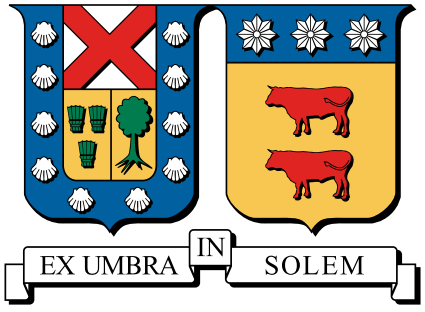
Universidad Técnica Federico Santa María



Distribución de Tópicos Emergentes en Conceptos Formales

Profesores Guía

Marcelo Mendoza

José Luis Martí

Memorista

Pablo Ortega Mesa

25 de febrero de 2016

Introducción

- ✓ Fuerte **crecimiento** de generadores de información.
- ✓ **Facilidad** de acceso a los datos.
- ✓ Generación de **comunidades** en torno a tópicos.

Introducción

- ✓ Fuerte **crecimiento** de generadores de información.
- ✓ **Facilidad** de acceso a los datos.
- ✓ Generación de **comunidades** en torno a tópicos.
- ✗ Capacidad **limitada** de procesamiento.
- ✗ Búsquedas **ineficientes** de datos.
- ✗ **Desinformación** producto de resultados errados.

Introducción

Problema de Exploración

¿Cuál data es **relevante** para las necesidades requeridas?

Introducción

Problema de Exploración

¿Cuál data es **relevante** para las necesidades requeridas?

Problema de Explotación

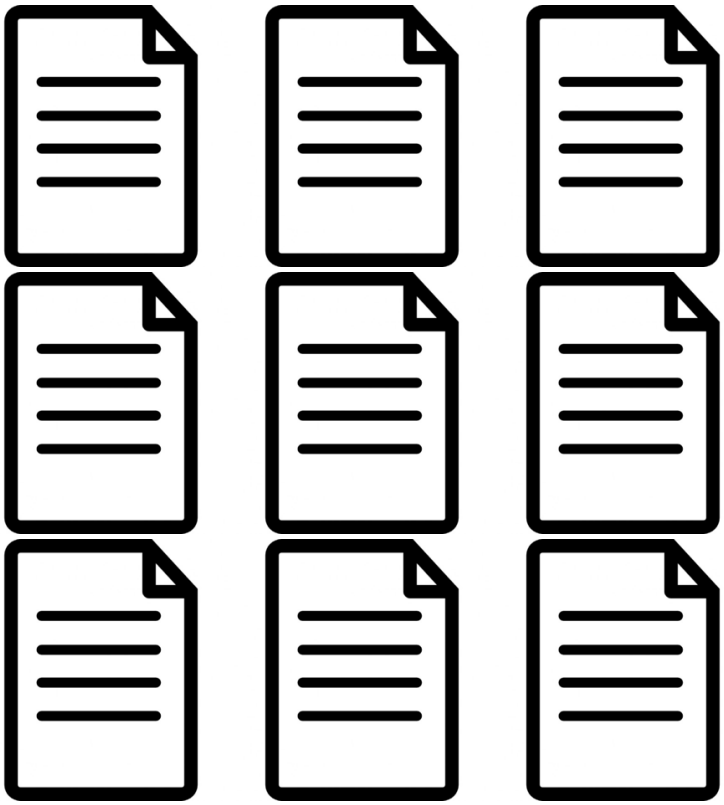
¿Cómo utilizar esa información **eficientemente**?

Problema

Con 1 documento...

Ningún problema señalado ocurre.

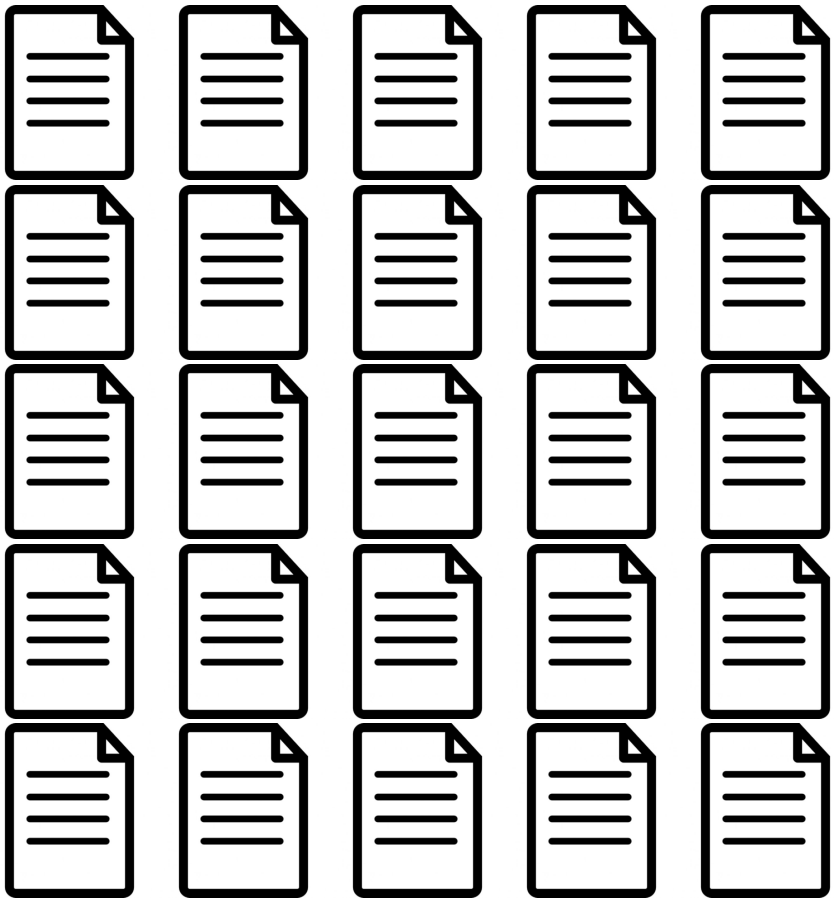
Problema



Con 9 documentos...

Todavía ningún problema señalado ocurre.

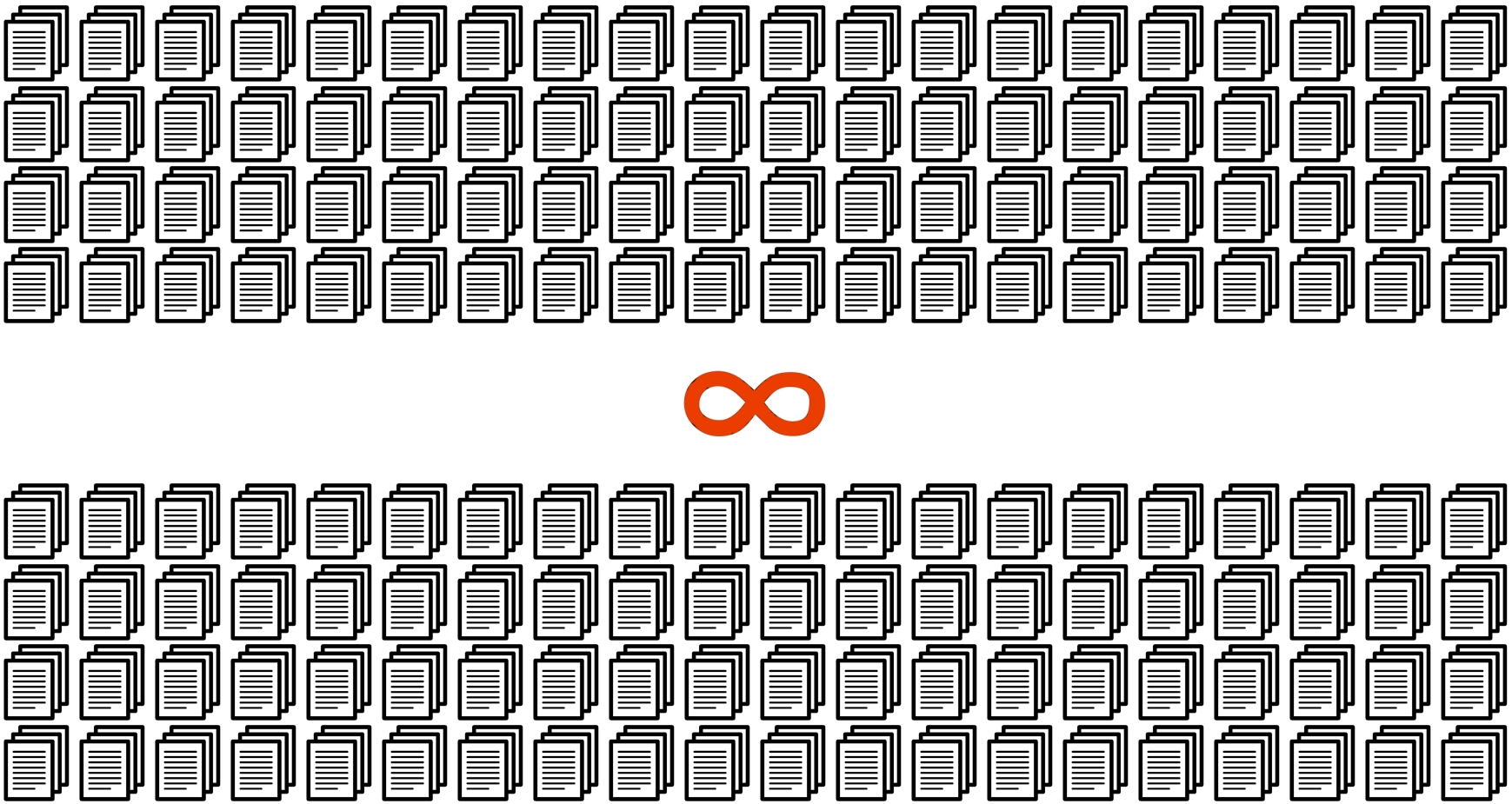
Problema



Con 25 documentos...

Quizás tomé un poco de tiempo, pero ningún problema señalado ocurre.

Problema



Problema

Con infinitos documentos...

- ✗ ¿Cómo puedo detectar los documentos relevantes a mis necesidades? (*Information Retrieval*).
- ✗ ¿De qué habla esta gran colección de documentos? (*Natural Language Processing*).
- ✗ Detección de *milestones* dentro de la colección.
- ✗ Navegación a través de los contenidos de la colección y no a través de los documentos que la componen (*Linked Data*).
- ✗ Generar una estructura que sea capaz de almacenar estos documentos de forma lógica y eficiente (*Information Retrieval*).
- ✗ Automatizar la tarea de mantener la colección de documentos vigente.

Problema

Comunidad Científica

- ¿Cómo puedo obtener aquellos artículos relacionados con la investigación que estoy haciendo o que deseo realizar?
- ¿Cómo puedo detectar quiénes han sido los precursores de las ideas detrás de las técnicas?
- Actualmente, resolver estás preguntas **consume gran parte** del tiempo de un investigador científico.

Solución

Procesamiento de la información

- La colección documental utilizada es la base de datos bibliográfica **DBLP**.
- Procesar enormes colecciones documentales a través de técnicas de *Information Retrieval* como lo es *Formal Concept Analysis* y dentro del ámbito de *Topic Modeling* se utilizará la técnica *Latent Dirichlet Allocation*.

Solución

Resumen Interactivo

Generar una visualización que permita

- Navegar a través de los distintos “conceptos formales”.
- Analizar la distribución de “tópicos emergentes”.

Marco Teórico - *DBLP*

- Plataforma Web alojada en Alemania que contiene artículos científicos relacionados con ciencias de la computación.
- En los años 80's fue una base de datos pequeñas relacionada a través de programación lógica.
- Contiene artículos de las revistas *VLDB*, *IEEE*, *ACM*, además de distintas conferencias.

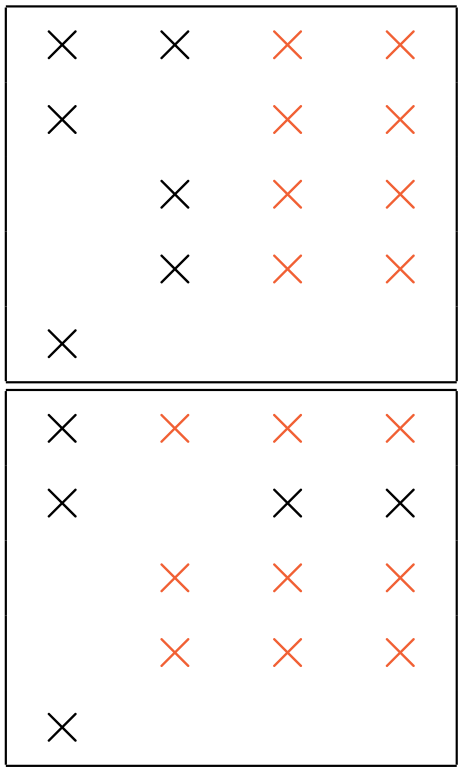
Marco Teórico - *FCA*

- Método de análisis de datos.
- Analiza la información que describe la relación entre un particular conjunto de objetos y atributos.
- Produce dos salidas
  - *Concept Lattice*
  - Implicaciones de atributos

Marco Teórico - *FCA*

I	y_1	y_2	y_3	y_4
x_1	×	×	×	×
x_2	×		×	×
x_3		×	×	×
x_4		×	×	×
x_5	×			

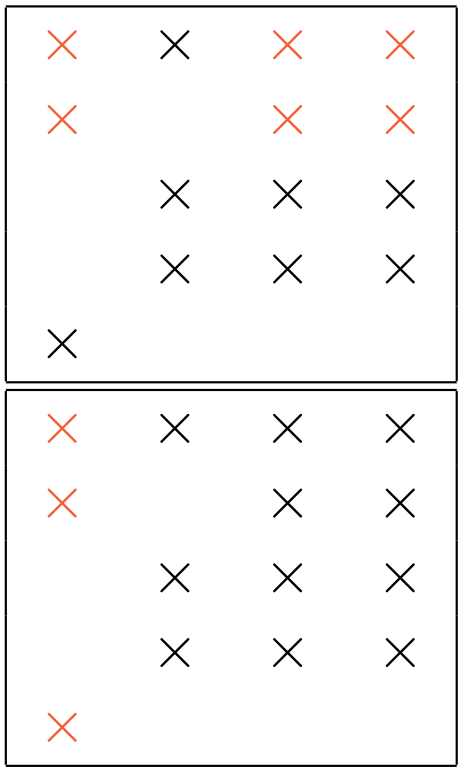
Cuadro: Contexto Formal



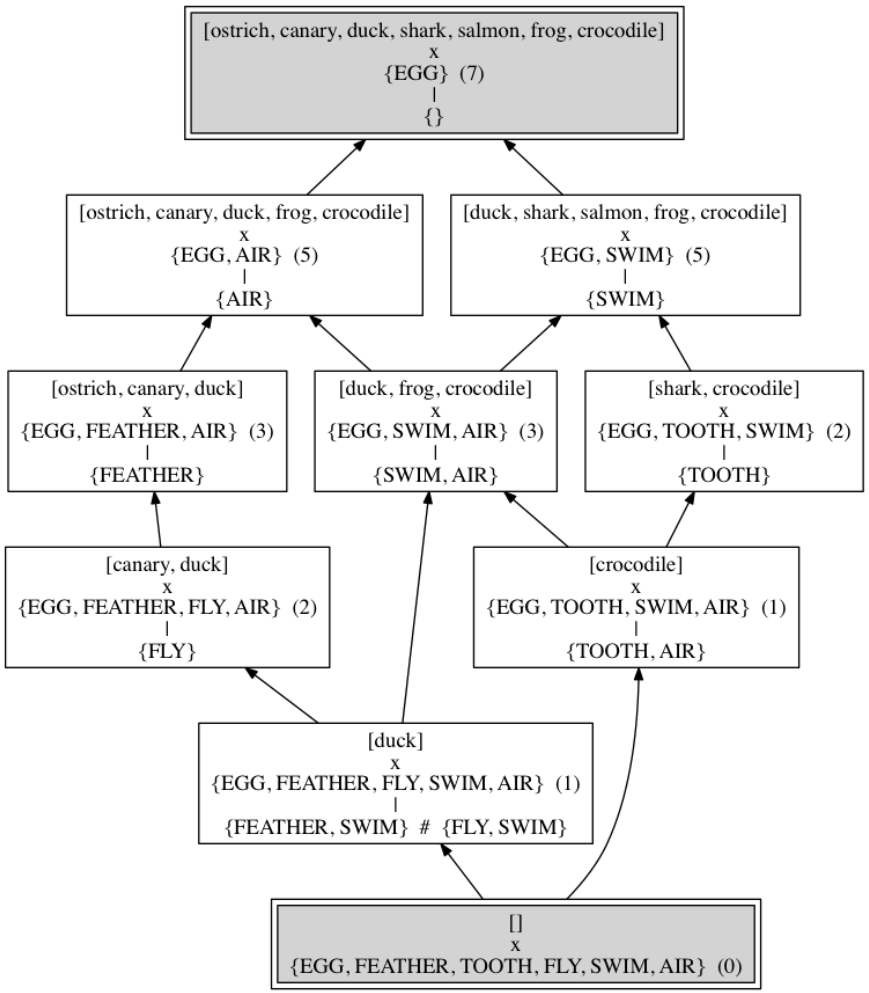
Marco Teórico - *FCA*

I	y_1	y_2	y_3	y_4
x_1	×	×	×	×
x_2	×		×	×
x_3		×	×	×
x_4		×	×	×
x_5	×			

Cuadro: Contexto Formal



Marco Teórico - *FCA*



Marco Teórico - *FCA*

Soporte Mínimo

El **support** de un concepto formal dado por $\langle A, B \rangle$, donde $A \subseteq X$ y $B \subseteq Y$ está definido por:

$$supp(\langle A, B \rangle) = \frac{|A|}{|X|}$$

Frequent Concept

Dado un umbral $minsupp \in [0, 1]$, entonces el concepto $\langle A, B \rangle$ es llamado *Frequent Concept* si y sólo si $supp(\langle A, B \rangle) \geq minsupp$.

Marco Teórico - *FCA*

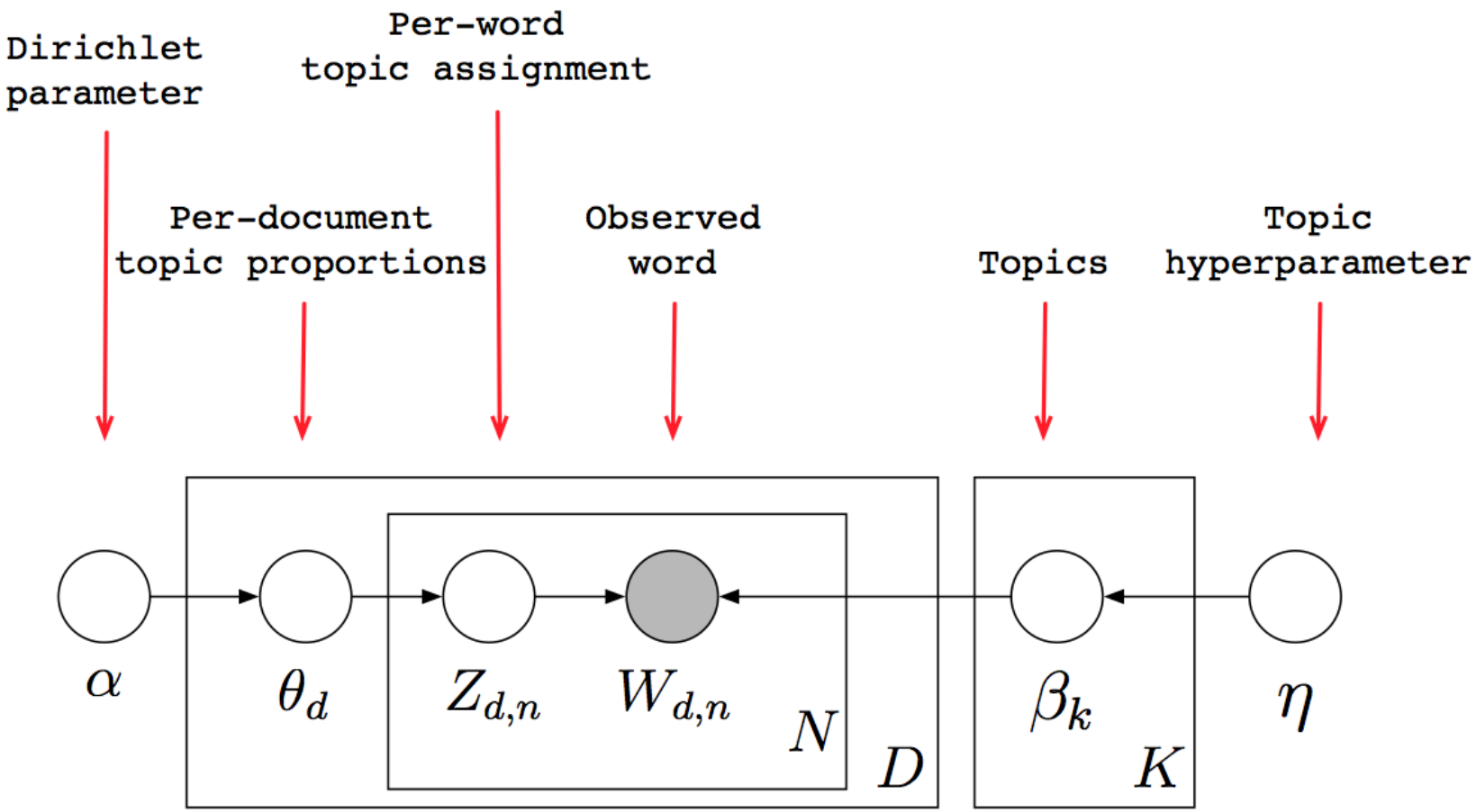
Iceberg Lattice

Un **Iceberg Lattice** es el conjunto de todos los *Frequent Concepts* dado un *minsupp*

Marco Teórico - *LDA*

- Modelo perteneciente al área *Topic Modeling*
- Busca descubrir tópicos a partir de una gran colección de documentos.
- **LDA** asume que:
  - Un documento *D* habla sobre un conjunto limitado de **Tópicos**.
  - Un *tópico* se compone a través de un **vocabulario fijo**.
- LDA es un proceso generativo que utiliza técnicas de **Inferencia Estadística** para detectar los tópicos de una gran cantidad de datos.

Marco Teórico - *LDA*

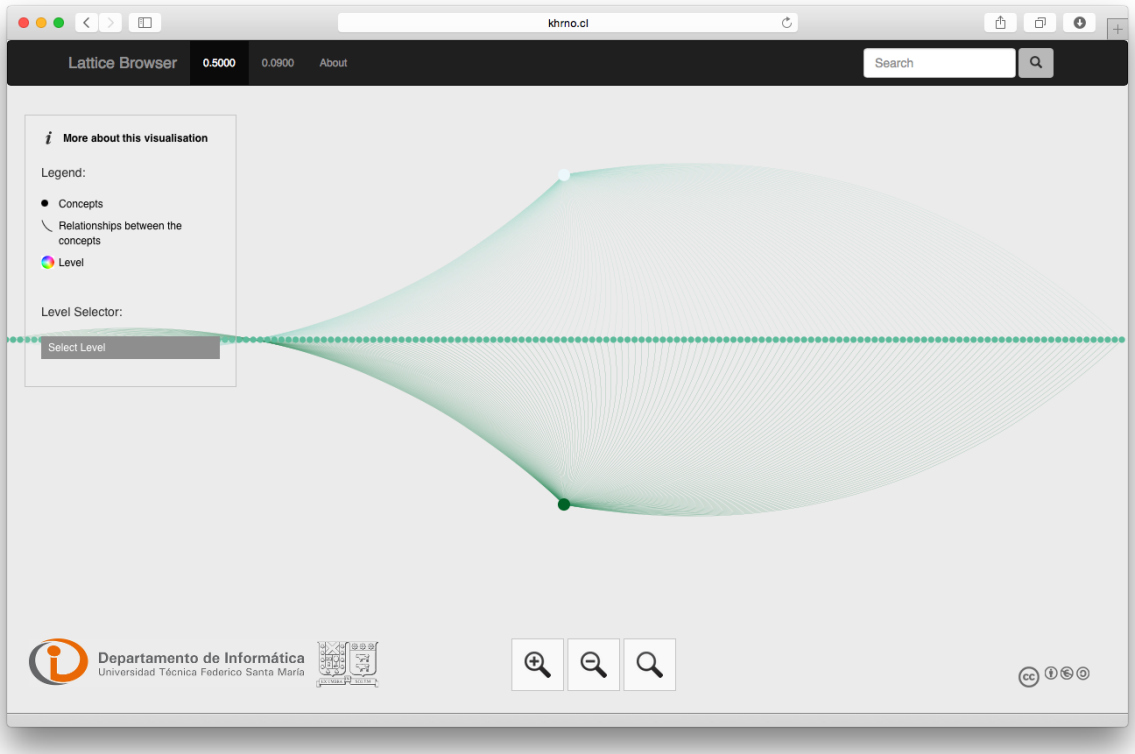


Marco Teórico - *D3.js*

- Librería *Open Source* de *Javascript*
- Ideada para manipular documentos basados en información.
- Componente fuerte en manipulación del *DOM* de un sitio web.
- Ideal para generar **herramientas interactivas**

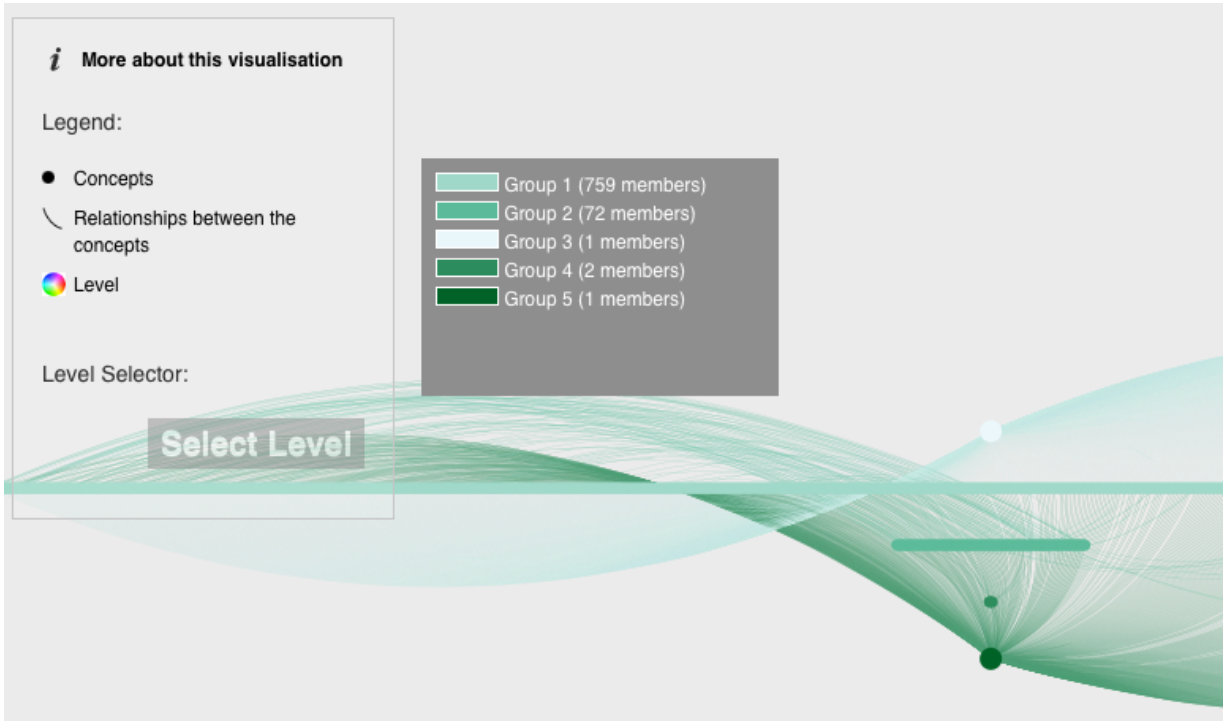
Resultados

Visión Global del Lattice



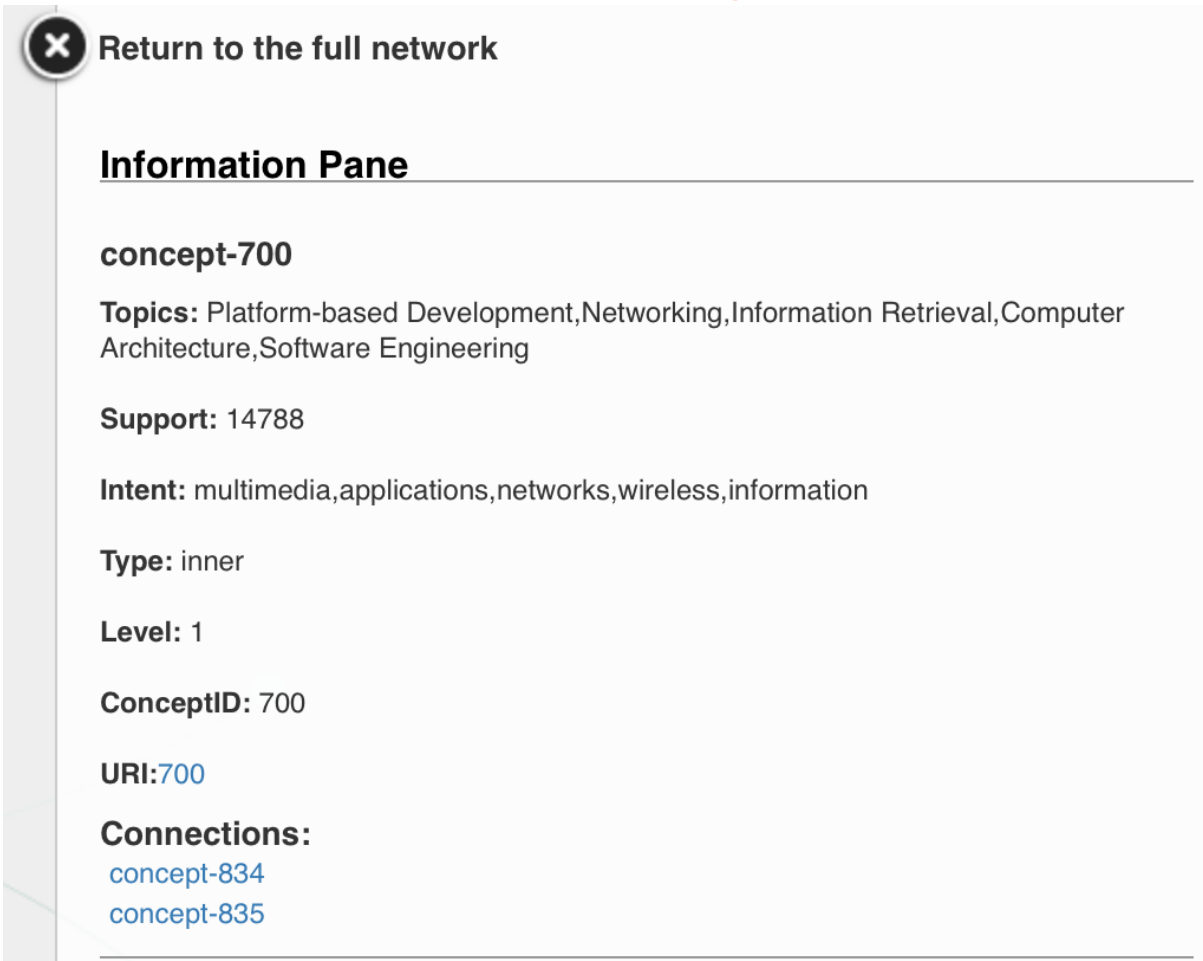
Resultados

Selector de nivel



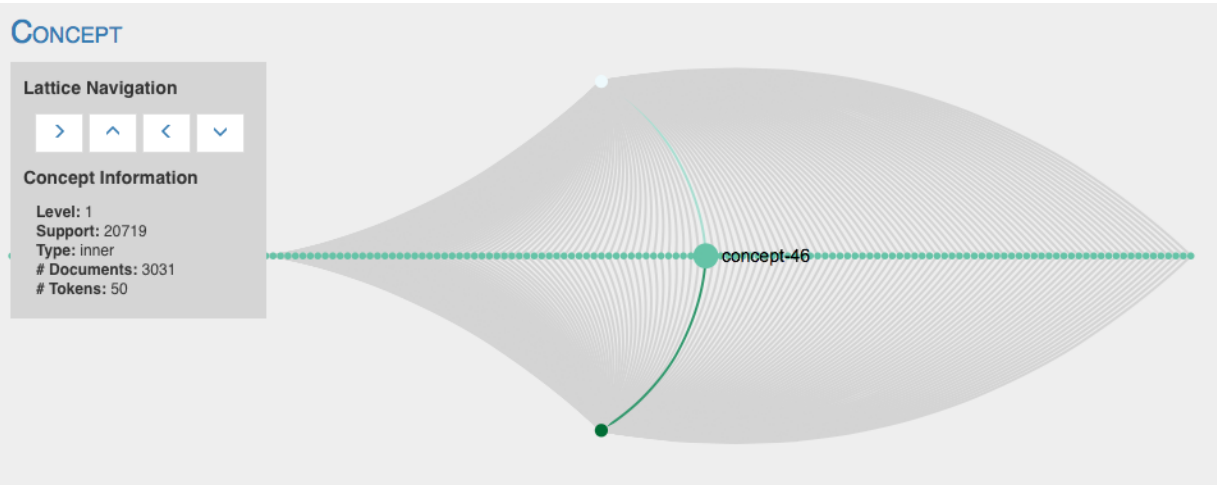
Resultados

Preview del Concepto Formal



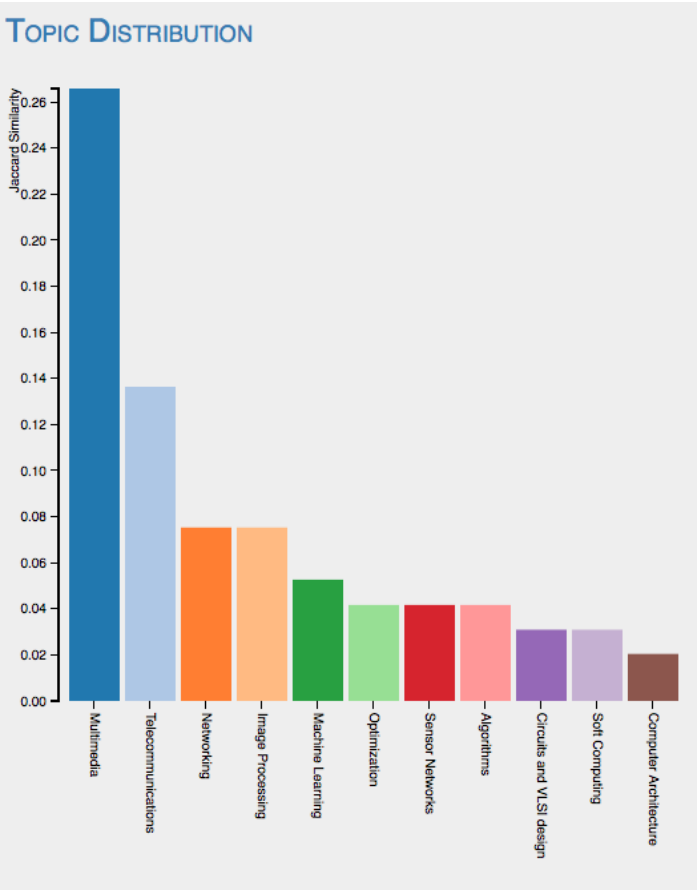
Resultados

Detalle del Concepto Formal



Resultados

Detalle del Concepto Formal



Resultados

Detalle del Concepto Formal

Resultados

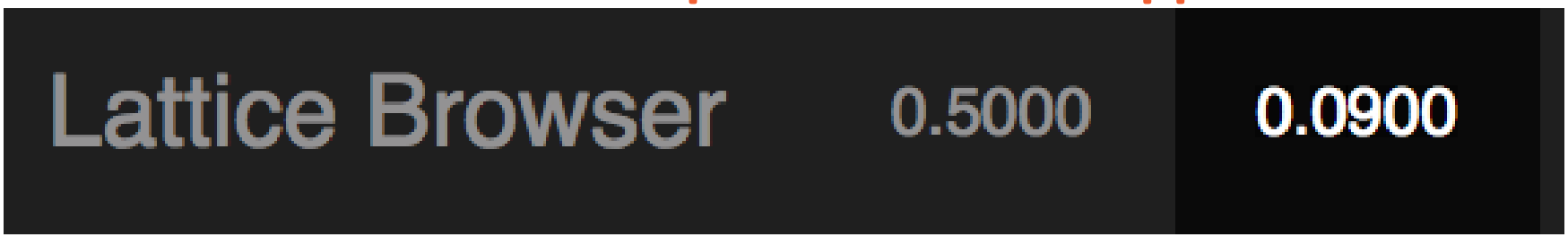
Detalle del Concepto Formal

Document List

- Estimation of short-term predictor parameters for coding and enhancement of noisy speech. 1982
- An embedded image coding system based on tarp filter with classification. 1982
- Channel coding considerations for digital speech encoded by linear prediction. 1982
- A new mode selection technique for coding Depth maps of 3D video. 1982
- Directional coding of audio using a circular microphone array. 1982
- Universal coding for quasi-stationary processes. 1982
- Adaptive lifting for multicomponent image coding through quadtree partitioning. 1982
- A New Bidirectionally Motion-Compensated Orthogonal Transform for Video Coding. 1983
- Mixed-domain coding of speech at 3 kb/s. 1983
- Efficient coding of high resolution typographic characters. 1983

Resultados

Selector de parámetro minsupp



Conclusiones

1. Problema de la **dimensionalidad**.
2. Fuerte relación entre las técnicas utilizadas.
3. Librerías gráficas flexibles entregando al usuario una gran capacidad de interactuar / navegar.
4. Resumen visual, interactivo y navegable de una gran colección de datos.

Trabajo Futuro

1. Extender este trabajo para analizar las **redes sociales** que forman los autores / consumidores en torno a los tópicos descubiertos.
2. Extender el análisis para incluir un **análisis de sentimientos**.
3. Crear componente para la tarea de la **recolección** de datos.
4. Monitoreo de redes sociales.
5. Alertas tempranas de eventos específicos.
6. Muchas otras aplicaciones...

Preguntas

Gracias