Searching the Web in Context: Genetic Algorithms for Exploring Query Space

Rocío L. Cecchini - Carlos M. Lorenzetti
Ana G. Maguitman - N. Beatriz Brignole

Bahía Blanca

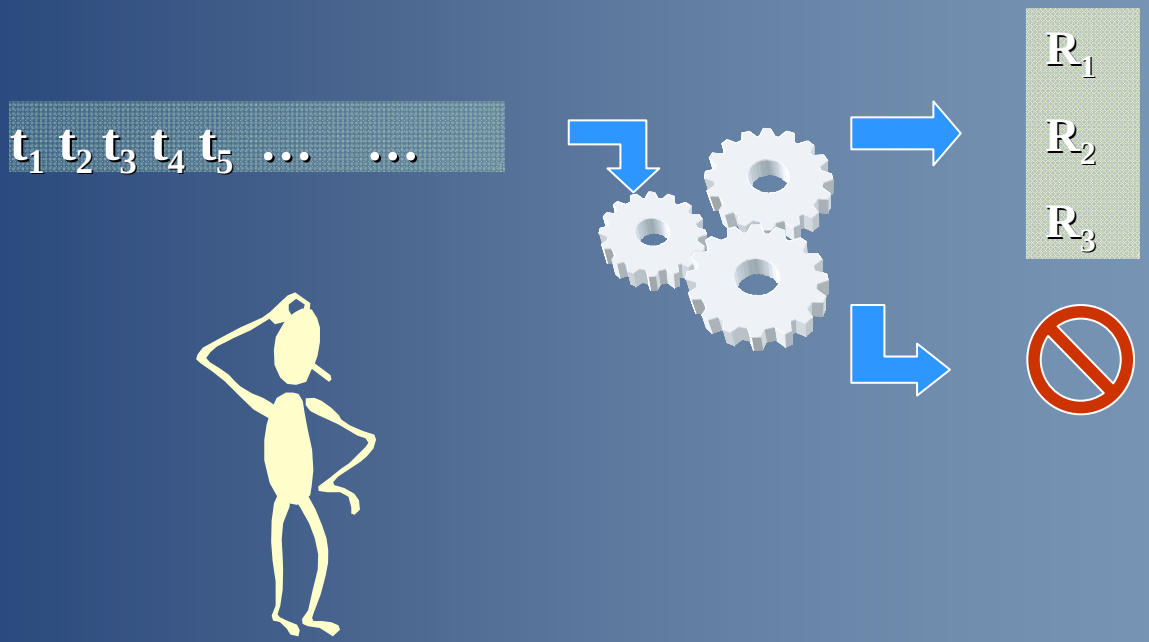
LIDeCC - *LIDIA* - Universidad Nacional del Sur

Planta Piloto de Ingeniería Química

Introducción

Objetivo del trabajo

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



Introducción

Objetivo del trabajo

Utilidad de generar buenas consultas

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

Técnica basada en AGs

Objetivo del trabajo

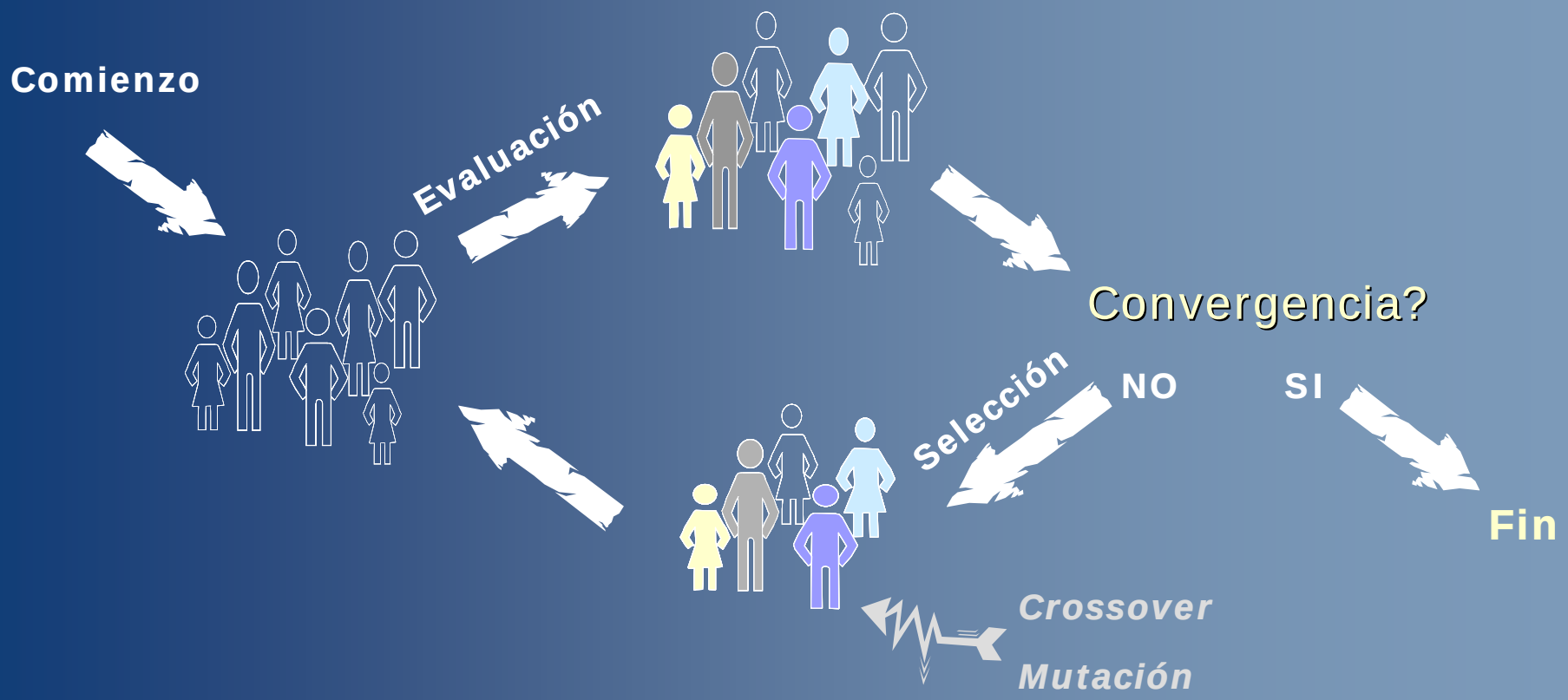
Algoritmos Genéticos - Revisión

Definición

- Los Algoritmos Genéticos son métodos estocásticos de búsqueda de soluciones cuasi-óptimas. En ellos se mantiene una población de potenciales soluciones, la cual es sometida a ciertas transformaciones con las que se trata de obtener nuevos candidatos. Además, esta población es sometida a un proceso de selección sesgado a favor de los mejores candidatos.

Algoritmos Genéticos - Revisión

Algoritmo



AGs para la Búsqueda web basada en contexto

Sistemas de información basados en contexto

- Los sistemas de información basados en contexto crean un modelo del contexto del usuario, infieren necesidades del usuario y buscan documentos relevantes en la web o en otras librerías electrónicas.

Porque AGs?

- Búsqueda web basada en contexto como problema de optimización.
  - Espacio de búsqueda?
  - Función objetivo?

AGs para la Búsqueda web basada en contexto

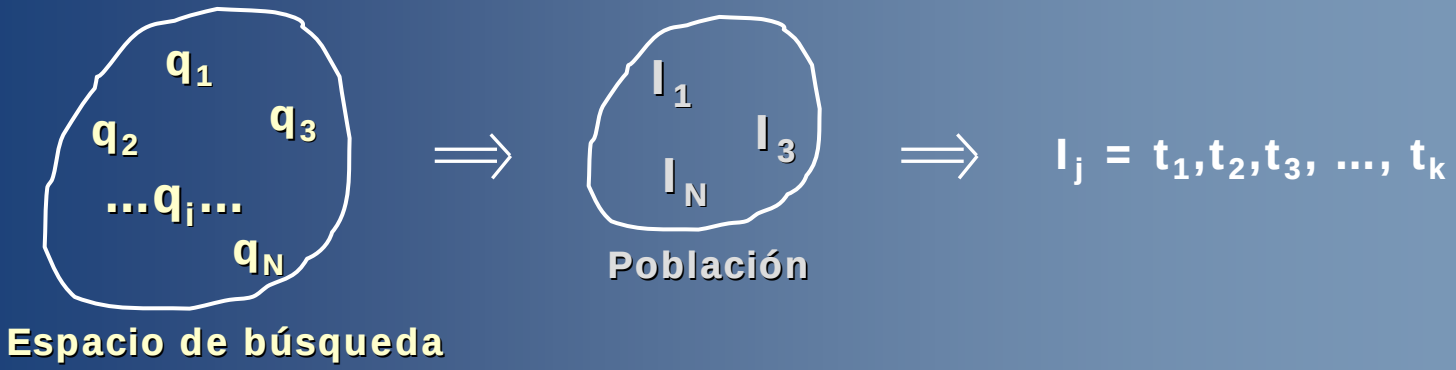
Porque AGs?

- Espacio multidimensional
- Soluciones subóptimas
- Múltiples soluciones. Podemos tener:
  - Resultados satisfactorios en distintos individuos
  - Interés en encontrar más de una consulta.
- Exploración
  - Crossover
  - Mutación
- Explotación
  - Selección

Esquema genético para obtener buenas consultas

Población

- Representación



Población

- Inicialización

N° de términos?

Esquema genético para obtener buenas consultas

Función de fitness

- Dado un espacio de búsqueda Q y un contexto temático inicial *c*, definimos la función de fitness *F* de la siguiente manera:

$$F(q) = \max_{d_i \in A_q}(\sigma(c, d_i))$$

Donde

- A_q es el conjunto de resultados obtenidos para la consulta q
- σ : D×D → [0…1] es la medida de similitud para un par de documentos.

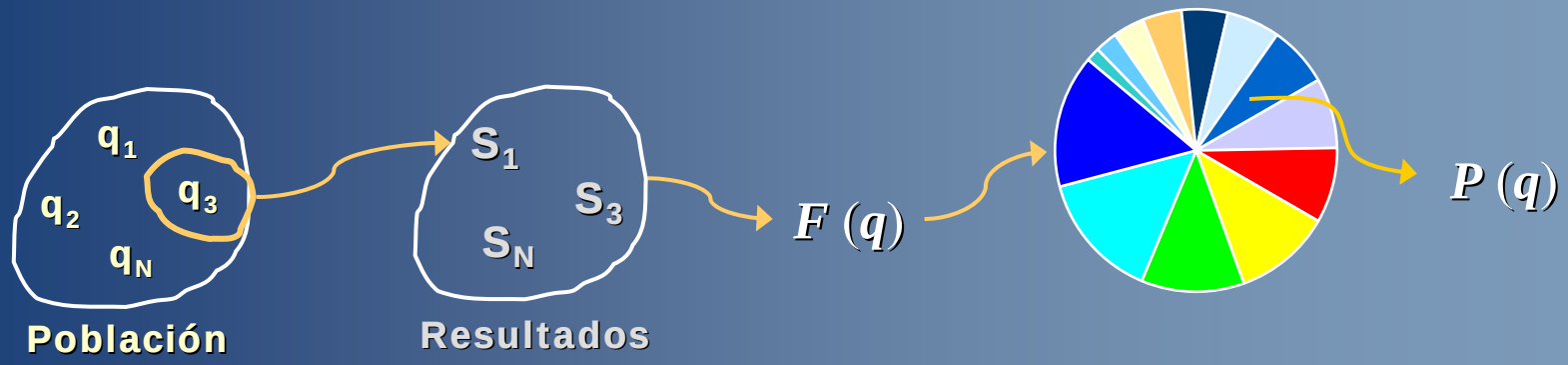
Obs:

- En σ puede usarse cualquier medida de similitud.
- En A_q no se usa el total de resultados obtenidos → 10 Snippets.

Esquema genético para obtener buenas consultas

Operadores genéticos

- Método de Selección

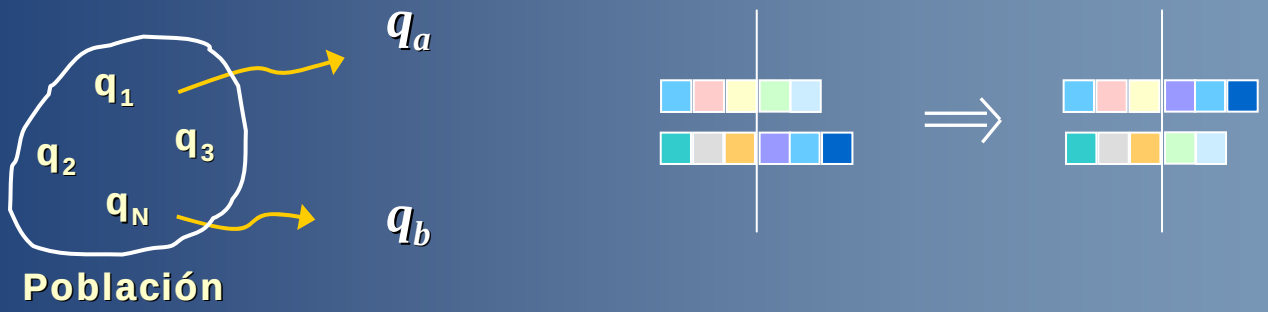


- Selección directa: un porcentaje de individuos de la población seleccionada, pasa directamente a formar parte de la nueva generación.

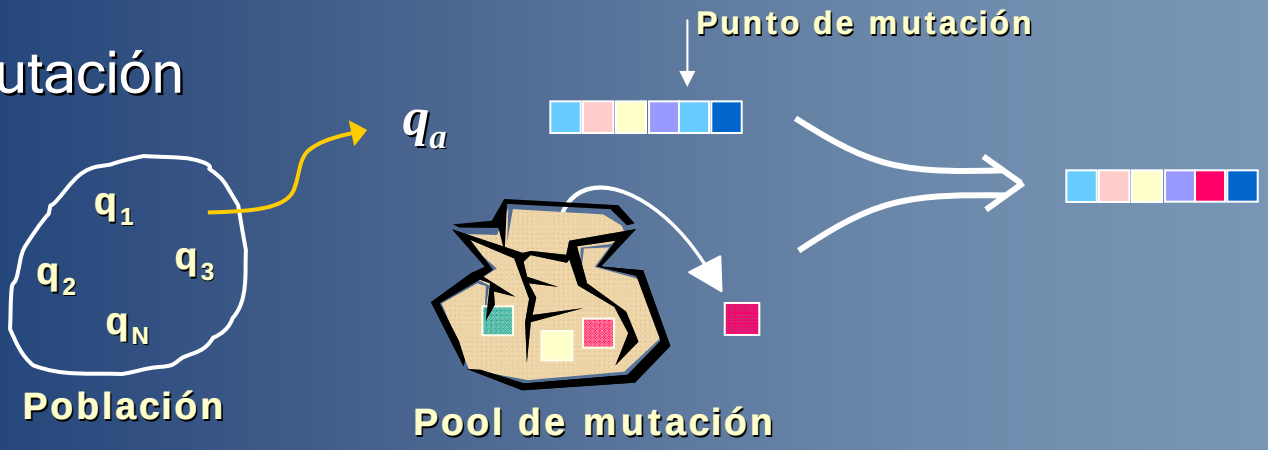
Esquema genético para obtener buenas consultas

Operadores Genéticos

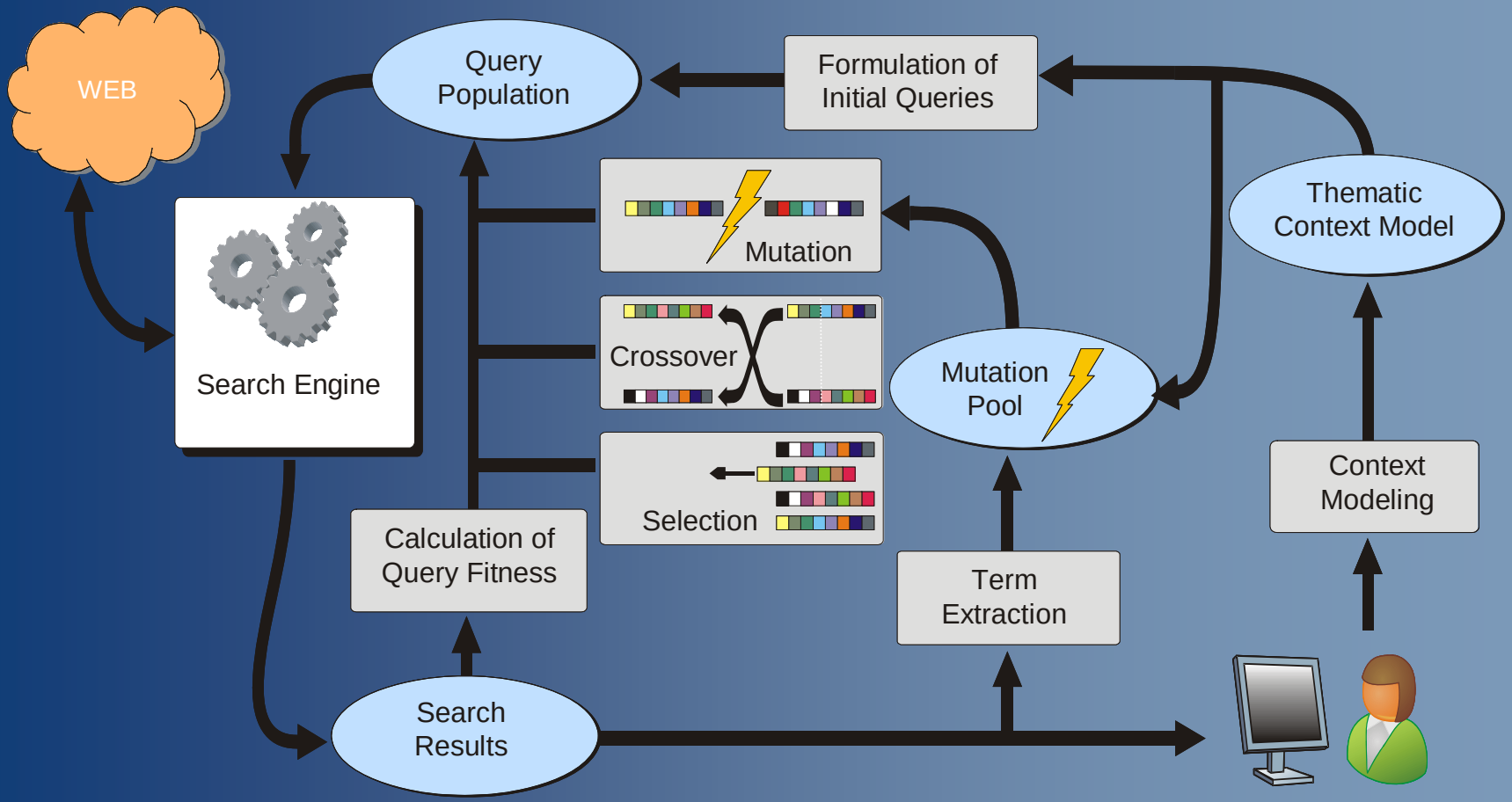
- Crossover



- Mutación



Arquitectura propuesta



Evaluación del funcionamiento

Criterio de evaluación

Definición:

- Dado un contexto temático *c*, una consulta *q* y el conjunto $A_q=\{a_1,\ldots,a_n\}$ de recursos recuperados para *q*. Una medida de similitud entre *c* y un recurso a_i puede ser computada usando *similitud por coseno* definida como:

$$\sigma(c,a_i)=\frac{\vec{c}\cdot\vec{a}_i}{\left\|\vec{c}\right\|\cdot\left\|\vec{a}_i\right\|}$$

Donde:

- $\vec{c}$ es la representación vectorial del contexto temático *c*
- $\vec{a}_i$ es la representación vectorial de a_i

Evaluación del funcionamiento

Calidad de *q* basada en similitud máxima:

$$Quality\_Max(q) = \max_{a_i \in A_q}(\sigma(c, a_i))$$

Calidad de *q* basada en similitud promedio:

$$Quality\_Mean(q) = \frac{\sum_{a_i \in A_q}(\sigma(c, a_i))}{|A_q|}$$

Evaluación del funcionamiento

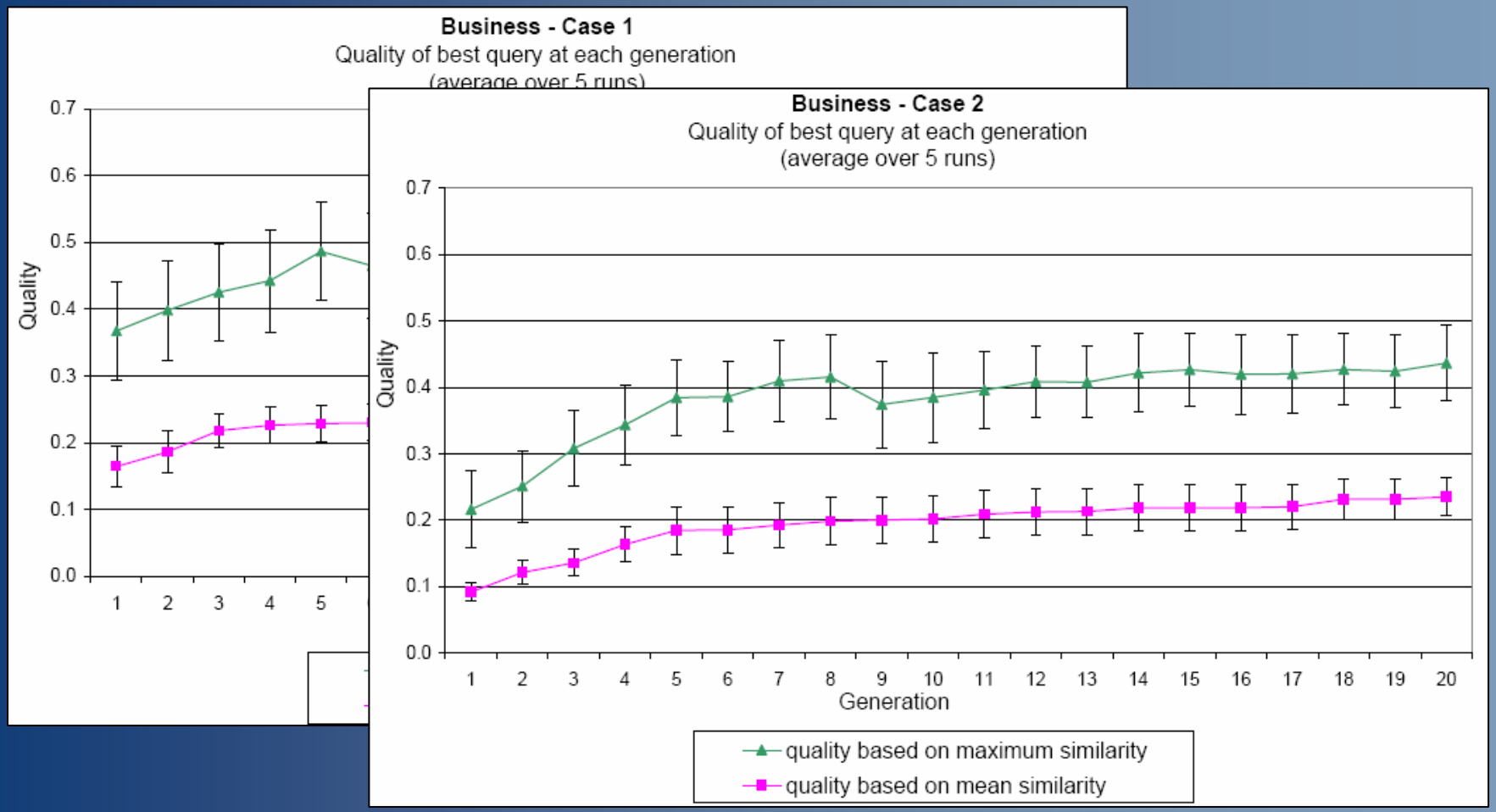
Diseño de experimentos

Para cada uno de los testeos se realizaron 5 corridas con los siguientes parámetros genéticos:

- n = 60
- g = 20
- Probabilidad de cruzamiento = 0,7
- Probabilidad de mutación = 0,03
- Longitud máxima inicial de las consultas = 32 palabras

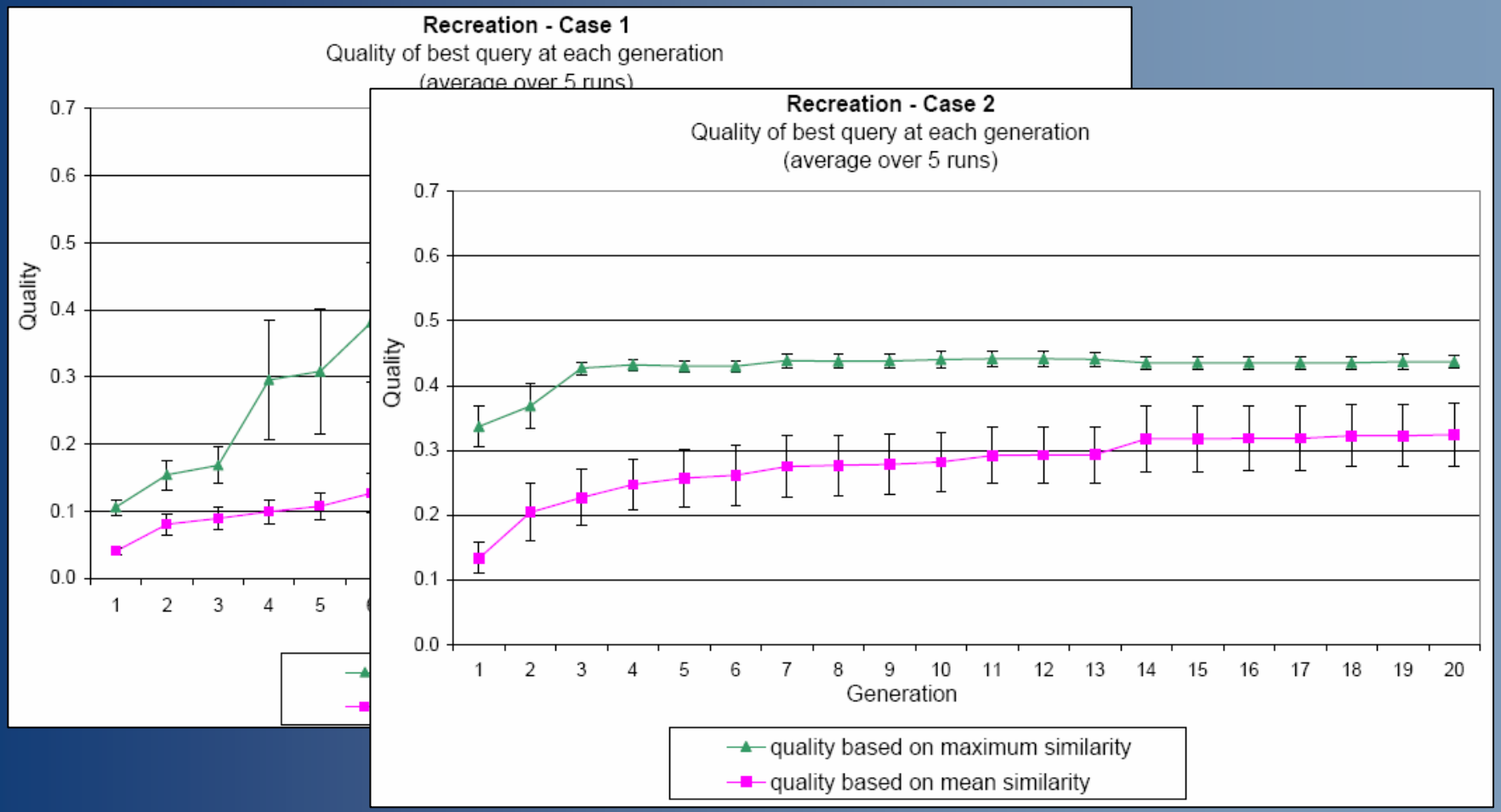
Evaluación del funcionamiento

Resultados



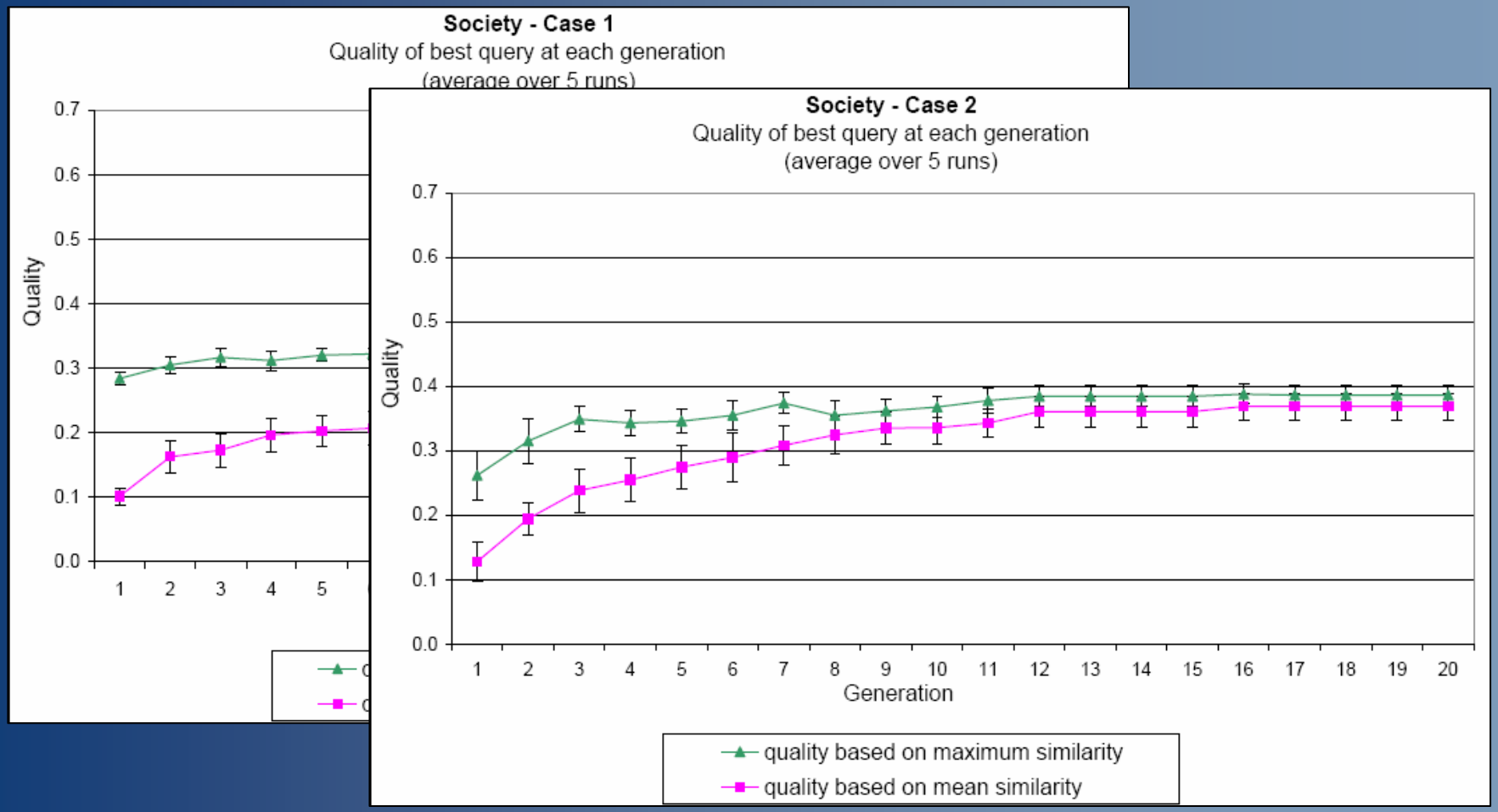
Evaluación del funcionamiento

Resultados



Evaluación del funcionamiento

Resultados



Conclusiones y trabajo futuro

Conclusiones

- Esta herramienta es aplicable a diferentes dominios en los cuales sea posible construir un modelo de términos a partir de un contexto.
- La adaptación de las consultas tiene un costo.
- Efectividad del AG en generación y refinamiento de consultas.

Trabajo futuro

- Experimentos con ≠ valores para los parámetros del AG.
- Método de selección.
- Aplicación de Programación Genética.
- Función de fitness

AGENCIA

CONICET

Searching the Web in Context: Genetic Algorithms for Exploring Query Space

Rocío L. Cecchini - Carlos M. Lorenzetti
Ana G. Maguitman - N. Beatriz Brignole

LIDeCC - *LIDIA* - Universidad Nacional del Sur

Planta Piloto de Ingeniería Química