Sistemas de Recuperación de Información

Introducción a los SRI

Lic. Carlos León González

Facultad de Matemática y Computación
Universidad de La Habana

6 de febrero de 2026

https://t.me/SRI_25_26

Objetivos de la asignatura

- Asimilar las características de los Sistemas de Recuperación de Información, su modelación matemática-computacional y su desarrollo e implementación.
- Asimilar y desarrollar herramientas para la búsqueda, extracción, almacenamiento y recuperación de información.
- Aplicar la modelación matemática al análisis y evaluación de los sistemas y fuentes de información.

Cada minuto de Internet en 2023

Tomado de https://ediscoverytoday.com/2023/04/20/2023-internet-minute-infographic-by-ediscovery-today-and-ltmg-ediscovery-trends/

Cada minuto de Internet en 2023

¿Toda la información puede almacenarse en una base de datos relacional?

Cada minuto de Internet en 2023

¿Toda la información puede almacenarse en una base de datos relacional?

Entonces, ¿dónde y cómo puede almacenarse la información?

Cada minuto de Internet en 2023

¿Toda la información puede almacenarse en una base de datos relacional?

Entonces, ¿dónde y cómo puede almacenarse la información?

¿Cómo el sistema es capaz, de forma automática, detectar “nueva” información?

Cada minuto de Internet en 2023

¿Toda la información puede almacenarse en una base de datos relacional?

Entonces, ¿dónde y cómo puede almacenarse la información?

¿Cómo el sistema es capaz, de forma automática, detectar “nueva” información?

¿Cómo se inserta esa información en la “base de datos” para que sea coherente con la realidad?

Cada minuto de Internet en 2023

¿Toda la información puede almacenarse en una base de datos relacional?

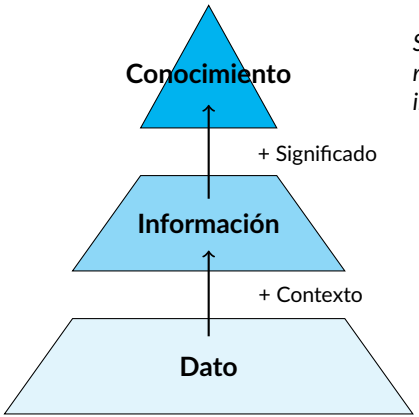
Entonces, ¿dónde y cómo puede almacenarse la información?

¿Cómo el sistema es capaz, de forma automática, detectar “nueva” información?

¿Cómo se inserta esa información en la “base de datos” para que sea coherente con la realidad?

¿Cómo recuperar la información requerida, si la consulta es definida en lenguaje natural?

¿Qué es la información?

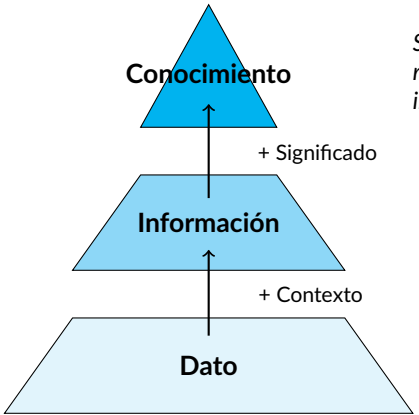


Saber cómo, comprensión, experiencia, percepción, intuición e información contextualizada.

Datos contextualizados, categorizados y calculados.

Hechos y cifras que reflejan algo específico, pero que no están organizados de ninguna manera.

¿Qué es la información?

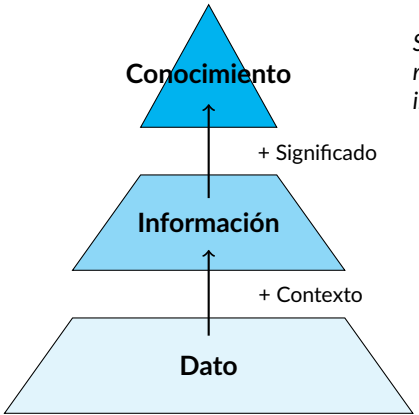


Saber cómo, comprensión, experiencia, percepción, intuición e información contextualizada.

Datos contextualizados, categorizados y calculados.

78.73

¿Qué es la información?

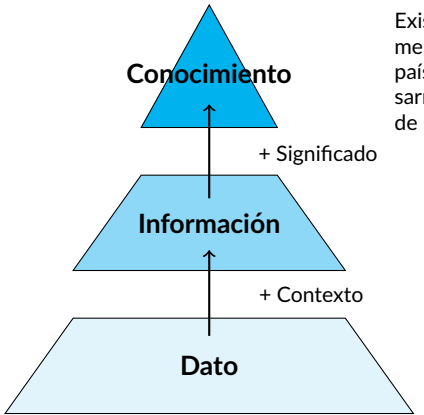


Saber cómo, comprensión, experiencia, percepción, intuición e información contextualizada.

78.73 es la esperanza de vida en Cuba.

78.73

¿Qué es la información?



Existe una tendencia sostenida al aumento de la esperanza de vida en el país posiblemente provocado por el desarrollo tecnológico y las condiciones de vida de la población.

78.73 es la esperanza de vida en Cuba.

78.73

Recuperación de Datos VS Recuperación de Información

Criterio	Datos	Información
Lenguaje de consulta	Lenguaje artificial	Lenguaje natural
Pregunta	Completa	Difusa
Respuesta	Adecuada	Por relevancia
Correspondencia	Exacta	Parcial
Representación	Por registros	Variada
Procesamiento	Transacciones, SQL	Análisis semántico, NLP
Almacenamiento	Estructurado (tablas)	No estructurado (textos, imágenes)
Objetivo	Almacenar y recuperar datos exactos	Proporcionar conocimiento útil
Herramientas	MySQL, PostgreSQL	Google, Elasticsearch
Contexto	Menos dependiente	Altamente dependiente

¿Qué es un Sistema de Recuperación de Información?

La **Recuperación de Información** es la localización de materiales (generalmente documentos) de naturaleza no estructurada (generalmente texto) para satisfacer una necesidad de información en una larga colección (generalmente almacenada en computadoras).

Manning, C. D., Ïntroduction to Information Retrieval", Cap. 1, pág. 1

¿Qué es un Sistema de Recuperación de Información?

La **Recuperación de Información** es la localización de materiales (generalmente documentos) de naturaleza no estructurada (generalmente texto) para satisfacer una necesidad de información en una larga colección (generalmente almacenada en computadoras).

Manning, C. D., Ïntroduction to Information Retrieval", Cap. 1, pág. 1

Luego, aquellos sistemas que implementan la recuperación de información para obtener documentos, registros, imágenes, sonidos, etc., e interaccionan con usuarios o servicios que requieren obtener información, son conocidos como **Sistemas de Recuperación de Información**.

Definición formal

Un **modelo de recuperación de información** es un cuádruplo $[D, Q, F, R(q_j, d_j)]$ donde:

D es un conjunto de representaciones lógicas de los datos de la colección.

Q es un conjunto compuesto por representaciones lógicas de las necesidades del usuario, denominadas “consultas”.

F es un framework para modelar las representaciones de los datos, consultas y sus relaciones.

R es una función de ranking que asocia un número real a una consulta $q \in Q$ y un documento $d \in D$. La evaluación de esta función establece un cierto orden entre la información de acuerdo a la consulta.

Evolución

DÉCADA 1960

Nacimiento de los primeros sistemas que utilizaban índices invertidos y técnicas de recuperación booleana

DÉCADA 1970

Aparecen sistemas comerciales como MEDLINE y SMART

DÉCADA 1980

Mayor atención a la representación de la información y al procesamiento del lenguaje natural

DÉCADA 1990

Desarrollo de motores de búsqueda web como Yahoo! y AltaVista, marcando la transición hacia la RI en entornos web

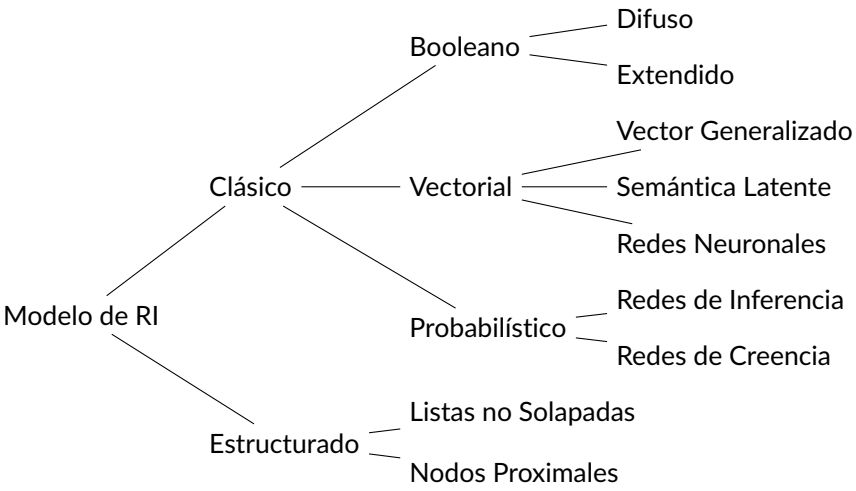
DÉCADA 2000

Énfasis en algoritmos de ranking, personalización y búsquedas más eficientes

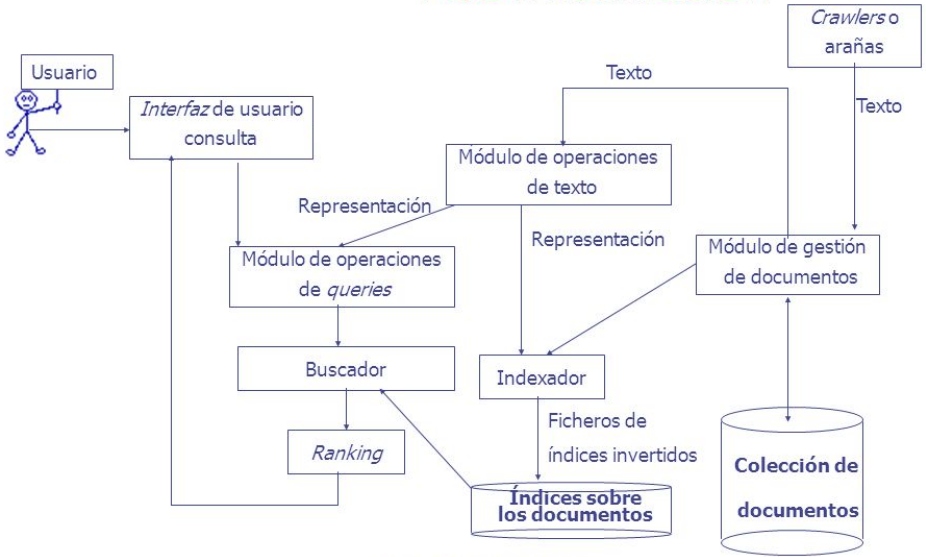
HASTA EL PRESENTE

Utilización de algoritmos de aprendizaje profundo para mejorar la precisión y relevancia en la RI

Tipos básicos de SRI



Estructura de un SRI



Tomado de `https://slideplayer.es/slide/1434895/`

Aplicaciones

Ejemplos prácticos actuales

- **Motores de búsqueda empresariales**
- **Búsqueda en repositorios de código**
- **Asistentes de IA**

Motores de búsqueda empresariales

- Búsqueda en documentos internos de la organización
- Índices corporativos con control de acceso y seguridad
- Integración con sistemas empresariales existentes
- Ejemplos: Elasticsearch, Solr, Microsoft SharePoint Search, Google Workspace Search
- Desafíos: escalabilidad, relevancia contextual, privacidad de datos

Búsqueda en repositorios de código

- **GitHub Code Search**: búsqueda semántica y por sintaxis en código fuente
- Filtros por lenguaje, repositorio, organización, etc.
- Aplicaciones:
  - Encontrar ejemplos de uso de APIs
  - Localizar bugs similares
  - Descubrir patrones de código
  - Aprender de código existente
- Desafíos técnicos: índices de código fuente, búsqueda por estructura (AST), contexto semántico del código

Asistentes de IA como SRI

- Los asistentes de IA (ChatGPT, Claude, Copilot) utilizan técnicas avanzadas de RI
- Combinan recuperación de información con generación de texto
- Acceso a bases de conocimiento actualizadas

Retrieval-Augmented Generation (RAG)

1. Consulta del usuario
2. Búsqueda en base de conocimiento
3. Recuperación de documentos relevantes
4. Generación de respuesta basada en contexto recuperado

¿Cómo han cambiado los LLMs el paradigma de la recuperación?

- **Comprensión semántica mejorada**: Los LLMs entienden mejor el contexto y la intención
- **Generación de respuestas**: No solo recuperan, sino que sintetizan información
- **Búsqueda conversacional**: Interacción más natural con el sistema
- **Vectorizaciones avanzadas**: Representaciones vectoriales más ricas

¿Cómo han cambiado los LLMs el paradigma de la recuperación?

- **Comprensión semántica mejorada**: Los LLMs entienden mejor el contexto y la intención
- **Generación de respuestas**: No solo recuperan, sino que sintetizan información
- **Búsqueda conversacional**: Interacción más natural con el sistema
- **Vectorizaciones avanzadas**: Representaciones vectoriales más ricas

Pero los LLMs **no reemplazan** los sistemas de RI tradicionales, sino que los **complementan** y **mejoran**.

¿Por qué la RI sigue siendo fundamental?

Limitaciones de los LLMs:

- Alucinaciones: generan información incorrecta
- Conocimiento estático: entrenados con datos históricos
- Costo computacional: muy costosos para búsquedas masivas
- Transparencia: difícil rastrear fuentes

¿Por qué la RI sigue siendo fundamental?

Limitaciones de los LLMs:

- Alucinaciones: generan información incorrecta
- Conocimiento estático: entrenados con datos históricos
- Costo computacional: muy costosos para búsquedas masivas
- Transparencia: difícil rastrear fuentes

Ventajas de la RI tradicional:

- Precisión verificable: resultados rastreables a documentos
- Eficiencia: búsquedas rápidas en grandes colecciones
- Actualización continua: índices actualizables en tiempo real
- Control: algoritmos interpretables y ajustables

¿Por qué la RI sigue siendo fundamental?

Limitaciones de los LLMs:

- Alucinaciones: generan información incorrecta
- Conocimiento estático: entrenados con datos históricos
- Costo computacional: muy costosos para búsquedas masivas
- Transparencia: difícil rastrear fuentes

Ventajas de la RI tradicional:

- Precisión verificable: resultados rastreables a documentos
- Eficiencia: búsquedas rápidas en grandes colecciones
- Actualización continua: índices actualizables en tiempo real
- Control: algoritmos interpretables y ajustables

La combinación de **RI tradicional** (para recuperación precisa) + **LLMs** (para comprensión y generación) ofrece los mejores resultados.

El futuro de la RI

- **RAG (Retrieval-Augmented Generation)**: Combina recuperación con generación
- **Embeddings multimodales**: Texto, imágenes, audio, video
- **Búsqueda híbrida**: Combinación de búsqueda léxica y semántica
- **RI personalizada**: Adaptación a preferencias del usuario
- **RI en tiempo real**: Actualización continua de índices

La Recuperación de Información sigue siendo **fundamental** en la era de la IA. Los LLMs han **transformado** cómo interactuamos con la información, pero los principios y técnicas de RI siguen siendo la **base** de estos sistemas modernos.

Sistema de evaluación

- Trabajos de Control (2)
- Proyecto
- Preguntas en clase
- Ejercicios propuestos
- Participación en clase

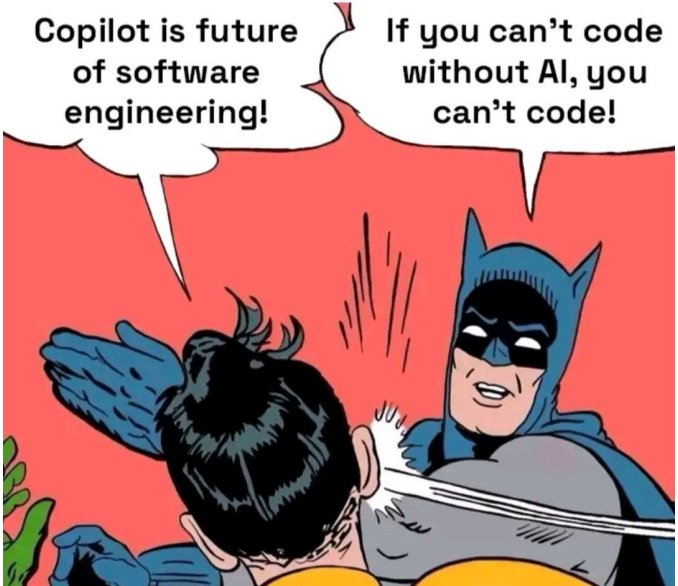
- Posibles cambios a notificar ...

Dudas, preguntas, sugerencias ...

El profe: alguna duda

Yo:

¡Cuidado! No confundas ...



Bibliografía

- Baeza-Yates, R., Ribeiro-Neto, B (2002) Modern Information Retrieval.
- Manning, C. D. (2009). An Introduction to Information Retrieval . Cambridge UP
- Baeza-Yates, R. a. (s.f.). Information Retrieval: Data Structures & Algorithms.

Sistemas de Recuperación de Información

Introducción a los SRI

Lic. Carlos León González

Facultad de Matemática y Computación
Universidad de La Habana

6 de febrero de 2026