Visualització i exploració de textos

Una aproximació conceptual i una proposta tecnològica per a la recuperació d'informació textual

Projecte de tesi doctoral

UNIVERSITAT DE BARCELONA FACULTAT DE BIBLIOTECONOMIA i DOCUMENTACIÓ
DOCTORAT EEES: H0J01 INFORMACIÓ i DOCUMENTACIÓ EN LA SOCIETAT DEL CONEIXEMENT

Curs Acadèmic 2013/2014

Doctorand: Jaume Nualart i Vilaplana

Director de tesi: Dr.Mario Pérez-Montoro. Universitat de Barcelona

Data de lliurament: 28 de Maig del 2014

Índex

1 Introducció **3**

2 Estat de la qüestió **3**
2.1 Visualització de dades, un camp jove 4
2.2 Visualització de textos 7
2.3 Anàlisi de textos 8
2.4 Conclusions de la revisió de la literatura 9

3 Objectius de la recerca **10**
3.1 Elaboració d'un estat de l'art en la visualització i exploració de textos. 11
3.2 Estudi per a textos individuals 11
3.3 Estudi per a col·leccions de textos 12

4 Metodologia **12**
4.1 Elaboració d'un estat de l'art en la visualització i exploració de textos. 12
4.2 Estudi per a textos individuals 13
4.3 Estudi per a col·leccions de textos 13

5 Resultats provisionals **14**
5.1 Elaboració d'un estat de l'art en la visualització i exploració de textos. 14
5.2 Estudi per a textos individuals 14
5.3 Estudi per a col·leccions de textos 14

6 Calendari de treball **14**

7 Bibliografia **16**

1 Introducció

La visualització de textos és un camp que avança ràpidament i diversificada. Ràpidament, perquè es nodreix de fonts creixents de documents tipus text, com són els portals de dades obertes i la proliferació de continguts accessibles via APIs. Diversificadament perquè el camp s'ha desenvolupat paral·lelament en una àmplia gamma de disciplines. A més, s'estan començant a consolidar punts de trobada per a la comunitat de visualització de dades tant en publicacions com en conferències (taules 1, 2 i 3).

En aquest projecte de recerca sostenim que la visualització i exploració de textos és un subcamp de la visualització de dades; i que està alimentada pel creixent avanç en l'anàlisi de textos, i per la creixent quantitat de dades accessibles en format text. És per això que pretenem mostrar que hi ha molt terreny per córrer en el desenvolupament i l'ús d'eines de visualització i exploració específiques per a treballar amb textos.

Conceptualment, aquest projecte de recerca se centra en dos aspectes articulats: d'una banda proposarem un marc teòric de classificació d'eines de visualització i exploració de textos que permeti un diàleg formal entre investigadors i una gramàtica específica per a comparar eines i mètodes de visualització i exploració. D'altra banda, aquest projecte investiga les possibilitats pràctiques en el desenvolupament d'eines per a la visualització i exploració de textos i per això es crearan eines que segueixin el marc teòric definit i que mostrin les àmplies oportunitats de desenvolupament d'aquest camp.

2 Estat de la qüestió

Donada la interdisciplinarietat del camp d'estudi, per recollir els casos que presentem hem hagut de buscar en diferents contextos i fonts: des de les ciències fins a les humanitats, des de revistes acadèmiques fins a blocs, des d'informes d'universitats fins a estudis de freelance, des d'institucions de dades obertes fins a comunitats de dades obertes. Camps diferents implica diferents filosofies i punts de vista.

Per tot això, aquesta revisió té com a objectiu ajudar les persones, no només del món acadèmic, que treballen amb dades i, especialment, amb textos, utilitzant tècniques de visualització. La ciència de les dades és també l'art de detectar patrons, conductes i evidències en la representació de la realitat, millorant, així, la forma, la velocitat i la claredat amb què es mostren i es descobreixen fets ocults, a simple vista, dins les dades.

S'ha recollit quaranta-nou casos per a revisar. Una vegada recollits i estudiats, els hem dividit en dos grans grups:

- Representació de textos individuals: en particular, tècniques per extreure el significat de cada text basades en l'estil d'escriptura, l'estructura del document i el registre de llenguatge, en lloc de les tècniques basades en estadístiques simples. Estem interessats en la representació del significat

de textos perquè una visualització adequada pot accelerar i/o millorar la selecció i la gestió personal de textos. L'avanç en camps com natural language procesing (NLP), la lingüística computacional i machine learning ofereixen tècniques per representar textos complexos amb dades d'alta qualitat. Proposem que es combinin aquestes tècniques amb les visualitzacions adequades per tal de millorar la forma d'examinar i comprendre textos.

- Representació de col·leccions de textos: explorar, seleccionar, navegar i analitzar col·leccions de textos és una tasca diària per a moltes persones que treballen amb ordinadors i dades. La recuperació d'informació és un punt crític en un entorn d'excés d'informació (Baeza-Yates et al., 1999). Quan un usuari realitza una cerca, els sistemes de recuperació d'informació responen amb una llista de resultats. En molts casos, la presentació dels resultats pot jugar un paper important en la satisfacció de les necessitats d'informació de l'usuari. Una presentació dolenta o inadequada pot obstaculitzar la satisfacció de les necessitats d'informació (Baeza-Yates et al., 2011). Normalment, els sistemes de recuperació d'informació presenten els resultats d'una consulta com una llista plana, d'una dimensió. i aquestes llistes solen ser opaques quant a l'ordre, és a dir, els usuaris no saben per què la llista té un ordre particular.

Nosaltres sostenim que les tècniques per representar col·leccions de textos com a resultat de cerques poden contribuir a millorar la navegació, l'exploració i la recuperació d'informació.

2.1 Visualització de dades, un camp jove

Avui dia la visualització de dades és un camp acadèmic consolidat.(Strecker and International Development Research Centre (IDRC), 2012). A continuació presentem una breu llista dels principals arguments i referències que defineixen aquesta disciplina relativament nova.

- Set de les deu primeres universitats en el rànquing Times Higher Education (TSL Education Ltd., 2012), tenen departaments o grups de recerca relacionats amb la visualització de dades. La visualització de dades s'ha desenvolupat en una àmplia gamma de departaments, des de ciència i estadística, fins a informàtica i lingüística; des de disseny Gràfic i química fins física, genètica i història. Recentment la visualització de dades s'ha convertit en un camp diferent, amb programes de màster i doctorat propis i departaments dedicats (taula 1).

- Les conferències dels últims cinc anys classificades per nombre de participants i dedicades principalment a la visualització de dades (taula 2).

- Les revistes especialitzades (taula 3).

institució	rank 2012	Department/Curs	URL
Harvard University	1	Broad Institute of Harvard and MIT	http://www.broadinstitute.org/vis
Massachusetts Institute of Technology	2	Broad Institute of Harvard and MIT	http://www.broadinstitute.org/vis
University of Cambridge	3	—	—
Stanford University	4	Stanford Vis Group	http://vis.stanford.edu/
University of California, Berkeley	5	VisualizationLab	http://vis.berkeley.edu/
University of Oxford	6	Visual Informatics Lab at Oxford	http://oxvii.wordpress.com/
Princeton University	7	PrincetonVisLab	http://www.princeton.edu/researchcomputing/vis-lab
University of Tokyo	8	—	—
University of California, Los Angeles	9	IDRE GIS and visualization	https://idre.ucla.edu/visualization
Yale University	10	—	—

Table 1: Principals universitats i departaments de visualització de dades

Conferència	Lloc	Temàtica	Participants	URL
NICAR 2013	USA	data journalism	149	http://ire.org/conferences/nicar-2013/
dd4d 2009	France	information visualization	52	http://www.dd4d.net
FutureEverything 2013	UK	Technology / society / art	52	http://futureeverything.org/
resonate 2013	UK	Creative code	44	http://www.thisisresonate.co.uk/resonate-13/
graphical web 2012	Switzerland	open web / datavis	38	http://www.graphicalweb.org/2012/
IEEEVis - VisWeek 2012	USA	information visualization	-	http://ieeevis.org/
EuroVis 2013	Germany	Computational Aesthetics	-	http://www.eurovis2013.de
Siggraph 2013	USA	computer graphics and interactive techniques	-	http://s2013.siggraph.org
OzViz 2012	Australia & NZ	workshops for visualisation practitioners, academics and researchers		http://www.ozviz2012.org/

Table 2: Conferències dedicades principalment a la visualització ordenada pels participants (font: Stefaner, M., 2013)

Nom	URL
Parsons Journal for Information Mapping	http://pjim.newschool.edu/issues/index.php
Journal of Visualization	http://springer.com/materials/mechanics/journal/12650
IEEE Transactions on Visualization and Computer Graphics (TVCG)	http://www.computer.org/portal/web/tvcg
Information Visualization	http://ivi.sagepub.com/
International Journal of Image Processing and Data Visualization (IJIPDV)	http://iartc.net/index.php/Visualization
IEEE Vis (Former Visweek)	http://ieeevis.org/
EuroVis	http://www.eurovis2013.de/
ACM CHI	http://chi2013.acm.org/
EG CGF	http://www.eg.org
IVS	http://www.graphicslink.co.uk/IV2013/

Table 3: Les publicacions més importants dedicades a visualització de dades.

- Finalment volem destacar el paper de difusió realitzat per importants llocs web. Aquest és el cas de infosthetics, visualcomplexity i visualizingdata.com.

2.2 Visualització de textos

Segons la classificació de Shneiderman (Shneiderman, 1996), els textos són dades unidimensionals. Un text és una dada seqüencial que sòl anar de dreta a esquerra o d'esquerra a dreta i línia per línia, de dalt a baix. No obstant això, un text pot tenir altres estructures, per exemple, segons la morfologia pot tenir paràgrafs, frases i paraules. Segons l'estructura de la informació pot ser ordenat per capítols, parts, seccions, subseccions, etc. Si el text té un format com HTML, llavors pot ser ordenat per una jerarquia d'etiquetes (¡body¿, ¡div¿, ¡p¿, etc). En aquests exemples, el text inclou estructures d'arbre juntament amb l'estructura unidimensional abans citada. A més, és habitual que els textos comportin una estructura abstracta que és difícil d'analitzar per als ordinadors. Aquests diferents tipus de dades en un mateix text, mostren les especificitats que els textos tenen com a estructures de dades que són.

La majoria dels casos revisats no representen les dades en brut, és a dir, el text tal com és, sinó que el text es divideix en trossos més petits, normalment extraient una part representativa del text. Tal procés és un procés de transformació de dades i succeeix, per exemple, quan un text es redueix a una llista

de paraules d'acord amb la seva freqüència. En aquest cas, el mètode escollit per representar les dades pertany a una família de mètodes que s'adapta a aquest tipus de dades, no específic a textos. En aquesta revisió repassarem les estratègies més referenciades per representar textos o col·leccions de textos, amb especial atenció a les estratègies que busquen representar textos rics en complexitats, irregularitats i abstraccions.

Existeix certa controvèrsia en la consideració de la visualització de textos com un subcamp específic de la visualització de dades: Illinski (Noah Iliinsky, 2013) afirma que el text per si sol, no pot ser considerat com un tipus de dades. Silić (Silic and Basic, 2010) diu que "el text no estructurat no és adequat per a la visualització". De fet, com s'ha dit més amunt, la majoria de les visualitzacions de textos transformen les dades originalment de tipus text o no-estructurades en un nou conjunt de dades estructurades i reduïdes respecte a l'original. Aquest nou conjunt de dades ja no és unidimensional, sinó que està ordenat per categories o com a dades de xarxa. El grup de dades resultant es pot representar amb una àmplia gamma d'eines no específiques a la representació de textos (Hearst, 2009, Grobelnik and Mladenic, 2002).

En qualsevol cas, és un fet que la quantitat de dades a les quals tenim accés creix dia a dia, i la major part d'aquestes dades està en format text. Viegas i Wattenberg en una entrevista amb Heer, argumenten: "Una de les coses que crec que és realment prometedor és la visualització de textos. Això ha estat ignorat fins ara, en termes d'eines de visualització d'informació i, en canvi, una gran quantitat de la informació més rica a la qual tenim accés està en format text" (Heer, 2010).

Hem comentat que l'anàlisi de textos és un camp clau per a la visualització de textos. A continuació presentem un breu comentari sobre aquest camp i la seva relació amb la visualització de textos.

2.3 Anàlisi de textos

L'anàlisi de dades és el limitant o coll d'ampolla de la visualització de dades. L'avanç en l'anàlisi de textos comporta, a diversos nivells, la comprensió del text per part dels ordinadors i la modificació del text original. És per aquestes raons que la visualització no pot anar més enllà dels resultats de l'anàlisi quant a l'etiquetatge o codificació del text original.

L'anàlisi de textos, és gairebé un sinònim de mineria de textos (Feldman and Sanger, 2006), i també és un camp interdisciplinari que inclou la recuperació d'informació, la mineria de dades en general, machine learning, l'estadística, la lingüística i natural language processing. Segons M. Hearst (Hearst, M., 2003) en la mineria de textos, l'objectiu és detectar informació en els textos desconeguda fins a aquest moment, informació que ningú sabia encara i que, per tant, no es podria haver escrit encara.

Revisar gramaticalment del text, l'anomenada mineria de textos, és un subcamp de la mineria de dades, les aplicacions típiques de la qual són: analitzar o comparar textos literaris, analitzar seqüències de dades de la biologia i la

genètica i, més recentment, modelar patrons de comportament dels consumidors o descobrir el frau en l'ús de targetes de crèdit. Hearst diferencia aquests casos de les pures operacions d'extracció d'informació, com són l'extracció de noms de persones, adreces o habilitats professionals. En aquest tipus de tasques s'obté un 80% de precisió. En canvi, en el primer grup de casos definits més amunt, la interpretació completa del llenguatge natural mitjançant un programa informàtic -continuant amb Hearst- sembla que no serà possible durant "una llarga temporada".

Per tant, per estudiar la visualització i l'exploració de textos és tan important seguir la literatura de visualització de dades, així com la literatura d'anàlisi de textos. Tots dos camps s'interrelacionen. D'una banda, l'anàlisi de text pot limitar les possibilitats de visualització i la interacció amb el text. D'altra banda, les tècniques de visualització milloren els resultats obtinguts amb l'anàlisi, quant a usabilitat i interacció. A més, hi ha una forta evidència empírica que les persones aprenen millor amb text i visualitzacions que només amb text (Anglin et al., 2004, Levie and Lentz, 1982).

2.4 Conclusions de la revisió de la literatura

Certament, la diversitat d'enfocaments des de diferents disciplines i publicacions, representa un repte per fer un estudi complet de l'estat de l'art del camp de la visualització i exploració de textos. Alguns dels casos estudiats han estat trobats en publicacions molt específiques, per exemple Joel Deshaye i Peter Stoicheff i les seves obres en les visualitzacions de novel·les de Faulkner (Deshaye and others.). En la lectura de les notes de Stoicheff es pot observar que van desenvolupar les visualitzacions només per a ajudar en un estudi molt específic de narrativa relacionada amb línies de temps de William Faulkner. No hi ha referències a les aplicacions d'aquestes idees interessants a altres textos, el que suggereix que hi ha d'haver més obres ocultes en les profunditats d'altres camps.

La visualització de textos, tal com la plantegem en en aquesta recerca, pot ser considerada un subcamp de la visualització de dades; tanmateix, els límits del camp de vegades no estan clarament definits. Aquest és el cas de Search Clock (Harrison, 2008) , en què els textos representats provenend'una enorme base de dades de consultes en els motors de cerca. Pot, aquest conjunt de dades, ser considerat una col·lecció de textos si cada un d'ells, en la majoria dels casos, és només una o dues paraules? Existeix una longitud mínima d'un text a ser considerat com a text? Per aquest cas vam decidir tractar-lo com una col·lecció de textos, curts, però, en última instància, textos.

Una decisió important en aquesta revisió ha estat la classificació dels casos trobats. Atès que hi ha pocs treballs que revisen únicament visualització de textos, hem hagit de fer referència a les revisions clàssiques de visualització de dades (Shneiderman, 1996), així com d'altres més recents (Collins et al., 2009). En aquests casos, les classificacions es basen en la resolució de tasques mitjançant la visualització de dades, i no tant en els aspectes explícits de la

visualització. És per això que hem decidit proposar una classificació pròpia basada en característiques visuals.

Aquesta és la llista dels punts destacats i deficiències que hem trobat revisant la literatura Transformar en text:

- La visualització de textos individuals s'aplica principalment a textos literaris. La literatura, a més de complexes combinacions de paraules, conté alts nivells d'abstracció humana i estructures lliures i experimentals. Potser seria més eficaç aplicar tècniques de visualització a altres tipus de textos amb un registre més formal i/o amb un esquema predefinit i un vocabulari molt conegut, com ara: textos legals, articles científics, textos i comunicacions basades en plantilles, etc.

- Hi ha molt pocs casos de visualització de text individual en parts que sigui també seqüencial (Wattenberg, 2002). La majoria de les visualitzacions de parts de textos extreuen l'essència del text en funció d'alguns criteris i la seqüència original del text es perd. Tot i que la visualització seqüencial té alguns avantatges, sembla que hi ha espai per desenvolupar estratègies de visualització de parts de textos que mantenen la seqüència de text original.

- Les visualitzacions de col·leccions de textos utilitzen mètodes que es poden trobar a la visualització de dades en general. Aquesta idea convida a l'experimentació en portar mètodes de visualització de dades més estàndards i enfocaments per al subcamp específic de la visualització de textos.

- Col·leccions de textos amb agregats és la categoria que ha desenvolupat dissenys i idees més específiques. Cal més investigació per tal de trobar alguns patrons en aquest tipus de visualització.

3 Objectius de la recerca

Com s'ha assenyalat anteriorment, l'objectiu general d'aquest projecte de recerca és introduir una sèrie de tècniques i propostes conceptuals en el camp de la visualització i exploració de dades textuals per a millorar el treball tant amb textos individuals com amb col·leccions de textos.

Més concretament, pretenem fer contribucions tant en l'àmbit teòric com en el pràctic d'aquest camp. En el camp teòric proposarem una classificació de tècniques de visualització i exploració de textos basada en la parametrització visual de les dades. En el camp pràctic, volem fer dos estudis sobre les dues categories de visualització i exploració de textos que proposem: una que representi textos individuals i una altra que representi col·leccions de textos.

Per a estudiar aquests dos tipus de dades hem escollit un context i una eina per a cadascun. El context que hem escollit com a proposta de recerca és el dels resultats dels sistemes de recuperació d'informació. Volem contribuir a

demostrar que la forma en què es presenta la informació en els estàndards de recuperació d'informació textual i documental és millorable amb eines complementàries de visualització de dades. Per a l'estudi de textos individuals desenvoluparem una eina específica i per a l'estudi de col·leccions de textos en desenvoluparem una altra. Ambdues eines les hauré creades jo mateix i es publicaran amb llicències lliures. Pensem que d'aquesta manera serà més fàcil l'adaptació de les eines als requeriments d'aquesta recerca.

Els objectius, per tant, queden agrupats en tres:

3.1 Elaboració d'un estat de l'art en la visualització i exploració de textos.

En aquesta fase recollim casos de visualització i exploració de textos i elaborarem una revisió de les estratègies seguides en cada cas, classificant-les i comparant-les.

L'objectiu d'aquesta revisió de l'estat de l'art en eines de visualització i exploració de textos és doble. Per un costat reunir els treballs més destacats en aquest camp per tal de donar-los a conèixer i poder compartir referents amb la comunitat. Per l'altre costat, proposarem la classificació de casos exposada més amunt, amb l'objectiu de facilitar la comunicació entre investigadors, així com també contribuir a la universalització i expansió del camp de la visualització de dades.

Per presentar els casos estudiats de manera acurada, es proposarà una classificació original dels casos revisats a partir de llurs característiques visuals, com a resultat d'un procés inductiu d'anàlisi. Com a base de la classificació proposem agrupar els casos en les dues categories citades: visualització de textos individuals i visualització de col·leccions de textos. Coherents amb el projecte, els casos es podran explorar i comparar mitjançant una eina de visualització i exploració en línia.

3.2 Estudi per a textos individuals

En aquesta fase volem saber si una eina de visualització de textos individuals pot aportar un valor afegit, als llistats de resultats dels sistemes de recuperació d'informació.

Per a respondre a aquesta qüestió, es desenvoluparà una eina complementària a la representació tradicional unidimensional de resultats de sistemes de recuperació d'informació. Cal que aquesta eina representi les essències dels continguts de cada ítem d'una llista retornada per un sistema de recuperació d'informació i ajudar, així, l'usuari a poder identificar el contingut més adient per a satisfer la seva necessitat d'informació abans d'abordar intel·lectualment cadascun dels resultats. Dit en paraules planeres, volem fer una representació gràfica d'un text donat que permeti fer una molt ràpida lectura visual en diagonal del text en qüestió.

A nivell gràfic ens influencien les tècniques usades per Keim (Keim and Oelke, 2007) en reconeixement d'autories, molt relacionat amb la identificació de plagis, usant estratègies de visualització de textos. La tècnica usada per Keim té una aplicació molt diferent de la que cerquem, doncs s'hi representa l'allargada de les frases de cada text. I això es fa amb petits quadrats de diversos grups de color. En el cas que ens ocupa volem mostrar la densitat de conceptes.

3.3 Estudi per a col·leccions de textos

En aquesta fase volem saber si una eina de visualització i exploració de col·leccions de textos pot proporcionar informació valuosa per als usuaris d'un sistema de recuperació d'informació aplicat a un arxiu digital determinat.

Per tal de demostrar això, com s'ha comentat més amunt, es desenvoluparà una eina de visualització i exploració capaç de donar una visió conjunta d'un d'arxiu digital. Una visió, tant de les dimensions de les dades, com de les categories presents en les dades.

L'eina ha de permetre el filtratge de dades i la representació de diferents paràmetres.

4 Metodologia

Donades les tres fases del projecte, cal descriure diferents metodologies per a assolir els objectius definits en cada cas.

4.1 Elaboració d'un estat de l'art en la visualització i exploració de textos.

Aquesta part conté dos objectius: recollir i fer una anàlisi dels casos més destacats de visualització i exploració de textos. I proposar una classificació d'aquests casos. Per tal d'assolir aquests objectius s'ha seguit una metodologia inductivista. Així, el procés de raonament pel qual s'arriba a cada conclusió bé de la generalització d'aspectes trobats en cada cas estudiat.

Les tasques que farem per a assolir aquests objectius, per ordre, són:

- Recollida de casos. Degut a la novetat d'aquest camp i a la interdisciplinarietat, els casos s'han trobat en diferents disciplines i, per tant, diferents revistes, conferències i arxius web.
- Estudi de cada cas: revisarem cada cas recollint dades i normalitzant-les per tal de poder-les comparar.
- Agrupació de casos similars. Observació de les característiques compartides entre casos. Observació de possibles jerarquies de característiques.

- Creació de la classificació de tècniques de visualització i exploració de textos: partint dels dos tipus base de visualitzacions i exploracions de textos segons el tipus de dades, agruparem els casos en característiques visuals i en la natura de les dades representades.
- Anàlisi de cada categoria basada en els casos estudiats.
- Representació dels casos amb una eina en línia per tal de poder navegar-los i tenir-ne una visió de conjunt.
- Resultats i conclusions de l'experiència.

4.2 Estudi per a textos individuals

Les tasques que farem per a estudiar aquest tipus de visualitzacions, per ordre, són:

- Desenvolupament i adaptació d'una eina de programari de visualització de textos individuals
- Aplicació del programari a un cas real. Elecció de textos per a representar: accessibles i amb llicències que en permetin l'ús.
- Estudi de l'eina desenvolupada com a base teòrica per a generalitzar i enunciar els principis de la visualització de textos individuals. Estudi de les característiques cognitives i psicològiques de les tècniques usades en l'eina.
- Validació de l'eina per comparació amb tècniques clàssiques i referencials de representació de dades.
- Comparació d'interfícies: amb i sense l'eina de visualització
- Resultats i conclusions de l'experiència.

4.3 Estudi per a col·leccions de textos

Les tasques que farem per a estudiar aquest tipus de visualitzacions, per ordre, són:

- Desenvolupament i adaptació d'una eina de programari de visualització i exploració de col·leccions de textos.
- Estudi de les característiques que aporta l'eina respecte al sistema tradicional tipus text de representació d'arxius digitals.
- Aplicació del programari a un cas real. Elecció d'una col·lecció de textos accessibles.
- Avaluació d'usabilitat de l'eina: qüestions que es volen investigar, disseny de l'experiment d'avaluació, proves a usuaris i anàlisi de dades.
- Resultats i conclusions de l'experiència.

5 Resultats provisionals

El llistat de resultats i productes de la recerca són:

5.1 Elaboració d'un estat de l'art en la visualització i exploració de textos.

El principal resultat de la revisió de l'estat de l'art d'estratègies per a la visualització i exploració de textos serà un document que llisti els casos més destacats dels darrers anys, així com les referències a casos i autors clàssics. Donada la novetat d'aquest subcamp, la classificació de visualització de textos que proposarem formarà també part dels resultats d'aquesta recerca que presentem.

5.2 Estudi per a textos individuals

Es publicarà una versió completa d'un paquet de programari amb llicència lliure per tal que tothom que hi tingui interès pugui usar-la i modificar-la.

I també es publicarà un estudi de l'eina desenvolupada, aportant una anàlisi teòrica de les característiques i de l'adaptació de l'eina a un entorn real de textos. Amb una comparació de l'eina amb eines clàssiques de representació de dades.

5.3 Estudi per a col·leccions de textos

Es publicarà una versió completa d'un paquet de programari amb llicència lliure per tal que tothom que hi tingui interès pugui usar-la i modificar-la.

Es publicarà una avaluació d'usuaris per a comprovar les possibilitats de l'eina desenvolupada aplicada a una col·lecció real de textos.

6 Calendari de treball

Curs 2013-14

- Estudi de l'estat de l'art de la visualització i exploració de textos (Revisió de la literatura)
- Desenvolupament de l'eina de visualització de textos individuals
- Elaboració i presentació del projecte de recerca de tesi doctoral
- Desenvolupament de l'eina de visualització de col·leccions de textos
- Avaluació de l'eina de col·leccions de textos

Curs 2014-15

- Estudi sobre visualització de textos individuals
- Estudi sobre visualització de col·leccions de textos
- Elaboració i presentació de l'informe d'activitat de recerca
- Procés administratiu
- Escriure i presentar resultats de la tesi
- Dipositar la tesi

Curs 2015-16

- Exposició de la tesi
- Procés administratiu

7 Bibliografia

Gary J Anglin, Hossein Vaez, and Kathryn L Cunningham. Visual representations and learning: The role of static and animated graphics. *Handbook of research on educational communications and technology*, 2:865–916, 2004. 9

Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999. 4

Ricardo Baeza-Yates, Andrei Z Broder, and Yoelle Maarek. The new frontier of web search technology: seven challenges. In *Search computing*, pages 3–9. Springer, 2011. 4

Christopher Collins, Sheelagh Carpendale, and Gerald Penn. Docuburst: Visualizing document content using language structure. In *Computer Graphics Forum*, volume 28, pages 1039–1046. Wiley Online Library, 2009. 9

Muri Deshaye and others. The sound and the fury: a hypertext edition. URL `http://www.usask.ca/english/faulkner`. 9

Ronen Feldman and James Sanger. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press, 2006. 8

M. Grobelnik and D. Mladenic. Efficient visualization of large text corpora. In *Proceedings of the seventh seminar. Dubrovnik, Croatia*, 2002. URL `http://ailab.ijs.si/dunja/SiKDD2002/papers/GrobelnikSep02.pdf`. 8

Chris Harrison. Search clock, 2008. URL `http://www.chrisharrison.net/index.php/Visualizations/SearchClock`. 9

Marti a Hearst. Search user interfaces. *Search User Interfaces*, 54(Ch 1):404, November 2009. ISSN 00010782. doi: 10.1145/2018396.2018414. URL `http://searchuserinterfaces.com/book/`. 8

Hearst, M. What is text mining? http://people.ischool.berkeley.edu/˜hearst/text-mining.html, 2003. URL `http://people.ischool.berkeley.edu/{~}hearst/text-mining.html`. 8

Jef Heer. A conversation with jeff heer, martin wattenberg, and fernanda viegas, 2010. 8

DA Keim and D Oelke. Literature fingerprinting: A new method for visual literary analysis. *Visual Analytics Science and Technology, \ldots*, 2007. URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4389004`. 12

W Howard Levie and Richard Lentz. Effects of text illustrations: A review of research. *ECTJ*, 30(4):195–232, 1982. 9

Noah Iliinsky. *Choosing visual properties for successful visualizations*. s IBM Software - Business Analytics, 2013. URL http://public.dhe.ibm.com/common/ssi/ecm/en/ytw03323usen/YTW03323USEN.PDF. 8

Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pages 336–343, 1996. 7, 9

Artur Silic and Bojana Dalbelo Basic. Visualization of text streams: A survey. In Rossitza Setchi, Ivan Jordanov, RobertJ. Howlett, and LakhmiC. Jain, editors, *Knowledge-Based and Intelligent Information and Engineering Systems*, volume 6277 of *Lecture Notes in Computer Science*, pages 31–43. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-15389-1. doi: 10.1007/978-3-642-15390-7_4. URL http://dx.doi.org/10.1007/978-3-642-15390-7_4. 8

Stefaner, M. Gender balance visualization. http://moritz.stefaner.eu/projects/gender-balance/, 2013. URL http://moritz.stefaner.eu/projects/gender-balance/. 6

Jacqueline Strecker and International Development Research Centre (IDRC). *Data visualization in review: summary*. IDRC, Ottawa, ON, 2012. 4

TSL Education Ltd. World university rankings 2012-2013 - times higher education. http://www.timeshighereducation.co.uk/world-university-rankings/2012-13/world-ranking, 2012. URL http://www.timeshighereducation.co.uk/world-university-rankings/2012-13/world-ranking. 4

Martin Wattenberg. Arc diagrams: Visualizing structure in strings. In *Information Visualization, 2002. INFOVIS 2002. IEEE Symposium on*, pages 110–116, 2002. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1173155. 10