Llistat de descriptors lliures i llistat de paraules clau

Manela Juncà Campdepadrós

PID_00144346

Universitat Oberta de Catalunya
www.uoc.edu

Índex

Introducció

Aquest mòdul us introdueix en l'ús dels llenguatges lliures. Comprèn dos llenguatges: els llistats de descriptors lliures i les llistes de paraules clau.

Itinerari d'estudi

El mòdul comença amb les característiques dels llenguatges lliures, ja que els dos darrers llenguatges del curs comparteixen aquesta tipologia.

A continuació, es descriuen les llistes de descriptors lliures, la seva creació i el procés d'indexació i recuperació. Finalment, el curs acaba amb les llistes de paraules clau i la indexació automàtica.

Taula. Conceptes més importants

Concepte	Vegeu
Llenguatge lliure	1. Llenguatges documentals lliures
Descriptor lliure	2. Llistat de descriptors lliures
Paraula clau	3. Llistat de paraules clau
Paraula buida	3.1.2. Funcionament i evolució dels programes d'indexació automàtica
Càlcul de freqüència	3.1.2. Funcionament i evolució dels programes d'indexació automàtica
Freqüència inversa	3.1.2. Funcionament i evolució dels programes d'indexació automàtica
Discriminació	3.1.2. Funcionament i evolució dels programes d'indexació automàtica
Anàlisi morfològica	3.1.2. Funcionament i evolució dels programes d'indexació automàtica
Anàlisi sintàctica	3.1.2. Funcionament i evolució dels programes d'indexació automàtica
Anàlisi semàntica	3.1.2. Funcionament i evolució dels programes d'indexació automàtica

Objectius

Amb l'estudi dels materials associats a aquest mòdul assolireu els objectius següents:

1. Conèixer l'aplicació dels llenguatges lliures en l'anàlisi de contingut.

2. Valorar els avantatges dels llenguatges lliures en la indexació.

3. Valorar els inconvenients dels llenguatges lliures en la recuperació.

4. Quant als **llistats de descriptors lliures**:
- Definir els llistats de descriptors lliures i la seva tipologia.
- Conèixer el procés de creació del llenguatge.
- Adquirir una certa habilitat en la indexació de documents.

5. Pel que fa als **llistats de paraules clau**:
- Introduir-se en els mecanismes de la indexació automàtica.
- Reconèixer les aplicacions estadístiques i lingüístiques aplicades a la indexació automàtica.
- Saber simular/entendre el procés de selecció i tria de paraules clau.

1. Llenguatges documentals lliures

Aquest darrer mòdul està dedicat als dos llenguatges documentals lliures: els llistats de descriptors lliures i els llistats de paraules clau.

Tipologia segons el nivell de control

Recordem que la tipologia segons el nivell de control classifica els llenguatges documentals en lliures o controlats, en funció de si els termes d'indexació corresponen a un llenguatge natural o a un llenguatge artificial construït per a garantir la indexació i recuperació.

A diferència dels llenguatges controlats, que neutralitzaven les mancances dels llenguatges naturals i requerien coneixements elevats per a ser aplicats, els llenguatges lliures no ofereixen tantes garanties en la recuperació, però són amigables. Són llenguatges molt usats, ja que són:

- **Barats**: les despeses de construcció són mínimes. Els elabora una persona o un programa informàtic.

- **Ràpids**: tant en la construcció, que és immediata, com en l'actualització del seu vocabulari, que es va incorporant al llistat a mesura que va ingressant en el fons documental.

- **Terminològicament rics**: de resultes de l'actualització immediata de les seves llistes.

- **Fàcils d'usar**: no necessiten coneixements previs sobre llenguatges documentals, control de vocabulari o precoordinació, etc.

- **Coherents**: ofereixen coherència màxima, de manera que dos serveis d'informació i documentació (SID) amb el mateix document i programari d'indexació automàtica arribaran a una indexació idèntica (aquesta característica només és aplicable al llistat de paraules clau).

Tanmateix, els llenguatges lliures presenten inconvenients en la recuperació, ja que en treballar amb llenguatge natural lliure arrosseguen tots els problemes derivats de l'ambigüitat (sinonímia, polisèmia, homonímia), cosa que provoca silenci i soroll documentals i presenta un nombre excessiu de resultats per cada petició. És el procés invers dels llenguatges controlats, en què la dificultat raïa en el procés d'indexació i assegurava, en canvi, una recuperació unívoca.

Llenguatges controlats →	Dificultat en la indexació →	Facilitat en la recuperació
Llenguatges lliures →	Facilitat en la indexació →	Dificultat en la recuperació

Actualment moltes bases de dades combinen la utilització dels llenguatges controlats amb els lliures, de manera que podem cercar i recuperar per diverses vies. És habitual que els catàlegs bibliogràfics i altres bases de dades ofereixin la possibilitat de cercar la matèria de dues maneres:

a) Pel **camp *matèria***: indexant amb algun llenguatge controlat (tipus llista d'encapçalaments de matèria, tesaurus o una llista d'autoritats).

b) Pel **camp *paraula clau***: indexant amb un programa d'indexació automàtica o llistat de paraules clau.

2. Llistat de descriptors lliures

Un **llistat de descriptors lliures** és un vocabulari monolingüe de termes d'indexació ordenats alfabèticament.

Aquests termes són escollits per l'analista, sense verificar si existeixen ni com s'introdueixen en una llista prèviament establerta. Per tant, no és un llenguatge controlat, sinó lliure, i per això té aquest nom.

2.1. Tipologia

Les llistes de descriptors lliures són un llenguatge documental que té les **característiques** següents:

Vegeu també

Vegeu la tipologia dels llenguatges documentals en l'apartat 5 del mòdul "Anàlisi de contingut: resum i indexació" d'aquesta assignatura

1) Es tracta d'un tipus de **llenguatge analític**. En el procés d'indexació s'identifiquen els conceptes que conformen el contingut del document i es representen mitjançant una sèrie de descriptors. Per aquesta raó, la seva aplicació està lligada a l'existència de sistemes automatitzats que permeten la recuperació de la informació mitjançant la combinació d'aquests descriptors.

Exemple

Posem com a exemple el resum següent:

Dalmau Ribalta, A. (2002). *Els càtars*. Editorial UOC (Biblioteca Lectus Universitaria).

El catarisme és un moviment cristià dissident de la baixa edat mitjana. Es va estendre per diversos punts d'Europa, amb una incidència especial al Llenguadoc. Aquesta obra dóna a conèixer què va ser exactament el catarisme i la seva trajectòria.

L'obra *Els càtars* procura distingir l'anomenada *església de Déu* de tota una literatura esotèrica i llegendària que l'ha deformat de manera significativa i que, al mateix temps, s'allunya d'una concepció historiogràfica tradicional, ja superada, que la considerava com una simple prolongació de l'antic maniqueisme.

Al llarg d'aquesta obra s'explica en què consistia la litúrgia, els ritus i la doctrina religiosa d'aquesta església, considerada herètica i perseguida per l'Església catòlica. Resumeix la seva tràgica història amb fets tan rellevants com l'anomenada *croada albigesa* i el naixement de la Inquisició.

L'analista indexa més d'un terme[1]:

```
Baixa edat mitjana
Càtars
Cristianisme
Dissidència religiosa
Europa
Litúrgia
Llenguadoc
```

[(1)]En contraposició a la LEMAC, que és un llenguatge sintètic i seria un únic terme:

```
LEMAC: Càtars–Història
```

2) Es tracta d'un tipus de **llenguatge natural** perquè la indexació es basa en l'ús de paraules i/o expressions lliures que pertanyen al llenguatge o discurs comú i que es troben en els mateixos documents (títol, resum, text, etc.) i en els conceptes de la consulta de l'usuari. És a dir, en cap moment no es fa servir un codi (numèric, alfanumèric) que representi aquest contingut. El llenguatge natural presenta el problema de la manca d'univocitat, és ambigu, i possibilita els fenòmens de la sinonímia, polisèmia i homonímia.

Exemple

Llistat de descriptors lliures[2]:

```
Càtars
```

[(2)]En contraposició indexat amb un sistema de classificació com la CDU, que és l'únic llenguatge documental codificat, seria:

```
CDU any 2004 27-789.67.
```

3) Es tracta d'un **llenguatge lliure** perquè és format per termes d'indexació extrets del llenguatge natural i que són emprats o suggerits en el mateix document. El llistat de descriptors resultant és constituït per una col·lecció no ordenada de conceptes (només per ordre alfabètic). Aquests conceptes són expressats per paraules o per expressions extretes dels documents, o proposats pels mateixos documentalistes, sense verificar si són en una llista establerta *a priori* (és també, per tant, un llenguatge construït *a posteriori*).

4) Es tracta d'un tipus de **llenguatge postcoordinat**, ja que permet la coordinació de conceptes en el moment de la recuperació i la utilització d'un gran nombre de termes d'indexació o punts d'accés. La combinació precisa d'aquests diferents descriptors, en la recuperació, ha de remetre al document cercat.

5) Es tracta d'un tipus de **llenguatge d'estructura combinatòria o associativa**, en què els conceptes o descriptors s'organitzen d'una manera independent i només es produeix la combinació o intersecció en les operacions d'indexació i recuperació. Per tant, es tracta d'un llenguatge en què els termes que el componen s'organitzen en una llista per ordre alfabètic sense respectar cap altre tipus d'estructura que situï cada descriptor en una cadena lògica, jeràrquica i/o estructurada de conceptes.

Exemple

En aquest cas, l'indexador fa:

```
Càtars AND Europa
```

2.2. Llistat de descriptors lliures en la indexació

El llistat de descriptors lliures és un llenguatge documental que dóna llibertat total a l'analista, la qual cosa el fa molt amigable, ràpid i fàcil d'usar.

Tots nosaltres en algun moment de la nostra vida hem ordenat fotos en àlbums, classificat els llibres de la biblioteca, ordenat articles i fotocòpies sota algun tema o subtema, etc. No n'érem conscients, però estàvem indexant amb un llistat de descriptors lliures.

2.2.1. Creació d'un llistat de descriptors lliures

El **procés de creació** d'una llista de descriptors lliures és molt senzill: l'analista va llegint els documents que ha d'indexar i va prenent nota dels termes que són interessants al seu parer. Els va anotant i els posa per ordre alfabètic. Els dóna un mínim de forma, i procura eliminar duplicitats de nombre i gènere (singular-plural, masculí-femení), usa una única llengua i elimina alguns sinònims.

Els descriptors lliures tenen una despesa de construcció gairebé nul·la. Són immediats, permeten un vocabulari actualitzat i s'adapten a la realitat del SID.

En alguns casos, el llistat de descriptors lliures sol ser el primer pas en la tasca de crear un llenguatge documental. Els analistes comencen per redactar un llistat de descriptors lliures amb la intenció de controlar-lo i avançar cap a un llenguatge documental realment controlat. Si l'analista vol exercir una mica de control sobre els seus descriptors, sense arribar a treballar amb un vocabulari del tot controlat, com serien les llistes d'encapçalaments o els tesaurus, ha de prendre algunes mesures com per exemple:

- Establir una llengua o idioma de la llista.
- Controlar les formes singular/plural.
- Establir els sintagmes nominals preferents:
  - Substantiu + adjectiu.
  - Substantiu + preposició + substantiu.

- Controlar els sinònims.
- Utilitzar formes usuals abans que molt tècniques.

Vegeu també

Les mesures de control de vocabulari detallats en l'apartat dedicat a les llistes d'encapçalaments de matèria del mòdul "Llistes d'encapçalaments de matèria i llistes d'autoritats" són adequats també en aquest cas.

2.2.2. Indexació

L'analista llegeix el document, determina la matèria del document i acte seguit proposa un descriptor. Representant-ho verbalment, l'analista diria "aquest document tracta de tal tema". No consulta cap llista, no comprova si el terme que pensa està acceptat o no, no ha de seguir les regles de combinació de cap llenguatge. Llibertat total a l'hora d'indexar.

El fet de no necessitar una traducció del llenguatge natural a un d'artificial controlat té avantatges i comporta inconvenients. El primer inconvenient i més evident és la disparitat d'indexacions: tres analistes davant el mateix document poden arribar a tres resultats ben diferents.

Disparitat d'indexacions

Seguint amb l'exemple sobre el llibre dels càtars, veiem el resultat de la indexació feta per tres analistes diferents:

Disparitat d'indexacions

Analista a	Analista b	Analista c
Baixa edat mitjana Càtars França Moviments cristians dissidents	Catarisme Heretgies Ordes monàstics Religió	Albigesos Cismes Història medieval Llenguadoc

Com a conseqüència d'això, la taxa de coherència és la més baixa de tots els llenguatges documentals.

Exemple

En l'exemple, l'ús de sinònims i termes genèrics ha alterat tant el resultat que no han coincidit en cap descriptor (la taxa és del 0%).

Un segon aspecte que cal destacar és el grau d'exhaustivitat. Recordem que el grau d'exhaustivitat pot ser alt, mitjà o baix i que afecta els llenguatges postcoordinats com aquest, en què l'analista pot escollir el nombre de termes per a representar un document.

D'aquesta manera amb una llista de descriptors lliures podem indexar el llibre dels càtars amb tres graus diferents:

Exhaustivitat profunda o alta	Exhaustivitat intermèdia o mitjana	Exhaustivitat genèrica o baixa
Baixa edat mitjana Catarisme Càtars Càtars, història Cristianisme Croada albigesa Dissidència religiosa Doctrina religiosa Església de Déu Europa Inquisició Litúrgia Llenguadoc Ritus religiosos	Baixa edat mitjana Càtars Cristianisme Dissidència religiosa Europa Litúrgia Llenguadoc	Baixa edat mitjana Càtars Llenguadoc

a) Els **avantatges** d'aquest llenguatge en la indexació giren entorn de la facilitat, la simplicitat i llibertat d'ús:

- No necessita traducció dels conceptes del llenguatge natural dels documents a un llenguatge artificial.

Exemple

Si en el text dels càtars que estem usant com a exemple es parlava de *literatura esotèrica i llegendària*, l'analista pot indexar **llegendes** o **literatura esotèrica**, sense comprovar-ne la forma en cap llista, si és acceptat o no, si es prefereix la forma singular/plural, si es pot adjectivar o no, si té relacions semàntiques que n'ajudin la indexació.

- Es tracta d'un tipus de llenguatge ràpid i fàcil d'actualitzar. Aquest avantatge és molt apreciat en entorns tecnològics en què la terminologia va més avançada que la seva normalització en llistes controlades. Recordem que entitats com el Termcat són la font adient per a consultar la forma correcta d'escriure el terme.

Exemple

Si el text que cal indexar parla del *Maniqueisme* o qualsevol altre tema, no hi ha problema; s'afegeix a la llista alfabètica de descriptors **Maniqueisme**.

- S'adapta perfectament al nivell d'usuaris i tipus de SID, ja que és un llenguatge fet a la mida del seu centre.

Exemple

Si un SID de tipus genèric rep aquest document l'indexarà amb algun terme del tipus **Religió**; en canvi, si és un SID especialitzat extraurà termes més concrets com, per exemple, **Croada albigesa**.

- No cal una formació prèvia dels analistes. Precisament l'absència de regles i principis fa innecessària la formació.

b) Els **inconvenients** giren entorn de l'elecció del terme d'indexació entre tots els termes possibles. Un segon problema derivat és preveure de quina manera l'usuari farà la cerca i així coincidir.

Exemple

En l'exemple dels càtars, l'analista pot decidir:

- Pel marc geogràfic: Llenguadoc, França, Europa.
- Pel marc temporal: baixa edat mitjana, edat mitjana, història medieval.
- Pel concepte *càtars*: càtars, catarisme, albigès.
- Pel concepte *dissidència*: dissidència, heretgia, cisma...

Alguns analistes opten per indexar el terme tal com apareix en el document (per exemple, Llenguadoc) i d'altres fan l'acció d'escollir termes que creuen que tenen més possibilitats de ser recuperats (per exemple, França).

2.2.3. Aplicacions del llistat de descriptors lliures

Ja hem comentat l'ús generalitzat de descriptors lliures en l'àmbit domèstic i que també pot ser un primer pas en el camí cap a la creació d'un llenguatge controlat, però també troba **aplicacions** en els àmbits següents:

1) En la **indexació d'informació temporal**: quan els ítems d'informació tenen una vida determinada, lligada a un esdeveniment (polític, esportiu, cultural, etc.) concret que fa massa costós invertir en un llenguatge controlat i la seva formació posterior.

2) En la **indexació de documents audiovisuals (imatge fixa o en moviment i àudio) del web**: els sistemes d'indexació automàtica del web no poden, de moment, indexar documents que no continguin text, com són les fotografies o els vídeos de banda ampla, tan nombrosos a la xarxa. La solució passa per indexar-los manualment amb la col·laboració dels internautes.

3) En la **indexació social** els internautes assignen etiquetes amb descriptors als recursos web. Mathes (2004) parla de *metadades generades per l'usuari*. Poden indexar fotografies, webs, blogs i compartir els seus descriptors per a col·laborar en la indexació de tota mena de continguts en l'espai web compartit i obert. Cada recurs (per exemple, el web de la Viquipèdia) és indexat per un grup d'usuaris que poden coincidir o no amb les etiquetes: un pot indexar *wiki*, un altre *enciclopèdia*, un altre *obra de referència*, etc.

Lectura complementària

Si voleu ampliar la informació sobre les metadades generades per l'usuari, podeu consultar la pàgina web d'Adam Mathes:

- Folksonomies

L'assignació de les etiquetes es fa sense ànim de lucre. Els internautes no busquen un guany econòmic, sinó beneficiar-se de cerques millors. Els descriptors escollits no se supervisen ni tenen cap estructura semàntica.

Sinònims

Altres noms que rep el fenomen de la indexació social: *tagging*, *etiquetatge col·laboratiu*, *classificació social*, *etiquetatge social*, *folksonomies*. L'origen del neologisme *folksonomia* el devem a Thomas van der Wal, que va fusionar *folk* ("gent", "popular") i *taxonomia* ("gestió de la classificació"), cosa que va donar com a resultat una "indexació gestionada popularment".

La indexació social participa de les característiques de les llistes de descriptors lliures en la filosofia de la indexació, ja que cada participant indexa uns descriptors lliures seleccionats per un procés intel·lectual a partir de l'examen del recurs, sense verificar si els descriptors proposats són o no en una llista establerta.

Els intents per millorar la recuperació van en la línia d'aplicar algoritmes de ponderació i eliminació d'etiquetes de significat buit per a la comunitat. Aquests mecanismes són totalment invisibles a l'usuari. Alguns autors proposen alfabetitzar l'usuari donant-li instruccions per a indexar, però es dubta de la seva eficàcia, ja que una de les raons de l'èxit del *tagging* és la llibertat que dóna a l'indexador internauta.

2.3. Llistat de descriptors lliures en la recuperació

Els **avantatges** que tenen els llistats de descriptors lliures en la recuperació són els següents:

- Com a llenguatge analític, permet la coordinació de diversos termes en la consulta i això permet cerques precises.

- El llenguatge s'actualitza contínuament.

- Facilitat d'ús en la realització de consultes, ja que no cal dominar prèviament cap llenguatge especialitzat, sofisticat ni artificial.

Tanmateix també tenen **inconvenients**. Si en la indexació tot eren facilitats, en la recuperació trobem tota mena de problemes derivats del silenci i el soroll documentals.

El principal problema d'aquest llenguatge documental es troba precisament en aquest punt, la recuperació, ja que l'abundància de sinònims i homònims fa difícil predir amb quin terme cercarà l'usuari i si coincidirà amb l'escollit per l'indexador.

Les dificultats més notables es presenten a l'hora de formular la petició:

- No elimina la polisèmia del llenguatge natural, cosa que provoca soroll documental.

- Presencia de sinònims, la qual cosa genera silenci documental i impossibilita recuperar els documents que tracten sobre els conceptes de la consulta si aquests han estat indexats pels analistes amb formes diferents.

Exemple de sinònims

Un exemple de cerca amb tota mena de sinònims seria: CEE OR UE OR Unió Europea.

- Variants ortogràfiques del mateix concepte (disc, disquet, disket).

- Variants idiomàtiques (bafles i altaveus).

- Manca d'una estructura semàntica que ajudi a la localització de termes. Aquesta manca provoca buits quan demanem per un terme genèric, amb el qual no ha estat indexat el document.

Exemple de manca d'estructura semàntica

Cerquem, per exemple, sobre llenguatges documentals i no recuperem un document sobre tesaurus perquè està indexat només com a tesaurus, no amb un terme conceptualment més genèric com pot ser *llenguatge documental*. En un llenguatge controlat, aquesta vinculació es faria evident (amb sigles del tipus *TG*, *TA*) i ajudaria a la recuperació.

- Coordinacions falses. Entenem per *coordinacions falses* recuperacions inesperades i errònies que, malgrat contenir els elements de la cerca, són temes diferents.

Exemple de coordinacions falses

Cerquem, per exemple, sobre pintura catalana i demanem Pintura AND Catalunya i recuperem:

- Pintors catalans (exemple, M. Fortuny).
- Catalunya en la pintura (exemple, la visió J. Sorolla sobre el litoral català).
- Pintura a Catalunya (tots aquells pintors que han pintat a Catalunya).
- Industrials de la pintura catalans (pintors de parets).

Les solucions a aquests inconvenients passen per recopilar tota la llista que siguem capaços d'elaborar sobre el concepte: termes genèrics i específics, polisèmics, sinònims, sigles, variants idiomàtiques, canvis de gènere i nombre. Les estratègies de cerca solen ser complexes.

3. Llistat de paraules clau

El **llistat de paraules clau** és un vocabulari ordenat alfabèticament dels termes amb càrrega significativa extrets d'un document mitjançant un programa informàtic.

Llistats de paraules clau

Considerem *llistats de paraules clau* els llistats resultants de la indexació automàtica.

Una llista de paraules clau és constituïda per una col·lecció disposada en ordre alfabètic, de les paraules significatives, anomenades també *no buides* (és a dir, totes les paraules que no són articles, conjuncions, pronoms, preposicions, numerals i certs verbs i adverbis), extretes d'una manera automàtica per l'ordinador a partir del títol, del resum i cada cop més sovint del text complet dels documents (van Slype). Acostumen a ser llistes monolingües.

Lectura complementària

Si voleu ampliar el vostre coneixement sobre la informació donada per van Slype, podeu llegir l'obra següent.

Slype, van G. (1991). *Los lenguajes de indización: concepción, construcción y utilización en los sistemas documentales*. Madrid: Pirámide, pàg. 23 (Fundación Germán Sánchez Ruipérez. Biblioteca del Libro).

Els llistats de paraules clau tenen la mateixa tipologia que els llistats de descriptors lliures. Els dos llenguatges són:

- Analítics.
- Naturals.
- Lliures.
- Postcoordinats.
- D'estructura combinatòria.

Les **diferències** entre ells, i són diferències grans, rauen en el fet que:

- Els llistats de descriptors lliures es basen en la indexació humana i els llistats de paraules clau, en la indexació automàtica.
- Els llistats de descriptors lliures indexen conceptes i els llistats de paraules clau indexen majoritàriament unitermes.

3.1. La indexació automàtica

La **indexació automàtica** és el mètode pel qual un ordinador aplica un algoritme (o programa) al títol, resum o text complet del document per tal d'identificar-hi els termes que puguin representar la matèria i ser usats com a termes d'indexació i recuperació en un índex o llista.

Hi ha dos **tipus** d'indexació automàtica:

a) Totalment automatitzada (en anglès, *automatic indexing*).

b) Semiautomàtica (en anglès, *machine-aided indexing*): el programa selecciona possibles descriptors procedents d'un tesaurus o una llista controlada i un documentalista accepta o denega la proposta.

Els factors que fan possible la indexació automàtica són l'elevada inversió de temps que comporta la indexació intel·lectual, l'augment de la documentació electrònica i a text complet, l'extensió de sistemes de gestió documental a les institucions i empreses, i l'evolució del processament del llenguatge natural (PLN[3]) (Méndez i Moreiro, 1999).

(3)PLN és la sigla de *processament del llenguatge natural.* A partir d'ara denotem només PLN.

Lectura complementària

Per ampliar la vostra informació sobre la indexació automàtica, podeu llegir l'obra següent:

E. Méndez; J. A. Moreiro (1999). "Lenguaje natural e Indización automatizada". *Ciencias de la Información* (pàg. 11-24).

Tot i que podem trobar aplicacions dedicades exclusivament al resum i la indexació automàtica, el més habitual és que aquests programes formin part del mateix programa de gestió documental del centre.

3.1.1. Ciències implicades en la indexació automàtica

En la indexació automàtica intervenen diverses disciplines científiques, la informàtica, la lingüística i l'estadística:

1) La **informàtica** proporciona l'algoritme que representa, processa, emmagatzema i recupera les paraules clau.

2) La **lingüística** assigna etiquetes de tipus morfològic i fa anàlisis sintàctiques. Els programes d'indexació automàtica més evolucionats incorporen també anàlisis de tipus semàntic:

- **Morfologia**: treballa a nivell de paraula. Assigna la categoria gramatical de les paraules i l'arrel comuna (o lema). Recordem que les categories gramaticals són substantiu, verb, adjectiu, adverbi, etc., i que morfològicament poden ser de diversos gèneres, nombres, temps i mode. La lematització permet comptar totes les vegades que un concepte surt en el text encara que s'expressi amb formes diferents.

Exemple de desambiguació

Per exemple, *columna* en un text d'arquitectura o *columna* en un text de periodisme.

- **Sintaxi**: el seu camp de treball és la frase. Identifica els components de la frase: subjecte, verb, objectes i la seva concordança. Gràcies a aquesta anàlisi pot desambiguar gramaticalment els termes no resolts en l'etapa morfològica.

- **Semàntica**: s'ocupa del text. Desambigua termes polisèmics, ja que el text permet optar per un significat o l'altre.

3) L'**estadística** té diverses aplicacions en documentació, però en indexació automàtica intervé especialment en el càlcul de la freqüència relativa de les paraules en un text. Aquesta tècnica calcula quantes vegades surt una paraula

en un text i en funció d'uns barems (ni les paraules més repetides, ni les menys) se seleccionen les paraules clau del document. Els teòrics d'aquest mètode són G. K. Zipf (1949), H. P. Luhn (1957), Sparck Jones (1972) i G. Salton (1989).

Tractarem amb més detall d'aquests processos en el pròxim subapartat.

3.1.2. Funcionament i evolució dels programes d'indexació automàtica

Els primers estudis en indexació automàtica són de les dècades de 1950-1960. Es basaven en principis estadístics i probabilístics. En la dècada de 1970, G. Salton incorpora el model de valor de discriminació i de rellevància dels termes. En la dècada de 1980 apareixen els criteris lingüístics i en la de 1990 es planteja la indexació automàtica d'informació multimèdia (imatges i so). Des de la dècada de 1980 fins a l'actualitat, els programes d'indexació automàtica combinen el càlcul estadístic amb l'etiquetatge morfològic i l'anàlisi sintàctica.

Evolució dels programes d'indexació automàtica

1950-1960	1970	1980	1980-actualitat
Càlculs estadístics i probabilístics	Rellevància dels termes	Anàlisi lingüística	Model mixt: estadístic + lingüístic

Els programes informàtics d'indexació automàtica busquen emular el procés mental humà. La qüestió central que ens interessa és "com poden detectar les paraules amb més càrrega semàntica, més significat del text?". Amb aquest objectiu, segueixen els punts següents:

1) Lectura dels documents

El primer pas és llegir el text. Per a fer-ho cal que el document es trobi en format electrònic. Actualment, molts documents es troben en suport electrònic. També són fàcilment accessibles els títols, resums, sumaris o bé el text complet en les bases de dades de publicacions periòdiques i catàlegs bibliogràfics. Si es tracta de documentació en suport paper, de primer cal escanejar-la i aplicar un programa de tipus OCR (reconeixement òptic de caràcters) per a transformar la imatge de la pàgina escanejada en text electrònic.

Programes OCR

Aquests programes OCR estan implementats en la majoria d'escàners domèstics. Recomanem fer la prova d'escanejar una pàgina i convertir-la en text.

Com pot un programa detectar una paraula? El programa detecta la paraula perquè la considera una cadena de caràcters entre espai en blanc i espai en blanc.

2) Disseny

L'equip de documentalistes i programadors ha de decidir:

- Les parts del document que seran indexables. Són obligatoris el títol i el resum i desitjable el text sencer, sobretot en articles.

- El tractament que donaran a xifres, signes de puntuació, guions, majúscules/minúscules i accents. Habitualment són caràcters que no aporten significat, però en determinats contextos poden ser determinants.

Exemple

Hi ha vegades que aquests caràcters sí que tenen importància:

- Xifres: TV3, 1492.
- Punts, guions, signes: www.uoc.edu, Canal+, e-mail.
- Accents (útils per a diferenciar diacrítics): deu/Déu en català, te/té en castellà.

Un cop tenim el document en format electrònic i el programa n'ha llegit les parts seleccionades, el pas següent consisteix a destriar les paraules que tenen significat de les que no en tenen, i entre les que en tenen, seleccionar les de més rellevància en la indexació i recuperació posterior. Aquest pas es resol aplicant diversos mètodes estadístics i lingüístics, o bé en un ordre seqüencial, o bé alternant els mètodes.

3) Paraules amb contingut i paraules buides

En llenguatge natural hi ha paraules que tenen molt de significat (substantius, adjectius, etc.) i paraules que no en tenen tant (articles, conjuncions, etc.). Aquestes darreres es coneixen com a *paraules buides*.

Les **paraules buides** són paraules sense significat, com els articles, els pronoms, les preposicions, les conjuncions, els adverbis, etc., que són filtrades abans o després del processament del text.

Les paraules buides són molt freqüents però aporten poc valor de contingut semàntic al text i també poca ajuda en la recuperació, ja que l'usuari no les utilitza en la cerca i tampoc no són recollides en el fitxer invers de la base de dades. Es coneixen com a *llistes de paraules buides* en català, *listas de detención* en castellà i *stop word list* en anglès (Hans Peter Luhn va ser el creador del terme i del concepte de *stop word*).

Es fa una llista d'aquestes paraules buides i s'introdueixen en el programa, que va llegint el text paraula per paraula i les contrasta amb el fitxer de paraules buides. Si les troba a la llista, vol dir que són paraules buides i no les indexa. Si no apareixen en la llista, vol dir que tenen significat.

Paraules buides

Articles	el, la, un, una
Pronoms	meu, nostre...
Verbs	menjar, ésser, riure... (en totes les formes verbals)
Adverbis	tranquil·lament, eficaçment...
Preposicions	a, amb, de, sense...

Numerals	quart, vuitè...
Adjectius	alt, baix, gran...
Conjuncions	i, doncs, però, perquè...

Hi ha sistemes en què la llista de paraules buides:

a) Està **predeterminada**: el sistema disposa, de bon començament, de la llista de paraules buides del seu idioma/idiomes. De fet, la seva elaboració és fàcil, ja que només cal afegir-hi les categories buides d'una base de dades de terminologia en l'idioma que es vulgui. Els articles sempre són els mateixos, les conjuncions també, fins i tot els verbs es poden arribar a comptabilitzar i flexionar en tots els temps verbals.

b) S'**evita expressament** per a possibilitar al sistema la cerca per frases i sintagmes. Els sistemes que els eviten disposen d'altres eines per a reduir significativament el nombre de paraules indexades, com tècniques de *stemming* o lematització que veurem més endavant.

Exemple

Un concepte com el diari *El País*, en què l'article té un paper important.

c) Està **contextualitzada** (*stop word context-dependent*). Cada sistema elabora la llista de paraules buides segons el seu àmbit temàtic. Contextualitzar la llista permet evitar dos inconvenients greus:

- Paraules amb significat que esdevenen buides.

Exemple

Si un usuari cerca al catàleg d'una biblioteca especialitzada en astronomia, no buscarà pel concepte "Astronomia", ja que tota la col·lecció de la biblioteca farà referència a aquest concepte. El concepte *Astronomia* esdevé buit en aquest context.

- Paraules buides que esdevenen importants en la indexació.

Exemple

En un text d'història els nombres (1914, 1936...), numerals (Jaume I) i els adjectius poden tenir molta càrrega significativa (alta i baixa edat mitjana, sobrenoms adjectivats de personatges com Jaume I el Conqueridor i esdeveniments com guerra freda). En l'exemple següent, les paraules subratllades són buides:

Paraules castellanes que són buides o no depenen de l'àmbit temàtic

Paraula	En el text	Economia	Dret	Educació	Multidisciplinari
ESO	"Eso es lo importante..." "La reforma de la ESO"	ESO	ESO	ESO	ESO
Cabo	"Se precisa llevar a cabo..." "La composición rocosa del cabo..."	Cabo	Cabo	Cabo	Cabo (geografia)
Tuya	"Yo traeré la mía pero tú trae la tuya." "La tuya es un árbol perenne."	Tuya	Tuya	Tuya	Tuya (árbol)

Font: extracte d'I. Gil Leiva (2008). *Manual de indización. Teoría y pràctica.* Gijón: Trea (pàg. 333).

En el cas de la contextualització, les paraules buides poden ser introduïdes manualment o bé ser el resultat d'algun còmput estadístic de freqüència.

Exemple

Suposem el text "Osiris és el primer de tots els humans i déus que viu una segona vegada".

Descomponem el text com ho faria un programa d'indexació automàtica. Partim del supòsit que el programa té una llista de paraules buides predeterminada que inclou preposicions, conjuncions, numerals, verbs i articles.

Text	Comprovació a la llista de paraules buides		Resultat de la indexació
Osiris	No hi és	És un nom propi	Osiris
és	És una paraula buida	És una forma verbal	–
el	És una paraula buida	És un article	–
primer	És una paraula buida	És un numeral	–
de	És una paraula buida	És una preposició	–
tots	És una paraula buida		–
els	És una paraula buida	És un article	–
humans	No hi és	És un substantiu	humans
i	És una paraula buida	És una conjunció	–
déus	No hi és	És un substantiu	Déus
que	És una paraula buida	És una conjunció	–
viu	És una paraula buida	És una forma verbal	–
una	És una paraula buida	És un article	–
segona	És una paraula buida	És un numeral	–
vegada	No hi és	És un substantiu	vegada

El text queda indexat amb aquestes paraules clau:

```
Osiris
humans
déus
vegada
```

Són molt nombroses les paraules buides en un text? La presència de paraules buides no és gens menyspreable. Els estudis estadístics calculen que el 50% de les paraules d'un text són buides. Així, un article de 8.000 paraules en té unes 4.000 de buides i 4.000 amb significat. En l'exemple d'Osiris les quinze paraules de la frase han quedat reduïdes a quatre paraules clau (en aquest cas, l'exemple buscava pretesament moltes paraules buides per a il·lustrar aquest subapartat).

El procediment en PLN per a destriar les paraules buides de les de significat consisteix en l'aplicació d'una llista prèvia de paraules buides o d'una combinació de mètodes estadístics (freqüència) i lingüístics (anàlisi morfològica,

sintàctica i semàntica). Des que H. P. Hans Meter Luhn el 1957 va eliminar les paraules buides, aquesta tècnica ha estat una constant en els posteriors programes d'indexació automàtica.

4) Mètodes estadístics

Els mètodes estadístics han estat la primera aproximació a la indexació automàtica i encara avui en dia en són una part consubstancial. La teoria de fons és el càlcul del pes (ponderació) de les paraules: ni les paraules més repetides (per buides) ni les menys repetides (per específiques) són adequades per a ser seleccionades.

Els mètodes estadístics aplicats en PLN són de tres tipus: freqüència, freqüència inversa i discriminació; es poden usar sols o en combinació:

a) Freqüència d'aparició (llei de Zipf): George Kingsley Zipf (1949) psicòleg i psicolingüista nord-americà va proposar que el nombre de vegades que un terme surt en un text (freqüència) i la posició que ocupa en una llista de paraules més o menys freqüents (rang) és constant.

Per a Zipf els escriptors usen un ventall reduït de paraules, preferentment curtes, i la seva longitud és inversament proporcional a la freqüència.

Exemple

Al Web hi ha disponible una còpia de les paraules franceses i llur freqüència. En aquesta llista veiem que paraules d'ús quotidià tenen una freqüència més alta i paraules menys quotidianes a la inversa.

Freqüència d'aparició de les paraules franceses

aliments	13,97		grand	674,68
aller	234,58		laisse	120,97
ami	95,16		les	16.011,00
avec	3.019,71		neutre	11,52
collège	24,42		permutable	0,03
comité	73,71		pour	5.332,48
être	1.986,23		translateur	0,06
goût	98,77		vitamines	3,84

Font: extracte tret de http://www.lexique.org/listes/liste_mots.txt

Hans Meter Luhn (1957) aplica la llei de Zipf al camp de la indexació automàtica. Luhn veu que la freqüència amb què es repeteix una paraula pot ser un bon indicador per a seleccionar les paraules clau amb més significat. Les que tenen una freqüència molt alta no acostumen a tenir càrrega significativa (en

l'exemple anterior serien *avec, les, pour*, però recordem que en funció del context també ho podria ser *aliments*) i fan que es recuperin molts documents. Les que tenen una freqüència baixa solen ser paraules molt específiques en el text i, per tant, amb càrrega significativa, però en la recuperació presenten pocs resultats. Per a Luhn el millor per a indexar eren els termes amb una freqüència mitjana.

Luhn proposa els passos següents:

- Calcular la freqüència de totes les paraules del text o col·lecció.
- Classificar-les en ordre decreixent.
- Eliminar-ne les de freqüència més alta.
- Eliminar-ne les de freqüència més baixa.
- Indexar amb la resta de paraules.

Luhn aplica aquesta tècnica per a eliminar les paraules buides. A partir d'ell, tots els sistemes utilitzen aquesta tècnica per a crear la llista de paraules buides.

b) Freqüència inversa: Sparck Jones (1972) va posar de manifest la capacitat de discriminació d'un terme enfront d'un altre. Aquesta discriminació ha de ser vista en el conjunt de la col·lecció, no en un sol document. Cal comparar les paraules clau entre els documents del fons per a detectar quines són realment discriminatòries.

Exemple

Imaginem un fons de quatre documents:

- Document 1: "Visiteu el Museu Picasso a Barcelona al barri del Born".
- Document 2: "A Figueres podeu visitar el Museu dels joguets".
- Document 3: "El museu de Figueres i el Museu Picasso són els més famosos".
- Document 4: "A Figueres no deixeu de visitar el Museu Dalí".

Lectura recomanada

Per a la resolució del càlcul vegeu:

Francisco Javier Martínez Méndez. *Recuperación de información en la red*.

Eliminem les paraules buides *el*, *els*, *a*, *al*, *del*, *dels*, *més*, *no* i totes les formes del verb *visitar* (*visiteu*, *visitar*). Ens queda:

Doc.	Museu	Picasso	Barcelona	Barri	Born	Figueres	Joguets	Famosos	Dalí
1	1	1	1	1	1	0	0	0	0
2	1	0	0	0	0	1	1	0	0
3	2	1	0	0	0	1	0	1	0
4	1	0	0	0	0	1	0	0	1

Seguint l'exemple, la paraula *museu* que apareix a cada document no té gaire utilitat per a discriminar el document 1 de la resta; en canvi Picasso, Barcelona i totes les paraules que apareixen poc serien discriminatòries. Per a mesurar aquest valor de discriminació es proposa la *freqüència inversa*. El pes d'una paraula augmenta si apareix poques vegades en un document i disminueix si apareix sovint en la resta de documents.

Pesos de cada paraula

Museu	0		Born	0,602
Picasso	0,301		Joguina	0,602
Figueres	0,301		Famosos	0,602
Barcelona	0,602		Dalí	0,602
Barri	0,602			

c) Valor de discriminació del terme: G. Salton (1989), a partir de la idea que els vocables d'un text es classifiquen segons la seva capacitat per a discriminar uns documents dels altres en una col·lecció, va idear un sistema d'indexació, conegut com el **model de valor de discriminació**, que atribueix el pes o valor més alt a aquells termes que causen la màxima separació possible entre els documents d'una col·lecció. És a dir, el valor d'un terme depèn de com varia la separació mitjana entre els documents quan a un terme es fixa una identificació de contingut. Per tant, les millors paraules són aquelles que aconsegueixen la distància més gran. L'anàlisi del **valor de discriminació** consigna una funció específica en l'anàlisi de contingut a les paraules simples, a les juxtaposades, a les frases i a grups de paraules.

El valor de discriminació d'un terme es defineix com la mesura dels canvis en la separació espacial, que es manifesta quan una paraula qualsevol és assignada a una col·lecció com a terme d'indexació per a representar millor les diferències que hi pugui haver entre els documents. Precisament l'assignació fa reduir la densitat espacial, i per contra, un discriminador pobre incrementa la densitat espacial. D'aquesta manera, si primer es calculen les densitats espacials i s'atribueixen a cada terme, és possible especificar els termes en ordre decreixent segons els seus valors de discriminació (Gil i Rodríguez, 1996).

Lectura complementària

Si voleu ampliar la vostra informació sobre el sistema d'indexació ideat per Salton, podeu llegir l'obra següent:

I. Gil Leiva; J. V, Rodríguez Muñoz (1996). "Tendencias en los sistemas de indización automática. Estudio evolutivo". *Revista Española de Documentación Científica* (vol. 19, núm. 3, pàg. 273-291).

5) Mètodes lingüístics

Els primers analitzadors lingüístics són de la dècada de 1960-1970. La seva aportació a l'anàlisi del contingut és cabdal, ja que permeten analitzar el text en tres nivells de profunditat: paraula, frase i text.

Cadascun d'aquests nivells és analitzat per mòduls del programa basats en diferents disciplines:

- Paraula: morfologia.
- Paraula dins la frase: sintaxi.
- Paraula dins el text. semàntica.

a) Anàlisi morfològica: és la branca de la gramàtica tradicional que estudia la forma de les paraules, independentment de les seves relacions o funcions dins la frase. La forma de les paraules es distribueix en diverses categories anomenades *parts de l'oració*: noms, adjectius, verbs, etc. També estudia les variacions que poden experimentar per motius dels accidents gramaticals com són el gènere, nombre, declinació i conjugació.

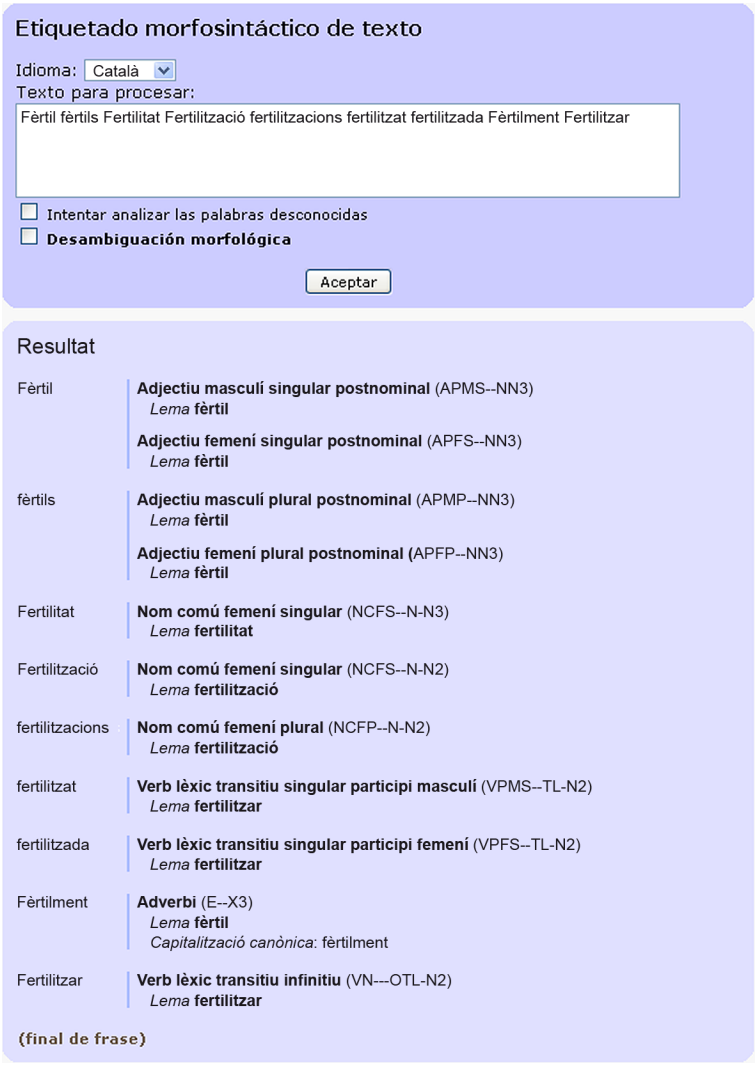
L'anàlisi morfològica en indexació automàtica es basa en regles gramaticals i diccionaris de sufixos i afixos.

La seva intervenció en la indexació automàtica és molt rellevant en la formació de la llista de paraules buides i una primera agrupació de les paraules, ja que:

- Assigna la categoria gramatical de cadascuna de les paraules del text (nom comú, verb, adjectiu, adverbi, article, etc.), segons quin en sigui el gènere, nombre, temps i mode. Això redueix el ventall de paraules susceptibles de contenir informació, ja que en detecta les buides i les flexions derivades dels accidents gramaticals (gènere, nombre, etc.). A la Xarxa hi ha aplicacions gratuïtes que analitzen morfològicament un text. Els exemples més comuns són *Daedalus* en castellà o *Thera* en català i castellà.

Exemple

Analitzarem un text sobre fertilitat amb Daedalus:



- Assigna l'arrel comuna de diverses paraules. La mateixa anàlisi morfològica procedeix la lematització (*stemming*, en anglès), que consisteix a localitzar l'arrel comuna d'un grup de paraules (fèrtil, fertilitat, fertilitzar, fertilització). La lematització permet comptar totes les vegades que un concepte surt en el text, encara que s'expressi amb formes diferents.

Exemple

En la recuperació, quan l'usuari cerqui per *fèrtil* recuperarà tots els altres conceptes, ja que es troben indexats conjuntament.

Sense lematitzador	Amb lematitzador
Fèrtil – 1 resultat Fèrtils – 1 resultat Fèrtilment – 1 resultat	Fèrtil – 3 resultats

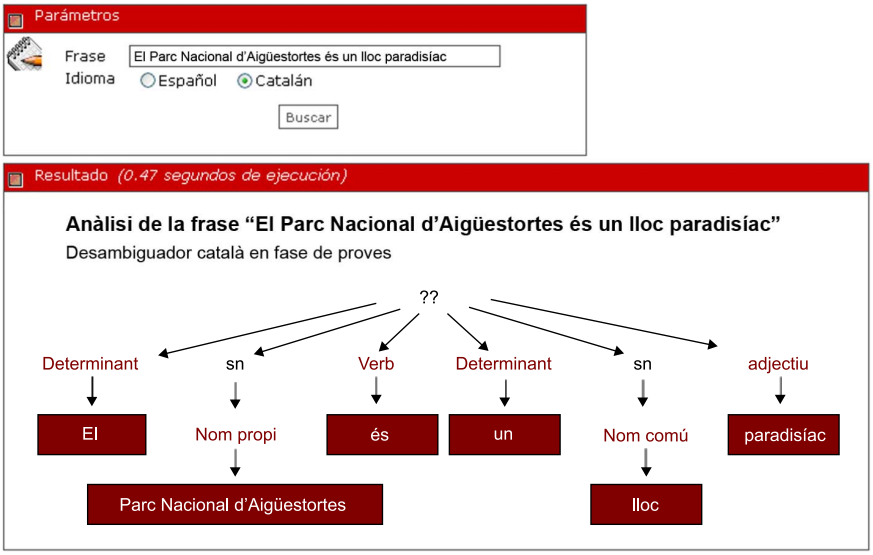
Gràcies a aquesta anàlisi, els càlculs estadístics es poden fer sobre un conjunt ja seleccionat de termes.

b) Anàlisi sintàctica: estudia la funció o les associacions de les paraules dins la frase, la qual cosa s'anomena *estructura sintàctica*. Identifica els components de la frase: subjecte, predicat, complements o objectes i la seva concordança. El programa d'anàlisi sintàctica agafa la frase i intenta reconèixer l'estructura amb la qual s'ha construït dins un conjunt d'anàlisis possibles conegudes com a *gramàtica*.

Gràcies a aquesta anàlisi, pot desambiguar gramaticalment els termes no resolts en l'etapa morfològica. Els analitzadors sintàctics (*parsing*) es basen en analitzadors morfològics de l'etapa anterior, gramàtiques i diccionaris.

Exemple

Anàlisi sintàctica de la frase "El Parc Nacional d'Aigüestortes és un lloc paradisíac" amb Thera:



L'anàlisi sintàctica és útil per als polisèmics o alguna paraula que per la seva similitud podria ser considera buida pel programa.

Exemple

Analitzem la frase en castellà: "Esta nota para mañana[4]", en què morfològicament cada paraula té diverses opcions, però que l'anàlisi sintàctica pot aclarir. Per exemple, *nota* pot ser un nom comú femení (la nota escrita) o una forma verbal (3a. persona singular del present indicatiu o 2a. persona singular de l'imperatiu del verb *notar*), però dins la frase, per la seva posició, només pot ser considerada com a nom comú i, en conseqüència, seria una paraula que caldria indexar, no una de buida. Cal dir, però, que actualment no tots els analitzadors sintàctics poden arribar a aquestes prestacions i cometen errors.

[4]L'exemple està tret d'I. Gil (2008). *Manual de indización. Teoría y pràctica*. Gijón: Trea.

c) Anàlisi semàntica. La semàntica és la branca de la lingüística que estudia el significat de les paraules. La seva utilitat rau a desambiguar sinònims, polisèmics i vincular anàfores amb el terme al qual fan referència. Tot i així, hem de dir que aquesta part de l'anàlisi és la més complexa per als programaris actuals. Els analitzadors semàntics es basen en diccionaris lèxics, xarxes semàntiques, tesaurus i ontologies.

Les xarxes semàntiques

Aquestes xarxes són bases de dades que representen el coneixement d'una llengua o àmbit d'una manera estructurada.

Per a la indexació automàtica són molt interessants les bases de dades tipus WordNet, que donen informació lèxica i semàntica dels termes. Un algoritme vinculat a WordNet pot resoldre casos de sinonímia o hiperonímia, ja que la base de dades proporciona relacions bàsiques de significat.

WordNet i Euro WordNet són bases de dades multilingües en diversos idiomes (Euro WordNet comprèn l'holandès, l'italià, el castellà, l'alemany, el francès, el txec i l'estonià). Cada idioma dissenya la seva WordNet però manté la mateixa estructura que la inicial de Princeton quant als termes sinònims amb relacions semàntiques de relació d'equivalència entre ells. Els diferents WordNet estan connectats entre ells i proporcionen accés a una ontologia compartida. En accés gratuït només es troba el Wordnet de Princeton.

Els tesaurus participen en l'anàlisi semàntica aportant les relacions semàntiques d'equivalència, jerarquia i associació de cada paraula clau seleccionada. Les ontologies aporten, a més a més, el detall en cada tipus de relació. Un exemple és tota mena de relacions d'associació entre dos descriptors.

3.2. El llistat de paraules clau en la indexació

El programa d'indexació automàtica reconeix paraules, per a ser més exactes, cadenes de caràcters entre un espai i el següent. Per tant, reconeix paraules, no pas expressions formades per més d'una paraula. Diem, doncs, que el llistat de paraules clau indexa unitermes.

Uniterme

Recordem que els unitermes són l'element amb significat més petit d'un terme d'indexació. *Pluja* és un uniterme. *Pluja d'estels* podria ser un descriptor d'un tesaurus format per dos unitermes: *pluja* i *estels*.

Els **avantatges** que tenen les llistes de paraules clau en la indexació són els següents:

- Indexació immediata, en qüestió de segons.
- Coherència d'indexació exacta entre diversos SID. Dos SID amb el mateix programa indexador arribaran al mateix resultat.
- Indexació exhaustiva i tan específica com ho sigui el contingut del document indexat.
- Actualització ràpida de la base de dades.
- Riquesa terminològica extraordinària: evolució automàtica de la terminologia, paral·lela a la mateixa evolució del coneixement i de la ciència.
- Costos mínims de construcció i manteniment.

Un dels **inconvenients** que tenen les llistes de paraules clau en la indexació és que les paraules amb significat s'indexaran tal com surtin en el text. Per tant, pot indexar:

- El mateix terme en diferents idiomes, per exemple: *fonts d'informació*, *fuentes de información*, *information resources*, en funció de la llengua del text. I no té relacions d'equivalència entre els termes, com tenen els tesaurus multilingües.
- Les errades ortogràfiques (*tesauris*, *thesaurus*).

- Singular, plural, masculí, femení... (encapçalament/encapçalaments, llista/llistat...).
- Sigles i conceptes sense distinció (SID/servei d'informació i documentació; UE/Unió Europea).
- Sinònims (llista de paraules buides/*stop word list*).
- Homònims (*Hierro*, i no sabrem si fa referència a l'illa o l'element químic).

Com que no hi ha control sobre el vocabulari, no controla l'ambigüitat del llenguatge natural.

3.3. El llistat de paraules clau en la recuperació

La indexació automàtica sustenta els cercadors; en conseqüència, els usuaris de Google i Yahoo coneixen d'una manera implícita els avantatges i els inconvenients de la recuperació de documents amb aquest llenguatge. Tal com diu Isidoro Gil (2008, pàg. 108), l'ús d'Internet està convertint cada usuari en un *paradocumentalista* en potència. Els usuaris són conscients que cal escollir amb cura les paraules clau, que siguin com més específiques millor, posar el text entre cometes, restringir la cerca amb operadors booleans, especificar dates, anar afegint termes, etc., si volen evitar el soroll documental.

Els **avantatges** que tenen les llistes de paraules clau en la recuperació són els següents:

- Les cerques poden ser molt precises (tots els termes significatius hi són presents).
- Molt fàcil d'usar per l'usuari.
- És el llenguatge més actualitzat, evoluciona automàticament al mateix temps que la terminologia del document. Hi podem buscar *iPhone*, *telèfons 3G*, *iPod*, *spam*, *spyware*, etc., termes que no constarien en cap altre llenguatge documental, atès que el procés d'actualització és lent.
- La immediatesa: el document arriba i, en qüestió de segons, està indexat i disponible per a la cerca.

Quant als **inconvenients**, la indexació automàtica es caracteritza per la possibilitat d'indexar, si es vol, el text sencer del document, la qual cosa genera molts punts d'accés per matèria, moltes paraules clau. Quan fem una cerca, el cercador ens pot tornar milers de registres i és un problema molest (soroll documental). Ara bé, aquesta característica també minimitza problemes potencialment perillosos (el silenci documental), com ara la pèrdua de conceptes i la sinonímia.

Els principals inconvenients són els següents:

- Pèrdua de conceptes. Com que el programa indexa unitermes, els conceptes expressats d'una manera composta se separen en dues unitats i perden el sentit.

Exemple de pèrdua de conceptes

Universitat de Barcelona, per exemple, se separa en universitat i Barcelona. Separar conceptes pot donar lloc a coordinacions falses en la recuperació com en aquest cas, en què recuperem totes les universitats de la ciutat de Barcelona: UB (Universitat de Barcelona), UPF (Universitat Pompeu Fabra), URL (Universitat Ramon Llull), etc.

- No reconeix la sinonímia. El programa no pot reconèixer els sinònims del mateix concepte. Si en el text apareixen tots aquests sinònims, el programa no detectarà cap semblança entre ells i els indexarà tots: *captivar*, *encisar*, *embruixar*, *hipnotitzar*, *suggestionar*.

- No reconeix l'anàfora i l'el·lipsi.

- No permet la univocitat.

Exemple d'estratègia de cerca

Per a buscar documents sobre la Unió Europea, per exemple, caldria pensar en tots els seus sinònims i sigles com ara *UE* / *Unió Europea* / *European Union* / *CEE* / *Comunitat Econòmica Europea*.

- Soroll i silenci documentals produïts per l'ambigüitat pròpia del llenguatge natural (fenòmens de sinonímia, polisèmia i homonímia).

- No apareix cap mena de control sobre variants ortogràfiques, sobre plurals, adjectius, verbs, sigles, etc., ni es fa cap control sobre els errors ortogràfics o d'impressió que poden aparèixer en els documents.

- Complexitat en el moment de plantejar l'estratègia de cerca; cal una agrupació de sinònims completa, ús de truncaments i d'operadors de proximitat.

Precisament per a minimitzar els inconvenients en la recuperació, els SID acostumen a complementar-ho amb altres llenguatges documentals preferentment controlats, com un tesaurus o una llista d'encapçalaments de matèria.

Actualment la majoria de catàlegs en línia de biblioteques universitàries ens permeten fer cerques per diversos camps, dels quals destaquem especialment el de *paraula clau de matèria*. Aquesta opció descompon el text (fins i tot l'encapçalament construït amb una llista d'encapçalament de matèria, és a dir, amb un llenguatge controlat i humà) en una successió de paraules clau, és a dir, en un llenguatge lliure i automàtic.

Activitats

Llistat de descriptors lliures

1. Indexeu amb un llistat de descriptors lliures el resum següent amb els tres graus d'exhaustivitat.

López Alonso C.; Séré A. (ed.) (2003). *Nuevos géneros discursivos: los textos electrónicos*. Madrid: Biblioteca Nueva.

Esta obra describe la transformación de determinados discursos sociales ante la revolución de las nuevas tecnologías de la información y de la comunicación (TIC). En efecto, Internet forma parte de nuestra vida diaria y ya no es fácil prescindir de este medio tecnológico. Esta obra, que consta de tres partes, analiza estos nuevos géneros discursivos:
En la primera parte, se describen los distintos textos electrónicos –el correo, los chats, foros– y otros escritos en la red –prensa, newsletters, E-zinc, etc.
En la segunda, se introduce al lector en la génesis y evolución de los lenguajes informáticos y sus diferentes tipos, especialmente los lenguajes de marcado y los modelos hipermedia.
La tercera, finalmente, se centra en los nuevos productos hipertextuales en el campo del discurso de transmisión de conocimientos con muy variadas utilizaciones: la enseñanza a distancia o textos de consulta como los diccionarios.

2. Indexeu el resum anterior per a dues bases de dades diferents: l'una genèrica i l'altra especialitzada en documentació.

3. Analitzeu detingudament l'exemple dels tres analistes que indexen el document sobre els càtars i digueu per quins motius són tan diversos els resultats.

Analista a	Analista b	Analista c
Baixa edat mitjana	Catarisme	Albigès
Càtars	Heretgies	Cismes
França	Ordes monàstics	Història medieval
Moviments cristians dissidents	Religió	Llenguadoc

4. Elaboreu una llista amb les operacions que cal dur a terme sobre el lèxic per a transformar una llistat de descriptors lliures en un llenguatge controlat no codificat.

Llistes de paraules clau

1. Indexeu el text següent simulant el funcionament d'una llista de paraules clau:

Piero De Benedetto Dei Franceschi (Sansepolcro, Toscana, 1416-1492) va ser un pintor del quattrocento italià conegut pel seu àlies *Piero della Francesca*. També va ser geòmetra i matemàtic.

És un dels personatges principals i fonamentals del Renaixement. Piero della Francesca és un personatge itinerant: és una figura que trobarem a Ferrara entre el 1447 i el 1448 treballant amb Lionello d'Este; a Rimini per a Sigismundo Malatesta entorn el 1450; a Roma per al papa Pius II entre el 1458 i 1459, i a Urbino per a Federico de Montefeltro en diverses ocasions. La seva actitud itinerant se sol comparar amb la de Leon Batista Alberti. L'autor neix a Borgo San Sepolcro (nord de Florència, a la zona de l'Úmbria) l'any 1416. L'autor prové d'una família de mercaders i, a causa d'això, sabia aritmètica, càlcul, àlgebra, geometria i comptar amb l'àbac. Després de la formació que va tenir per a portar el negoci familiar es va formar amb un mestre local que s'anomenava Antonio di Anghiari per a ser pintor. En aquells moments, hi havia molts estils: encara era vigent el linealisme i el lirisme de Fra Angelico, Panozzo Gonzolo o Filippo Lipi i, d'altra banda, hi havia el realisme geomètric de Paolo Uccelo.

Té un estil pictòric molt particular i, per tant, és fàcil d'identificar. Fa servir una llum molt diürna i uns colors molt brillants, que li provenen del seu contacte amb Domenico de Venezziano. En la seva obra sempre és present el pes de la geometria, que implica, d'una banda, utilitzar la perspectiva lineal i, de l'altra, reduir les figures a l'essència. Les figures de Piero della Francesca són molt estàtiques, poc nervioses, cosa que va al revés que en la resta de pintures renaixentistes de Florència, que a mesura que avancem en el temps cada cop són més dinàmiques. També veiem que les seves figures són poc expressives i monolítiques. Longhi, quan parla de Piero, diu que les seves figures són "columnes".

"Piero della Francesca", *Viquipèdia*

a) Dissenyeu les opcions del programa que simulareu (parts del text, signes i símbols).

b) Quantes paraules buides hi heu localitzat?
c) Quins inconvenients plantejarà en la recuperació?

Glossari

anàlisi morfològica *f* Branca de la gramàtica tradicional que estudia la forma de les paraules, independentment de les seves relacions o funcions dins la frase. La forma de les paraules es distribueix en diverses categories anomenades *parts de l'oració*: noms, adjectius, verbs, etc. També estudia les variacions que poden experimentar per motius dels accidents gramaticals com són el gènere, nombre, declinació i conjugació.

anàlisi semàntica *f* Branca de la lingüística que estudia el significat de les paraules. La seva utilitat rau a desambiguar sinònims, polisèmics i vincular anàfores amb el terme al qual fan referència.

anàlisi sintàctica *f* Estudi de la funció o les associacions de les paraules dins la frase, cosa que s'anomena *estructura sintàctica*. Identifica els components de la frase: subjecte, predicat, complements o objectes i la seva concordança. El programa d'anàlisi sintàctica agafa la frase i intenta fer evident l'estructura amb la qual s'ha construït dins un conjunt d'anàlisis possibles conegudes com a *gramàtica*.

automatic indexing *m* Vegeu **indexació automàtica**.

classificació social *f* Vegeu **indexació social**.

coordinació falsa *f* Recuperació inesperada (i errònia) que malgrat contenir els elements de la cerca corresponen a temes diferents.

descriptor lliure *m* Terme d'indexació propi del llenguatge documental. Llista de descriptors lliures.

etiquetatge col·laboratiu *m* Vegeu **indexació social**.

etiquetatge social *m* Vegeu **indexació social**.

folksonomia *f* Neologisme (Thomas van der Wal) resultat de la fusió de *folk* ('gent', 'popular') i *taxonomia* ('gestió de la classificació'), cosa que dóna com a resultat una "indexació gestionada popularment".
sin. **indexació social**.

freqüència d'aparició *f* Llei de Zipf. George Kingsley Zipf (1949) psicòleg i psicolingüista nord-americà va proposar que el nombre de vegades que un terme surt en un text (freqüència) i la posició que ocupa en una llista de paraules més o menys freqüents (rang) és constant. Per a Zipf els escriptors usen un ventall reduït de paraules, preferentment curtes i la seva longitud és inversament proporcional a la freqüència.

freqüència inversa *f* Sparck Jones (1972) va posar de manifest la capacitat de discriminació d'un terme enfront d'un altre. Aquesta discriminació ha de ser vista en el conjunt de la col·lecció, no en un sol document. Cal comparar les paraules clau entre els documents del fons per a detectar quines són realment discriminatòries.

indexació automàtica *f* Mètode pel qual un ordinador aplica un algoritme (o programa) al títol, resum o text complet del document per tal d'identificar els termes que puguin representar la matèria i ser usats com a termes d'indexació i recuperació en un índex o llista.
en *automatic indexing*.

indexació social *f* Tipus d'indexació en xarxa en què els internautes assignen etiquetes amb descriptors als recursos web. L'assignació de les etiquetes es fa sense ànim de lucre, no es busca un guany econòmic, sinó beneficiar-se de cerques millors. Els descriptors escollits no se supervisen ni tenen cap estructura semàntica.
sin. **classificació social, etiquetatge col·laboratiu, etiquetatge social, folksonomia**
en *tagging*.

lematització *f* Localització de l'arrel comuna d'un grup de paraules.
en *stemming*.

llista de detenció *f* Vegeu **paraula buida**.

llista de descriptors lliures *f* Llenguatge documental. Vocabulari monolingüe de termes d'indexació ordenats alfabèticament. Aquests termes són escollits per l'analista, sense verificar si existeixen o com s'introdueixen en una llista prèviament establerta.

llista de paraules clau *f* Llenguatge documental. Vocabulari ordenat alfabèticament dels termes amb càrrega significativa extrets d'un document mitjançant un programa informàtic.

machine-aided indexing *m* Vegeu **mètode d'indexació semiautomàtica**.

mètode d'indexació semiautomàtica *m* El programa selecciona possibles descriptors procedents d'un tesaurus o una llista controlada i un documentalista accepta o denega la proposta.
en *machine-aided indexing*.

mètode estadístic *m* Primera aproximació a la indexació automàtica. Calcula el pes (ponderació) de les paraules: ni les paraules més repetides (per buides) ni les menys repetides (per específiques) són adequades per a ser seleccionades. Els mètodes estadístics aplicats són de tres tipus: freqüència, freqüència inversa i discriminació. Es poden usar sols o en combinació.

mètode lingüístic *m* Mètode que permet analitzar el text en tres nivells de profunditat: paraula (anàlisi morfològica), frase (anàlisi sintàctica) i text (anàlisi semàntica).

paraula buida *f* Paraula sense significat com, per exemple, articles, pronoms, preposicions, conjuncions, etc., que són filtrades abans o després del processament del text. Són paraules molt freqüents però que aporten poc valor de contingut semàntic al text i també poca ajuda en la recuperació, ja que l'usuari no les utilitza en la cerca i tampoc no són recollides en el fitxer invers de la base de dades.
sin. **llista de detenció**
en *stop word list*.

paraula clau *f* Terme d'indexació propi del llenguatge documental. Llista de paraules clau.

reconeixement òptic de caràcters *m* Programa que transforma la imatge de la pàgina escanejada en text electrònic. De l'expressió anglesa optical character recognition (OCR).

OCR *m* Vegeu **reconeixement òptic de caràcters**.

stemming *m* Vegeu **lematització**.

stop word list *f* Vegeu **paraula buida**.

tagging *m* Vegeu **indexació social**.

valor de discriminació del terme *m* Mètode estadístic aplicat a la indexació automàtica, desenvolupat per Jalton (1989). La mesura dels canvis en la separació espacial que es manifesta quan una paraula qualsevol és assignada a una col·lecció com a terme d'indexació per a representar millor les diferències que hi pugui haver entre els documents. L'assignació en disminueix la densitat espacial, i al contrari, un discriminador pobre incrementa la densitat de l'espai. D'aquesta manera, si es calculen primer les densitats espacials i s'atribueixen a cada terme, és possible especificar els termes en ordre decreixent pels seus valors de discriminació.

Bibliografia

Alonso Berrocal, J. L.; Figuerola, C. G.; Zazo Rodríguez, A. (2009). *Recuperación Avanzada de la Información*. Salamanca. [Data de consulta: 31-07-2009].

Cid, P.; Cuadrado, M.; Aguiriano, C. (1999). *Fonaments de llenguatges documentals* [document electrònic]. Barcelona: UOC.

Gil Leiva, I. (2008). *Manual de indización. Teoría y pràctica*. Gijón: Trea.

Gil Leiva, I.; Rodríguez Muñoz, J. V. (1996). "Tendencias en los sistemas de indización automática. Estudio evolutivo". *Revista Española de Documentación Científica* (vol. 19, núm. 3, pàg. 273-291).

Hassan Montero, Y. (2006). "Indización Social y Recuperación de Información". *No Solo Usabilidad* (núm. 5, ISSN 1886-8592).

Méndez, E.; Moreiro, J. A. (1999). "Lenguaje natural e Indización automatizada". *Ciencias de la Información* (pàg. 11-24).

Quintarelli, E. (2005). "Folksonomies: power to the people". ISKO Italy-UniMIB meeting: Milan. [Data de consulta: 31-07-2009].

Salton, G. (1988). *Automatic text processing : the transformation, analysis, and retrieval of information by computer*. Reading (Mass.) [etc.]: Addison-Wesley.

Slype, van G. (1991). *Los lenguajes de indización: concepción, construcción y utilización en los sistemas documentales*. Madrid: Pirámide (Fundación Germán Sánchez Ruipérez. Biblioteca del Libro).

Pàgines web interessants

Cibermetría

Cybermetrics. Recursos presentados por el departamento de CINDOC dentro del Consejo Superior de Investigaciones Científicas; artículos, seminarios, etc.

Blog spot. Acceso a algunos artículos sobre cibermetría.

Cybermetrics. Software disponible relativo a la cibermetría.

Webometrics and organizations. Enlaces hacia publicaciones y grupos de trabajo sobre cibermetría.

Applications of informetrics to information retrieval research. Artículo que relaciona la cibermetría con la recuperación de información.

The diameter of the world wide web. Artículo sobre la medición de la web realizada por los autores.

Inteligencia artificial

World Scientific: Connecting great minds. Vínculos a varias publicaciones periódicas sobre campos especializados como inteligencia artificial o sistemas neuronales.

AI international. Vínculos a entidades investigadoras sobre el tema como universidades, laboratorios, etc, así como vínculos a congresos que versen sobre el mismo como AAAI o PRICAI.

Altres pàgines web d'interès

The Open Archives Initiative.

GOOGLE-WATCH. Una página crítica sobre el sistema de recuperación de Google (el aparente uso comercial del orden de la recuperación).

RDF: Un modelo de metadatos flexible. Página sobre metadatos, con recursos web relacionados.

DUBLINCORE: Metadata Initiative (DCMI). Iniciativa de metadatos para paginas web, conversión de formatos... Proyectos, software, grupos de trabajo.

Pew Internet & american life. Web dependiente del Pew Research Center, con acceso a publicaciones de actualidad sobre Internet.

z39.50 International Standard Maintenance Agency. Página sobre el mantenimiento de este protocolo de búsqueda e intercambio de información.