Técnicas avanzadas de RI

Juan Manuel Fernández Luna

DECSAI
Departamento de Ciencias de la Computación e I.A.
Universidad de Granada

jmfluna@decsai.ugr.es

Sistemas de Recuperación de Información y Recomendación
Máster en Ciencia de Datos e Ingeniería de Computadores

Presentación basada en los capítulos 21, 22 y 23 del libro
Recuperación de información. Un enfoque práctico y multidisciplinar. Ra-Ma. 2011.

Sumario

PLN para RI

Sistemas de búsqueda de respuestas

Recuperación de información patrocinada

Detección y seguimiento de temas de actualidad

Detección de novedad

Construcción automática de resúmenes

Búsqueda social

Investigación en Recuperación de Información

PLN para RI

- Estudio y desarrollo de técnicas computacionales capaces de procesar textos excritos en lenguajes humanos.
- Mayor solapamiento entre RI y PLN es la *expansión de consultas*.
- Ayuda a salvar la brecha existente entre el vocabulario usado en una consulta y el del documento.

PLN para RI

- Expansión de consultas basada en inventarios de sinónimos.
- Palabras polisémicas o monsémicas dependiendo del dominio (abierto o cerrado).
- Por ejemplo, banco en una colección financiera o en la Web.
- Polisemia → sinónimos válidos sólo para algunos significados.
- Solución: emplear contexto.

PLN para RI

- Problema: consulta muy pequeña.
- Solución: crear contexto en tiempo de indexación.
- **Desambiguación de sentidos**: descubrir con qué sentido se está utilizando una palabra en un contexto dado.

Sumario

- PLN para RI
- Sistemas de búsqueda de respuestas
- Recuperación de información patrocinada
- Detección y seguimiento de temas de actualidad
- Detección de novedad
- Construcción automática de resúmenes
- Búsqueda social
- Investigación en Recuperación de Información

Sistema de búsqueda de respuestas

- Tratan de proporcionar la respuesta exacta a una pregunta, no el documento que la contenga.

  ¿Cuándo nació Charles Darwin?

  12 de febrero de 1890

  ¿Cuál es el PNB de Japón?

  Tabla con los datos históricos

- Respuesta a pregunta explícita ó lista de términos.

Sistema de búsqueda de respuestas

- Dos tipos de preguntas:
  - **Factuales**: obtener un dato muy concreto (nombre d de una persona, fecha, lugar o definición).
  - **No factuales**: obtener respuesta a una pregunta compleja (causa o consecuencia de un evento, ventajas ó desventajas de emplear un producto con respecto a otro,...).

Sistema de búsqueda de respuestas

- Dos tipos de preguntas:
  - **Factuales**:
    - Métodos superficiales:
    - Identificación del tipo de pregunta (clasificación: qué, cuándo, quién, dónde,...).
    - Extracción de contextos (lanzar la respuesta a un SRI para obtener párrafos).
    - Selección del dato concreto a partir de los diferentes contextos.

Sistema de búsqueda de respuestas

- Dos tipos de preguntas:
  - **Factuales**:
    - Problemas en consultas como:
      - ¿Qué actor dobló al personaje Kogawa Toshimi?
      - ¿Quién es el presidente de Francia? (respuesta cambia con el tiempo).
      - ¿Cuál es la última obra de Arturo Pérez Reverte? (respuesta cambia con el tiempo).
      - ¿Cuántas personas han fallecido en terremotos en América entre 2003 y 2010? (agragación de resultados).
      - Métodos superficiales no suficientes.

Sumario

- PLN para RI
- Sistemas de búsqueda de respuestas
- Recuperación de información patrocinada
- Detección y seguimiento de temas de actualidad
- Detección de novedad
- Construcción automática de resúmenes
- Búsqueda social
- Investigación en Recuperación de Información

RI Patrocinada

- Proporcionar a los usuarios anuncios altamente relevantes a sus necesidades de información en un momento concreto.
- La mayor parte de motores de búsqueda de la web proporcionan este servicio.
- Forma de financiación.
- Resultados orgánicos vs patrocinados.
- ¿Diferenciados o integrados?

RI Patrocinada

- **Documento = información suministrada por el anunciador**:
  - Título, breve descripción, página web destino y posibles consultas para las cuales el anuncio es relevante (palabras clave), zona geográfica, idioma, ...
  - ¿Dónde se puede hacer mejor RI Patrocinada?
  - es decir, ¿dónde existe muchas más información del usuario útil para la RI Patrocinada?

RI Patrocinada

- **Caso 1: Igual consulta del usuario que la(s) palabra(s) clave.**
  - Recuperación directa.
  - Palabras clave del anunciante: reproductor mp3, reproductores mp3, marca1, marca2,...
  - Consulta de usuario: reproductor mp3.
  - Se recupera el anuncio directamente.

RI Patrocinada

- **Caso 2: Anuncios contextuales.**
  - **Acciones**: leyendo una crítica sobre una película, una noticia sobre un móvil, ó viendo el temario de un curso de internet sobre Android...
  - El SRI Patrocinada calcula la relevancia entre la página abierta por el usuario y determina qué anuncio incluir en ella.

RI Patrocinada

- **Diferencias RI tradicional vs patrocinada.**
  - En RI tradicional → un único actor, el usuario
  - Objetivo: maximizar la relevancia de la información.
  - En RI patrocinada, cuatro:
    - **Usuario**: maximizar la relevancia de los anuncios.
    - **Anunciantes**: atraer usuarios interesados en sus productos o servicios → acciones o conversiones.
      - Ventas, darse de alta en una asociación, suscriptores,...
    - **Dueños o editores** de las páginas web donde aparecen los anucios: reciben parte de los beneficios.

RI Patrocinada

- Diferencias RI tradicional vs patrocinada.
  - En RI patrocinada, cuatro:
    - **Motores de búsqueda**: cuanto mejor lo hagan, más clientes tendrán.
    - **Objetivo global**: maximizar la relevancia de los anuncios, basándose en:
      - Calidad de la información de anuncio y métodos de RI.

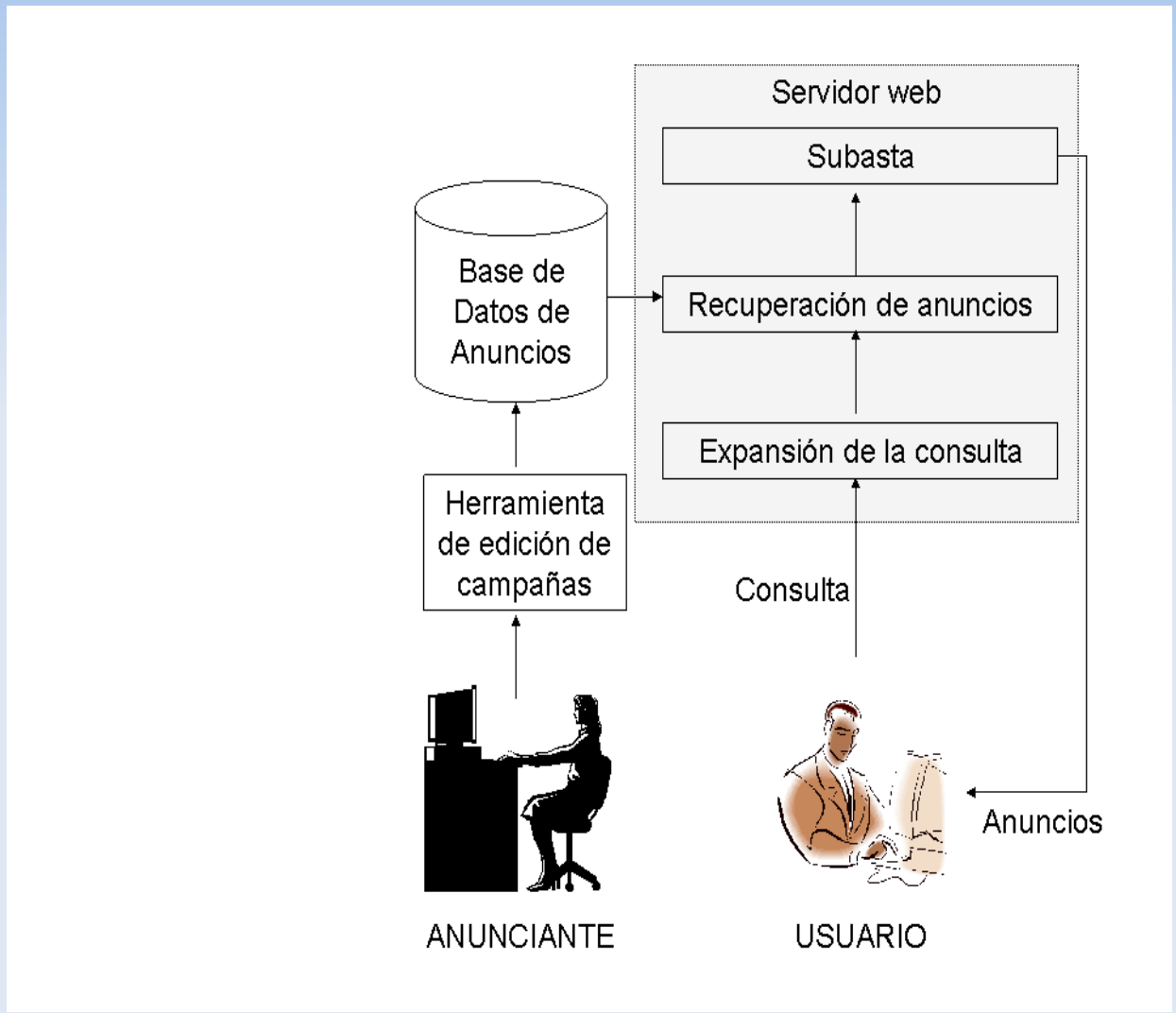
RI Patrocinada

- Diferencias RI tradicional vs patrocinada.
  - Cada anuncio lleva asociado un **precio** → ordenación de éstos al presentarlos al usuario.
    - **Coste por impresión**: valor para el anunciante haber expuesto su marca o mensaje a los clientes. Barato. Coste por mil impresiones.
    - **Coste por clic**: impresión gratuita y los anunciantes pagan por clics recibidos. Clic más caro que impresión.
    - **Coste por acción**: pagan por una acción concreta (compra, registro de usuario, rellenar un formulario,...).

RI Patrocinada

- ¿Varios anunciantes interesados en las mismas palabras clave o productos?
  - **Subasta**.
  - Ejemplo, AdWords Google.
  - Pujas separadas por cada palabra clave.
  - Se ordenarían por valor decreciente de la puja.
  - El ganador pagaría el valor del segundo.
  - Modificación: incorporar la relevancia del anuncio para calcular el ranking y el valor a pagar.

RI Patrocinada



Arquitectura de un SRI Patrocinada

Sumario

PLN para RI

Sistemas de búsqueda de respuestas

Recuperación de información patrocinada

Detección y seguimiento de temas de actualidad

Detección de novedad

Construcción automática de resúmenes

Búsqueda social

Investigación en Recuperación de Información

Detección y seguimiento temas actualidad

- **Contexto**: noticias → documento de texto donde se informa sobre un determinado suceso, con narrativa que responde el qué, dónde, cómo y porqué del mismo.
- **Topic Detection and Tracking (TDT)**.
- Línea de investigación que permite al usuario realizar un seguimiento de los temas de actualidad a partir de varias fuentes.

Detección y seguimiento temas actualidad

- **Objetivo**:
  - Organización de sucesos y detectar aquellos nuevos.
  - Identificar noticias que conforman un suceso al que hay que darle seguimiento.

Detección y seguimiento temas actualidad

- **Noticia o historia:**
  - Un artículo periodístico o un segmento de transmisión de un medio de comunicación con un enfoque coherente. Narración de un hecho.
- **Suceso**:
  - Algo que ocurre en un instante de tiempo y lugar específicos, junto con todas sus precondiciones necesarias y consecuencias inevitables (elecciones nacionales, accidentes aéreos concretos, crímenes y desastres naturales específicos).

Detección y seguimiento temas actualidad

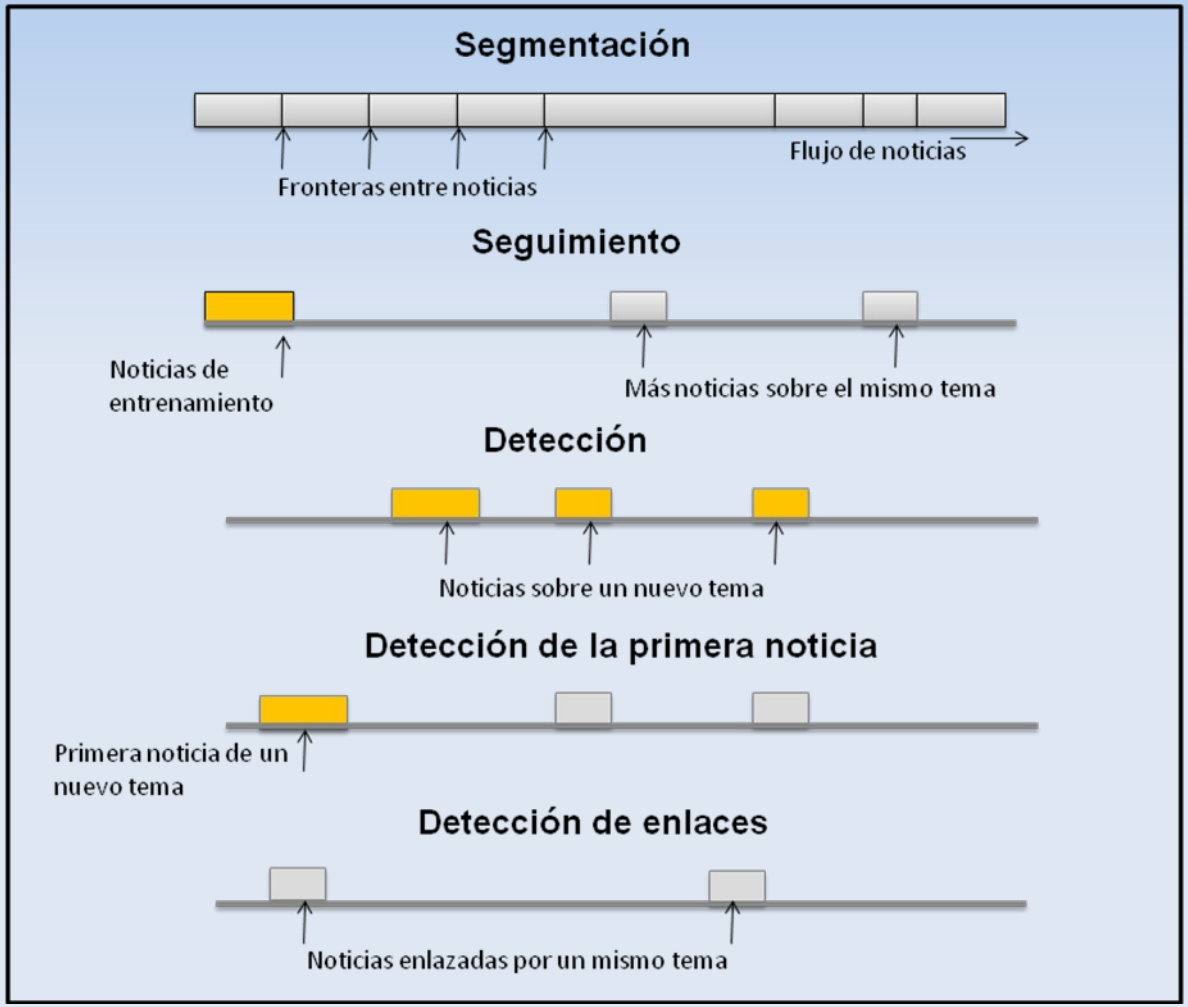
- **Tema**:
  - actividad o suceso de especial relevancia, junto con todos los sucesos, hechos o actividades directamente relacionados con él.

    El conjunto de acciones conectadas que tienen un enfoque común se define como actividad (campañas electorales, investigaciones, labores de socorro en desastres).

    Por ejemplo, las noticias sobre el accidente del avión en que viajaba el presidente polaco Lech Kaczynski el 10 de abril de 2010, la búsqueda de los sobrevivientes y el funeral de las víctimas abordan el mismo tema.

Detección y seguimiento temas actualidad

Tareas asociadas a TDT



Wayne, 2000

Detección y seguimiento temas actualidad

- **Detección**:
  - Reconocer en un flujo de noticias aquellas que pertenecen a un nuevo tema o a otro conocido previamente.
  - **Agrupar** noticias que abordan el mismo tema.
  - Aprendizaje no supervisado.
  - Detección en línea: toman la decisión conforme llegan las noticias.
  - Detección retrospectiva: se tiene toda la colección completa y el objetivo es estructurarla.

Detección y seguimiento temas actualidad

- **Detección**:
  - Detección de la primera noticia.
  - Detección del resto de noticias en grupos apropiados.
- **Seguimiento**:
  - Partiendo de un conjunto de noticias que el usuario ha identificado previamente, se vigila el flujo de noticias para encontrar otras sobre el mismo tema.
  - **Clasificar** noticias.
  - Aprendizaje supervisado.

Detección y seguimiento temas actualidad

- **Elementos** de un sistema TDT:
  - Modelo de representación de las noticias.
  - Medida de similitud entre noticias.
  - Gestor de propiedades temporales de las noticias

    (para diferenciar entre las elecciones de 2004 y 2008).
  - Algoritmo de agrupamiento que permita organizar las noticias en temas.

Detección y seguimiento temas actualidad

- Considerando las propiedades temporales:
  - Orden cronológico (problema con noticias que se refieran a sucesos pasados o futuros).
  - Ventana temporal y comparando cada noticia sólo con las que pertenecen a dicha ventana.
  - Incluyendo en las funciones de similitud parámetros que tengan en cuenta estas propiedades temporales.

Sumario

PLN para RI

Sistemas de búsqueda de respuestas

Recuperación de información patrocinada

Detección y seguimiento de temas de actualidad

Detección de novedad

Construcción automática de resúmenes

Búsqueda social

Investigación en Recuperación de Información

Detección de novedad

- Hasta ahora un sistema es eficaz si su módulo de recuperación es eficaz.
- La relevancia depende sólo de la consulta.
- ¿Habría otros factores que influirían?
- El **contexto del usuario**: ubicación, la estación del año, la meteorología, la hora, el estado de ánimo del usuario,...
- La **novedad** de la información también afecta a la relevancia.

Detección de novedad

- Hasta ahora se ha asumido un modelo basado en la suposición de la **independencia** de la relevancia: **la relevancia de un documento es independiente de la relevancia del resto**.
- Problema: las ordenaciones podrían tener documentos en las posiciones altas casi iguales, con mucha **redundancia**.
- Esto resulta molesto a los usuarios.

Detección de novedad

- **Objetivo**:
  - No seleccionar individualmente documentos sino **escoger globalmente la mejor combinación de elementos que satisfaga la necesidad de información del usuario**.
  - Dado un conjunto de documentos (por ejemplo, el conjunto ordenado de documentos que un sistema de recuperación de documentos ha estimado como relevantes), la detección de novedad consiste en **filtrar los documentos de dicho conjunto que contienen información redundante, conservando tan sólo aquellos que proporcionan nueva información (información novel)**

Detección de novedad

¿Novedad?

- *"Novedad o nueva información significa nuevas respuestas a las preguntas potenciales que representan una petición del usuario o necesidad de información" (Li y Croft, 2008).*
- Dos vertientes:
  - Necesidad de información → consultas.
  - Información novel → detectando documentos que contienen preguntas que no han sido respondidas en respuestas previas (documentos vistos por el usuario).

Detección de novedad

¿Novedad?

- **Tipos**:
  - **Novedad directa**: los usuarios pueden querer seguir buscando documentos relacionados con un tema que se ha detectado previamente como novel.
  - **Novedad indirecta**: los usuarios podrían estar interesados en buscar documentos que no contienen información ya vista (evita información redundante).

Detección de novedad

- **Salida** de un TDT:
  - Listado ordenado de documentos que son tanto novedosos como relevantes.
  - El sistema TDT parte de una ordenación de relevancia y realiza un proceso de reordenación y/o filtrado promoviendo las piezas de información novedosas.

Detección de novedad

- **Diversidad**:
  - Cada una de las posibles interpretaciones de una consulta.
  - Irak → conflictos, información geográfica, histórica, etc.
  - Subtema o faceta: cada una de las interpretaciones o subtemáticas relacionadas con una consulta.
  - Es interesante introducir diversidad en una salida del SRI.

Detección de novedad

- **Diversidad**:
  - Cada una de las posibles interpretaciones de una consulta.
  - Irak → conflictos, información geográfica, histórica, etc.
  - Subtema o faceta: cada una de las interpretaciones o subtemáticas relacionadas con una consulta.
  - Es interesante introducir diversidad en una salida del SRI.

Detección de novedad

- Tipos:
  - **Diversidad extrínseca**: Hay incertidumbre sobre la necesidad de información (el SRI no sabe qué faceta es la correcta)
  - **Diversidad intrínseca**: aunque la consulta sea clara y definida el usuario no quiere revisar información redundante, por tanto, el sistema debe cubrir el mayor número de aspectos.
  - Útil en:
    - Consultas navegacionales: mundo deportivo.
    - Interesado en diferentes visiones u opiniones.
    - Necesidad de información = visión general sobre un tema.
    - Resultados de diferentes fuentes para asegurar corrección y autoridad.

Detección de novedad

- **Diversidad**:
  - Ambigüedad: ante la duda, ofrecerle en los primeros puestos una representación de las temáticas existentes.
  - El usuario podrá reformular la consulta para centrarse en la que le interesa.

Detección de novedad

¿Novedad = diversidad?

Detección de novedad

¿Novedad = diversidad?

Normalmente no intercambiables, pero...

… al buscar novedad se promueve la diversidad.

- **Detección de novedad**: encontrar información novel. Dado un documento y un conjunto de documentos vistos previamente, un documento deberá tratar de un tema o interpretación que es diferente a cualquier otra ya tratado previamente o, al menos, hacer referencia a una faceta o subtema distinto de los cubiertos anteriormente.

Detección de novedad

- **Aproximaciones** a nivel de documento:
  - Integrar detección de novedad con topicalidad:

    Por ejemplo, el Modelo de Relevancia Marginal:

$$MMR = \underset{d_i \in C-S}{\operatorname{argmax}} \left[\left(\lambda \cdot sim\left(d_i, q\right) - \left(1-\lambda\right) \cdot \max_{d_j \in S} sim\left(d_i, d_j\right) \right) \right]$$

Detección de novedad

- **Aplicaciones** de la detección de novedad:
  - Respuesta automática de preguntas.
  - Filtrado adaptativo de documentos.
  - Extracción de subtópicos.
  - Generación automática de resúmenes.

Sumario

PLN para RI

Sistemas de búsqueda de respuestas

Recuperación de información patrocinada

Detección y seguimiento de temas de actualidad

Detección de novedad

Construcción automática de resúmenes

Búsqueda social

Investigación en Recuperación de Información

Construcción de resúmenes

- La **construcción automática de resúmenes de documentos** (summarization) consiste en, dados una fuente de información (uno o más documentos textuales) y un demandante (usuario o aplicación), extraer el contenido de la fuente de información y presentarlo al demandante de forma condensada, comprensible, y que satisfaga sus necesidades.
- **Resumen**= corto, preservar información y estar desprovisto de redundancia.

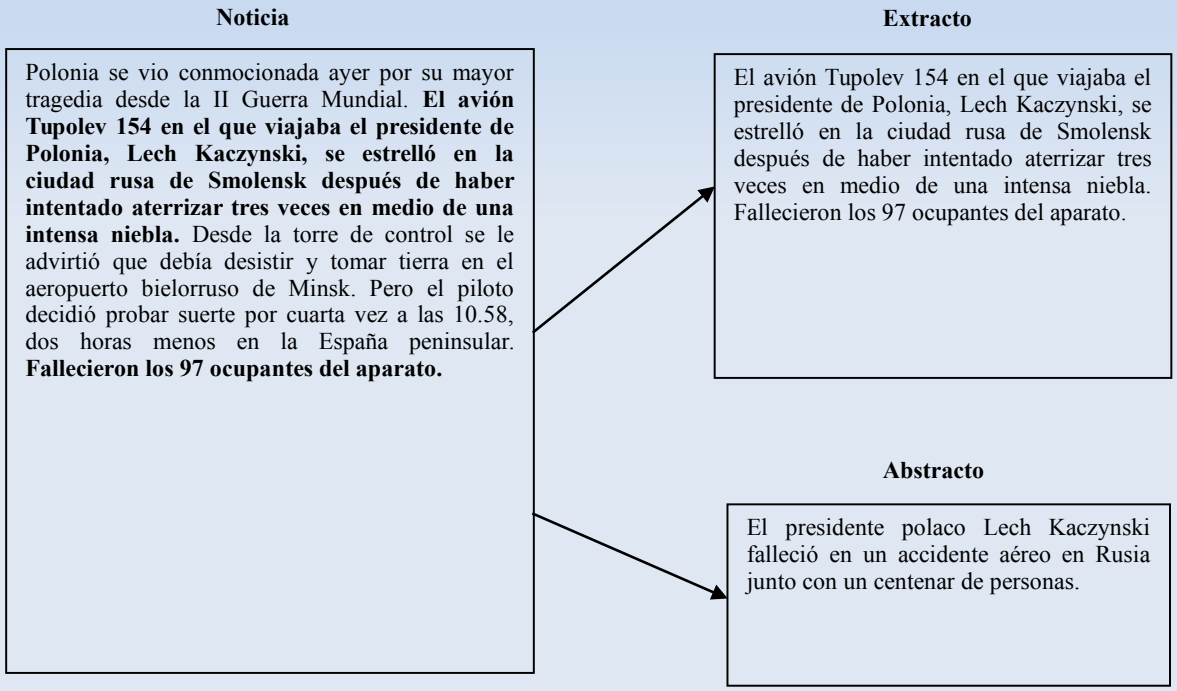
Construcción de resúmenes

Clasificación de los resúmenes:

Criterio de clasificación		
Técnica	Extracto	Abstracto
Contenido	Genérico	Enfocado a un usuario
Función	Indicativo	Informativo
Fuente	Un solo documento	Múltiples documentos

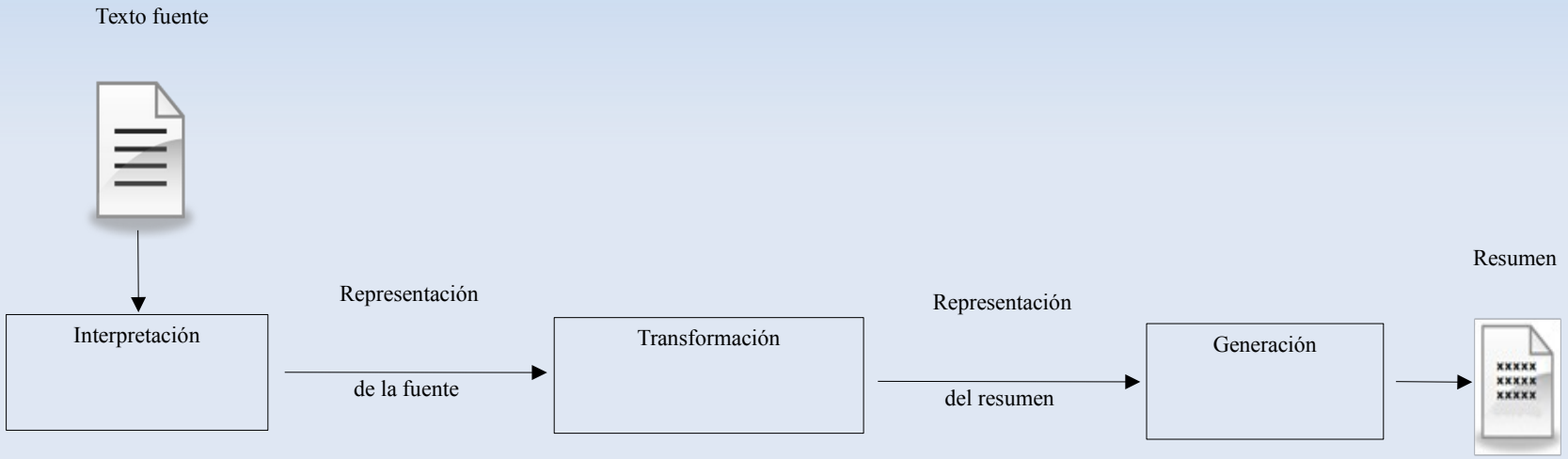
Construcción de resúmenes

Clasificación de los resúmenes:



Construcción de resúmenes

Proceso de construcción:



Construcción de resúmenes

Construcción de extractos:

- Se basan en la **selección de oraciones**.

1. Preprocesamiento del texto fuente:
    1. Segmentación de oraciones,
    2. lematización,
    3. eliminación de palabras vacías,
    4. análisis sintáctico,...
2. Extracción de los rasgos significativos de las oraciones.
3. Creación del listado de oraciones (se ordenan según dichos rasgos).
4. Seleccionar las oraciones del listado hasta que se alcance un umbral de longitud.
5. Conformación del resumen (reordenar oraciones según el orden de aparición, resolución de anáforas, reparar partes estructuradas (tablas o listas),...).

Construcción de resúmenes

Rasgos para representar oraciones:

- **Presencia de palabras clave**: considera la ocurrencia en la oración de las palabras más frecuentes en el documento o las palabras del título, pues éstas suelen indicar el tema principal.
- **Posición de la oración**: las oraciones que expresan el tema del documento suelen ocurrir en posiciones específicas dependiendo del tipo de documento (por ejemplo, el título para noticias, el resumen para artículos científicos o un epígrafe para libros).
- **Presencia de frases pista**: considera su presencia en frases como "importante", "por lo tanto", "en resumen", "el objetivo es", etc.
- **Longitud de la oración**: las oraciones cortas no suelen considerarse.
- **Presencia de palabras con mayúsculas**: considera la ocurrencia de siglas y nombres de entidades (personas, lugares, empresas, etc.)
- **Cohesión**: considera las co-ocurrencias de palabras frecuentes y cadenas léxicas, presencia de las palabras en posiciones sintácticas importantes (por ejemplo, el sujeto), penaliza la presencia de anáforas (repetición de una misma palabra al comienzo de una oración), etc.

Construcción de resúmenes

Rasgos para representar oraciones:

Dada una oración *s*, su puntuación se calcula como

$$v(s)=\alpha\cdot C(s)+\beta\cdot K(s)+\gamma\cdot T(s)+\delta\cdot L(s),$$

donde $C(s)$, $K(s)$, $T(s)$ y $L(s)$ denotan, respectivamente, los valores de los rasgos ***palabras-pistas***, ***palabras-claves***, ***palabras-título*** y ***posición*** de *s*, y α, β, γ y δ son sus pesos asociados.

- Rasgo palabras-pistas: pondera las oraciones de acuerdo con la frecuencia de aparición de un conjunto de palabras pistas obtenido a partir de resúmenes de una colección de entrenamiento.
- Rasgos *palabras-título* y *palabras-claves* favorecen a las oraciones que presentan palabras del título y palabras muy frecuentes en el documento, respectivamente.
- El rasgo *posición* favorece a las oraciones que se encuentran cerca del comienzo y del final del texto.

Construcción de resúmenes

Si se desea construir un extracto de un conjunto de documentos...

- Igual esquema de trabajo.
- Se incluye el tema del conjunto de documentos.
- Las oraciones relacionadas con este tema obtienen mayor puntuación.
- Eliminación de información redundante.
- Detección de novedad.
- Normalización temporal (”hoy”).

Construcción de resúmenes

Construcción de resúmenes:

- Más cercano al PLN.
- Extracción de elementos importantes del documento (oraciones).
- Aplicación de determinadas operaciones de generación del lenguaje:
  - Reducción de oraciones: La tecnología de CD-ROM, que emplea las mismas técnicas láser utilizadas para crear los discos compactos de sonido, permite capacidades de almacenamiento del orden de varios cientos de megabytes (millones de bytes) de datos" → "La tecnología de CD-ROM permite capacidades de almacenamiento del orden de varios cientos de megabytes de datos".

Construcción de resúmenes

Construcción de resúmenes:

- Aplicación de determinadas operaciones de generación del lenguaje:
  - Paráfrasis: reemplaza una frase por su paráfrasis. Por ejemplo, reemplazar la frase “Las medidas anti-inflación son defendidas por el gobierno” por “El gobierno apoya la política anti-inflación”.
  - Transformación sintáctica del verbo: por ejemplo, cambiar el verbo a su forma impersonal.
  - Transformación sintáctica de la oración: por ejemplo, la posición del sujeto puede moverse del final al comienzo de una oración.
  - Combinación de oraciones: unir lo expresado en dos oraciones. Por ejemplo, unir dos oraciones próximas que compartan el mismo sujeto, eliminando el sujeto de la segunda y añadiendo la conjunción “y”.

Construcción de resúmenes

Construcción de resúmenes:

- Aplicación de determinadas operaciones de generación del lenguaje:
  - Generalización o especificación: reemplazar cláusulas o frases por sus descripciones más generales o más específicas.

    Por ejemplo, “una nueva ley que exige a los editores web obtener consentimiento de los padres antes de recoger información personal de los niños” → “legislación para proteger la privacidad de los niños en línea”.
  - Reordenación: cambiar el orden de las oraciones extraídas. Por ejemplo, colocar la oración final de un artículo científico como la primera del abstracto.

Sumario

PLN para RI

Sistemas de búsqueda de respuestas

Recuperación de información patrocinada

Detección y seguimiento de temas de actualidad

Detección de novedad

Construcción automática de resúmenes

Búsqueda social

Investigación en Recuperación de Información

Búsqueda social

- Medios sociales, o contenido generado por el usuario – User Generated Content, UGC.
- Alto desarrollo de la Web 2.0 → el usuario medio puede publicar contenidos en línea de forma sencilla.
- Por ejemplo, blogs o plataformas como Twitter o Facebook.
- Búsqueda social: participación activa de comunidades de usuarios en procesos de búsqueda.

Búsqueda social

- Búsqueda apoyada en **folksonomías** (Del.iocio.us, Flickr,...).
- Búsqueda en **medios sociales**:
  - Búsqueda opinada.
  - Búsqueda por blog clave.
  - Búsqueda en Twitter.

Búsqueda social

Búsqueda opinada:

- Blogueros escriben entradas en su blog.
- Otros usuarios comentan y se debate.
- Objetivo: buscar **tendencias de opinión**.
- Por ejemplo, rumores positivos o negativos sobre diferentes temas.
- Identificar entradas de blog que expresen opinión sobre un objetivo determinado (persona, lugar, organización, producto).
- ¿Qué piensa la gente de X?

Búsqueda social

Búsqueda opinada:

- Depende de tres factores:
  - Relevancia de las entradas de blog con respecto a la consulta.
  - Calidad del opinante.
  - Opinión (análisis de sentimientos).
- Larga tradición en la comunidad PLN.
- Aproximación básica: obtener una lista de entradas relevantes y reordenarlas según la opinión.

Búsqueda social

Búsqueda opinada:

- Detección de opinión → **Análisis de sentimientos**:
  - Técnicas de clasificación.
  - Basadas en diccionarios y listas de palabras que expresan sentimientos conocidos.

Búsqueda social

Búsqueda por blog clave:

- Búsquedas de blogs para, normalmente, suscribirse por RSS (síntesis de blogs).
- Dada una consulta, obtener un conjunto de blogs relevantes dedicados fundamentalmente al tópico de la consulta.
- Selección de recursos vs búsqueda de expertos.

Búsqueda social

Búsqueda por blog clave:

- Selección de recursos:
  - Blogs = colecciones de sus entradas.
  - Problema de RI distribuido.
  - Objetivo: seleccionar las colecciones que más probablemente contengan documentos relevantes para la consulta.
- Búsqueda de expertos:
  - Objetivo: encontrar gente con experiencia relevante en un tema concreto.
  - Agregar relevancia de las entradas a su creador.
  - Recuperar entradas relevantes y considerarlas como votos para el bloguero.

Búsqueda social

Búsqueda en Twitter:

- Diferencia básica: entradas de máximo 140 caract.
- Normalemente noticias sobre el usuario, comentarios de enlaces, debates e información de ubicación.
- El usuario puede emplear los tuits para encontrar información.
- Búsquedas en Twitter para monitorizar contenido mientras que las de la Web para desarrollar y aprender sobre un tema.

Búsqueda social

Búsqueda en Twitter:

- Búsqueda de personas → importante.
- Búsqueda de opiniones → ordenar tuits
  - Importante la actualidad
  - Pero también puede considerarse los seguidores de un autor de un tuit,
  - número de retuits,
  - presencia de URLs,
  - presencia de hashtags.

Sumario

PLN para RI

Sistemas de búsqueda de respuestas

Recuperación de información patrocinada

Detección y seguimiento de temas de actualidad

Detección de novedad

Construcción automática de resúmenes

Búsqueda social

Investigación en Recuperación de Información

Investigación en RI

- RI una ciencia **empírica** desde sus inicios (Cranfield, década de los 60).
- **Investigación dirigida por los datos** → los conocimientos se adquieren a partir de grandes cantidades de datos).
- Avances significativos al evolucionar los datos.

Investigación en RI

Método científico aplicado a la RI

- **Revisar** el estado del arte.
- **Implementar** el estado del arte (uno o varios valores de referencia, o baselines).
  - Incremental: se añade una característica a un valor de referencia que lo mejora.
  - Comparativa: se demuestra que la técnica propuesta es mejor que el baseline.
- **Diseñar e implementar** nuestra nueva técnica.
  - Hay que mostrar los fundamentos y los conceptos en que se basa.

Investigación en RI

Método científico aplicado a la RI

- **Experimentar** para establecer parámetros (entrenamiento) y posteriormente para evaluar la técnica y compararla.
- **Reproducibilidad**: explicar todos los aspectos de la propuesta y configuración experimental. Uso de datos públicos → Colecciones de evaluación estándar.
- **Relevancia**: pruebas de significación estadística al comparar con los valores de referencia.

Investigación en RI

Colecciones de prueba de RI

Colección	Contenidos	Año	Documentos	Tamaño (GB)
Disk 1 & 2	Noticias	1992	740K	2,0
Disk 4 & 5	Noticias	1998	500K	1,9
WT2G	Páginas web	1999	240K	2,0
WT10G	Páginas web	2000	1.6M	10,0
GOV	Págs. web del Gob.US	2002	1.8M	18,0
W3C	Págs. web de W3C.org	2005	330K	5,7
CERC	Págs. web de CSIRO.au	2006	330K	4,1
Blogs06	Entradas de blogs	2006	3M	13,0
GOV2	Págs. web del Gob. US	2004	25M	425,0
Blogs08	Entradas de blogs	2009	28M	1.445,0
ClueWeb09	Páginas web	2009	1,2B	25.600,0

Investigación en RI

Plataformas de RI

- **Terrier (Universidad de Glasgow)**
  - Soporta todas las colecciones TREC.
  - Implementa muchos modelos de RI.
  - Indexa colecciones de gran tamaño.
- **Lemur/Indri (Universidad de Massachussetts)**
  - Soporta colecciones TREC.
  - Indexa colecciones de gran tamaño.
  - Utiliza un lenguaje de consultas estructurado.
  - Implementa modelos del lenguaje.

Investigación en RI

Plataformas de RI

- **Lucene (Apache Consortium)**
  - No soporta TREC.
  - Implementa los modelos clásicos.
  - Indexa colecciones de gran tamaño.
  - Apoyo de la comunidad científica y empresarial.

Investigación en RI

Foros de evaluación en RI

- Text REtrieval Conference (TREC).
- Cross-Language Evaluation Forum (CLEF).
- NII Test Collection for IR Systems Project (NTCIR).
- Forum for Indian Retrieval Evaluation (FIRE).
- Initiative for the Evaluation of XML Retireval (INEX).

Investigación en RI

Foros de evaluación en RI

- Sesiones de TREC:
  - Búsqueda ad hoc.
  - Búsqueda de respuestas.
  - Realimentación por relevancia.
  - Búsqueda de un elemento conocido (página principal ó correo electrónico).
  - Búsqueda de expertos.
  - Búsqueda de opiniones.
  - Síntesis de blogs.
  - Identificación de noticias importantes.
  - Búsqueda en dominios específicos (biomédico, patentes químicas,...).

Investigación en RI

¿Dónde publicar?

- Congresos:
  - Advances in Research & Development in Information Retrieval (SIGIR).
  - Conference in Information & Knowledge Management (CIKM).
  - European Conference in IR (ECIR).
  - Web Search and Data Mining (WSDM).
  - World Wide Web Conference (WWW).
  - Congreso Español de Recuperación de Información (CERI).
  - Dutch Information Retrieval (DIR) Workshop.
  - Asia Infromation Retrieval Societies Conference (AIRS).
  - String Processing and Information Retrieval Symposium (SPIRE).

Investigación en RI

¿Dónde publicar?

- Revistas:
  - Information Retrieval Journal.
  - Information Processing and Management.
  - Transactions on Information Systems
  - Journal of the American Society for Information Science and Technology.

Investigación en RI

Herramientas para trabjar con muchos datos

- Nuestro sistema de RI.
- Datos del orden de Terabytes → Más de un programa.
- Una solución: paradigma **MapReduce** de Google.
  - Localidad de datos: al distribuir en muchas máquinas se mejora la tasa de transferencia global.

    Sistema de archivos distribuido.
  - Operaciones de mapear, ordenar y reducir para realizar la tarea correspondiente.
- **Hadoop** → Implementación de MapReduce del Apache Consortium.

Investigación en RI

Herramientas para trabjar con muchos datos

- Otras herramientas complementarias a Hadoop:
  - Bases de Datos no relacionales: HBase, MongoDB,...
  - Bases de Datos relacionales: Hive.
  - Otras bases de datos: Pig.

FIN

Esto es todo amgios :-)