

Elements d'estadística i de matemàtica de la informació

Infometria, bibliometria, mediametria,
cienciometria, museometria, webmetria

Thierry Lafouge
Yves-François Le Coadic
Christine Michel

Prefaci de Leo Egghe

PID_00187820

Material docent de la UOC

Thierry Lafouge

Professor titular de l'Escola Nacional Superior de Ciències de la Informació i de Biblioteques de Lió (ENSSIB), membre del laboratori de recerca RECODOC (Représentation des Connaissances et Documentation) de la Universitat de Lió 1 i professor HDR (habilitat per a dirigir recerques).

Yves-François Le Coadic

Professor de l'Escola Nacional d'Arts i Oficis de París (CNAM). Autor de les obres *La Science de l'information* (PUF), *Usages et usagers de l'information* (Nathan) i *Besoin d'information* (ADBS) i membre del Comitè de Redacció del *Dictionnaire encyclopédique de l'information et de la documentation* (Nathan).

Christine Michel

Professora titular de la Universitat de Bordeus 3, membre del Grup de Recerca sobre l'Estudi de Sistemes Informatitzats de Comunicació (GRESIC) de la Universitat de Bordeus 3 i membre associada del REGODOC.

L'encàrrec i la creació d'aquest material docent han estat coordinats pel professor: Víctor Cavaller Reyes (2012).

Primera edició: febrer 2012

© Thierry Lafouge, Yves-François Le Coadic, Christine Michel

Tots els drets reservats

© d'aquesta edició, FUOC, 2012

Av. Tibidabo, 39-43, 08035 Barcelona

Disseny: Manel Andreu

Material realitzat per Eureka Media, SL

Dipòsit legal: B-3.148-2012

Cap part d'aquesta publicació, incloent-hi el disseny general i de la coberta, no pot ser copiada, reproduïda, emmagatzemada o transmesa de cap manera ni per cap mitjà, tant si és elèctric, com químic, mecànic, òptic, de gravació, de fotocòpia, o per altres mètodes, sense l'autorització prèvia per escrit dels titulars del copyright.

Prefaci

Les disciplines de la infometria, la bibliometria i les altres “-metries” que s'esmenten en el títol d'aquesta assignatura tenen com a objectiu mesurar i avaluar els diferents aspectes dels fenòmens informacionals i, en aquest sentit, són comparables a altres disciplines molt populars, com l'econometria, la sociometria, la psicometria, la quimiometria o la biometria. Les disciplines estudiades en aquesta assignatura mesuren, en primer lloc, l'evolució (és a dir, el creixement o l'envelliment) de les informacions, però a causa del caràcter explosiu d'aquesta evolució, els autors proposen mesurar també la “qualitat” de la informació, i per això ens ofereixen alguns mitjans de selecció; el que ens interessa actualment no és tant tenir “tota” la informació, sinó tenir la informació “bona”.

El desenvolupament ràpid de les biblioteques electròniques, d'Internet i de la informació en general té unes conseqüències importants per a moltes d'aquestes “-metries”. Disciplines com la cienciometria, la bibliometria i la infometria estan experimentant un creixement exponencial (o potser fins i tot més que exponencial) en termes d'investigadors i d'articles publicats. A més, sorgeixen disciplines noves, com la webmetria. Totes aquestes disciplines tenen moltes coses en comú, però també desenvolupen els seus problemes propis i les seves solucions pròpies. També comparteixen les tècniques estadístiques i les distribucions probabilistes que s'apliquen de manera habitual en tots aquests dominis, com per exemple el creixement exponencial o el creixement logístic, les lleis de Lotka i de Zipf, o les tècniques estadístiques unidimensionals o multidimensionals que permeten cartografiar i fabricar un atlas de les xarxes imbricades, com les xarxes de citacions o la xarxa mateixa WWW. Però el fet que el factor temps estigui normalment absent de l'estudi del Web és un element que indica que la webmetria té uns problemes propis com, per exemple, el fet que les pàgines web no tenen una data concreta de publicació, amb la qual cosa els seus hipervincles no són comparables a les citacions o a les referències utilitzades en la cienciometria per als estudis d'obsolescència, per exemple.

Així, doncs, com es poden abordar aquestes diferents disciplines i les seves tècniques dins d'una mateixa obra? Una primera manera de procedir és agafar una “-metria” concreta (o un conjunt de “-metries” semblants) i subdividir l'assignatura en mòduls dedicats a problemes “-mètrics” específics. Aquest és el camí escollit per la majoria dels llibres que tracten d'aquest tema. És una manera de procedir molt satisfactòria: cada capítol estudia un problema (per exemple, l'anàlisi de les citacions) aplicant els mètodes estadístics o matemàtics (o probabilistes) necessaris. Aquest mètode ofereix una visió de conjunt excel·lent dels diferents problemes, però no fa una classificació clara de les tècniques estadístiques i matemàtiques; les mateixes tècniques apareixen i tornen a aparèixer en els diferents capítols d'un llibre organitzat així. Una altra manera de procedir

(que és la que s'ha emprat en aquesta assignatura) és presentar per endavant les diferents tècniques estadístiques i matemàtiques possibles i després estudiar on i com es poden emprar (en totes les "-metries" factibles). Això ofereix una idea general clara de la ciència comuna a totes les "-metries" diferents i permet que el lector pugui aplicar posteriorment aquestes tècniques a altres problemes.

Aquesta assignatura també és molt completa en el sentit que conté dos mòduls dedicats a les aplicacions de l'estadística i les matemàtiques a les diverses "-metries" precedits per un mòdul sobre la mesura de la informació. Aquest mòdul és molt important, ja que obliga l'estudiant a reflexionar sobre el que poden ser "les mesures de la informació" abans de mesurar-les realment. Brinda la possibilitat de comprendre realment què representa aquesta mesura de la informació en totes les seves diferents interpretacions (és a dir, tipus de mesura).

El fet que els autors hagin estat capaços d'incloure exemples i problemes d'infometria, bibliometria, cienciometria, mediametria, museometria i webmetria és realment destacable i, pel que sé, únic en aquest camp. També cal destacar la constatació que les "velles" aplicacions vinculades a les biblioteques clàssiques o als museus són semblants a les noves aplicacions vinculades al Web o a les biblioteques electròniques.

Convido els estudiants a apreciar aquest treball (com jo mateix he fet) i a utilitzar-lo com a ajuda per a resoldre les nombroses situacions pràctiques amb les quals es trobaran. Estic segur que els ajudarà a resoldre-les o que, com a mínim, els oferirà indicacions per a trobar la solució.

Introducció

“Per mitjà de l'estadística, però sense un excés de matemàtica, es tracta d'abordar els problemes generals de la forma i del volum de la informació i de les lleis fonamentals que en regeixen la producció, el creixement, la comunicació i l'ús.”

Derek J. de Solla Price

Fa ja molt de temps que s'intenten descriure quantitativament els fenòmens informacionals. L'entrada de l'estadística i de la matemàtica en la ciència de la informació es va produir en la dècada de 1920. També van aparèixer les primeres lleis científiques que posaven de manifest unes relacions quantitatives relativament constants i expressables mitjançant funcions matemàtiques que estableixen unes relacions universals i necessàries entre l'aparició d'un fenomen i les condicions que en provoquen l'aparició, que permetien fer previsions.

La primera d'aquestes lleis va ser enunciada el 1926 per Alfred Lotka. En estudiar l'índex decennal 1907-1916 de *Chemical Abstracts*, va constatar l'existència d'una relació simple entre el nombre d'autors i el nombre d'articles que havien publicat. Alguns anys més tard, el 1934, Samuel G. Bradford, un bibliotecari britànic, va elaborar, amb l'ajuda d'un model matemàtic simple, un mètode d'organització de la documentació que permetia determinar la distribució dels articles sobre un tema concret en les revistes científiques i tècniques. Aquest descobriment va ajudar els gestors a definir la política d'adquisicions d'un centre de documentació. Aquest mètode també permet localitzar la informació pertinent dins de la massa de totes les referències disponibles. D'això es derivaran nombrosos i cèlebres treballs.

L'any següent, el 1935, va ser un lingüista americà, George Zipf, qui es va dedicar a mesurar la freqüència d'aparició de les paraules i va establir la llei que porta el seu nom. Més tard, el 1948, com a complement al desenvolupament de la comunicació a distància, un enginyer americà, Claude Shannon, va elaborar, seguint els passos d'R. V. L. Hartley (1928), una teoria estadística de la transmissió dels senyals elèctrics (anomenada, inexplicablement, *teoria de la informació*). Aquesta teoria explica la probabilitat de transmissió d'un senyal elèctric a través d'una línia de transmissió. El 1968, Philip M. Morse, un matemàtic americà, aplica els models de previsió sorgits de la recerca operativa a la gestió de la circulació de les obres en les biblioteques. Finalment, el 1976, Derek J. de Solla Price construeix un model probabilista que explica diferents fenòmens característics de les “produccions bibliomètriques”. Aquest model i les seves conseqüències continuen essent, segons el nostre parer, i de bon tros, la més important de les contribucions, precursora dels estudis futurs en bibliometria, cienciometria, infometria, mediametria, museometria i webmetria, és a dir, de l'aplicació de l'estadística i de la matemàtica als llibres, a les biblioteques, a la recerca científica i tècnica, a la informació científica i tècnica, als mitjans de comunicació, als museus i a Internet.

Tots aquests treballs confirmen l'existència, en el camp de la informació, de regularitats, de distribucions i de relacions mesurables i universals que permeten fer previsions i que es poden mobilitzar tant en el sector cultural com en el sector comercial. Però no ha estat fins fa molt poc que tot aquest cos de coneixements estadístics i matemàtics s'ha reunit en algunes obres, bàsicament en anglès, a les quals fem referència, com per exemple la de Léo Egghe i Ronald Rousseau (*Introduction to Informetrics: Quantitative Methods in Library, Documentation and Information Science*, Elsevier, 1990). I fa encara menys que tot aquest cos de coneixements s'ha començat a aplicar també als sectors mercantils de la informació dins dels sectors culturals.

Actualment, en els sectors de la cultura, l'educació i la recerca, una bona gestió dels serveis públics exigeix com més va més l'ús d'una àmplia gamma d'eines de gestió adaptades als contextos culturals, educatius i científics, i a les dimensions i el caràcter del servei. Són eines d'anàlisi de les necessitats d'informació de la comunitat a la qual serveixen, eines de control i d'avaluació i eines de mesura del rendiment que permeten que l'establiment disposi d'una bateria d'indicadors de rendiment.

L'oferta de serveis d'informació ha de respondre a les necessitats d'informació del públic al qual serveixen. Com que les necessitats canvien amb el temps, els seus processos de determinació s'han de repetir a intervals regulars. Aproximadament cada cinc anys, les biblioteques públiques han de recollir dades sociodemogràfiques i informació sobre els organismes públics i les empreses de la zona, sobre les zones habitades, sobre les xarxes de transport i sobre els altres serveis d'informació. A més, han de fer, de manera regular, enquestes entre els usuaris per conèixer les seves necessitats d'informació i de serveis, però també per saber com valoren la informació obtinguda i els serveis utilitzats. Per a obtenir uns resultats objectius, aquestes anàlisis han de ser a càrrec de persones o d'organismes externs al servei d'informació que s'analitza.

Tot seguit, per a garantir que les estratègies aplicades permetran aconseguir els objectius fixats, el control del servei recorre de manera contínua a les estadístiques procedents de les anàlisis de les necessitats d'informació, de les enquestes de satisfacció i dels indicadors de rendiment. Tots els programes i els serveis també s'avaluaran de manera regular per a comprovar si s'han aconseguit els objectius.

Finalment, els indicadors de rendiment que subministren la informació sobre el rendiment del servei són unes eines especialment necessàries. Estem parlant d'indicadors d'ús (i de no-ús) obtinguts per mitjà de sondejos, enquestes, etc., indicadors de recursos humans i materials, indicadors de qualitat i indicadors de costos. Per la seva banda, les mesures de rendiment serveixen per a mesurar l'eficàcia del servei mitjançant un estudi dels béns *intermediaris* (recursos) i dels *productes* (resultats) d'any en any. Les mesures poden ser comparatives.

En l'altre extrem, la gestió comercial que envaeix Internet d'una manera particular implica fer anàlisis estadístiques de la "relació amb el client":

- Quadre d'indicadors del seguiment de l'activitat dels llocs; audiència per hora, data (dia, setmana, mes), nombre de sessions, de màquines, de pàgines visitades, de clics, etc. Però davant la falta d'unitats de mesura coherents, d'eines fiables i de mètodes comprovats, resulta difícil conèixer el nombre real de consultes dels llocs web.
- Exploració de la informació per a actualitzar les pràctiques dels usuaris i la identificació dels comportaments típics; l'objectiu d'aquests mètodes és arribar a convertir els visitants en compradors.
- Mesura permanent de la qualitat dels serveis oferts gràcies als qüestionaris que es presenten de manera aleatòria als usuaris dels llocs o bé en resposta al seu comportament en el lloc.
- Sondejos regulars entre els usuaris per a conèixer-ne els usos, els costums, les expectatives i les preferències.

Als qui aquesta matematització pugui resultar sorprenent, només direm que l'aplicació de la matemàtica a l'anàlisi dels fenòmens socials i humans no és, ni de bon tros, una cosa recent. Aquesta via ja la van obrir Georges L. L. Buffon en l'aritmètica moral, Marie-Jean A. Condorcet parlant sobre el problema de l'interès general, l'examen detingut dels escrutinis, etc. Però van ser l'economia (a la fi del segle XIX), la demografia, la psicologia (1910) i, més tard, la sociologia (1950) les primeres que van donar lloc a recerques matemàtiques. I d'aquí van sorgir l'econometria, la psicometria i la sociometria. Però tot això no ha estat gens fàcil. Així, en la biologia, els treballs quantitius han necessitat molt temps per a adquirir el dret de ser citats. Per a poder publicar i, en conseqüència, legitimar els treballs quantitius en biologia, l'estadístic Karl Pearson, famós pel seu coeficient, va crear al principi del segle una nova revista científica, la revista *Biometrika*.

Així, doncs, en les ciències de la informació avui és possible enumerar, classificar, distribuir i mesurar usant eines i objectes estadístics i matemàtics: els primers, però no els menys importants, són els nombres.

Continguts

Mòdul 1

La mesura de la informació

Thierry Lafouge; Yves-François Le Coadic; Christine Michel

1. El nombre
2. La mesura
3. Sistema d'informació i de mesura
4. Mesures i lleis

Mòdul 2

Estadística de la informació

Thierry Lafouge; Yves-François Le Coadic; Christine Michel

1. Escales i unitats de mesura estadística
2. L'estadística unidimensional
3. L'estadística bidimensional
4. L'estadística multidimensional
5. L'estadística probabilista

Mòdul 3

Matemàtica de la informació

Thierry Lafouge; Yves-François Le Coadic; Christine Michel

1. Precisió i límit en matemàtiques: el zero i l'infinit
2. Les sèries
3. Les funcions
4. Les equacions
5. Els conjunts

Annexos

Thierry Lafouge; Yves-François Le Coadic; Christine Michel

1. Les proves en estadística
2. Taules estadístiques
3. Càlcul de moments
4. Derivada i integral, exemple de resolució d'equacions diferencials