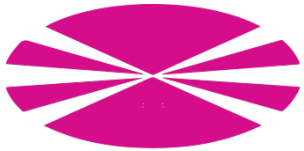


ENXEÑARÍA TELEMÁTICA
UNIVERSIDADE DA CORUÑA

Recuperación de información desde diferentes perspectivas

Grupo de Ingeniería Telemática

Facultad de Informática

Universidade de A Coruña

Diego Fernández, Víctor Carneiro, Francisco Novoa , Xacobe Macía & Fidel Cacheda

Contenidos

- Recuperación de Información en entornos distribuidos (motores de búsqueda):
  - Obtención de información (crawling process)
  - Indexación de la información (indexing process)
  - Recuperación de Información y Ranking (RR process)
- Enfoques:
  - Eficacia
  - Eficiencia
  - Escalabilidad
  - Dispersión de datos
  - …

victor.carneiro@udc.es

Obtención de información (I)

Crawling

- Arquitectura eficiente de obtención de información mediante sistemas de *crawling*
  - Escalabilidad
  - Diversidad
  - Dinamismo
  - Web oculta
  - Cambios tecnológicos
- Aportación:
  - Problema del web spam
  - Páginas soft-404
  - Arquitectura escalable y eficiente

ENXEÑARÍA TELEMÁTICA
UNIVERSIDADE DA CORUÑA
victor.carneiro@udc.es

Obtención de información (II)

Páginas soft-404

- Muchos web server envían códigos 200 HTTP como respuesta a documentos no encontrados.
- Aportación: construcción de un sistema denominado soft404Detector, basado en análisis de contenido para filtrado de estas páginas.
- Uso de heurísticas como ratio bytes contenido vs total, tamaño, imágenes, words, keywords, … con lo que consigue una precisión de 0.992.

Víctor M. Prieto, Manuel Álvarez, Fidel Cacheda. ”Soft-404 Pages, a Crawling Problem“. Journal of Digital Information Management (JDIM). Vol. 12, issue 2. pp. 73-92, April 2014

victor.carneiro@udc.es

Obtención de información (III)

Páginas Web Spam

- Páginas sin contenido válido con keywords y enlaces a otras páginas para aumentar el pageRank y por tanto los beneficios.
- Aportación:
  - Combinación de heurísticas utilizando árboles de decisión.
  - Selección de propiedades del conjunto global en función de relevancia, recursos, eficiencia, …

victor.carneiro@udc.es

Arquitectura escalable y eficiente (I)

Crawling :: Dinamismo web (i)

- Buscadores web usan *Crawlers* para descargar páginas e indexarlas, pero estas cambian constantemente y de manera impredecible.
- Aportación: Construcción de un sistema distribuido y colaborativo de detección de cambios en páginas web que reduce significativamente este tiempo y su tratamiento e indexación por parte del buscador.

Víctor M. Prieto, Manuel Álvarez, Víctor Carneiro, Fidel Cacheda. ”Distributed and Collaborative Web Change Detection System“. Computer Science and Information Systems Journal (ComSIS). ComSIS Consortium. Volume 12, issue 1. pp. 91-114, 2015.

victor.carneiro@udc.es

Arquitectura escalable y eficiente (II)

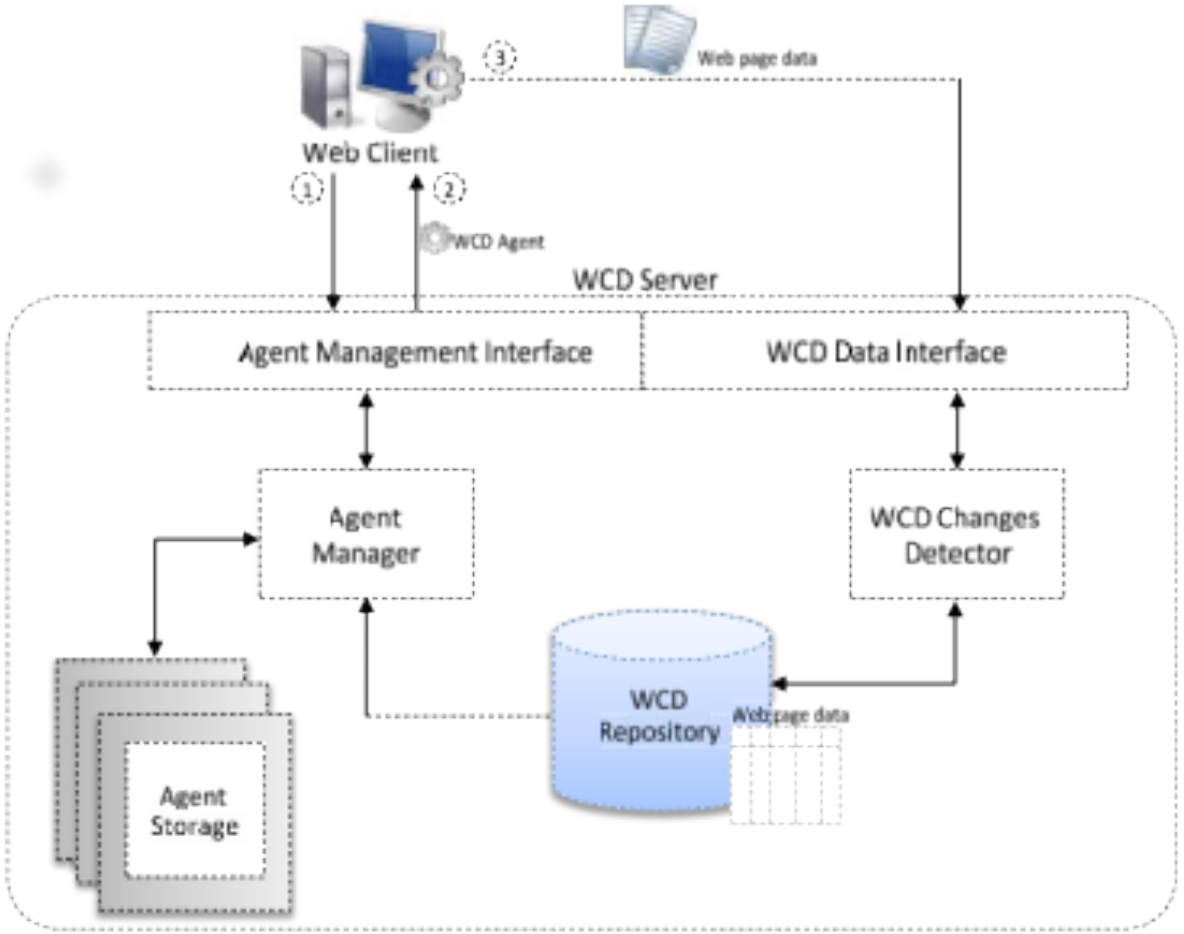
Crawling :: Dinamismo web (ii)

- Características del WCD:
  - Uso de un agente distribuido
  - Parte del procesamiento se realiza en la parte cliente
  - Actúa en modo push; no es necesario visitas a los web servers.
  - Los cambios en páginas web son notificados inmediatamente, lo que mejora la “experiencia de búsqueda”.
- Se consigue una media de 12 minutos para low PageRank y 1 minuto para high PageRank frente a las 24 horas de media en buscadores tradicionales.

victor.carneiro@udc.es

Arquitectura escalable y eficiente (III)

Crawling :: Dinamismo web (iii)



victor.carneiro@udc.es

Contenidos

- Recuperación de Información en entornos distribuidos (motores de búsqueda):
  - Obtención de información (crawling process)
  - Indexación de la información (indexing process)
  - Recuperación de Información y Ranking (RR process)
- Enfoques:
  - Eficacia
  - Eficiencia
  - Escalabilidad
  - Dispersión de datos
  - …

victor.carneiro@udc.es

Indexación de información (I)

Problemas a analizar

- Problemas de eficiencia y escalabilidad
- Optimización de técnicas de indexación: ficheros invertidos, ficheros de firmas, modelo vectorial, …
- Comportamiento de búsqueda del usuario e importancia del contexto, comunidades, …
- Dispersión y dinamismo de la información
- Medidas de eficiencia, cobertura, diversidad, …

Sistemas recomendadores (I)

- Técnicas de IR en sistemas de recomendación con filtrado colaborativo (FC) basados en memoria.
- Aportación:
  - Nuevas métricas para la medida de la precisión MAE vs GIM.
  - Algoritmos basados en tendencias o diferencias entre usuarios e items para mejorar la eficiencia.
  - Mejora de la eficiencia en algoritmos tipo kNN (vectorial model, preselección de vecinos).
  - Mitigación de la dispersión: similitud, profile expansion,…

Fidel Cacheda, Víctor Carneiro, Diego Fernández, Vreixo Formoso. “Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems”. ACM Transactions on the Web (TWEB), Volume 5, Issue 1, February 2011.

Sistemas recomendadores (II)

Algoritmos kNN

- Estudio de factores que afectan a los algoritmos de vecinos más cercanos (kNN) usados en FC
- Aportación:
  - Arquitectura distribuida que mejora el rendimiento y tiempo de respuesta de algoritmos kNN en la selección de vecinos, mediante técnicas de Big Data.
  - Técnicas para sistemas recomendadores distribuidos:
    - User partition
    - Item partition

Vreixo Formoso, Diego Fernández, Fidel Cacheda, Víctor Carneiro. “Distributed architecture for k-Nearest Neighbors recommender systems”. World Wide Web. Internet and Web Information Systems. Volume 18, Issue 4, pp 997-1017, 2014. Springer.

victor.carneiro@udc.es

Sistemas recomendadores (III)

Problema del Cold Start

- Técnicas de expansión de perfil permiten mejorar la precisión con nuevos items y obtener buenas recomendaciones a nuevos usuarios.
- Uso de técnicas de query expansion de IR
- Aportación:
  - Técnicas de item-global vs item-local
  - Técnicas de user-local

Vreixo Formoso, Diego Fernández, Fidel Cacheda, Víctor Carneiro. “Using profile expansion techniques to alleviate the new user problem”. Information Processing & Management, Volume 49, Issue 3, May 2013.

victor.carneiro@udc.es

Sistemas recomendadores (IV)

Compresión de matriz

- Además de eficientes los algoritmos tienen que ser eficaces.
- La matriz de ratings puede ser indexada para facilitar la rapidez de las recomendaciones
- Aportación:
  - Técnicas de reducción de la matriz de indexación
  - Basado en técnicas de IR sobre CF.
  - Reducción de hasta el 75% del tamaño de la matriz.

Vreixo Formoso, Diego Fernández, Fidel Cacheda, Víctor Carneiro. “Using rating matrix compression tecniques to speed up collaborative recommendations”. Information Retrieval, Volume 16, Issue 6, December 2013.

victor.carneiro@udc.es

Sistemas recomendadores (V)

Aplicación a la seguridad

- Uso de técnicas de filtrado colaborativo para predicción de tráfico:
  - Categorización de tráfico mediante PCAP DATASETs
  - Construcción de matriz users-items
  - Aplicación de algoritmos de CF
  - Evaluación de resultados…
- Aplicación:
  - Predicción de tráfico de red
  - Predicción de ataques
  - …

ENXEÑARÍA TELEMÁTICA
UNIVERSIDADE DA CORUÑA
victor.carneiro@udc.es

Contenidos

- Recuperación de Información en entornos distribuidos (motores de búsqueda):
  - Obtención de información (crawling process)
  - Indexación de la información (indexing process)
  - Recuperación de Información y Ranking (RR process)
- Enfoques:
  - Eficacia
  - Eficiencia
  - Escalabilidad
  - Dispersión de datos
  - …

victor.carneiro@udc.es

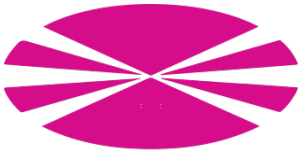
Recuperación de información

Green IR

- A partir del estudio del comportamiento de usuario en queries, optimización de la gestión de consultas y eficiencia energética.
- Aportación:
  - Modelo matemático de predicción del comportamiento.
  - Aumento/reducción del número de nodos en función de la contención y dificultad de las queries.
  - Mejora de la eficiencia energética mediante reducción automática de recursos sin comprometer la eficiencia.

Ana Freire, Craig Macdonald, Nicola Tonellotto, Iadh Ounis and Fidel Cacheda. “A self-adapting latency-Power Trade-off Model for Replicated Search Engines“. 7th ACM Web Search and Data Mining, New York, February 2014.

victor.carneiro@udc.es



ENXEÑARÍA TELEMÁTICA
UNIVERSIDADE DA CORUÑA

Gracias por su atención

Grupo de Ingeniería Telemática

Facultad de Informática de A Coruña

victor.carneiro@udc.es