Peter. An Investigation of the Robustness of the Tobit Estimator to Non-Normality.
  - 0.44 Olsen, Randall J. Note on the Uniqueness of the Maximum Likelihood Estimator for the Tobit Model.
  - 0.75 Fair, Ray C. A Note on the Computation of the Tobit Estimator.
- Fair, Ray C. A Note on the Computation of the Tobit Estimator.
  - 0.45 Greene, William H. On the Asymptotic Bias of the Ordinary Least Squares Estimator of the Tobit Model.
  - 0.75 Arabmazar, Abbas & Schmidt, Peter. An Investigation of the Robustness of the Tobit Estimator to Non-Normality.
- Park, Yung C. The Transmission Process and the Relative Effectiveness of Monetary and Fiscal Policy in a Two-Sector Neoclassical Model.
  - 1 Benavie, Arthur. Monetary and Fiscal Policy in a Two-Sector Keynesian Model.
  - 1 Tower, Edward. Monetary and Fiscal Policy under Fixed and Flexible Exchange Rates in the Inter-run.
  - 0.67 McMillin, W Douglas. A Dynamic Analysis of the Impact of Fiscal Policy on the Money Supply: A Note.

- Francesca Cornelli & David Goldreich. Bookbuilding and Strategic Allocation
    - 1 Francesca Cornelli & David Goldreich. Bookbuilding and Strategic Allocation

## Referencias


[1] Jose M. Barrueco. Reference linking. *El profesional de la información*, 11(4):278–282, 2002.

[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] Johan Bollen and Rick Luce. Evaluation of digital libraries impact and user communities by analysis of usage patterns. *D-Lib Magazine*, 8(6), 2002.

[4] Robert R. Korfhage. *Information storage and retrieval*. John Wiley and Sons, 1997.

[5] Carl Shapiro and Hal Varian. *Information rules*. Harvard Business School Press., 1999.