

## **CAPÍTULO 2**

### **Recuperación de información, sistemas de apoyo para grupos y agentes**

Junto a los desafíos de cualquier biblioteca tradicional, las bibliotecas digitales enfrentan también otros retos de diferente índole, especialmente de tipo administrativo, tecnológico y de impacto social. Resalta por ejemplo, la necesidad de brindar atención personalizada a sus múltiples usuarios, considerando sus intereses y preferencias [Sánchez et al. 2001]. Por otro lado, las bibliotecas digitales han puesto su atención en técnicas automáticas que faciliten el manejo de los enormes espacios de información.

Se han incorporado tecnologías de información de diferentes áreas para responder a estos retos, como son los sistemas de búsqueda y recuperación de información, algunas características de ambientes que apoyan el trabajo en grupo y la implementación de tareas mediante agentes. En este capítulo se tratan estos temas y se enuncian algunos sistemas que los incorporan.

#### **2.1 Recuperación de información**

El objetivo de esta sección es delinear los conceptos y técnicas de sistemas de recuperación de información considerados en esta investigación, algunos de ellos implementados en EGRAI.

El área de recuperación de información, se encarga principalmente del estudio de sistemas y técnicas para asignar índices, buscar y devolver datos valiosos a los usuarios. Este conjunto de procedimientos se denomina también

*filtrado de información.* Por lo general se procesa texto o cualquier otra forma no estructurada [Salton y McGill 1983] ; [Faloutsos y Oard 1995].

Los aspectos de recuperación de información en este documento, cubren únicamente datos en formato tipo texto. En este sentido, se emplea la palabra *consulta* para denotar una solicitud del usuario, misma que expresa sus requerimientos de información. Por ejemplo, los enunciados "congresos de bibliotecas digitales" o "métodos de recuperación de información", se consideran consultas en un contexto de recuperación de información.

A los libros, revistas, cartas, telegramas, tesis, notas, disertaciones, mapas, fichas bibliográficas, páginas electrónicas o a cualquier otro recurso de información, se les denomina *documentos*. Un grupo de documentos del mismo tipo forma una *colección*.

Las consultas y documentos contienen texto en lenguaje natural, aunque en estos últimos, el texto puede representar su contenido o contiene únicamente datos que lo describen, (*meta datos*) [Faloutsos y Oard 1995]. El texto requiere de un procesamiento previo a la aplicación de los métodos de búsqueda y recuperación.

En el proceso previo, con frecuencia se emplean sólo letras minúsculas y se descartan palabras con poco contenido semántico, las cuales no son consideradas útiles en la recuperación porque no aportan información para describir la consulta o el documento. Estas palabras se conocen en inglés como *stopwords* y se eliminan del proceso de asignación de índices [Witten et al. 1999]. En español, a falta de una traducción exacta, en este documento se les denominan "*palabras irrelevantes*".

Aunque las palabras irrelevantes varían de acuerdo al contexto, muchos sistemas de recuperación de información cuentan con una lista de ellas. En su

mayoría, tienen la categoría gramatical de preposiciones, adverbios, sustantivos comunes o artículos. El apéndice II contiene dos listas de este tipo de palabras.

Una vez eliminadas las palabras irrelevantes, las palabras restantes se conocen como *términos*. Cuando el número de términos que describen a las consultas y documentos se reduce, la eficiencia de un sistema de recuperación se incrementa, porque disminuye la cantidad de almacenamiento requerida y el tiempo de respuesta [Witten et al. 1999].

Para reducir el número de términos en una colección, se emplea comúnmente el algoritmo de Porter [Porter 1980], el cual consiste en eliminar los sufijos de los términos, asignando un índice únicamente a su raíz. Por ejemplo: supongamos que un documento contiene los términos: biblioteca, bibliotecas, bibliotecario, bibliotecaria, bibliotecarios, bibliotecarias. Si el sistema de recuperación de información emplea el algoritmo de Porter, únicamente asignaría un solo índice en lugar de seis, este índice apuntaría al término "bibliotec".

Originalmente el algoritmo de Porter empleaba términos en inglés, pero dada la eficiencia que se obtiene en un sistema de recuperación de información que lo implementa y su uso ampliamente difundido en ámbitos de investigación y sistemas comerciales, se ha modificado para ser utilizado en otros idiomas, como el español [Gómez y de Buenaga 1996]. Los detalles del algoritmo, son tema de la siguiente sección.

### **2.1.1 Descripción del algoritmo de Porter**

El objetivo del algoritmo de Porter es mejorar la eficiencia de un sistema de recuperación de información al reducir el número de términos que requieren de la asignación de un índice, lo cual facilita el manejo de las colecciones

aunque resta importancia a consideraciones del procesamiento de lenguaje natural, como ciertos aspectos lingüísticos [Salton y McGill 1983].

Para ello, emplea una lista de sufijos junto con un criterio para indicar bajo qué condiciones el sufijo de la palabra puede ser eliminado, obteniendo únicamente su raíz. En la literatura en inglés, este proceso se conoce también como "*stemming*" [Witten et al. 1999]. En español, comúnmente se denomina "*obtención o extracción de la raíz*".

La traducción de la descripción del algoritmo citada en [Porter 1980] para términos en inglés es la siguiente:

---

Una *consonante c* en una palabra, es una letra diferente de *A, E, I, O, y U, o Y* precedida por una consonante. Si una letra no es consonante, entonces es una *vocal*. Una vocal se representa con la letra *v*.

Una lista de *consonantes* de longitud mayor que cero se denota por *C*, una lista de vocales de longitud mayor a cero se representa con *V*.

Cualquier palabra o parte de una palabra, puede representarse por:

[C] (VC) **m** [V],

donde **m** es parte de una palabra, conocida como *medida*.

*Observación:*

Los corchetes indican que sus argumentos son opcionales, y los paréntesis que la secuencia de sus argumentos aparece una vez.

Las reglas para remover un sufijo son de la forma:

**(condición) s1 ? s2,**

interpretada como: si una palabra tiene el sufijo s1 y la raíz de esa palabra, (desde la primer letra hasta la letra anterior a s1), satisface la condición dada, el sufijo s1 se reemplaza por el sufijo s2.

La condición puede ser alguna de las siguientes:

- \*S,** la raíz termina con "S", (similarmente para otras letras).
- \*v\*,** la raíz contiene una vocal.
- \*d,** la raíz termina con una consonante doble.
- \*o,** la raíz termina con *cvc*, donde la segunda *c* es no es "W" ni "Y".

Se emplean también condiciones con los conectivos lógicos (*and*, *or* y *not*):

1) **(m>1 and (\*S or \*T)),**

prueba una raíz con medida mayor a 1 que termina en "S" o "T".

2) **(\*d and not (\*L or \*S or \*Z)),**

prueba una raíz que termina con una consonante doble diferente de "L", "S" o "Z".

---

La eliminación de palabras irrelevantes y el algoritmo de Porter son métodos dirigidos a mejorar la ejecución de los sistemas de recuperación de información, la cual puede ser calculada por medidas estándares, mismas que se describen a continuación.

### 2.1.2 Medidas para evaluar la ejecución de los sistemas de recuperación de información

Una vez que se han recuperado documentos, es común asociarles un cierto valor que indique qué tanto han satisfecho las necesidades de información de los usuarios. Este valor se conoce comúnmente como *relevancia*, y a pesar de ser una medida subjetiva, se toma en cuenta en cualquier método de recuperación de información.

Utilizando la relevancia como punto de partida, existen medidas estándares para evaluar la ejecución en los sistemas de recuperación de información [Witten et al. 1999]. Una de ellas es la *precisión*, definida como el número de documentos relevantes entre el total de documentos recuperados.

Por ejemplo, supongamos que la colección *X* contiene 40 documentos, y que el sistema de recuperación de información *Y* devuelve 30, 20 de los cuales son relevantes. Entonces, la precisión del sistema *Y* es del 67% (20/30).

Otra medida utilizada ampliamente es la *devolución*, (del inglés "recall"), la cual se calcula cuando el número de documentos relevantes que satisfacen una consulta, se divide entre el total de documentos relevantes de la colección.

Como ejemplo, supongamos que en la colección *X* existen 40 documentos relevantes, y que el sistema de recuperación *Y* regresa 30 de los cuales sólo 20 son relevantes, por tanto, la devolución es del 50% (20/40).

Los sistemas de recuperación de información tienden a maximizar la devolución y la precisión de forma simultánea [Salton y McGill 1983] y [Faloutsos y Oard 1995], para ello se valen de diferentes métodos, (algunos de los cuales se describen en la siguiente sección), que han propiciado que los sistemas actuales puedan atender las solicitudes de los usuarios cada vez en menos tiempo.

### **2.1.3 Métodos para recuperar información**

Mediante la consulta, un sistema de recuperación de información recibe los datos necesarios para iniciar su función. Las consultas pueden ser expresadas de diferentes formas, por ejemplo, unas incluyen sólo palabras clave, otras consultas aceptan palabras clave y conectivos lógicos, y otras se expresan en lenguaje natural [Watstein y Kesselman 1988].

Los métodos de recuperación de información utilizados con mayor frecuencia son el modelo de espacios vectoriales, la recuperación booleana y variaciones de estos dos [Meadow 1992], donde cada uno de ellos establece criterios de similitud para comparar qué tan parecidos son dos términos, dos documentos, o un documento de una consulta. Se describen en las siguientes secciones.

#### **2.1.3.1 Recuperación booleana**

En este método, un documento está representado por un conjunto de palabras clave, (palabras con un valor semántico importante), las cuales pueden ser extraídas del contenido del documento, de una parte de éste o de sus meta datos. De forma análoga, la consulta es un grupo de palabras clave [Salton et al. 1983].

Para que un documento sea considerado relevante, debe contener los términos de la consulta [Meadow 1992], por ello se requiere que el usuario tenga cierto conocimiento del tema a buscar, de tal manera que las palabras de la consulta representen de la mejor forma posible y de manera concisa, su solicitud de información; de lo contrario, los resultados no serán satisfactorios.

Por lo general, en un sistema que emplea recuperación booleana, una colección se representa mediante un *archivo invertido*, el cual es un índice que relaciona cada término con una lista de documentos donde éste aparece [Witten et al. 1999].

La recuperación booleana extendida [Salton et al. 1983] es una variación de este método, la cual agrega conectivos lógicos que enlazan a las palabras clave. Debido a su rapidez, ambos métodos se usan con frecuencia en Internet [Meadow 1992].

Otro método que provee mayor flexibilidad que la respuesta afirmativa o negativa que ofrecen la recuperación booleana, es el modelo de espacios vectoriales, cuya descripción se presenta a continuación.

### **2.1.3.2 Modelo de espacios vectoriales**

Uno de los modelos automáticos más utilizado para recuperar información es el modelo de espacios vectoriales [Salton y McGill 1983]. En él, las consultas, los términos y los documentos se representan como vectores en un espacio con múltiples dimensiones. Un vector tiene tantas dimensiones como términos en el espacio del documento [Witten et al. 1999].

A diferencia de la recuperación booleana, en el modelo de espacios vectoriales se considera el carácter semántico de los documentos, mediante la asignación de pesos a los términos, que indica su presencia o importancia en el documento o en la colección [Balabanović 1998].

Aunque el modelo en sí no señala cómo asignar pesos a los términos, las técnicas más comunes son la *frecuencia del término*, entendida como el número de veces que aparece el término en el documento, la *frecuencia inversa del término*, calculada como la unidad entre el número de documentos que

contienen al término, o el producto de la *frecuencia del término por su frecuencia inversa* [Salton y McGill 1983]. Este último caso se conoce como la regla *TF X IDF* por sus siglas en inglés, (Term Frequency, Inverse Document Frequency) [Witten et al. 1999]. Frecuentemente, el peso para los términos de las consultas es uno.

En ocasiones los pesos de los términos de un documento se normalizan, de tal forma que sumen uno o se dividen entre la raíz cuadrada de la suma de sus cuadrados. La normalización se implementa para evitar que los documentos grandes tengan mayor probabilidad de ser recuperados en comparación con documentos pequeños, debido a que los documentos grandes serían representados por un número mayor de términos [Balabanović 1998] y [Witten et al. 1999].

Independientemente de la técnica utilizada para asignar pesos a los términos, debe tomarse en cuenta que un término que aparece en muchos documentos no debe considerarse más importante que un término que aparece en unos cuantos, y que un documento con muchas ocurrencias de un término, no debe ser tratado menos importante en relación a un documento con unas pocas [Witten et al. 1999].

A la cercanía de cualesquiera dos vectores que representan elementos del modelo de espacios vectoriales se le denomina *similitud*. Existen varias medidas de similitud, sin embargo, la más popular es el coseno del ángulo entre los vectores, donde un menor ángulo representa mayor similitud. Por ejemplo, si Q representa el vector de la consulta y D el vector de un documento, la similitud entre Q y D se calcula como:

$$\cos(Q,D) = (Q \cdot D) / (|Q||D|)$$

donde  $Q \cdot D$  es el producto interno de  $Q$  y  $D$ , y las barras verticales indican la norma euclidiana de los vectores [Witten et al. 1999].

Cuando el resultado de la aplicación de la fórmula anterior se aproxima a la unidad, quiere decir que los vectores son muy similares. En la práctica, dado que el número de términos de la consulta no siempre es igual a la dimensión de los vectores que representan los documentos, en la recuperación se elige un valor cercano a uno conocido como *coeficiente de correlación* [Salton y McGill 1983], a partir del cual, si el resultado de aplicar la medida de similitud a la consulta y a un documento  $X$  es mayor o igual al coeficiente de correlación, el documento  $X$  se considera relevante.

La colección puede ser descrita como una matriz  $A$  de  $t \times d$ , con  $t$  términos y  $d$  documentos [Berry et al. 1999]. Un componente  $a_{t,d}$  de la matriz, se interpreta como el peso del término  $t$  en el documento  $d$ , en la colección o en ambos. En el ejemplo siguiente se muestra una aplicación del modelo de espacios vectoriales.

### **Ejemplo:**

Supóngase una colección  $C$  de cuatro documentos ( $d_1, d_2, d_3, d_4$ ). Se emplean los términos del título para representar cada documento. Sea "agentes de software" la consulta  $c$  del usuario, como se cita a continuación:

<i>Documento</i>	<i>Título</i>
1	Agentes de software en bibliotecas digitales
2	Agentes digitales, agentes móviles
3	Lecturas de agentes de software
4	Taxonomía de agentes

Después de eliminar las palabras irrelevantes, aplicar el algoritmo de Porter y asignar el peso a cada término como el producto de su frecuencia en el documento por su frecuencia inversa, se obtiene la tabla de la figura 2.1. Cada fila es la representación vectorial del documento.

Términos →	LECTUR	AGENT	SOFTWARE	BIBLIOTEC	DIGITAL	MOVL	TAXONOM
Documentos ↓							
Consulta:	0	1	1	0	0	0	0
1	0	0.2	0.5	1	1	0	0
2	0	0.4	0	0	1	1	0
3	1	0.2	0.5	0	0	0	0
4	0	0.2	0	0	0	0	1

Figura 2.1 Ejemplo de la representación vectorial documentos

Si la medida de similitud es el coseno del ángulo, redondeando se obtiene:

$$\text{similitud}(c,d1) = 0.4626$$

$$\text{similitud}(c,d2) = 0.2722$$

$$\text{similitud}(c,d3) = 0.6163$$

$$\text{similitud}(c,d4) = 0.1961$$

Estableciendo un coeficiente de correlación igual a 0.6, el tercer documento resulta ser el más relevante.

El modelo de espacios vectoriales ha dado resultados de precisión y devolución satisfactorios cuando la asignación de pesos emplea el producto de la frecuencia del término por la frecuencia inversa del documento y la normalización de la longitud del vector [Meadow 1992].

Sin embargo, los resultados de los sistemas de recuperación de información no siempre son del todo favorables, y se hace necesario implementar algún proceso para construir una nueva consulta a partir de la anterior. Esto se conoce como la *expansión de la consulta o consulta por ejemplo*, y básicamente la nueva consulta puede ser generada al agregar términos de otros documentos mediante cierta retroalimentación o mediante el empleo de sinónimos de los términos de la consulta y [Faloutsos y Oard 1995].

En bibliotecas digitales, es importante identificar a qué colecciones debe aplicarse la medida de similitud de la consulta, para obtener los resultados esperados. Por otro lado, la efectividad del sistema de recuperación también se ve afectada por el número de colecciones disponibles [Golvochinsky et al.1999].

Además de las técnicas automáticas de recuperación de información, el bibliotecario referencista sigue teniendo un papel fundamental para asegurar la entrega en tiempo y forma de la información que requieren los usuarios, es él quien en muchas ocasiones limita el alcance de la búsqueda e indica en base a su experiencia, las colecciones en donde se va a buscar.

El trabajo de los referencistas puede verse apoyado por la funcionalidad y características de los sistemas para grupo, los cuales son tema de la siguiente sección.

## **2.2 Sistemas que apoyan el trabajo en grupo**

Como entes individuales, los humanos tenemos asignadas determinadas tareas. Como seres sociales, nos relacionamos con individuos cuyos intereses son comunes a los nuestros, haciendo del trabajo en grupo una actividad cotidiana, donde la comunicación y la coordinación de los integrantes permiten que el grupo se mantenga como tal.

Por la propia naturaleza del trabajo en grupo, resulta difícil separar los aspectos sociales de los técnicos cuando se trata de diseñar sistemas que afectan las actitudes y el comportamiento de los individuos [Oravec 1996]. Por otro lado, el comportamiento del grupo puede verse afectado por los elementos del ambiente en donde se desenvuelve, y por el concepto de pertenencia que poseen sus integrantes, mismo que se refleja en su participación [Murray 1991].

A pesar de que en diferentes contextos existen controversias sobre la realización de actividades de forma individual o en grupo, para el caso de los referencistas esto no representa ninguna dificultad, debido a que gran parte de su labor es realizada en equipo [Ingram et al. 2000], de ahí la importancia de contar con tecnologías que les permita fortalecer este estilo de laborar. Es importante señalar que generalmente los referencistas se especializan en algún área en particular, y la cooperación en este sentido es indispensable.

Para referirse al trabajo en grupo cuando se utilizan computadoras, en la literatura se emplean diferentes términos, entre los más comunes están groupware, CSCW (Computer Supported Cooperative Work), soporte tecnológico para la colaboración de un grupo de trabajo, cómputo de un grupo de trabajo, computación colaborativa, tecnología de la coordinación, grupos apoyados por computadora y sistemas que apoyan el proceso de grupo. En este documento se adopta la palabra *groupware*.

Dependiendo del tiempo y del lugar en donde se realiza el trabajo en grupo, Johansen como se cita en [Olson et al. 1993] propone la siguiente clasificación del groupware: a) mismo tiempo - mismo lugar, b) mismo tiempo - diferente lugar, c) mismo lugar - diferente tiempo y d) diferente lugar - diferente tiempo.

Por ejemplo, los sistemas de soporte de decisiones pertenecen a la categoría de mismo tiempo - mismo lugar, el "chat" y las videoconferencias, a mismo tiempo - diferente lugar, mientras que el correo electrónico pertenece a la última categoría.

Otra clasificación del groupware se propone en [Wheeler, et al. 1999], la cual depende del tipo de software utilizado. El groupware se considera que pertenece a alguna de las siguientes generaciones:

- Primera Generación: Software propietario, previamente instalado.
- Segunda Generación: Emplea un navegador Web.
- Tercera Generación: Componentes dinámicos y características de las otras dos generaciones.

Las condiciones tecnológicas y de la organización, son aspectos dignos de consideración cuando se pretende incorporar este tipo de tecnología. Sin embargo, la importancia del groupware es grande cuando se observa que con medios tecnológicos, cada vez es más común que las personas trabajen en equipo atravesando fronteras y organizaciones [Ingram et al. 2000]. Los referencistas son un ejemplo típico de este hecho. Por otro lado, gran parte del trabajo que se realiza en equipo se ha vuelto digital [Mills 1999].

Cuando las interacciones de los integrantes de un grupo se realizan en diferente tiempo, la comunicación se considera *asíncrona*. Entre las ventajas de este tipo de comunicación está permitir que los que intervienen en ella consideren completamente sus ideas antes de comunicarlas, y cuando esta comunicación se propicia para alcanzar un objetivo común, los integrantes del grupo pueden establecer sus propios modelos mentales que les permiten realizar la parte del trabajo que les corresponde a su manera [Skamoto y Kuwana 1993]. A su vez, la comunicación síncrona facilita la organización de la

información generada en el grupo por periodos de tiempo prolongados [Murray 1991].

La mayoría de los sistemas para grupo que emplean la comunicación asíncrona tienen algunas de las características siguientes:

- mensajes de texto,
- no emplean un protocolo de comunicación especial,
  
- cuentan con una lista de los integrantes del grupo y,
  
- no existen consideraciones de tipo personal o de subgrupos.

El correo electrónico es una de las aplicaciones más comunes de groupware que emplea comunicación asíncrona. Está considerado como una de las maneras más cotidianas para formar grupos. Su éxito se atribuye entre otras cosas a su rapidez, a su capacidad de enviar un mismo mensaje a muchos usuarios y a que está basado en texto [Baecker 1993].

Aunque desarrollado en la década de los 70's en el proyecto ARPANET y a pesar de que tomó más de una década en llegar a ser un producto comercial en grandes organizaciones y otra década más en alcanzar las primeras escuelas y hogares [Holtham 94], en la actualidad el uso del correo electrónico se ha extendido a una gran variedad de ámbitos.

En la educación, el correo electrónico fomenta la comunicación entre colegas, estudiantes y profesores. Junto con los pizarrones de anuncios, [Raymond 2000] identifica a la comunicación asíncrona como una abstracción factible para la colaboración.

Las características de los sistemas para grupos con comunicación asíncrona descritas, sugieren que un sistema que apoye el trabajo en equipo de los referencistas debe contemplarlas. Sin embargo, existen otras tareas de los referencistas que pueden ser realizadas de forma automática por agentes de software, los cuales se enuncian en seguida.

## 2.3 Agentes

El concepto de agente tiene diferentes connotaciones. En el contexto de este trabajo, se retoma la propuesta de [Sánchez 1997a], al considerar como agentes a aquellos programas empleados para apoyar las actividades de los usuarios, los cuales pueden llevar a cabo algunas de sus tareas o asistirlo en el manejo de grandes espacios de información u otras aplicaciones, cuya característica principal es poseer cierto grado de *autonomía*.

En la literatura, se pueden encontrar diferentes tipos de agentes. Por ejemplo, se dice que un agente es estático si vive en una sola máquina y específico si se trata de un programa pequeño encargado de alguna tarea en beneficio del usuario. Otra clase de agentes son los agentes móviles, los cuales son procesos de software capaces de vagar en redes. Realizan tareas en beneficio de sus propietarios y regresan a su lugar de partida cuando han cumplido con esas tareas. Se conocen también como agentes de red o agentes de Internet [Jennings 1998].

En [Papazoglou et al. 1992] hace referencia a un agente de información, como un agente capaz de manipular datos obtenidos de una o más colecciones, para atender solicitudes de los usuarios o de otros agentes.

[Sánchez 1997a] propone una taxonomía de agentes, la cual considera que los agentes pueden ser denominados *agentes de programador*, si se trata de abstracciones de ingenieros de software o científicos informáticos para diseñar e implementar sistemas complejos, *agentes de red* cuando se habla de entidades autónomas que migran en los nodos de una red en un ambiente distribuido, *agentes de usuario* para referirse a abstracciones de los usuarios finales cuando estos interactúan con los sistemas de cómputo.

En esta taxonomía, los *agentes de usuario* se consideran *agentes de información* si ayudan a los usuarios a manejar espacios de información, *agentes de tareas* cuando apoyan la realización de tareas de cómputo desarrolladas por uno o más usuarios y *agentes sintéticos* si son introducidos en la interfaz del usuario como personajes.

Desde el punto de vista de la interacción entre el agente y el usuario, [Foner 93] describe algunas características que distinguen a un agente de otro programa, en particular, le asocia una representación, cierta autonomía, una forma de comunicarse con el usuario, un dominio de interés y considera la forma en que puede recuperarse de errores para responder a ciertas expectativas del usuario.

En [Maes 1994], las características que definen a un agente son las siguientes: actúan en beneficio del usuario, útiles en ambientes no estructurados, vastos, dinámicos y abiertos, se utilizan cuando el usuario delega tareas porque éstos conocen sus intereses, hábitos o preferencias, pueden hacer sugerencias y están disponibles a cualquier hora revisando la ocurrencia de eventos aún cuando el usuario no está presente.

En el contexto de las bibliotecas digitales, [Sánchez y Leggett 1997] identifican como agentes a los medios que realizan diferentes tareas en

beneficio de los usuarios, percibidos como procesos autónomos o casi autónomos.

Los agentes pueden ser utilizados para asistir en las tareas de recuperación de información [Rhodes 2000]. Además de un método de búsqueda, suelen emplear técnicas que les permiten buscar términos relacionados o conceptos, emplean métodos para expandir o corregir las consultas en base a la información sobre los intereses y preferencias del usuario.

Cuando el agente tiene asociada una representación, ésta suele ser utilizada para transmitir información sobre su estado, fomentando la retroalimentación mediante un canal visual [Scherer 1980], aunque existen sistemas que emplean enunciados en lenguaje natural.

Existen diferentes puntos de vista en relación a la representación de agentes. Por ejemplo, [Maes y Koda 1996] afirman que "la representación del agente en la interfaz no determina la eficiencia del mismo, pero propicia que ésta sea más agradable y que se establezca una mejor interacción entre el humano y la computadora". De la misma manera, [André et al. 1998] aseveran que un personaje ayuda a proveer una representación más cómoda del paradigma de agentes.

Actualmente, una gran variedad de aplicaciones muestran en su interfaz personajes autónomos, principalmente en contextos relacionados con el entretenimiento, la medicina, la investigación, el comercio y la educación.

El capítulo siguiente establece el contexto de este proyecto, mediante la descripción de los principales trabajos relacionados.