Técnicas de Text Mining para mapeos bibliográficos

Federico Ferrero

Seminario - Universidad Santiago de Compostela
Julio de 2020

Generación de datos (de tipo textual) en investigación

Producción de datos

- Instrumentos tradicionales para producir datos.
    - Entrevistas
    - Encuestas
    - ...

Recolección de datos

- de la Web
    - Tweets
    - Referencias bibliográficas
    - Comentarios en foros...

Objetivos

1. Descargar referencias bibliográficas de la Web of Science vinculadas a un tema específico.

   Luego...

2. Trabajar con técnicas de Text Mining para realizar un mapeo bibliográfico sobre la producción científica sobre dicho tema.

Big Data

¿Cómo analizar grandes volúmenes de datos?

¿Teoría Fundamentada? ...

Text Mining:

"is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources" (Marti Hearst, 2003)

http://people.ischool.berkeley.edu/~hearst/text-mining.html

Descubrimiento?

Objetivismo y big data?

Trabajar con Minería de Texto

Softwares:

- **WEKA: Waikato Environment for Knowledge Analysis** (**Weka**), developed at the University of Waikato, New Zealand. It is free software licensed under the GNU General Public License, and the companion software to the book "Data Mining: Practical Machine Learning Tools and Techniques"

  https://www.cs.waikato.ac.nz/ml/weka/

  https://en.wikipedia.org/wiki/Weka_(machine_learning)

  Libro: https://www.cs.waikato.ac.nz/~ml/weka/book.html

- **SCIMAT: Science Mapping Analysis Tool**

Uso de lenguajes de programación:

Por ejemplo **R**

SCIMAT tool

Science Mapping Analysis Tool

- **SciMAT** es un software de **código abierto** desarrollado por la Universidad de Granada, España (Cobo, López-Herrera, Herrera-Viedma and Herrera, 2011)

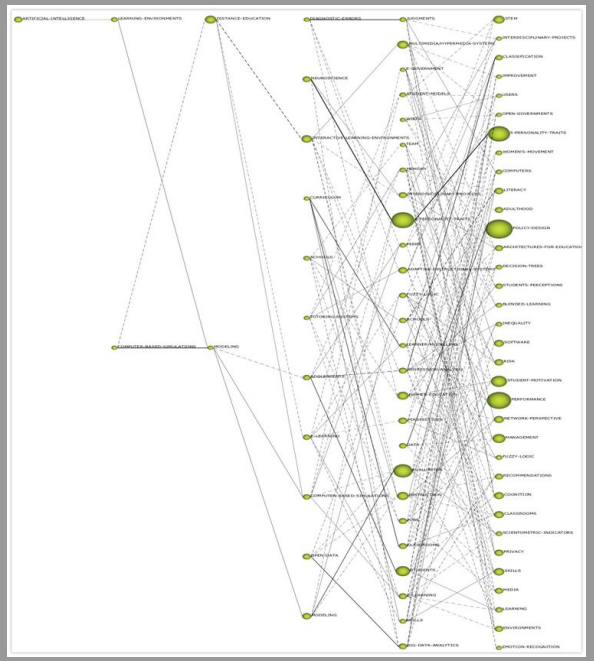
- Provee un modo de llevar adelante **Systematic Literature Review**.

- Permite realizar **mapeos de la producción científica** sobre un determinado tema desde un punto de vista longitudinal.

- 3 módulos:
  1. **Manejo y administración** de la información.
  2. **Análisis** (mapeo científico). **Visualización** de los resultados en mapas.
- https://sci2s.ugr.es/scimat/

4 tipo de diagramas

Evolución temática



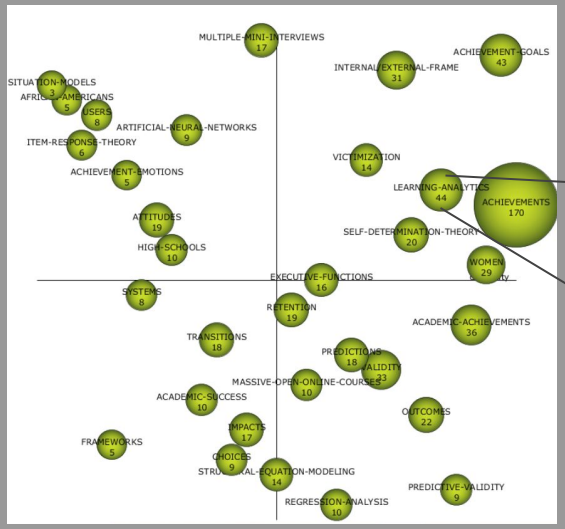
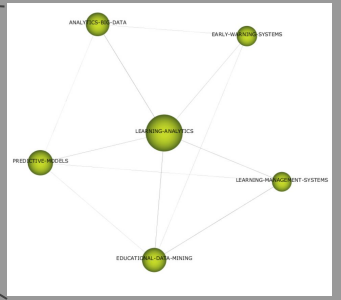


Diagrama estratégico

Diagrama temático



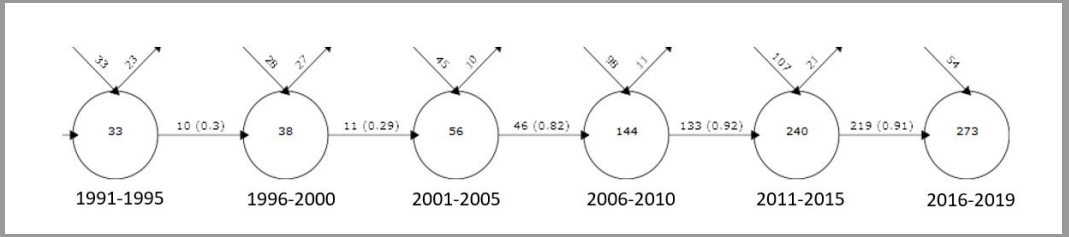


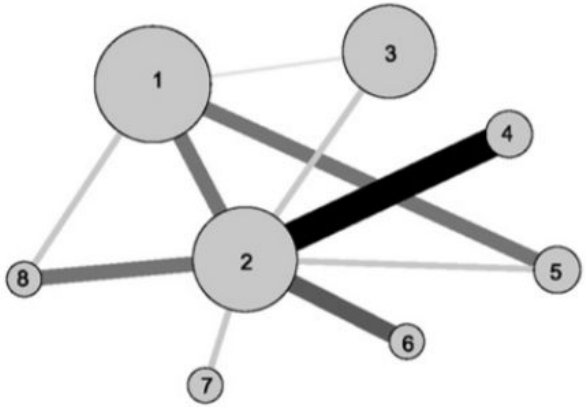
Diagrama de estabilidad

4 tipos de diagramas

Equivalence Index = When the keywords always appear together, the equivalence index equals unity; when they are never associated, it equals zero.

1) Thematic diagram

- **Clusters** de keywords.
- Grupos de nodos que representan diferentes keywords conectadas.
- **Nodo** = keyword.
- El grupo entero de keywords es llamado “tema”.
- **Nombre del tema** = nodo/keyword central.
- **Volumen nodo** = proporcional al número de documentos correspondientes a cada keyword.
- **Grosor de las línea** entre nodos = proporcionales al *Equivalence Index*.



Example of Thematic diagram extracted from Cobo, López-Herrera, Herrera-Viedma and Herrera (2011)

4 tipos de diagramas

Thematic diagram para LEARNING-ANALYTICS

Periodo 2015-2020

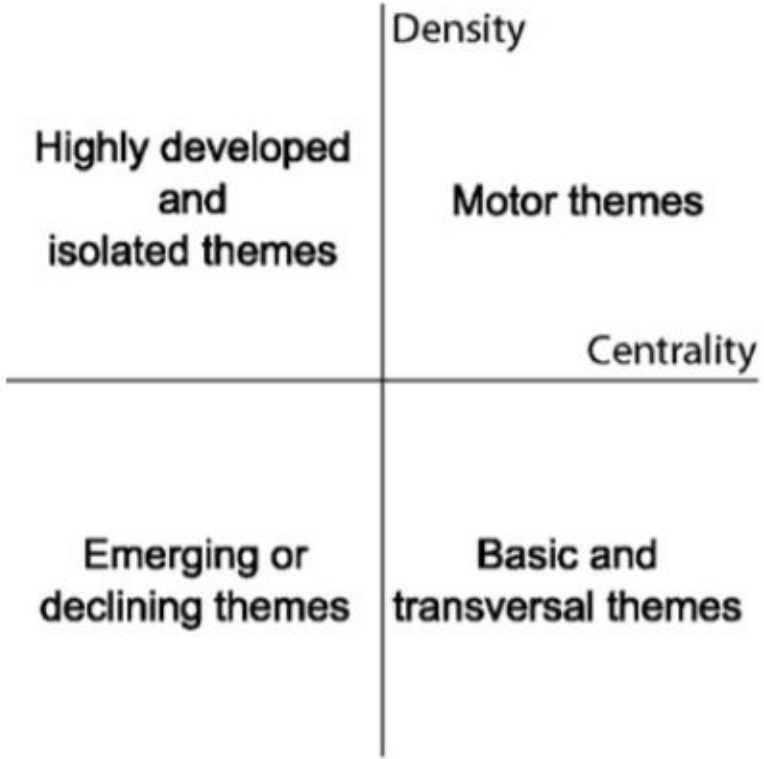
Cluster info:

- Name: LEARNING-ANALYTICS
- Density: 5.17
- Density range: 0.69
- Centrality: 10.68
- Centrality range: 0.84

4 tipos de diagramas

2) Strategic diagram

- Localiza “temas” de acuerdo a dos parámetros:
  - **CENTRALIDAD** = mide el grado de interacción de una red con otras redes. Mide la fuerza de los vínculos externos con otros temas. Es medida de la importancia de un tema en el desarrollo del campo de investigación analizado.
  - **DENSIDAD** = mide la fuerza de vínculos internos entre todas las keywords que forman parte de un tema de investigación. Es medida de el desarrollo interno de un determinado tema.



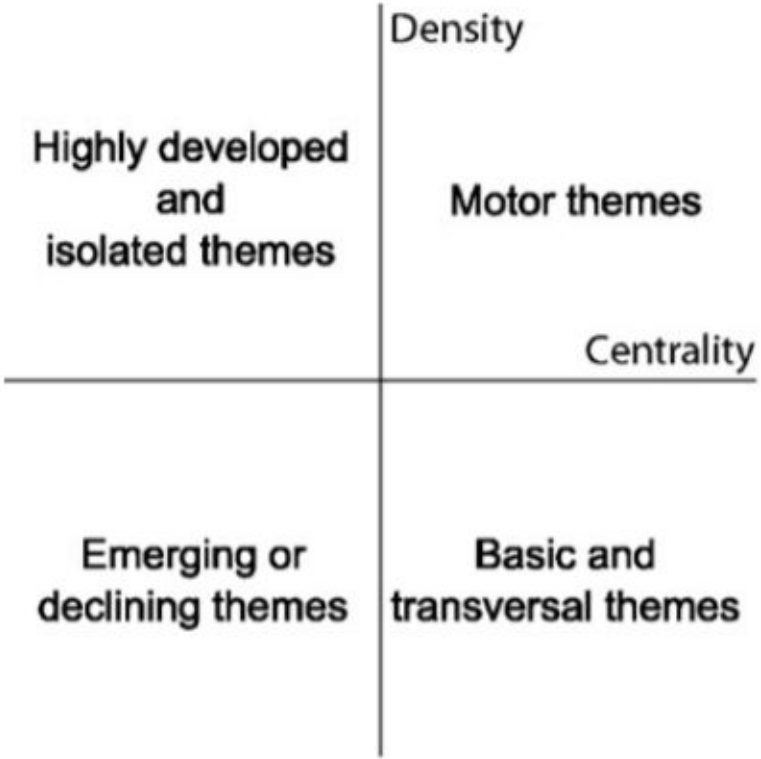
Example of Thematic diagram extracted from Cobo, López-Herrera, Herrera-Viedma and Herrera (2011)

4 tipos de diagramas

2) Strategic diagram

- Este diagrama se organiza en 4 cuadrantes:

1. **TEMAS EMERGENTES O EN DESAPARICIÓN** (abajo-izquierda) = baja centralidad y densidad
2. **TEMAS ALTAMENTE DESARROLLADOS PERO AISLADOS** (arriba-izquierda) = alta densidad pero baja centralidad
3. **TEMAS BÁSICOS Y TRANSVERSALES** (abajo-derecha) = alta centralidad pero baja densidad
4. **TEMAS MOTORES** (arriba-derecha) = alta centralidad y densidad: son temas importantes, conectados con otros temas y bien desarrollados internamente

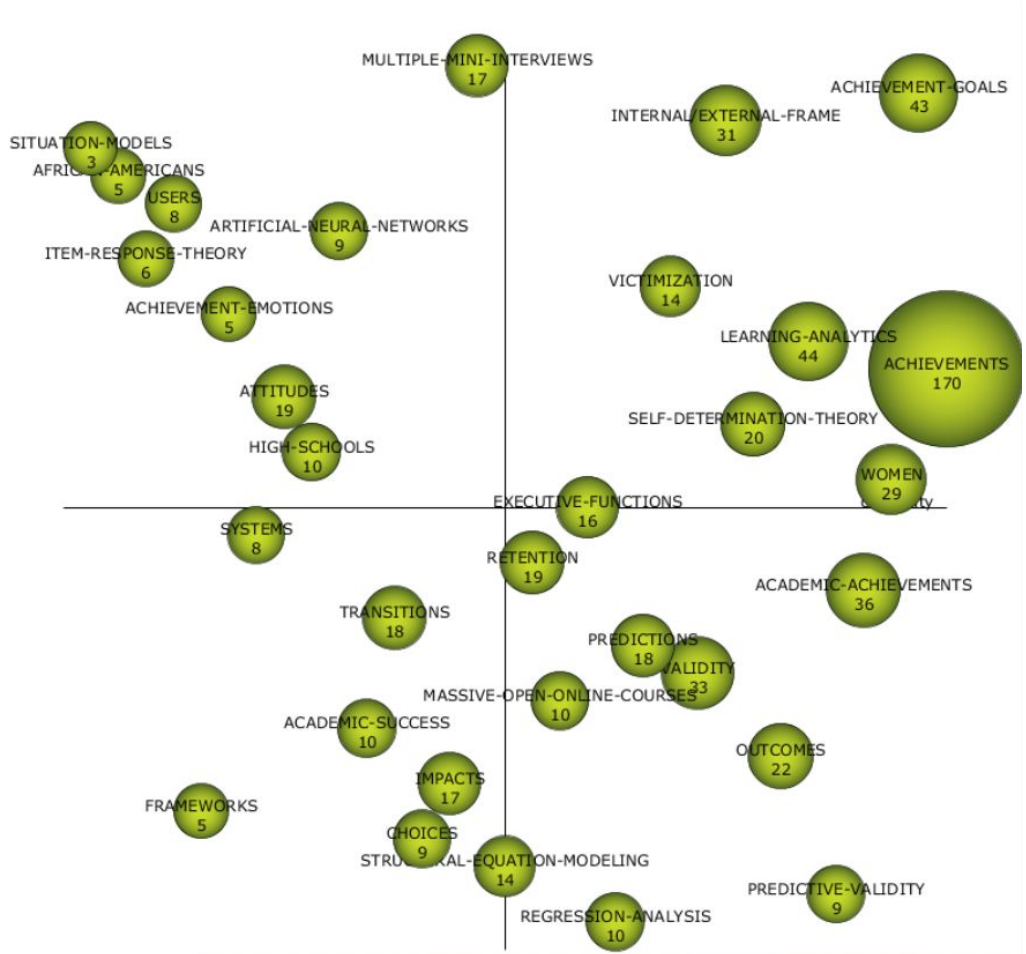


Example of Thematic diagram extracted from Cobo, López-Herrera, Herrera-Viedma and Herrera (2011)

4 tipos de diagramas

Ejemplo de diagrama estratégico.
Temas vinculados a “predicción en educación” en periodo 2015-2020.

Puedo obtener un diagrama estratégico para cada periodo y analizar cómo los temas van evolucionando.



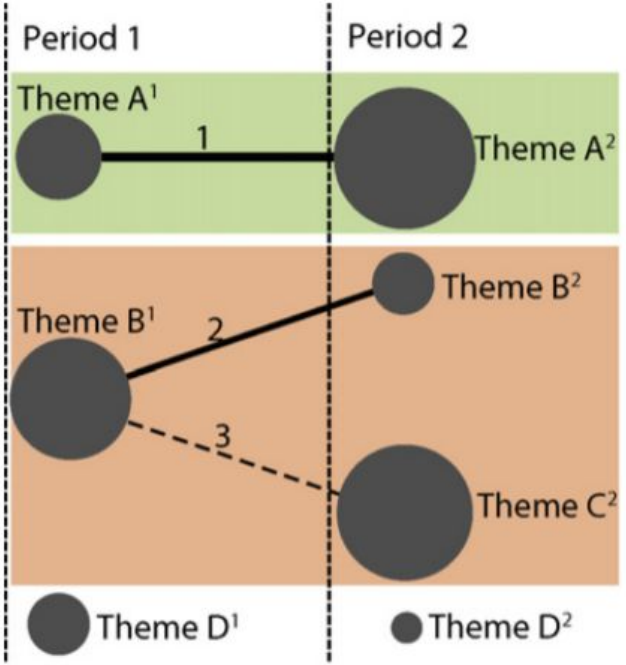
4 tipos de diagramas

Inclusion Index = será igual a 1 si las keywords de un tema están totalmente contenidas en el tema siguiente.

3) *Thematic evolution diagram*

Muestra la evolución temática del campo de investigación bajo análisis.

- **Periodos**: organizados verticalmente.
- **Temas**: nodos se ubican en cada uno de los periodos y se pueden vincular con temas del siguiente periodo.
- **Línea sólida**: los temas vinculados comparten el mismo nombre (both themes have the same name or the name of one of them is part of the other one).
- **Línea de puntos**: los temas vinculados comparten elementos que no son el nombre de los temas.
- **Grosor de la línea**: proporcional al *Inclusion Index*.
- **Volúmenes de los nodos**: proporcional al número de documentos publicados asociados con cada tema.

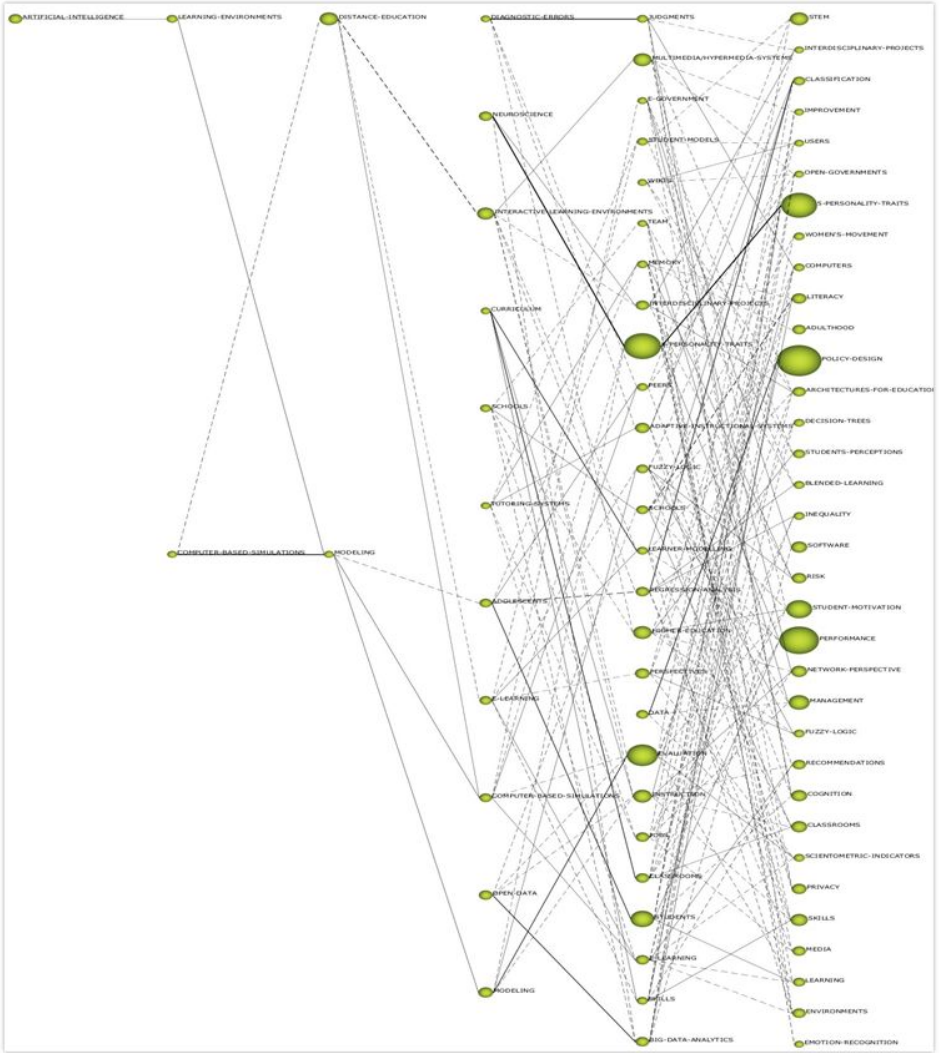


Example of Thematic Evolution diagram extracted from Cobo, López-Herrera, Herrera-Viedma and Herrera (2011)

4 tipos de diagramas

Ejemplo de Thematic evolution diagram

Evolución de temas vinculados a producciones académicas sobre Inteligencia Artificial en educación (1991-2019)

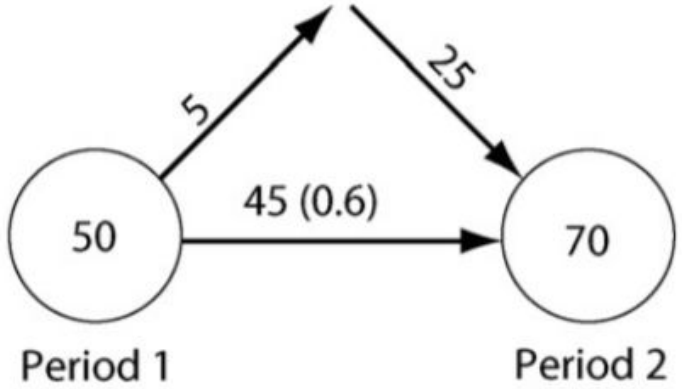


4 tipos de diagramas

4) Stability between periods diagram

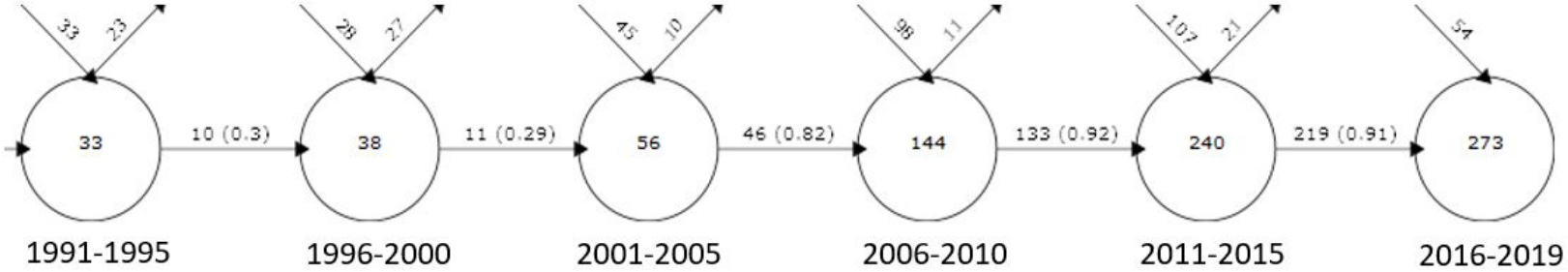
Muestra medidas de estabilidad entre periodos consecutivos (o cuánto se superponen).

- **Círculos** = periodos.
- **Números en círculos** = cantidad de keywords asociadas a cada periodo.
- **Flecha horizontal** = keywords compartidas por los dos periodos consecutivos.
- **Número en paréntesis** = *Similarity Index*.
- **Flecha hacia arriba** = número de keywords salientes.
- **Flecha hacia abajo** = número de nuevas keywords entrantes en el periodo.



Example of Stability diagram extracted from Cobo, López-Herrera, Herrera-Viedma and Herrera (2011)

4 tipos de diagramas

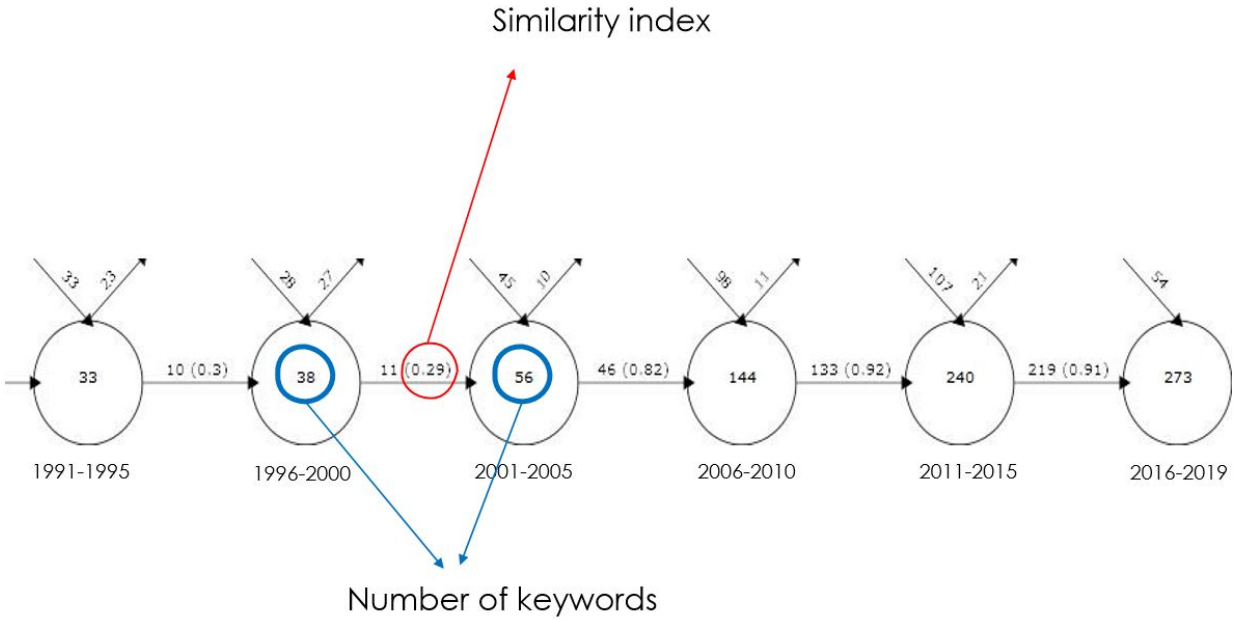


Ejemplo de Stability between periods diagram

Evolución de temas vinculados a producciones académicas sobre Inteligencia Artificial en educación (1991-2019)

4 tipos de diagramas

- **# of keywords increases**: **33** (1991-1995) to **273** (2016-2019).
- **Jump in 2006-2010**: 56 to 144.
- **Similarity Index increases** from 0.3 to 0.91: terminology is shared and maintained while the research field is consolidated.

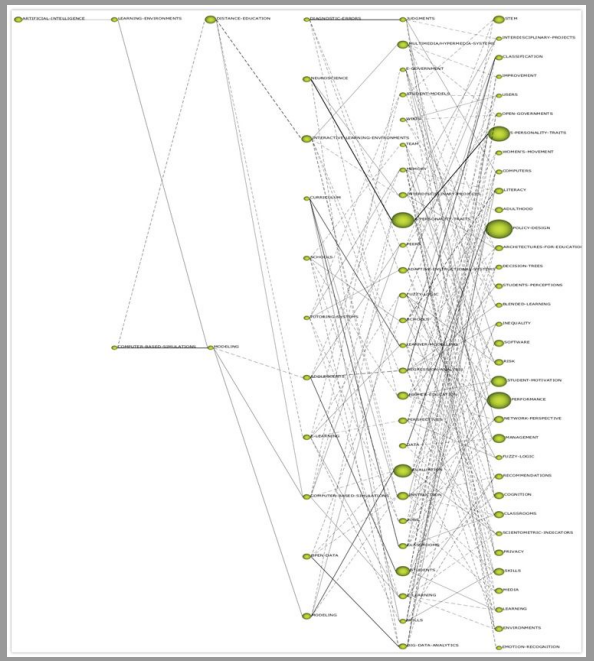


Ejemplo de Stability between periods diagram

Evolución de temas vinculados a producciones académicas sobre Inteligencia Artificial en educación (1991-2019)

En resumen:
4 tipo de diagramas

Evolución temática



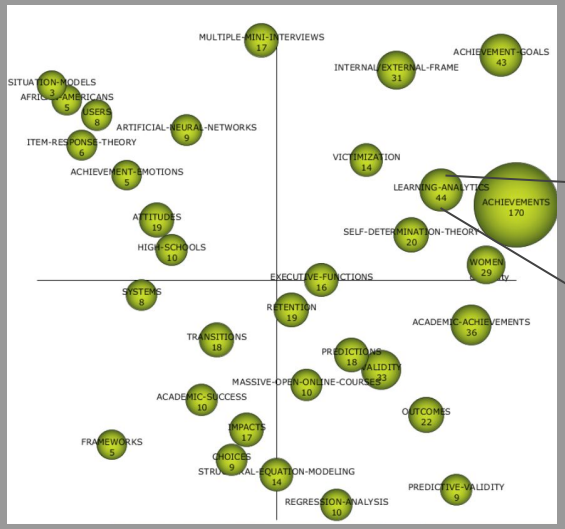


Diagrama estratégico

Diagrama temático

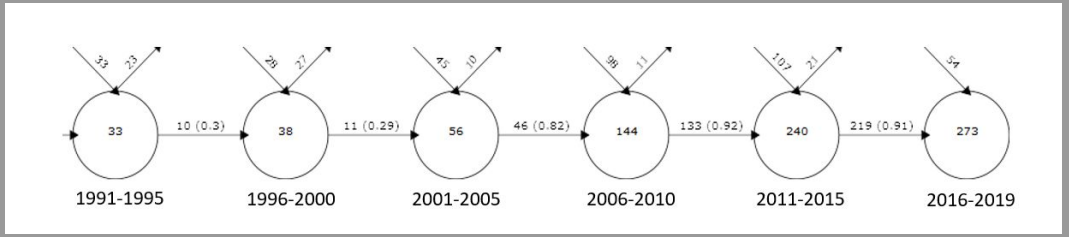


Diagrama de estabilidad

¿Cómo construir estos diagramas?

1) Descargar datos bibliográficos de la Web of Science
2) Normalizar los datos y analizarlos con SCIMAT

Web of Science: descargar referencias bibliográficas

Web of Science Core Collection (ISI WoS)

(TS= (big data AND education)) *AND* IDIOMA: (English) *AND* TIPOS DE DOCUMENTOS: (Article)
Refinado por: CATEGORÍAS DE WEB OF SCIENCE: (EDUCATION EDUCATIONAL RESEARCH)
Período de tiempo: Todos los años. Índices: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI.

Tener en cuenta que Web of Science sólo permite descargar grupos de hasta 500 referencias por vez. Descargar cada grupo de referencias y luego combinar todos los archivitos .txt en uno solo. https://www.youtube.com/watch?v=k2bT_hHcbgE

Descargar SCIMAT

Descargar el software

https://sci2s.ugr.es/scimat/download.html

Verificar tener descargados:

1. JAVA
2. SCIMAT

Guía elaborada por los autores:
https://sci2s.ugr.es/scimat/software/v1.01/SciMAT-v1.0-userGuide.pdf

Ya trabajando con SCIMAT

---

1. **Crear un proyecto**.
2. **Abrir archivo de texto** (en nuestro caso descargado desde Web of Science).
3. **Normalización** of keywords (automática y manual).
4. Crear **periodos**.
5. **Análisis**.
6. **Interpretar** diagramas:
    - Thematic diagram
    - Strategic diagram
    - Stability between periods diagram
    - Thematic evolution diagram

Ejemplo de análisis preliminar: Inteligencia Artificial en la producción científica del campo educativo

https://federico-jf.github.io/work_samples/EPPS%206302%20Presentation%20Federico%20Ferrero.pdf

2 Referencias

M.J. Cobo, A.G. López-Herrera, E. Herrera-Viedma and F. Herrera, **An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field**, *Journal of Informetrics*, vol. 5, num. 1, pp. 146-166, 2011. https://www.sciencedirect.com/science/article/abs/pii/S1751157710000891

M.J. Cobo, A.G. López-Herrera, E. Herrera-Viedma and F. Herrera, **SciMAT: A new Science Mapping Analysis Software Tool**. *Journal of the American Society for Information Science and Technology*, 63:8 (2012) 1609-1630 doi: **10.1002/asi.22688** https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.22688