Agentes Inteligentes: Recuperación Autónoma de Información en la Web

Emanuel García Pérez

November 24, 2014

La web como fuente de información

El constante crecimiento de Internet durante los últimos años ha sido, dentro del campo de la información, uno de los desarrollos más importantes. Respecto al ámbito científico, muchas de las fuentes de información tradicionales ya se encuentran en la red, lo que ha propiciado poner en evidencia el problema concerniente a la **recuperación de información** en la web.

Recuperación de información en la web

Actualmente los sistemas de recuperación de información en la web utilzian dos mecanismos, no excluyentes entre si y que se pueden combinar, para solventar este problema:

- **Búsqueda mediante palabras clave** Se pueden aplicar técnicas que mejoren los resultados; tesauros de similitud, análisis de cluster, utilización de redes neuronales.
- **Clasificación de páginas en categorías** La clasificación suele realizarse manualmente, aunque también se ha realizado de forma automática utilziando mapas autoorganizados para establecer las categorías de los términos, los cuales serán empleados para definir vectores de páginas web.

Sistemas generales de búsqueda

Ya sea que se utilice una búsqueda por palabras clave o a través de la clasificación previa de páginas, o una combinación de ambos, estos métodos parten de la existencia de una **base de datos**, de un tamaño considerablemente grande, que contenga la colección de páginas web, lo que conlleva a que la experiencia de los usuarios se vea mermada por la **baja presición** de las respuestas devueltas y la enorme cantidad de resultados asociados a una búsqueda particular.
Es debido a este tipo de problemas que se pretende obtener resultados más precisos, no necesariamente en gran cantidad, basándose en el uso de **agentes inteligentes**.

Agentes Inteligentes

AGENTES INTELIGENTES

Es un sistema situado en y a parte de un entorno que se siente ese entorno y actúa sobre él, a traves del tiempo, persiguiendo sus propios objetivos de forma que afecte lo que siente en el futuro.

Son definidas por mecanismos de cooperación, cordinación y negociación; por lo tanto sus mecanismos explicitos se establecen las relaciones entre susbsistemas y componentes

Interacciones

Propiedades

- **Continuidad Temporal:** Se considera un agente un proceso sin fin, ejecutándose continuamente y desarrollando su función.
- **Autonomia:** Es capaz de adaptarse aunque el entorno cambie severamente.
- **Socialidad:** Permite a un agente comunicar con otros agentes o incluso con otras entidades.
- **Racionalidad:** Siempre realiza lo correcto a partir de los datos que percibe del entorno.
- **Reactividad:** Actúa como resultado de cambios en su entorno.
- **Pro-actividad:** Es capaz decontrolar sus propios objetivos a pesar de cambios en el entorno.
- **Adaptibilidad:** Está relacionado con el aprendizaje que un agente es capaz de realizar y si puede cambiar su comportamiento basándose en ese aprendizaje.
- **Movilidad:** Capacidad de un agente de trasladarse a través de una red telemática.
- **Veracidad:** Asunción de que un agente no comunica información falsa a propósito.
- **Benevolencia:** Asunción de que un agente está dispuesto a ayudar a otros agentes si esto no entra en conflicto con sus propios objetivos.

Caracteristicas

- **Autonomia:** Capacidad de actuar sin intervención humana directa o de otros agentes.
- **Reactividad:** Estan inmersos en un determinado entorno, en el que persive estimulos y reacciona en un tiempo preestablecido.
- **Racionalidad:** Tiene unos objetivos especificos y siempre intenta llevarlos a cabo
- **Iniciativa:** Reaciona a los cambios que se producen en su entorno del cual debe tomar la iniciativa para actuar guiado por los objetivos que debe satisfacer.

Tipos

Según su funcionamiento

- **Cooperativos:** Se coordinan con otros agentes para conocer el entorno de información en el que se mueven.
- **Móviles:** Realizan búsquedas basadas en la programación remota.

Según sus prestaciones

- **Agentes de búsqueda:** Son agentes especializados en la búsqueda de datos, que realizan búsquedas conceptuales en base a los términos solicitados por los usuarios
- **Agentes de consulta:** Realizan las consultas originando más agentes como respuesta las peticiones del usuario.
- **Agentes de bases de datos:** Especializados en hacer consultas en una base de datos concreta, que conocen la información almacenada en ella.
- **Agentes de consulta de bases de datos:** Son el resultado de la interacción de los dos anteriores
- **Agentes mediadores de consulta de bases de datos:** Equivalen al anterior, aunque además cuentan con un sistema mediador entre el agente de consulta y el de base de datos.
- **Agentes de usuario:** Asociados a un individuo, a quien aportan información sobre las búsquedas de los usuarios.
- **Agentes de interfaces web:** Son gestores de consulta capaces de almacenar conocimientos y procesar información incompleta o de alto nivel.
- **Agentes compradores web o shopbots** buscan información sobre productos en venta y comparan características y precios entre ellos para ofrecer al usuario los productos más adecuados a sus necesidades.

Características típicas(1)

- **Autonomía**: Trabajar sin supervisión humana, una vez fijadas las condiciones y restricciones necesarias, se espera que el agente intente cumplir sus objetivos.
- **Inteligencia**: Existen diferentes conceptos que pueden cubrir este rasgo, siendo cualesquiera de ellos el empleado para aseverar que el agente es inteligente.
- **Cooperación**: Ser capaz de colaborar con otros agentes, intercambiando informaciones resultados de acciones propias. La cooperación requiere que exista algún mecanismo de negociación entre agentes.

Características típicas(2)

- **Comunicación**: Implica tanto la capacidad de comunicarse con el usuario, así como también tener conocimiento sobre el mundo o dominio sobre el cual opera el agente.
- **Reactividad**: Poder responder ante eventos, tomando sus propias decisiones, inclusive modificando su manera de operar, teniendo en cuenta siempre lograr sus objetivos.
- **Adaptatividad**: Aprender de experiencias pasadas, o de otros agentes, asÃ como de las reacciones del usuario antes resultados previos. (Aprendizaje automático)

Agentes inteligentes y recuperación de información

La exploración automática de la web puede ser una tarea bien ejecutada por un agente. Esto se puede haciendo que las necesidades informativas requeridas sean parte de las especificaciones iniciales del agente; éste exploraría la red, eligiendo los enlaces más prometedores, accediendo a nuevas páginas, recopilando aquellas que puedan satisfacer las especificaciones iniciales. Este tipo de enfoque para abordar el problema de la exploración web tiene ciertas limitaciones, siendo una de estas el tiempo, dado que, aún de forma automática, explorar la web requiere mucho tiempo, por lo cual se espera que el agente entregue buenos resultados en un intervalo de tiempo razonable, definido algunas veces por el usuario. También se restringe la cantidad de resultados recuperados por el agente, compensando esto al incrementar la precisión de los mismos.

Puntos de partida(1)

Dado que el agente explorara una gran cantidad de páginas, es necesario establecer un punto de partida; típicamente la web se representa como un grafo dirigido, donde las diferentes páginas son los nodos y los enlaces son los arcos del grafo.
La exploración parte de un nodo, y utilizando los arcos explorar los demás. Previo a esto, es necesario localizar los nodos o puntos de partida que puedan estar lo más cercanos posible a las páginas relevantes para las necesidades de información del usuario.

Puntos de partida(2)

Un enfoque usado para elegir buenos puntos de partida es comenzar el trabajo del agente con una búsqueda clásica en bases de datos de diferentes buscadores convencionales, lo cual implica que dicha búsqueda sea enviada a metabuscadores. Por tanto, inicialmente el agente enviará la consulta al metabuscador, recogiendo las páginas que le sean devueltas, tales páginas son las candidatas a ser puntos de inicio de la exploración. Estos puntos se pueden manejar secuencialmente, empezando la exploración en cada uno o bien en paralelo, utilizando varios agentes para ello. En este último caso, los agentes deben hacer uso de la cooperación tanto para compartir criterios de selección de páginas como para evitar exploraciones en los mismos nodos.

Puntos de partida(3)

Explorar con varios agentes diferentes puntos de partida tiene la ventaja de permitir utilizar procesamiento paralelo, además de obviar en cierta medida problemas de comunicaciones, como cuellos de botella, servidores lentos, etc., conllevando a una mejora en el tiempo de respuesta.
También se debe considerar que, al tener varios puntos de inicio para explorar, podemos previamente priorizar una parte de ellos, esto puede ser de varias maneras, eligiendo los **n** primeros, aplicar medidas de similitud entre las especificaciones del usuario y el contenido de las páginas, o bien dejando que el usuario seleccione aquellos que considere como mejores puntos de partida.

Activación de enlaces

Dado un punto de partida, el agente debe extraer los enlaces que ese punto contenga y guardarlos en una lista, posteriormente irá tomando enlaces de la lista, recuperando las páginas a las que apuntan y así sucesivamente. Si la exploración se realiza entre varis agentes, la lista debe ser compartida de alguna forma para evitar duplicar exploraciones. El almacenamiento y seguimiento de todos los enlaces llevaría, teóricamente, a la exploración de la web. Debido a las limitaciones de recursos, capacidad de procesamiento y tiempo, es necesario establecer un orden de prioridad para los enlaces a explorar, atendiendo a dos premisas fundamentales: la relevancia del enlace respecto a la necesidad del usuario, y las posiblidades de acceder a mayores espacios de exploración en ciertos enlaces respecto a otros.

Selección de enlaces prometedores

Para determinar la importancia de una página, una posibilidad consiste en utilizar los *blacklinks* de la misma, esto es, las páginas que tienen enlaces hacia la página en cuestión. La forma más simple es contar el número de *blacklinks*, pero el problema es disponer de dicha información. Un método más sofisticado que este conteo de *blacklinks* es el algoritmo **PageRank**, cuya idea básica es que la importancia de una página es directamente proporcional al número de *blacklinks* que ésta tiene, pero no todos los *blacklinks* pesan lo mismo, sino que su valor está en función de la importancia de la página que proceda, y ésta página a su vez esta valorada de igual forma.

PageRank

Según este algoritmo, el cálculo del **PageRank** se debe hacer de forma iterativa, asignando de antemano pesos determinados a los nodos, ya sea de manera aleatoria o en función de algún otro criterio. Este método a pesar de sus ventajas, implica un cálculo costoso en términos de tiempo de proceso, un problema que también está presente al calcular otros tipos de coeficientes requeridos para estimar la importancia de unos nodos frente a otros.

Selección de páginas por contenido

Más allá de la importancia de una página, es de mayor relevancia disponer de medios para estimar la proximidad de un nodo a las necesidades del usuario, permitiendo esto seleccionar páginas para que el agente las entregue al usuario. Pero además de la importancia, determinar cuales enlaces son más prometedores para seguir la exploración.

Técnicas de recuperación de información

Al considerar cada página web como un documento en si, podemos aplicar técnicas habituales en recuperación información para estimar la semejanza entre una página explorada y las necesidades del usuario, comparando a través de estas técnicas el texto de la página web. Uno de los procedimientos más conocidos es el modelo vectorial, el cual opera con palabras o términos y calcula para cada uno de éstos un peso que trata de expresar la importancia de la palabra en cuestión. Este cálculo se basa en la estimación de la frecuencia de aparición de las palabras en el documento ponderandolo respecto a la frecuencia de aparición de las palabras en toda la colección de documentos. Debido a que en la exploración directa de la web se desconoce el segundo factor, se pondera la estimación de las palabras únicamente en virtud de su frecuencia dentro de la página donde ocurre.

Estudio de enlaces

La similitud documental se puede abordar desde el punto de vista de los enlaces, obviando el contenido de las páginas a las que apuntan, además la recuperación basada en exclusivamente los enlaces de las páginas web parece tener una efectividad a tener en cuenta; la similitud dependiente de los enlaces se define como:

$$\text{sin}_y^{link} = \frac{link_{ij}}{\sum_{i=1}^{n} link_{ij}}$$

$link_{ij}$: número de enlaces desde D_i a D_j en una colección de N documentos de la web.