Introducció a la recuperació d'informació

PID_00271437

Blas Torregrosa García

Temps mínim de dedicació recomanat: 2 hores

Universitat Oberta de Catalunya

Blas Torregrosa García

Enginyer en Informàtica i màster universitari en Seguretat de les Tecnologies de la Informació i de les Comunicacions (MISTIC) per la Universitat Oberta de Catalunya (UOC). Especialitzat en ciberseguretat. Professor col·laborador del màster de Ciència de Dades de la UOC i professor associat a la Universitat de Valladolid (UVA).

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats pel professor: Ferran Prados Carrasco (2020)

Primera edició: febrer 2020

Av. Tibidabo, 39-43, 08035 Barcelona
Realització editorial: FUOC

Índex

Introducció

La **recuperació d'informació** és una disciplina que s'ocupa de la representació, l'emmagatzematge, l'organització i l'accés a elements d'informació. L'objectiu de la recuperació d'informació és obtenir informació que pugui ser útil o rellevant per a l'usuari.

Presentarem la recuperació d'informació com a disciplina científica, proporcionant una caracterització formal basada en la noció de rellevància. Exposarem els criteris d'avaluació, la forma com es processen els documents per obtenir un índex i els diferents models clàssics de recuperació.

1. Recuperació d'informació

La recuperació d'informació[1] no és un àrea nova, sinó que s'ha anat desenvolupant des de les acaballes de la dècada dels cinquanta del segle passat. No obstant això en l'actualitat, el fet de disposar de la informació necessària dintre del termini i en la forma escaient pot implicar l'èxit o el fracàs d'un projecte.

(1)Sovint abreujada com a IR, per les seves sigles en anglès, *information retrieval*.

La **recuperació d'informació** és una «disciplina que s'ocupa de la representació, l'emmagatzematge, l'organització i l'accés a elements d'informació», segons Baeza-Yates. Anys abans Salton va proposar la definició «un camp relacionat amb l'estructura, l'anàlisi, l'organització, l'emmagatzematge, la cerca i la recuperació d'informació».

Les definicions següents incideixen explícitament en el paper de l'usuari com a font de consultes i destinatari de les respostes. Així, Croft considera la recuperació d'informació com «el conjunt de tasques mitjançant les quals l'usuari localitza i accedeix als recursos d'informació que són pertinents per a la resolució del problema plantejat. En aquestes tasques exerceixen un paper fonamental els llenguatges documentals, les tècniques de resum, la descripció de l'objecte documental, etc.». D'altra banda, Korfhage indica que «la localització i la presentació a un usuari d'informació rellevant a una necessitat d'informació expressada mitjançant una consulta».

La recuperació d'informació intenta resoldre el problema de «trobar i organitzar documents rellevants que satisfacin la necessitat d'informació d'un usuari, expressada en un determinat llenguatge de consulta».

1.1. Recuperació d'informació i recuperació de dades

Encara que puguin semblar conceptes molt semblants, hi ha diferències significatives quant als objectes amb què es tracta i la seva representació, però també pel que fa a l'expressió de les consultes i els resultats.

En la **recuperació de dades** els objectes considerats són estructures de dades conegudes. La seva representació es basa en un format previ ben definit i amb un significat implícit per a cada element. Per exemple, una taula en una base de dades que emmagatzema instàncies de clients d'una organització posseeix un conjunt de columnes que defineixen els atributs de tots els clients i cada fila correspon a un client en particular. Cada element (atribut) té un domini conegut i la seva semàntica està clarament establerta.

D'altra banda, en la **recuperació d'informació** l'objecte de tractament és bàsicament un document de text, en general sense estructura.

Quant a les consultes, la recuperació de dades compta amb una estructura ben definida donada per un llenguatge de consulta que permet la seva especificació de manera exacta. Les consultes no són ambigües i consten d'un conjunt de condicions que han de complir els elements.

Bases de dades relacionals

En bases de dades relacionals, les consultes s'especifiquen utilitzant el llenguatge SQL (*Structured Query Language*), la semàntica del qual és precisa.

Finalment, en un sistema de recuperació de dades els resultats consisteixen en un conjunt complet d'elements que satisfan totes les condicions de la consulta. Com que la consulta no admet errors, el resultat ha de ser exacte. I l'ordre d'aparició dels resultats és l'especificat per la consulta (o cap), però no té cap influència en la importància del resultat.

Taula 1. Diferències entre recuperació de dades i recuperació d'informació

	Recuperació de dades	**Recuperació d'informació**
Estructura	Informació estructurada amb una semàntica ben definida.	Informació no estructurada.
Recuperació	**Determinística**. Tot el conjunt «Solució» és rellevant per a l'usuari.	**Probabilística**. Una part dels documents recuperats pot no ser rellevant.
Consulta	**Especificació precisa**. Llenguatge formal, precís i estructurat	**Especificació imprecisa**. Llenguatge natural, ambigu i no estructurat.
Resultats	Encerts exactes.	Encerts parcials.

1.2. El problema de la recuperació d'informació

De forma general, el problema de la recuperació d'informació s'ha d'abordar des de dos punts de vista: el computacional i l'humà. El **computacional** té a veure amb la construcció d'estructures de dades i algorismes eficients que millorin la qualitat de les respostes. En el cas **humà** es refereix a l'estudi del comportament i de les necessitats de l'usuari.

Des d'un punt de vista general, el problema de la recuperació d'informació consta dels elements següents:

- Un **conjunt de documents** que contenen informació d'interès (sobre un o diversos temes).

- Uns **usuaris** amb unes necessitats d'informació que plantegen al sistema de recuperació d'informació en forma d'una **consulta**.

- Com a resposta, el sistema retorna referències a **documents rellevants**, és a dir, aquells que satisfan la necessitat d'informació sol·licitada, generalment en forma d'una llista ordenada.

Per complir amb els seus objectius, aquests sistemes han de realitzar algunes tasques bàsiques, que estan formulades com a procediments computacionals:

- Representació lògica dels documents. Alguns sistemes solament emmagatzemen parts dels documents, mentre que altres ho fan de manera completa.

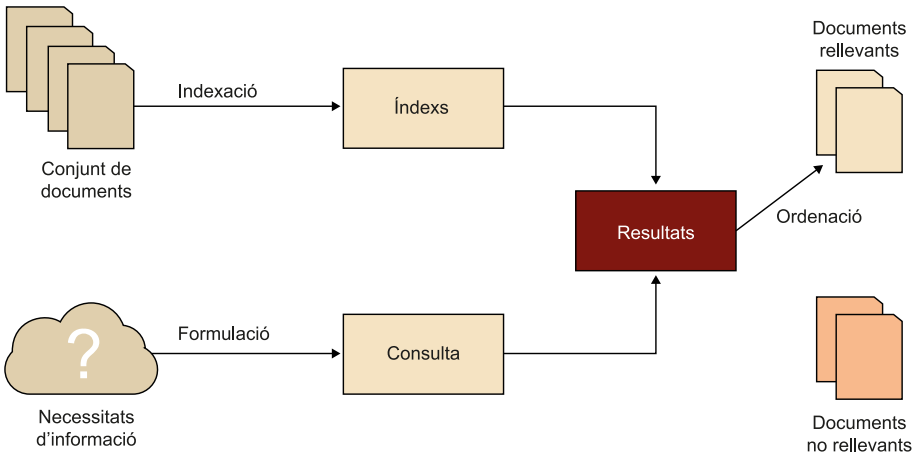
- Representació de la necessitat d'informació de l'usuari en forma de consulta.

- Avaluació dels documents respecte d'una consulta per establir la rellevància de cadascun.

- Ordenació (rànquing) dels documents considerats rellevants per formar el conjunt «Solució» o «Resposta».

- Presentació de la resposta a l'usuari.

Figura 1. El problema de la recuperació d'informació



Com es pot observar, es parteix d'un conjunt de documents, els quals estan compostos per successions de paraules amb la seva estructura gramatical (paràgrafs, seccions, etc.). Aquests documents solen estar escrits en llenguatge natural. El conjunt de tots els documents tractats pel sistema es denomina **corpus** o **col·lecció**. Per poder realitzar operacions sobre un corpus, cal obtenir en primer lloc una representació lògica de tots els seus documents, que pot consistir en un conjunt de termes, frases o altres unitats (sintàctiques o semàntiques) que permetin caracteritzar-los.

A partir d'aquesta representació lògica dels documents, un procés d'**indexació** durà a terme la construcció d'estructures de dades, denominades **índexs**, que emmagatzemin aquesta representació. Aquestes estructures permeten realitzar cerques eficients.

L'**algorisme de cerca** accepta com a entrada una expressió de consulta de l'usuari i comprovarà en l'índex quins documents poden satisfer-la. A continuació un **algorisme d'ordenació** o rànquing determinarà la rellevància de cada document i retornarà una llista amb la resposta. Se sobreentén que el primer element d'aquesta llista correspon al document més rellevant respecte de la consulta.

En general, un sistema de recuperació d'informació no dona una resposta directa a una consulta, sinó que permet localitzar referències a documents que poden contenir informació útil.

1.3. Definint la rellevància

La recuperació d'informació es defineix com la disciplina que en una consulta troba documents rellevants en lloc de simples coincidències de patrons. Això revela un aspecte fonamental: la **rellevància** dels resultats s'avalua en relació amb la necessitat d'informació, no amb la consulta.

Vegem un exemple. Suposem la necessitat d'informació per determinar si menjar xocolata és beneficiós per reduir la pressió arterial. Això es podria expressar en una consulta a un motor de cerca: «xocolata efecte pressió». S'avaluarà un document resultant com a rellevant si respon a la necessitat d'informació, i no solament perquè conté totes les paraules de la consulta.

Cal destacar que la rellevància és un concepte amb propietats interessants:

1) La primera és que és **subjectiu**: dos usuaris poden tenir la mateixa necessitat d'informació i emetre judicis diferents sobre el mateix document recuperat.

2) Una altra característica és la seva **naturalesa dinàmica**, tant a l'espai com en el temps: els documents recuperats i mostrats a l'usuari en un moment donat poden influir en la rellevància dels documents que es mostraran més endavant. A més, d'acord amb el seu estat actual, un usuari pot expressar diferents judicis sobre el mateix document (donada la mateixa consulta).

3) I finalment, la rellevància és **multifacètica**, ja que està determinada no únicament pel contingut d'un resultat recuperat, sinó també per factors com ara l'autoritat, la credibilitat, l'especificitat, l'exhaustivitat, l'actualitat i la claredat de la font.

Cal assenyalar que la rellevància no és coneguda pel sistema abans del judici de l'usuari. De fet, podríem dir que la tasca d'un sistema de recuperació d'informació és «endevinar» un conjunt D de documents rellevants pel que fa a una consulta donada q calculant una funció de rellevància R per a cada document de la col·lecció.

1.4. Tractant amb dades no estructurades

Una de les dificultats fonamentals en el procés de recuperació d'informació és la naturalesa no estructurada (generalment text) dels documents i una altra és la grandària ingent de les col·leccions de documents.

Un punt clau en la recuperació d'informació pel que fa a la recuperació de dades és la seva naturalesa no estructurada. La recuperació de dades, tal com la realitzen les bases de dades relacionals, es refereix a recuperar tots els objectes que satisfan unes condicions clarament definides i expressades per mitjà d'un llenguatge de consulta formal. En aquest context, les dades tenen una estructura ben definida i s'hi accedeix per mitjà de llenguatges de consulta com ara SQL. A més, els resultats tenen coincidències exactes, és a dir, no es retornen les coincidències parcialment correctes com a part de la resposta. Per tant, la rellevància no s'aplica a la recuperació de dades.

D'altra banda, la «societat digital» està produint una gran quantitat de continguts. De fet, mentre que el 2006 es van crear o replicar a tot el món al voltant de 10^{18} bytes (10K petabytes) d'informació, el 2010 aquest nombre va augmentar en un factor de 6 (988 exabytes, és a dir, gairebé un zettabyte). Llavors va començar l'**era del zettabyte**, que és un període de la història de la humanitat i de la informàtica en què el tràfic a internet global va superar per primera vegada un zettabyte (fet que va ocórrer el 2016) i la quantitat de dades digitals al món va superar per primera vegada el zettabyte (cosa que va succeir el 2012).

Petabytes i zettabyte

Un **petabyte** és una unitat d'emmagatzematge d'informació que equival a 10^{15} bytes.

Un **zettabyte** és una unitat d'emmagatzematge d'informació que equival a 10^{21} bytes (= 1.099.511.627.776 gigues).

1.5. Definició formal

Un sistema de recuperació d'informació (SRI) es pot caracteritzar per un quatern:

$$\text{SRI} = \{D, Q, F, R(q_k, d_j)\}$$

on:

- D és el conjunt de representacions dels documents en la col·lecció (cadascun referenciat com a d_j).

- Q és el conjunt de les representacions de les necessitats d'informació de l'usuari, denominades **consultes** (referenciades individualment com a q_k).

- F és un marc de treball o estratègia per modelar la representació dels documents, les consultes i les seves relacions.

- $R(q_k, d_j)$ és una funció d'ordenació o rànquing que associa un nombre real a cada document d_j segons la seva rellevància respecte a la consulta q_k.

La funció $R(q_k, d_j)$ determina l'ordre de rellevància dels documents i és la peça clau en tot el procés de recuperació d'informació.

1.6. Aplicacions de la recuperació d'informació

L'aplicació dels sistemes de recuperació d'informació més coneguda i estesa són els motors de cerca, però les tècniques de recuperació d'informació també són fonamentals per a altres tasques.

Un **motor de cerca** és un sistema de recuperació d'informació dissenyat per ajudar a trobar informació emmagatzemada en sistemes informàtics. Els resultats de cerca generalment es presenten com una llista i es denominen **resultats**. Els motors de cerca ajuden a minimitzar el temps necessari per trobar informació.

Motor de cerca

Un exemple són els cercadors d'internet (generalment a la web).

Els **sistemes de filtratge** eliminen informació redundant o no desitjada que apareix en un flux d'informació, mitjançant mètodes automàtics, abans de presentar-la als usuaris. Una aplicació clàssica de filtratge d'informació són els filtres d'*spam*, que aprenen a distingir entre correus electrònics útils i correus nocius en funció del contingut.

Els **sistemes de recomanació** (es poden considerar com una forma de filtratge d'informació) presenten a l'usuari elements d'informació interessants, com ara cançons, pel·lícules, llibres, etc. en funció del seu perfil o de les eleccions d'elements semblants per proximitat geogràfica, coneixement o interessos comuns.

El **resum de documents** és una altra aplicació consistent a crear una versió abreujada d'un text per reduir la sobrecàrrega d'informació. El resum sol ser generalment extractiu, és a dir, se seleccionen les frases més rellevants d'un document i es recopilen per formar una versió més compacta del document.

L'agrupació i la categorització de documents també són aplicacions importants en la recuperació d'informació. L'**agrupació** consisteix a reunir documents en funció de la seva afinitat, mentre que la **categorització** utilitza una sèrie

predefinida de classes i assigna cada document a la classe més rellevant. Les aplicacions típiques de la categorització són la identificació de categories en articles i en notícies.

Els **sistemes de resposta a preguntes**[2] s'ocupen de la selecció de documents rellevants per respondre a les consultes de l'usuari, formulades en un llenguatge natural. La característica principal d'aquests sistemes és que proporcionen respostes en forma d'oracions, frases o, fins i tot, paraules rellevants, depenent del tipus de pregunta formulada.

(2)QA, per les seves sigles en anglès, *Question Answering*.

Finalment, un assumpte interessant es refereix a la recuperació en diversos idiomes, és a dir, la recuperació de documents en un idioma diferent de l'idioma en què es va formular la consulta de l'usuari. Una aplicació notable d'aquesta tecnologia es refereix a la recuperació de documents legals.

2. Avaluació d'un sistema de recuperació d'informació

Fins a aquest moment s'ha considerat la rellevància com el criteri clau per determinar la qualitat d'un sistema de recuperació d'informació, destacant el fet que es refereix a una necessitat implícita de l'usuari.

Per formalitzar aquesta qüestió, es diu que un sistema de recuperació d'informació serà **mesurable** en termes de rellevància quan es disposa de la informació següent:

1) Una col·lecció D de documents de referència,

2) un conjunt Q de consultes de referència, i

3) una terna $t_{jk} = \langle d_j, q_k, r* \rangle$ per a cada consulta $q_k \in Q$ i cada document $d_j \in D$ que conté un judici binari de rellevància $r*$ del documento d_j pel que fa a la consulta q_k, judici emès per una autoritat de referència.

En termes generals, la **precisió** (P) és la fracció de documents recuperats que són rellevants per a una consulta i proporciona una mesura de la «solidesa» del sistema.

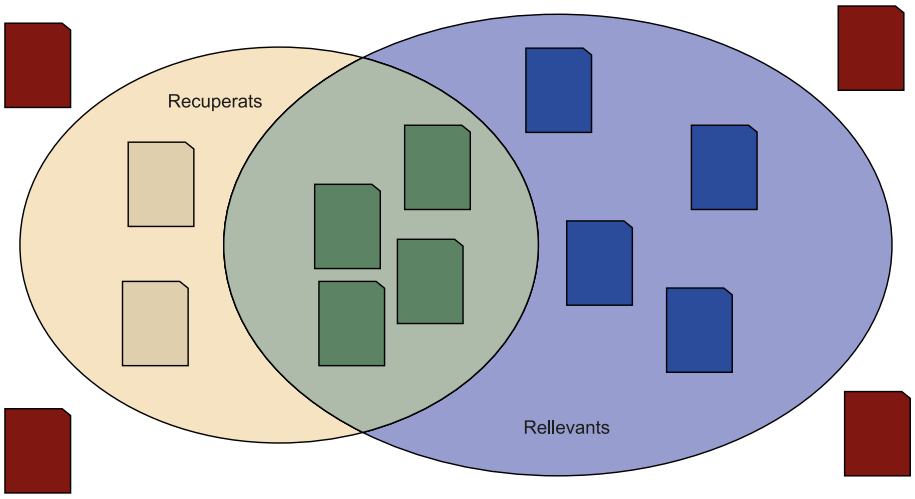
La precisió no té a veure amb el nombre total de documents que el sistema considera rellevants. Aquest detall s'explica mitjançant l'**exhaustivitat**[3] (R), que es defineix com la fracció de documents «veritablement» rellevants que es recuperen efectivament i, per tant, proporciona una mesura de la «integritat» del sistema.

(3)S'ha traduït el terme en anglès *recall* com a exhaustivitat per ser el més aproximat conceptualment.

Taula 2. Matriu de confusió per a l'avaluació

	Rellevant	**No rellevant**
Recuperat	**Veritables positius (TP)**	**Falsos positius (FP)** Error de tipus I
No recuperat	**Falsos negatius (FN)** Error de tipus II	**Veritables negatius (TN)**

Figura 2. Mesures de rendiment



Més formalment, donat el conjunt complet de documents D i una consulta q, es defineix el subconjunt $TP \subseteq D$ com el conjunt de resultats **veritables positius**, és a dir, documents recuperats que són realment rellevants per a la consulta q. I es defineix $FP \subseteq D$ com el conjunt de **falsos positius**, és a dir, el conjunt de documents recuperats que no són rellevants per a la consulta q. També el subconjunt $FN \subseteq D$ com el conjunt de documents que corresponen a les necessitats de l'usuari però que el sistema no recupera. Amb aquesta notació es defineixen la precisió i l'exhaustivitat:[4]

[(4)]Anomenada *R*, de l'anglès, *recall*.

$$\text{Precisió}(P) = \frac{TP}{TP+FP}$$

i

$$\text{Exhaustivitat}(R) = \frac{TP}{TP+FN}$$

Aquestes dues mesures estan altament correlacionades. Empíricament s'ha comprovat que una alta exhaustivitat va acompanyada d'una precisió molt baixa i viceversa, és a dir, es compleix una relació inversa entre les dues mesures.

Hi ha un compromís entre l'exhaustivitat i la precisió, de manera que augmentar l'exhaustivitat (recuperant una major quantitat de documents) fa que disminueixi la precisió (augmentant el nombre de documents no rellevants). Per contra, si recuperem uns quants documents i tots són rellevants, es tindrà una precisió màxima, però segurament hi haurà documents rellevants que no es recuperaran. El sistema ideal és aquell que sempre recupera tots els documents rellevants i solament aquests.

Com que la precisió i l'exhaustivitat tenen diferents avantatges i desavantatges, s'ha definit una única mesura d'avaluació que equilibra els dos mesuraments. Es denomina **Mesura-F1** (*Score-F1*) o **mitjana harmònica** de P i R:

$$F_1 = 2\frac{PR}{P+R} = \frac{2 \times TP}{2 \times TP + FN + FP}$$

Aquesta mesura combina la precisió i l'exhaustivitat en un únic valor comprès entre 0 i 1. L'interessant d'aquesta mètrica és que el valor màxim d'F_1 correspon al millor compromís entre P i R, i el seu valor solament serà alt quan ambdues mesures tinguin valors alts. Si $F_1 = 0$, no s'han recuperat documents rellevants, mentre que si $F_1 = 1$ s'han recuperat tots els documents rellevants i solament aquests.

L'**exactitud**[5] (A) és la mesura de rendiment més intuïtiva i és simplement una relació entre el recuperat correctament (TP i TN) i el total de documents:

[5] Abreujada com a A, de l'anglès *accuracy*.

$$\mathrm{A} = \frac{TP + TN}{TP + TN + FP + FN}$$

El **soroll**[6] (N) determina la proporció de documents no rellevants trobats en els documents recuperats:

[6] Abreujat com a N, de l'anglès *noise*.

$$N = \frac{FP}{TP + FP}$$

3. Preprocessament

És evident que recórrer tots els documents d'una col·lecció cada vegada que es fa una consulta és una solució poc pràctica i, en general, impossible. Per evitar això, cal indexar els documents amb antelació.

Un **índex** és una vista lògica que representa els documents d'una col·lecció mitjançant un conjunt de termes o paraules clau, això és, qualsevol paraula que aparegui en el text del document. Un **terme** és una instància d'una seqüència de caràcters agrupats per al seu processament i que tenen una unitat semàntica.

La idea que subjau en la indexació és que tant la semàntica dels documents com les necessitats d'informació de l'usuari es poden expressar adequadament mitjançant conjunts de termes i les relacions entre ells. Això es deu al fet que no tots els termes que componen un document són igualment representatius del seu contingut. Qüestions com ara la seva posició, la quantitat d'aparicions o la seva funció lingüística defineixen el grau d'importància de cadascun dels termes.

El resultat és una representació de la col·lecció computacionalment adequada per als processos següents i es denomina **indexació de la col·lecció**.

3.1. Procés d'indexació

La **indexació** és una operació que té com a propòsit la identificació dels termes que representen el contingut d'un document i la traducció d'aquests en una forma computacionalment manejable.

El concepte d'indexació inclou la construcció d'estructures de dades que permetin emmagatzemar els termes representatius per possibilitar posteriorment la recuperació eficient dels documents.

3.1.1. Operacions amb el text

Quan considerem un text en llenguatge natural és fàcil notar que no totes les paraules siguin igualment representatives de la semàntica del document. En general, els substantius (o grups de paraules que contenen substantius) són els components més representatius d'un document en termes de contingut.

Segons això, el sistema de recuperació d'informació també processa prèviament el text dels documents per determinar els termes més «importants» que s'utilitzaran per construir l'índex. Per tant, se selecciona un subconjunt de totes les paraules per representar el contingut d'un document.

Per seleccionar les paraules clau, la indexació ha de complir dos objectius diferents i potencialment oposats:

1) Ser **exhaustiu**, és a dir, usar un nombre suficientment gran de termes del document.

2) L'**especificitat**, és a dir, excloure els termes genèrics que tenen poca semàntica i que engrandeixen la grandària de l'índex.

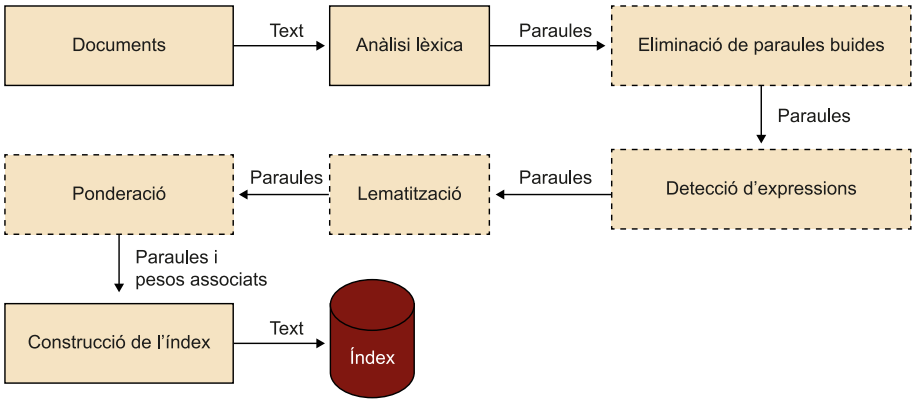
Certs termes molt genèrics, com ara els articles, les conjuncions o les preposicions, tenen un baix poder discriminant, ja que la seva freqüència en qualsevol document tendeix a ser alta. En altres paraules, els termes genèrics tenen una **freqüència de terme** alta. Per contra, els termes específics tenen un poder discriminant més alt, a causa de la seva escassa aparició: tenen una **freqüència de document** baixa.

Freqüència de terme i de document

Freqüència de terme (*term frequency*) definida com el nombre de vegades que apareix el terme en el document.

Freqüència de document (*document frequency*) definida com el nombre de documents en què apareix el terme.

Figura 3. Procés d'indexació



A continuació es descriuen les fases de preprocessament realitzades per un sistema de recuperació d'informació, prenent com a entrada un document i produint com a sortida els termes per a l'índex.

1) Anàlisi del document. Els documents estan en tot tipus d'idiomes, jocs de caràcters i formats. Fins i tot el mateix document pot contenir múltiples idiomes o formats. L'anàlisi s'ocupa del reconeixement i de la descomposició de l'estructura del document en els seus components.

2) Anàlisi lèxica. L'anàlisi lèxica converteix el document en un conjunt de paraules o *tokens*. Hi ha una sèrie de dificultats relacionades amb l'anàlisi lèxica que inclouen la identificació correcta dels separadors de les paraules (si l'idioma els té), abreviatures o dates. Aquesta complexitat depèn molt

de l'idioma del document. El reconeixement d'abreviatures i, en particular, d'expressions de temps presenta força complexitat i hi ha diversos estudis sobre aquest fet.

3) Eliminació de paraules buides (*stop-words*). Un pas posterior és l'eliminació de paraules buides, és a dir, l'eliminació de paraules d'alta freqüència amb poca càrrega semàntica, encara que amb sentit gramatical. Hi ha llistes de paraules buides per a cada idioma, que inclouen, normalment, articles, preposicions, conjuncions, etc. No obstant això, com que aquest procés pot disminuir l'exhaustivitat, en alguns motors de cerca no la implementen.

4) Detecció d'expressions. Aquest pas intenta capturar el significat del text més enllà d'una llista de paraules mitjançant la identificació de grups de substantius i altres expressions. La detecció de frases es pot abordar de diverses maneres, com ara l'ús de regles (per exemple, la retenció de termes que no estan separats per signes de puntuació), l'anàlisi morfològica o l'anàlisi sintàctica. Un enfocament comú per a la detecció d'expressions es basa en l'ús de **tesaurus**. Alternativament hi ha tècniques d'aprenentatge automàtic, com l'algorisme d'extracció de claus (KEA),[7] que identifica expressions candidates utilitzant mètodes lèxics.

[7] Per les seves sigles en anglès *Key Extraction Algorithm*.

Tesaurus

Un **tesaurus** és una llista de paraules o termes controlats, emprats per representar conceptes. Els tesaurus fets manualment són generalment jeràrquics i contenen termes relacionats, exemples d'ús i casos particulars.

5) Lematització. El pas següent intenta normalitzar les paraules mitjançant l'eliminació de sufixos. L'objectiu és obtenir la forma bàsica (diccionari) de cada paraula eliminant la part flexiva d'aquesta que s'usa per formar els plurals, el gènere, les conjugacions verbals, les formes adverbials, etc. El mètode clàssic d'abordar això va ser ideat per Porter mitjançant un algorisme basat en regles.

6) Ponderació. La fase final del processament és la ponderació dels termes. Com s'ha esmentat anteriorment, cada paraula en un text té un poder descriptiu diferent i, per tant, els termes es poden ponderar de manera diferent per tenir en compte la seva importància dins d'un document o en tota la col·lecció de documents.

Un dels mètodes de ponderació més usat és el denominat **TF*IDF** (*Term Frequency*, *Inverse Document Frequency*), que estableix una relació entre la freqüència d'un terme dins d'un document i la seva freqüència en tots els documents de la col·lecció (*N*), és a dir, s'obté la freqüència del terme t_i en el document $d_j(TF)$ i es multiplica pel recíproc de la quantitat de documents (*n*) de la col·lecció en què apareix $t_i(IDF)$:

$$TF * IDF_{ij} = TF_{ij} \times \log_2(N/n)$$

En el càlcul de l'*IDF* els valors propers a 0 indiquen que el terme posseeix poc pes i, per tant, un valor baix de discriminació. Per contra, els valors allunyats de 0 indiquen que el terme és poc freqüent i resulta més adequat per caracteritzar els documents en què es troba.

En general, la indexació es basa en l'anàlisi de la freqüència dels termes i la seva distribució en els documents. Aquesta anàlisi té com a objecte establir criteris que permetin determinar si una paraula és un terme d'indexació vàlid, fonamentalment perquè permet discriminar el contingut dels documents i, d'alguna manera, aporta informació.

3.1.2. Lleis empíriques sobre el text

Hi ha algunes propietats interessants en el llenguatge i el seu ús que poden ser útils per comprendre el procés d'indexació, ja que determinen com es distribueixen les freqüències d'aparició de les diferents paraules en una col·lecció i com creix la grandària del vocabulari a mesura que creix aquesta col·lecció.

Llei de Zipf

Formulada en la dècada dels quaranta del segle passat per George K. Zipf, lingüista de la Universitat de Harvard, que va realitzar una sèrie d'estudis empírics que van demostrar que la gent, quan escriu, tendeix a preferir paraules més conegudes a les menys conegudes. Va denominar a això *llei del mínim esforç*.

La **llei de Zipf** estableix que, donada una llista de paraules juntament amb la freqüència d'aparició de cadascuna ordenades de major a menor, es compleix que la freqüència $f(w)$ d'una paraula multiplicada per la seva posició $r(w)$ en una llista ordenada, és igual a una constant C que depèn de l'idioma, és a dir:

$$C = r(w) \times f(w)$$

Paraules més freqüents

En espanyol, les deu paraules més freqüents segons la RAE són *de*, *la*, *que*, *el*, *en*, *y*, *a*, *los*, *se*, *del*. En concret *la* apareix la meitat de vegades que *de*, *que* un terç de vegades que *de*, i així successivament.

La llei de Zipf és una llei de potències, la qual cosa vol dir que és igual la grandària del text que estiguem estudiant i que aquesta proporció en la freqüència d'aparició de les paraules sempre es compleix.

A més la llei de Zipf s'aplica a tots els idiomes, independentment de la família a la qual pertanyen. Per tant, té a veure amb la forma com el cervell processa el llenguatge.

Llei de Heaps

Aquesta llei planteja la relació entre la grandària del text (nombre de paraules) i el creixement del vocabulari (nombre de paraules úniques). En particular, determina que la grandària del vocabulari V (i el seu creixement) és una funció de la grandària del text N (mesurat en paraules):

$$V = K\,N^{\beta}$$

on K és una constant (entre 10 i 100) i β és una altra constant entre 0 i 1 (normalment entre 0,4 i 0,6).

Aquesta troballa és molt important per l'escalabilitat del procés d'indexació, atès que estableix que la grandària del vocabulari (i la grandària de l'índex) presenta un creixement sublineal pel que fa al creixement del nombre de documents.

3.2. Estructures de dades per a la indexació

En aquest apartat es presenten les estructures de dades bàsiques per a la implementació de sistemes de recuperació d'informació. A partir dels conceptes i de les tècniques exposades en l'apartat anterior sobre el preprocessament, resulta necessari comptar amb estructures de dades eficients que suportin les estratègies de les cerques. La justificació de la indexació és que el cost (en termes de temps i espai d'emmagatzematge) dedicat a la creació de l'índex es recupera en l'execució de múltiples consultes.

Per tant, la primera pregunta que cal abordar quan s'afronta la indexació és quina estructura d'emmagatzematge hauria d'usar per maximitzar l'eficiència de la recuperació. Una primera solució simplista utilitzaria una matriu de documents i termes, és a dir, una matriu on les files corresponen als termes i les columnes corresponen als documents de la col·lecció. D'aquesta manera, cada cel·la w_{ij} representa el pes del terme t_i en el document d_j.

Taula 3. Matriu d'associació terme-document

	d_1	d_2	d_3	...	d_n
t_1	w_{11}	w_{12}	w_{13}	...	w_{1n}
t_2	w_{21}	w_{22}	w_{23}	...	w_{2n}
t_3	w_{31}	w_{32}	w_{33}	...	w_{3n}
...	...	...	...	...	...
t_m	w_{m1}	w_{m2}	w_{m3}	...	w_{mn}

No obstant això, en el cas de grans col·leccions de documents, aquest criteri donaria com a resultat una matriu amb molts pocs valors (i molts buits), ja que la probabilitat que cada paraula aparegui en un document de la col·lecció disminueix amb el nombre de documents.

A continuació es presenten algunes millores a aquest plantejament.

3.2.1. Índexs invertits

El fonament d'un índex invertit és molt senzill. Primer cal crear un diccionari de termes V (també denominat *vocabulari*) que conté tots els termes únics de la col·lecció de documents.

A continuació, per a cada terme $t_i \in V$ es crea una llista L_i que conté la referència a cada document d_j en què t_i apareix. Aquesta llista L_i es denomina **llista de publicació** o **llista invertida** i pot contenir informació addicional, com ara la freqüència (TF*IDF) o la posició de t_i dins de d_j.

El conjunt del diccionari de termes i totes les seves llistes invertides es denomina **índex invertit**.

Taula 4. Índex invertit amb informació de freqüència

Vocabulari	Llistes invertides
t_1	(d_1, 1), (d_3, 4), (d_5, 2), ..., (d_n, 1)
t_2	(d_3, 2)
t_3	(d_1, 1), (d_2, 6)
...	...
t_m	(d_2, 3)

Els índexs invertits no tenen rival quant a eficiència: de fet, com que un terme generalment apareix en molts documents, es redueixen les necessitats d'emmagatzematge. A més, totes aquestes estructures admeten la compressió de manera que puguin cabre en la memòria.

Donada aquesta estructura d'índex invertit, el procés de cerca consta de quatre passos principals:

1) Accedir al diccionari de termes per identificar els termes de la consulta.

2) Per a cada terme de la consulta es recuperen les llistes invertides.

3) Filtrar els resultats: si la consulta es compon de diversos termes (possiblement connectats pels operadors lògics), cal fusionar les llistes de resultats parcials.

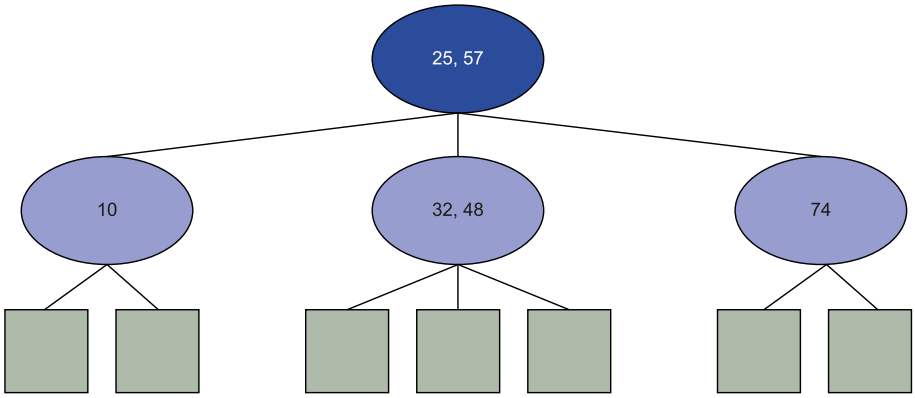
4) Lliurar la llista de resultats.

3.2.2. Arbres B i B+

Els **arbres B** constitueixen una categoria molt important d'estructures de dades que permeten una implementació eficient de conjunts i diccionaris per a les operacions de consulta. Hi ha una gran varietat d'arbres B: els arbres **B**, **B+** i **B***, encara que tots ells estan basats en la mateixa idea, inclouen la utilització d'arbres de cerca no binaris i amb la condició de balanceig.

Els **arbres B+**, en concret, són àmpliament utilitzats en la representació d'índexs en bases de dades. De fet, aquest tipus d'arbres està dissenyat específicament per a aquestes aplicacions, en què la característica fonamental és el temps en les operacions d'accés a les dades.

Figura 4. Arbre B balancejat



Els **arbres B** s'usen molt en bases de dades a causa del seu temps d'accés reduït: de fet, el nombre màxim d'accessos per a un arbre B està limitat a la profunditat de l'arbre. Un inconvenient dels arbres B és el seu baix rendiment en les cerques seqüencials. Aquest problema pot ser mitigat per la variant de l'arbre B+, en què els nodes fulla o finals estan vinculats formant una cadena que segueix un ordre.

Un altre problema que solen plantejar els arbres B és que poden perdre el balanceig després de moltes insercions. Això es pot modificar adoptant procediments de regeneració del balanceig de l'arbre.

4. Models de recuperació d'informació

En aquest apartat es presenten tres models clàssics de recuperació d'informació: el **booleà**, el d'**espai vectorial** i el **probabilístic**. Aquests models proporcionen els fonaments de l'avaluació de les consultes, que és el procés que recupera els documents rellevants segons la consulta d'un usuari.

4.1. Model booleà

El **model de recuperació booleana** és un model de recuperació basat en la teoria de conjunts i en l'àlgebra de Boole, mitjançant el qual les consultes es defineixen com a expressions booleanes amb termes d'índex (i utilitzant els operadors booleans `AND`, `OR` i `NOT`). Per exemple, «foto AND muntanya OR neu».

En el model booleà la representació de la col·lecció de documents es realitza sobre una matriu binària document-terme, on els termes han estat extrets manualment o automàticament dels documents i representen el seu contingut.

Taula 5. Matriu binària terme-document

	d_1	d_2	d_3	...	d_n
t_1	0	0	1	...	0
t_2	1	0	1	...	1
t_3	0	1	0	...	0
...	...	...	...	...	...
t_m	0	0	0	...	1

Una consulta booleana q es pot resoldre recuperant tots els documents que contenen els termes de la consulta i creant una llista per a cada terme. Una vegada que aquestes llistes estiguin disponibles, els operadors booleans s'han de manejar de la manera següent:

- q_1 `OR` q_2: requereix construir la **unió** de les llistes de q_1 i q_2.
- q_1 `AND` q_2: requereix construir la **intersecció** de les llistes de q_1 i q_2.
- q_1 `AND NOT` q_2: requereix construir la **diferència** de les llistes de q_1 i q_2.

Es tracta d'un model molt senzill, però mitjançant algunes extensions es permet usar el caràcter comodí * per indicar l'acceptació de coincidències de termes parcials (`mont* OR nie*`). Altres extensions inclouen l'operador de proximitat `NEAR`, és a dir, una forma d'expressar que dos termes en una consulta han d'aparèixer a prop un de l'altre en un document (*rock NEAR roll*).

4.2. Model d'espai vectorial

El **model d'espai vectorial** representa documents i consultes com a vectors d'un espai vectorial en què cada dimensió correspon a un terme del vocabulari.

El fonament d'aquest model és que cada terme t_i del diccionari V es representa com un vector d'un espai vectorial euclidià de dimensió $|V|$, en què cada t_i té tots els seus components iguals a 0 excepte el corresponent a la dimensió associada a t_i en aquest espai que val 1.

Així, en un espai vectorial de dimensió 5 tindríem $t_{\text{rock}} = [1, 0, 0, 0, 0]$, $t_{\text{pop}} = [0, 1, 0, 0, 0]$, $t_{\text{jazz}} = [0, 0, 1, 0, 0]$, $t_{\text{heavy}} = [0, 0, 0, 1, 0]$ i $t_{\text{dansi}} = [0, 0, 0, 0, 1]$.

D'aquesta forma, qualsevol consulta q i qualsevol document $d_j \in D$ es poden representar a l'espai vectorial com:

$$q = \sum_{i=1}^{|V|} w_{iq} \cdot t_i$$

$$d_j = \sum_{i=1}^{|V|} w_{ij} \cdot t_i$$

on w_{iq} i w_{ij} són els pesos assignats al terme t_i per la consulta q i per a cada document d_j.

Aquest model implica que dos vectors document que estiguin «a prop» a l'espai vectorial «tracten» del mateix tema. Per tant, per resoldre els vectors consulta buscarà vectors document propers a l'espai vectorial. Aquesta similitud es pot representar intuïtivament com la projecció d'un vector sobre un altre, idea que s'expressa matemàticament en termes del **producte escalar**:

$$sim(d_j, q) = d_j \cdot q$$

La mètrica de similitud més utilitzada és la **similitud del cosinus**, que és una mesura de la similitud existent entre dos vectors en què s'avalua el valor del cosinus de l'angle que formen. Dit matemàticament:

$$d_j \cdot q = \|d_j\| \times \|q\| \times \cos(\alpha)$$

La similitud del cosinus és una funció de l'angle α format entre d_j i q a l'espai vectorial:

$$Sim_{\cos}(d_j, q) = \cos(\alpha) = \frac{d_j \cdot q}{\|d_j\| \times \|q\|} = \frac{\sum_{i=1}^{|V|}(w_{ij} \times w_{iq})}{\sqrt{\sum_{i=1}^{|V|}(w_{ij})^2}\sqrt{\sum_{i=1}^{|V|}(w_{iq})^2}}$$

La mesura del cosinus realitza una normalització de la longitud del vector (representada com a $\|\cdot\|$), la qual cosa permet descartar errors quan els vectors de longituds molt diferents generen projeccions insignificants.

Alguns avantatges del model d'espai vectorial pel que fa al model booleà resideixen en la seva interpretació geomètrica més intuïtiva i la possibilitat de ponderar les representacions de les consultes i els documents.

4.3. Model probabilístic

La recuperació d'informació és un procés incert: tot el procés en si mateix dista molt de ser exacte.

Un model probabilístic intenta representar la probabilitat de rellevància d'un document donada una consulta, és a dir, calcula la similitud entre les consultes i els documents com la probabilitat que un document d_j sigui rellevant per a una consulta q.

En altres paraules, sigui r la rellevància (en binari) pel que fa a un conjunt de documents D en relació amb una consulta q, el model probabilístic calcula la similitud:

$$Sim(q, d_j) = P(r = 1|\mathrm{q}, d_j), \forall d_j \in D$$

El document que maximitza aquesta probabilitat es recuperarà com el millor resultat i, a continuació, es recuperaran altres documents en ordre decreixent de probabilitat de rellevància.

Cal assenyalar que el conjunt de documents amb rellevància màxima és desconegut *a priori*; per tant, estimar aquesta probabilitat no és una tasca trivial. Una estratègia consisteix a calcular la probabilitat de rellevància mitjançant la

coincidència de termes en la consulta i en els documents. En aquest punt, es pot dur a terme una fase de processament iteratiu (opcionalment utilitzant la retroalimentació de l'usuari) amb la finalitat de millorar el conjunt de resposta.

El model de recuperació probabilístic clàssic és el model d'independència binari, en què els documents (i consultes) es representen com a vectors, és a dir, $d_j = [w_{1j}, ..., w_{|V|}]$ de manera que $w_{ij} = 1$ si i solament si el document d_j conté el terme t_i, i $w_{ij} = 0$ en cas contrari. El model significa que les aparicions d'un terme en els documents són independents, una hipòtesi que generalment funciona a la pràctica.

El model probabilístic té l'avantatge de classificar els documents d'acord amb la seva probabilitat decreixent de ser rellevants. Si a més compta amb la retroalimentació de rellevància (de l'usuari), és certament un avantatge. No obstant això, cal estimar la rellevància inicial dels documents, una tasca que podria no ser fàcil ni precisa.

Bibliografia

Baeza-Yates, R. A.; Ribeiro-Neto, B. A. (2011). *Modern Information Retrieval—the Concepts and Technology Behind Search* (2a. ed.). Harlow: Pearson Education.

Ceri, S.; Bozzon, A.; Brambilla, M.; Della Valle, E.; Fraternali, P.; Quarteroni, S. (2013). *Web Information Retrieval*. Heidelberg: Springer-Verlag.

Croft, W. B. (1987). «Approaches to intelligent information retrieval». *Information Proccessing & Management* (vol. 23, núm. 4, pàg. 249-254).

Grossman D. A.; Frieder O. (2004). *Information Retrieval: Algorithms and Heuristics* (vol. 15). Kluwer Academic, Norwell.

Korfhage, R. R. (1997). *Information Storage and Retrieval*. Nova York: Wiley Computer Publishing.

Manning, C. D.; Raghavan, P.; Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. Disponible a: https://nlp.stanford.edu/IR-book/

Witten, I.; Moffat, A.; Bell, T. (1994). *Managing Gigabytes: Compressing and Indexing Documents and Images*. Massachusetts: Morgan Kaufmann.