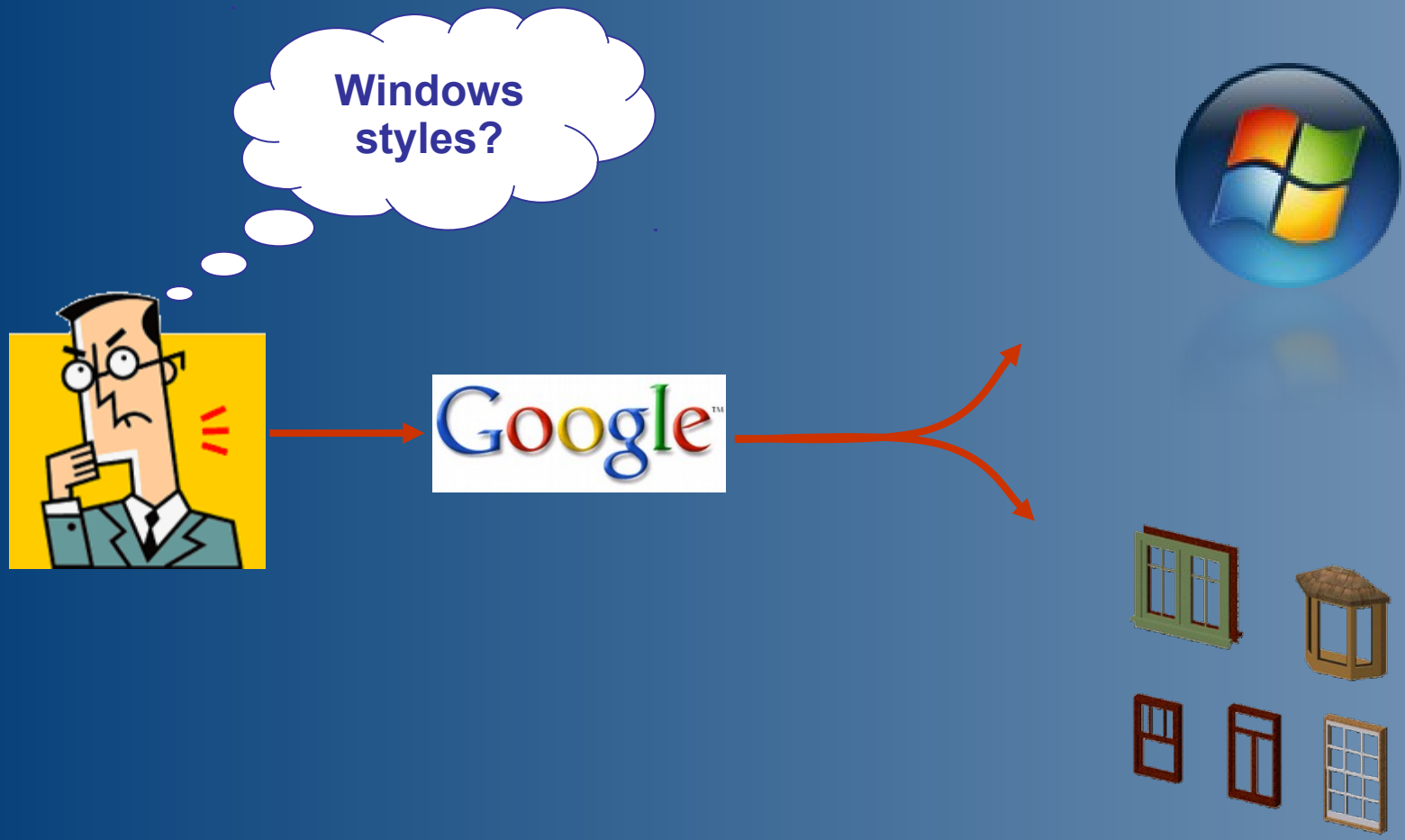
Incremental Methods for Context-Based Web Retrieval

Carlos M. Lorenzetti – Fernando M. Sagui
Ana G. Maguitman – Guillermo R. Simari
Carlos I. Chesñevar

LIDIA – Universidad Nacional del Sur

Artificial Intelligence Research Group – Universidad de Lleida

Problemas: ambigüedad



Una solución: CONTEXTO

Proponemos:

- identificar términos específicos
- encontrar fuentes relevantes
- generar automáticamente consultas

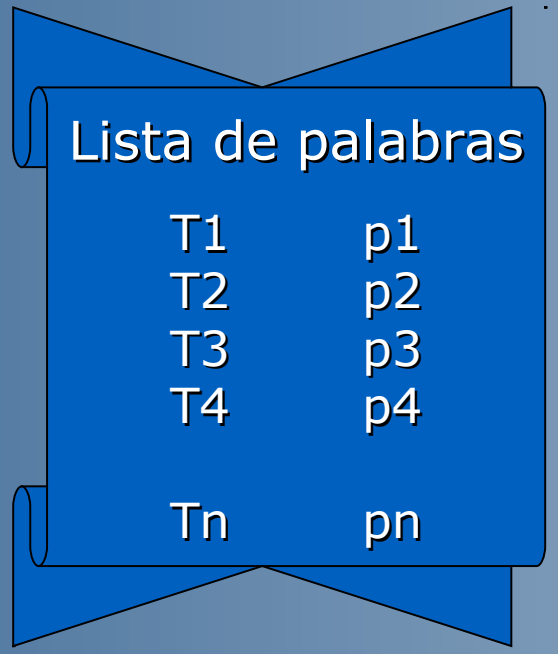
Una solución: CONTEXTO

Contexto

Artículos
Diario
Otros

Lista de palabras

Una solución: CONTEXTO



Importancia de los términos

Método tradicional: TF-IDF

emplea la forma más simple

$$TFIDF(d,t) = TF(d,t) \times IDF(t)$$

Importancia de los términos

Método tradicional: TF-IDF

emplea la forma más simple

$$TFIDF(d,t) = TF(d,t) \times IDF(t)$$

Cuenta las apariciones de un término en el documento

Penaliza a aquella palabras que son muy comunes

Importancia de los términos

Método Propuesto: Incremental

- *Descriptores*

  Términos que aparecen muchas veces en documentos de un mismo tópico:

  ¿Sobre qué trata este tema?

- *Discriminadores*

  Términos que sólo aparecen en documentos de un mismo tópico:

  ¿Qué palabras utilizo para encontrar información similar?

Descriptores y Discriminadores

Java

Máquina

Lenguaje

Virtual

Ruby

Programación

Applets

NetBeans

Computadoras

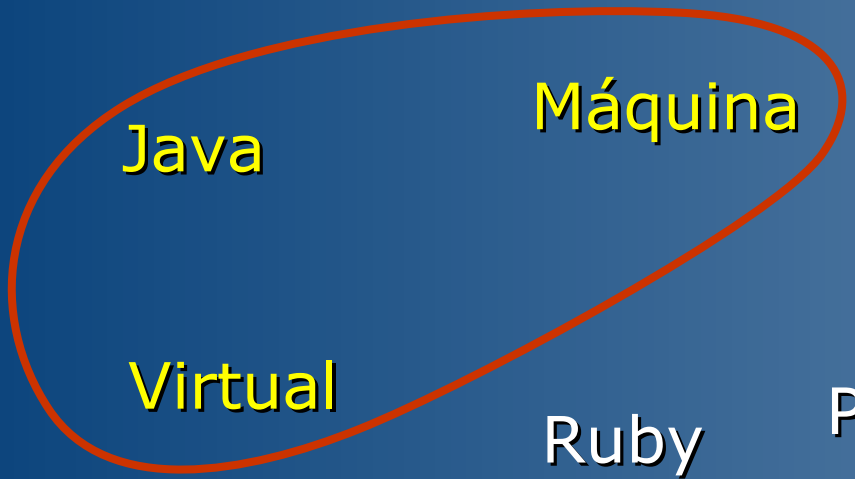
JVM

Código

JDK

Tópico: Máquina Virtual de Java

Descriptores y Discriminadores



Buenos descriptores

Lenguaje

Programación

Ruby

Applets

NetBeans

Computadoras

JVM

Código

JDK

Tópico: Máquina Virtual de Java

Descriptores y Discriminadores

Java

Máquina

Lenguaje

Virtual

Ruby

Programación

Applets

NetBeans

Computadoras

JVM

JDK

Buenos discriminadores

Código

Tópico: Máquina Virtual de Java

Cálculo de Descriptores y Discriminadores

Descriptores y Discriminadores en Documentos

Contexto Inicial		H (1)	(2)	(3)	(4)
java	4	2	5	5	2
máquina	2	6	3	2	0
virtual	1	0	1	1	0
lenguaje	1	0	2	1	1
programación	3	0	2	2	0
café	0	3	0	0	3
isla	0	4	0	0	2
provincia	0	4	0	0	1
jvm	0	0	2	1	0
jdk	0	0	3	3	0

Tópico: Máquina Virtual de Java

(1) espressotec.com
(2) netbeans.org
(3) sun.com
(4) wikitravel.org

$$\mathbf{H}[d_i, t_j] = k$$

Cantidad de ocurrencias del término *k* en el documento *i*

Descriptores de Documentos

Contexto Inicial		$\lambda(d_0, t_j)$
java	4	0,718
máquina	2	0,359
virtual	1	0,180
lenguaje	1	0,180
programación	3	0,539
café	0	0,000
isla	0	0,000
provincia	0	0,000
jvm	0	0,000
jdk	0	0,000

Tópico: Máquina Virtual de Java

Poder descriptivo de un término de un documento

$$\lambda(d_i, t_j) = \frac{\mathbf{H}[i, j]}{\sqrt{\sum_{k=0}^{n-1} (\mathbf{H}[i, k])^2}}$$

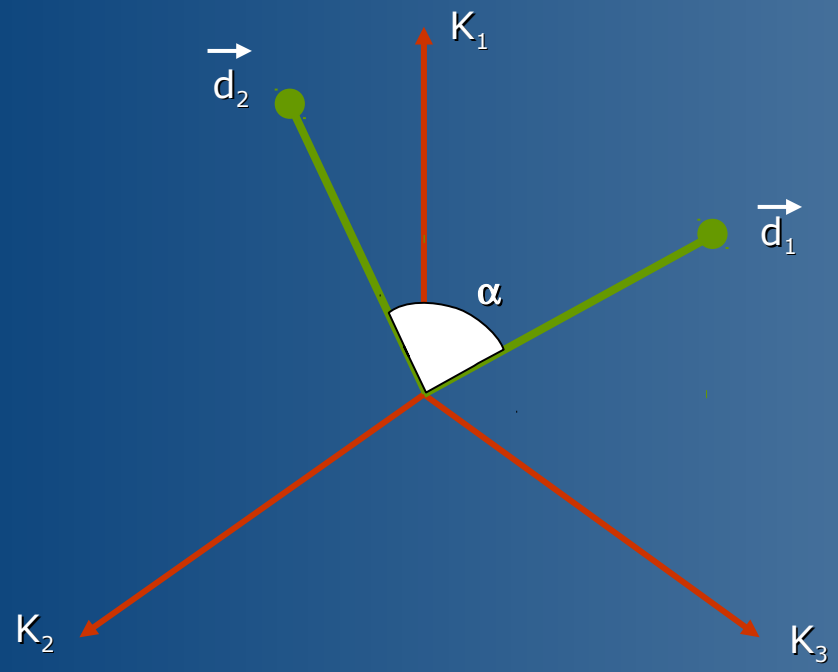
Discriminadores de Documentos

Contexto Inicial		$\delta(t_i, d_0)$
java	4	0,447
máquina	2	0,500
virtual	1	0,577
lenguaje	1	0,500
programación	3	0,577
café	0	0,000
isla	0	0,000
provincia	0	0,000
jvm	0	0,000
jdk	0	0,000

Tópico: Máquina Virtual de Java

Poder discriminante de un término de un documento

$$\delta(t_i, d_j) = \frac{s(\mathbf{H}^T[i,j])}{\sqrt{\sum_{k=0}^{m-1} s(\mathbf{H}^T[i,k])}}$$



Similaridad por coseno

$$\sigma(d_i, d_j) = \sum_{k=0}^{n-1} (\lambda(d_i, t_k) \cdot \lambda(d_j, t_k))$$

Similaridad entre documentos

Descriptores de Tópicos

Contexto Inicial		$\Lambda(d_0, t_j)$
java	4	0,385
máquina	2	0,158
jdk	0	0,124
café	0	0,089
isla	0	0,064
programación	3	0,055
lenguaje	1	0,040
provincia	0	0,040
jvm	0	0,032
virtual	1	0,014

Tópico: Máquina Virtual de Java

Poder descriptivo de un término en el tópico de un documento

$$\Lambda(d_i, t_j) = \frac{\sum_{k=0,\ k \neq i}^{m-1} (\sigma(d_i, d_k) \cdot \lambda(d_k, t_j)^2)}{\sum_{k=0,\ k \neq i}^{m-1} \sigma(d_i, d_k)}$$

Discriminadores de Tópicos

Contexto Inicial		$\Delta(t_i, d_0)$
jvm	0	0,848
jdk	0	0,848
virtual	1	0,566
programación	3	0,566
máquina	2	0,524
lenguaje	1	0,517
java	4	0,493
café	0	0,385
isla	0	0,385
provincia	0	0,385

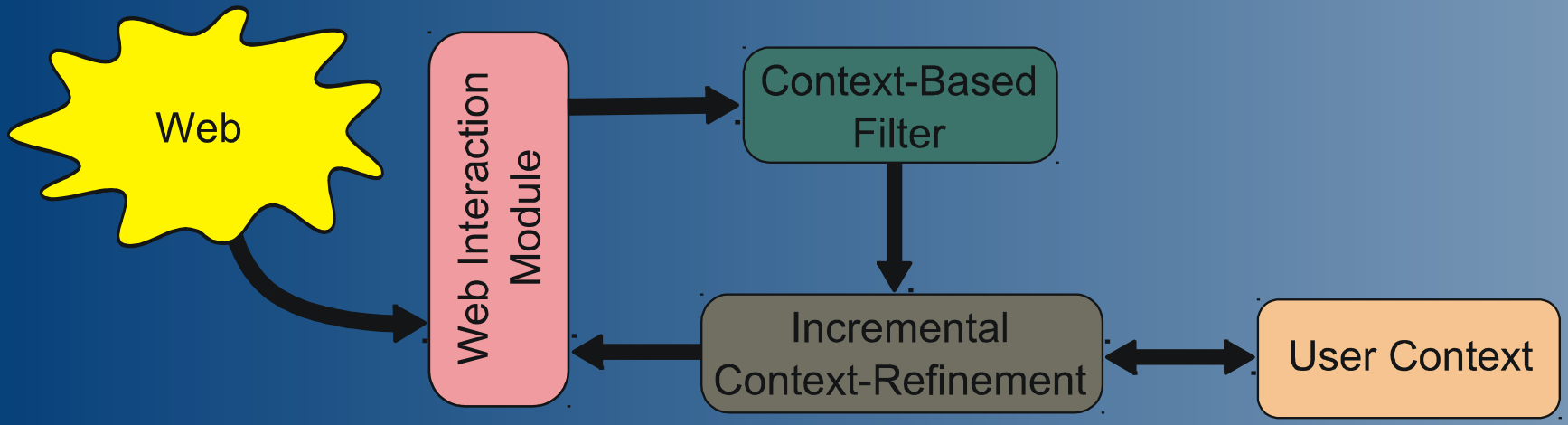
Tópico: Máquina Virtual de Java

Poder discriminante de un término en el tópico de un documento

$$\Delta(t_i, d_j) = \sum_{k=0,\ k \neq j}^{m-1} (\sigma(d_k, d_j) \cdot \delta(t_i, d_k)^2)$$

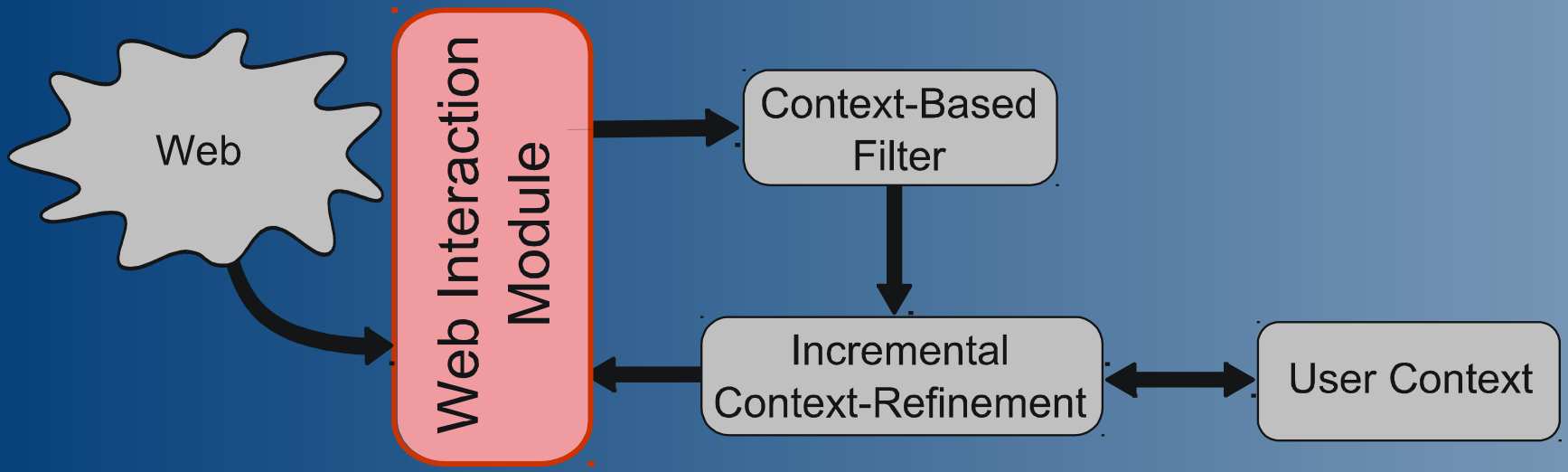
Implementación

Framework

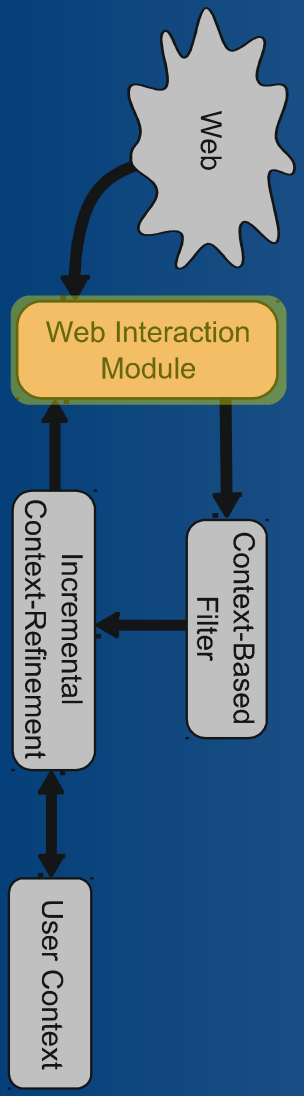

Web
Web Interaction Module
Context-Based Filter
Incremental Context-Refinement
User Context

Framework

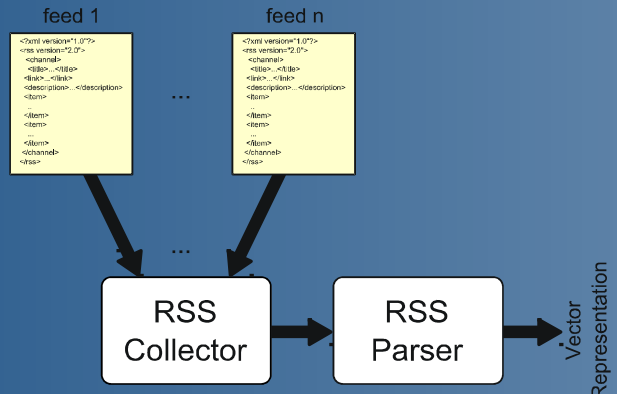
Se encarga de la comunicación con la Web



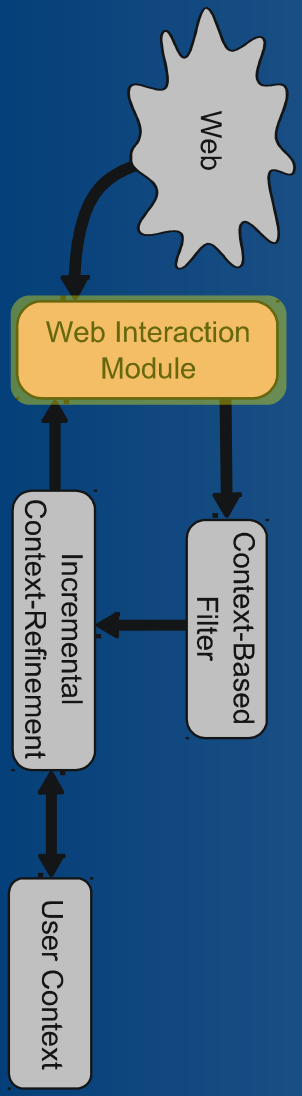
Web Interaction Module



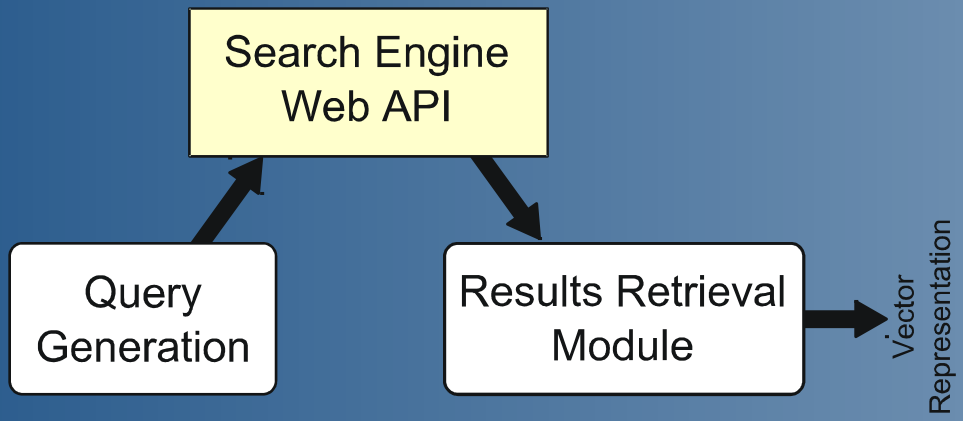
- Feed-based
  - RSSs = resúmenes XML de sitios web
  - Tienen ayudantes metadatos
    - fecha
    - autor
    - descripción
    - lenguaje



Web Interaction Module

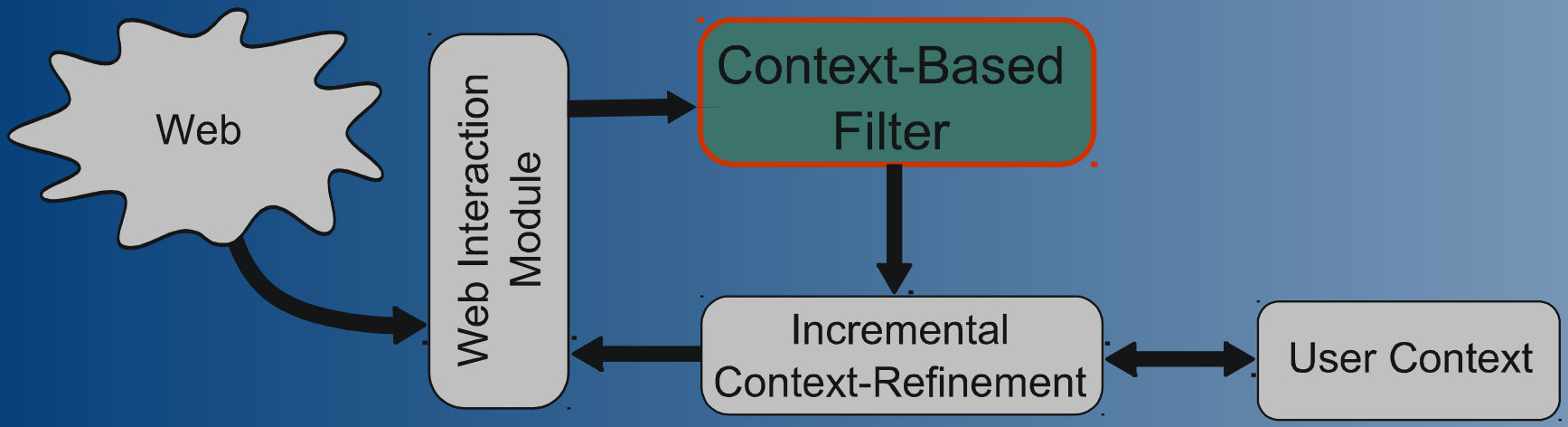


- Query-based
  - Información previamente indexada
  - Hay que generar la consulta



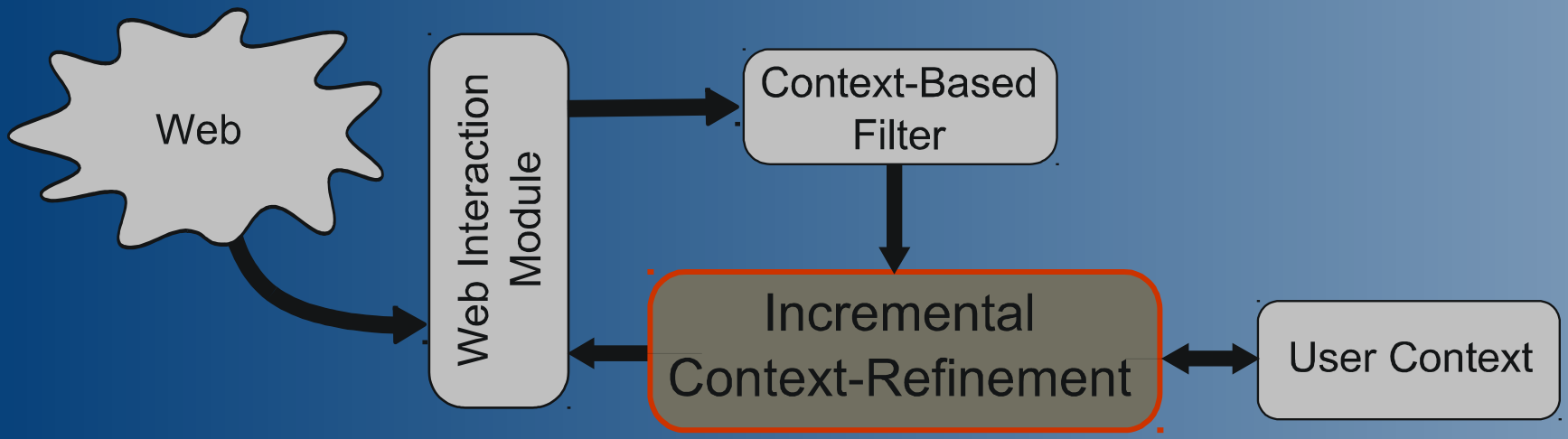
Framework

Estima la importancia del contenido que recibe

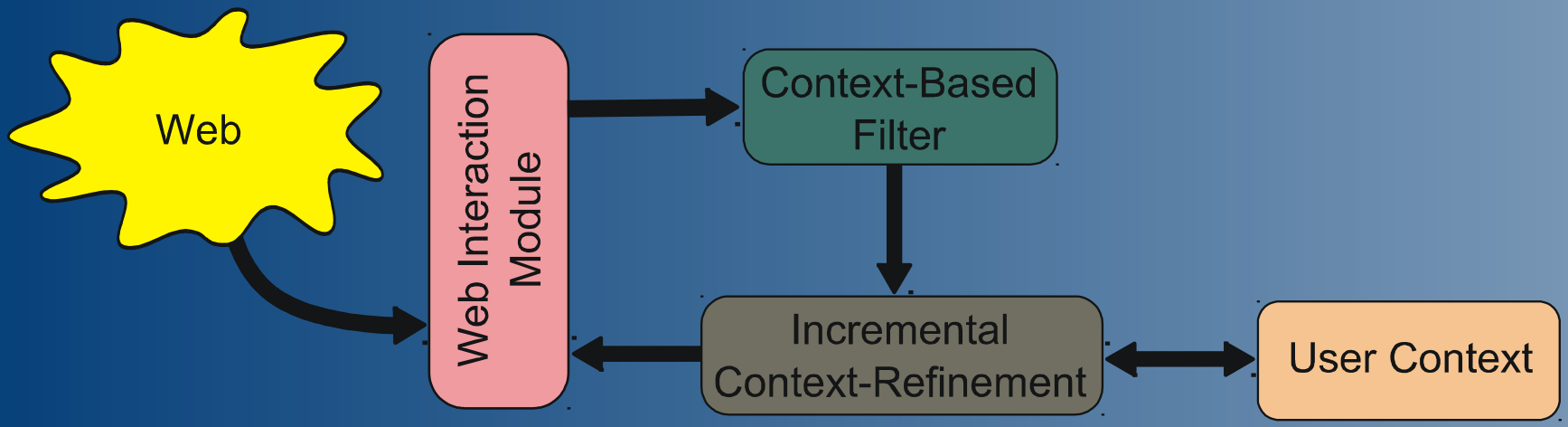


Framework

Refina la representación que tiene del contexto del usuario



Framework


Web
Web Interaction Module
Context-Based Filter
Incremental Context-Refinement
User Context

Evaluación

Evaluación

Método Base

1. Generar la Q(0) usando los términos con TFs más altos
2. i $\leftarrow$ 0
3. Enviar la Q(i) al motor de búsqueda
4. Obtener las respuestas y convertirlas a repres. vectorial
5. Generar una lista ordenada términos L_{TF} por frecuencia
6. i $\leftarrow$ i + 1
7. Q(i) $\leftarrow$ n términos de L_{TF}
8. ir al paso 3

Evaluación

Query–Based

1. Generar la Q(0) usando los términos con TFs más altos
2. i $\leftarrow$ 0
3. Enviar la Q(i) al motor de búsqueda
4. Obtener las respuestas y convertirlas a repres. vectorial
5. Generar una lista ordenada de descriptores, L_Λ
6. Generar una lista ordenada de discriminadores, L_Δ
7. i $\leftarrow$ i + 1
8. Q(i) $\leftarrow$ una combinación de L_Λ y L_Δ
9. ir al paso 3

Evaluación

Contexto Inicial

- 5 páginas en inglés del DMOZ
- Tópico: Recreación

Consulta

- 3 términos L_{Δ} + 2 términos L_{Λ}

Resultados analizados

- Similaridades promedio por iteración

Evaluación

Campamentos en Costa Rica

Consultas generadas

- Método base
  - world costa rica experience programs
  - costa rica abroad program programs
  - study abroad programs costa rica

- Método incremental
  - world costa rica experience programs
  - costa rica **couples families individuals**
  - costa rica adventures **austin cater**
  - adventure costa **hike kick lodge**
  - costa rica **below click damas**

Evaluación

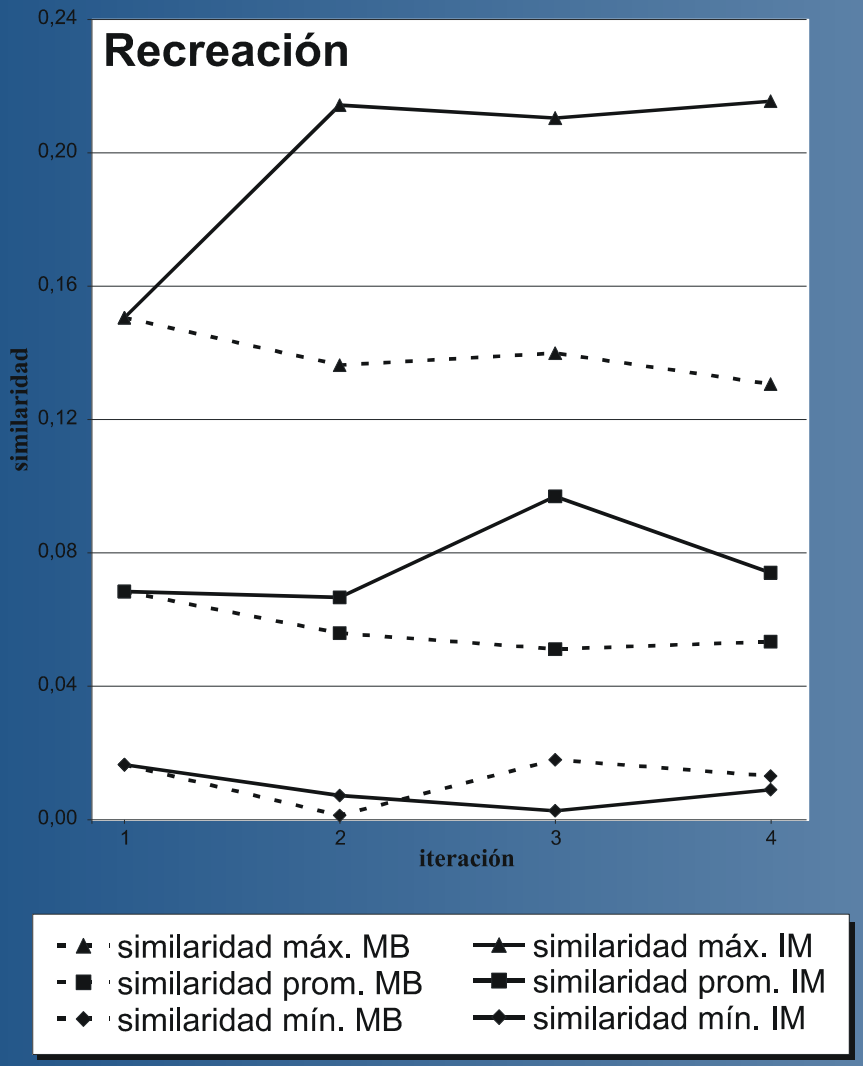
Publicidad acerca de campamentos

Consultas generadas

- Método base
  - koa rv camping soda america
  - camping soda koa rv campgrounds
  - camping koa soda rv campgrounds

- Método incremental
  - koa rv camping soda america
  - soda camping containers copy disposing
  - containers disposing america campgrounds chance
  - camping america **campground** campgrounds **directory**
  - camping **directory cabins usa** koa

Evaluación


Recreación
0,24
0,20
0,16
0,12
0,08
0,04
0,00
similaridad
1
2
3
4
iteración
similaridad máx. MB
similaridad prom. MB
similaridad mín. MB
similaridad máx. IM
similaridad prom. IM
similaridad mín. IM

Trabajo a Futuro

- Adaptación
- Métodos cualitativos
- Evaluaciones intensivas

Incremental Methods for Context-Based Web Retrieval

Carlos M. Lorenzetti – Fernando M. Sagui
Ana G. Maguitman – Guillermo R. Simari
Carlos I. Chesñevar

LIDIA – Universidad Nacional del Sur

Artificial Intelligence Research Group – Universidad de Lleida