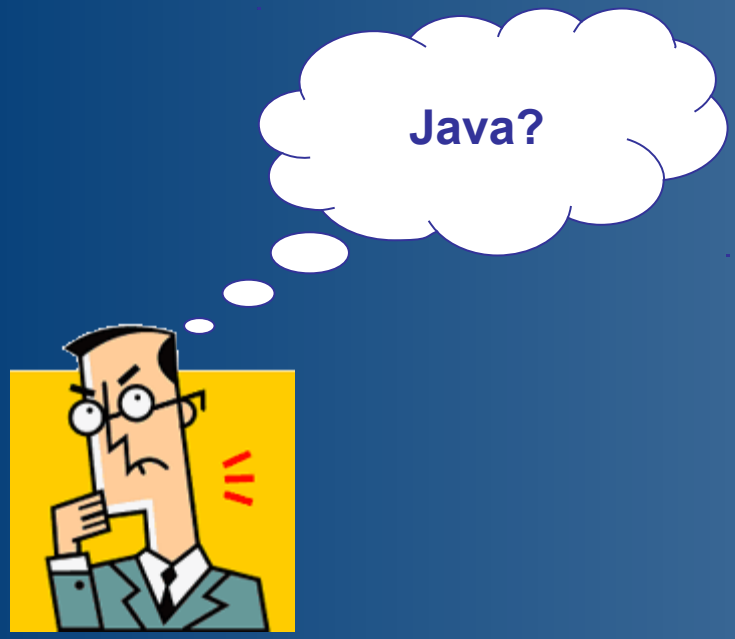
Incremental Methods for Information Access in Context: The Role of Topic Descriptors and Discriminators

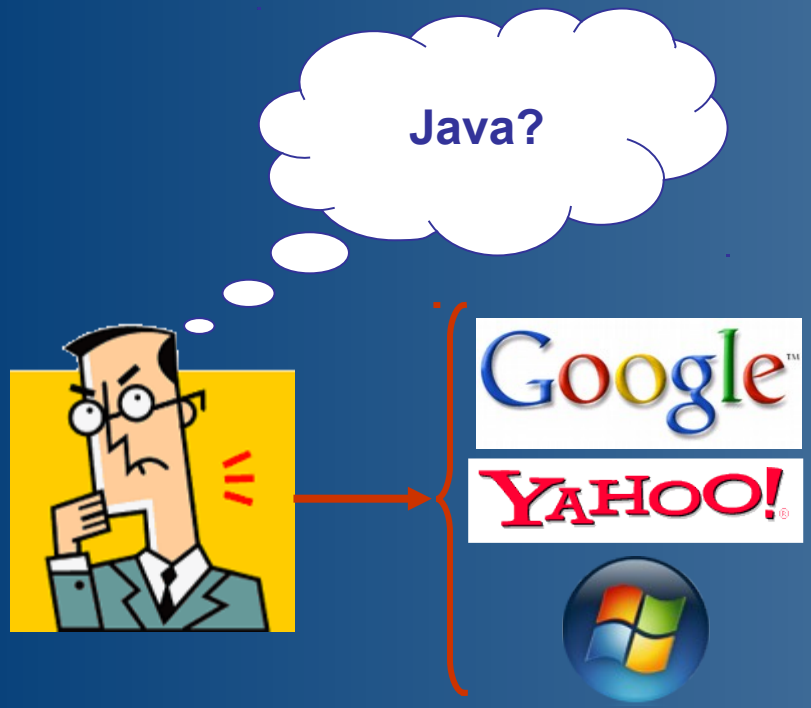
Carlos M. Lorenzetti – Rocío L. Cecchini
Ana G. Maguitman

Laboratorio de Inv. y Des. en IA – Universidad Nacional del Sur

Problemas: ambigüedad

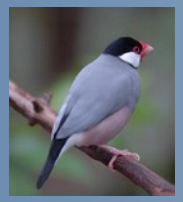


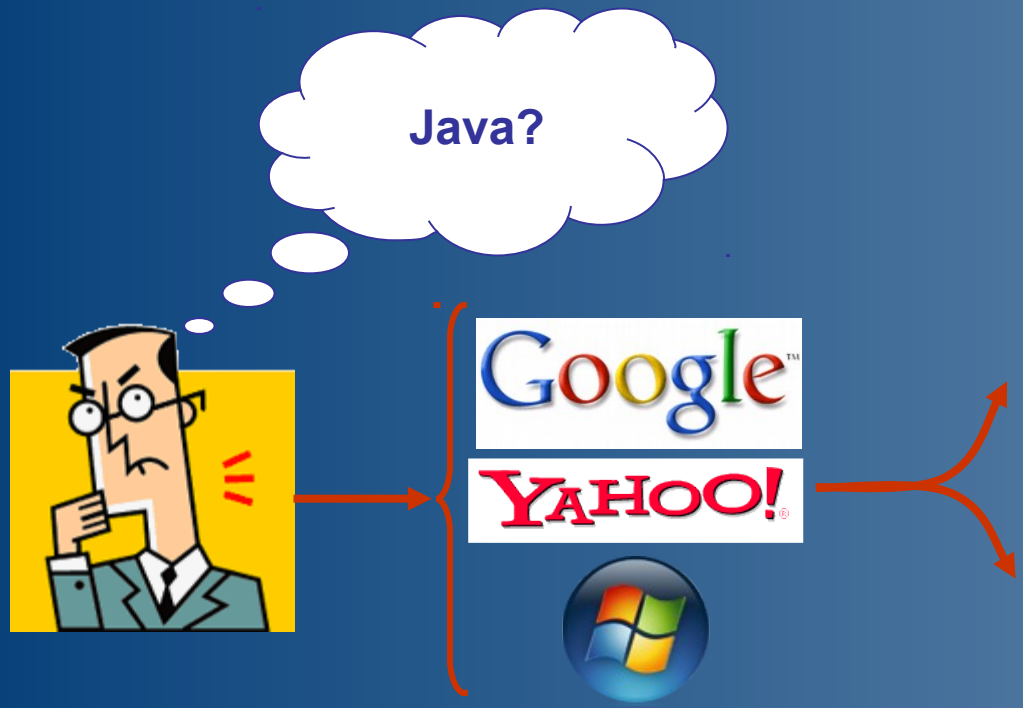
Problemas: ambigüedad



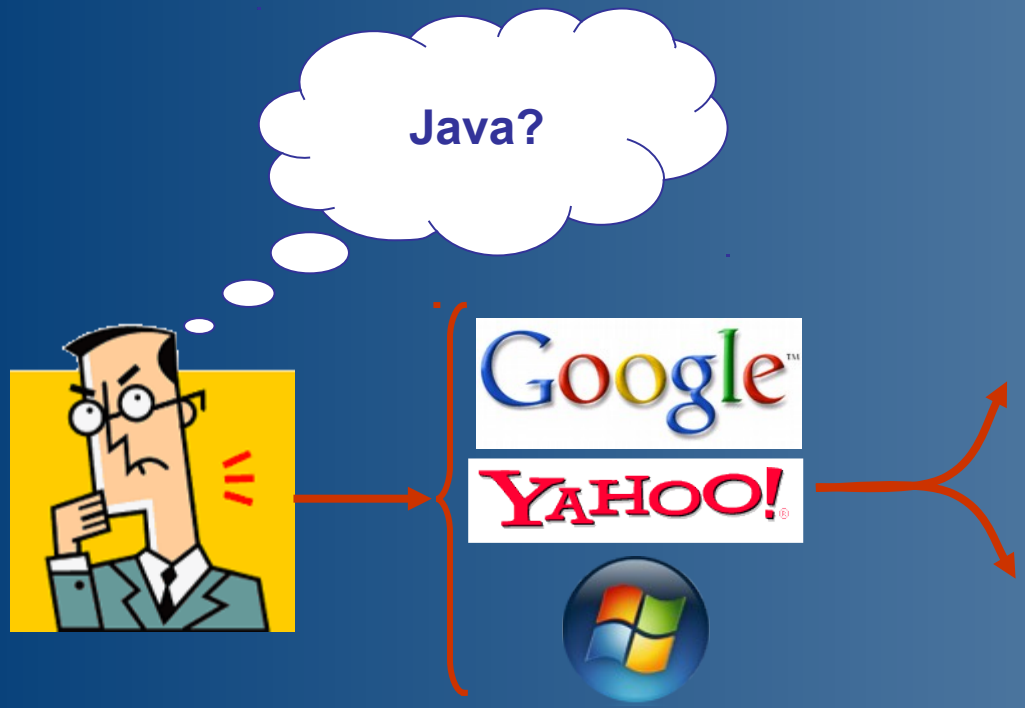
Problemas: ambigüedad

Animales

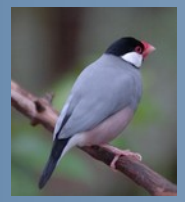




Problemas: ambigüedad

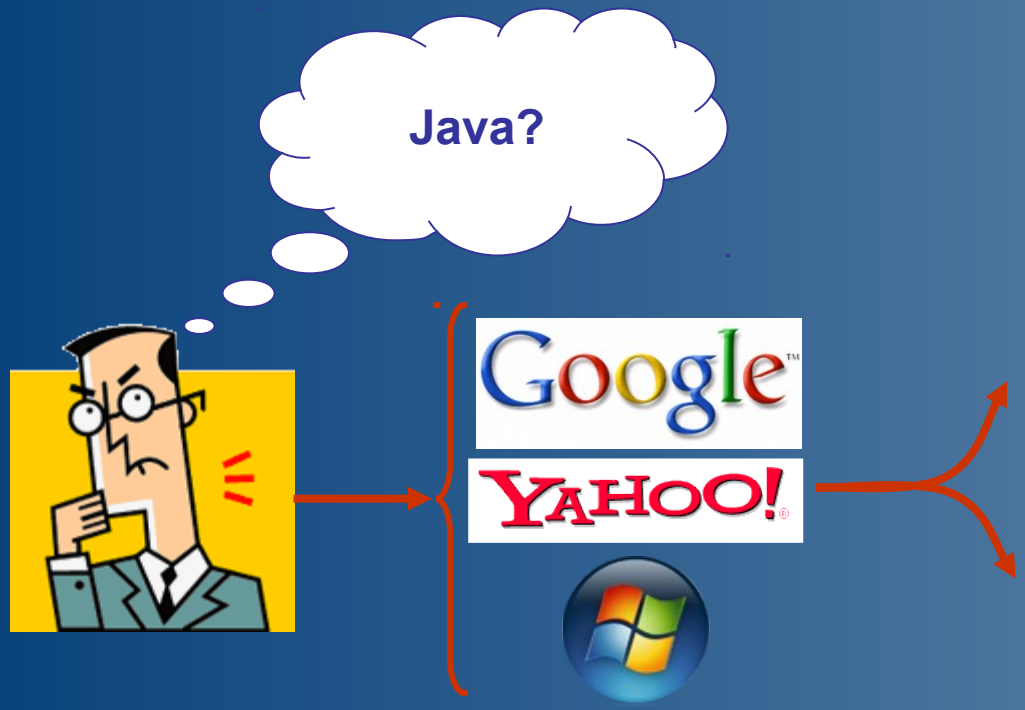


Animales

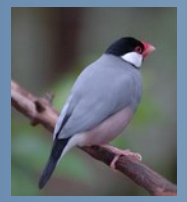


Computación

Problemas: ambigüedad



Animales

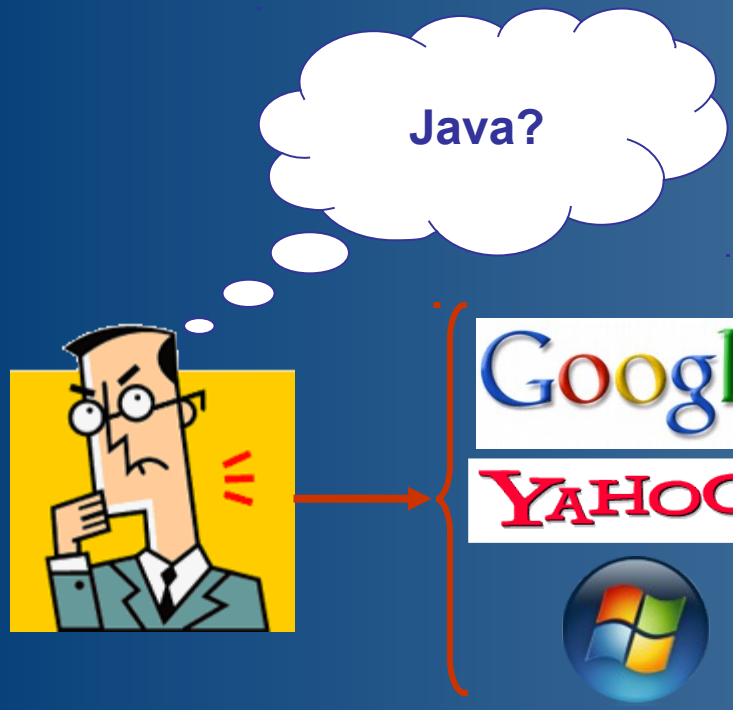


Computación

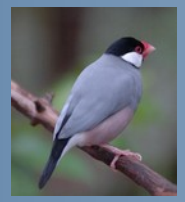
Alimentos

BRITISH AMERICAN TOBACCO

Problemas: ambigüedad



Animales



Computación

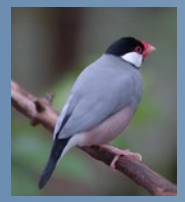
Alimentos

BRITISH AMERICAN TOBACCO

Entretenimiento

Problemas: ambigüedad

Animales



Computación

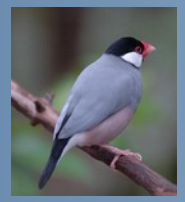
Alimentos

Entretenimiento

Geografía

Problemas: ambigüedad

Animales

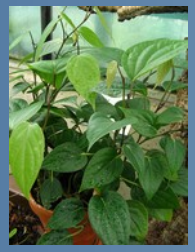


Computación

Alimentos

Entretenimiento

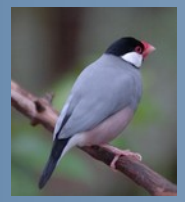
Geografía



Plantas

Problemas: ambigüedad

Animales



Computación

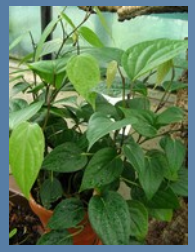
Java

Alimentos

BRITISH AMERICAN TOBACCO

Entretenimiento

Geografía



Plantas

Barcos

Soluciones

Proponemos:

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

Una solución: CONTEXTO

Contexto

Artículos
Diario
Otros

Lista de palabras

Una solución: CONTEXTO

Contexto

Artículos
Diario
Otros

Lista de palabras

T1	p1
T2	p2
T3	p3
T4	p4
Tn	pn

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

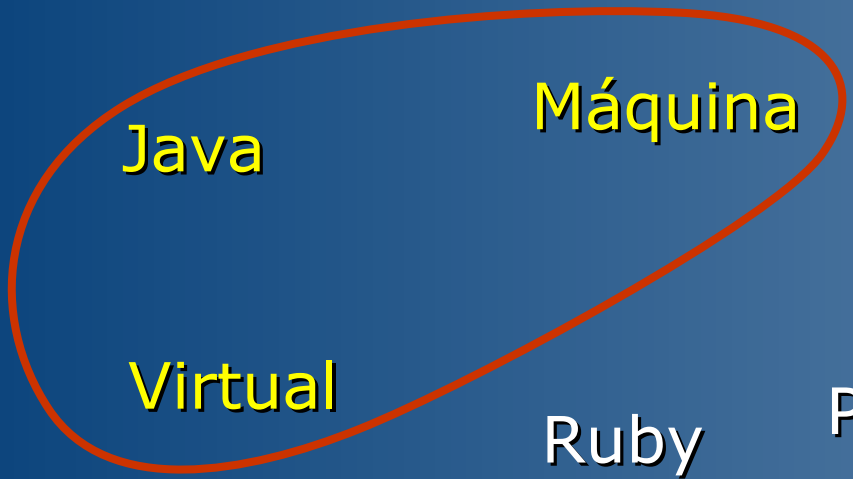
JVM

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

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

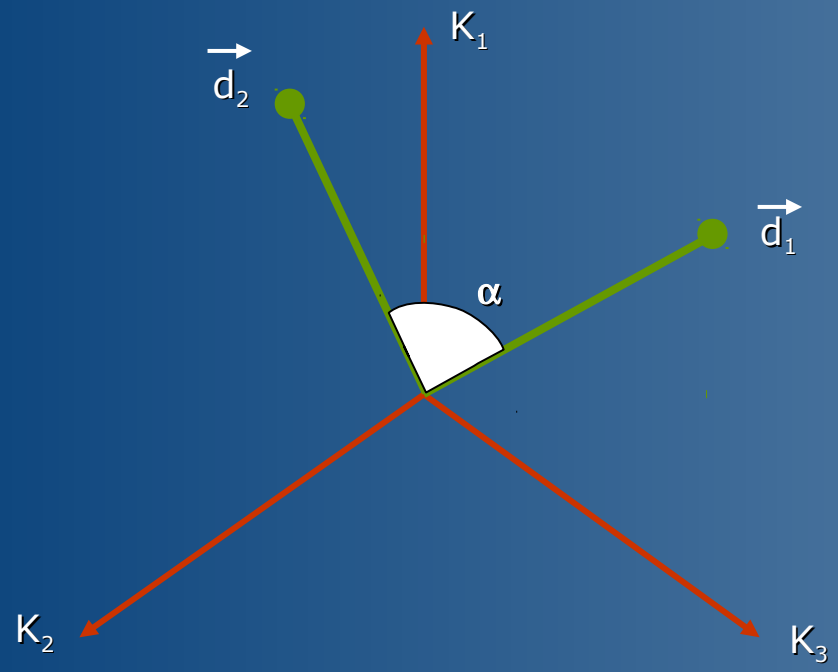
Discriminadores de Documentos

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

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



Criterio de comparación de documentos: *Similitud por coseno*

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

Similitud 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

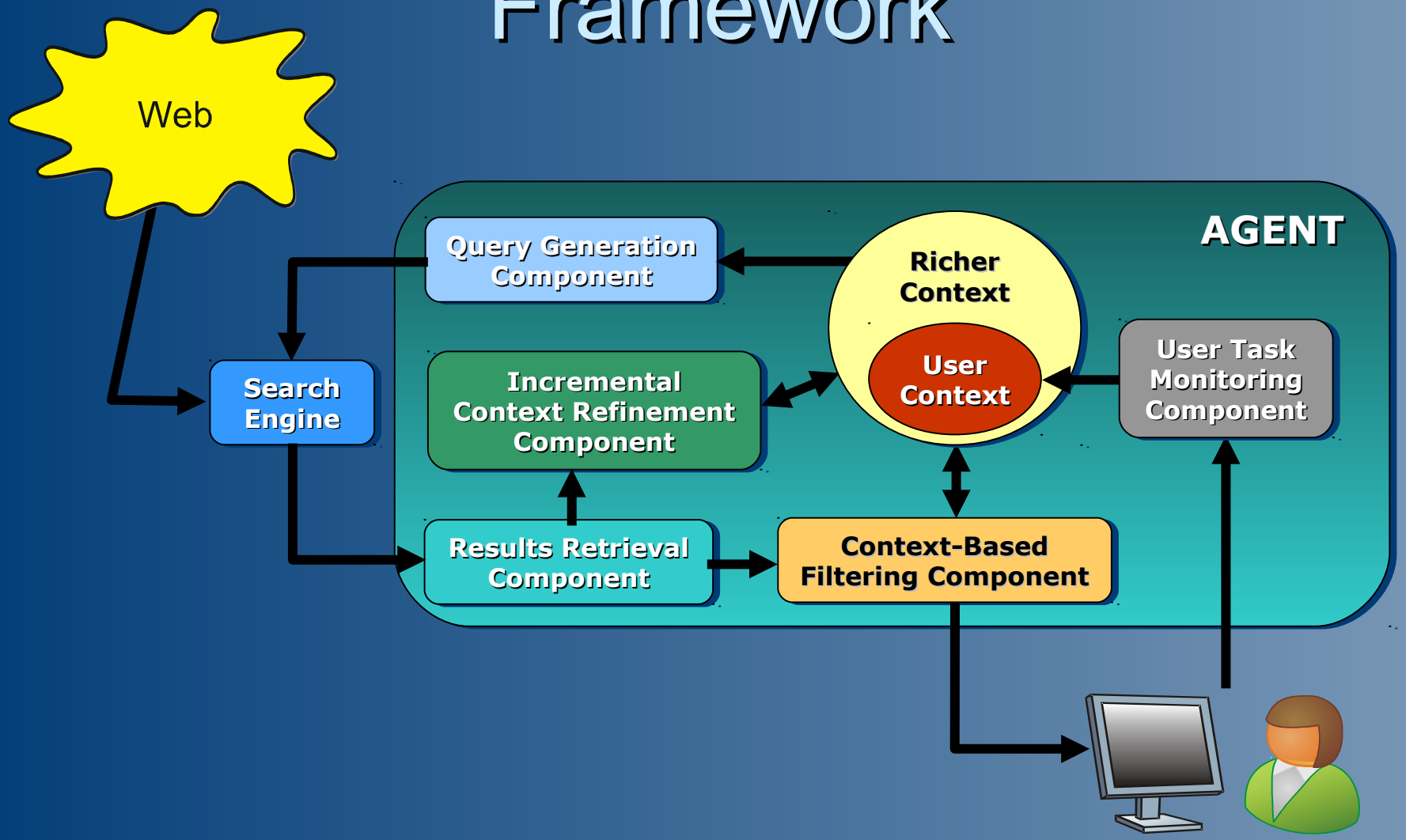
Tópico: Máquina Virtual de Java

Poder discriminante de un término en el tópico de un documento

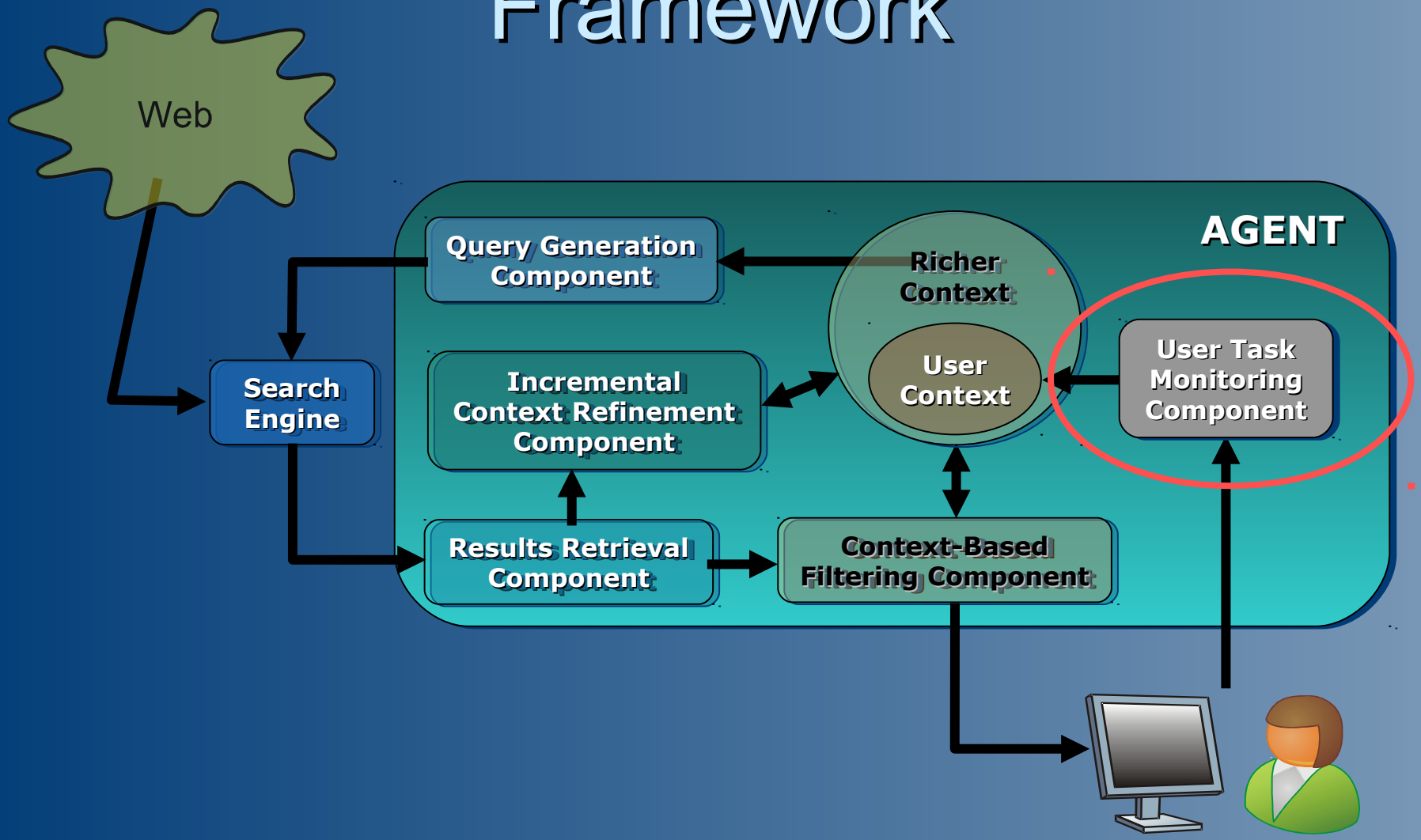
$$\Delta(t_i, d_j) = \sum_{k=0,\ k \neq j}^{m-1} (\sigma(d_k, d_j) \cdot \delta(t_i, d_k)^2)$$

Implementación

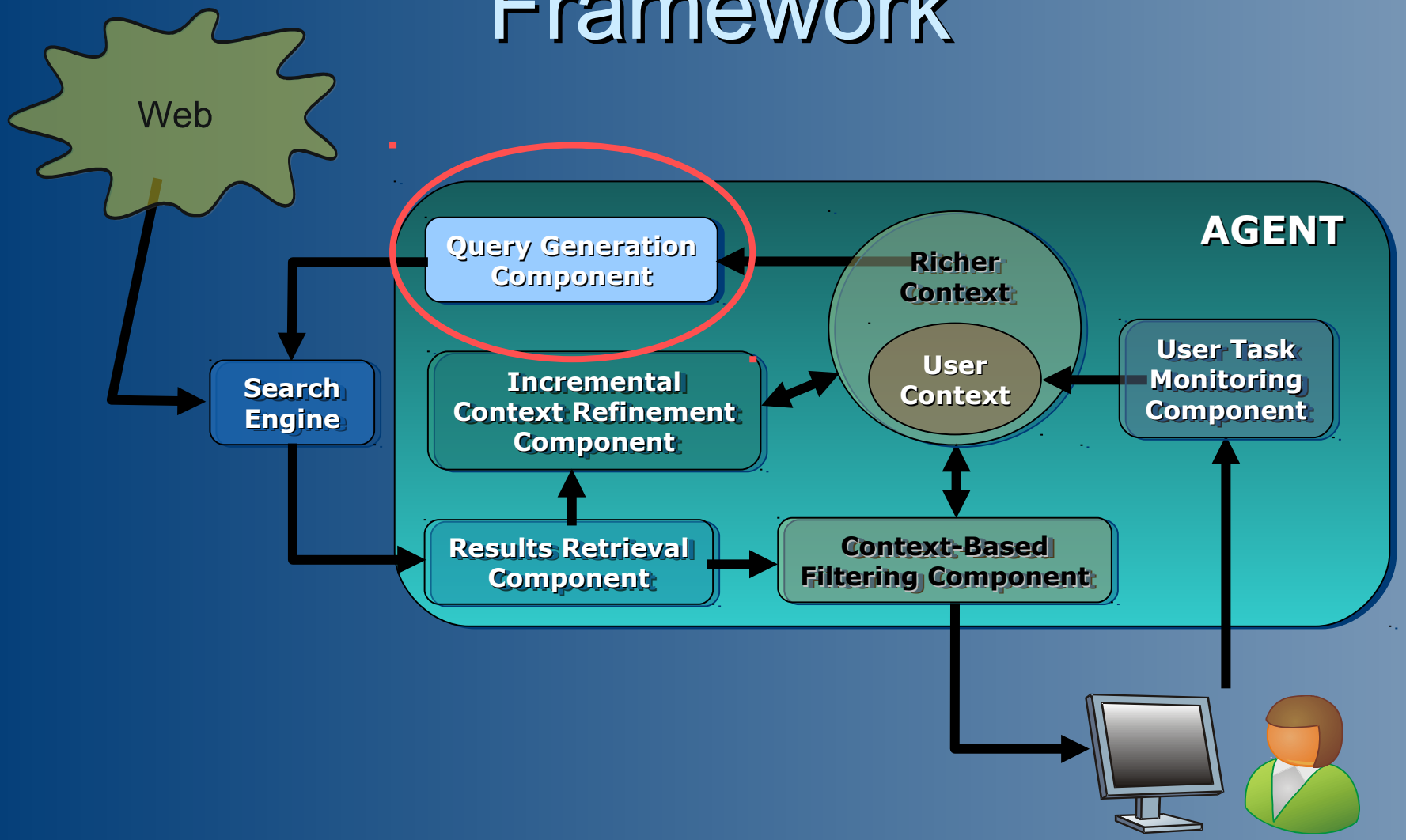
Framework


Web
AGENT
Query Generation Component
Richer Context
User Context
User Task Monitoring Component
Search Engine
Incremental Context Refinement Component
Results Retrieval Component
Context-Based Filtering Component

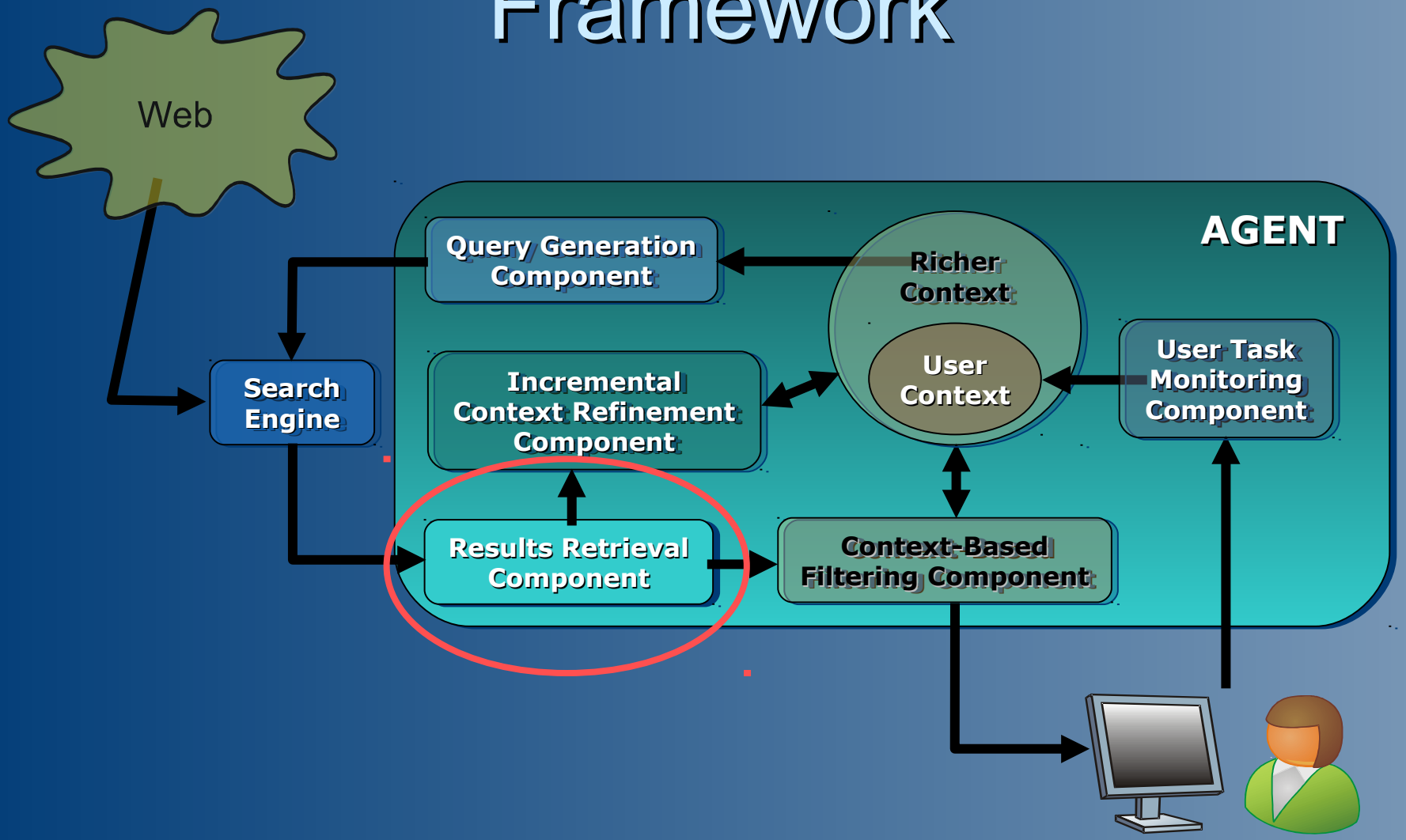
Framework


Web
AGENT
Query Generation Component
Richer Context
User Context
User Task Monitoring Component
Search Engine
Incremental Context Refinement Component
Results Retrieval Component
Context-Based Filtering Component

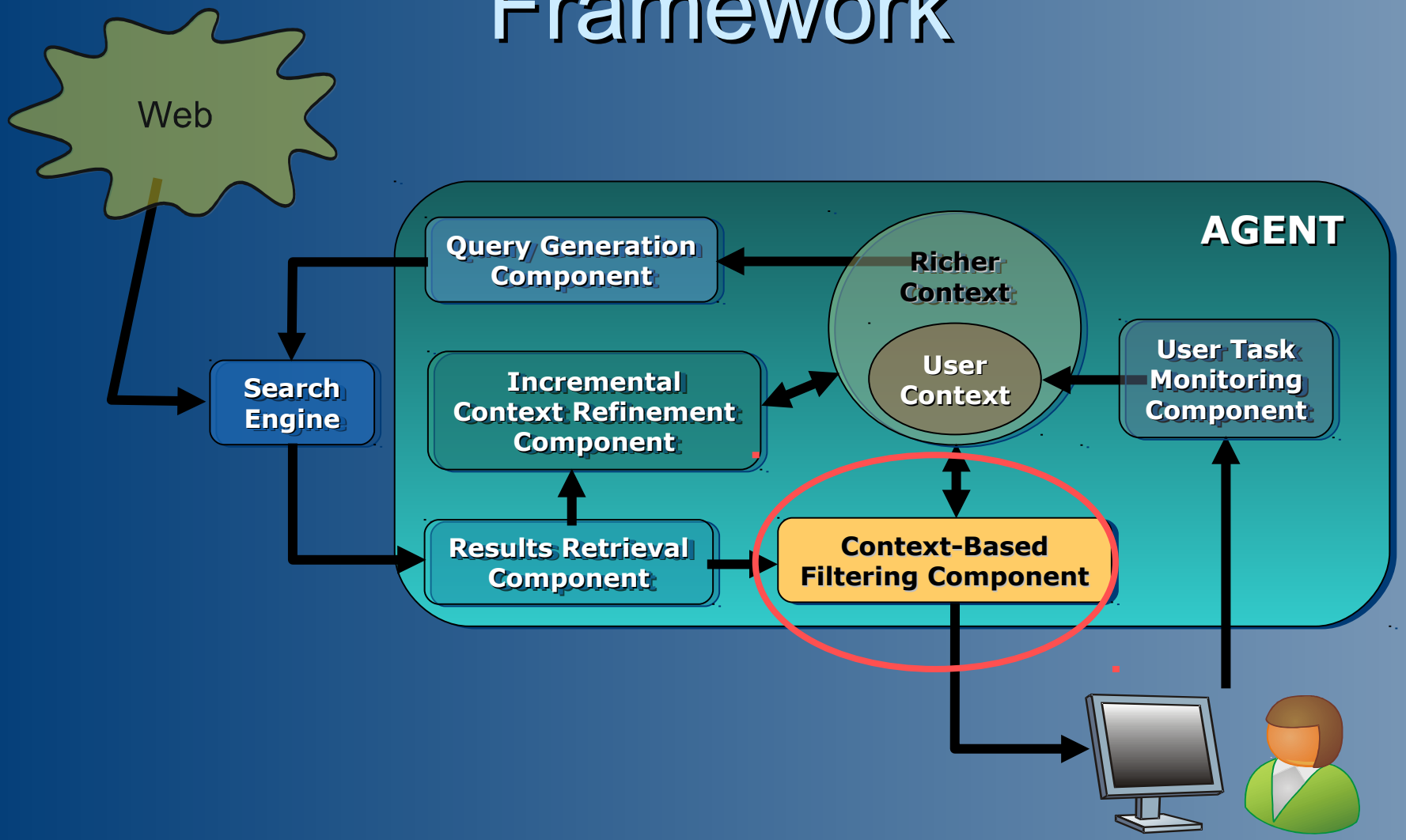
Framework


Web
AGENT
Query Generation Component
Richer Context
User Context
User Task Monitoring Component
Search Engine
Incremental Context Refinement Component
Results Retrieval Component
Context-Based Filtering Component

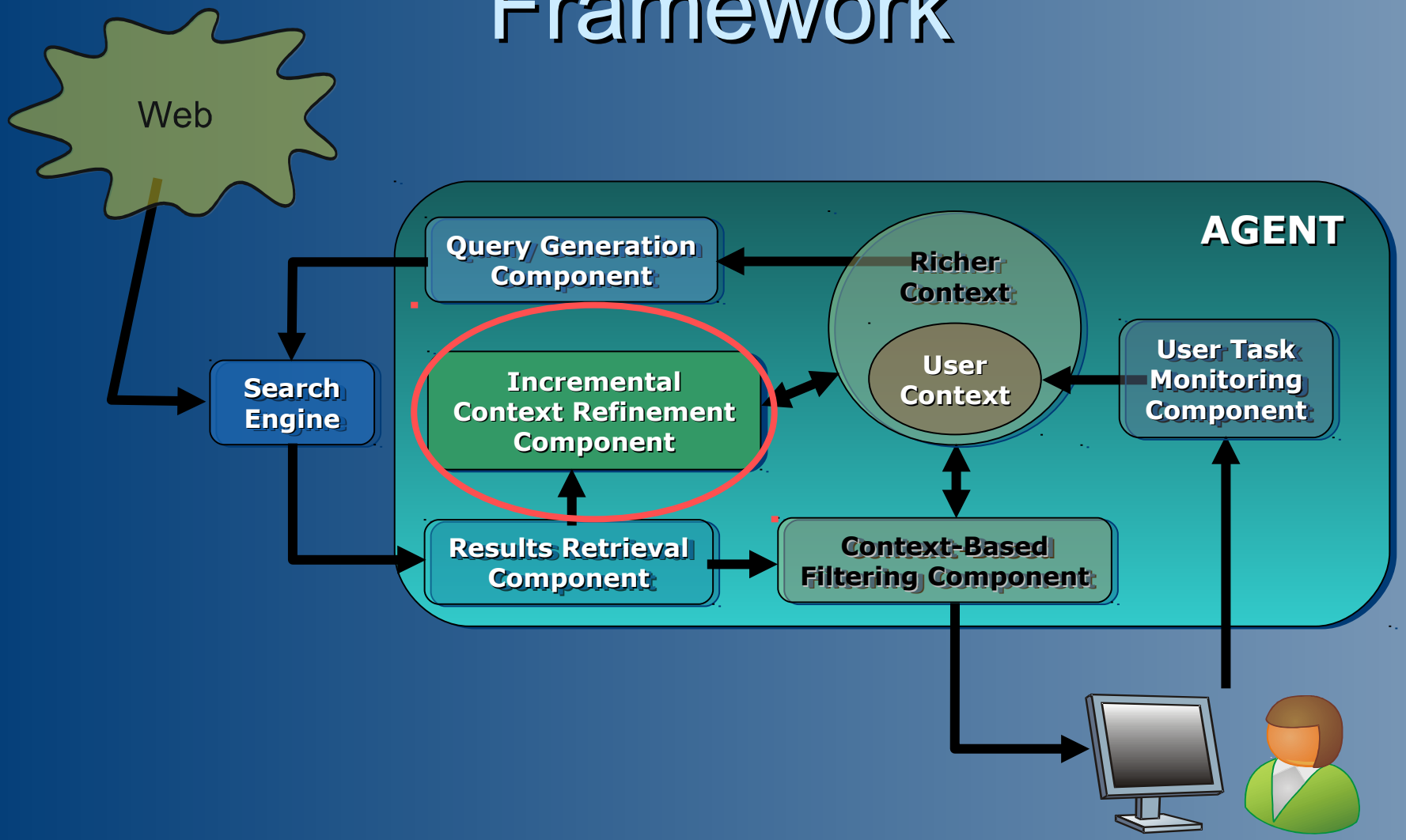
Framework


Web
AGENT
Query Generation Component
Richer Context
User Context
User Task Monitoring Component
Search Engine
Incremental Context Refinement Component
Results Retrieval Component
Context-Based Filtering Component

Framework


Web
AGENT
Query Generation Component
Richer Context
User Context
User Task Monitoring Component
Search Engine
Incremental Context Refinement Component
Results Retrieval Component
Context-Based Filtering Component

Framework


Web
AGENT
Query Generation Component
Richer Context
User Context
User Task Monitoring Component
Search Engine
Incremental Context Refinement Component
Results Retrieval Component
Context-Based Filtering Component

Evaluación

Evaluación

Intelligent Incremental Method

1. Generar *k* consultas usando los términos del contexto
2. Enviar las Q(i) al motor de búsqueda
3. Obtener las respuestas y convertirlas a repres. vectorial
4. Generar una lista ordenada de descriptores, L'_{Λ}
5. Generar una lista ordenada de discriminadores, L'_{Δ}
6. Actualizar incrementalmente las listas L_{Λ} y L_{Δ}
7. Generar *k* consultas, Q(i) ← una combinación de L_{Λ} y L_{Δ}
8. Enviar las Q(i) al motor de búsqueda
9. Obtener las respuestas y convertirlas a repres. vectorial
10. Para cada respuesta, comprobar si es una buena consulta
11. Para cada mala consulta, tratar de reformularla
12. ir al paso 4

Evaluación

Naïve Method

1. Generar *k* consultas usando los términos del contexto
2. Enviar las Q(i) al motor de búsqueda
3. Obtener las respuestas y convertirlas a repres. Vectorial
4. Generar una lista ordenada términos L_{TF} por frecuencia
5. Generar *k* consultas, Q(i) ← una combinación de L_{TF}
6. ir al paso 2

Evaluación

Contexto Inicial

- 15 páginas en inglés del DMOZ
- Tópicos: Recreación, Negocios, Sociedad

Consulta

- 5 términos L_Δ + 1 término L_Λ
- 20 consultas simultáneas

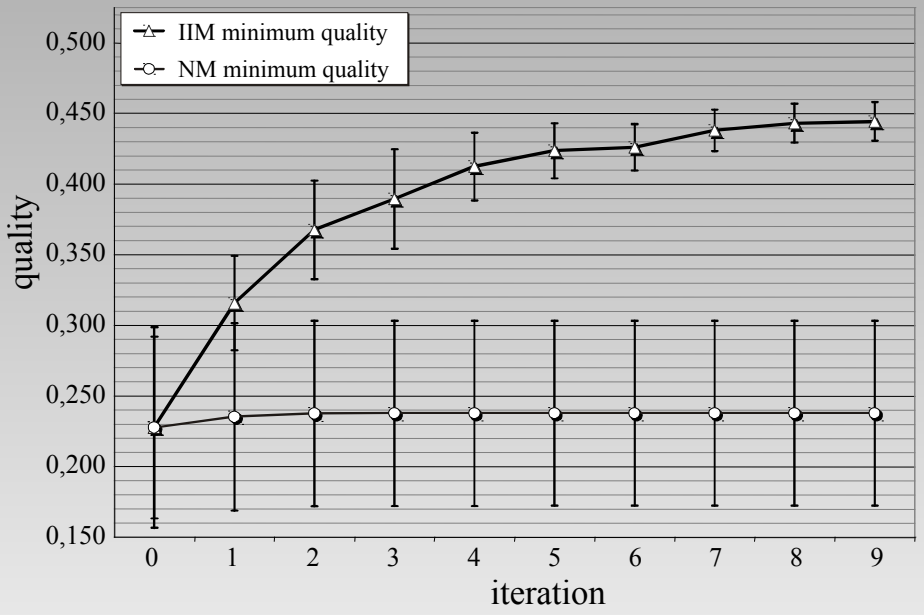
$\forall$ $\alpha = 0.4$

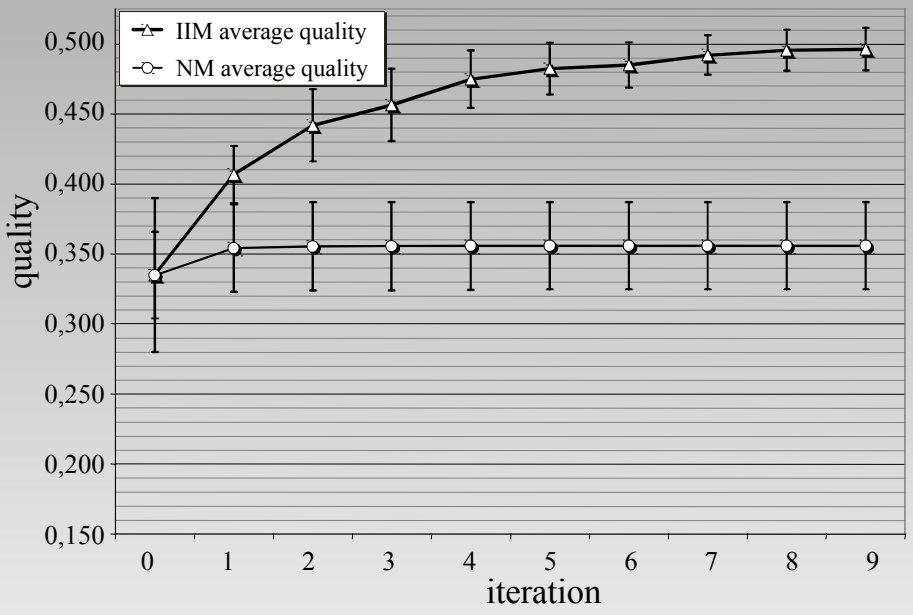
- Motor de búsqueda: Google

Resultados analizados

- Similitudes promedio mínima y promedio por iteración

Evaluación


IIM minimum quality
NM minimum quality
quality
0,500
0,450
0,400
0,350
0,300
0,250
0,200
0,150
0 1 2 3 4 5 6 7 8 9
iteration


IIM average quality
NM average quality
quality
0,500
0,450
0,400
0,350
0,300
0,250
0,200
0,150
0 1 2 3 4 5 6 7 8 9
iteration

Trabajo a Futuro

- Evaluaciones intensivas con colecciones estándar (TREC, etc)
- Mejorar la pérdida del foco
- Evaluar métodos cualitativos para reordenar los resultados basándose en preferencias
- Pruebas con usuarios

Incremental Methods for Information Access in Context: The Role of Topic Descriptors and Discriminators

Carlos M. Lorenzetti – Rocío L. Cecchini
Ana G. Maguitman

Laboratorio de Inv. y Des. en IA – Universidad Nacional del Sur