Revista de la Societat Catalana de Química · 5 / 2004 · p. 78-81

Químic@ en pantalla

Pere Alemany
Departament de Química Física, Universitat de Barcelona, a/e: alemany@qf.ub.es

Un dels tòpics més repetits sobre Internet és el de l'anomenada *revolució de la informació*, segons el qual avui dia la humanitat té al seu abast una quantitat d'informació gairebé infinita, disponible de manera pràcticament immediata usant un mitjà relativament senzill per obtenir-la. Aquesta afirmació està, però, en clara contradicció amb l'exclamació de molta gent que després d'una cerca infructuosa acaba reconeixent que «tampoc n'hi ha per tant, jo no hi trobo mai res, a Internet!». Quin dels dos extrems és cert? Realment tenim tota la informació que necessitem a Internet i no som capaços de trobar-la, o ja ens podem esforçar tot el que vulguem, que n'hi ha molta que no és a Internet i a la qual hi podem tenir un accés molt més eficient visitant, per exemple, una biblioteca? En aquest article intentarem descriure breument algunes de les principals característiques de la informació publicada a Internet per tal de contestar preguntes com aquestes.

La primera qüestió que podem plantejar és evident: Quanta informació hi ha a Internet? La resposta és difícil, en primer lloc pel volum inmens d'informació que hi ha, i en segon lloc, per la seva naturalesa dinàmica, és a dir, per la rapidesa amb què canvia. Una pregunta amb una resposta *a priori* més senzilla és saber quants ordinadors hi ha connectats a la Xarxa. Per estimar aquest nombre podem usar estadístiques del nombre d'adreces IP o de noms de domini registrats, que ens donen una fita superior per al nombre d'ordinadors connectats (igual que no tots els cotxes matriculats circulen simultàniament, no tots els ordinadors registrats han d'estar connectats alhora a la Xarxa). Per trobar aquest tipus d'informació un bon punt d'entrada és la pàgina web de la Societat d'Internet (*http://www.isoc.org*). Aquesta societat és una associació no governamental i sense afany de lucre dedicada exclusivament al desenvolupament mundial d'Internet. A la seva pàgina principal hi ha un enllaç a la secció «All about the Internet», on trobareu enllaços que ens connecten amb empreses i/o organitzacions que es dediquen a elaborar estudis estadístics de l'activitat a la Xarxa. Un d'aquests enllaços ens porta a la pàgina del Consorci de Sistemes d'Internet (*http://www.isc.org*), una altra organització no governamental amb finalitats molt semblants a la Societat d'Internet, que publica un estudi periòdic sobre el nombre de noms de domini registrats des del 1993 (*http://www.isc.org/index.pl?/ops/ds*). Aquí podem llegir que, dels 1.313.000 noms registrats el gener de 1993, s'ha passat a 233.101.481 el gener d'enguany. Com hem esmentat abans, aquest nombre és, però, una estimació del nombre màxim d'ordinadors connectats. El nombre real és bastant menor. Així doncs, al mateix estudi hi veiem que el gener de 1999, intentant accedir als 43 milions d'ordinadors registrats, només se n'obtenia resposta d'aproximadament el 20 %, és a dir, entre vuit i nou milions d'ordinadors. Extrapolant aquest resultat a dades actuals podem calcular que actualment hi ha aproximadament 47 milions d'ordinadors en xarxa. El que és impossible de saber és, però, la quantitat d'informació que contenen aquests ordinadors, i quin percentatge d'aquesta informació és accessible.

De tota manera, amb l'ordre de magnitud de les xifres exposades anteriorment, és fàcil calcular que el nombre de pàgines web disponibles a Internet és, com a mínim, de milers de milions, una quantitat d'informació totalment indigerible si no es disposa d'eines eficients de classificació i localització d'aquesta informació. Si esteu interessats a conèixer en detall les opcions de cerca que teniu a la web i els trucs per optimitzar la recuperació d'informació, us pot ser interessant una visita a *http://www.searchenginewatch.com*, una pàgina web dedicada exclusivament a recopilar dades relacionades amb aquests temes. En concret, a l'hora de cercar informació a la web disposem de dos grans categories d'eines: els índexs temàtics o directoris i els motors de cerca o cercadors.

Els primers són pàgines web on es presenta una col·lecció d'enllaços catalogats per temes que s'han estructurat jeràrquicament. N'hi ha de generals, per exemple el de Yahoo! (*http://www.yahoo.com*), el d'Open Directory (*http://www.dmoz.org*) o el de LookSmart (*http://www.search.looksmart.com*) que contenen informació sobre diferents temes, o d'especialitzats, com per exemple Links for Chemists (*http://www.liv.ac.uk/Chemistry/Links/links.html*), que recullen només pàgines web dedicades a una temàtica concreta. La característica principal dels índexs és que, en haver estat confeccionats manualment, la informació a la qual ens permeten accedir és de força qualitat (evidentment això dependrà del criteri dels editors que hagin seleccionat i classificat les pàgines web que surten a l'índex) i són força útils com a punts d'entrada quan s'estan buscant dades de caire general (per exemple: pàgines que parlin de química atmosfèrica). Els inconvenients que tenen són, per una banda, que en estar elaborats manualment han de contenir, per força, una quantitat limitada d'informació i, per altra banda, que no són útils per

TAULA 1. Principals característiques (en la data indicada a la darrera columna) dels tres directoris generals més populars. A la columna «Enllaços» s'indiquen el nombre de milions d'enllaços que conté cada directori

Directori	Editors	Categories	Enllaços	Data
Open Directory	36.000	361.000	2,6	Abril 2001
LookSmart	200	200.000	2,5	Juliol 2001
Yahoo!	100	No disponible	1,8	Juliol 2000

buscar dades específiques (per exemple els paràmetres de cel·la del cristall de clorur sòdic), ja que el nombre de categories en què s'estructura l'índex sol limitar-se a uns quatre o cinc nivells jeràrquics com a màxim i això obliga a una classificació poc detallada de la informació. Per fer-vos una idea de la quantitat de dades que podeu trobar en un directori, a la taula 1 s'especifiquen les característiques dels tres directoris generals més populars.

A banda de l'esmentat Links for Chemists, hi ha d'altres índexs temàtics adreçats específicament als químics. Entre aquests, els més exhaustius són Chemdex (*http://www.chemdex.org*), Information Retrieval in Chemistry (macedonia.nrcps.ariadne-t.gr), Chemie.de (*http://www.chemie.de*) i ChemIndustry (*http://www.chemindustry.com/index.asp*).

En cas de voler cercar més exhaustivament la informació continguda a Internet o de voler obtenir informació sobre un tema específic, haurem de recórrer als motors de cerca. Un motor de cerca és una aplicació informàtica que ens permet buscar informació en una base de dades, de pàgines web en el nostre cas. Els més populars en l'actualitat són Google (*http://www.google.com*), amb uns 250 milions de cerques diàries al febrer de 2003, AlltheWeb.com (*http://www.alltheweb.com*), Yahoo! (*http://www.yahoo.com*) o Altavista (*http://www.altavista.com*). Alguns cercadors, com ara Yahoo!, són en l'actualitat híbrids on disposem en una mateixa pàgina d'un índex temàtic i d'un motor de cerca.

Un motor de cerca està constituït per tres elements. En primer lloc hi ha el robot web (*spider* o *crawler* en anglès) que és una aplicació informàtica que visita de manera automàtica diverses pàgines web, les llegeix i va seguint els enllaços que hi troba. El robot va visitant periòdicament les pàgines web per detectar els possibles canvis. El segon component d'un motor de cerca és la base de dades o catàleg, que és simplement una col·lecció de còpies de totes les pàgines web que ha trobat el robot web. Fins que no s'ha inclòs una còpia d'una pàgina en el catàleg, la pàgina en qüestió no està disponible per al motor de cerca. El tercer component és el programari de cerca. Aquest permet a l'usuari especificar les condicions que han de complir els documents que busca (per exemple, que continguin una paraula determinada, que estiguin en un idioma determinat o que s'hagin publicat a la web en un interval de temps determinat) i convertir aquesta informació en el que s'anomena *l'equació de cerca*. A continuació el programari «pentina» tota la base de dades buscant els documents que compleixen l'equació de cerca per presentar-los posteriorment classificats segons la rellevància respecte del tema que s'està buscant.

L'eficiència d'un motor de cerca depèn en gran mesura de tres factors, i un és, naturalment, la grandària del catàleg. En principi, com més pàgines tingui indexades el catàleg, més exhaustiva serà la nostra cerca. El segon factor és la freqüència amb què es va revisant el catàleg. La informació a la web és molt dinàmica i un bon motor de cerca ha d'actualitzar la base de dades freqüentment si no vol acumular informació obsoleta o un gran nombre d'enllaços a pàgines que han deixat d'existir. Per acabar, és especialment important l'algorisme que usa el programari per assignar la rellevància als documents que recupera. Atès el volum d'informació contingut al catàleg, si l'equació de cerca no és gaire precisa es poden recuperar milions de documents que la compleixin. Davant la impossibilitat de llegir-los tots, és important que els primers de la llista siguin els que continguin la informació que més s'ajusti al que estem buscant. L'estratègia d'ordenació consisteix bàsicament en la localització i la freqüència de les paraules clau que busquem. Pel que fa a la localització, és evident que si volem, per exemple, pàgines relacionades amb la química i busquem pàgines que continguin el terme *química*, les que continguin aquest terme al títol o als encapçalaments principals tenen més probabilitats de ser pàgines que parlin de química que no pàgines on simplement hi aparegui la paraula *química* per casualitat (per exemple, en una pàgina on es parli de la *química* entre els dos actors principals d'una pel·lícula). L'altre criteri, el de la freqüència, es basa a suposar que si apareix moltes vegades un terme en una pàgina és perquè la pàgina tracta de qüestions relacionades amb aquest terme. Hem de recordar, però, que el càlcul de la rellevància es realitza automàticament i que sovint dóna lloc a sorpreses.

Un cas molt típic és el de pàgines de contingut eròtic, que per treure profit dels algorismes dels motors de cerca repeteixen en alguna part de la pàgina una gran quantitat de vegades alguns dels termes més buscats en la Xarxa (sovint sense cap relació amb el contingut de la pàgina on es troben) i aconsegueixen sortir així en els primers llocs a les cerques. Un altre cas que va aixecar força polseguera als Estats Units és el que va passar al cercador Google, en el qual en buscar el terme *jew* s'obtenia com a enllaç més rellevant el d'un grup d'ideologia neonazi. L'explicació donada pels responsables de Google és que en les pàgines dedicades al judaisme hi solen aparèixer referències a *judaism*, *jewish* o *jewish people*, mentre que el terme *jew* és el més freqüent en pàgines amb continguts antisemites.

Pel que fa a la quantitat d'informació emmagatzemada en el catàleg dels motors de cerca, en els més grans, com ara Google o AlltheWeb, hi ha indexats al voltant de quatre mil milions de pàgines web, mentre que per als mitjans com ara Altavista aquest nombre es redueix a uns mil milions. Per als qui vulguin una cerca més exhaustiva de la informació que hi ha a la web, hi ha unes eines específiques, els metabuscadors, que permeten realitzar simultàniament la mateixa cerca amb diferents motors. Si esteu interessats a obtenir més informació sobre aquest tipus d'eines, una pàgina que us pot ser útil és *http://www.searchenginewatch.com/links/index.php*. Igual que en el cas dels directoris, hi ha també motors de cerca especialitzats, dedicats, per exemple, a pàgines web per a infants, a pàgines relacionades amb la ciència, etc. En el camp de la química hi ha Chemie.de *(http://www.chemie.de*), un híbrid que inclou un directori i un motor de cerca especialitzat en química.

Hem vist fins ara, doncs, que una de les característiques principals d'Internet és la gran quantitat d'informació que hi ha a l'abast de l'usuari i que, per tant, és imprescindible usar alguna de les eines esmentades per intentar trobar el que busquem. Dues preguntes que sorgeixen en aquest context són, per una banda, saber si realment hi ha tanta informació com sembla o si aquesta es concentra de manera preferent en uns pocs temes i, per altra banda, saber de quina qualitat és. Malgrat que ambdues preguntes són de resposta difícil, podem assajar en el primer cas de donar una resposta utilitzant els motors de cerca. Sense cap ànim de ser exhaustius, a la taula 2 donem el nombre d'enllaços obtinguts a partir d'una sèrie de cerques realitzades amb diferents buscadors.

TAULA 2. Nombre aproximat d'enllaços (en milions) trobats per als termes indicats a la primera columna emprant quatre motors de cerca diferents

Terme	Google	AlltheWeb	Altavista	Yahoo!
Web	434	131	147	697
Sex	228	18	20	92
Science	105	28	29	146
Religion	8	12	12	56
Medicine	40	10	10	52
Physics	22	5	4	22
Biology	16	4	4	22
Chemistry	15	4	4	18
Mathematics	15	3	4	17
Geology	5	1	1	8
Astrology	5	2	2	10
Tarot	6	1	1	5
Alternative medicine	2	1	1	4

Els nombres recollits a la taula ens poden donar alguna idea de la quantitat d'informació que podem trobar a Internet sobre diferents temes. El primer terme cercat, *web*, ens dóna una idea de la grandària relativa dels catàlegs dels diferents motors de cerca, atès que *web* és un terme que se suposa que sortirà a moltes pàgines (només cal que al títol hi posi alguna cosa com ara «pàgina web dels amics de la cria de canaris») escrites en qualsevol idioma. Per als altres termes ens hem limitat a l'anglès, llengua predominant a Internet. És curiós observar que mentre que per a alguns termes, sobretot els relacionats amb la ciència, el nombre d'enllaços (en comparació als obtinguts per al terme *web*) és força semblant, hi ha altres termes en els quals s'observen diferències notables. El cas més xocant és la diferent proporció entre *sex* i *religion* que es troba a Google en comparació amb els altres motors.

Pel que fa a les pàgines web relacionades amb temes científics, veiem que, a part de la medicina, les disciplines per a les quals es troba major volum d'informació a la web són la física i la biologia, amb la química en tercer lloc. En el cas de la química hem realitzat una cerca més detallada (taula 3) i hem trobat que la bioquímica és l'àrea sobre la qual hi ha més quantitat d'informació, mentre que la química inorgànica és la menys present a la web. És interessant comprovar, tot comparant aquestes dades amb les de la taula anterior, que algunes disciplines esotèriques com ara l'astrologia o el tarot

TAULA 3. Nombre aproximat d'enllaços (en milions) trobats per als termes indicats a la primera columna emprant quatre motors de cerca diferents

Terme	Google	AlltheWeb	Altavista	Yahoo!
Organic chemistry	1,10	0,18	0,19	1,00
Inorganic chemistry	0,30	0,06	0,06	0,30
Analytical chemistry	0,50	0,10	0,10	0,50
Physical chemistry	0,80	0,09	0,10	0,60
Biochemistry	3,70	0,90	0,90	5,20
Chemical engineering	2,00	0,30	0,30	1,70

són capaces d'acumular una quantitat de pàgines web considerablement més gran que qualsevol camp de la química per separat!

La possibilitat que ens ofereixen els buscadors de limitar les cerques a pàgines web escrites en un idioma determinat ens permet també tenir una idea del volum d'informació que hi ha sobre una matèria en qualsevol idioma. En la taula 4 es recullen els resultats trobats a Google buscant pàgines web que fessin referència a la química en algunes llengües europees. Les dades indiquen que, tal com era d'esperar, l'anglès és la

TAULA 4. Nombre aproximat d'enllaços (en milions) trobats per als termes indicats a la primera columna emprant Google. A la tercera columna s'indica quants d'aquests enllaços corresponen a pàgines web escrites en un idioma determinat i, a la darrera, el nombre d'enllaços de química que hi ha en relació amb els parlants de cadascuna de les llengües analitzades

Terme	Total	De les quals en:	Enllaços/parlants
Chemistry	15,40	Anglès: 6,50 (42 %)	0,020
Química	0,99	Castellà: 0,85 (86 %) Català: 0,06 (6 %)	0,002 0,015
Chemie	5,20	Alemany: 2,10 (40 %)	0,021
Chimie	1,27	Francès: 0,73 (43 %)	0,010
Chimica	1,18	Italià: 0,65 (55 %)	0,017

llengua predominant. Si dividim el nombre d'enllaços pel de parlants natius de cadascun del idiomes veiem que la proporció de pàgines dedicades a la química és del mateix ordre en tots els idiomes escollits, excepte el castellà, llengua per a la qual és excepcionalment baixa. Des d'aquest punt de vista podem constatar que la química en català té una presència a Internet comparable a la d'altres llengues europees com ara l'italià o el francès.