

**Texty: estudi per a millorar la
recuperació d'informació
amb visualització de textos**

**TREBALL D'INVESTIGACIÓ
PROGRAMA DE DOCTORAT EN SEGURETAT I PREVENCIÓ**

**ESCOLA DE POST GRAU
UNIVERSITAT AUTÒNOMA DE BARCELONA (UAB)**

Jaume Nualart i Vilaplana

Direcció tesis

Dra Roser Martínez Quirante

Escola de Prevenció i Seguretat Integral
Coordinadora de Recerca i professora
Universitat Autònoma de Barcelona

Dr Mario Pérez-Montoro

Departament de Biblioteconomia i Documentació
Professor i coordinador del programa de doctorat Informació i
Documentació en la Societat del Coneixement
Universitat de Barcelona

Índex

Introducció

1. Objectius

2. Presentació i descripció tècnica de l'eina proposada: texty

3. Estudis aplicats:

3.1. Textos humanistes de crítica d'art (media art): arxiu d'Ars Electronica

3.2. Literatura científica: Information Research papers

3.2.1 Fonts dels corpus de vocabularis

3.2.2 Tractament dels termes per a cada vocabulari

3.2.3 Corpus de textos per a presentar

3.2.4 Creació dels textys

3.2.5 Exemple aplicat a un número de la revista

3.2.6 Conclusions

3.3. Estudi per a 1 texty

3.3.1. Comparativa de texty amb diagrama de barres:

3.3.2. Comparativa de texty i diagrama de línies:

3.3.3. Conclusions de la comparació entre texty i els diagrames de barres i de línies

4. Estudi proposat:

4.1. Textos legals: sentències

5. Conclusions

6. Futur de texty

Introducció

La recuperació d'informació és un factor crític en un entorn d'excés d'informació ¹.

Aquest és el punt de partida d'una investigació que vol contribuir a la millora del treball informacional, especialment del treball que implica textos.

Fa uns 15 anys es parlava de l'excés d'informació i dels perills d'aquest ². La gent que hem escollit dedicar-nos a la tecnologia de la informació sempre hem pensat que hem de ser capaços de veure l'excés d'informació com una avantatge que, això si, requereix de l'ús intel·ligent de la tecnologia per tal que aquest pensament es transformi en una realitat d'ús quotidià.

La tecnologia de dades amb ordinadors, fora del món especialitzat, es veu com un obstacle. La visualització de dades és dedica a dotar-nos d'eines gràficament accessibles que tradueixin, representin o interpolin; en aquest sentit és que la visualització de dades aspira a tenir excessos d'informació, dels quals, poder extreure'n les dades que ens interessin en cada cas i cada ús.

Partint de la idea de representar les essències d'un textos, en aquest treball d'investigació presentem primer un estudi d'aquesta idea, que anomenem texty; i després presentem tres aplicacions a tres camps de coneixement, és a dir, a tres diferents corpus de text que tenen, en cada cas, un mateix registre. Ens referim als camps del media art, la literatura científica i els textos legals.

Aquesta es la primera part del doctorat que vol posar eines concretes a l'abast de tothom per a poder treballar amb més efectivitat i comoditat amb textos.

Recuperació d'informació

Wikipedia:(http://en.wikipedia.org/wiki/Information_retrieval)

Recuperació d'informació(IR) és l'àrea d'estudi relacionada amb la recerca de documents, la recerca d'informació dins dels documents i per les metadades d'aquests documents, així com la recerca d'emmagatzematge de dades estructurat, bases de dades relacionals i dades a la Internet. Hi ha un solapament en l'ús de les expressions de recuperació de documents, recuperació d'informació, i la recuperació de textos, però cada una té un cos propi en la literatura, la teoria, la praxi i la tecnologia. IR és interdisciplinari, basat, en la informàtica, les matemàtiques, la ciència bibliotecària, les ciències de la informació, l'arquitectura d'informació, la psicologia cognitiva, la lingüística,

1. Baeza-Yates, Ricardo, and Berthier Ribeiro-Neto. Modern Information Retrieval. New York: Addison-Wesley, 1999.

2. David Bawden, Clive Holtham, Nigel Courtney, (1999) "Perspectives on information overload", Aslib Proceedings, Vol. 51 Iss: 8, pp.249 - 255

Davant d'una cerca realitzada per un usuari, els sistemes de recuperació d'informació solen respondre amb un llistat de resultats basats en rànquings, i aquests resultats solen ser presentats en llistats plans d'una dimensió. Normalment, aquests llistats són opacs quant l'ordre; l'usuari no sap perquè el llistat té un ordre determinat. I, per a refinar, l'usuari ha d'interactuar de nou, normalment, fent un filtre de la primera recerca.

S'ha treballat molt sobre el tema de la representació de resultats de cerca. Les noves propostes presenten representacions en 2-D o 3-D dels resultats. Aquests sistemes plantegen, per tant, una nova metàfora, més complexa que el tradicional llistat 1-D.³, la qual cosa requereix d'un entrenament extra de l'usuari per a llegir o interactuar plenament amb la visualització.

El camp de la visualització de la recuperació d'informació es creixent i ja des dels anys 90 s'ha proposat visualitzacions 2-D i 3-D per a resultats de cerques⁴. En treballs com Lighthouse⁵, es proposa una interessant combinació entre *ranking list* i *clustering visualization*. La representació de resultats de cerques usant gràfics presenta una manera no habitual de veure aquests resultats. Tant mateix, no és evident l'ús i l'aprofitament de moltes d'aquestes representacions per a un usuari estàndard sense una explicació o entrenament previs.

1. Objectius

Volem demostrar que la forma en que es presenta la informació en els estàndards de recuperació d'informació és millorable amb eines complementàries de visualització de dades.

Concretament presentem *texty* com a eina de representació de textos basat en vocabularis predefinits segons els interessos de l'usuari en el moment de la recerca.

Primer descrivim tècnicament l'eina proposada i en fem la història del desenvolupament. Després presentem tres estudis que s'han fet amb *texty* per a demostrar-ne la utilitat i coneixe'n les habilitats.

El primer estudi descriu la creació de *texty* i demostra visualment la lectura dels *textys*.

El segon es un estudi qualitatiu comparat amb sistemes clàssics de representació de dades. On demostrem que *texty* dona més informació i de manera més clara que els sistemes clàssics (diagrama de barres, de línies)

I el tercer és una proposta d'aplicació de *texty* a textos de sentències judicials. On presentem un petit estudi fet a professionals del dret per extreure'n les necessitats de cerca en el treball diari amb sentències judicials.

3. Information Retrieval Visualization. CPSC 533c Class Presentation. Qixing Zheng. March 22, 2004

4. Design of 3-D Visualization of Search Results: Evolution and Evaluation. John Cugini*, Sharon Laskowski. National Institute of Standards and Technology, Gaithersburg, MD 20899. Marc Sebrechts. The Catholic University of America, Washington, DC 20064

5. Information Visualization, 2000. InfoVis 2000. IEEE Symposium on In Information Visualization, 2000. InfoVis 2000. IEEE Symposium on (2000), pp. 125-129.

L'objectiu d'aquest treball d'investgació és demostrar que l'ús d'eines de visualització en la cerca i estudi de textos és una realitat aplicable i, potser, una necessitat inevitable.

2. Presentació i descripció tècnica

En aquest treball presentem una eina complementària a la representació tradicional unidimensional de resultats de sistemes de recuperació d'informació. Aquesta eina representa les essències dels continguts de cada ítem d'una llista retornada per un sistema de recuperació d'informació i ajuda a l'usuari a poder identificar el contingut més adient per a satisfer la seva necessitat d'informació abans d'abordar intel·lectualment cadascun del resultats.

Dit en paraules planeres, texty és una representació gràfica d'un text donat que permet fer una molt ràpida *lectura en diagonal* del text en qüestió.

Un texty és una imatge que representa només algunes paraules clau d'un text agrupades en colors/categories i distribuïdes com a punts de colors i posicionats proporcionalment respecte el text original. Texty és, per tant, una representació física d'alguns grups de paraules que determinen una característica sobre tipus de contingut del text.

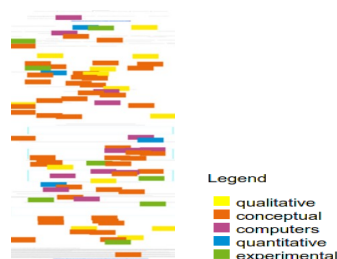


Fig 1. Un texty amb llegenda de colors

A nivell gràfic, hem usat tècniques molt semblants a les usades per Keim en el reconeixement d'autories ⁶. La tècnica usada per Keim té una aplicació molt diferent, doncs representa l'allargada de les frases de cada text com a quadrets amb una gradació de colors. A texty el conjunt de punts d'un color mostren la densitat de conceptes referents o representatius d'un camp lingüístic determinat, que anomenem *vocabulari*. Les visualitzacions de Keim les analitza l'ull humà de manera molt semblant a texty, on els factors a determinar visualment són: zones de color, densitat de punts, posició i distribució al pla dels punts.

Un texty és, per tant, una imatge-icona que representa la distribució física de paraules clau d'un text com una imatge plana. Aquestes paraules clau s'agrupen en vocabularis; a cada vocabulari se li associa un color. En pocs

6. Literature Fingerprinting: A New Method for Visual Literary Analysis, by Daniel A. Keim http://inforvis.uni-konstanz.de/papers/2007/2007_va

segons es pot saber sobre l'estructura, la densitat conceptual i la temàtica del text.

Texty és una tècnica no intrusiva doncs no interfereix en el *ranking* original del sistema de recuperació d'informació en concret. En aquest treball demostrem que aquesta eina de representació de textos enriqueix els llistats 1-dimensionals de recursos.

Els estudis comparatius amb sistemes clàssics de representació que presentem més endavant, mostren que Texty és una tècnica més efectiva que els diagrames clàssics (diagrama de barres i de línies) quant a temps que es necessita per a analitzar-ne un, i quant a la complexitat de la informació espacial que conté. El cervell humà és capaç de detectar molt ràpidament zones amb major densitat de colors que representen un camp lingüístic o un vocabulari concrets.

3. Estudis aplicats

Presentem primerament dues aplicacions i, a continuació, una proposta d'aplicació de texty. Els corpus de textos escollits pertanyen tres camps de coneixement diferents:

Estudis aplicats:

- Textos humanistes de crítica de *media art* (arxiu d'Ars Electronica)
- Literatura científica (Information Research Journal)

Estudi proposat:

- Textos legals I (sentències)

Estudis aplicats

3.1 Textos humanistes de crítica d'art (*media art*): arxiu d'Ars Electronica

El 2008-9, com a investigador a l'Institut Ludwig Boltzman (Linz, Austria), des del departament de visualització d'informació⁷ ens vam proposar el repte de fer eines de visualització amb les dades de l'arxiu d'Ars Electronica, un arxiu de cultura digital, *media art* i tecnologia que recull dades des del 1987. Es va identificar una mancança en la representació de col·leccions de textos amb un mateix registre lingüístic. Es cercava la manera de representar un text abans de llegir-lo. Una manera per a discriminar en un llistat de textos i per a poder comparar-los.

Ars Electronica⁸ és un festival, segurament el més important d'Europa, sobre cultura digital que es mou en en la intersecció entre *media art*, ciència i tecnologia. Gairebé ininterrompudament des del 1987 cada setembre s'entreguen a Linz (Àustria) el premis Golden Nica. Aquests premis s'han

7. <http://vis.mediaartresearch.at/webarchive/public/view/mid:1>

8. <http://www.aec.at>

agrupat en categories que, en passar els anys, han anat canviant per adaptar-se a nous conceptes.

D'aquest immens i asimètric arxiu vam escollir treballar amb gairebé els cent textos que els jurats van escriure entre els anys 1987 i 2007 per a anunciar cada premi atorgat. Aquests textos tenen una extensió d'entre 1.500 i 4.000 caràcters i, amb *texty*, podem veure la diferent natura i estructura que tenen.

Inicialment, a més dels textos, es va disposar de 5 vocabularis (*Art work, Person or Institution, Date, Keyword i Award*) de molta qualitat sobre la història del Media Art elaborades per Gerhard Dirmoser⁹. La feina de Dirmoser va proporcionar la base per a desenvolupar una eina que representés aquests 5 vocabularis amb 5 colors diferents en una imatge proporcional i representativa del text físic. De la feina al LBI en va sortir les primeres representacions intuïtives amb la tècnica de *Texty*.

La llegenda de color i categories la podem veure a la Fig.2

Legend



Fig. 2: Llegenda de colors/categories dels vocabularis per a l corpus de textos d'Ars Electronica.

A la Fig. 3 veiem dos *textys* per a dos textos del corpus. Ambdós de l'edició d'Ars Electronica del 2007 però de categories/premis diferents. Veiem que el "*Statement of the computer animation*" és un text de caire social, a on se citen moltes persones i institucions. A la segona meitat és llisten obres d'art i persones relacionades. En canvi el "*Statement of the digital communities*" es un text de caire clarament teòric (el violeta indica "keywords" -paraules clau- en el camp del *media art*). A la part final, en mig del discurs teòric continuat, es citen obres d'art. Només a l'inici es citen certes persones o institucions (*punts taronges*) i es donen algunes referències històriques (*punts blaus*).



Fig. 3 Dos *textys* de dos textos del corpus de decisions dels jurats dels Golden Nica Awards. Ars

9. http://90.146.8.18/en/archives/picture_ausgabe_03_new.asp?iAreaID=145&showAreaID=149&iImageID=38289

Si en canvi comparem textos de les mateixes categories, trobem estils molt semblants. Això en mans d'un historiador/a de l'art pot ser analitzat en funció de les diferents disciplines que representen les categories d'aquests premis històrics.

En la Fig. 4 podem veure aquest cas exemplificat. La majoria de textos de la categoria Computer Animation a l'inici del text fan algun tipus de referència teòrica, i la segona meitat del text sol contenir moltes referències a persones i institucions, relacionant-les amb obres d'art.

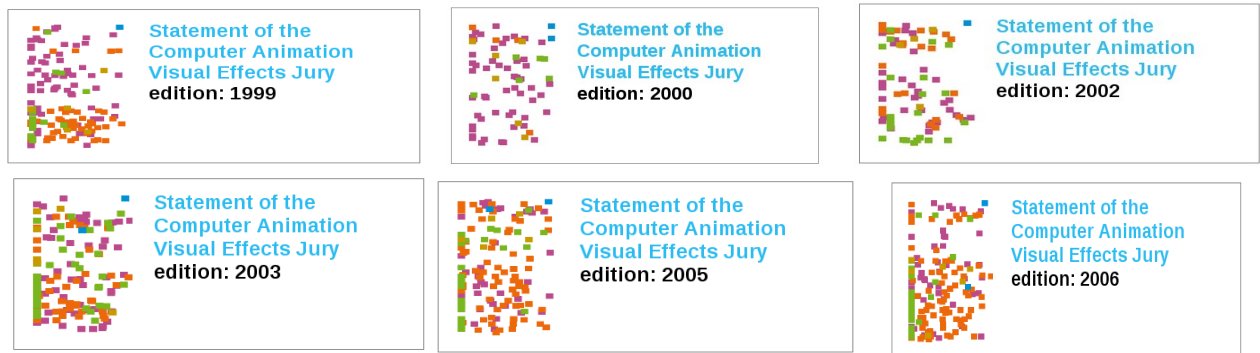


Fig. 4: Sis textys del premi d'Animació per ordinador (entre els anys 1999 i 2006) dels Golden Nica Awards. Ars Electronica. Linz (Àustria).

Aquest prototipus va formar part del catàleg oficial del 30è aniversari d'Ars Electronica el setembre del 2009, dins l'exposició "Mapping the Archive: Prix Ars Electronica"¹⁰

3.2 Literatura científica: Information Research papers

Després de la primera aplicació de texty i els primers dissenys fets al Ludwig Boltzman Institute de Linz, i ja dins del programa de doctorat a la UAB, vam voler seguir amb la investigació i fer un estudi científic sobre els prometedors beneficis dels textys, i demostrar-ne, així, la validesa.

Per fer aquest estudi vam cercar una col·lecció no massa gran de textos amb un registre semblant i un camp semàntic específic. A més, els textos havien de ser accessibles lliurement per facilitar-ne l'estudi.

Per totes aquestes raons vam escollir els articles publicats a *Information Research Journal*¹¹. Una revista oberta publicada i editada pel Professor T. D. Wilson (***** <http://informationr.net/tdw/biog.html>).

Els articles pertanyen a un mateix fons documental, són una unitat, comparteixen el registre acadèmic, tenen una estructura semblant (*intro, method, analisys, results*) i tenen qualitat estandarditzada (*peer reviewed*).

10. <http://www.aec.at/humannature/de/history-lounge/mapping-the-archive-prix-ars-electronica>

11. <http://informationr.net/ir/>

A més, desenvolupar programari amb llicències lliures, com és el cas de Texty, té més sentit quan s'aplica a dades també lliures i accessibles. Les dades accessibles, com els articles del *Information Research Journal*, conviden a investigadors i usuaris a millorar o complementar les eines tecnològiques que tenim per llegir-les, usar-les i entendre-les.

La web d'*InformationR* disposa de sistema d'exploració: per tema, per número, per autor. Disposa d'un llistat de *reviews* separat. I dos sistemes de recuperació d'informació: *Atomzsite search* i *Google*.

Una vegada triat el corpus de textos, vam identificar les següents categories temàtiques que podrien ajudar a classificar els continguts d'aquets textos: *Conceptual Approach*, *Experimental Approach*, *Qualitative Methodology*, *Quantitative Methodology and Computers/Technology*. Es va triar aquestes categories perquè representen una part important de les característiques principals i essencials dels articles publicats en aquesta revista científica: tipus d'*enfoc*, tipus de metodologia i grau de tecnologia emprada.

3.2.1- Fonts dels corpus de vocabularis

El pas següent va ser identificar les fonts d'informació d'on poder extreure els vocabularis que es desenvoluparen posteriorment. La tria d'aquestes fonts es va basar en dos criteris; el prestigi intel·lectual: (Stanford Encyclopedia of Philosophy¹² i l'enciclopèdia Britànica¹³; i llur popularitat: Wikipedia¹⁴. La distribució de fonts per temàtica es va realitzar de la següent manera:

12. <http://plato.stanford.edu/>

13. <http://www.britannica.com/>

14. <http://www.wikipedia.org/>

	Definicions				
	Qualitative Methodology	Conceptual Approach	Computers/Tec hnology	Quantitative Methodology	Experimental Approach
Stanford Encyclopedia of Philosophy	Aristotle's Categories Intrinsic vs. Extrinsic Properties	Concepts Category Theory		Mathematics Statistics	Experiment in physics
Britanica	Qualitative states Qualitative tests to distinguish alternative theories			Mathematics Statistics	
Wikipedia	qualitative data Quantitative property Qualitative properties Qualitative research Quality (philosophy)	Terminology Theory Vocabulary Concept	list of programing languages list of popular computers list of hardware componets, software glossary		Test method Case study Experiment

Table 1. Conceptes i fonts dels conceptes pels 5 vocabularis escollits.

3.2.2- Tractament dels termes per a cada vocabulari

A continuació es va definir els 5 vocabularis partint dels 5 corpus de textos dels conceptes escollits (Veure Taula 1). Primer es va passar un filtre *stopwords*, per a treure'n les paraules buides. Després es va eliminar les paraules que estaven presents menys de 4 vegades, considerades poc significatives per a cada temàtica. Després es van eliminar les paraules que estan presents en més d'un vocabulari, és a dir, les interferències entre vocabularis. Vam obtenir, així, un número de termes per a cada vocabulari.

	Número de termes	Termes /1.000 paraules
Conceptual Approach	610	21,16
Experimental Approach	510	23,29
Qualitative Methodology	451	19,71
Quantitative Methodology	700	22,24
Computers/Technology	312	18,91

Taula 2. Número de paraules seleccionades per a cada vocabulari abans de fer la revisió intel·lectual.

Finalment, es va fer una revisió intel·lectual per a detectar;

- Termes inconsistents respecte la temàtica
- Termes no unívocs.
- Termes incoherents respecte a cada vocabulari

Cal dir que l'objectiu d'aquest article es presentar les potencialitats de l'eina Texty i que en futures investigacions es realitzaran estudis acurats de les millors estratègies per a definir les paraules que millor representen un camp concret de coneixement o, com l'anomenem en aquest article, vocabulari.

Finalment, els nostres vocabularis experimentals van quedar configurats així:

	Número de termes
Conceptual Approach	76
Experimental Approach	57
Qualitative Methodology	67
Quantitative Methodology	69
Computers/Technology	410

Taula 3. Número final de termes per a cada vocabulari..

El vocabulari *Computers/Technology* es descriptiu, per això hem deixat un gran número de termes, doncs tots ells fan clara referència a *ordinadors o tecnologia*.

El llistat final de paraules es pot trobar online : <http://v.subvideo.tv/texty/terms.php>

3.3.3- Corpus de textos per a presentar

Escollits els articles d'Information Research Journal com a corpus de textos als quals aplicar l'eina Texty i per a poder fer l'estudi amb la comoditat del laboratori, es va fer una rèplica no pública del web d'InformationR. S'ha pogut manipular, així, amb tota llibertat els textos i el codi HTML de cada article.

A continuació es van tractar tots els textos dels articles: captura de llistat de fitxers; cada article es troba en un fitxer HTML. En el nom d'aquest fitxer (paperXXX.html) trobem números naturals que comencen per l'1 i al final del

2010 arriben al 452 i que responen a un ordre cronològic de publicació. Dins de cada número de la revista els números tenen menys sentit, doncs l'ordre és un ordre editorial.

3.3.4- Creació dels texty

El procés de creació d'un texty no és complicat, tot i amb això, ha anat evolucionant i simplificant-se molt.

En la primers versió texty usava el *framework javascript jquery*¹⁵ que recollia les coordenades de les paraules del text en el navegador i pantalla de l'usuari. Al costat del servidor Image Magick¹⁶ creava les imatges PNG i amb PHP les imatges vectorials SVG.

Actualment l'estratègia es més directa: primer es parseja el text HTML i se li aplica un CSS que fa el text blanc i les paraules dels vocabularis i el seu fons del color corresponent de cada vocabulari. I s'aplica a cada text el programari html2image¹⁷. Finalmet amb Inage Magick (mogrify) ajustem les mides del texty final (300x450px)

El sistema que es proposa és no intrusiu i és un possible camí per a implantar texty en una web open access.

En el nostre estudi hem creat textys amb 5 colors. A l'hora d'escollir els colors s'ha tingut en compte les principals restriccions que es solen recomanar per aquest tipus d'atributs gràfics: per una banda, utilitzar colors bàsics que el cervell humà pugui distingir amb facilitat¹⁸. I, per un altre, que com a humans, no recordem més de 9 colors diferents amb significats diferents¹⁹.

Cal aturar-se un moment per analitzar la informació continguda en les zones en blanc d'un texty. Donat que texty és una representació física de dades, és a dir, que els punts de colors apareixen en posicions relatives a les posicions reals de determinats termes en el text, l'absència de tinta dóna informació rellevant sobre el text representat.

Considerant la teoria de E. Tufte sobre el rati de Tinta i Dades²⁰, per a texty s'hauria de considerar que hi ha "dades sense tinta". L'absència de colors significa una menor densitat de termes dels vocabularis emprats. Si considerem les zones blanques com a zones amb dades, la fórmula de Tuffte pel cas de texty quedaria

15. <http://jquery.com/>

16. <http://www.imagemagick.org>

17. HTML2Image For Linux and Unix <http://www.guangmingsoft.net/htmlsnapshot/html2image.htm>

18. Kay, Paul. 1969 Basic Color Terms: Their Universality and Evolution. Berkeley: University of California Press y de Ware, Collin (2004). Information Visualization: Perception for Design. San Francisco: Morgan Kauffman

19. Few (2006). Show me the numbers. Oakland: Analytics Press

20. REF: The Visual Display of Quantitative Information. Edward R. Tufte. 2001 (p.93)

$$\text{Data-ink ratio} = \frac{100}{100} = 1$$

Fig 5. Tufte's data-ink ratio equation

Donant el màxim de proporció de tinta dedicada a representar dades.









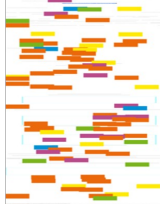


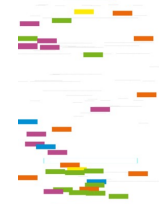





3.3.5- Exemple aplicat a un número de la revista

Hem produït textys per a tots els articles del Journal Information Research, des del Volum 1, #1 (1995/96) fins el Volum 15 #4, amb un total de 454 textys²¹. I presentem aquí una comparativa de tots els textys del Volum 15, No 4 December 2010 del Journal Information Research²².

Presentem els 17 articles del volum 15, no 4 Desembre 2010 del *Information Research Jour*

21. <http://o.subvideo.tv> user=texty / password=texty

22. <http://informationr.net/ir/15-4/infres154.html>

<p>1. Proceedings of ISIC</p>  <p><i>Cultural differences in the health information environments and practices between Finnish and Japanese university students</i> Graeme Baxter, Rita Marcella and Laura Illingworth</p>	<p>2. Proceedings of ISIC</p> <p><i>Organizational information behaviour in the public consultation process in Scotland</i> Leanne Bowler</p> 	<p>3. Proceedings of ISIC</p>  <p><i>Talk as a metacognitive strategy during the information search process of adolescents</i> Jenny Bronstein</p>	<p>4. Proceedings of ISIC</p>  <p><i>Selecting and using information sources: source preferences and information pathways of Israeli library and information science students of your paper.</i> Donald O. Case</p>
<p>5. Proceedings of ISIC</p>  <p><i>A model of the information seeking and decision making of online coin buyers</i> Kreetta Askola, Toshimori Atsushi and Maija-Leena Huotari</p>	<p>6. Proceedings of ISIC</p>  <p><i>Local versus global information relevance in Website use: a case study with the information literacy portal AlfinEEES .Francisco Javier García Marco and María Pinto</i></p>	<p>7. Proceedings of ISIC</p>  <p><i>Information behaviour research and information systems development: the SHAMAN project, an example of collaboration</i> Elena Maceviciute and T.D. Wilson</p>	<p>8. Proceedings of ISIC</p>  <p><i>Avoiding health information in the context of uncertainty management</i> Anu Sairanen and Reijo Savolainen</p>
<p>9. Proceedings of ISIC</p>  <p><i>A study of labour market information needs through employers' seeking behaviour</i> Sonia Sanchez-Cuadrado, Jorge Morato and Yorgos Andreiadakis</p>	<p>10. Proceedings of ISIC</p>  <p><i>'Information in context': co-designing workplace structures and systems for organizational learning</i> Mary M. Somerville and Zaana Howard</p>	<p>11. Proceedings of ISIC</p>  <p><i>"We have a lot of information to share with each other". Understanding the value of peer-based health information exchange</i> Tiffany C. Veinot</p>	<p>12. Proceedings of ISIC</p>  <p><i>Information sharing: an exploration of the literature and some propositions</i> T.D. Wilson</p>
<p>13. Proceedings of ISIC</p>  <p><i>Applying McKenzie's model of information practices in everyday life information seeking in the context of the menopause transition</i> Alison Yeoman</p>	<p>14. Regular article</p>  <p><i>Double or nothing: is redundancy of spatial data a burden or a need in the public sector of Uganda?</i> Walter T. de Vries and Beatrice Winnie Nyemera</p>	<p>15. Regular article</p>  <p><i>Analysis of automatic translation of questions for question-answering systems</i> Lola García-Santiago and María-Dolores Olvera-Lobo</p>	<p>16. Regular article</p>  <p><i>Dietary blogs as sites of informational and emotional support</i> Reijo Savolainen</p>
<p>17</p>  <p><i>Information and information science: an address on the occasion of receiving the award of Doctor Honoris Causa, at the University of Murcia, 30 September, 2010.</i> T.D. Wilson</p>			

Una primera ullada ens diu (veure figures 6 i 7) el següent:

- El to predominat en aquesta *issue* és experimental, si bé seguit d'a prop per l'enfoc qualitatiu.
- Els articles 3 i 13 tenen un caire experimental molt clar, si bé també els 4, 11, 14 tenen un to experimental.
- 5 dels 17 articles (38.5%) tenen presència remarcable de tecnologia amb ordinadors.
- Els articles amb una càrrega conceptual més gran són el 7 i el 9, si bé també tenen contingut conceptual els 8, 15 i 16.
- L'article amb una metodologia quantitativa remarcable és el 15.

Aquí podem veure com Texty pot ser útil per a l'exploració i la navegació dels textos abans de llegir-los. Es proposa usar texty en l'index de cada *issue* i saber quin tipus d'article ens trobarem abans de llegir-lo.

3.3.5.1- Conclusions

En les conclusions, primerament, presentem les comparatives que hem fet entre texty i dues tècniques gràfiques tradicionals, com són el diagrama de barres i el diagrama de línies.

Demostrem que texty aporta claredat, velocitat i visió de l'estructura del l'article en qüestió.

Aquests resultats són extrapolables a qualsevol sistema de localització/recuperació d'informació que gestioni una col·lecció de documents textuais.

3.3.6- Estudi per a 1 texty

3.3.6.1 Comparativa de texty amb diagrama de barres:

Hem comparat la representació texty amb els diagrames de barres. La llegenda de colors per als cinc vocabularis descrits és:

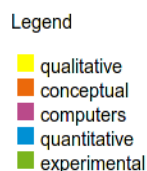


Fig 7. Llegenda de vocabularis

Per il·lustrar tot això hem escollit articles del Volum 15, número 4 de desembre del 2010 que és el volum més nou en el moment de començar aquesta investigació.

Presentem aquí algunes comparacions de les dues tècniques destacant per a cada cas els avantatges que aporta texty.

- Cas 1, paper 441: A study of labour market information needs through employers' seeking behaviour. Sonia Sanchez-Cuadrado; Jorge Morato, and Yorgos Andreadakis and Jose Antonio Moreiro²³

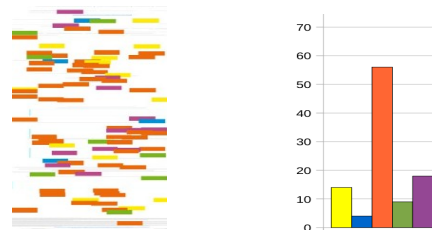


Fig 8: texty i diagrama de barres per a l'article 441 (InformationR)

Ambdós mètodes permeten identificar ràpidament el vocabulari amb més presència. En aquest cas el "conceptual". En aquest article es descriuen *knowledge representation techniques* amb suport computacional. Això ens ho mostren també ambdues representacions, però amb Texty, a diferència del diagrama de barres, s pot veure que es parla d'aquestes tècniques a la part mitja de l'article (color violeta).

- Cas 2, paper 445: Information behaviour research and information systems development: the SHAMAN project, an example of collaboration. Elena Maceviciute and T.D. Wilson²⁴

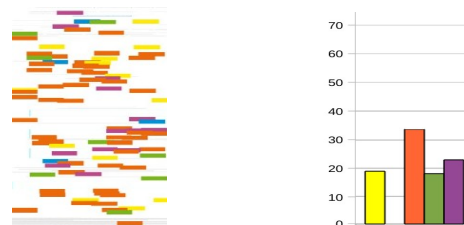


Fig 9: texty i diagrama de barres per a l'article 445 (InformationR)

Observant el texty d'aquest article podem aventurar-nos a dir que aquest article té un to conceptual. Inicialment, en el *background* sobre *long-term digital preservation*, podem dir que es parla de tècniques que requereixen ordinadors (per exemple: "e-mail, word-processed documents and spreadsheets, but e-books, sound recordings, films, scientific data sets, social science data archives". En la part mitja de l'article veiem una concentració de punts vers pertanyents al vocabulari "experimental". Això coincideix amb l'explicació de les dades usades pel programa SHAMAN a partir d'entrevistes a usuaris. Total aquesta informació no es pot deduir a partir del diagrama de barres.

23. <http://informationr.net/ir/15-4/paper441.html>

24. <http://informationr.net/ir/15-4/paper445.html>

- Cas 3, paper 450: Analysis of automatic translation of questions for question-answering systems. Lola García-Santiago and María-Dolores Olvera-Lobo²⁵

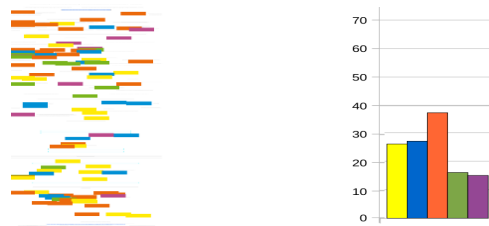


Fig 10: texty i diagrama de barres per a l'article 450 (InformationR)

En aquest cas tenim un article amb presència considerable dels 5 vocabularis. Aquí s'aprecia, potser amb més claredat, la importància de poder veure la distribució física de termes en l'article. Així podem dir que l'article comença amb un to "conceptual" per a després explicar el mètode, de to "experimental". L'article no requereix massa coneixements informàtics, tot hi que n'hi ha referències a la primera meitat de l'article. Al final hi ha referències de caire "conceptual". En general l'article fa un *approach* qualitatiu, doncs el color groc es reparteix per tot l'article. De nou, tota aquesta informació és impossible extreure-la a partir del diagrama de barres.

3.3.6.2 Comparativa de texty i diagrama de línies:

Comparem ara texty amb un diagrama de línies on:

$$\text{numero termes} = F(\text{línies de text: } 0, 1, \dots, N).$$

- Cas 4, paper 438: Dietary blogs as sites of informational and emotional support, by Reijo Savolainen²⁶

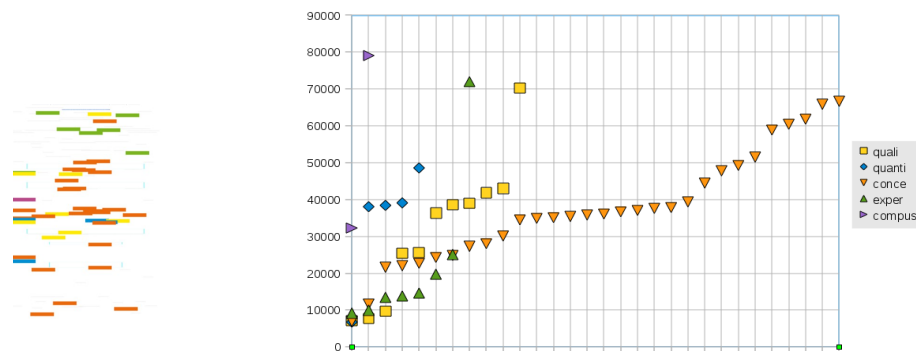


Fig 11. texty i diagrama de línies per a l'article 438 (InformationR)

El diagrama de línies representa a l'eix Y la posició del terme respecte al número de caràcters total del text; en aquest cas, l'article té 81767 caràcters. L'eix X representa el número de termes per a cada vocabulari i evita

25. <http://informationr.net/ir/15-4/paper450.html>

26. <http://informationr.net/ir/15-4/paper438.html>

l'overlapping que faria intel·ligible el diagrama.

La lectura del diagrama de línies ens aporta més informació sobre l'estructura i la distribució de termes en el text que el diagrama de barres, però encara no arriba a la senzillesa de Texty. La lectura del diagrama de línies és més feixuga i requereix d'un entrenament superior en temps al que texty requereix.

També s'observa que per a veure de quina manera es barregen els vocabularis cal superposar mentalment els punts, reduint el recorregut de l'eix X, cosa que no és intuïtiva i que també requereix d'un entrenament previ.

Per a textos amb números més alts de termes, *l'overlapping* pot ser un gran problema en l'ús de diagrames de línies

3.3.6.3 Conclusions de la comparació entre texty i els diagrames de barres i de línies:

Tant texty com els diagrames mostren la quantitat de termes de cada vocabulari, és a dir, l'enfoc general d'un article en un cop d'ull. Les principals millores diferenciadores que introdueix texty són:

- 1- Texty diu a on apareix cada terme (a l'inici, al mig, al final)
- 2- Amb texty es pot veure l'estructura conceptual de l'article; p.e. a l'inici hi ha una explicació conceptual i després es desenvolupa la part experimental i finalment els càlculs a on hi ha un ús intens de tecnologia i ordinadors).
- 3- Amb texty es pot veure l'estructura física de l'article (es poden identificar taules i llistats).

Queda demostrat que un diagrama de barres o un de línies mai ens pot donar aquestes informacions que texty aporta.

4. Estudi proposat

Un cop fet l'estudi tècnic de texty hem començat a estudiar possibles aplicacions de texty a altres camps del coneixement. Relacionant aquest treball d'investigació amb el departament a on se'm permet estudiar, hem escollit textos legals de sentències judicials; textos que aconsegueixen tenir un mateix registre i que els advocats i fiscals solen haver de cercar.

Aquest estudi suggereix la importància potencial de texty també en aquest camp. De fet, actualment estem en contacte amb l'editorial Aranzadi en espera de mostrar-los texty i veure si tenen interès en aplicar-lo a la seva àmplia base de dades.

4.1 Textos legals: sentències

De la pràctica quotidiana d'advocats i fiscals en la cerca de sentències surt aquest estudi proposat per a una futura aplicació de texty a documents legals.

4.1.1 Fonts de cerca de sentències i millores que hi pot aportar texty

Què ens donen els sistemes de recuperació d'informació sobre sentències?

Hem provat Aranzadi²⁷

I també VLEX²⁸ que permet cercar sentències directament, amb un sistema de navegació per arbre de categories amb 2 nivells de profunditat.

Mostra dels resultats de la recerca

La imatge mostra dues captures de pantalla de resultats de cerca. La part esquerra és una captura de la pàgina de resultats de 'BUSCADOR' d'Aranzadi, que mostra una llista de resultats amb títols i descripcions breus de sentències. La part dreta és una captura de resultats de VLEX, que mostra una llista de sentències amb títols, dates i estrelles de classificació.

Fig. 12: Captures de resultats de "sentencia" a Aranzadi (esq.) I de llistat de sentències categoria Dret Penal > Faltes a VLEX (dreta)

A la Fig. 12 esq. mostrem una recerca de la paraula "sentencia" al cercador general d'Aranzadi. Aranzadi ofereix serveis de pagament i segurament és possible cercar només en sentències. És una cerca general i trobem molts documents que contenen la paraula cercada, incloent-hi notícies. Per tant és una cerca massa poc específica i la variabilitat de tipus i registres de text que obtenim no ens sembla millorable amb texty, que només ha estat comprovat per textos d'un mateix registre.

Millora que aporta texty:

Texty podria millorar aquests llistats incorporant un texty al costat de cada ítem dels resultats. (Fig. 13)

27. <http://www.aranzadi.es/index.php/informacion-juridica/jurisprudencia-actual/civil/sentencia-del-juzgado-de-lo-mercantil-n-1-de-bilbao-de-26-enero-2010>

28. http://vlex.com/search/index?query%5Bcoleccion_id%5D=2&query%5Bvoz_id%5D=951744



[Sentència núm. 28/2011 de TS, Sala 2ª, de lo Penal, 26 de gener de 2011](#)

★ ★ ★

DELITO DE AGRESIÓN SEXUAL. PRESUNCIÓN DE INOCENCIA. VALORACIÓN DE LA PRUEBA. Las preguntas transcritas por la defensa reflejan que la iniciativa del Presidente sólo se orientaba a puntualizar algunas de las respuestas ofrecidas por los testigos a las preguntas formuladas por el Ministerio Fiscal. Ni las cuestiones planteadas a la testigo -la médico de la residencia geriátrica en la que se hallaba la víctima-, las que fueron dirigidas a la cuidadora, ni, en suma, las aclaraciones solicitadas a la doctora -médico psiquiatra del acusado-, sugieren una toma de postura del Tribunal a favor de la acusación, ni un prejuicio anticipado acerca de la autoría del acusado. No se hace lugar al recurso de casación.

Fig. 13: Simulació d'un texty representant la sentència descrita textualment extreta de VLEX.

Al punt següent discutim quines característiques seria útil representar, és a dir, quines aspectes pot representar cada color del texty.

Mostra d'una fitxa de sentència:

The screenshot shows a legal database interface. On the left, there is a sidebar with a search bar and a list of categories, including 'CIVIL'. The main content area displays the details of a case: 'Sentencia nº 28/2011 de TS, Sala 2ª, de lo Penal, 26 de Enero de 2011'. The case summary includes the ponente (MANUEL MARCHENA GÓMEZ), the number of the appeal (1798/2010), and the procedure (RECURSO CASACIÓN). Below the summary, there is a section for 'Resumen' (Summary) which repeats the text from Figure 13. There is also an 'Extracto' (Extract) section. On the right side of the interface, there is a yellow box with a document icon and the text 'Accede a este documento y prueba vLex GRATIS durante 3 días'. Below this, there is a search bar for 'Email' and a link to 'acceda aquí' if the user is a vLex client. At the bottom, there are social media sharing options for Facebook and Twitter.

Fig. 14: Captures de la presentació de cada sentència; la fitxa per una sentència a Aranzadi (esq.) i VLEX (dreta)

Millora que aporta texty:

Texty també pot millorar la navegació per la sentència, un cop la tenim localitzada i decidir si val la pena fer-li una ullada. A la Fig. 15 mostrem com un simple texty-scroller pot proporcionar un texty sensible als clics que permet desplaçar-se pel text.



Fig. 15: un texty amb scroll per a navegar el text representat. El texty es sensible als clics que ens porten a punts del document.

4.1.2 Mètode de la investigació

La manera de treball ha estat partir de tres entrevistes amb advocades en actiu que treballen a l'estat espanyol, especialistes en dret penal i dret laboral.. Amb aquestes entrevistes hem volgut saber:

- Amb quina assiduitat cerqueu en sentències? Quant de temps us consumeix?
Per a preparar un judici sempre es revisen sentències, per tant, quasi diàriament.
- Com de llargues són les sentències mireu?
Varia molt però podem dir entre 3 i 80 pàgines.
- Què voleu saber sobre la sentència?
La majoria de preguntes son bàsicament booleanes, és a dir que tenen dues respostes possibles. Son del tipus:
 - Qui guanya?
 - Hi ha lesions? Greus o lleus?
 - Cas consumat?
 - Hi ha armes?

Quins paràmetres no booleanes podem identificar?

Aquest és el punt crític, doncs amb texty podem representar característiques generals, i les advocades entrevistades no estan habituades a poder cercar d'aquesta manera. Els hem preguntat si els aniria bé veure certes paraules pertanyents a grups semàntics; també hem proposat representar característiques físiques del text, com les seccions o les referències.

D'aquestes entrevistes concluïm que un bon conjunt de característiques que seria interessant representar usant la tècnica texty, és:

Físics:

- referències a altres textos
- Dates
- Llocs, indrets: països, ciutats, pobles, carrers

Semàntics:

- Violència / lesions
- Finances / diners
- Immigració / papers / fronteres

També proposem incorporar separadors de les seccions estàndards de les sentència. Normalment:

1. Fets provats
2. Fonaments del dret
3. Fallo

Per a representar els paràmetres booleans incorporem una banda superior amb un codi de colors i lletres. Per a una proposta acabada, caldria internacionalitzar aquests indicadors, segurament amb l'ús d'icones simples.

Finalment el resultat del prototipus de texty aplicat a sentències judicials es pot veure a la Fig 16.



Fig. 16: un texty prototipus per a la representació d'una sentència. Les barres horitzontals mostren la separació de le seccions de la sentència. La banda superior indica que ell resultat es d'innocent (bola verda), que no s'ha consumat (C vermella), que hi ha armes (A verda), que hi ha lesions (L verda(i que el cas es de l'any 2007 a Espanya.

Naturalment, desenvolupar un text adaptat a les sentències judicials requeriria de la col·laboració estreta amb professionals que estiguin interessats/des en aquesta idea.

5. Conclusions

La feina diària amb informació digitalitzada necessita de millores. Cal que les màquines treballin més al nostre servei. En el principi de l'era de les dades obertes, texty és una eina que, per un costat, contribueix a la recerca en aquesta direcció. I, per l'altre, es mostra com una eina amb una gran potència d'utilitat en diversos camps.

6. Futur de texty

Com hem dit, texty no és un substitut dels sistemes de cerca clàssics sinó que és proposa com a complement. És exportable a d'altres fons i amb altres vocabularis adequats a cada cas. Texty és pot personalitzar fins el punt que ens mostri uns o altres vocabularis (colors) en funció del lector, dels textos representats o de la intenció en cada cas.

Com dèiem, les llegendes dinàmiques i els vocabularis personalitzats poden augmentar l'eficàcia de texty, així com usar diferents capes per a representar, segons es desitgi, uns o altres vocabularis.

L'ús d'imatges interactives (sensibles a clics de ratolí) permeten usar texty per a desplaçar-se pel text representat.

L'adaptació de texty per a textos en múltiples llengües també és una implementació possible; només cal disposar de traduccions dels vocabularis.