Nous sistemes de recuperació de la informació

per Dr. Peter Willett (*)

L'article pretén donar una visió panoràmica de la investigació que s'ha realitzat d'aquesta nova generació de sistemes de recuperació de la informació, tot describint-ne els seus components més importants i il·lustrant-ho amb exemples basats en aquests nous principis que ja s'estiguin utilitzant.

New information retrieval systems

by Dr. Peter Willett (*)

This article offers an overall view of the research that has been conducted, through descriptions of the main components of this new generation of information retrieval systems. Contains examples of systems currently in ise that are based upon these principles.

1. INTRODUCCIÓ

El terme «recuperació de la informació» és utilitzat cada vegada més com a expressió general per designar un ampli ventall de disciplines en l'àrea del processament de la informació no numèrica. Inclou temes com els sistemes d'organització de bases de dades, el processament automàtic de llenguatge natural, els algorismes de recerca i comparació de dades, els sistemes d'ofimàtica i el nou hardware informàtic. Tanmateix, el terme ha estat utilitzat històricament per referir-se a tècniques d'emmagatzematge i recuperació de la informació de bases de dades textuals, sobretot d'aquelles que contenen ci-

(*) Professor en cap de documentació a la Universitat de Sheffield.

(*) Senior Lecturer in Information Science. University of Sheffield.

(**) Traduït per Maria Saborit i Margarida Sanjaume.

tacions bibliogràfiques, per exemple, articles de premsa, patents, informes, etc. Així doncs, els sistemes de recuperació de la informació es podrien descriure més adequadament com a sistemes de recuperació de documents o, encara amb més precisió, com a sistemes de recuperació de referències, ja que, normalment, són dissenyats per processar, d'una manera eficient i efectiva, les referències més que no pas els documents pròpiament dits. Aquesta és la definició de recuperació de la informació que utilitzarem en aquest article, però cal advertir que la ràpida difusió de bases de dades de text complet (Tenopir, 1984; Blair, 1986) implica que les tècniques aquí descrites aviat podran ser aplicades als documents pròpiament dits.

Els sistemes automatitzats de recuperació de la informació existeixen des de fa gairebé tres dècades. Després del treball pioner de Luhn sobre Keyword In Context (KWIC), sobre indexació i els SOI, sistemes de difusió selectiva de la informació (Luhn, 1957; Salton, 1975b), els anys seixanta es van desenvolupar una sèrie de sistemes de recerca retrospectiva que utilitzaven un processament no interactiu («batch»). Es basaven en la recerca de fitxers estructurats en sèrie, guardats generalment en cintes magnètiques, adquirides de productors de bases de dades, com la National Library of Medicine (NLM) i el Chemical Abstracts Service (CAS). El desenvolupament de les xarxes de telecomunicació de llarga distància, de cost baix, els dispositius d'emmagatzematge d'accés directe i els sistemes operatius en temps real i multiusuari, encapçalaren els primers sistemes de recerca en línia que permeteren a usuaris llunyans d'accedir a bases de dades bibliogràfiques per mitjà del telèfon. Actualment hi ha prop de tres mil bases de dades disponibles per a recerques en línia (Williams, 1985), a part dels centenars o milers d'organitzacions que donen facilitats d'accés per a la recuperació d'informació interna usant un dels paquets disponibles per a recuperació de text lliure (FTX) (Ashford, 1984; Hamilton et al., 1985). Cal suposar que aquests nombres s'incrementaran encara més amb la proliferació de bancs de dades numèriques i fitxers de textos complets, i amb la difusió dels sistemes integrats d'automatització d'oficines.

Amb tot, malgrat els massius desenvolupaments tecnològics que hi ha hagut, les tècniques de recerca disponibles per a l'usuari han canviat molt poc aquests anys. Concretament tots els sistemes es basen en tècniques de recerca booleana desenvolupades primer per a sistemes no interactius SDI. En un àmbit com aquest, qui fa la recerca és responsable de la selecció dels termes de recerca i dels operadors booleans —AND, OR i NOT— que els connecten. A fi de resoldre vaguetats i complexitats en les recerques de text lliure es proporciona assistència mecanitzada al responsable de la recerca, en forma de disponibilitats, per poder fer la recerca per truncament, i per proximitat, juntament amb la possibilitat de veure en línia diccionaris de termes en el fitxer invers. (Meadow and Cochrane, 1981). L'experiència en la utilització d'aquests sistemes ens ha ensenyat que cal una extensa preparació per portar a terme recerques d'una manera eficaç, per tant els sistemes en línia han de ser emprats per intermediaris preparats que ajudin els que necessiten la informació. Amb l'increment ràpid dels sistemes de recuperació de text hi ha una consciència

creixent que cal donar més atenció al proveïment de mitjans de recerca que permetin a l'usuari final de fer la recerca d'una manera prou eficient i efectiva. Una manera de poder aconseguir-ho és utilitzant sistemes intermediaris en què s'empra una complicada connexió per fer augmentar les capacitats de recuperació bàsiques d'un sistema en línia, amb una àmplia varietat de mecanismes d'ajut i suport. Exemples d'aquests tipus de sistemes són descrits per Marcus (1983), Pollitt (1984) i Wales (1984), entre altres, i alguns d'ells són ara assequibles com a paquets comercials de software.

Un mètode força diferent per donar mitjans de recerca a l'usuari final es troba en els sistemes de recuperació de la informació producte de la investigació de les últimes dues dècades. Aquest treball ha implicat el desenvolupament de procediments algorísmics que permeten a l'ordinador d'assumir moltes de les funcions realitzades normalment pels intermediaris mitjançant processos que utilitzen un ampli ventall de tècniques estadístiques. Seguint els treballs pioners de Salton i els seus col·laboradors (1971), el focus primari d'interès ha estat centrat en el disseny de sistemes de recuperació més útils, per exemple, sistemes que poden recuperar quantitats més grans d'informació que no pas els sistemes convencionals (van Rijsbergen, 1979; Sparck Jones, 1982; Salton and McGill, 1983). Tanmateix, les tècniques que s'han desenvolupat poden ser també més eficients en tant que requereixen un grau menor de participació i de competència de l'usuari; aquest augment d'eficiència s'obté per la transferència de responsabilitat, en moltes de les operacions bàsiques de recuperació, de l'usuari a l'ordinador.

El treball d'investigació comença ara a reflectir-se en els sistemes operacionals d'índole diversa (Stibic, 1980; Porter, 1982; Doszkocs, 1982; Doszkocs, 1983; Brzozowski, 1983; Koll *et al.*, 1984; Hendry *et al.*, 1986). Amb els progressos en la tecnologia del hardware i del software i amb l'ús creixent dels sistemes de recuperació de text, sembla segur que aquesta tendència continuarà augmentant en el futur. Aquest article pretén donar una visió panoràmica de la investigació que s'ha realitzat amb la descripció dels principals components d'aquesta nova generació de sistemes de recuperació de la informació, i amb alguns exemples de sistemes de recuperació basats en aquests principis i que ja funcionen. La secció següent descriu l'ús de tècniques d'indexació automàtica per obtenir models de documents i de consultes, incloent-hi mètodes per a la selecció de termes a partir de textos de documents i l'ús de tècniques de fusió («conflation») per poder reunir algunes de les variants de forma dels mots que caracteritzen les bases de dades de text lliure. La secció 3 tracta de l'ús de tècniques de recuperació de la informació per coincidència òptima («best match retrieval») com a alternativa als mecanismes de recuperació booleans que caracteritzen els sistemes actuals FTX. La secció 4 descriu les tècniques de ponderació de termes («term weighting») que es poden utilitzar per fer ressaltar la capacitat dels termes de recerca per descriminar entre documents rellevants i no rellevants, seguit d'una breu descripció d'altres tècniques que, amb un desenvolupament addicional, poden figurar en els futurs sistemes de recuperació de la informació.

El motiu d'escriure aquesta introducció general no ha estat donar una visió crítica exhaustiva de la investigació feta fins a l'actualitat, sinó el de resumir els resultats més importants d'una manera relativament poc tècnica; les descripcions més detallades de les qüestions tractades en aquest article i moltes de les àrees d'investigació ressenyades es poden trobar a Sparck Jones (1974, 1982), Robertson (1977a), Salton (1979), van Rijsbergen (1979), Raghavan i Deogun (1982), Gerrie (1983), Salton i McGill (1983) i a Kraft (1985).

2. INDEXACIÓ AUTOMÀTICA DE DOCUMENTS I DE CONSULTES

La indexació de documents consisteix a fer la descripció del contingut de cada document d'una base de dades utilitzant una sèrie de descriptors, termes d'indexació, mots-clau que actuen com a claus secundàries per a la recuperació del document en resposta a consultes posteriors.

La indexació s'ha fet històricament d'una manera manual per indexadors experts que coneixen el tema de la base de dades i que estan familiaritzats amb les tècniques d'indexació, com ara la quantitat de termes o les relacions de thesaurus, usats en la seva organització. De tota manera, encara que la indexació manual pot donar excel·lents resultats en principi, aquests, a la llarga, no són gaire satisfactoris. Cleverdon (1984) indica que si dues persones o grups de persones elaboren un thesaurus en un camp temàtic determinat, únicament un 60%, més o menys, dels termes d'índex, són comuns en tots dos thesaurus; també observa que si dos indexadors amb experiència, indexen el mateix document utilitzant un determinat thesaurus, només un 30% aproximadament dels termes assignats són comuns als dos conjunts de descriptors. A més, altres estudis esmentats per Cleverdon i Salton (1975b) demostren que els que fan les recerques varien considerablement en la manera amb què les fan, fins i tot per a un mateix tema d'una determinada base de dades, i en la manera, també, de jutjar la importància dels documents recuperats.

En aquestes circumstàncies no és gens sorprenent que molta gent hagi suggerit el desenvolupament de tècniques automàtiques d'indexació, en què la tasca de seleccionar descriptors de contingut es fa d'una manera automàtica en lloc de manualment. S'ha descrit un ampli ventall de tècniques d'indexació automàtica (Sparck Jones, 1974; Salton, 1975a; Salton, 1975b; Harter, 1978; Salton and McGill, 1983), i hi ha un consens general a considerar que el sistema d'indexació automàtica ha d'estar format per un mòdul de selecció de termes, responsable de la selecció dels descriptors sobre la base d'una anàlisi del text del document; un procediment de fusió, utilitzat per assegurar que diferents formes d'una determinada paraula siguin reconegudes com a equivalents, i un mecanisme de ponderació, que assigna les mesures referents a la importància relativa del termes que s'han seleccionat per descriure un document o una consulta. La selecció i la fusió de termes es descriuen en aquesta secció mentre que l'exposició sobre els sistemes de ponderació es troba a la secció 4.

2.1 SELECCIÓ DE TERMES

La indexació manual es basa en les anàlisis sintàctica i semàntica dels textos o de les consultes. Hi ha hagut diversos intents d'aplicar tècniques lingüístiques al problema de la selecció dels termes d'índex. A pesar dels molts estudis realitzats durant els anys 60 i principis dels 70, els resultats obtinguts en l'ús d'aquestes tècniques són generalment poc satisfactoris. Per aquesta raó, la major part de la investigació sobre indexació automàtica s'orienta cap a les tècniques estadístiques, en lloc de les lingüístiques.

Els primers estudis seriosos sobre indexació automàtica es deuen a Luhn (Luhn, 1957; Salton, 1975b), al final dels anys 50, el qual va suggerir que els termes que han de descriure el contingut d'un document es podien obtenir seleccionant mots del text del document; d'aquesta manera, els termes d'indexació derivarien de les mateixes paraules de l'autor del text, tal com apareguessin en el text complet, en el títol o en el resum (aquesta última font de descriptors és àmpliament utilitzada, per la gran disponibilitat de dades imprescindibles que es llegeixen en màquina). Així, l'operació d'indexació requereix l'extracció d'un cert nombre de mots clau del total de mots que componen el text.

La contribució principal de Luhn va ser suggerir que el procediment d'extracció es podria basar en la informació de freqüència. Luhn apuntava que un mot-clau que apareixia molt sovint en una base de dades difícilment discriminava entre documents significatius i no significatius en el procés d'especificació del mot-clau en consulta; d'altra banda, un mot-clau que no aparegués gaire sovint discriminava molt bé, però per la seva naturalesa era molt improbable que fos especificat en una consulta. En conseqüència, els mots-clau més útils per a la recuperació d'informació són aquells que apareixen amb freqüències intermèdies i els termes que representen el contingut d'una base de dades s'obtenen simplement comptant la freqüència d'aparició de cada mot-clau; és a dir, la freqüència d'aparició en una base de dades, i aleshores utilitzant com a termes d'indexació per a cada document els mots-clau que conté i que tenen freqüències d'aparició intermèdies.

El model de *discriminació per termes* (Salton, 1975a) utilitza com a base teòrica els termes de freqüència mitjana. Aquest model suggereix que els termes que indexen bé són els que millor separen els documents quan són representats com a punts en l'espai multidimensional definit pel conjunt de tots els possibles termes d'indexació. Les possibilitats discriminants dels diferents termes es poden mesurar pel canvi de separació a dins el document quan el terme és utilitzat per indexar i quan no ho és. Una investigació sobre la relació entre els valors de la discriminació dels termes i les freqüències d'aparició revela que termes de freqüència baixa-mitjana provoquen les separacions més grans, i en conseqüència s'ha suggerit que aquests termes són els que han de ser emprats per representar el contingut dels documents i de les consultes, d'acord amb l'estratègia pragmàtica d'indexació de Luhn. Treballs més recents, però, indiquen que la relació entre la discriminació de termes i la freqüència d'apa-

rició pot dependre de la manera com és calculada la separació dins el document (Willet, 1986). A més, la identificació de termes de freqüència intermèdia com els més adequats per indexar, discrepa de l'esquema de freqüències ponderades, descrit en detall a la secció 4, que suggereix que els termes d'aparició menys freqüents són els més útils. Cal destacar que el model de discriminació per termes descriu la selecció de termes per indexar, mentre que l'esquema de freqüència ponderada descriu el pes relatiu que s'ha de donar als termes quan es fa la recerca d'un document. Una anàlisi de les coincidències i les complementarietats de les diferents teories de la recuperació d'informació va ser estudiat per Harter (1978).

L'ús de les dades de freqüència d'aparició com a base de la selecció dels mots-clau es pot estendre clarament a altres tipus d'informació estadística. Un exemple il·lustratiu el trobem en l'ús de la *freqüència de document* per cada mot-clau, és a dir, amb quina freqüència apareix el mot-clau en el text d'un document concret. En aquest cas generalment els termes més útils són els que tenen alta freqüència de document, però de baixa a mitjana freqüència en el conjunt de tots els documents, o sigui termes l'aparició dels quals està restringida a un nombre relativament petit de documents. Salton (1975b) ressenya un seguit de mesures de freqüència que han estat utilitzades per avaluar el valor dels mots-clau com a termes d'indexació. Cal remarcar que després de dues dècades d'investigació encara hi ha desacord en la forma més adequada de seleccionar termes (Luhn, 1957; Stevens, 1965; Salton, 1975a; Salton, 1975; Harter, 1978; Salton, 1986). De fet, més que l'aplicació de criteris de selecció complicats, la tendència actual és utilitzar tots els mots-clau d'un document o d'un text de consulta, i llavors diferenciar-los mitjançant un esquema de ponderació apropiat. Així, la idea de seleccionar alguns dels mots-claus ha estat substituïda per la simple extracció de tots ells, amb l'excepció d'alguns termes de freqüència molt alta que són eliminats mitjançant una llista de mots-clau a rebutjar, que conté els mots amb més freqüència d'aparició; cent o dos-cents mots correntment, que no serien útils per a l'objectiu de recuperació de la informació. Els mots rebutjats («stopword») són primordialment mots sense significat específic, com ara conjuncions, preposicions, articles, —I, PER, A, LA, etc.— però també pot contenir mots d'expressions utilitzades molt sovint en les consultes com ara: «QUALSEVOL COSA SOBRE», «TENS CAP COSA SOBRE», i mots d'especificació redundant, com ara: «INFORMACIÓ» o «PROGRAMA», en una base de dades sobre computadores. Així, una consulta com:

«VOLDRIA DOCUMENTS SOBRE SISTEMES INTERMEDIARIS EXPERTS PER A RECERCA BIBLIOGRÀFICA EN LÍNIA»

seria processada per una rutina automàtica d'indexació amb els mots-clau de consulta: «BIBLIOGRÀFIC», «EXPERT», «INTERMEDIARI», «SISTEMES», «RECERCA», «EN LÍNIA».

Una llista semblant de mots-clau seria utilitzada per representar el contingut dels documents corresponents a la base de dades. Les formulacions d'una consulta són generalment bastant concises i en conseqüència qualsevol mot-clau apareixerà probablement una sola vegada. Si es disposa d'un resum de

document o d'un text complet, els mots-clau poden aparèixer moltes vegades i la mostra del document pot contenir no només cada un dels mots-clau seleccionats, sinó també la freqüència amb què apareixen en el document.

Podem veure que el model de consulta anterior es compon només de mots-clau simples, sense cap de les frases substantives que caracteritzen molts dels sistemes d'indexació manual, per exemple: «RECUPERACIÓ DE LA INFORMACIÓ» o «SISTEMA DE GESTIÓ DE BASES DE DADES». La identificació automàtica d'aquest tipus de frases mitjançant tècniques lingüístiques s'ha demostrat que és extremament difícil (Damerau, 1970; Salton, 1975b) i en conseqüència hi ha hagut diverses temptatives d'identificar-les utilitzant tècniques estadístiques. Un d'aquests mètodes deriva del model de discriminació de termes esmentat abans, i implica la connexió dels mots-clau de freqüència elevada que apareixen alhora en un document. D'aquesta manera, les freqüències d'aparició són més baixes i, per tant, milloren les característiques de discriminació (Salton, 1975a; Salton and McGill 1983). Aquest tipus d'anàlisi de frase, intenta que augmenti la precisió del sistema de recuperació de la informació (s'entén com a precisió d'un sistema la fracció dels documents recuperats que són importants). Alternativament, es disposen de tècniques perquè augmenti l'índex de recuperació («recall») d'un sistema (entenem com a «recall» la fracció del nombre total de documents importants que són recuperats). Això s'aconsegueix, normalment, en sistemes manuals, utilitzant un thesaurus, en el qual per cada terme, se n'indiquen els sinònims i d'altres de relacionats. Els thesaurus poden ser creats automàticament, utilitzant informació de com apareixen els termes conjuntament en el text (Stiles, 1961; Stevens, 1965); tot i així, com es demostra en la secció 5, quan parlem sobre l'agrupació de termes, l'ús d'aquesta tècnica no ha resultat gaire eficaç en la pràctica. Un mètode alternatiu perquè augmenti l'índex de recuperació és l'ús de les tècniques de fusió que es descriuen tot seguit.

■ 2.2 FUSIÓ

Una vegada s'ha identificat el conjunt de mots que representen una consulta o un document, s'ha de trobar alguna manera de solucionar el problema de les diferents variants dels mots que, probablement, es trobaran en sistemes de text lliure. Aquestes variants són degudes a múltiples causes; per exemple, requeriments gramaticals com ara:«BIBLIOGRAPHY» «BIBLIOGRAPHIC»,
ortografia alternativa acceptada, per exemple:
«RECOGNISE» «RECOGNIZE»
antònims, com per exemple:
«ABILITY» «DISABILITY»
i problemes deguts a faltes ortogràfiques, de transcripció a un altre alfabet i abreviacions. El problema de les variants dels mots es pot alleugerir, però no eliminar, amb l'ús d'un algorisme de *fusió*, consistent en un procediment de càlcul que redueix les variants d'un mot en una forma única a fi de facilitar

l'objectiu de recuperació de la informació. La fusió compleix dues funcions útils en els sistemes de recuperació de la informació. Primera, pot reduir el nombre total de termes diferents, amb la conseqüent reducció de la mida del diccionari, facilitant-ne alhora l'actualització, segona, encara més important, els mots similars tenen generalment significats similars i així l'efectivitat de la recuperació pot augmentar quan els mots similars poden ser identificats pel procediment de fusió.

En els sistemes en línia d'avui es fa normalment fusió de termes durant la recerca utilitzant un mètode de truncament per la dreta que el que fa la consulta pot especificar en lloc d'utilitzar mètodes automàtics. Es necessita una experiència considerable per fer el truncament de manera eficaç, ja que és molt freqüent que es produeixin dos tipus d'errors. El truncament excessiu, per una banda, provoca que l'arrel que queda després del truncament sigui massa curta, i el resultat és que mots sense cap mena de relació resulten fusionats, com per exemple l'arrel «MED» faria recuperar conjuntament els mots «MEDIA» i «MEDICA». Per altra part, el truncament deficient, que apareix quan es treuen sufixos massa curts, provoca que mots relacionats siguin descrits amb arrels diferents, com «COMPUTERS» truncat a «COMPUTER*» en lloc de «COMPUT» (que inclouria també «COMPUTING» i «COMPUTATIONAL»). Aquests problemes, per descomptat, es troben en qualsevol procediment automàtic o manual que intenti fusionar variants morfològiques.

El procediment més corrent de fusió es basa en l'ús d'un *algorisme troncal* («stemming algorithm») que redueix tots els mots, amb la mateixa arrel a una forma única, mitjançant el procediment de treure de l'arrel els afixos derivacionals i inflexionals. En la majoria dels casos només es treuen els sufixos, de manera que l'algorisme té una funció comparable a la de truncament manual per la dreta. Hi ha molts tipus diferents d'algorismes troncals que han estat publicats en estudis; encara que es diferencien en els detalls, la majoria operen sobre la base d'un diccionari de finals comuns de mots, com ara «-SES», «-ATION», «-ING», etc. Quan un mot és presentat per extreure'n l'arrel es comença buscant aquests sufixos per l'extrem dret del mot. Si es troba el sufix, es treu, excepte en un conjunt d'ocasions prohibides, com ara treure «-ABLE» de «TABLE» o «-S» de «GAS». A més, hi ha un ventall de comprovacions que es poden fer, per exemple: eliminar el doblament de consonants terminals que es donen quan s'utilitza el gerundi, com «FORGETTING» i «FORGET». Exemples d'algorismes troncals típics són descrits per Lovins (1968), Porter (1980) i Ulmschneider i Doszkocs (1983). Un estudi de Lenon *et al.* (1981) que compara un ampli ventall de procediments de fusió explica que hi ha relativament poca diferència en l'efectivitat de recuperació de la informació entre els diversos algorismes comparats, encara que tinguin diferents mecanismes operatius. El procés de buscar el tronc comú d'un conjunt de mots, és fàcil d'apli-car i és de gran efectivitat en el sentit d'agrupar mots amb sufixos. Però hi ha molts altres tipus de variants de mots que apareixen en bases de dades de text lliure, i ha hagut diversos intents per poder proporcionar

mecanismes de fusió per a totes elles. Així, el paquet 3RIP FTX conté un diccionari en línia de mots inversos, de manera que, quan es realitza un truncament, els mots amb finals diferents són fusionats, és a dir, permet la recerca per truncament a l'esquerra. Un exemple més general d'aquesta tècnica, el donen Bartley i Choueka en el context de les bases de dades de text complet (Bartley and Choueka, 1982). Una aproximació alternativa, i més general, consisteix a calcular *el grau de similitud* entre un terme específic de consulta i cada un dels termes en el diccionari del fitxer invers. Els mots similars poden ser mostrats en el terminal perquè siguin inclosos en la consulta, si l'usuari així ho desitja. Un exemple d'aquest sistema el descriu en Freund i Willet (1981), que empren una mesura de similitud basada en el nombre de trigrams (subcadenes de 3 caràcters) comuns entre un parell d'arrels de mot. Porter (1983) descriu un exemple de sistema de recuperació que es basa en aquesta idea.

■3. RECERCA PER COINCIDÈNCIA ÒPTIMA

■3.1. COMPARACIÓ ENTRE EL SISTEMA DE RECERCA BOOLEANA I EL DE COINCIDÈNCIA ÒPTIMA

La immensa majoria de sistemes de recuperació utilitzats es basen en la recerca booleana, encara que hi ha greus problemes relacionats amb l'ús d'aquest model de recuperació. El principal desavantatge és la dificultat que comporta la formulació de la consulta utilitzant els operadors booleans AND, OR i NOT, ja que els usuaris finals no són, en general, capaços de formular bé les consultes i necessiten l'ajuda d'intermediaris preparats. Un segon problema és el total descontrol de la dimensió de la informació generada per una determinada consulta. Sense un coneixement detallat del contingut del fitxer, la persona que fa la recerca és incapaç de preveure quants documents seran identificats per satisfer la consulta. Podrien ser centenars, si la consulta ha estat feta en termes generals o bé cap, si ho ha estat en termes massa concrets. En ambdós casos, la persona encarregada de la recerca haurà de replantejar la consulta, a fi d'aconseguir, en una segona recerca, recuperar un nombre més útil de documents. Una tercera limitació de la recerca boolena és que les operacions de recuperació resulten una simple divisió de la base de dades en dos subconjunts diferents: el dels documents que satisfan la consulta i el dels que no la satisfan. Tots els documents recuperats tenen presumiblement la mateixa utilitat per al recercador i no hi ha cap mecanisme que els pugui col·locar en ordre decreixent

per probabilitat d'importància. Finalment, no hi ha una possibilitat clara a partir de la qual es pugui observar la importància relativa dels diferents components de la consulta, des del moment que la recerca booleana assumeix implícitament que totes les claus tenen ponderacions d'1 o 0, segons si són presents a la consulta o no hi són.

Un concepte fonamental dels nous sistemes descrits en aquest article és el de la recerca per *coincidència òptima*, anomenada també *de màxima proximitat* («nearest neighbour») o també *de resultat ordenat jeràrquicament* («ranked output»). Una recerca per coincidència òptima compara el conjunt d'arrels de la consulta amb el conjunt d'arrels corresponent a cada un dels documents de la base de dades, calcula una mesura de similitud entre la consulta i cada document i llavors classifica els documents en ordre decreixent de similitud amb la consulta. Una mesura típica de similitud és el nombre de termes comuns, l'anomenada recerca de *nivell de coordinació*, que ha estat molt promoguda per Cleverdon (1984). El resultat de la recerca és una llista ordenada, en la qual aquells documents que el sistema considera més similars a la consulta són al principi de la llista i per tant són els primers mostrats a l'usuari. En general, si s'ha emprat una mesura adequada de similitud, els primers documents inspeccionats seran els que tinguin la màxima probabilitat de ser importants per a la consulta que ha estat plantejada (Robertson, 1977b).

Els sistemes de recuperació de la informació en línia basats en el mètode de recerca per coincidència òptima poden alleugerir molts dels problemes associats amb el mètode de recerca booleana. En aquests tipus de sistemes no cal especificar les relacions booleanes entre els mots-clau de la consulta, ja que el mètode de recerca per coincidència òptima requereix només una llista no estructurada de mots-clau. Això fa que aquests tipus de sistemes siguin molt interessants per als usuaris finals. Normalment no hi ha problemes associats amb el volum de resultats produït, ja que l'usuari pot fer la recerca amb un nombre raonable de termes de la llista. Una recerca ràpida i ben orientada pot implicar només la revisió dels primers cinc o deu documents de la llista ordenada, mentre que es pot aconseguir un percentatge més alt de recuperació augmentant el nombre de documents inspeccionats. Finalment, és molt fàcil de tenir en compte informació de ponderació per calcular la funció de coincidència usada per determinar el grau de similitud entre la consulta i els documents del fitxer. A més, aquests factors de ponderació poden ser modificats a criteri de l'usuari, segons l'eficàcia en recerques prèvies, i per això, si cal fer una segona recerca, hi ha un mètode interessant que possibilita la incorporació automàtica d'informació rellevant per retroacció («feedback»).

3.2. APLICACIÓ DE LA RECERCA PER COINCIDÈNCIA ÒPTIMA

Després dels avantatges del mètode de recerca per coincidència òptima, descrits en la secció anterior, el lector es pot preguntar per què generalment

no trobem aquest mètode de recerca en els sistemes actuals de recuperació de la informació. Sens dubte, una raó important és la inèrcia. Els sistemes booleans han estat utilitzats des de fa molts anys i hi ha una resistència natural, d'usuaris i proveïdors de sistemes de recuperació de la informació, a desenvolupar noves tècniques. D'altra banda, la raó principal és probablement el cost de càlcul, o més aviat el cost de càlcul *aparent*, ja que la manera de realitzar el mètode de la recerca per coincidència òptima requereix fer la comparació de cada consulta amb cada un dels documents de la base de dades. Aquesta comparació seqüencial de la col·lecció sencera de documents és massa lenta per fer recerques interactives, i per aquesta raó hi ha un interès considerable en el desenvolupament d'algorismes eficients per fer la recerca basada en la coincidència òptima, que poden establir un ordre sense que calgui revisar tots els documents. L'ús d'una organització de fitxer invers, similar a la utilitzada pels sistemes convencionals de recuperació booleana, augmenta l'eficiència de l'operació.

Un grup d'algorismes de recerca per coincidència òptima és ideat per a la recerca amb fitxers inversos en bases de dades externes; en aquest cas l'algorisme de recerca fa servir els recursos proveïts pel sistema principal de recuperació de la informació en línia, i la recerca per coincidència òptima s'ha d'aplicar fent servir únicament les operacions de recuperació booleana disponibles en el software principal (Morrissey, 1983; Bovey and Robertson, 1984; Robertson *et al.*, 1986). Poden ser desenvolupats algorismes més flexibles si es coneix el contingut de la base de dades, ja que en aquest cas és possible de construir funcions addicionals en el software de recuperació. Aquesta metodologia ha estat extensament estudiada.

Un algorisme de recerca per coincidència òptima, basat en fitxer invers, comporta el processament dels documents en cada una de les llistes de fitxers inversos corresponents als termes de la consulta. Aquestes llistes són anomenades d'ara endavant com a llistes de consulta. Si un document, en un punt qualsevol de la recerca, encara no ha estat comparat amb la consulta, es recupera de la memòria de suport i se'n calcula la similitud amb la consulta; si el valor de similitud que en resulta és més gran que el valor normal de coincidència òptima, el document esdevé el més aproximat. L'eficàcia d'aquest mètode es pot incrementar substancialment mitjançant l'ús de tècniques especials que permeten la identificació dels documents amb coincidència òptima, sense haver de revisar tots els documents de les llistes de consulta (Smeaton and Rijsbergen, 1981; Murtagh, 1982). Mentre aquests algorismes minimitzen el nombre d'operacions per determinar la coincidència entre la consulta i el document, requereixen un gran nombre d'accessos a disc i, per tant, no són adequats per a la recerca interactiva en un context multiusuari (Perry and Willet, 1983). Aquest problema pot ser superat per l'algorisme de recerca per coincidència òptima proposat per Noreault *et al.* (1977), que fa servir un mètode d'addició de les llistes de consulta. L'addició genera una nova llista que conté els identificadors de tots aquells documents que tenen com a mínim un terme en comú amb la llista de la consulta, juntament amb el nombre actual de termes comuns per cada un dels documents. Per altra part, els sistemes booleans usen les ope-

racions lògiques, AND, OR i NOT, sobre les llistes de consulta per crear noves llistes que contenen els identificadors d'aquells documents que satisfan la relació lògica de la consulta. L'addició de llistes de consulta es pot obtenir de la manera següent: quan l'identificador d'un document es troba per primera vegada en una llista de consulta, s'assigna un comptador al document i se li dóna el valor 1. Aquest comptador és incrementat en una unitat cada vegada que es troba el document en les llistes de consulta subsegüents. Quan s'hagin processat d'aquesta manera totes les llistes de consulta, cada comptador contindrà el nombre de termes que són comuns a la consulta, i a un dels documents de la base de dades. Així doncs, el càlcul de totes les similituds implica accessos a disc només per a les llistes de consulta, tal com succeiria en un sistema booleà convencional; les despeses addicionals són la memòria necessària per a l'acumulació dels termes coincidents i la selecció de les similituds finals, que calen per generar la classificació dels documents. El procediment es pot generalitzar a fi de permetre la ponderació dels termes de consulta, tot augmentant els comptadors amb el valor de ponderació assignat a cada terme de consulta en lloc de la unitat. Així, al final del processament de les llistes de consulta, cada comptador contindrà la suma de les ponderacions d'aquells termes que són comuns al document i a la consulta. Una vegada hagin estat identificats els documents que presenten millors coincidències, podran ser mostrats a l'usuari en el terminal. Les anàlisis i les subtileses d'aquest algorisme bàsic es poden trobar a Perry i Willett (1983), Buckley i Lewit (1985) i a Vorhees (1985), mentre que Porter (1982) descriu una versió alternativa en la qual totes les llistes de consulta es processen en paral·lel, i no pas en seqüència, com s'ha indicat més amunt.

3.3. RECERCA COMBINADA

L'exposició sobre la recerca per coincidència òptima feta en els paràgrafs anteriors pot haver fet la impressió que presenta tots els avantatges i que la recerca booleana té poc a fer-hi. De fet, hi ha diverses característiques útils de la recuperació booleana que no són compartides en la recuperació per coincidència òptima. Així, els operadors booleans permeten definir de manera molt explícita subjectes complexes i de moltes facetes. Per altra part l'ús de l'operador AND facilita la identificació parcial de frases substantives (Kraft, 1985). Per aquestes raons, alguns investigadors han proposat l'ús d'un sistema híbrid que permet a l'usuari d'especificar restriccions booleanes que han de ser satisfetes pels resultats d'una recerca per coincidència òptima. Això és formalment equivalent a ordenar els resultats d'una recerca booleana tal com ho trobem en el paquet STAIRS FTX. Aquest tipus de mecanisme híbrid de recuperació proporciona una eina útil de recuperació a tota mena de recerques i es pot aplicar fàcilment quan es disposa dels dos tipus de recerca. Hi ha alguns problemes teòrics associats amb l'ús de la recuperació booleana classificada (Bookstein, 1978; Radecki, 1982), encara que s'han pogut resoldre parcialment en

un treball recent de Salton *et al.* (1983). Aquest autor suggereix l'ús d'una nova tècnica de recerca combinada, amb resultats substancialment superiors als obtinguts utilitzant individualment els dos tipus de recerca.

■4. PONDERACIÓ DELS TERMES DE RECERCA

El mètode de recerca per coincidència òptima implica el càlcul d'una mesura quantitativa de la similitud entre la consulta i cada un dels documents d'un fitxer, i aleshores les similituds calculades constitueixen la base per fer la classificació.

Una mesura de similitud comprèn dos components principals: primer, un *esquema de ponderació dels termes*, que és usat per assignar valors numèrics a cada terme de l'índex en una consulta o en un document per destacar-ne la importància relativa; el segon component és el *coeficient de similitud*, que fa servir aquests valors per calcular el grau de similitud entre una consulta i cada un dels documents en un fitxer.

S'han fet estudis extensos per determinar la incidència d'aquests dos factors en l'efectivitat dels sistemes de recuperació de documents. La majoria d'aquests treballs són de tipus empíric, ja que fins fa molt poc no hi havia cap base teòrica per a la selecció de les mesures de similitud (Sparck Jones, 1973; Sparck Jones and Bates, 1977). Aquesta situació ha canviat radicalment amb el desenvolupament de models probabilístics (van Rijsbergen, 1979) de recuperació de la informació, que han permès la millora de l'eficàcia de la recuperació i a més han donat una base teòrica ferma per poder seleccionar una àmplia gamma de tècniques de recuperació. Concretament, s'han descrit models probabilístics que inclouen ponderació de termes d'índex, retroacció de la rellevància, classificació dels mots-clau, i la recerca de fitxers de documents agrupats.

Avui dia, el resultat de la investigació en aquest camp indica que l'esquema de ponderació utilitzat té més importància que la selecció del coeficient de similitud per a la comparació de documents i consultes. En l'explicació que segueix s'assumeix que el coeficient de similitud és el *producte escalar* («*dot product*»). Anomenarem Q la consulta i D el document, que serà representat en la recerca per un vector d'M components, on M és el nombre de termes diferents usats per indexar els documents de la base de dades, i on i ($1 <= i <= M$) representa la ponderació del terme i en la consulta o en el document. Aleshores el *producte escalar* s'obté del producte d'aquests dos vectors, és a dir:

$$\sum q_i * d_i$$

O la suma es fa per tots els termes M. Si algun terme, i, no és a la consulta

o al document, o sigui $q_i = 0$ o $d_i = 0$, llavors aquest terme no tindrà cap contribució en el coeficient de similitud. Una contribució que no sigui zero, d'altra banda, es fa només si el terme *i* és present tant en la consulta com en el document. En el cas de termes no ponderats, o sigui $q_i = 1$ o $d_i = 1$, el *producte escalar* correspon exactament al nombre de termes comuns a la comanda i al document. Salton (1971) va suggerir l'ús del *coeficient cosinus* en el qual el *producte escalar* és normalitzat amb referència a l'arrel quadrada de les sumes dels quadrats dels components de la comanda i del document, és a dir:

$$\frac{\sum q_i * d_i}{N \sum q_i^2 * d_i^2}$$

Aquest coeficient ha estat utilitzat extensament i sembla que dóna bons resultats pràctics.

Es poden utilitzar esquemes de ponderació de termes tant per als termes de consulta, com per als termes de documents o per a ambdós tipus. La majoria dels estudis d'indexació contenen mètodes per a ponderar els termes de consulta i els documents caracteritzats per termes d'indexació binària, és a dir present o absent.

Sparck Jones (1972) va introduir el concepte de ponderació *freqüència conjunta*, o bé *freqüència inversa de document* (Invers Document Frequency, IDF). Aquest concepte implica l'assignació de ponderacions als termes d'una consulta de manera que són inversament proporcionals a la freqüència d'aparició del terme en el conjunt de documents en què s'ha de fer la recerca. La justificació d'aquest mètode és que les persones tendeixen a expressar les seves necessitats d'informació amb termes de definició àmplia, utilitzats sovint, i els termes més específics, o sigui, de baixa freqüència d'aparició seran els que probablement tindran més importància en la identificació del material rellevant. Això és perquè el nombre de documents importants per satisfer una consulta és probable que sigui relativament petit i, en conseqüència, termes que apareixen molt sovint han de pertànyer a molts documents irrellevants. Així doncs, termes d'aparició infreqüent tenen una probabilitat més alta d'apareixer en documents rellevants, i per tant s'ha de considerar que tenen una importància potencial més gran a efectes de la recuperació d'informació. Aquestes consideracions porten a l'ús de la ponderació:

$$\text{Log}_2 N/n_i$$

per a algun terme *i* que té n_i aparicions en N documents del conjunt. L'ús de la funció logarítmica ha estat justificat amb arguments de teoria de la informació (Robertson, 1974) i, més recentment, com un cas límit de les ponderacions probabilístiques de la rellevància (Croft and Harper, 1979); aquest darrer treball és descrit més endavant. Proves fetes amb l'esquema de freqüència inversa de document (IDF) demostren que constantment dóna resultats que són

superiors, en la recerca per coincidència òptima, als que s'obtenen amb l'ús de termes de consulta no ponderats, quan de fet tots els termes de consulta s'han de considerar amb la mateixa importància per als objectius de la recuperació. Una vegada la recerca inicial ja ha estat feta i l'usuari ha revisat uns quants documents, hi ha la possibilitat de fer una recerca per retroacció de rellevància («relevance feedback»). El que es fa amb aquest mètode és utilitzar opinions donades prèviament per l'usuari sobre la rellevància dels documents revisats per calcular un nou conjunt de valors que reflecteixin de manera més acurada la importància de cada terme de la consulta. Robertson i Sparck Jones (1976) van suggerir l'ús de la informació de rellevància com a base de ponderació dels termes de consulta. Utilitzant la teoria de la probabilitat i fent ús de la hipòtesi que les aparicions dels termes d'índex en els documents serien estadísticament independents, van formular teòricament l'ús de la ponderació de termes:

$$\log \frac{p_i\,(1 - q_i)}{q_i\,(1 - p_i)}$$

on p i q són les probabilitats que el terme i aparegui en un document rellevant i en un de no rellevant, respectivament. A la pràctica, és convenient de fer la substitució de les probabilitats per freqüències d'aparició. Si el terme i apareix en n dels documents N en un conjunt i apareix en r vegades en els documents rellevants R d'una consulta, la ponderació anterior pot ser expressada en la forma:

$$\log \frac{r\,(N - n_i - R + r_i)}{(R - r_i)\,(n_i - r_i)}$$

La teoria que porta a aquest esquema de ponderació també genera un coeficient de similitud particular. Aquest coeficient de similitud és el *producte escalar*, definit prèviament, que correspon a la suma de les ponderacions dels termes de la consulta que apareixen dins del document. Així si d indica la presència ($d_i = 1$) l'absència ($d_i = 0$) del terme de consulta i en un document, la funció de coincidència usada és:

$$\sum d_i \; log \; \frac{p_i\,(1 - q_i)}{q_i\,(1 - p_i)}$$

on la suma es fa per tots els termes de la consulta. S'ha demostrat que aquestes ponderacions comporten una eficàcia excel·lent de recuperació en els anomenats experiments retrospectius, en què es disposa d'informació de rellevància completa, és a dir, quan r_i i R són coneguts per cada terme de la consulta. A fi d'usar les ponderacions de manera predictiva quan no es disposa de dades completes sobre rellevància, cal valorar p_i i q_i. Això ho pot fer el responsa-

ble de la recerca donant criteris de rellevància sobre el resultat de la recerca inicial, feta amb els primers 10 o 20 documents. Les dades de rellevància poden ser utilitzades pel sistema per valorar r_i i R, i per tant establir les ponderacions, ja que n_i es coneix per cada terme del fitxer invers. Aquest procediment és descrit en detall per Robertson i Sparck Jones (1975) i per Sparck Jones (1979a, 1979b).

Aquestes tècniques de retroacció de la rellevància no es poden utilitzar fins que l'usuari ha tingut l'oportunitat de donar alguna informació de rellevància al sistema. Aquest aspecte ha estat estudiat per Croft i Harper (1979), que suggereixen que en absència de dades de rellevància els valors de totes les p_i havien de suposar-se iguals, és a dir que a tots els termes de la consulta se'ls suposa la mateixa probabilitat d'aparició en documents rellevants, i que l'aparició d'un terme en un document no rellevant pot ser aproximada per la seva aparició en tota la col·lecció sencera. Aquestes dues suposicions corresponen a l'establiment de p_i / (1 - p_i) igual a una constant, que anomenarem C, i a l'establiment de q_i igual a n_i / N, respectivament. Si substituïm aquests valors en l'expressió de ponderació donada més amunt, obtindrem:

$$\sum d_i \log \frac{C\,(N - n_i)}{n_i}$$

que també es pot expressar així:

$$\sum d_i \log C + \sum d_i \log \frac{N - n_i}{n_i}$$

on la suma es fa igual que abans sobre tots els termes de la consulta. La primera part d'aquesta expressió correspon simplement a un nivell de coincidència coordinada. Amb altres paraules, el nombre de termes comuns entre el document i la consulta multiplicat per log C correspon a la primera part de l'expressió, mentre que la segona part és molt similar a la ponderació IDF que s'ha descrit abans. Una anàlisi similar que demostra també la forta relació entre IDF i la ponderació de rellevància ha estat recentment presentat per Robertson (1986).

Per emprar la ponderació anterior a la recerca inicial, cal trobar la manera de valorar C. Croft i Harper suggereixen que p_i hauria de tenir un valor de O.9, que correspon a un valor de 9 per a C. Així, quan una llista específica de consulta s'ha de processar durant una recerca de coincidència òptima, n_i s'obté del nombre d'identificadors de documents de la llista pre-establerta i aquest valor és el que es fa servir per calcular el terme log (N - n_i) / n_i). La ponderació s'obté aleshores incrementant aquest logaritme amb la quantitat log C. La suma de les ponderacions per a aquells termes de la consulta que apareixen en cada document constitueix la base de l'ordenació presentada a l'usuari.

Els documents es mostren a l'usuari l'un darrere l'altre en ordre de similitud decreixent amb la consulta, i se li pregunta a l'usuari si cada document

presentat és rellevant o no ho és per a la consulta. Els criteris de rellevància són usats, llavors, per valorar les probabilitats p_i i q_i i les corresponents ponderacions usades en el mètode de recerca de rellevància per retroacció. D'aquesta manera els interessos de l'usuari són reflectits més fidelment que en la recerca inicial basada en el mètode IDF. La recerca per retroacció pot ser repetida a criteri de l'usuari, i evidentment no ho farà si obté prou material important en la recerca inicial.

Els treballs de Sparck Jones (1972), Croft i Harper (1979) i Robertson (1986) han donat una base eficaç per a la ponderació dels termes d'una consulta; tanmateix, encara hi ha un debat considerable referent al fet de si els termes s'han de ponderar. Salton i els seus col·laboradors (Salton 1971; Salton and McGill, 1983) han proposat des de fa temps l'ús d'un esquema de ponderació de documents en què la ponderació IDF d'un terme coincident és multiplicada per la freqüència del terme en el document (tal com s'ha definit a la secció 2.1). Un desenvolupament més elaborat d'aquesta mateixa idea ha estat publicat per Croft (1983), el qual dóna una justificació teòrica per a l'ús d'un tipus particular d'esquema de ponderació per freqüència de documents. Aquests esquemes, però, no són generalment eficaços en la recuperació d'informació, tal com demostren els estudis experimentals que s'han portat a terme (Sparck Jones, 1973; Sparck i Bates, 1977; Croft, 1983; Salton 1986).

■ 5. ALTRES ÀREES D'INVESTIGACIÓ

Els treballs sobre indexació automàtica, recuperació per coincidència òptima i ponderació de termes, descrits en les seccions anteriors, constitueixen la base d'una nova generació de sistemes FTX. Hi ha diverses àrees que han estat i continuen essent investigades activament i que configuraran els sistemes del futur. Aquesta secció fa una breu repassada d'alguns d'aquests treballs. Altres àrees d'interès per a la investigació actual es poden trobar en les actes anuals de la *International Conference on Research and Development in Information Retrieval* i en revistes com ara *Journal of Documentation* i *Information Processing and Management* inter alia.

Les tècniques descrites més amunt són de caràcter pròpiament estadístiques, encara que recentment hi ha un interès creixent en l'ús de tècniques basades en l'enginyeria del coneixement per dissenyar sistemes de recuperació d'informació. Hi ha dues àrees que són particularment investigades: l'ús de tècniques de processament automàtic del llenguatge natural (ANLP) per a la indexació automàtica i l'ús de sistemes experts per a la recuperació d'informació en línia.

La major part de la investigació inicial en el camp de la indexació automàtica es centrava en l'ús de tècniques lingüístiques en lloc d'estadístiques, per identificar termes d'indexació i frases en textos de llenguatge natural. Els resultats d'aquests treballs, però van ser tots decebedors ja que es va descobrir que tècniques senzilles d'extracció de mots-clau com les descrites en la secció

2 funcionaven millor regularment que els mètodes complicats d'anàlisi de frases (Salton 1975b). En part, aquests resultats es van obtenir majoritàriament de l'anàlisi de textos de documents més que no pas dels textos de consultes (Sparck Jones i Tait, 1984). El fet més important és que les tècniques lingüístiques utilitzades tenien un àmbit molt limitat, i només després de molts anys d'investigació s'ha fet evident el grau de problemàtica que presenta el processament automàtic del llenguatge natural (ANLP). L'estudi en aquesta àrea ha comportat alguns sistemes de tractament de bases de dades amb llenguatge natural satisfactoris (Salton and McGill, 1983; Cercone and McCalla, 1986). I actualment s'ha renovat l'interès en l'ús de tècniques de processament automàtic de llenguatge natural, de cara a la indexació. En trobem un exemple a l'estudi recent de Sparck Jones i Tait (1984), que parla de la utilització d'un processador de llenguatge natural per generar frases substantives gramaticalment acceptables a partir de consultes redactades en llenguatge natural. Aquestes frases es poden buscar en els resums dels documents anb la idea que la presència d'una frase de consulta serà un indicador més substancial de la importància del document que la presència només dels mots-clau que hi hagi. Malgrat tot, els resultats experimentals de què disposem fins ara no són suficients per determinar si el complicat processament lingüístic que s'utilitza incrementa efectivament l'eficàcia del sistema, i també queda per demostrar la millor manera d'utilitzar aquests processaments en els sistemes FTX.

Els sistemes d'intermediaris experts pretenen donar a l'usuari mecanismes senzills d'accés als sistemes actuals de recuperació bibliogràfica en línia. Les possibilitats d'aquests tipus de sistemes intermediaris inclouen, no només detalls de tipus mecànic, com ara la formulació de consultes *offline* i processos automàtics d'inici d'ús del sistema informàtic, sinó que tenen funcions de guiatge i donen suport a les tècniques de retroacció de la rellevància que poden ajudar l'usuari durant el procés de recerca. Hi ha molts d'aquests sistemes disponibles actualment (Marcus, 1983; Pollitt, 1984; Wales, 1984), i hi ha una anàlisi d'aquesta línia d'investigació de Keyhoe (1985), però són sistemes concebuts principalment per reduir alguns dels problemes que comporten els sistemes en línia convencionals. Un estudi inicial de l'aplicació de tècniques de sistemes experts en la recuperació de documents va ser fet per Croft (1986), el qual suggereix que un sistema expert podria rebre coneixement dels requeriments típics de l'usuari que fa la recerca i coneixement del domini sobre el contingut de la base de dades —o de les bases de dades— disponible per fer la recerca. El sistema expert podria aleshores ajudar a seleccionar els termes més adequats i l'estratègia de recerca més apropiada.

Recentment, van apareixent en el mercat més exemples de l'ús de tècniques de recuperació de la informació basades en l'enginyeria del coneixement. Així, Zarri (1984) i Tong i Shapiro (1984) descriuen mètodes basats en regles de producció que permeten deduir fets de la informació guardada en bases de dades textuals. Amb l'interès creixent de sistemes de tot tipus basats en l'enginyeria del coneixement, sembla probable que els resultats d'aquest tipus d'investigació aportaran en el futur força metodologies noves per a la recuperació d'informació.

La classificació automàtica o l'anàlisi per agrupació («cluster») és una tècnica estadística molt variada que permet la identificació automàtica de grups o agrupacions d'objectes similars. La similitud es detecta calculant totes les similituds entre parelles d'objectes. Hi ha dos tipus d'agrupació de documents que són interessants des del punt de vista de la recuperació de la informació: l'agrupació de documents basada en els termes que tenen comuns i l'agrupació de termes basada en els documents en els quals apareixen simultàniament (van Rijsbergen, 1979).

La raó per agrupar termes és que augmenta l'eficàcia de l'índex de recuperació en els sistemes FTX. Donada una classificació per termes, cada terme d'un document i/o d'una consulta pot ser substituït per l'identificador del grup que el conté. D'altra banda, la classificació pot ser usada per ampliar una consulta amb l'addició de termes de cada una de les agrupacions que contenen un dels termes de la consulta original. En tots dos casos l'ús de la classificació permet introduir un mecanisme per identificar coincidències addicionals entre les sèries de termes de document i de consulta, i així proporciona un mecanisme que incrementa la capacitat del sistema de recuperació de documents. L'entusiasme inicial per aquestes tècniques va ser seguit per la constatació que els procediments d'agrupació empírics normalment utilitzats no comportaven un augment de l'eficàcia de recuperació i indubtablement reduïen l'eficàcia de recerca (Sparck Jones, 1971). Treballs posteriors fets per van Rijsbergen (1977) van estendre els mètodes de ponderació de termes basats en la rellevància, descrits en la secció 4, que permeten d'incloure les dependències estadístiques conegudes entre termes. Aquest treball va donar una base teòrica sòlida a l'ús de la informació d'aparició conjunta de termes, però els experiments van demostrar que aquesta informació no podia ser emprada per augmentar el funcionament del sistema en àmbits útils de recuperació de la informació (Smeaton i van Rijsbergen, 1983). Conseqüentment hi ha algun dubte sobre la utilitat de les agrupacions de termes basats en l'aparició conjunta en la informació a l'hora de la recuperació.

La situació, pel que fa referència a l'agrupació de documents, és més esperançadora. La justificació de l'agrupació de documents és que un sistema de recuperació de la informació que fa la recerca per agrupacions de documents, en lloc de documents individuals, pot aconseguir nivells millors d'eficàcia, ja que l'organització de l'arxiu i la recerca tenen en compte les relacions que s'estableixen entre els documents de la base de dades. Jardine i van Rijsbergen (1971) van ser els primers a demostrar que la recerca basada en agrupacions podia donar millors resultats que la recerca convencional per coincidència òptima, i almenys hi ha prou evidència per suggerir que això pot passar a la pràctica (Croft, 1984), sobretot en recerques programades amb precisió que recuperen pocs documents rellevants. Hi ha dos problemes associats a l'ús dels mètodes d'agrupació de documents. El primer és l'elevat cost del còmput relacionat amb el càlcul de les similituds internes dels documents, les quals poden ser calculades usant una modificació de l'algorisme de coincidència òptima amb fitxers inversos descrits en la secció 3 (Willet, 1981), però la càrrega del còmput encara és alta. El segon problema és que hi ha molts mètodes diferents

d'agrupació disponibles, i no és gens clar quin mètode s'ha d'emprar exactament. Hi ha un considerable interès en aquests problemes (Griffiths *et al.*, 1984; Vorhees, 1985). Sembla possible almenys que la recerca basada en l'agrupació pot ser una alternativa útil o un complement al mètode convencional de recerca per coincidència òptima ja que dóna la informació rellevant inicial prèvia a la recerca per retroacció.

Els sistemes de software complicats requeriran increments continus de la potència del processament de l'ordinador. Concretament el hardware per fer procés paral·lel sembla que és un camí adequat perquè augmenti l'eficiència dels sistemes FTX, ja que permet que algunes o moltes de les operacions de coincidència de les consultes —documents es puguin fer simultàniament. Un exemple d'un sistema de recuperació de la informació operacional basat en el disseny de sistema paral·lel és el CAS ONLINE, en el qual es pot fer la recerca d'un fitxer que té uns 8 milions de substàncies químiques, partit en subconjunts d'aproximadament tres quarts de milió de substàncies; cada un d'aquests subconjunts és buscat per un parell de minicomputadors DEC, i aquesta recerca paral·lela global és controlada per un ordinador principal IBM (Dittmar *et al.*, 1983). L'ús de hardware paral·lel per donar suport a operacions de recuperació d'informació en sistemes de gestió de bases de dades, especialment els basats en models relacionals de dades, ha estat àmpliament descrit per Hsiao (1983). Hi ha diverses màquines paral·leles comercialitzades per fer recerca en bases de dades. Hi ha també força interès en l'ús del hardware paral·lel per a les aplicacions de recuperació d'informació (Hollaar, 1979; Teskey, 1985), però els resultats obtinguts sovint són restringits al tipus particular de màquina emprada, i no hi ha encara consens sobre el tipus d'ordinador paral·lel més adequat per al processament de les bases de dades bibliogràfiques. Amb la ràpida reducció dels costos de hardware i amb la introducció de nous tipus de sistemes paral·lels, com el sistema multiprocessador basat en microprocessador de la casa INMOS, anomenat Transputer (Walker, 1985), sembla probable que aquesta serà un àrea activa d'investigació en el futur.

■ REFERÈNCIES

Ashford, J. (1984) Information storage and retrieval systems on mainframes and minicomputers: a comparison of text retrieval packages available in the UK. *Program,* vol. 18, pp. 124-146.

Bernstein, L.M. and Williamson, R.E. (1984) Testing of a natural language retrieval system for a full text knowledge base. *Journal of the American Society for Information Science,* vol. 35, pp. 235-247.

Blair, D.C. (1986) Full text retrieval: evaluation and implications. *International Classification,* vol. 13, pp. 18-23.

Bookstein, A. (1978) Oh the perils of merging Boolean and weighted retrieval systems. *Journal of the American Society for Information Science,* vol. 29, pp. 156-158.

Bourne, C.P. (1977) Frequency and impact of spelling errors in bibliographic data bases. *Information Processing and Management,* vol. 13, pp. 1-12.

Bovey, J.D. and Robertson, S.E. (1984) An algorithm for weighted searching on a Boolean system. *Information Technology: Research and Development*, vol. 3, pp. 84-87.

Bratley, P. and Choueka, Y. (1982) Processing truncated terms in document retrieval systems. *Information Processing and Management*, vol. 18, pp. 257-266.

Brzozowski, J.P. (1983) MASQUERADE: searching the full text of abstracts using automatic indexing. *Journal of Information Science*, vol. 6, pp. 67-73.

Buckley, C. and Lewit, A.F. (1985) Optimization of inverted vector searches. *Proceedings of the Eighth International Conference on Research and Development in Information Retrieval.* Washington: ACM, pp. 97-110.

Cercone, N. and McCalla, G. (1986) Accessing knowledge through natural language. *Advances in Computers*, vol. 25, pp. 1-99.

Cleverdon, C. (1984) Optimizing convenient online access to bibliographic databases. *Information Services and Use*, vol. 4, pp. 37-47.

Croft, W.B. (1980) A model of cluster searching based on classification. *Information Systems*, vol. 5, pp. 189-195.

Croft, W.B. (1983) Experiments with representation in a document retrieval system. *Information Technology: Research and Development*, vol. 2, pp. 1-21.

Croft, W.B. (1986) User-specified domain knowledge for document retrieval. *Proceedings of the Ninth International Conference on Research and Development in Information Retrieval.* Washington: ACM, pp. 201-206.

Croft, W.B. and Harper, D.J. (1979) Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, vol. 35, pp. 285-295.

Damerau, F.J. (1970) Automatic parsing for content analysis. *Communications of the ACM*, vol. 13, pp. 356-360.

Dittmar, P.G., Farmer, N.A., Fisanick, W., Haines, R.C. and Mockus, J. (1983) The CAS ONLINE search system. Part 1. General desing and selection, generation and use of search screens. *Journal of Chemical Information and Computer Sciences*, vol. 23, pp. 93-102.

Doszkocs, T.E. (1982) From research to application: the CITE natural language information retrieval system. *Lecture Notes in Computer Science*, vol. 146, pp. 251-262.

Doszkocs, T.E. (1983) CITE NLM: natural language searching in an online catalog. *Information Technology and Libraries*, vol. 2, pp. 364-380.

Dyson, B. (1984) Data input standards and computerization at the University of Hull. *Journal of Librarianship*, vol. 16, pp. 246-261.

Frakes, W.B. (1984) Term conflation for information retrieval. *In*: C.J. van Rijsbergen (*ed*). *Research and Development in Information Retrieval.* Cambridge: CUP, pp. 383-390.

Freund, G.E. and Willet, P. (1982) Online identification of word variants and arbitrary truncation searching using a string similarity mesure. *Information Tech-*

nology: Research and Development, vol. 1, pp. 177-187.

Gerrie, B. (1983) *Online Information Systems. Use and Operating Characteristics, Limitations and Design Alternatives.* Arlington, Ma.: Information Resources Press.

Griffiths, A., Robinson, L.A. and Willett, P. (1984) Hierarchic agglomerative clustering methods for automatic document classification. *Journal of Documentation*, vol. 40, pp. 175-205.

Hamilton, C.D., Kimberley, R. and Smith, C.H. (1985) *Text Retrieval: a Directory of Sotfware.* Aldershot: Gower.

Harter, S.P. (1978) Statistical approaches to automatic indexing. *Drexel Library Quarterly*, vol, 14, pp. 57-74.

Hendry, I.G., Willett, P. and Wood, F.E. (1986) INSTRUCT: a teaching package for experimental methods in information retrieval. Part 1. The users' view. *Program*, vol. 20, pp. 245-263; Part 2. Computational aspects. *Program*, vol. 20, pp. 382-393.

Hollaar, L.A. (1979) Unconventional computer architectures for information retrieval. *Annual Review of Information Science and Technology*, vol. 14, pp. 129-151.

Hsiao, D.K. (1983) *Advanced Database Machine Architecture.* Englewood Cliffs, N.J.: Prentice-Hall.

Jardine, N. and van Rijsbergen, C.J. (1971) The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval*, vol. 7, pp. 217-240.

Keyhoe, C.A. (1985) Interfaces and expert systems for online retrieval. *Online Review*, vol. 9, pp. 489-505.

Koll, M.B., Noreault, T. and McGill, M.J. (1984) Enhanced retrieval techniques on a microcomputer. *Proceedings of the National Online Meeting, April 10-12, 1984.* New York: Learned Information, pp. 165-170.

Kraft, D.H. (1985) Advances in information retrieval: where is that / # *& $ record? *Advances in Computers*, vol. 24, pp. 277-318.

Lennon, M., Peirce, D.S., Tarry, B.D. and Willett, P. (1981) An evaluation of some conflation algorithms for information retrieval. *Journal of Information Science*, vol. 3, pp. 177-183.

Lovins, J.B. (1968) Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, vol. 11, pp. 22-31.

Luhn, H.P. (1957) A statistical approach to mechanised encoding and searching of library information. *IBM Journal of Research and Development*, vol. 1, pp. 309-317.

Marcus, R.S. (1983) An experimental comparison of the effectiveness of computers and humans as search intermediaries. *Journal of the American Society for Information Science*, vol. 34, pp. 381-404.

Meadow, C. and Cochrane, P. (1981) *Basics of Online Searching.* Chichester: Wiley.

Morrissey, J. (1983) An intelligent terminal for implementing relevance feedback on large operational retrieval systems. *Lecture Notes in Computer Science*, vol. 146, pp. 38-50.

Murtagh, F. (1982) A veri fast, exact nearest neighbour algorithm for use in information retrieval. *Information Technology: Research and Development*, vol. 1, pp. 275-283.

Noreault, T., Koll, M. and McGill, M.J. (1977) Automatic ranked output from Boolean searches in SIRE. *Journal of the American Society for Information Science*, vol. 28, pp. 333-339.

Perry, S.A. and Willett, P. (1983) A review of the use of inverted files for best match searching in information retrieval systems. *Journal of Information Science*, vol. 6, pp. 59-66.

Pollitt, A.S. (1984) A 'front-end' system: an expert system as an online search intermediary. *Aslib Proceedings*, vol. 36, pp. 229-234.

Porter, M.F. (1980), An algorithm for suffix stripting. *Program*, vol. 14, pp. 130-137.

Porter, M.F. (1982) Implementing a probabilistic retrieval system. *Information Technology: Research and Development*, vol. 1, pp. 131-156.

Porter, M.F. (1983) Information retrieval at the Sedgwick Museum. *Information Technology: Research and Development*, vol. 2, pp. 169-186.

Radecki, T. (1982) Reducing the perils of merging Boolean and weigted retrieval systems. *Journal of Documentation*, vol. 38, pp. 207-211.

Raghavan, V.V. and Deogun, J.S. (1982) Information retrieval research; strategies and user inplications. *Information Technology: Research and Development*, vol. 1, pp. 157-171.

Robertson, S.E. (1974) Specificity and weighted retrieval. *Journal of Documentation*, vol. 30, pp. 41-46.

Robertson, S.E. (1977a) Theories and models in information retrieval. *Journal of Documentation*. vol. 33, pp. 126-148.

Robertson, S.E. (1977b) The probability ranking principle in information retrieval. *Journal of Documentation*, vol. 33, pp. 294-304.

Robertson, S.E. (1986) On relevance weight estimation and query expansion. *Journal of Documentation*, vol. 42, pp. 182-188.

Robertson, S.E. and Sparck Jones, K. (1976) Relevance weighting of search terms. *Journal of the American Society for Information Science*, vol. 27, pp. 129-146.

Robertson, S.E., Thompson, C.L., Macaskill, M.J. andBovey, J.D. (1986) Weighting, ranking and relevance feedback in a front-end system. *Journal of Information Science*, vol. 12, pp. 71-75.

Salton, G. (1971) *The SMART Retrieval System*. Englewood Cliffs: Prentice-Hall.

Salton, G. (1975a) *A Theory of Indexing*. Philadelphia: Society for Industrial and Applied Mathematics.

Salton, G. (1975b) *Dynamic Information and Library Processing*. Englewood Cliffs: Prentice-Hall.

Salton, G. (1979) Mathematics and information retrieval. *Journal of Documentation*, vol. 35, pp. 1-29.

Salton, G. (1986) Recent trends in automatic information retrieval. *Proceedings of the Ninth International Conference on Research and Development in*

Information Retrieval. Washington: ACM, pp. 1-10.

Salton, G., Fox, E.A. and Wu, H. (1983) Extended Boolean information retrieval. *Communications of the ACM*, vol. 26, pp. 1.022-1.036.

Salton, G. and McGill, M.J. (1983) *Introduction to Modern Information Retrieval.* New York: McGraw-Hill.

Smeaton, A.F. and van Rijsbergen, C.J. (1981) The nearest neighbour problem in information retrieval. An algorithm using upperbounds. *Proceedings of the Fourth International Conference on Research and Development in Information Retrieval.* Washington: ACM, pp. 83-87.

Smeaton, A.F. and van Rijsbergen, C.J. (1983) The retrieval effects of query expansion on a feedback document retrieval system. *Computer Journal*, vol. 26, pp. 238-246.

Sparck Jones, K. (1971) *Automatic keyword Classification for Information Retrieval.* London: Butterworth.

Sparck Jones, K. (1972) A tatistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, vol. 28, pp. 11-21.

Sparck Jones, K. (1973) Index term weighting. *Information Storage and Retrieval*, vol. 9, pp. 619-633.

Sparck Jones, K. (1974) Automatic indexing. *Journal of Documentation*, vol. 30, pp. 393-432.

Sparck Jones, K. (1979a) Search term relevance weighting given little relevance information. *Journal of Documentation*, vol. 35, pp. 30-48.

Sparck Jones, K. (1979b) Experiments in relevance weighting of search terms. *Information Processing and Management*, vol. 15, pp. 133-144.

Sparck Jones, K. (1982) *Information Retrieval Experiment.* London: Butterworth.

Sparck Jones, K. and Bates, R.G. (1977) *Research on Automatic Indexing 1974-6.* Cambridge: Computer Laboratory, University of Cambridge.

Sparck Jones, K. and Tait, J.I. (1984) Automatic search term variant generation. *Journal of Documentation*, vol. 40, pp. 50-66.

Stevens, M.E. (1965) *Automatic Indexing - a State of the Art Report.* Washington: National Bureau of Standards Monograph nº 91.

Stibic, V. (1980) Influence of unlimited ranking on practical online search strategy. *Online Review*, vol. 4, pp. 273-278.

Stiles, H.E. (1961) The association factor in information retrieval. *Journal of the ACM*, vol. 8, pp. 271-279.

Tenopir, C. (1984) Full-text databases. *Annual Review of Information Science and Technology*, vol. 19, pp. 215-246.

Teskey, F.N. (1985) *Novel Computer Architectures for Data Storage and Retrieval.* London: British Library Research and Development Department Report nº 5845.

Tong, R.M. and Shapiro, D.G. (1985) Experimental investigations of uncertainty in a rule-based system for information retrieval. *International Journal of Man-Machine Studies*, vol, 22, pp. 265-282.

Ulmschneider, J.E. and Doszkocs, T. (1983). A practical stemming algorithm for online search assistance. *Online Review*, vol. 7, pp. 301-318.

Van Rijsbergen, C.J. (1977) A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation*, vol. 33, pp. 106-119.

Van Rijsbergen, C.J. (1979) *Information Retrieval*. London: Butterworth.

Vorhees, E. (1985) *The Effectiveness and Efficiency of Agglomerative Hierarchic Clustering in Document retrieval*. PhD thesis, Cornell University.

Wales, J.L. (1984) Using a microcomputer to access bibliographic data bases: experience with Userlink software in the ICI Organics Division Information and Library Services Unit. *Program*, vol. 18, pp. 247-257.

Walker, P. (1985) The Transputer. *Byte*, vol. 10(5), pp. 219-235.

Willett, P. (1981) A fast procedure for the calculation of similarity coefficients in automatic classification. *Information Processing and Management*, vol. 17, pp. 53-60.

Willett, P. (1985) An algorithm for the calculation of exact term discrimination values. *Information Processing and Management*, vol. 21, pp. 225-232.

Williams, M.E. (1985) Electronic databases. *Science*, vol. 228, pp. 445-456.

Zarri, G.P. (1984) Expert systems and information retrieval: an experiment in the domain of bibliographical data management. *International Journal of Man-Machine Studies*, vol. 20, pp. 87-106.