An Entropy-Based Approach for preserving Diversity in Evolutionary Topical Search

CECILIA BAGGIO – ROCÍO L. CECCHINI

CARLOS M. LORENZETTI – ANA MAGUITMAN

Introducción

Objetivo del trabajo

Introducción

Objetivo del trabajo

- Diseñar nuevas técnicas capaces de refinar consultas de manera automática para acumular recursos relevantes y diversos respecto a un contexto temático.

Introducción

Objetivo del trabajo

- Diseñar nuevas técnicas capaces de refinar consultas de manera automática para acumular recursos relevantes y diversos respecto a un contexto temático.
- Problema común en enfoques anteriores: uso de estrategias evolutivas que resultan en un Recall pobre debido a la pérdida de diversidad genética.

Introducción

Objetivo del trabajo

- Diseñar nuevas técnicas capaces de refinar consultas de manera automática para acumular recursos relevantes y diversos respecto a un contexto temático.
- Problema común en enfoques anteriores: uso de estrategias evolutivas que resultan en un Recall pobre debido a la pérdida de diversidad genética.
- Proponemos una nueva estrategia inspirada en la noción de entropía basada en teoría de la información para favorecer la diversidad de la población con el objetivo de alcanzar **buen recall global**.

Utilidad de generar buenas consultas

- Búsqueda basada en la tarea del usuario

Utilidad de generar buenas consultas

- Búsqueda basada en la tarea del usuario
- Recolección de recursos web para portales temáticos

Utilidad de generar buenas consultas

- Búsqueda basada en la tarea del usuario
- Recolección de recursos web para portales temáticos
- Búsqueda en la Deep-web

Utilidad de generar buenas consultas

- Búsqueda basada en la tarea del usuario
- Recolección de recursos web para portales temáticos
- Búsqueda en la Deep-web
- Soporte para gestión de conocimiento

Introducción

Objetivo del trabajo

Introducción

Objetivo del trabajo

Introducción

Objetivo del trabajo

Introducción

Objetivo del trabajo

Contexto
Artículos
Diarios
Otros

$q_1 = t_1, t_2$

$q_2 = t_1, t_3, t_4$

...

$q_n = t_5, t_8, .., t_p$

Consultas
Temáticas

Introducción

Objetivo del trabajo

Contexto
Artículos
Diarios
Otros

$q_1 = t_1, t_2$

$q_2 = t_1, t_3, t_4$

...

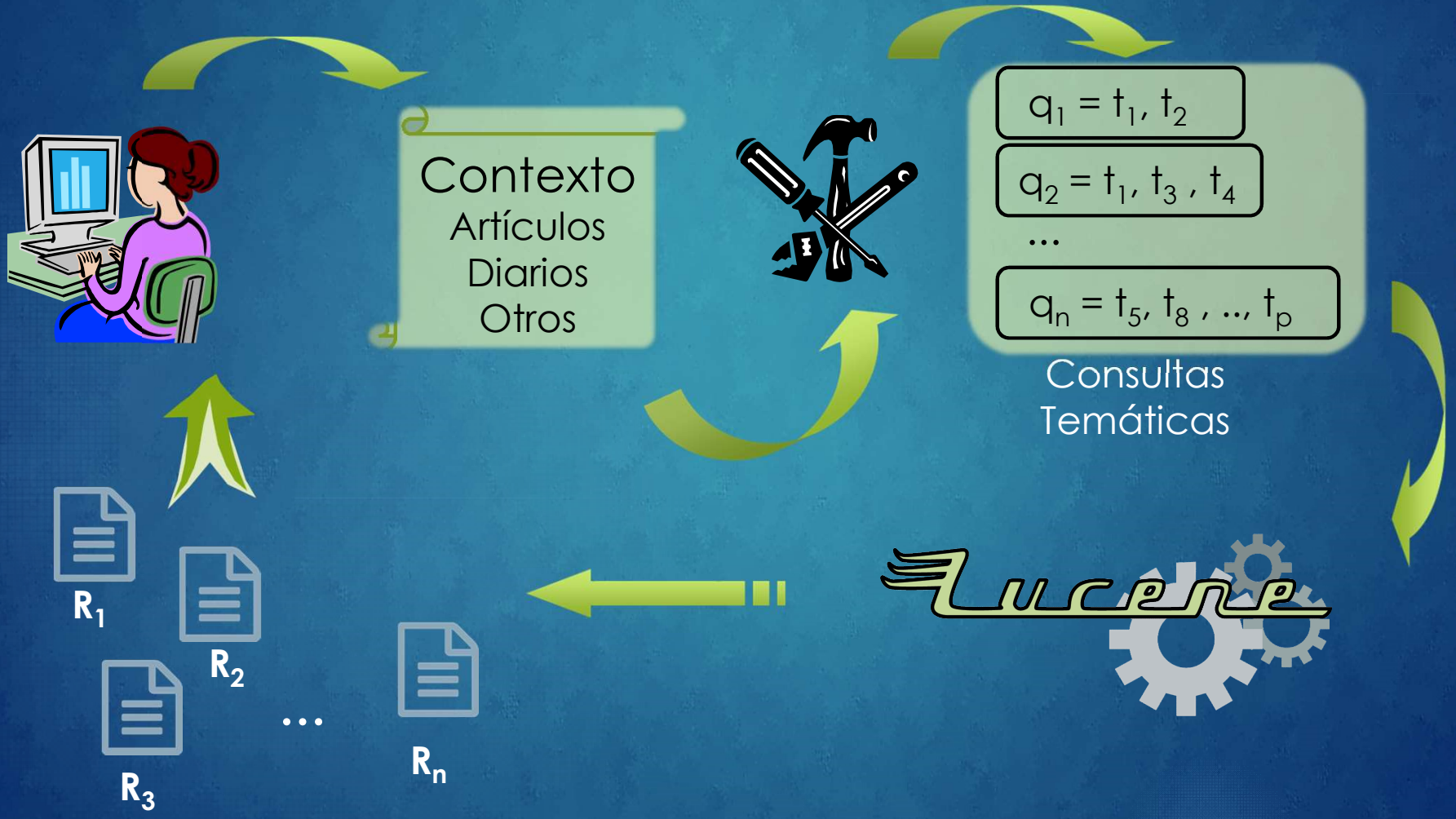
$q_n = t_5, t_8, .., t_p$

Consultas
Temáticas

Lucene

Introducción

Objetivo del trabajo



Introducción

Objetivo del trabajo

Contexto
Artículos
Diarios
Otros

MOEA

$q_1 = t_1, t_2$

$q_2 = t_1, t_3, t_4$

...

$q_n = t_5, t_8, .., t_p$

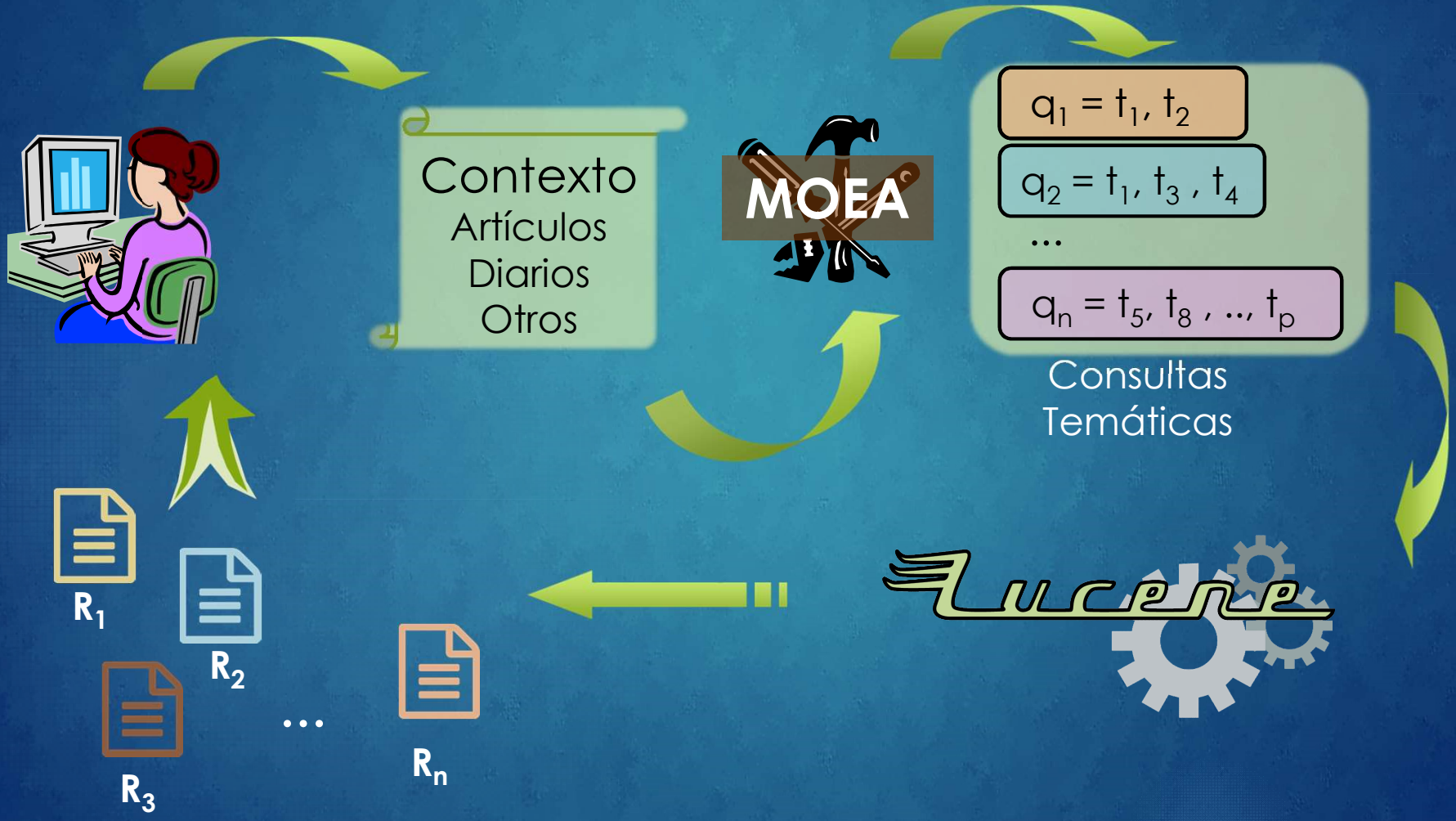
Consultas Temáticas

Lucene

R_1 R_2 R_3 ... R_n

Introducción

Objetivo del trabajo



Motivación del uso de **MOEAs** para el problema de generación de consultas

Motivación del uso de **MOEAs** para el problema de generación de consultas

- Espacio multidimensional de búsqueda.

Motivación del uso de **MOEAs** para el problema de generación de consultas

- Espacio multidimensional de búsqueda.
- Soluciones subóptimas pueden considerarse efectivas.

Motivación del uso de **MOEAs** para el problema de generación de consultas

- Espacio multidimensional de búsqueda.
- Soluciones subóptimas pueden considerarse efectivas.
- Múltiples soluciones. Podemos tener:
  - Resultados satisfactorios en distintos individuos.
  - Interés en encontrar más de una consulta.

Motivación del uso de **MOEAs** para el problema de generación de consultas

- Espacio multidimensional de búsqueda.
- Soluciones subóptimas pueden considerarse efectivas.
- Múltiples soluciones. Podemos tener:
  - Resultados satisfactorios en distintos individuos.
  - Interés en encontrar más de una consulta.
- Exploración y Explotación.

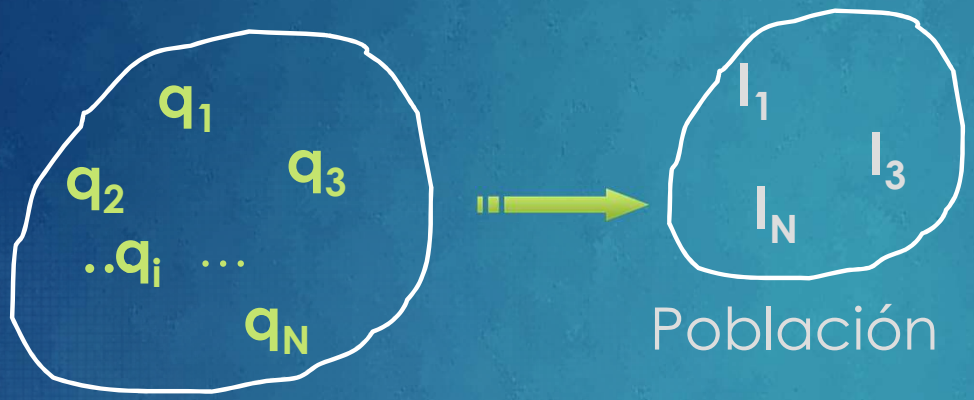
Motivación del uso de **MOEAs** para el problema de generación de consultas

- Espacio multidimensional de búsqueda.
- Soluciones subóptimas pueden considerarse efectivas.
- Múltiples soluciones. Podemos tener:
  - Resultados satisfactorios en distintos individuos.
  - Interés en encontrar más de una consulta.
- Exploración y Explotación.
- Múltiples objetivos y posiblemente conflictivos entre sí.

Evolucionando Consultas Temáticas

Población

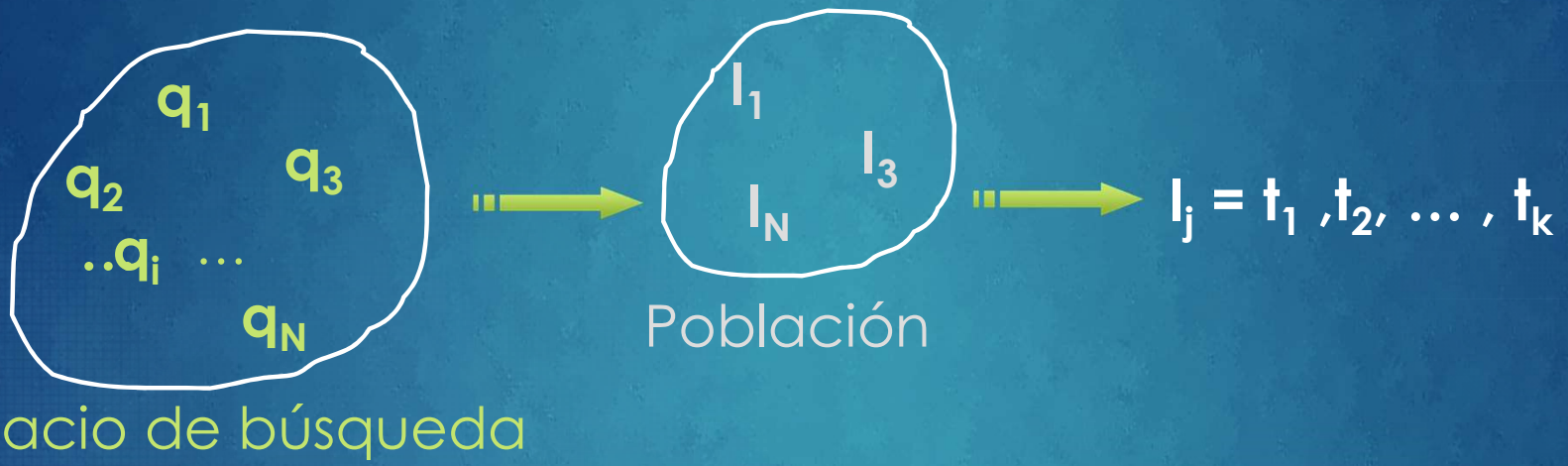
- Representación



Evolucionando Consultas Temáticas

Población

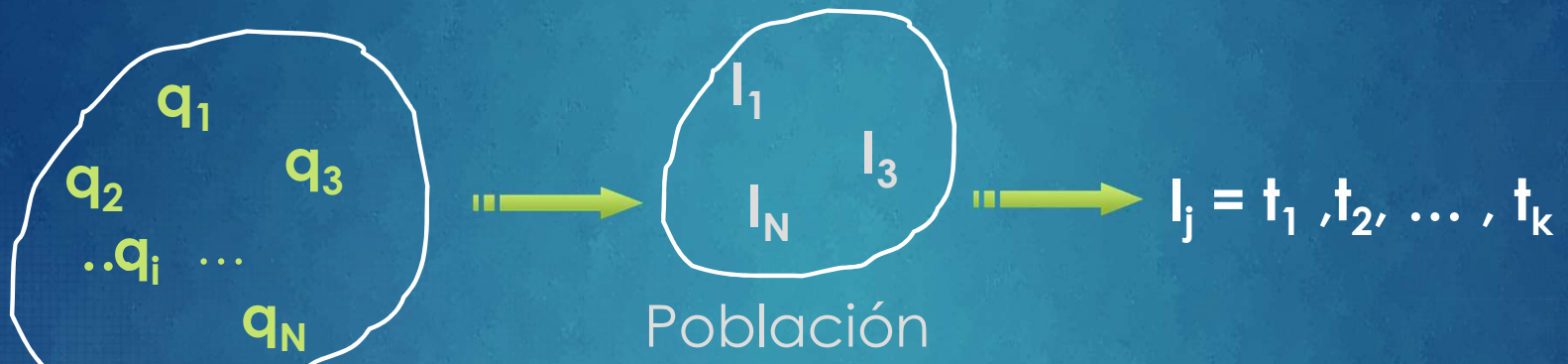
- Representación



Evolucionando Consultas Temáticas

Población

- Representación



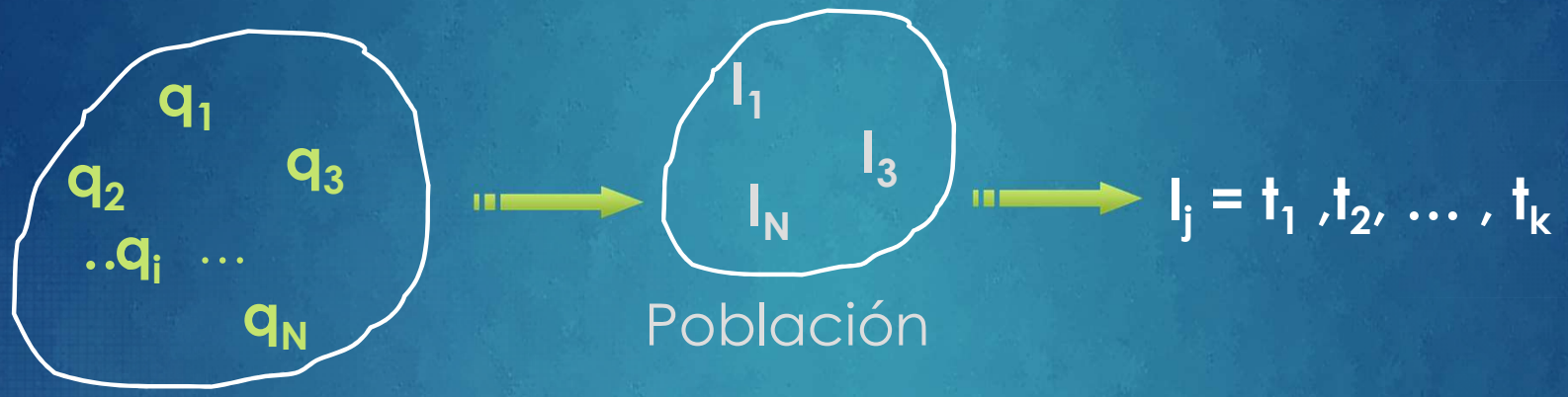
Población

- Inicialización

Evolucionando Consultas Temáticas

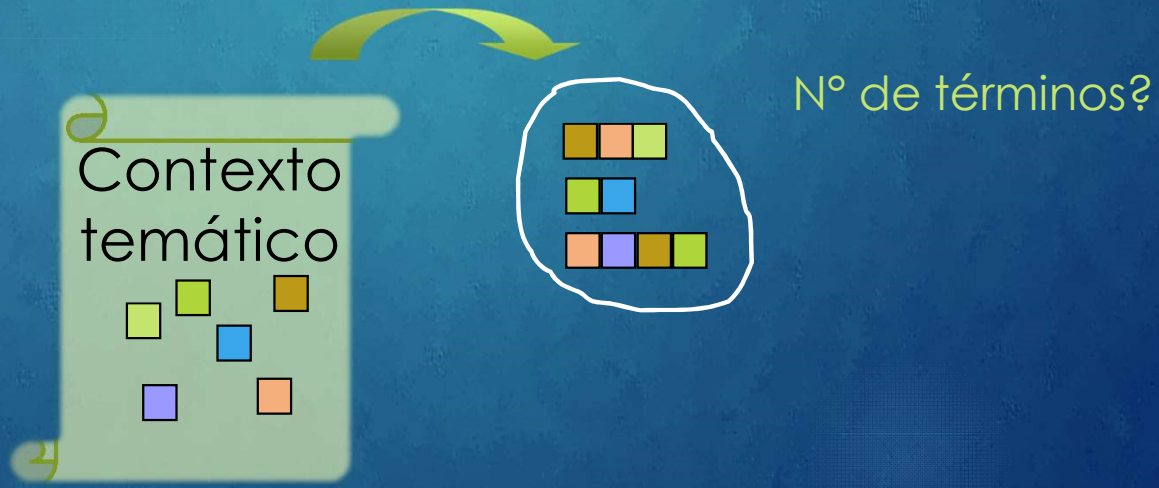
Población

- Representación



Población

- Inicialización



Evolucionando Consultas Temáticas

Función de Fitness

Evolucionando Consultas Temáticas

Función de Fitness

- Sea Q un espacio de búsqueda y T un conjunto de tópicos. Definimos:

$$Precision@10(q,t) = \frac{\left|A_q^{10} \cap D_t\right|}{\left|A_q^{10}\right|}$$

$$Recall(q,t) = \frac{\left|A_q \cap D_t\right|}{\left|D_t\right|}$$

- $\boldsymbol{D_t}$: conjunto que contiene a todos los documentos asociados al tópico **t**.
- $\boldsymbol{A_q}$: conjunto de resultados devuelto por el motor de búsqueda cuando **q** se usa como consulta.
- $\boldsymbol{A_q^{10}}$: es el conjunto de los primeros 10 documentos del ranking $\boldsymbol{A_q}$.

Evolucionando Consultas Temáticas

Función de Fitness

- Sea Q un espacio de búsqueda, T un conjunto de tópicos y P una población de consultas:

$$Recall\_Entrópico(q,t,P) = \frac{\sum_{d_i \in A_q \cap D_t} (\mathrm{IQF}(d_i,P)/\log(|P|+1))}{|D_t|}$$

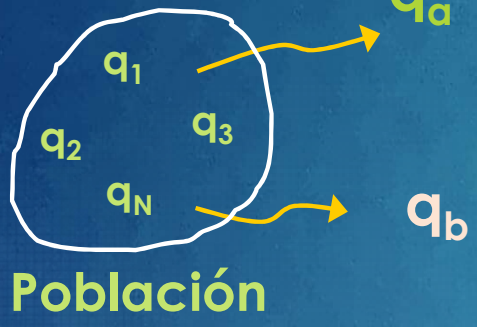
- $\mathrm{IQF}(d_i,P) = \log((|P|+1)/n_i)$

Es la frecuencia inversa en la población de consultas P del documento d_i . Representa la especificidad del documento d_i, donde n_i es el número de consultas que recuperan al documento d_i

Evolucionando Consultas Temáticas

Operadores Genéticos

- Recombinación



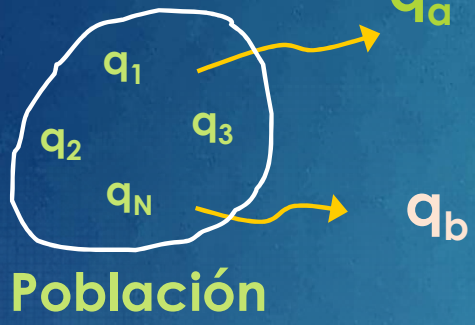
$q_a = t_1, t_2, \ldots, t_k$

$q_b = t_1, t_2, t_3, \ldots, t_m$

Evolucionando Consultas Temáticas

Operadores Genéticos

- Recombinación



$q_a = t_1, t_2, \dots, t_k$

$q'_a = t_1, t_2, t_3, \dots, t_m$

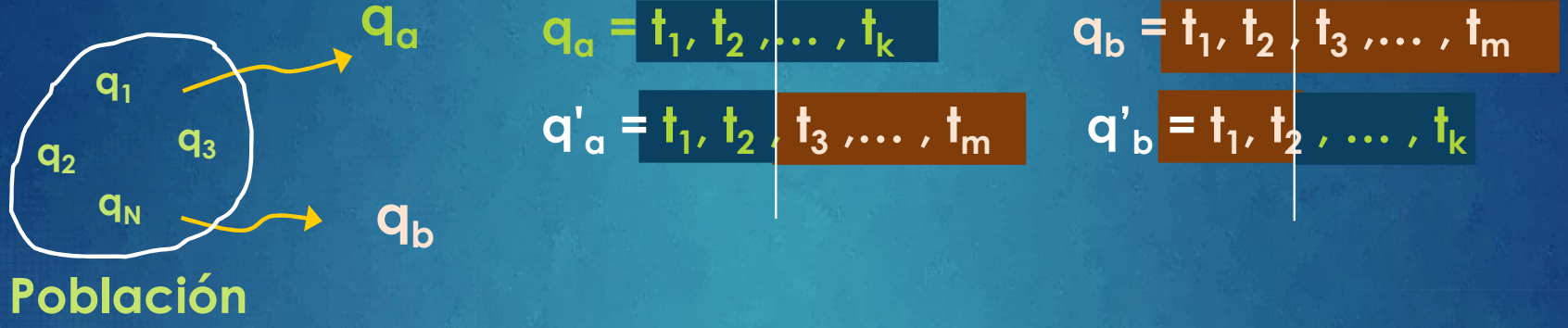
$q_b = t_1, t_2, t_3, \dots, t_m$

$q'_b = t_1, t_2, \dots, t_k$

Evolucionando Consultas Temáticas

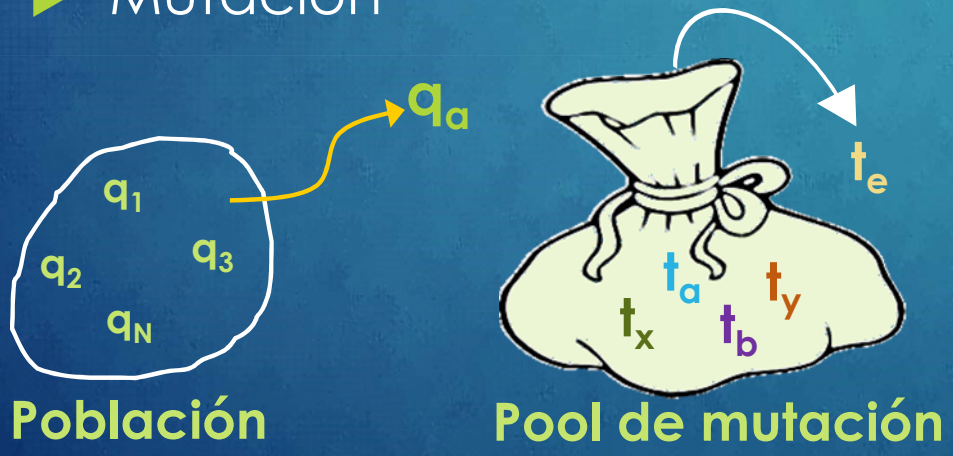
Operadores Genéticos

- Recombinación



Población

- Mutación

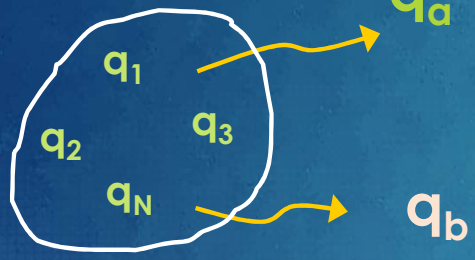


Población **Pool de mutación**

Evolucionando Consultas Temáticas

Operadores Genéticos

- Recombinación



Población

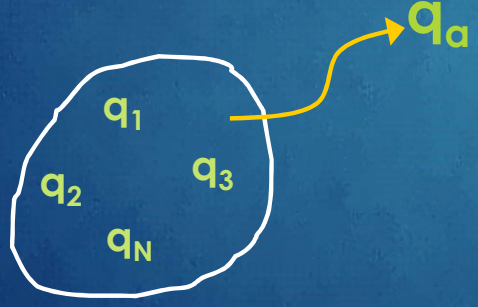
$q_a = t_1, t_2, \ldots, t_k$

$q'_a = t_1, t_2, t_3, \ldots, t_m$

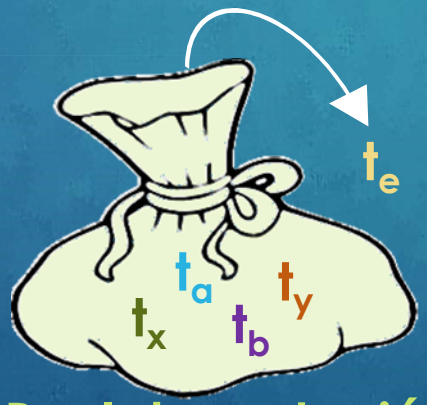
$q_b = t_1, t_2, t_3, \ldots, t_m$

$q'_b = t_1, t_2, \ldots, t_k$

- Mutación



Población



Pool de mutación

Punto de mutación

$q_a = t_1, t_2, t_3, t_4, \ldots, t_p$

$q'_a = t_1, t_2, t_e, t_4, \ldots, t_p$

Evolucionando Consultas Temáticas

Operadores Genéticos

- Selección

Evolucionando Consultas Temáticas

Operadores Genéticos

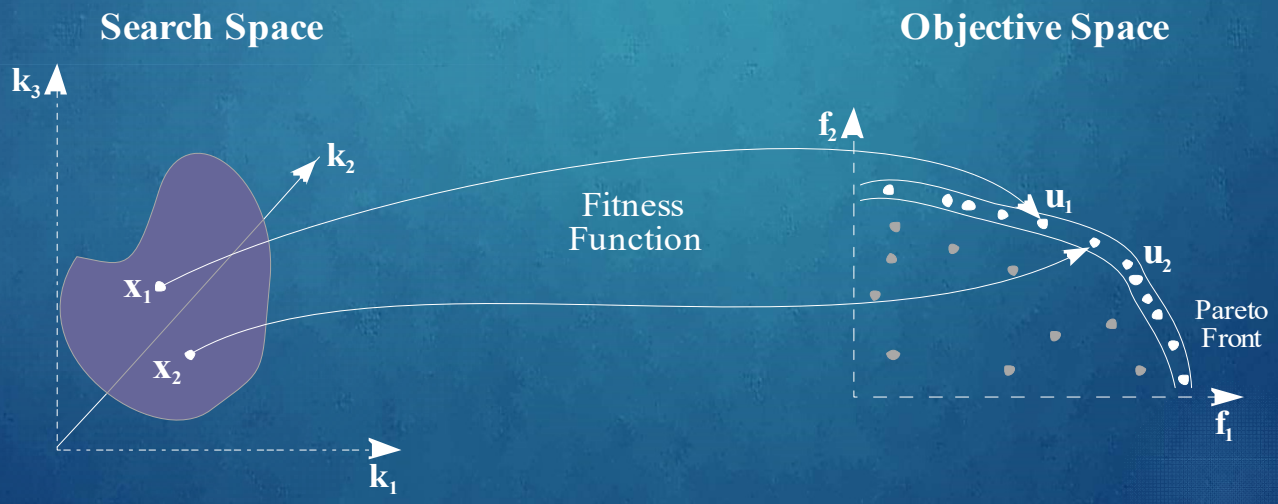
- Selección

  Método **NSGA-II** (Non-dominated Sorting Genetic Algorithm) para evolucionar consultas guiado por los los objetivos dados en las combinaciones:

  Co1: Precision@10 y Recall

  Co2: Precision@10 y Entropic-Recall

  Basado en el concepto de **dominancia de Pareto**



Análisis de la performance sobre conjuntos de entrenamiento y prueba

Análisis de la performance sobre conjuntos de entrenamiento y prueba

Métricas utilizadas

Dado un tópico t, una población de consultas P, una consulta $q \in P$ y el conjunto $A_q=\{a_1,\ldots,a_n\}$ de recursos recuperados para q:

Análisis de la performance sobre conjuntos de entrenamiento y prueba

Métricas utilizadas

Dado un tópico t, una población de consultas P, una consulta $q \in P$ y el conjunto $A_q=\{a_1,\ldots,a_n\}$ de recursos recuperados para q:

- $\overline{Precision@10}(P,t) = \frac{\sum_{q \in P} Precision@10(q,t)}{|P|}$

Análisis de la performance sobre conjuntos de entrenamiento y prueba

Métricas utilizadas

Dado un tópico t, una población de consultas P, una consulta $q \in P$ y el conjunto $A_q=\{a_1,\ldots,a_n\}$ de recursos recuperados para q:

- $\overline{Precision@10}(P,t) = \frac{\sum_{q \in P} Precision@10(q,t)}{|P|}$

- $Global\_Recall(P,t) = \frac{|A(P) \cap D_t|}{|D_t|}$

$$donde\ A(P) = \bigcup_{q_i \in P} A_{q_i}$$

Análisis de la performance sobre conjuntos de entrenamiento y prueba

Métricas utilizadas

- $$\overline{Jaccard\_Similarity\_Index}(P) = \frac{\sum_{q_i, qj \in P, i \neq j} Jaccard\_Similarity\_Index(A^*_{q_i}, A^*_{q_j})}{|P| \,.\, (|P| - 1)}$$

Análisis de la performance sobre conjuntos de entrenamiento y prueba

Métricas utilizadas

- $\overline{Jaccard\_Similarity\_Index}(P) = \frac{\sum_{q_i, qj \in P, i \neq j} Jaccard\_Similarity\_Index(A^*_{q_i}, A^*_{q_j})}{|P| \,.\, (|P| - 1)}$

- Donde:

$$Jaccard\_Similarity\_Index(A^*_{q_i}, A^*_{q_j}) = \frac{|A^*_{q_i} \cap A^*_{q_j}|}{|A^*_{q_i} \cup A^*_{q_j}|}$$

y $A^*_{q_i} = A_{q_i} \cap D_t$

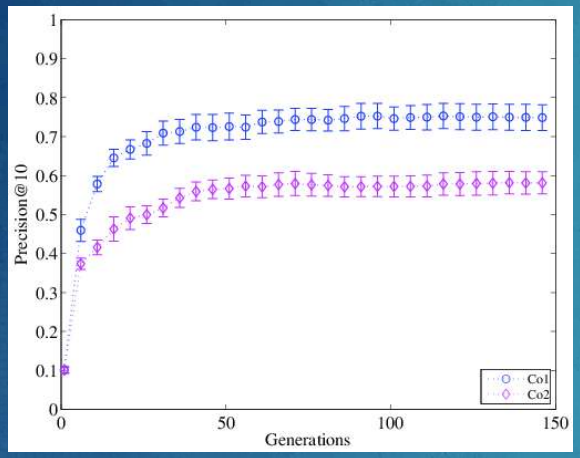
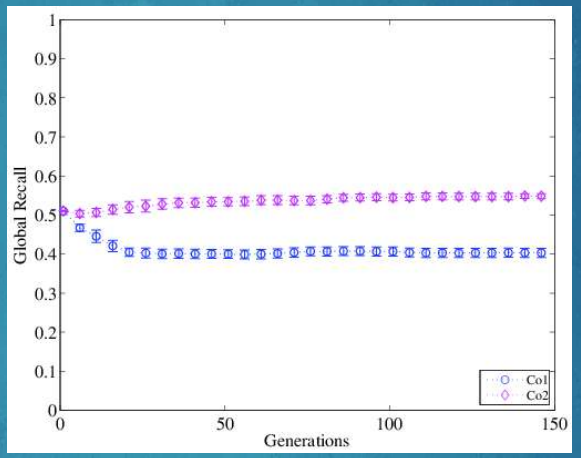
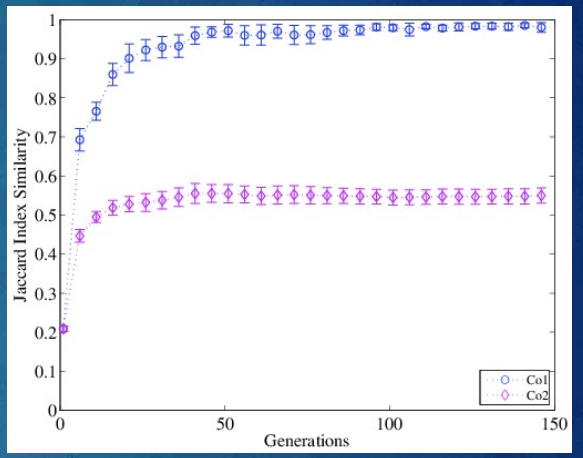
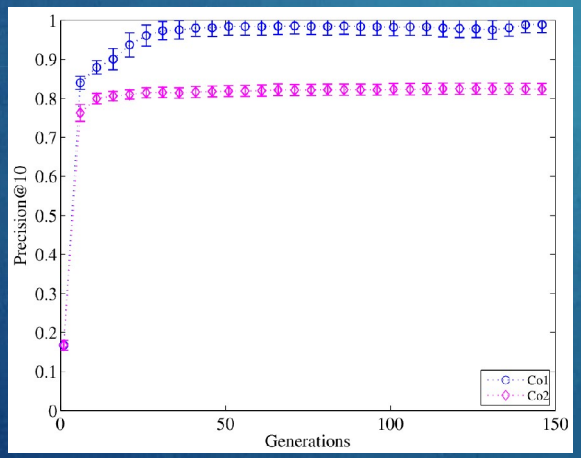
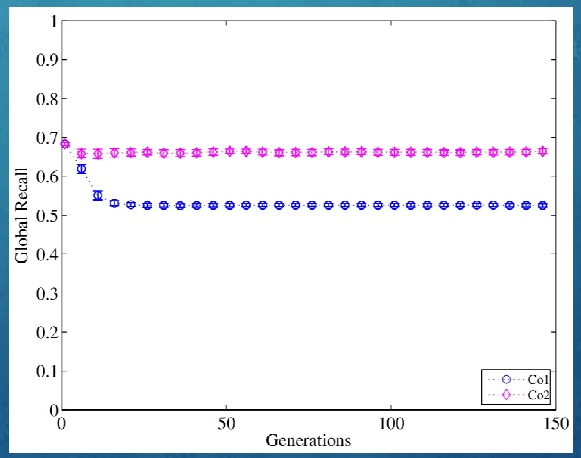
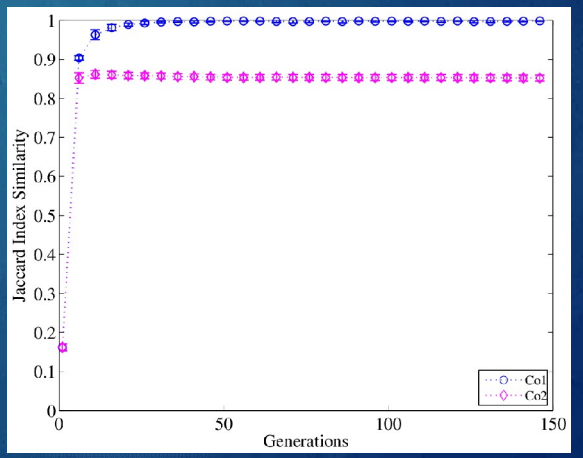
La similitud promedio de Jaccard se calcula promediando la similitud de Jaccard entre todos los pares de conjuntos de documentos recuperados, restringido solo a los documentos relevantes.

Corpus de Documentos y Diseño de Experimentos

- 448 tópicos del tercer nivel de DMOZ.
  - Reportamos resultados para los tópicos Bodypainting y Aquaculture.
- Más de 100 URLs por tópico.
- Más de 350.000 páginas.
- **Lucene** fue usado para indexar los documentos y para crear un motor de búsqueda.
- Cada tópico fue dividido en 2/3 para entrenamiento y 1/3 para testing.
- Para cada uno de las pruebas se realizaron 20 corridas con los siguientes parámetros genéticos:
  - n = 100, g = 150
  - Probabilidad de cruzamiento = 0,7
  - Probabilidad de mutación = 0,03
  - Longitud máxima inicial de las consultas = 32 palabras

Evaluación de la performance durante el entrenamiento

Evaluación de la performance durante el entrenamiento

	Evolución $Precision@10$	Evolución *Global Recall*	Evolución $Jaccard - Sim. - Index$
Tópico: *Body Painting*			
Tópico: *Aquaculture*			

Evaluación sobre un nuevo corpus

Evaluación sobre un nuevo corpus

- Cálculo de los I.C (al 95%) de las métricas de Precision@10, Global Recall y Mean Jaccard Index para cada tópico sobre el conjunto de test, usando las **poblaciones de consultas evolucionadas** en comparación con las **consultas obtenidas a partir de la descripción del tópico.**

TESTING C.I.				
Topics	Metrics	Co1 - Co2	Co1	Co2
		First Generation	Last Generation	Last Generation
BODY PAINTING	$\overline{Precision@10}$	[0.058, 0.075]	[0.432, 0.509]	[0.342, 0.395]
	Global Recall	[0.482, 0.482]	[0.362, 0.388]	[0.480, 0.525]
	$\overline{Jaccard\text{-}Similarity\text{-}Index}$	[0.237, 0.272]	[0.975, 0.999]	[0.580, 0.662]
AQUACULTURE	$\overline{Precision@10}$	[0.158, 0.186]	[0.773, 0.845]	[0.735, 0.775]
	Global Recall	[0.696, 0.708]	[0.511, 0.529]	[0.672, 0.686]
	$\overline{Jaccard\text{-}Similarity\text{-}Index}$	[0.191, 0.210]	[0.985, 0.997]	[0.822, 0.835]

Conclusiones y trabajo futuro

- Hemos propuesto una estrategia para evolucionar consultas temáticas con énfasis en preservar la diversidad.

Conclusiones y trabajo futuro

- Hemos propuesto una estrategia para evolucionar consultas temáticas con énfasis en preservar la diversidad.
- Formulamos una función de aptitud inspirada en la noción de entropía basada en teoría de la información que mejora la cobertura a nivel poblacional.

Conclusiones y trabajo futuro

- Hemos propuesto una estrategia para evolucionar consultas temáticas con énfasis en preservar la diversidad.
- Formulamos una función de aptitud inspirada en la noción de entropía basada en teoría de la información que mejora la cobertura a nivel poblacional.
- Esta función de aptitud es utilizada en combinación con precisión a 10, dando lugar a una estrategia MOEA que alcanza buena cobertura global con pérdida moderada de la precisión.

Conclusiones y trabajo futuro

- Hemos propuesto una estrategia para evolucionar consultas temáticas con énfasis en preservar la diversidad.
- Formulamos una función de aptitud inspirada en la noción de entropía basada en teoría de la información que mejora la cobertura a nivel poblacional.
- Esta función de aptitud es utilizada en combinación con precisión a 10, dando lugar a una estrategia MOEA que alcanza buena cobertura global con pérdida moderada de la precisión.
- Trabajo futuro:
  - Experimentos con ≠ valores para los parámetros del AE.
  - Método de selección.
  - Aplicación de Programación Genética.
  - Función de aptitud

¡Muchas Gracias!

CECILIA BAGGIO – ROCÍO L. CECCHINI

CARLOS M. LORENZETTI – ANA MAGUITMAN

{cb, rlc, cml, agm}@cs.uns.edu.ar

ICIC CONICET UNS

¡Muchas Gracias!

CECILIA BAGGIO – ROCÍO L. CECCHINI

CARLOS M. LORENZETTI – ANA MAGUITMAN

{cb, rlc, cml, agm}@cs.uns.edu.ar

ICIC CONICET UNS

Arquitectura propuesta