CLUSTERS: UNA ALTERNATIVA PARA LA VISUALIZACION DE RECUPERACION DE INFORMACIÓN

****

Nidia Patricia Espíndola
Jack Diana L. Zambrano G.
Estudiantes Sistemas de Información y Documentación
UNIVERSIDAD DE LA SALLE
2007

****

RESUMEN

Representar grandes cantidades de datos de manera gráfica usando técnicas de clasificación (cluster) facilita la generación de nuevo conocimiento como resultado de catalogar diferentes tipos de documentos de texto como correo electrónico, búsquedas de Internet, historiales médicos, en general. El entorno actual de los negocios es algunas veces impredecible y requiere que la toma de decisiones se realice de manera ágil, por lo que habitualmente se deben analizar altos volúmenes de información acerca de desarrollo de nuevos productos para consumo, gestión de calidad, o análisis de mercados, entre otros, en busca de patrones y relaciones.

ABSTRACT

Representing large amount of data graphically using cluster techniques aids to discover knowledge as result of analyzing different kind of text documents as e-mails, web search results, medical data, and so on. Today business environment is some times unpredictable and require making tough decisions quickly and this involve search large volume of data about consumer product development, quality control management, or marketing, determining patterns and relationships.

PALABRAS CLAVE:
CLUSTER, CLUSTERING / CLASIFICACIÓN DE DATOS / RECUPERACIÓN DE INFORMACIÓN / REPRESENTACIÓN GRÁFICA DE INFORMACIÓN.

KEYWORDS:
CLUSTER / CLUSTERING / DATA CLASSIFICATION / INFORMATION RETRIEVAL / GRAPHIC DATA REPRESENTATION

CLUSTER COMO SOLUCION PARA EL PROBLEMA DE RECUPERACION DE INFORMACIÒN

Vivimos en un mundo donde tener la información adecuada en el lugar y tiempo preciso, define la pérdida o ganancia de un negocio, miles de datos llenan a diario nuestra red y se presentan grandes inconvenientes al momento de recuperar información:

- Demasiada información para consultar en poco tiempo
- Fuentes poco confiables
- Recuperación de información no solicitada
- Falta de orden en la información
- Listas interminables de artículos y textos con títulos que no explican nada

Cuando una persona busca información espera hacerlo en el menor tiempo posible y obtener documentos con un alto nivel de relevancia y que su pertinencia corresponda a las necesidades de cada usuario.

Cluster ofrece una posibilidad de obtener información de manera estructurada, clasificada por sus particularidades y similitudes, agilizando estas tareas.

Clustering es una de las principales técnicas de Data Mining(*) que consiste en particionar un conjunto de datos (dataset) en colecciones de objetos o instancias de manera que dentro de cada colección los objetos sean “similares” entre sí y a su vez se “diferencien” de los objetos contenidos en otras colecciones[1]. Mari Carmen Marcos Mora define Clustering como: Acción de agrupar objetos similares mediante algoritmos matemáticos[2].

Algunas características de clustering son:

1. Proporcionar como descriptores de una agrupación el conjunto de palabras más características de la misma.
2. Crear listas de palabras desde las que se identifica la información relevante.
3. Elaborar extractores que identifican materias.

Para determinar la similitud o disimilitud de un documento frente a otro y conformar una colección, se utilizan funciones algorítmicas cuya aplicación define los grupos documentales o clusters. Si a estas técnicas de agrupación se suma una representación gráfica de su resultado se puede obtener una forma más sencilla de ojear y obtener la información solicitada. (LAMARCA LAPUENTE – 2007- [3])

∗ Data Mining es una pequeña parte del proceso de descubrimiento de conocimiento que consiste en la aplicación de algoritmos matemáticos para la extracción de patrones utilizando los datos disponibles.

[1] NAVAS, María Daniela. Un modelo de Clustering temporal, [en línea], Buenos Aires. 2004. 190 h. Trabajo de grado (Ingeniería en Informática). Universidad de Buenos Aires. Facultad de Ingeniería, [Citado 15 de marzo de 2007], Página Web Formato de archivo: PDF/Adobe Acrobat, Disponible en Internet <http://www.fi.uba.ar/materias/7500/navas-tesisdegradoingenieriainformatica.pdf>

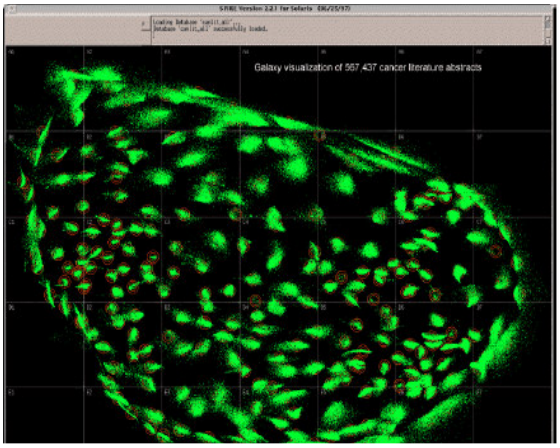
[2] DE SAMANIEGO, Javier. Recuperación y organización de la información: Clustering. [en línea], [Madrid-España?]. Universidad Carlos Tercero de Madrid, [Citado 15 de marzo de 2007], Página Web Formato de archivo: HTML, Disponible en Internet <http://es.geocities.com/clusteringrecuperacion/clustering.htm>

3 LAMARCA LAPUENTE, María Jesús. El nuevo concepto de documento en la cultura de la imagen, [en línea], Universidad Complutense de Madrid - Tesis doctoral, 2007, 01/04/2007, [Citado 15 de marzo de 2007], Formato HTML, Disponible en Internet: <http://www.hipertexto.info/documentos/busq_rec.htm>.

VISUALIZACIÓN EN LOS PROCESOS DE BÚSQUEDA CON CLUSTERING

La recuperación de información tomando como base clusters tiene diferentes formas de visualización, esto se debe a que son técnicas que utilizan algoritmos matemáticos que bien pueden variar dependiendo del diseño realizado por sus creadores y/o programadores. Algunos ejemplos de ellos son:

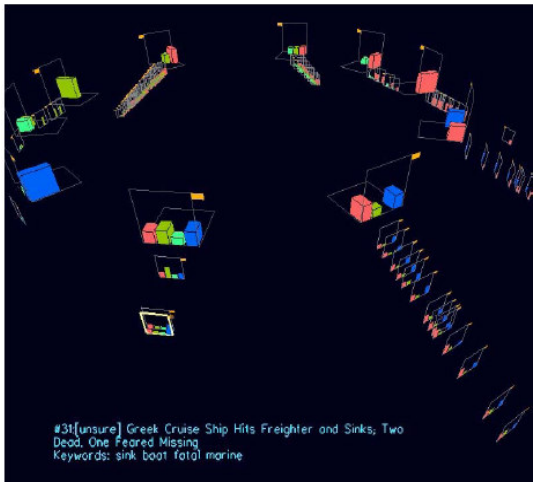
SPIRE (Spatial Paradigm for Information Retrieval)



En el proyecto SPIRE (Spatial Paradigm for Information Retrieval) del Pacific Nortwest Nacional Laboratorios (http://www.pnl.gov/infoviz) se desarrolló "Galaxias". Este programa de clasificación visual basado en clustering, hace que los documentos aparezcan como estrellas y estén agrupados entre sí como constelaciones teniendo como base la co-ocurrencia estadística, de esta forma, un vistazo sirve para ver fácilmente los temas y dirigirse al que interesa, y dentro de esa zona ahondar más para ver qué documentos incluye cada cluster[4].

Nirve (NIST Information Retrieval Visualization Engine)

Un ejemplo muy significativo de representación de documentos obtenidos en una consulta en un sistema de recuperación de información mediante agrupación (clustering) es el caso de NIRVE (NIST Information Retrieval Visualization Engine) (Cugini, Laskowski y Sebrechts, 2000) (http://zing.ncsl.nist.gov/~Cugini/uicd/NIRVE-home.html) [5].



En este sistema cada documento recuperado incluye el título del documento, un identificador único, la relevancia con respecto a la consulta, el tamaño del documento y el número de ocurrencias de cada descriptor que lleva asignado.

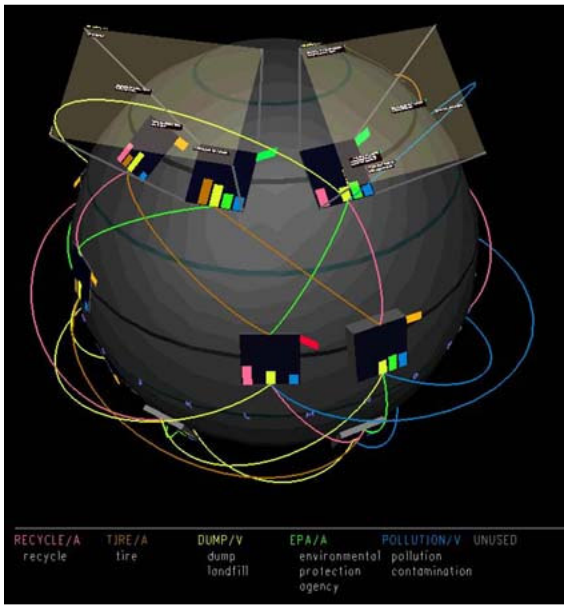
---

[4] MARCOS MORA, Mari Carmen. "La visualización en el proceso de búsqueda y recuperación de información" [en línea]. En Rovira, C.; Codina, L. (dir.). Documentación digital. Barcelona: Sección Científica de Ciencias de la Documentación del Departamento de Ciencias Políticas y Sociales de la Universidad Pompeu Fabra, 2004. Formato PDF, Disponible en Internet <http://www.mcmarcos.com/pdf/2004_visualizacion-modd.pdf>. ISBN: 84-88042-39-6.

[5] Ibid.

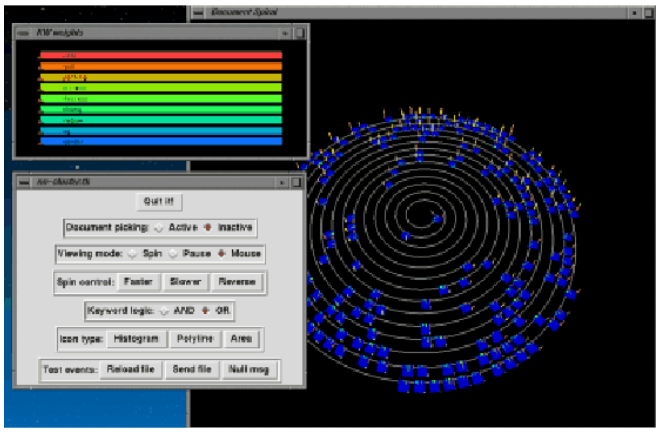
Cada cluster se representa como una caja con barras de colores que indican el perfil de concepto medio de sus documentos; cuantos más documentos contenga el cluster, más grande se representa la caja.

NIRVE puede representar el resultado a una consulta de variadas formas, por ejemplo en forma de globo (una esfera) donde la latitud viene determinada por el número de conceptos del cluster (cuantos más conceptos, más cerca del polo norte). Los iconos se colocan de manera que los clusters con más conceptos en común estén más cerca entre sí. Si dos clusters difieren entre sí por un solo término, entonces se dibuja un arco que los une y que tendrá el color asignado al concepto diferente.



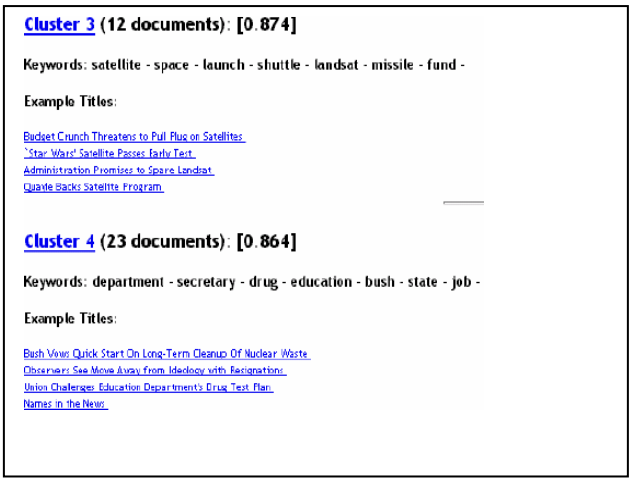
Prise

Diseñado en 1996, combina la presentación de los documentos por medio de clustering con una idea innovadora, la cual consiste en graficar sobre una línea en espiral los resultados obtenidos teniendo en cuenta su relevancia, en donde los cluster que se encuentran más retirados del centro son aquellos que contiene la información acorde a lo que solicitó el usuario[6].



[6] Ibid.

Clustifier



El método de consulta ofrecido por este programa requiere que se le establezcan los temas de interés a buscar solo en un 10%, mientras que el porcentaje restante es establecido automáticamente.

En versiones anteriores el programa Clustifier, era posible realizar una consulta web de contenidos bibliográficos obteniendo el resultado organizado por grupos[7].

BUSCADORES EN LA RED BASADOS EN CLUSTERING

Debido a la eficacia de esta forma de recuperación de información algunos buscadores han estructurado sus resultados para ser presentados al usuario utilizando clusters:

Vivísimo® Clusty iBoogie™ BETA Kart00 www.kartoo.com

Estos son solo cuatro ejemplos de buscadores a los cuales los usuarios de la red pueden acceder para hacer más ameno, rápido, eficiente y preciso la recuperación de información.

RESULTADOS ESPERADOS

Este documento busca invitar al usuario de la red al uso de herramientas de recuperación que le permitan moverse ágilmente en medio de grandes cantidades de información disponibles en el mundo de hoy. Esperamos sea una nueva guía para quienes lo consulten.

---

[7] DE SAMANIEGO, Javier. Op. Cit. Disponible en internet <http://es.geocities.com/clusteringrecuperacion/tipos.htm>

REFERENCIAS BIBLIOGRÁFICAS

CHALMERS, Matthew y CHITSON, Paul. Bead: Exploration in information visulization. [en línea], Cambridge, Rank Xerox Cambridge EuroPARC. [Citado 15 de marzo de 2007], Formato PDF, Disponible en Internet: <http://mmir.doc.ic.ac.uk/www-pub/npiv-sigir2000.pdf>.

DE SAMANIEGO, Javier. Recuperación y organización de la información: Clustering. [en línea], [Madrid-España?]. Universidad Carlos Tercero de Madrid, [Citado 15 de marzo de 2007], Página Web Formato de archivo: HTML, Disponible en Internet <http://es.geocities.com/clusteringrecuperacion/clustering.htm>

LAMARCA LAPUENTE, María Jesús. El nuevo concepto de documento en la cultura de la imagen, [en línea], Universidad Complutense de Madrid - Tesis doctoral, 2007, 01/04/2007, [Citado 15 de marzo de 2007], Formato HTML, Disponible en Internet: <http://www.hipertexto.info/documentos/busq_rec.htm>.

MARCOS MORA, Mari Carmen. La visualización en el proceso de búsqueda y recuperación de información [en línea]. En Rovira, C.; Codina, L. (dir.). Documentación digital. Barcelona: Sección Científica de Ciencias de la Documentación del Departamento de Ciencias Políticas y Sociales de la Universidad Pompeu Fabra, 2004. Formato PDF, Disponible en Internet <http://www.mcmarcos.com/pdf/2004_visualizacion-modd.pdf>. ISBN: 84-88042-39-6.

NAVAS, María Daniela. Un modelo de Clustering temporal, [en línea], Buenos Aires. 2004. 190 h. Trabajo de grado (Ingeniería en Informática). Universidad de Buenos Aires. Facultad de Ingeniería, [Citado 15 de marzo de 2007], Página Web Formato de archivo: PDF/Adobe Acrobat, Disponible en Internet <http://www.fi.uba.ar/materias/7500/navas-tesisdegradoingenieriainformatica.pdf>

PÁGINAS DE APOYO

http://in-spire.pnl.gov/
http://www.itl.nist.gov/iaui/vvrg/cugini/uicd/nirve-home.html
http://vivisimo.com/
http://clusty.com
http://www.iboogie.com/
http://kartoo.com