ARTÍCULOS

CÓMO DIBUJAMOS TEXTOS. REVISIÓN DE PROPUESTAS DE VISUALIZACIÓN Y EXPLORACIÓN TEXTUAL

Jaume Nualart-Vilaplana, Mario Pérez-Montoro y Mitchell Whitelaw

Note: This article can be read in its original English version on:
http://www.elprofesionaldelainformacion.com/contenidos/2014/may/02.pdf

Jaume Nualart-Vilaplana es doctorando en la *Faculty of Arts and Design*, *University of Canberra* (Australia), ingeniero de investigación en el *Nicta* (Australia), y doctorando en la *Facultat de Biblioteconomia i Documentació* de la *Universitat de Barcelona*. Posee DEA y licenciatura por la por la *Universitat Autònoma de Barcelona*
http://orcid.org/0000-0003-4954-5303

Machine Learning Research Group at NICTA, Canberra Research Laboratory
Tower A, 7 London Circuit, Canberra City ACT 2601, Canberra, Australia
jaume.nualart@canberra.edu.au

Mario Pérez-Montoro es doctor en filosofía y educación por la *Universitat de Barcelona*, master en gestión y sistemas de información de la Universitat Politècnica de Catalunya. Estudió en el *Istituto di Discipline della Comunicazione* de la *Università de Bologna* (Italia) y ha sido profesor visitante en el *Center for the Study of Language and Information* (*CSLI*) de la *Stanford University* (California, EUA) y en la *School of Information* de la *UC Berkeley* (California, EUA). Es profesor en el *Dept. de Ciències de la Informació* de la *Univ. de Barcelona*. Su trabajo se ha centrado en la arquitectura de información y visualización. Es autor del libro *Arquitectura de la información en entornos web* (*Trea*, 2010).
http://orcid.org/0000-0003-2426-8119

Facultat de Biblioteconomia i Documentació, Universitat de Barcelona
Melcior de Palau, 140. 08014 Barcelona, España
perez-montoro@ub.edu

Mitchell Whitelaw es profesor, escritor y profesional de los nuevos medios de arte y cultura, especialmente sistemas generativos y de estética de datos. Su trabajo ha aparecido en revistas como *Leonardo*, *Digital creativity*, *Fibreculture* y *Senses and society*. Su trabajo en *a-life art* se publicó en el libro *Metacreation: art and artificial life* (MIT Press, 2004). Su trabajo actual abarca el arte generativo y el diseño, la materialidad digital y la visualización de datos. Es profesor asociado en la *Faculty of Arts and Design* de la *University of Canberra*, donde dirige el *Master of digital design*. Bloguea en *The teeming void*.
http://orcid.org/0000-0001-9013-9732

Faculty of Arts and Design, University of Canberra
Bldg, Floor & Room: 9, C12. ACT 2617, Canberra, Australia
mitchell.whitelaw@canberra.edu.au

Resumen

En este trabajo se presenta una revisión de estrategias para la visualización y exploración de textos, argumentando que constituye un subcampo de la visualización de datos que se nutre de los avances en el análisis de textos y de la creciente cantidad de datos accesibles en formato texto. Proponemos una clasificación original para un total de cuarenta y nueve casos revisados. La clasificación está basada en las características visuales de cada caso, identificadas mediante un proceso inductivo de análisis. Agrupamos los casos (publicados entre 1994 y 2013) en dos categorías: las visualizaciones de texto individuales y la visualizaciones de colecciones de textos. Los casos revisados pueden ser explorados y comparados online.

Palabras clave

Visualización de texto, Visualización de datos, Exploración de datos, Visualización de información, Análisis de textos.

Title: How we draw texts: a review of approaches to text visualization and exploration

Artículo recibido el 19-01-2014
Aceptación definitiva 09-03-2014

Abstract

This paper presents a review of approaches to text visualization and exploration. Text visualization and exploration, we argue, constitute a subfield of data visualization, and are fuelled by the advances being made in text analysis research and by the growing amount of accessible data in text format. We propose an original classification for a total of 49 cases based on the visual features of the approaches adopted, identified using an inductive process of analysis. We group the cases (published between 1994 and 2013) in two categories: single-text visualizations and text-collection visualizations, both of which can be explored and compared online.

Keywords

Review, Text visualization, Data visualization, Data exploration, Data display, Information visualization, Text analysis.

Nualart-Vilaplana, Jaume; Pérez-Montoro, Mario; Whitelaw, Mitchell (2014). "Cómo dibujamos textos. Revisión de propuestas de visualización y exploración textual". *El profesional de la información*, mayo-junio, v. 23, n. 3, pp. 221-235.

http://dx.doi.org/10.3145/epi.2014.may.02

1. Introducción

El objetivo de esta revisión es aportar el contexto adecuado y proponer una clasificación de las herramientas de visualización y exploración de textos. Se enumeran, clasifican y analizan los trabajos más relevantes aparecidos en el campo de la visualización y la exploración de textos entre 1995 y 2013. Este es un campo que avanza rápido y de forma diversificada. Rápido, porque se nutre de las iniciativas de datos abiertos y *web scraping* (extracción de datos de webs) y de forma diversificada porque tradicionalmente el campo se desarrolla paralelamente en una amplia gama de disciplinas (figura 1). Además, están empezando a consolidarse puntos de encuentro dentro de la comunidad de visualización de datos tanto en publicaciones como en conferencias y encuentros científicos (ver tablas 1, 2 y 3).

Por este motivo, para recoger los casos que se presentan se ha buscado en diferentes contextos y fuentes: desde las ciencias experimentales hasta las humanidades, desde revistas académicas hasta blogs, desde universidades hasta estudios *freelance*, desde instituciones que ofrecen datos abiertos hasta comunidades que ceden sus datos abiertos. Diferentes campos implica diferentes filosofías y puntos de vista.

Esta revisión tiene como objetivo ayudar a las personas que trabajan con datos (no sólo del mundo académico), y en especial con textos utilizando técnicas de visualización. Técnicas que permiten detectar patrones, conductas y/o evidencias en la representación la realidad, mejorando, así, la forma, la velocidad o la claridad con la que se muestran o se descubren hechos ocultos en los datos.

Figura 1. Profesiones de los inventores de algunos métodos de investigación

Es difícil definir una línea conceptual clara que separe la visualización de la exploración de datos. En este trabajo se revisan herramientas que permiten ambas operaciones, tanto de forma independiente como simultánea. Sin embargo, por economía del lenguaje, a veces nos referiremos simplemente a visualización de textos incluyendo los dos ámbitos.

Los casos estudiados se dividen en dos grandes grupos:

1) Representación de textos individuales: en particular, técnicas para extraer el significado de cada texto basadas en el estilo de escritura, la estructura del documento y el registro de lenguaje; frente a las técnicas basadas en simples estadísticas. Estamos interesados en la representación del significado de textos porque una visualización adecuada puede acelerar y/o mejorar la selección y la gestión personal de textos. El avance de campos como el *natural language processing* (NLP), la lingüística computacional y *machine learning* ofrecen técnicas para representar textos complejos con datos de alta calidad. Proponemos que se combinen estas técnicas con las visualizaciones adecuadas para, así, mejorar la forma de examinar y comprender textos.

2) Representación de colecciones de textos: explorar, seleccionar, navegar y analizar colecciones de textos es una tarea diaria para muchas personas que trabajan con ordenadores y datos. Hay espacio para llevar a cabo nuevas herramientas e ideas. La recuperación de información es un punto crítico en un entorno de exceso de información (**Baeza-Yates** *et al*., 1999). Cuando un usuario realiza una búsqueda, los sistemas de recuperación de información responden normalmente con una lista de resultados. En muchos casos la presentación de los resultados puede jugar un papel importante en la satisfacción de las necesidades de información de ese usuario. Una presentación mala o inadecuada puede obstaculizar la satisfacción de esas necesidades de información (**Baeza-Yates** *et al*., 2011). Normalmente, los sistemas de recuperación de información

Tabla 1. Principales universidades y departamentos de visualización de datos

Institución	Puesto en 2012	Departamento/Curso	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	--	--

Tabla 2. Conferencias dedicadas principalmente a la visualización, ordenada por nº de participantes (**Stefaner**, 2013)

Conferencia	Localización	Tema	Nº de participantes	URL
Nicar 2013	EUA	Periodismo de datos	149	*http://ire.org/conferences/nicar-2013*
Dd4d 2009	Francia	Visulaización de información	52	*http://www.dd4d.net*
FutureEverything 2013	Reino Unido	Tecnología/sociedad/arte	52	*http://futureeverything.org*
Resonate 2013	Reino Unido	Código creativo	44	*http://www.thisisresonate.co.uk/resonate-13*
Graphical web 2012	Suiza	Web abierto/visualización de datos	38	*http://www.graphicalweb.org/2012*
IeeeVis - VisWeek 2012	EUA	Visualización de información	-	*http://ieeevis.org*
EuroVis 2013	Alemania	Estética computacional	-	*http://www.eurovis2013.de*
Siggraph 2013	EUA	Gráficos por ordenador y técnicas interactivas	-	*http://s2013.siggraph.org*
OzViz 2012	Australia y NZ	Workshops for visualisation practitioners, academics and researchers	-	*http://www.ozviz2012.org*

presentan los resultados de una consulta como una lista plana, de una dimensión, que suele ser opaca en términos de orden, es decir, los usuarios no saben por qué la lista tiene un orden particular. Para refinar su búsqueda, los usuarios tienen que interactuar de nuevo, normalmente filtrando la primera salida del resultados. Proponemos que las nuevas técnicas para representar colecciones de textos (resultados de búsquedas incluidos) pueden contribuir a mejorar la navegación, exploración y recuperación de información.

Hoy en día la visualización de datos puede considerarse como un campo académico consolidado (**Strecker**; *IDRC*, 2012). Argumentos que definen esta disciplina:

- -Siete de las diez primeras universidades en el *ranking Times Higher Education* (2012), tienen departamentos o grupos de investigación relacionados con la visualización de datos. Esta disciplina se ha desarrollado en una amplia gama de departamentos, desde las ciencias o la estadística, hasta informática o lingüística, o desde diseño gráfico o química hasta la física, la genética o la historia. Recientemente la visualización de datos se ha convertido en un campo independiente, con programas de master y doctorado, y departamentos centrados en esa disciplina (tabla 1).
- Existe un volumen importante de conferencias en los últimos cinco años dedicadas principalmente a la visualización de datos (tabla 2).
- Las contribuciones más importantes de visualización de datos se concentran en congresos y reuniones científicas, pero también se publican algunas revistas especializadas (tabla 3).
- Por último queremos destacar el papel de difusión realizado por importantes sitios web. Este es el caso, entre otros, de *Infosthetics*, *Visualcomplexity* o *Visualizingdata.com*.

1.1. Visualización de textos

Shneiderman (1996) clasifica los textos normales como datos unidimensionales. Un texto es un dato secuencial que va de derecha a izquierda o de izquierda a derecha y línea por línea, de arriba abajo. Sin embargo, un texto puede tener múltiples estructuras internas, por ejemplo, según la morfología puede tener párrafos, frases y palabras. Según la estructura de la información, un texto puede estar orde-

Tabla 3. Publicaciones más importantes dedicadas a visualización de datos.

Nombre	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*

nado por capítulos, partes, secciones, subsecciones, etc. Si el texto tiene un formato como html, entonces puede estar ordenado por las etiquetas <body>, , <p>, etc. En estos ejemplos el texto incluye estructuras de árbol junto con la estructura unidimensional. Además los textos pueden tener un componente subjetivo con una estructura abstracta que es difícil de analizar por los ordenadores. Estos diferentes tipos de datos en un mismo texto, muestran las especificidades que los textos tienen como estructuras de datos.

La cantidad de datos a los que tenemos acceso crece día a día. Y la mayor parte de estos datos está en formato de texto. **Fernanda Viégas** y **Martin Wattenberg** argumentan en una entrevista con **Jeff Heer**: "Una de las cosas que creo que es realmente prometedora es la visualización de textos. Esto ha sido ignorado hasta ahora, en términos de herramientas de visualización de información y, sin embargo, una gran cantidad de la información más rica que tenemos está en formato texto" (**Heer**, 2010).

> "Siete de las diez mejores universidades tienen departamentos o grupos de investigación en visualización de datos"

El análisis de datos ha definido los límites de la visualización de datos, esto es, la fina capa que separa las múltiples verdades de las mentiras que de esconden en los datos. En el caso de la visualización de textos, esta función la realiza la lingüística computacional, *natural language processing*, *machine learning* y la estadística. El avance en el análisis de textos conlleva, en varios niveles, la comprensión del texto por parte de los ordenadores, la modificación, en definitiva, del texto original, también llamado dato no-estructurado (ver apartado "Análisis de texto").

Existe cierta controversia en la consideración de la visualización de textos como un subcampo específico de la visualización de datos. Algunos autores no lo ven de esta manera. Así, por ejemplo, **Illinski** (2013) afirma que el texto por sí solo no puede ser considerado como un tipo de datos. En la misma línea, **Šilić** (2010) dice que "el texto no estructurado no es adecuado para la visualización". De hecho, como se mencionó anteriormente, la mayoría de las visualizaciones de textos transforman los datos originalmente de tipo textual o no estructurados en un nuevo conjunto de datos estructurados y reducidos respecto al original. Este nuevo conjunto de datos ya no es unidimensional, sino que está ordenado por categorías o en red. El grupo de datos resultante se puede representar con una amplia gama de herramientas no específicas a la representación de textos (**Hearst**, 2009; **Grobelnik**; **Mladenić**, 2002).

La mayoría de los casos revisados no representan los datos en bruto, es decir, el texto tal como aparece originariamente. En su lugar, el texto se divide en piezas de datos más pequeñas, normalmente extrayendo una parte representativa del texto. Se trata de un proceso de transformación de datos y se implementa, por ejemplo, reduciendo un texto a una lista de palabras de acuerdo con su frecuencia. En este caso, el método elegido para representar los datos pertenece a una familia de métodos que se adaptan a este tipo de datos, no específico para textos. En esta revisión se repasan las estrategias más referenciadas para representar textos o colecciones de textos, con especial atención a las estrategias que buscan representar textos ricos en complejidades, irregularidades y abstracciones.

1.2. Análisis de textos

El análisis de textos, entendido en cierta manera como sinónimo de la disciplina de la minería de textos (**Feldman**; **Sanger**, 2006), es un campo interdisciplinario que incluye la recuperación de información, minería de datos en general, *machine learning*, estadística, lingüística y *natural language processing*. Según **Marti Hearst** (2003), en la minería de textos el objetivo es detectar información representada en los textos y hasta ahora desconocida, información que nadie sabía aún y que, por tanto, no se podría haber descrito todavía.

La minería de textos es un subcampo de la minería de datos cuyas aplicaciones típicas son, entre otras, analizar o comparar textos literarios, analizar secuencias de datos de biología y genética o, más recientemente, descubrir patrones en el comportamiento de los consumidores o el fraude en el uso de tarjetas de crédito. **Hearst** diferencia estos casos de las puras operaciones de extracción de información, como son la extracción de nombres de personas, direcciones o habilidades profesionales. En este tipo de tareas se obtiene un 80% de precisión, pero en el primer grupo de casos definidos anteriormente, la interpretación completa del lenguaje natural mediante un programa informático parece que no será posible durante "una larga temporada" (**Hearst**, 2003).

Para estudiar la visualización y exploración de textos es importante seguir la bibliografía de visualización de datos, así como la de análisis de textos. Ambos campos se interrelacionan. Por un lado, el análisis de textos puede limitar las posibilidades de visualización y la interacción del propio texto. Por otro, las técnicas de visualización mejoran los resultados obtenidos con el análisis haciéndolos usables e interactivos. Además, existe una fuerte evidencia empírica de que las personas aprenden mejor con texto y visualizaciones que sólo con texto (**Anglin** *et al*., 2004; **Levie**; **Lentz**, 1982).

2. Revisión

En esta sección se presenta una posible clasificación de las herramientas de visualización y exploración de texto basada en la características visuales de cada caso y como fruto de un proceso de análisis inductivo de 49 casos. Después de la clasificación presentamos el análisis de los casos agrupados en categorías y subcategorías.

2.1. Clasificación de los casos

El nivel superior de la clasificación de visualizaciones de textos según el tipo de datos está formada por dos grandes categorías:

1) Textos individuales: donde el texto es una secuencia de palabras ordenadas de acuerdo con la jerarquía: documento > párrafos > frases > signos de puntuación > palabras > sílabas y fonemas. En los casos en que el texto es un libro, es posible tener más granularidades, como, por ejemplo, capítulos > secciones > subsecciones > etc. También tenemos los metadatos del propio texto u otros textos adjuntos, que pueden incluir títulos, autores, notas editoriales, notas de copyright, notas de

agradecimiento, dedicatorias, prefacios, tablas de contenido, glosarios, índices alfabéticos, bibliografía, entre otros.

2) Colecciones de textos: grupos de textos en los que cada elemento es una entidad claramente diferenciada. En general, cuando hablamos de colecciones de textos nos referimos a textos que tienen cierta similitud, ya sea de registro, de longitud o de estructura. Todos los casos revisados son colecciones de la misma clase de textos. También existen referencias de colecciones heterogéneas de textos (**Meeks**, 2011), sobre todo para ofrecer un análisis representativo de un área de conocimiento, en cuyos casos el objetivo de la colección es incluir la máxima variedad de expresiones y un amplio vocabulario. En esos casos el conjunto de datos es heterogéneo tanto por estructura como por registro.

A partir de estos dos tipos de datos se han añadido varias subdivisiones subjetivas para cada caso según las características visuales representadas en cada tipo revisado. El objetivo de esta clasificación es mostrar las características clave de cada visualización de texto revisada.

Textos individuales
- Todo <-> Parte
- Secuencial <-> No-secuencial
- Estructura del discurso <-> Estructura sintáctica
- Resultados de una búsqueda
- Línea de tiempo

Colecciones de textos
- Items <-> Agregados
- Capa de datos tipo *landscape*
- Resultados de una búsqueda
- Línea de tiempo

2.1.1. Visualizaciones de textos individuales

Clasificamos los casos utilizando tres indicadores: qué parte del texto está representada, cómo está representada la secuencia del texto y, finalmente, cuál de las estructuras del texto se utiliza en la representación de cada caso.

¿Todo el texto o sólo una parte?

En algunos casos, una parte del texto se considera esencial y se utiliza como materia prima para la visualización de todo el texto.

Sin embargo, hay representaciones donde participa todo el texto. Estos casos a menudo representan la totalidad del texto de forma implícita. Ejemplos de este tipo son, entre otros:
- capítulos de libro, pero no todo el texto;
- todas las frases del texto como líneas de color;
- verbos presentes, que representan el estilo del texto;
- personajes de una novela y su aparición en el texto;
- lugares o fechas presentes.

Cuando se representa de manera explícita el conjunto del texto son, por razones obvias, casos con textos cortos, como canciones, discursos y poemas.

En algunos casos, como en *Radial word connections* (caso 1) sólo se representan determinadas palabras, sin embargo, clasificamos este caso como una representación de todo el texto, porque cada capítulo de la novela se representa implícitamente a lo largo del círculo.

Los casos en los que está representada la totalidad del texto, implícitamente o no, como un elemento principal de la visualización, se han clasificado como la visualización de todo el texto.

¿Sigue la secuencia de texto?

¿Tiene la visualización la misma secuencia que el texto original y, por tanto, nos referimos al mismo orden que el texto original? Si es así, el caso será considerado secuencial, de lo contrario lo llamamos no-secuencial. Un caso típico que no sigue la secuencia del texto original es una nube de etiquetas o palabras (figura 1).

> La mayoría de las visualizaciones transforman textos no estructurados en un reducido conjunto de datos estructurados

¿Se representa la estructura del discurso o la estructura sintáctica?

Un texto puede tener dos tipos de estructura útiles para esta investigación. Hay una estructura subjetiva respecto al punto de vista del autor: estructura del discurso. En lingüística, discurso es un concepto amplio. Aquí lo usamos para referirnos a las partes de un texto y el esquema de un documento, como partes, capítulos, secciones o subsecciones. La estructura del discurso es ampliamente utilizada en la visualización de textos, ya que es una manera fácil de representar la secuencia de texto.

La segunda estructura que tenemos en cuenta es la sintáctica. El adjetivo "sintáctica" también puede hacer referencia a varios conceptos, pero aquí lo utilizamos para referirnos a oraciones, frases, y palabras como verbos, sustantivos y morfemas. Se trata de una estructura objetiva y, a diferencia de la estructura del discurso, no depende de la decisión del autor sino de las reglas de la lingüística. En la visualización de textos, elementos de esta estructura como las oraciones son muy comunes.

2.1.2. Visualizaciones de colecciones de texto

Clasificamos los casos según dos pares de indicadores: se representan items o agregados compuestos, y se representan los datos puros o hay una capa de datos de tipo *landscape* (paisaje).

Los elementos de la colección, ¿están diferenciados o se representan como agregados? ¿Cómo está representado gráficamente cada elemento de la colección? ¿Cada texto está representado como una entidad gráfica, es decir, como un punto o una palabra o frase corta? ¿Los elementos de la visualización pueden ser contados? ¿Se diferencian visualmente?

Hay casos en los que cada elemento no está representado por una entidad gráfica separada, sino, por ejemplo, por una zona de color. En otros casos los elementos se agrupan y se muestran como distribuciones de frecuencia. Cuando los elementos de la colección no están diferenciados gráficamente, es decir, no son contables, entonces visualmente hablamos de agregados o, en general, de una visualización de agregación en lugar de una visualización de items.

¿Se trata de datos puros o son un conglomerado de datos más una segunda capa tipo landscape?

¿Hay algún contexto gráfico que acompaña a los elementos de la colección? ¿Hay algún otro conjunto de datos, relacio-

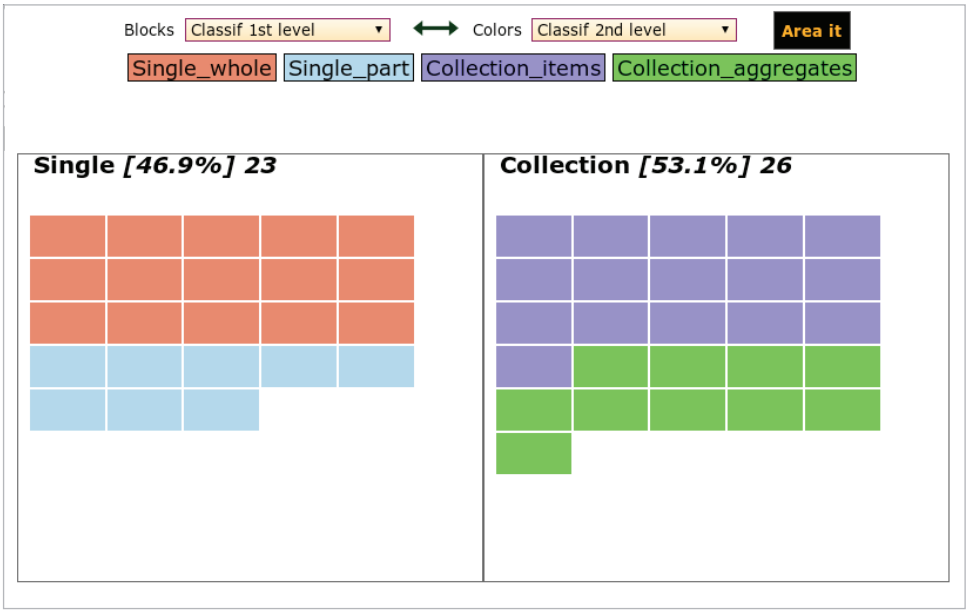

Figura 2: Los 49 casos visualizados con un software llamado *AREA* (captura de pantalla)

nados con el texto, también representados? Algunos casos presentan los items inmersos en un entorno gráfico utilizando, por ejemplo, un mapa. Este contexto puede ser un mapa geográfico real, una metáfora o, por ejemplo, un paisaje conceptual compuesto por palabras que forman una segunda capa que complementa la capa de datos de la colección donde cada distancia juega un papel: documento-documento (similitud entre documentos), palabra-documento (importancia de una palabra en un documento) o palabra-palabra (similitud entre las palabras de la colección).

Escalas y ejes no son considerados como parte de la capa tipo *landscape* de datos, como tampoco lo son los elementos de la interfaz de la representación. Esta capa de datos, al no considerarse como parte del conjunto principal de datos, reduce sustancialmente la ratio de tinta empleada para representar los datos, siguiendo la propuesta de **Tufte** (**Tufte**; **Graves-Morris**, 1983), en comparación con la representación de datos puros.

2.1.3. Visualizaciones combinadas de textos individuales y colecciones

Las propiedades aplicables de forma combinada a visualizaciones de textos individuales y a colecciones de texto que proponemos son las líneas temporales de tiempo, los resultados de una búsqueda y el tamaño del conjunto de datos.

¿Hay una línea temporal?

¿Cambian los textos en el tiempo? Hay un conjunto de casos que muestran los cambios de un conjunto de datos a través del tiempo. La mayoría de los enfoques de este tipo se han planteado en informática para representar la evolución del código de un programa o en *Wikipedia*, mostrando una serie de aspectos de las revisiones de cada artículo en el tiempo.

También se incluyen en esta categoría las visualizaciones en las que el conjunto de datos a través del tiempo varían. Éste es el caso de la visualización de últimas noticias, donde el conjunto de datos (las noticias) crece en volumen con el tiempo.

¿Se trata de una lista de resultados de una búsqueda?

Las visualizaciones de resultados de los sistemas de recuperación de información son un tipo de representaciones que se caracterizan por el número cambiante de items representados, que depende directamente del número de resultados obtenidos en la búsqueda. Éste es un subcampo creciente dentro de la visualización de datos y está relacionado con las disciplinas de sistemas de información y recuperación de información (**Mann**, 2002; **Hearst**, 2009).

¿Es válido para conjuntos de datos pequeños o grandes?

Es raro que una herramienta de visualización sea independiente del tamaño del conjunto de datos que se representan. En los casos en los que la herramienta está diseñada claramente para un tamaño específico del conjunto de datos el lector encontrará una explicación al respecto.

2.2. Análisis de casos de visualización

Se han revisado 49 casos utilizando la clasificación propuesta anteriormente. Tratando de cubrir los aspectos más importantes de la visualización de textos, esta revisión se centra en las ideas específicas para la visualización, en lugar de en los datos y el contexto de cada caso.

Cada caso se presenta con nombre, nombre corto, autor(s), año de publicación, URL para más información, descripción de los datos, disciplina relacionada con el trabajo, descripción del método de visualización, descripción general del caso, captura de pantalla, clasificación (individual o colección), clasificación (individual-todo el texto, individual-parte del texto, colección-items, colección-agregado), clasificación (línea temporal), clasificación (resultados de búsqueda), clasificación (conjunto de datos pequeño, conjunto de datos grande, N/A).

Los casos se agrupan en dos secciones y cuatro subsecciones:

Visualización de textos individuales (23 casos):
- Visualización de todo el texto (15 casos)
- Visualización de parte del texto (8 casos)

Visualización de colecciones de textos (26 casos):
- Colecciones de items (16 casos)
- Colecciones de agregados (10 casos)

Para cada apartado los casos están ordenados por año de publicación (descendente). Con el fin de ayudar al lector, esta colección se puede explorar con una visualización también incluida en la revisión, llamada *AREA* (**Nualart**, 2013).

2.2.1 Visualización de textos individuales

Se presentan los textos individuales agrupados como vi-

sualización de todo el texto, visualización de parte del texto y otras sub-categorías. Éstas incluyen: secuencial, no-secuencial, estructura del discurso, estructura sintáctica, resultados de búsqueda y datos con línea temporal. Cada sub-sección sigue la estructura: lista de los casos en el grupo, descripción del grupo y discusión.

a) Visualizaciones de todo el texto

1) Literatura. *Novel views: Les misérables, Radial word connections* por Jeff Clark (2013)

2) Literatura. *Novel views: Les misérables, Character mentions* por Jeff Clark (2013)

3) Literatura. *Poem viewer* por Katharine Coles *et al.* (2013)

4) Política. *State of the Union 2011, Sentence bar diagrams* por Jeff Clark (2011)

5) Literatura. *Visualizing lexical novelty in literature* por Matthew Hurst (2011)

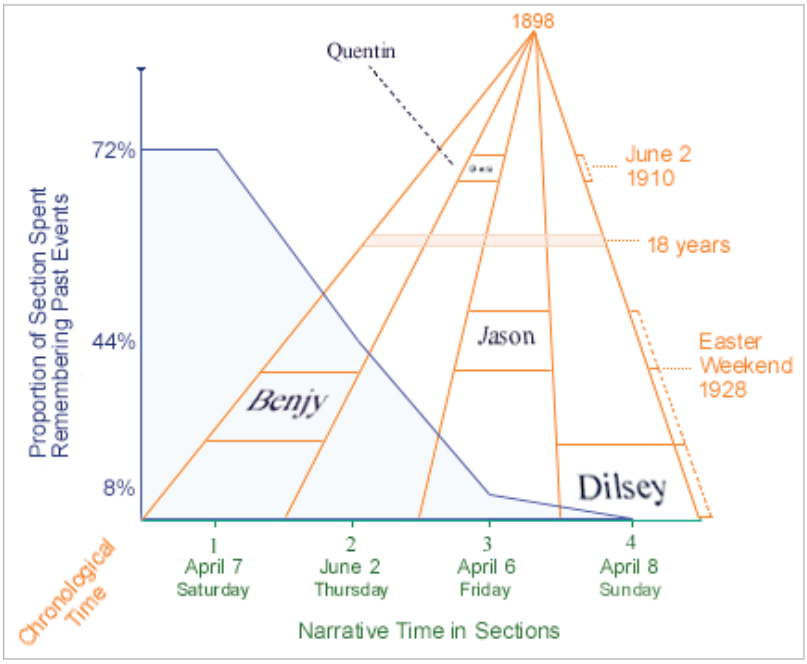

Figura 3. (Caso 13) *3-D display of time in the novel The sound and the fury* de William Faulkner, por Joel Deshaye y Peter Stoicheff (2003)

6) Artículos científicos. *On the origin of species: The preservation of favoured traces* por Ben Fry (2009)

7) Artículos científicos. *Texty* por Jaume Nualart (2008)

8) Religión. *Bible cross-references* por Chris Harrison (2008)

9) Literatura. *Literature fingerprint* por Daniel A. Keim y Daniela Oelke (2007)

10) Wikipedia. *History flow* por Fernanda Viégas y Martin Wattenberg (2003)

11) Literatura. *Colour-coded chronological sequencing* por Joel Deshaye y Peter Stoicheff (2003)

12) Literatura. *2-D display of time in the novel* por Joel Deshaye (2003)

13) Literatura. *3-D display of time in the novel* por Joel Deshaye (2003)

14) No específico. *Wattenberg's arc diagram* por Martin Wattenberg (2002)

15) Salud. *TileBars* por Marti A. Hearst (1995)

Descripción

- Número de casos: se encontraron 15 casos en la categoría de visualización de todo el texto.
- Años: entre 1995 y 2013 (dieciocho años).
- Autores: todos proceden de los ámbitos académicos. Los más prolíficos de esta categoría son Jeff Clark y Joel Deshaye (con tres casos cada uno), seguidos de Martin Wattenberg (con dos casos).
- Acerca de los datos: la mayoría de los corpus de texto utilizados pertenecen a la literatura (ocho casos). La mayoría de los autores utilizan los textos de literatura para mostrar nuevas estrategias de visualización; especialmente textos bien conocidos, como novelas clásicas.
- Acerca de los métodos: todos los casos, excepto *Arc diagram* (caso 14) utilizan el color como parte del método de visualización. Cinco casos utilizan variaciones de diagramas de barras (4, 5, 6, 9 y 11). Tres casos utilizan curvas que conectan partes de los textos: dos de ellos usan arcos y uno usa diagramas radiales (1, 8 y 14).

Discusión

No es posible establecer un método común para las visualizaciones de todo el texto. Todos los casos presentan un eje que representa implícitamente todo el texto. En 13 de los 15 casos la línea de texto se representa mediante una única línea horizontal o vertical. Las 2 excepciones representan la línea de texto como un círculo *Radial word connections* (caso 1) y como iconificación de un texto en una página, *Texty* (caso 7).

Puesto que la visualización de todo el texto incluye siempre una abstracción del mismo que llamamos línea de texto, surge una pregunta: ¿qué parte del texto se encuentra físicamente presente en las visualizaciones revisadas de textos completos? Es sorprendente que la mayoría de los casos (9 de los 15) no muestran ni una sola palabra (4, 5, 6, 7, 8, 9, 10, 11 y 15). Cuatro casos muestran un pequeño número de palabras (1, 2, 12, y 13). Sólo dos casos (3 y 14) muestran literalmente todo el texto.

El patrón más común es mostrar la ocurrencia de alguna característica como, por ejemplo, algún término concreto, referencias cruzadas o el carácter de personajes. Esto ocurre con todos los casos, excepto 3, 12, 13 y 15. Salvo en *Arc diagram* de Wattenberg las ocurrencias están representadas con un mismo color.

Es interesante observar cómo algunos casos representan datos similares de formas muy diferentes: *History flow* de Viégas y Wattenberg (caso 10) y de *The preservation of favoured traces* de Ben Fry (caso 6). Ambos muestran lo mismo: el historial de versiones de los documentos en cada

sección de cada documento. El segundo es una animación.

También podemos ver similitudes entre *TileBars* (caso 15) y *Texty* (caso 7). *Texty* utiliza la misma técnica que *TileBars*, destacando algunas palabras del texto dentro de una figura rectangular que representa la totalidad del texto.

También podemos observar casos que son opuestos o complementarios: *Arc diagram* de Wattenberg (caso 14) muestra la repetición, y de *Visualizing lexical novelty* de Hurst (caso 5), muestra sólo cadenas de texto nuevas, y no las repetidas.

La literatura y otros textos complejos, como los discursos políticos (caso 4) o la *Biblia* (caso 8) dominan el tipo de corpus de esta categoría (diez casos). Esto es sorprendente porque estos textos tienen un alto nivel de abstracción y poca estructura predefinida. Con el fin de ilustrar nuevas propuestas pensamos que es una buena idea trabajar con textos más estructurados y sencillos, con un vocabulario más regular, una longitud del texto estandarizada, con una clara estructura del discurso y corrección en el lenguaje (artículos científicos, textos de patentes, diagnósticos de salud, etc.). En literatura, debido a su inherente libertad de escritura, no es necesario seguir ningún patrón o regla que pueda ayudar a estructurar la no-estructura.

Sin embargo, dependiendo de como se trata y procesa, la naturaleza del texto no siempre es un punto clave. Por ejemplo, la obra de Matthew Hurst (caso 5) muestra una manera de realizar el seguimiento de la introducción de nuevos términos a lo largo de un texto. Esta herramienta puede ser utilizada por expertos en literatura y, al mismo tiempo, se puede aplicar a cualquier otro texto con resultados no relacionados con la complejidad debido a la ubicuidad del método. Sin embargo, sería interesante aplicar este método para trabajos científicos en los que el estilo es mucho más definido. Ideas similares se aplican en *Radial word connections* (caso 1), *Sentence bar diagrams* (caso 4) y *Literature fingerprint* (caso 9).

b) Visualización de parte del texto

16) Literatura. *Novel views: Les misérables. Characteristic verbs* por Jeff Clark (2013)

17) Cualquier texto. *Wordle* por Jonathan Feinberg (2009)

18) Libros. *DocuBurst* por C. Collins, S. Carpendale y G. Penn (2009)

19) Literatura. *Phrase nets* por Frank van Ham, Martin Wattenberg y Fernanda B. Viégas (2009)

20) Datos de *Google*. *Word spectrum: Visualizing Google's bi-gram data* por Chris Harrison (2008)

21) Datos de *Google*. *Word associations: Visualizing Google's bi-gram data* por Chris Harrison (2008)

22) Literatura/canciones. *Document arc diagrams* por Jeff Clark (2007)

23) Cualquier libro. *Gist icons* por P. DeCamp, A. Frid-Jimenez, J. Guiness, y D. Roy (2005).

Descripción

- Número de casos: 8
- Años: publicados entre 2005 y 2013 (ocho años).
- Sobre autores y datos: dos casos de Jeff Clark (16 y 22) y uno de la pareja creativa Wattenberg y Viégas en colaboración con Van Ham (caso 19) utilizan textos literarios. Los dos casos de Chris Harrison (20 y 21) utilizan muchos datos en formato *bi-gram* publicados por *Google*. Hay un caso, *Wordle* (17), que no depende de la naturaleza del texto que automatiza el popular método de nube de etiquetas introducido por Feinberg. Finalmente aparecen dos casos interactivos que permiten representar grandes colecciones de datos al mismo tiempo: *DocuBurst* (18) y *Gist icons* (23).
- Acerca de los métodos: en seis de los ocho casos (16, 17, 18, 19, 22 y 23) los datos se reducen a lo que se llama *bag-of-words* y sólo estas palabras están presentes en la visualización. Los casos 20 y 21 son representaciones de todos los *bi-grams* que comparten dos palabras comparadas entre sí.

Discusión

La visualización de parte del texto es una forma exitosa y popular de visualizar. Esto es debido a que es sorprendente que un texto largo se pueda representar con eficacia por un pequeño conjunto de palabras. Métodos estadísticos muy simples, como la frecuencia de palabras, pueden tener un resultado fácil de entender. Una lista de varios tamaños de palabras es una forma directa de comunicarse con cualquier usuario, desde principiantes hasta expertos. La mayoría de los métodos de visualización de partes de un texto que se encuentran

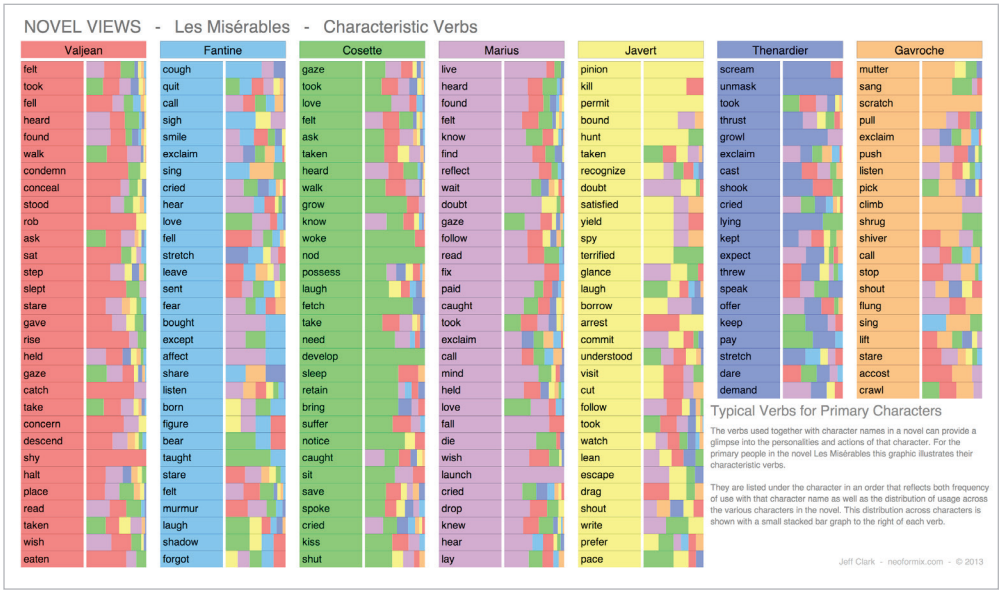

Figura 4. (Caso 16) *Novel views: Les misérables. Characteristic verbs* por Jeff Clark (2013)

online utilizan métodos estadísticos para extraer e identificar esa parte representativa de la totalidad.

En este trabajo sostenemos que la extracción de una parte del corpus puede afectar a la estructura del corpus completo del texto y su complejidad. En las visualizaciones revisadas, la mitad de los casos presentan un corpus de texto no-estructurado, sin embargo los criterios para extraer la parte del todo están muy bien definidos, como son las listas de verbos (*Characteristic verbs*, caso 16, figura 4), palabras que se encuentran en el texto según un patrón (*DocuBurst*, caso 18) y listas de palabras no incluidas en una lista predefinida de palabras vacías (*Google's bigram data*, caso 21).

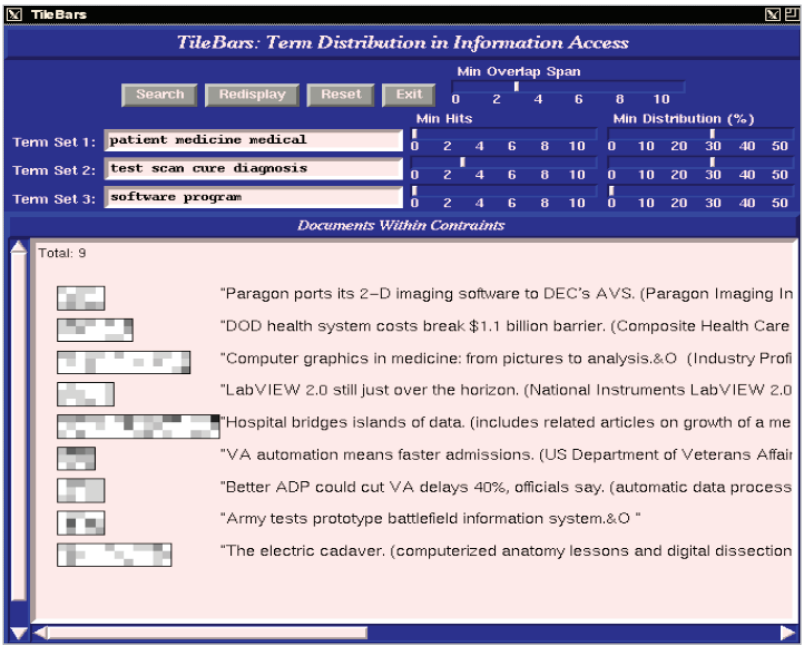

Figura 5. (Caso 15) *TileBar* representa resultados de búsquedas con distribución de coincidencias en cada documento, por Marti Hearts (1995).

De todas formas, los casos donde la extracción está basada en la palabra o frase en lugar de en la funcionalidad o la estadísticas pura tienen una mayor dependencia de la naturaleza del texto. Este es el caso de *Novel views: Les misérables - Characteristic verbs* (16), que sólo representa verbos, y el caso de *DocuBurst* (18), que utiliza la base abierta de datos léxica *Wordnet* como copia de seguridad. Y finalmente, los casos de *Phrase net* (19) y los dos de *Google bi-grams* (21).

Un comportamiento común que hemos identificado en la visualización de parte del texto es que una vez que se extrae una parte del texto, en todos los casos, excepto uno (*Document arc diagrams*, 22) no hacen ya ninguna referencia a la secuencia de texto original en la visualización. Trataremos este tema con más detalle en el siguiente punto (Visualización secuencial).

c) Otras subcategorías

Se incluyen los siguientes tipos de visualizaciones: secuencial, no secuencial, de estructura sintáctica, de resultados de búsqueda y con línea temporal.

Visualización secuencial

Encontramos 16 casos de un total de 23 en visualización de textos individuales donde se mantiene al menos una de las secuencias del texto original. En 7 de los 16 casos la secuencia sigue la estructura del discurso, sobre todo los capítulos. El resto, 9 casos, utilizan elementos sintácticos para representar la secuencia original, sobre todo palabras.

Es destacable que sólo un caso de visualización de parte del texto (*Document arc diagrams*, caso 22) sigue la secuencia del texto original. Al mismo tiempo, todos los casos de visualización de todo el texto son secuenciales. Parece que la característica secuencial es intrínseca a la visualización de todo el texto. Estas visualizaciones no representan literalmente cada palabra del texto, sino una metáfora gráfica de todo el texto: una línea. En cualquier caso, esta línea puede representar tanto la estructura discursiva como la sintáctica de forma gráfica con una línea o un área que representa la longitud total.

La secuencialidad de la visualización permite al lector retroceder y adelantar por la visualización como por el texto. En el caso de un texto largo, un libro (9 de 16 casos), la visualización puede actuar como un mapa o una guía para el texto.

Visualización no-secuencial

Hemos encontrado cinco casos: tres de ellos son nubes de etiquetas o palabras (casos 17, 20 y 21). Uno es tipo red (*Phrase nets*, caso 19) y otro es una representación todos los verbos del texto (*Characteristic verbs*, caso 16).

Visualización de la estructura del discurso

Casos: 1, 2, 5, 6, 8, 11, 12, y 13

Los 8 casos encontrados que siguen la estructura discursiva del texto son visualizaciones secuenciales. No hay ninguno en el que la estructura del discurso aparezca desordenada en relación con el texto. Esto no debe sorprender ya que cuando un texto se divide en capítulos y cada capítulo está representado como una entidad, se ha considerado como visualizaciones de colección de textos (*Sentence bar diagrams*, caso 4). Por eso, todos los casos de esta sección representan partes del texto ordenadas y alineadas en una línea o una curva. De estos 8 casos, 5 representan capítulos o secciones de un libro, 2 representan volúmenes, y hay uno especial (*Colour-coded chronological sequencing*, 11) en el que el texto se divide en colores de acuerdo con los temas puramente narrativos. Este es el único caso que utiliza elementos de la estructura del discurso más profundos que el de capítulos, secciones, libros y volúmenes.

Visualización de la estructura sintáctica

Casos: 3, 16, 4, 7, 18, 9, 22, y 23

La segunda mitad de los casos secuenciales, 8, utilizan elementos intrínsecos al texto: grupos de palabras (casos 7, 18, 22 y 23), verbos (16), oraciones (4 y 9) y un análisis completo del texto (3). Obtener la estructura sintáctica requiere

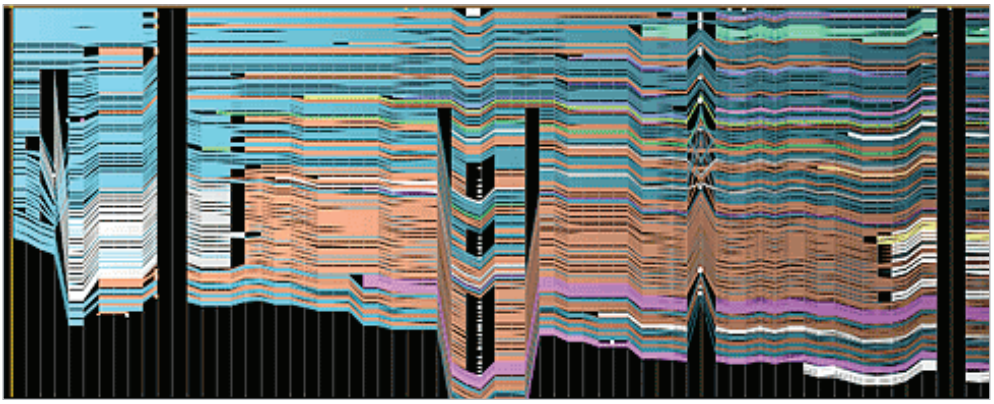

Figura 6. (Caso 10) *History flow* por Fernanda Viégas and Martin Wattenberg (2003)

el análisis del texto palabra por palabra, a veces utilizando una base de datos léxica o semántica para palabras, a veces análisis de oraciones y párrafos. La visualización usando la estructura sintáctica depende poco de la naturaleza del texto en el sentido de que la metodología es agnóstica a la complejidad del texto. Por lo general, el software extrae o marca los elementos sintácticos elegidos automáticamente.

Visualización de resultados de una búsqueda

Casos: 18, 23, y 15

Los 3 casos que visualizan resultados de búsquedas se presentan como aplicaciones web y son interactivos. El usuario puede consultar el sistema de visualización y obtener una representación única para cada petición. Los 3 casos que presentamos ya no están online. *DocuBurst* (caso 18) es una aplicación hecha con el software *Prefuse* que se puede descargar (**Collins** *et al*., 2009).

> La visualización de parte del texto es una forma exitosa de dibujar un texto, seguramente debido a la forma en que un texto largo puede representarse con un pequeño grupo de palabras

TileBars (figura 5) es un caso clásico (citado 625 veces según *Google Scholar*) por una autora de referencia en la visualización y las interfaces de los motores de búsqueda, Marti Hearst. *DocuBurst* y *Gist icons* son visualizaciones radiales interactivas. De hecho, *Gist icons* es una de las referencias y una de las principales influencias en el desarrollo de *DocuBurst*, como se explica en el citado documento de **Collins**.

En general, en los sistemas de recuperación de información no se implementa la visualización de resultados de búsqueda. La mayoría de éstos son listas unidimensionales de textos resumidos (**Nualart**; **Pérez-Montoro**, 2013). Estos tres casos tienen en común que se aplican a grandes colecciones de datos y, a partir de una consulta se muestra una visualización mejorada de los resultados que ayuda al usuario en el proceso de lectura y filtrado de los resultados. En los tres casos parece una tarea clave distinguir entre elementos similares. *TileBars* busca en *PubMed*, en más de 20 millones de entradas. *DocuBurst* utiliza la base de datos léxica abierta *WordNet* (155.287 palabras organizadas en 117.659 *synsets*, con un total de 206.941 pares de palabras) para clasificar el texto visualizado. Finalmente, los ejemplos de *Gist icons* representan, entre otros datos, una serie completa de aproximadamente 7 millones de patentes de la *USPTO* y 500.000 correos electrónicos de *Enron*.

En la categoría de colección de textos presentamos 9 casos de visualización aplicados a resultados de una búsqueda.

Visualizaciones con línea temporal

Se presentan 2 casos (9 y 10) donde la visualización se puede utilizar para comprender o seguir la evolución temporal del texto representado. La visualización de texto dinámico demuestra que la visualización de datos puede ser casi la única manera de resolver algunas tareas y no sólo una forma más. Por ejemplo, es muy difícil demostrar cómo una entrada de la *Wikipedia* evoluciona con el tiempo de acuerdo con la participación de los editores. *History flow* es una solución clara para este problema y ayuda a entender parte del proceso de colaboración compleja de la *Wikipedia* (*History flow*, caso 10, figura 6).

El segundo caso (*Favoured traces*, 6) es una visualización animada que muestra cómo cambiaron las teorías de Darwin a lo largo de las ediciones de su *Origen de las Especies*. En palabras de Ben Fry: "La primera edición inglesa era de aproximadamente 150.000 palabras y la sexta es una cantidad mucho mayor, 190.000 palabras. En la evolución encontramos tanto mejoras del texto como cambios de ideas, ya sea dando más peso a un razonamiento, agregando detalles o, incluso, cambiando la idea en sí misma".

2.2.2. Colecciones de textos

Presentamos colecciones de textos agrupados como: visualización de colecciones de items y de colecciones de agregados, además de otras subcategorías. Éstas incluyen datos más una segunda capa de datos tipo *landscape* y visualización de resultados de búsqueda. Cada subsección sigue la estructura: lista de casos, descripción del grupo y discusión.

a) Elementos de visualización

24) Literatura (Nota: convierte un texto individual en una colección de textos). *Novel views: Les misérables. Segment word clouds* por Jeff Clark (2013)

25) Literatura. *Grimm's fairy tale network* por Jeff Clark (2013)

26) Twitter. *Spot* por Jeff Clark (2012)

27) Ciencia. *Word storm* por Quim Castellà y Charles Sutton (2012)

28) Literatura. *Topic networks in Proust. Topology* por Elijah

Meeks y Jeff Drouin (2011)

29) Wikipedia. *Notabilia* por D. Taraborelli, G. L. Ciampaglia y M. Stefaner (2010)

30) Media art. *X by Y* por Moritz Stefaner (2009)

31) Motor de búsqueda. *Search clock* por Chris Harrison (2008)

32) Magazine en línea. *Digg rings* por Chris Harrison (2008)

33) Ciencia. *Royal Society Archive* por Chris Harrison (2008)

34) Wikipedia. *WikiViz: Visualizing Wikipedia* por Chris Harrison (2007)

35) Visualización. *AREA* por Jaume Nualart (2007)

36) *Chromograms* por M. Wattenberg, F. B. Viégas y K. Hollenbach (2004)

37) Motores de búsqueda. *KartOO/ Ujiko* por Laurent Baleydier y Nicolas Baleydier (2001)

38) Motores de búsqueda. *Touchgraph* por TouchGraph, LLC (2001)

39) Internet. *HotSauce* por Ramanathan V. Guha (1996)

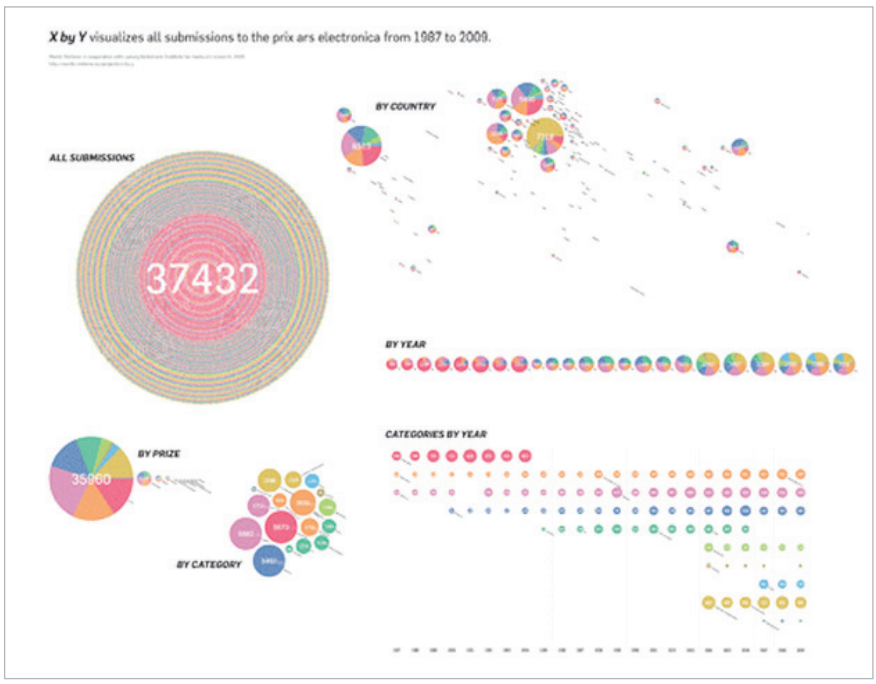

Figura 7. (Caso 30) *X by Y* por Moritz Stefaner (2009)

Descripción

- Número de casos: 16.
- Años: entre 1996 y 2013 (17 años).
- Autores: los más prolíficos son Chris Harrison (casos 13, 32, 33 y 34) y Jeff Clark (casos 24, 25 y 26), seguidos por Moritz Stefaner con dos casos (casos 29 y 30).
- Disciplinas y datos: es de destacar que nueve de los casos son datos provenientes de internet: *Wikipedia* (casos 29, 34 y 36), motores de búsqueda (casos 31, 37 y 38), *Twitter* (caso 26), medios de comunicación online (caso 32) y páginas web (caso 39). En esta categoría sólo hay tres casos que utilizan textos literarios (casos 24, 25 y 28). Por último dos casos representan artículos científicos (casos 27 y 33), uno de ellos representa datos de *Media art* (caso 30) y el otro representa colecciones no específicas (caso 35).

Discusión

La gran diferencia entre visualizaciones de texto individuales y de colecciones de textos es la propia naturaleza de los textos. En las colecciones la mayoría de éstos no son literarios y, además, son accesibles online. Probablemente la naturaleza del texto es menos importante cuando el objetivo de la representación es la colección.

Las visualizaciones de items utilizan métodos independientes de la naturaleza de los textos. Una vez determinadas las colecciones, los datos se pueden considerar un caso general de visualización de datos y no un caso puro de visualización de texto. Por esta razón, en esta categoría, podemos generalizar diciendo que los métodos que encontramos son bien conocidos y utilizados en otros campos de la visualización de datos. Siguiendo este razonamiento nos encontramos con 6 visualizaciones de red (casos 25, 28, 34, 37, 38 y 39), 3 líneas temporales (31, 32 y 33), y 3 que utilizan también

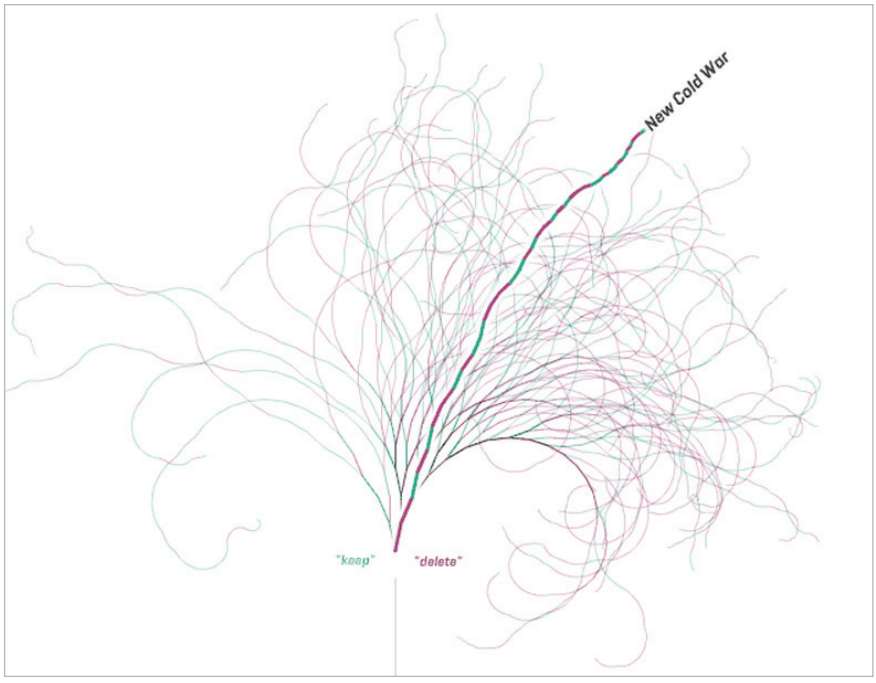

Figura 8. (Caso 29) *Notabilia. 100 discusiones sobre eliminacion de páginas en Wikipedia* por Dario Taraborelli, Giovanni-Luca Ciampaglia (datos y análisis) y Moritz Stefaner (visualización) (2010)

Figura 9. (Caso 43) *Web seer* por Fernanda Viégas y Martin Wattenberg (2009)

las líneas temporales pero que permiten la agrupación por categorización (26, 30 y 35) (figura 7).

Por último, hay 4 casos específicos de la visualización de textos. Dos de ellos están dedicados a la comparación de items de la colección: *Segment word clouds* (24) y *Word storm* (27). Sobre *Segment word clouds* es destacable la transformación de un texto individual en una colección de textos. Representa los capítulos de *Les misérables* como items de nubes de palabras; eso hace que sea fácil compararlos. También utiliza el color para detectar palabras nuevas en cada capítulo.

Word storm es una reinvención de la nube de etiquetas o palabras, más concretamente es una variación de *Wordle* (caso 17) que permite comparar nubes de palabras mediante la asignación de una posición fija a cada palabra. Esta simple idea hace que sea visualmente fácil comparar nubes de palabras, mientras se mantienen sus habituales características.

> Sería más efectivo aplicar técnicas de visualización a los textos más estructurados, con disposición y vocabulario definidos

Por último, los casos más originales que merecen una mención especial son *Notabilia* (29) y *Chromograms* (36). *Notabilia* muestra la evolución de las discusiones en *Wikipedia* llamadas "artículo para eliminación". Estas discusiones son a veces muy encendidas, pudiendo llegar a "flame wars", debido a la controversia sobre algunos artículos. *Notabilia* representa la evolución y la decisión final de las cien discusiones más largas. La visualización es de Moritz Stefaner y es un sistema tipo arbusto interactivo que destaca ramas cuando el ratón pasa por encima. La forma de las ramas informa acerca de la naturaleza de la discusión: cíclicas, restas, o de nunca acabar.

Chromograms también es un trabajo basado en los datos de *Wikipedia*. Analiza los comentarios de los editores para cada edición de una entrada de *Wikipedia*. Produce líneas de color codificado que en una zona corta explican los años de las ediciones de las entradas de *Wikipedia*.

b) Visualización de agregados

40) Literatura. *Grimm's fairy tale metrics* por Jeff Clark (2013)

41) Topic Models. *Termite* por J. Chuang, C. D. Manning y J. Heer (2012)

42) Wikipedia. *Pediameter* por Müller-Birn, Benedix y Hantke (2011)

43) Sugerencias de *Google*. *Web Seer* por Fernanda Viégas y Martin Wattenberg (2009)

44) *Google* n-grams. *Web tri-grams: visualizing Google's tri-gram data* por Chris Harrison (2008)

45) Discursos políticos. *FeatureLens* por A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. Clement, B. Shneiderman y C. Plaisant (2007)

46) Noticias en línea. *Newsmap* por Marcos Weskamp (2004)

47) Conversaciones por correo electrónico. *Themail* por Fernanda B. Viégas, Scott Golder, Judith Donath (2006)

48) Motores de búsqueda. *WebBook* por S. K. Card, G. G. Robertson, y W. York (1996)

49) Cualquier texto. *Dotplot applications* por Jonathan Helfman (1994)

Descripción

- Número de casos: 10.
- Años: publicados entre 1994 y 2013 (19 años).
- Los autores y los datos: sólo Fernanda B. Viégas tiene participación simultánea en dos de los diez casos que se presentan en esta categoría (casos 43 y 47). El resto de los autores participa con un caso cada uno. La naturaleza de los textos es muy similar a la categoría de visualización de items. Cinco casos de datos que se pueden encontrar online (*Wikipedia*, caso 42; *Google*, casos 43 y 44; noticias online, caso 46; resultados de motores de búsqueda, caso 48). Como textos no-estructurados hay uno literario (*Sentence bar diagrams*, caso 4), uno de discursos políticos (*FeaturedLens*, caso 45) y uno de un año de conversaciones por correo electrónico entre dos personas (*TheMail*, caso 47). Por último, hay dos casos bastante especiales: *Termite* (41) y *Dotplot* (49).

Discusión

Visualización de agregados es la categoría con mayor variabilidad de métodos. Los 10 métodos presentes en esta categoría,

además de representar colecciones de textos, sólo tienen en común que no están representando items específicos. Debido a esta variedad de casos, los comentamos uno por uno.

Sentence bar diagrams (caso 4) es una matriz en forma de tabla que permite ordenar filas haciendo clic en cada columna. Las columnas definen, en orden cuantitativo, 13 medidas sobre los 62 cuentos de hadas de los hermanos Grimm. Es una poderosa herramienta para comprender y comparar los relatos.

Termite (caso 41) representa un conjunto de datos transformados llamado *topic models* de temas. *Topic models* es una forma de extraer un conjunto de palabras más elaborada que el típico análisis estadístico de frecuencia de palabras. *Termite* no visualiza textos, sino que compara fragmentos de textos. Es una herramienta para comparar y obtener una primera evaluación de los *topic models*.

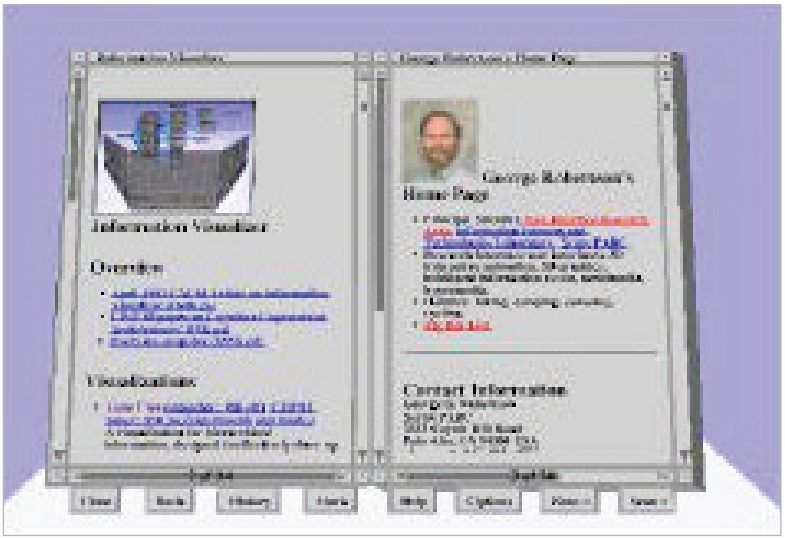

Figura 10. (Case 48) *WebBook* por S. K. Card, G. G. Robertson, y W. York (1996)

Pediameter (caso 42) es una interfaz específica que utiliza gráficos de barras para mostrar ediciones de *Wikipedia* en tiempo real. Es notable que el proyecto utiliza un dispositivo llamado *Arduino* para detectar las ediciones y transcribir a un indicador físico, fusionando mundos digitales y materiales.

Web seer (caso 43) es un método específico de visualización que muestra las consultas de búsqueda más populares a partir de sugerencias de *Google*. Permite la comparación de las consultas representando las sugerencias del buscador como árboles de palabras y expresiones. La simplicidad de este método contrasta con su poder de comunicación: rápido y fácil de entender.

Tri-gram de datos de *Google* (caso 44) utiliza una forma de representación similar a *Web seer*. En cuanto a datos, utiliza el enorme conjunto de datos de *Google* n-grams. Representa y compara tres frases (tri-grams).

FeaturedLens (caso 45) es una interfaz interactiva tipo panel de control que permite la comparación de textos. La zona central muestra una representación visual de conceptos frecuentes similar a *Texty* (caso 7) y *TileBars* (caso 15). Permite la navegación por el texto y muestra gráficos de líneas de palabras frecuentes a lo largo de los textos.

Newsmap (caso 46) utiliza la técnica *treemap* para proponer una nueva forma de leer las noticias en tiempo real basado en las noticias de *Google*. Es totalmente personalizable en cuanto a los temas de noticias, país, fecha y hora de publicación. Permite también la búsqueda de noticias. Este software está disponible online para su uso libre.

TheMail (caso 47) es un experimento para el que se desarrolló una interfaz muy específica que permite seguir y analizar la evolución de la correspondencia por correo electrónico entre dos personas en el transcurso de un año. Representa las palabras que caracterizan a cada uno de los escritores y su evolución en el tiempo.

WebBook (caso 48, figura 10) es una aplicación sorprendente en su época, 1996. Transformaba una lista de resultados de motor de búsqueda en un publicación multimedia (con texto e imágenes, básicamente) usando la metáfora de un libro. Esta fue una visualización pura de una colección dinámica de textos (páginas web) que se presentaban como agregados de texto e imágenes.

Finalmente se presenta *Dotplot* (caso 49) una gran idea de visualización que tiene múltiples usos. Es comparable a los *Arc diagrams* (caso 14). El uso principal de *dotplots* es la comparación de textos, incluyendo textos multi-idioma, comparación de versiones de texto y comparación de códigos de programación.

c) Otras subcategorías

Entre estas subcategorías se incluyen capas de datos tipo paisaje, visualización de resultados de búsqueda y de línea de tiempo.

El landscape de datos como una segunda capa de datos

Casos: 40, 26, 28, 33, 47, 37, 38 y 49

La idea típica de los datos tipo *landscape* (paisaje) es una visualización de datos en red con dos capas de datos, como en *Topic networks* (caso 28). En este caso, la primera capa es la lista de los textos de Marcel Proust representados como objetos, y la segunda capa es una red de *topic models* de los textos. Las posiciones de los nodos de ambas capas están optimizados. Proximidad significa nodos más relacionados. Esta definición de paisaje de datos se puede encontrar también en los casos, ya no accesibles, de los motores de búsqueda: *Kartoo* / *Ujiko* (caso 37) y *TouchGraph* (caso 38).

El resto de los casos clasificados dentro de esta categoría muestran colecciones de textos en combinación con otros datos. *Dotplot* representa la coincidencia o no de las cadenas de varios textos. Y lo mismo ocurre con *Grimm's fairy tale metrics* donde una tabla combina una lista de textos en las filas con varios parámetros en las columnas. Estos parámetros no son directamente parte del texto, sino características del texto transformadas para cada cuento: longitud, diversidad léxica y presencia de diferentes grupos de palabras que representan conceptos (por ejemplo: cuerpo -> mano, cabeza, corazón, ojos y pie).

Un tercer tipo de paisaje de datos se basa en la representación de metadatos dependientes del tiempo. Es el caso de *Spot* (26), *Royal Society Archive* (33) y *TheMail* (47).

Por su propia naturaleza, la característica común que tienen las visualizaciones con capa tipo paisaje de datos es la capacidad de comparar un conjunto de textos de forma simultánea con un segundo parámetro. A su vez, la limitación de estas visualizaciones es el número de elementos representados; un gran número de elementos crea problemas visuales de superposición.

Visualización de resultados de una búsqueda

Casos: 26, 43, 35, 45, 47, 46, 37, 38, y 48

En comparación con la visualización de textos individuales, las visualizaciones de colecciones de textos tienen muchos más casos de representación de resultados de búsquedas (3 casos frente a 9). El sentido común sugiere que en la presentación de una colección de textos una característica natural del enfoque puede ser una manera de seleccionar parte de la colección de acuerdo con algunos criterios, es decir, filtros y otras características de la búsqueda.

Todos los casos de esta categoría permiten las búsquedas y devuelven resultados con una visualización única para cada consulta. Todos los casos incluyen un cuadro de texto para la búsqueda y un botón de búsqueda.

Visualización con línea temporal

Casos: 42, 29, 36, y 46.

Permiten al usuario seguir la evolución de los textos de la colección a través del tiempo. Sólo uno está diseñado para ser en tiempo real (*Newsmap*, caso 46), pero potencialmente todos ellos pueden mostrar la colección para una fecha y hora específicas.

Una de las dificultades en una visualización que representa colecciones de textos que cambian con el tiempo es el acceso a una fuente actualizada o una API. Probablemente por esta razón, 3 de los 4 casos usan datos de *Wikipedia* y el cuarto caso usa *Google News*. Todos usan fuentes online que permiten el acceso público.

3. Conclusiones

La diversidad de enfoques de diferentes disciplinas, la difusión de publicaciones y, a veces, la falta de publicaciones oficiosas de nuevas ideas, suponen un reto a la hora de hacer un estudio completo del trabajo en el campo de la visualización de textos. Algunos de los casos que se presentan han sido encontrados en publicaciones muy específicas, por ejemplo Joel Deshaye y Peter Stoicheff con sus trabajos en las visualizaciones de Faulkner (casos 11, 12 y 13). La lectura de las notas de Stoicheff muestra que desarrollaron las visualizaciones sólo para ayudar en un estudio muy específico de narrativas y líneas temporales de William Faulkner. No hay referencias a las aplicaciones de estas interesantes ideas a otros textos, lo que sugiere que un mayor número de trabajos permanecen ocultos en las profundidades de otras disciplinas.

La visualización de textos, tal como se argumenta en este trabajo, puede ser considerada un subcampo de la visualización de datos. Sin embargo, los límites del campo a veces no están claramente definidos. Éste es el caso de *Search clock* (caso 31), en el que el corpus de texto es una enorme base de datos de consultas de motores de búsqueda. ¿Puede este conjunto de datos ser considerado como una colección de textos si cada uno de ellos, en la mayoría de los casos, consiste sólo en una o dos palabras? ¿Existe una longitud mínima de un texto para ser considerado como tal? Decidimos tratar este caso (caso 31) como una colección de textos, cortos, pero en última instancia, textos.

Una decisión importante en esta revisión ha sido la clasificación de los casos encontrados. Dado que hay pocos trabajos que revisan únicamente visualización de textos, hemos referenciado las revisiones clásicas de visualización de datos (**Shneiderman**, 1996), así como otras más recientes (**Collins** *et al.*, 2009). En estos casos, las clasificaciones se basan en tareas que la visualización puede resolver en lugar de en los aspectos explícitos de la visualización. Es por ello que hemos decidido proponer una clasificación propia que, aunque lejos de ser perfecta, es de esperar que sea útil para una clasificación basada en características visuales.

A modo de resumen, podemos destacar de forma concentrada los puntos de vista y las deficiencias que hemos encontrado en este análisis.

En primer lugar, la visualización de texto individual se aplica principalmente al ámbito de la literatura. Los textos literarios, además de complejas combinaciones de palabras, pueden tener altos niveles de abstracción, libertad de las estructuras y experimentación narrativa. Pensamos que puede ser más eficaz aplicar las técnicas de visualización a otros tipos de textos con un registro más formal y/o una estructura predefinida y un vocabulario más controlado. Este es el caso de: textos legales, artículos científicos, comunicaciones basadas en plantillas, etc.

Por otro lado, hemos encontrado solamente un caso de visualización individual de parte del texto que es secuencial (*Document arc diagrams*, caso 22). La mayoría de las visualizaciones de partes de textos extraen la esencia del texto en función de algunos criterios y la secuencia original del texto se pierde. Dado que las visualizaciones secuenciales tienen algunas ventajas, animamos al desarrollo de enfoques de visualización de partes de textos que mantengan la secuencia original.

En tercer lugar, las visualizaciones de colecciones de textos utilizan métodos que se pueden encontrar en la visualización de datos en general. Esta idea invita a la experimentación en la aplicación de métodos de visualización de datos estándares en el subcampo específico de visualización de textos.

Y, en cuarto lugar, colecciones de agregados es la categoría que ha desarrollado diseños e ideas más específicas. Es necesaria más investigación con el fin de encontrar patrones en este tipo de visualizaciones.

Por último, para concluir, es interesante introducir una reflexión abierta: ¿por qué la mayoría de los casos examinados con más de 5 años ya no están online? Si este software ya no está (o nunca ha estado) en uso, hay que cuestionar su eficacia. No hemos investigado cuántos casos son parte de productos de software comercial y cuántos, después de haber sido publicados, se han olvidado. En cualquier caso,

es cuestionable por qué algunas grandes ideas no se han convertido en nuevos estándares. En este sentido, animamos a otros investigadores a elaborar aplicaciones para que sean adoptadas en algún campo y puedan resolver tareas de algún grupo de usuarios. De hecho, como se observa en los casos revisados, la adopción de esas ideas parece ser un reto importante.

Agradecimientos

Este trabajo forma parte del proyecto *Audiencias activas y periodismo. Interactividad, integración en la web y buscabilidad de la información periodística*. CSO2012-39518-C04-02. Plan Nacional de I+D+i, *Ministerio de Economía y Competitividad* (España).

4. Bibliografía

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

Baeza-Yates, Ricardo; **Ribeiro-Neto, Berthier** *et al.* (1999). *Modern information retrieval*. New York: ACM press, vol. 463.

Baeza-Yates, Ricardo; **Broder, Andreiz**; **Maarek, Yoelle** (2011). "The new frontier of web search technology: Seven challenges". *Search computing*, v. 6585 of *Lecture notes in computer science*, pp. 3-9.
http://dx.doi.org/10.1007/978-3-642-19668-3_1

Benavides, David; **Segura, Sergio**; **Ruiz-Cortés, Antonio** (2010). "Automated analysis of feature models 20 years later: A literature review". *Information systems*, v. 35, n. 6, pp. 615-636.
http://dx.doi.org/10.1016/j.is.2010.01.001

Collins, Christopher; **Carpendale, Sheelagh**; **Penn, Gerald** (2009). "DocuBurst: Visualizing document content using language Structure". *Computer graphics forum* (Procs. of the *Eurographics/IEEE-VGTC Symposium on visualization*, EuroVis), v. 28, n. 3, pp. 1039-1046.
http://dx.doi.org/10.1111/j.1467-8659.2009.01439.x

Feldman, Ronen; **Sanger, James** (2006). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press. ISBN: 13 978 0 521 83657 9

Grobelnik, Marko; **Mladenić, Dunja** (2002). "Efficient visualization of large text corpora". In: *Procs of the 7th seminar*. Dubrovnik, Croatia.
http://ailab.ijs.si/dunja/SiKDD2002/papers/GrobelnikSep02.pdf

Hearst, Marti A. (2003). *What is text mining?*
http://people.ischool.berkeley.edu/~hearst/text-mining.html

Hearst, Marti A. (2009). "Search user interfaces", Chapter 1. ISBN: 9780521113793
http://searchuserinterfaces.com/book
http://searchuserinterfaces.com/book/sui_ch1_design.html

Hearst, Marti A. (2011). "Natural search user interfaces". *Communications of the ACM*, v., 54, n. 11, November, pp. 60-67.
http://cacm.acm.org/magazines/2011/11/138216-natural-search-user-interfaces/fulltext
http://dx.doi.org/10.1145/2018396.2018414

Heer, Jeff (2010). "A conversation with Jeff Heer, Martin Wattenberg, and Fernanda Viégas". *Queue*, v. 8, n. 3, 10 pp., March.
http://doi.acm.org/10.1145/1737923.1744741

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

Kitchenham, Barbara (2004). *Procedures for performing systematic reviews*. Keele, UK, Keele University, 33 pp.

Levie, W. Howard; **Lentz, Richard** (1982). "Effects of text illustrations: A review of research". *ECTJ*, v. 30, n. 4, pp. 195–232.

Mann, Thomas M. (2002). *Visualization of search results from the world wide web*.
http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.91.2535

Meeks, Elijah (2011). *Digital humanities specialist*. Documents.
https://dhs.stanford.edu/comprehending-the-digital-humanities/documents

Nualart-Vilaplana, Jaume (2013). *How we draw texts: a visualization of text visualization tools.*
http://research.nualart.cat/textvistools

Nualart, Jaume; **Pérez-Montoro, Mario** (2013). "Texty, a visualization tool to aid selection of texts from search outputs". *Information research*, v. 18, n. 2, June.
http://www.informationr.net/ir/18-2/paper581.html

Shneiderman, Ben (1996). "The eyes have it: A task by data type taxonomy for information visualizations". In: *Visual Languages*. Proceedings IEEE Symposium, pp. 336–343.
http://dx.doi.org/10.1109/VL.1996.545307

Šilić, Artur; **Dalbelo-Bašić, Bojana** (2010). "Visualization of text streams: A survey". *Knowledge-based and intelligent information and engineering systems*, v. 6277 of *Lecture notes in computer science*, pp. 31–43. Berlin, Heidelberg: Springer.
http://dx.doi.org/10.1007/978-3-642-15390-7_4

Stefaner, Moritz (2013). *Gender balance visualization*.
http://moritz.stefaner.eu/projects/gender-balance/#NUM/NUM

Strecker, Jacqueline (2012). *Data visualization in review: summary*. International Development Research Centre (IDRC), Ottawa, ON, Canada.
http://idl-bnc.idrc.ca/dspace/bitstream/10625/49286/1/IDL-49286.pdf

Times Higher Education. *World university rankings 2012-2013*.
http://www.timeshighereducation.co.uk/world-university-rankings/2012-13/world-ranking

Tufte, Edward R.; **Graves-Morris, Peter R.** (1983). *The visual display of quantitative information*, v. 2. Cheshire, CT: Graphics Press, 199 pp.