Computación y Sistemas Vol. 2 No. 1 pp. 14-23
© 1998, CIC - IPN. *ISSN 1405-5546* Impreso en México

Un Enfoque Integrado para Disminuir la Sobrecarga en la Búsqueda de Información Utilizando un Agente Adaptivo Guiado por Interacciones Dialógicas en Lenguaje Natural

John A. Atkinson-Abutridy
Departamento de Ingeniería Informática
Universidad de Concepción
Concepción, Chile
atkinson@inf.udec.cl

Anita Ferreira-Cabrera
Facultad de Humanidades y Artes
Universidad de Concepción
Concepción, Chile
aferreir@udec.cl

Artículo recibido el 4 de agosto de 1998; aceptado el 18 de septiembre de 1998

Resumen

El presente artículo describe un enfoque integrado para búsqueda/filtrado de información en la recuperación de documentos. El enfoque contribuye con algunos aportes fundamentales, entre ellos se destacan el uso de agentes adaptivos que ayudan en el proceso de filtrado y en la utilización de mecanismos de interacción con el usuario vía análisis del lenguaje, como un medio de obtener una retroalimentación más precisa sobre lo que se busca. Se describe un sistema implementado sobre la base de dicho enfoque, el cual genera un universo de salida más reducido que en enfoques tradicionales. Finalmente, se muestran algunas pruebas y resultados obtenidos.

Palabras clave:
Agentes, Búsqueda de Información, Diálogo, Lenguaje Natural.

1 Introducción

En la actualidad, el crecimiento de Internet y sus servicios, tales como WWW, han abierto un mundo de posibilidades en cuanto a la información que ellos contienen. Dicha información puede llegar a ser abrumadora debido al crecimiento diario que experimenta la red. A pesar de que actualmente existe un cierto número de buscadores de información, el acceso a ésta continúa presentando varias limitantes:

- A partir de la información entregada por los buscadores, el usuario debe invertir bastante tiempo en verificar si en dichas referencias está exactamente lo que él necesita.
- La cantidad de información relacionada directa o indirectamente a un tema solicitado por el usuario es tan grande que la mayoría de las veces se deshecha, manteniendo sólo las primeras referencias.
- La búsqueda/recuperación de información que normalmente se lleva a cabo a partir de palabras claves y es el único filtro existente en la mayoría de las aplicaciones actuales.
- La búsqueda de la información solicitada no toma en cuenta aspectos lingüísticos relacionados con el tema, lo cual puede ayudar a entregar una mayor precisión sobre éste.
- No existe interacción o retroalimentación en el proceso de búsqueda/filtrado de información. Las decisiones dependen directamente de las consideraciones del sistema.

Desde un punto de vista cognitivo, los sistemas y enfoques utilizados actualmente para la búsqueda/recuperación de información, en su mayoría no toman en

cuenta el modelo del usuario, para dar cuenta de su comportamiento, intereses, objetivos y/o deseos. Incluso, es difícil pensar que el sistema pudiera automatizar en forma inteligente ciertas tareas, ya que se desconoce la perspectiva del usuario.

En la actualidad, existen algunos sistemas y motores de búsqueda en Internet que tratan de solucionar parcialmente algunos problemas típicos: tales son los casos de *Webwatcher*, que aprende a partir de las reacciones del usuario, y *Letizi* [Maes, 93], que trata de anticipar las preferencias del usuario, en base a experiencias y comportamientos previos [Dent, 92; Belkin, 92; Etzioni, 95].

Otro grupo, se basa en el hecho de que se puede lograr una respuesta más óptima y un mejor rendimiento dando diferente tipo de recomendaciones al usuario sobre tareas a llevar a cabo, o clases relacionadas de información a buscar, tal es el caso de los así llamados *Sistemas basados en Recomendaciones* [Terveen, 97; Beerud, 94], en donde el aporte radica en el hecho que las recomendaciones se producen en forma bi-direccional: puede hacerlas el sistema al usuario, o puede hacerlas el usuario al sistema. Sistemas basados en este enfoque incluyen *Phoaks y Siteseer*, entre otros.

En general, la utilización de cualquier sistema actual en conjunto con los medios de entradas de datos que éstos poseen, trae consigo varios problemas prácticos para el usuario:

- Navegar sobre WWW es una tarea muy engorrosa para usuarios con poca experiencia, o cuando hay una gran sobrecarga de información de salida.
- Especificar qué es lo que se desea buscar no es una tarea fácil para el usuario, ya que ni el propio usuario sabe a veces lo que quiere, y por otro lado, no existe ayuda adicional.
- Obtener mejores resultados implica muchas veces proporcionar al sistema diferentes alternativas, lo cual redunda en una gran pérdida de tiempo.

Una característica común de la mayoría de estos sistemas o motores de búsqueda actuales es la carencia de un análisis lingüístico, que los ayude a tener una visión más específica y real de lo que desea el usuario y de la precisión/calidad de la información que se encuentra. Algunas excepciones actuales incluyen el sistema *AltaVista*, que ha considerado en el proceso de búsqueda y recuperación, tanto la variable estadística como la lingüística, a la hora de enfrentarse al usuario y a la información encontrada.

1.1 Hipótesis

Considerando la situación actual y sus problemas relacionados, nuestro enfoque como aporte para solucionar los problemas se sustentan en las siguientes hipótesis:

- Disminuir la sobrecarga de información entregada al usuario implica realizar un filtrado "inteligente" de ella utilizando el medio ambiente y la retroalimentación del usuario.
- Hacer uso de la retroalimentación del usuario puede lograr una mayor precisión en la información entregada.
- Adaptar el sistema a las diferentes preferencias del usuario y al ambiente en constante cambio, le permite explorar nuevas alternativas.
- Considerar la componente lingüística como aspectos básicos de operación puede ayudar en gran medida a especificar y delimitar los requerimientos del usuario y a detectar "conocimiento" del usuario que no es posible a través de otros medios.

2 Trabajo Previo

El trabajo de este artículo se circunscribe en la intersección de tres áreas de investigación distintas denominadas Recuperación y/o Filtrado de información, Agentes de Software y Análisis de lenguaje natural. Recuperación de información (IR) es un área de investigación bien establecida que se dirige a la recuperación de un gran conjunto de documentos en respuesta a consultas del usuario. La literatura de IR hace pocos años ha empezado a considerar aspectos de Filtrado de Información. Por comparación, la investigación en Agentes inteligentes es un campo relativamente nuevo de estudio emanado desde la Inteligencia Artificial [Maes, 94; Steels, 95]. La investigación en Agentes de Software se relaciona con aspectos del diseño de software autónomo e inteligente para una variedad de tareas. Por otro lado, el análisis de lenguaje natural se relaciona con el procesamiento de entradas en lenguaje natural que siguen ciertos modelos de lingüística/computacional. Desde esta última perspectiva el trabajo considera el enfoque de las influencias de las investigaciones llevadas a cabo por [Ferreira, 98], en el área de procesamiento de diálogos utilizando modelos de lingüística computacional. Este artículo describe un trabajo realizado en un enfoque integrado que reuna las tres áreas mencionadas.

2.1 Filtrado de Información

En contraste a IR, el filtrado de información (IF) ha surgido sólo recientemente. Se pueden distinguir tres tipos de sistemas de filtrado de información que dependen de la manera en que el usuario seleccione los documentos:

- *Cognitivos*: seleccionan los documentos basados en las características de su contenido.
- *Sociales*: seleccionan documentos a partir de recomendaciones de otros usuarios.
- *Económicos*: seleccionan documentos basados en algún cálculo de costo/beneficio para el usuario.

Se han utilizado varios enfoques para obtener los contenidos semánticos de los documentos. Algunos de estos sistemas incluyen *Oval*, que utiliza un enfoque basado en palabras claves, *Foltz* que realiza indexación semántica latente para filtrar artículos de noticias e INFOSCOPE que utiliza agentes basados en reglas que observan el comportamiento del usuario y hacen sugerencias [Beerud, 94; Belkin, 92].

Tanto los enfoques cognitivos como sociales son válidos para selección de documentos. La diferencia radica en el hecho que dependiendo del área de aplicación, uno es más ventajoso que el otro. Por ejemplo, si la información que se reúne para mantenerla actualizada en cierto ambiente, un filtrado social es la forma más adecuada para elegir. Sin embargo, si la información se reúne basada en un tópico, independiente de quiénes son los otros usuarios, son más apropiados los sistemas cognitivos.

Por otro lado, las capacidades de aprendizaje y adaptación adquieren mayor importancia en los contextos de IF que los de IR, esto debido a las características inherentes del ambiente: los sistemas de filtrado son utilizados por grandes grupos de personas, muchos de los cuales no son buscadores de información muy motivados. Los intereses no siempre están bien definidos o puede que no siempre se expresen. Por ésta y otras razones, los sistemas de filtrado deben entonces ser capaces de responder a intereses dinámicos de los usuarios.

2.2 Agentes de Software

En términos simples, un agente es un sistema que en base a ciertas entradas trata de satisfacer sus objetivos en un ambiente dinámico y complejo. Estos agentes están situados en el ambiente e interactúan con él a través de sensores y actuadores. Cuando son autónomos, éstos operan en forma totalmente autónoma, y mejoran a través del tiempo al alcanzar sus objetivos.

Uno de los enfoques que se puede emplear para construir agentes de software es utilizar técnicas de aprendizaje computacional. Su objetivo es construir agentes que adquieren su competencia y se adaptan a los requerimientos de los usuarios. El agente puede aprender observando al usuario e imitándolo, reaccionando a retroalimentación del usuario y aprendiendo de ejemplos provistos por el usuario.

2.3 Análisis del Lenguaje Natural

El análisis del lenguaje natural es un tópico de investigación bastante amplio y que desde hace bastantes años ha tomado mucha importancia en la comunidad de la ingeniería lingüística y de lingüística computacional. El análisis tiene que ver con el procesamiento sintáctico/semántico de entradas en lenguaje natural empleado por los usuarios, para finalmente generar ciertas estructuras de representación semántico/pragmática que permitan dar cuenta en forma no ambigua de los fenómenos y las características de la entrada para su posterior tratamiento computacional.

Hasta el momento, los procesos de recuperación /filtrado de información desde documentos (para nuestro caso, referencias a documentos en Internet) sólo han tomado en cuenta aspectos matemáticos/estadísticos.

Uno de los grandes avances en este sentido, desde el punto de vista de la integración de lingüística computacional con problemas de búsqueda de información, ha sido el llevado a cabo por [Ferreira, 98]. Dicha investigación ha ido más a fondo en las interacciones dialógicas vía lenguaje natural y han propuesto modelos cognitivos/computacionales para realizar la generación automática de diálogos en lenguaje natural por parte del sistema, lo cual constituye uno de los aportes fundamentales en las mejoras y la naturalidad de la interacción usuario-sistema en el contexto de la recuperación de información.

Para el desarrollo del actual trabajo se diseñó e implementó un sistema de interpretación de consultas en lenguaje natural que permitiera captar tanto aspectos lingüísticos básicos de la entrada como servir de alimentación a los procesos de inferencia y búsqueda posterior. La construcción del sistema de interpretación mencionado fue concebido utilizando la herramienta para generación de interfaces en lenguaje natural GILENA [Atkinson, 91; Atkinson, 95].

Para el presente caso, el corpus del dominio de entrada que se procesa, fue extraído a partir de un estudio experimental de casos considerado en la concepción del modelo descrito en [Ferreira, 98] y que se realizó a un cierto número de usuarios que dominan claramente el uso de herramientas de navegación y motores de búsqueda. Para completar el ciclo, se pensó el sistema de tal forma que las interacciones fueran en forma de diálogo bidireccional. Es decir, tanto el sistema como el usuario pueden realizar preguntas, afirmaciones u otro tipo de transacción, con el fin de acotar el dominio sobre el cual desea buscar la información el usuario.

3 El Sistema de Búsqueda Interactiva

La utilización de sistemas de búsqueda de información basados en WWW conlleva un gran flujo de información, siendo del usuario la responsabilidad de su revisión en forma manual. La propuesta considera un Agente capaz de filtrar la información al usuario, a través de diferentes interacciones dialógicas vía lenguaje natural [Zorrilla, 98] y se basa fuertemente en los desarrollos y modelos propuestos por [Ferreira, 98].

Como se muestra en la figura 1, nuestro modelo se compone de tres componentes fundamentales: Una interfaz en lenguaje natural capaz de procesar los requerimientos de búsqueda del usuario y un Agente Inteligente, encargado de realizar las búsqueda, filtrado y preprocesamiento de información, tomando para ello lo aprendido de la interacción con el usuario.

La operación del sistema está representada como se muestra en la figura 1. La interacción comienza con la petición del sistema de algún requerimiento del usuario, luego esta "consulta" es procesada por un sistema de lenguaje natural a partir del cual se extrae la información clave más importante. A continuación se representa "vectorialmente" la información extraída según los patrones comunes detectados en las búsquedas. Paralelamente un Agente adaptivo toma dicho vector y realiza la primera búsqueda real de información que se traducirá en la entrada para el resto de las interacciones del sistema con el usuario. Una vez que se encuentra la información posible se vectoriza y el agente por medio de un proceso de aprendizaje estadístico trata de encontrar los vectores de documentos (referencias) que son más similares a la información actualmente disponible. Por cada interacción positiva por parte del usuario, nuevos slots se llenan en su vector y se realiza un proceso de filtraje local, después del cual el sistema es capaz de generar otra pregunta en lenguaje natural para reducir el universo de posibles referencias encontradas.

La interacción termina cuando se obtiene un número de referencias a documentos pequeña, o cuando no se puede realizar mayores filtrados.

3.1 Representación

La representación utilizada para los documentos y el perfil del usuario se basa en la representación vectorial típica usada en el ámbito de recuperación de información. En la representación de espacios vectoriales, tanto los documentos (páginas) como las consultas se representan como vectores en algún espacio multidimensional. Cuando se procesa una consulta ésta se traduce a un patrón que representa su espacio vectorial de criterios y luego utilizando métricas de distancia se recuperan los documentos (páginas) en base a sus "distancias" a la búsqueda deseada.

Los vectores se estructuran en base al siguiente esquema:

Luego que se ha generado el patrón de criterios, éste se estructura en un vector unidimensional que servirá posteriormente para realizar el proceso de entrenamiento con la información encontrada en Internet, el cual se representa de la forma:

$$V_i = X_0 X_1 X_2 X_3 ... X_n$$

Donde cada X_i representa el valor que se ha extraído de la entrada del usuario para el criterio o campo i-esimo de la página o documento encontrado.

Es importante destacar como se mencionó antes, que esta vectorización es la misma que se utilizará para todas las referencias que el Agente encuentre en la red relativas a los requerimiento del usuario, de modo que las etapas de aprendizaje harán uso tanto del vector de entrada como los vectores de las referencias WWW obtenidas.

Los criterios corresponden a aspectos relevantes que poseen las referencias a páginas WWW que podrían ser de utilidad en el momento de entrenar los patrones y de filtrar la información. Inicialmente, el criterio X_0 corresponderá al *tema* o tópico de la consulta y el resto del vector estará vacío.

En la fase experimental llevada a cabo por [Ferreira, 98] se seleccionó un conjunto de 20 individuos con experiencia en WWW y en la búsqueda de información, para llevar a cabo un proceso de interacción simulada en la cual ellos, buscaban información, describían/explicaban al sistema lo encontraban, y mediante un proceso iterativo, finalmente expresaban su satisfacción o no satisfacción en relación a sus expectativas.

A partir de dichos estudios, se logró extraer y sintetizar los criterios más frecuentes, utilizados por los usuarios para seleccionar información útil en el WWW. Algunos de ellos son los siguientes:

- Dirección URL de la página referenciada.
- Idioma en el cual está escrita la información (página).
- País de procedencia de la página (o referencia)
- Tipo de página (comercial, educacional, etc)
- Referencias relacionadas a eventos
- Documentación técnica
- Grupos de investigación
- Productos y Servicios

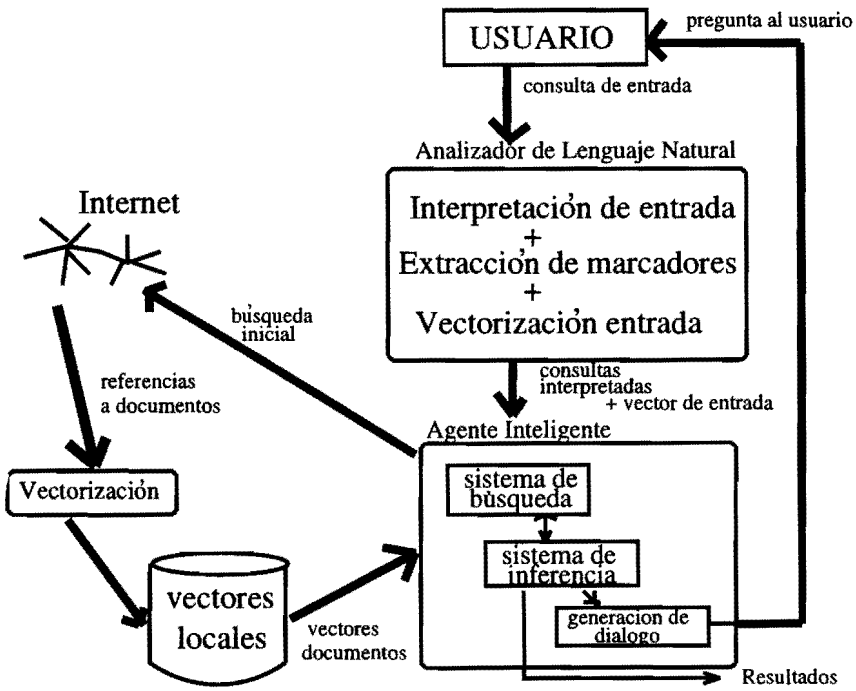

Figura No.1: El Sistema de Búsqueda/Filtrado de Información

Por ejemplo, en la componente X_4 podría indicar el criterio "idioma", por lo cual si en una posterior interacción con el usuario se determina que su búsqueda está relacionada con documentos en idioma *español*, el slot X_4 se llenará con el valor *español*.

Inicialmente, para el vector de entrada, sólo el slot X_0 estará lleno y corresponderá al criterio "tema" o tópico de la consulta. El resto de los criterios se consideran como valores sin importancia inicialmente.

Cada vez que el sistema genera una pregunta al usuario, existe algún criterio que se va a cumplir por lo que se llenarán más criterios hasta que exista información adecuada y suficiente para realizar una búsqueda/filtrado más específico.

Adicionalmente, cada criterio tiene asignado un cierto "peso" que representa la contribución de ese "término" en un documento determinado o bien la importancia que un cierto campo o criterio tiene sobre otro.

El valor de los criterios para los vectores de todos los documentos se obtiene a través de un análisis del texto completo de los documentos. El peso del criterio depende de su frecuencia de ocurrencia y de su importancia relativa en comparación al resto de los criterios o campos.

3.2 Interpretación de consultas en Lenguaje Natural

La operación del sistema comienza con la interacción del usuario. Este proporciona una cierta consulta que es procesada lingüísticamente. Según estudios de [Ferreira, 98] las acciones llevadas a cabo en el diálogo responden a *Actos de Habla* [Searle, 69] que para el caso de este tipo de interacción lo ha dividido básicamente en cuatro grupos, desde los cuales comenzará la posterior generación de respuestas en lenguaje natural por parte del sistema:

- Petición de información
- Confirmación o Negación de propuestas
- Especificación de cuantificaciones
- Especificación de tópicos

Con cada una de las cuales, el sistema es capaz de generar alguna respuesta adecuada dependiendo de interacciones previas, la información encontrada y las inferencias realizadas por el agente de búsqueda/filtrado de información.

El sistema de procesamiento de dichas entradas del usuario en lenguaje natural, se diseñó utilizando el sistema de generación de interfaces en lenguaje natural GILENA [Atkinson, 91].

Posterior al procesamiento sintáctico necesario se obtiene la representación semántica simplificada del requerimiento del usuario que representa entre otros aspectos los criterios presentes en la interacción. Esta representación será traducida por un sistema de vectorización a un espacio multidimensional de criterios en la forma como se describió anteriormente.

Un ejemplo de interacción típica usuario-sistema que depende de la información encontrada, el peso de sus criterios y las respuestas del usuario, es el siguiente:

```
<SISTEMA>: En que lo puedo ayudar?
<USUARIO>: Deseo conocer todas las paginas que
           tengan relacion con ontologia
<SISTEMA>: Encontre varias, en que idioma las
           prefiere?
<USUARIO>: Las prefiero en Ingles
<SISTEMA>: Esta buscando algun tipo de pagina
           en particular?
<USUARIO>: Si, las de conferencias
...
```

3.3 Agente adaptivo para filtrado de información

El proceso de filtrado consiste en traducir los documentos a sus correspondientes representaciones vectoriales, encontrar los documentos que son similares a los perfiles extraídos desde los requerimientos del usuario y seleccionar las mejores páginas para mostrarlas al usuario. En nuestra propuesta los resultados de las búsquedas no se visualizan hasta que el filtrado no haya alcanzado un punto máximo que está determinado por las características de la información encontrada y de la interacción con el usuario.

De este punto de vista, un agente adaptivo está continuamente recibiendo entradas del usuario, aprendiendo de los procesos de búsqueda/filtrado que lleva a cabo y generando salidas (interacciones) al usuario.

El agente es el encargado de tomar el vector del usuario y con él realizar las siguientes tareas:

- **Búsqueda/Filtrado:** Se realiza la búsqueda de información en la WWW (referencias a páginas) a partir del vector de criterios iniciales. Luego, un proceso de vectorización, al igual que con la entrada del usuario, vectoriza la información contenida en las referencias obtenidas, y los mantiene en bases de datos locales.

- **Aprendizaje:** Se realiza el proceso de aprendizaje con los vectores obtenidos. Posteriormente, se filtran aquellos vectores cuyos criterios coincidan en mayor porcentaje con los del vector de entrada. Debido a que las similitudes no siempre son exactas, se aplica un proceso de aprendizaje estadístico/bayesiano en el cual se determinan los vectores más cercanos dependiendo de los criterios que se estén utilizando.

- **Generación de Preguntas:** Se generan diferentes tipos de preguntas en lenguaje natural dependiendo de las características de la información encontrada por el agente, debido a que la retroalimentación del sistema se obtiene a través de diálogos con el usuario. De modo de disminuir la sobrecarga de información eventualmente mostrada al usuario, se tienen dos alternativas: si la cantidad de referencias es pequeña, se le muestra al usuario toda la información encontrada y él puede seguir interactuando con el sistema si no está conforme.

  Si la cantidad de referencias es muy grande, se deben filtrar a medida que el usuario especifica más detalles. Esto se logra preguntando al usuario por criterios que aún están vacíos, y que al llenarlos se filtren el número de referencias que posteriormente al entrenamiento logren los requisitos. Por ejemplo, en un instante se encontraron **N** referencias en diferentes idiomas (N vectores), por lo cual se le podría preguntar al usuario por algún criterio vacío tal como "idioma", lo cual producirá un filtrado de los vectores que "calcen" con ello, y ese número N podría ser mucho más pequeño. Los criterios por los cuales preguntar dependen de tres factores:

  1. Preferencias del usuario aprendidas por el agente.
  2. Información estadística de los requerimientos típicos.
  3. Estados de los valores de los criterios en un momento determinado.

4 Aspectos de Implementación

El proceso de interacción del sistema está separado en dos fases: una para la recuperación inicial de los documentos candidatos y otra interactiva que permite ir filtrando y aprendiendo de las referencias a documentos que sean las más adecuadas a los requerimientos del usuario.

El proceso de interacción entre el usuario y el sistema con el objetivo de buscar y filtrar información para cumplir los objetivos del usuario se puede dividir en dos etapas internas:

1. Recuperación inicial de los documentos factibles dada la primera interacción con el usuario.

2. Selección de los documentos más importantes en base al resultado de la retroalimentación del usuario en las interacciones siguientes.

Inicialmente, todos los documentos que se extraen son vectorizados y colocados en una memoria temporal para un acceso más expedito. El manejo de los datos presentes en los vectores se representa simbólicamente en grafos dirigidos que almacenan adicionalmente los pesos de las variables de los vectores y los factores de confidencia para poder realizar posteriores inferencias.

Las predicciones o inferencias relativas a qué acción tomar y qué salidas generar por el agente de interacción están dadas por dos elementos:

1. La información de slots disponibles en los vectores de cada referencia a documento.

2. La utilización de mecanismos automáticos de inferencia simples, cuando la información anterior no es suficiente o es incompleta.

El último factor se considera como complemento en la mayoría de los casos y opera en la forma de inferencia estadística a partir del peso o el porcentaje de confiabilidad de una cierta característica de los vectores.

Los resultados de la inferencia tienen dos consecuencias básicas importantes: en el filtrado de la información necesario para el usuario que interactúa y en la generación de las preguntas del sistema que forman parte de la interacción con dicho usuario.

La figura 2 muestra un algoritmo simple para realizar el cálculo de niveles de confidencia adecuados para las decisiones mencionadas anteriormente, y la ejecución de las acciones posibles. En el caso de la interacción, las acciones posibles se traducirán como un determinado tipo de pregunta o respuesta, todo lo cual se determina en base al historial de referencias de los documentos recuperados.

En términos generales, este mecanismo permite generar el tipo de interacción de acuerdo a las inferencias realizadas sobre el tipo de atributo más comúnmente referenciado en los documentos. Es decir, si un slot del vector de entrada tiene un atributo para el cual el factor de confidencia es grande (por ejemplo, autor del documento), implica que se generará una pregunta que filtre información deseada por el usuario, en relación a dicha frecuencia, la que posteriormente se actualizará dependiendo de la retroalimentación positiva o negativa que entregue el usuario. Este "feedback" queda representado por un lado por respuestas positivas del usuario, y por el otro, respuestas negativas, inconformidad de la información entregada o imprecisión de ella.

Al término de dicho algoritmo el sistema está en condiciones de mostrar el conjunto final de documentos referenciados que son apropiados para las preferencias del usuario.

```
ALGORITMO: Ejecutar_Accion_con_Nivel_de_Confi-
           dencia
ENTRADAS: slot_vector, situacion
SALIDAS: accion_a_realizar

 INICIO
  MIENTRAS (exista informacion para calcular
            NC)  HAGA
  NC <-- Probabilidad de Combinacion de la
         lista de "situacion"
  Donde Probabilidad = (referencias a
    situacion en slot_vector)/Suma total de
    referencias a criterios de slot_vector
  SI (NC >= 0.5) y (NC<=0.96) ENTONCES
    Agente sugiere con un grado de
      confiabilidad dado por NC
    SI (Usuario_Acepta?) ENTONCES
      Incrementar la frecuencia de repeticion
        del par situacion/accion
      Agregar nueva Situacion
      Realizar el filtrado y re-vectorizacion
        de los documentos dependiendo de
        "situacion"
    SINO
      Salir
    FIN-SI
  SINO
    SI (NC >0.96) ENTONCES
      Agente genera pregunta dada por la
        accion de mayor frecuencia en
        "situacion"
      Incrementar frecuencia de repeticion
        de la accion
      Situacion <-- accion de salida con
        mayor frecuencia de uso
    SINO
      SI (NC=0) ENTONCES
        Continua-interaccion
      SINO
        Generar-pregunta-por-mas-datos
    FIN-SI
   FIN-SI
  FIN-MIENTRAS
 FIN
```

Figura No.2: Un Algoritmo Simple para Generación de Acciones

5 Un Ejemplo

Para dar una visión más clara de como opera el sistema, se dará un ejemplo simple de una operación típica a

partir de la consulta del usuario:

```
"necesito informacion sobre orientacion
   a objetos"
```

Una vez analizado, se obtiene la información relativa al tema de la consulta: `orientacion a objetos`.

La etapa posterior considera la vectorización de la entrada. Si suponemos que la estructura del vector es:

$$X = \{X_{tema}, X_{dir\_URL}, X_{idioma}, X_{tipo\_evento}\}$$

el contenido del vector de entrada inicial será:

V_0={"orientacion a objetos",vacio,vacio,vacio}

Luego, se realiza el proceso de búsqueda en Internet después del cual se obtiene un conjunto de vectores, como los que se muestran a continuación:

```
V1=("orientacion a objetos","www.utfsm.cl",
     espanol,charla)
V2=("orientacion a objetos","www.cs.cmu.edu"
     ,espanol,seminario)
V3=("orientacion a objetos","www.cs.stanford
     .edu",espanol,charla)
V4=("orientacion a objetos","www.nacional.
     ar",espanol,conferencia)
```

El etapa de aprendizaje y entrenamiento siguiente, detecta que el criterio `idioma` es el mismo en todas las referencias, por tanto no es necesario generar una pregunta sobre el idioma al usuario. En caso contrario, se producirá una pregunta al usuario de modo de filtrar las páginas con diferentes idiomas (si es que le interesa en qué idioma esté). Para el caso actual, la inferencia produce que el vector de entrada quede de la siguiente forma:

```
V0={"orientacion a objetos",VACIO,espanol,
    VACIO}
```

En el siguiente paso se continúa analizando las referencias, para otra posibilidad de filtrado. Por medio de mecanismos internos se determina que aún es posible reducir el número de referencias, tratando de llenar el criterio `tipo_evento`, por lo cual se genera una pregunta al usuario que se traduce en una de las interacciones como la siguiente:

```
<SISTEMA>: En que tipo de evento esta
           interesado?
<USUARIO>: me gustaria una charla
```

Posteriormente, se analiza la entrada del usuario, se actualiza el vector de entrada y se obtiene:

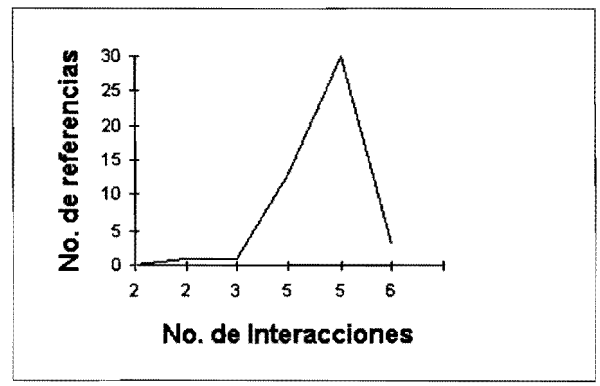

Figura No.3: Caso 1

```
V0={"orientacion a objetos",VACIO,espanol,
    charla}
```

Para finalmente, si es que el usuario está conforme, mostrarle la salida dada por la referencia a las direcciones URL, de los siguientes vectores:

```
V1=("orientacion a objetos","www.utfsm.cl",
     espanol,charla)
V4=("orientacion a objetos","www.cs.stanford
     .edu",espanol,charla)
```

6 Pruebas y Resultados

Con el fin de analizar y evaluar el rendimiento del sistema, se llevaron a cabo dos experimentos en los cuales el usuario interactúa con el sistema para encontrar información con la mayor precisión posible, sobre ciertos tópicos de interés. Estas sesiones miden el número de interacciones usuario-sistema versus el número de referencias a documentos finalmente entregadas al usuario. Inicialmente el universo de posibles documentos que calzan en las consultas ascienden a 30.000 referencias pero con fines prácticos del estudio y de un análisis cualitativo, dicho número se ha reducido de modo que el universo promedio con el que trabajarán los usuarios será de aproximadamente 50 referencias. El primer caso de prueba (figura 3) involucra la búsqueda de información para el concepto o clave: **Java**, y el segundo (figura 4), para el concepto: **Animaniacs**. Para fines de comprensión cada interacción se definirá como uno o más diálogos pregunta-respuesta entre el usuario y el agente.

En las interacciones del caso 1, se aprecia que hay un incremento en las referencias encontradas cuando hay más de tres interacciones. Ello no surge de un factor accidental o numérico, sino que para un mismo número de interacciones, por ejemplo; cinco, se aprecian diferentes salidas, esto debido a la forma adaptiva en la cual se va

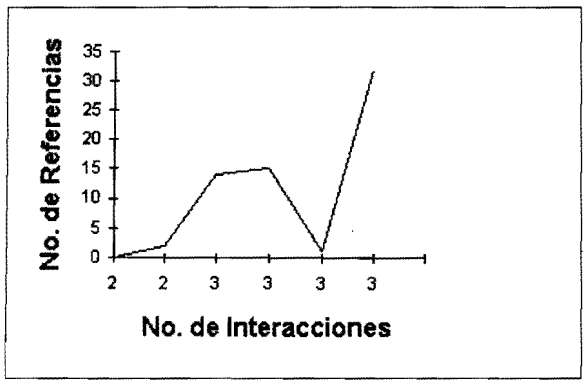

Figura No.4: Caso 2

dando el diálogo usuario-sistema, ya que va cambiando el contexto o el tipo de pregunta que realiza el agente dependiendo de la situación y contenido de los documentos encontrados. Por ejemplo, en dos sesiones con 5 interacciones se tienen diferentes resultados de salida, esto debido a que en una, se cambia el tipo de página deseada (dependiendo del análisis del contenido) y se restringen otros aspectos. Situación similar ocurrirá, si la restricción tiene que ver con el filtro relativo al lenguaje en el que está la mayor parte de las referencias, más aún, en ocasiones esto no produce ninguna referencia de salida.

En las interacciones del caso 2, ocurre una situación similar, incluso en el caso de diálogos con tres (3) interacciones se observan incrementos más "bruscos" en donde se produce un salto de 1 referencia a casi 35. Esto se debe a una inferencia realizada por el agente y a una restricción por parte del usuario relativa a la naturaleza del tipo de documento que desea.

De ambos casos se puede deducir que se producen grandes decrementos en las referencias de salidas finalmente entregadas, debido a restricciones en el tipo y naturaleza de las páginas asociadas a cada referencia electrónica entregada por los buscadores. El agente que está integrado en el modelo, toma en cuenta lo anterior, por lo cual existen cierto tipo de "preguntas" o tipos de "afirmaciones" que tienen mayor probabilidad que otras, dependiendo del contexto del diálogo.

7 Conclusiones

El trabajo presentado ha mostrado un enfoque y un conjunto de estrategias simples que se pueden llevar a cabo para resolver parcialmente el problema de la sobrecarga de información.

Las hipótesis iniciales de enfrentar el problema ayudándose de la retroalimentación del usuario y de las capacidades de inferencia del agente buscador, es una realidad que se ha comprobado experimentando en situaciones de mediana complejidad, cuyos resultados se han descrito.

Sin lugar a dudas se podría ir más a fondo en la recuperación de datos textuales, sin embargo, el objetivo del trabajo ha sido dar un enfoque más global e integrador de varios elementos, más que preocuparse cabalmente de uno de ellos.

Los experimentos y pruebas realizadas, dejan en explícito el hecho de que se puede ganar mucho tiempo en las interacciones, si hay un estudio previo que permita ponderar las diferentes características de las referencias a documentos dependiendo de su grado de importancia o uso. Lógicamente, las interacciones tanto en su forma como contenido dependerán fuertemente de estos factores, pero no se debe dejar de lado el aporte del usuario en las decisiones que podría generar el sistema.

El enfoque y el sistema experimental desarrollado nos brinda la ventaja de evitar el gran sobreflujo de información innecesaria e incompleta al usuario. A pesar de que el tamaño de los experimentos y del diseño de los sistemas, son de mediana complejidad, las consideraciones básicas que se han detectado no deberían cambiar drásticamente en desarrollos más avanzados.

Referencias

Atkinson, John. *GILENA: Generador de Interfaces en Lenguaje Natural*, Departamento de Informática, Universidad Técnica Federico Santa Maria, Valparaíso, Chile, 1991.

Atkinson, John. *A Flexible New Language for Specifying and Generating Natural-Language Systems*. Proceedings of the Third Natural Language Processing Pacific Rim Simposium, pp. 612-615, Korea, 1995.

Beerud Dilip S. *A Learning Approach to Personalized Information Filtering*. MSc thesis, MIT Dept. of EE and Computer Science, 1994.

Belkin, Nicholas. *Information filtering and Information retrieval: two sides of the same coin?*, CACM, pp. 29-37,Diciembre 1992.

Dent, Boticario. *A Personal Learning Apprentice*. In Proceedings of the National Conference on Artificial Intelligence. MIT Press, Julio 1992.

Etzioni, Oren. *Intelligent Agents on the Internet*. IEEE Expert, pp.44-49, Agosto 1995.

Ferreira, Anita. *Generación de Discursos Dialógicos*

Interactivos Explicativos y Descriptivos. PhD thesis, Departamento de Lingüística, Universidad Católica de Valparaíso, Valparaíso, Chile, 1998.

Maes, Pattie. *Evolving Agents for Personalized Information Filtering.* Proceedings of the Ninth Conference on Artificial Intelligence for Applications '93, Orlando, Florida, USA, Marzo 1993.

Maes, Pattie. *Social Interface Agents: Acquiring Competence by learning from users and others agents.* AAAI Spring Symposium, pp.71-58, Stanford University, Marzo 1994.

Searle, J.R.. *Speech Acts.* Cambridge University Press, Cambridge, 1969.

Steels, Luc. *The Artificial Life route to Artificial Intelligence.* Lawrence erlbaum Associates Pub. New Jersey, 1995.

Terveen, Loren. *PHOAKS: A System for sharing recommendations.* CACM, pp.59-62, Marzo 1997.

Zorrilla, Eduardo. *Diseño e Implementación de un Sistema de Búsqueda para Consultas en WWW vía Lenguaje Natural.* B.Eng. thesis, Departamento de Informática, Universidad Técnica Federico Santa María, Valparaíso, Chile, 1998.

***John Atkinson**, es Ingeniero y Master en Informática por la Universidad Técnica Federico Santa María de Chile. Tomó estudios especiales en el Royal Institute of Technology de Suecia.*
Sus áreas de investigación son: Agentes Inteligentes, Análisis del Lenguaje Natural, Computación Evolucionaria, Computación Paralela y Redes Neuronales. Actualmente es profesor asistente en el Departamento de Ingeniería Informática de la Universidad de Concepción, Concepción, Chile.

***Anita Ferreira**, es Master y Doctor en Lingüística por la Universidad de Concepción, y Universidad Catolica de Valparaiso, Chile. Sus áreas de investigación son: Procesamiento de Lenguaje Natural, Diseño de Multimedios para la Enseñanza de Lenguas y Aprendizaje de Lenguajes Asistido por computador (CALL). Actualmente es profesor asistente del Departamento de Lingüístico de la Universidad de Concepción, Concepción, Chile.*