AGENCIA

CONICET

# Genetic Algorithms for Topical Web Search: A Study of Different Mutation Rates

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

*LIDeCC* - *LIDIA* - Universidad Nacional del Sur

Planta Piloto de Ingeniería Química

# Búsquedas temáticas – ejemplos

- Búsqueda basada en la tarea del usuario

- Recolección de recursos web para portales temáticos

# Generación de las consultas

- Importancia de generar buenas consultas
- Generación de consultas como problema de optimización
  - Espacio de búsqueda
  - Función objetivo a optimizar
  - Medidas a emplear para la función objetivo

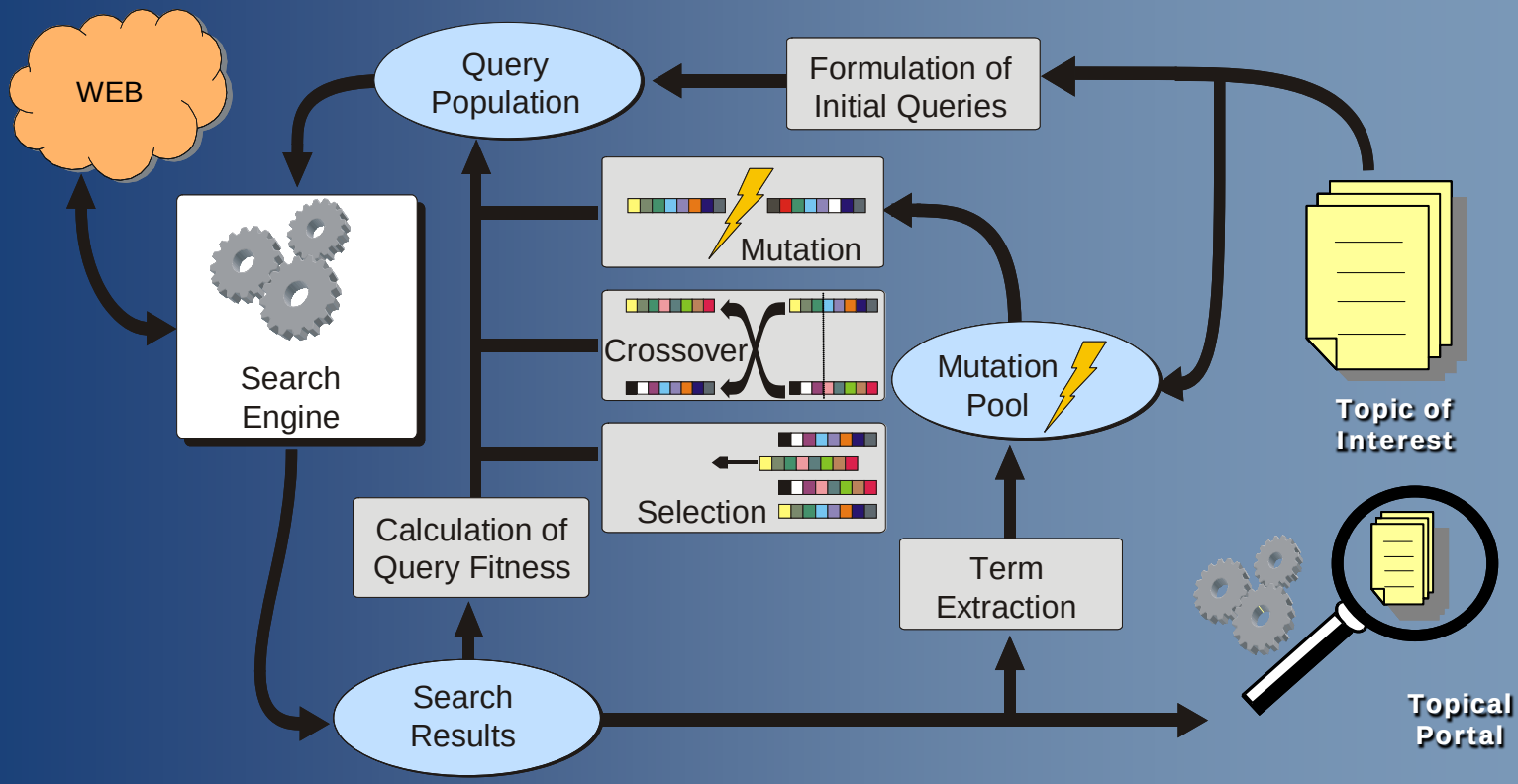
# Técnica basada en AGs

# AGs para explorar el espacio de úsqueda

## 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

# Arquitectura propuesta



# 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 $q$
- $\sigma$: D×D → [0…1] es la medida de similitud para un par de documentos.

Obs:

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

# 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:

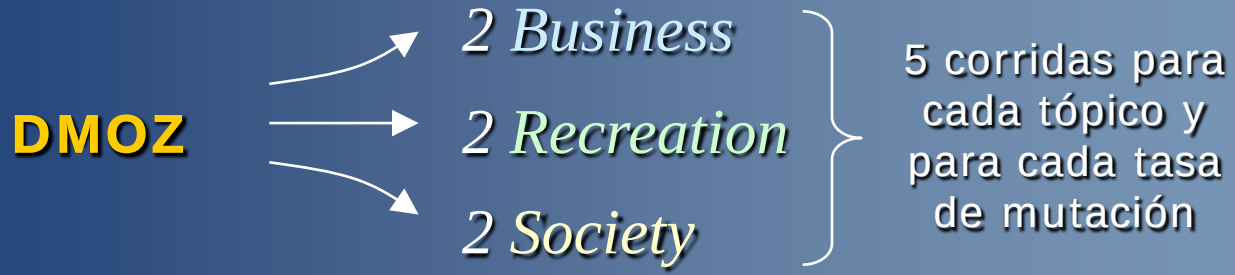
$$Quality\_Max(q) = \max_{a_i \in A_q}(\sigma(c, a_i))$$

Calidad de $q$ basada en similitud promedio:

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

# Efecto de aplicar diferentes tasas de mutación

## Diseño de experimentos

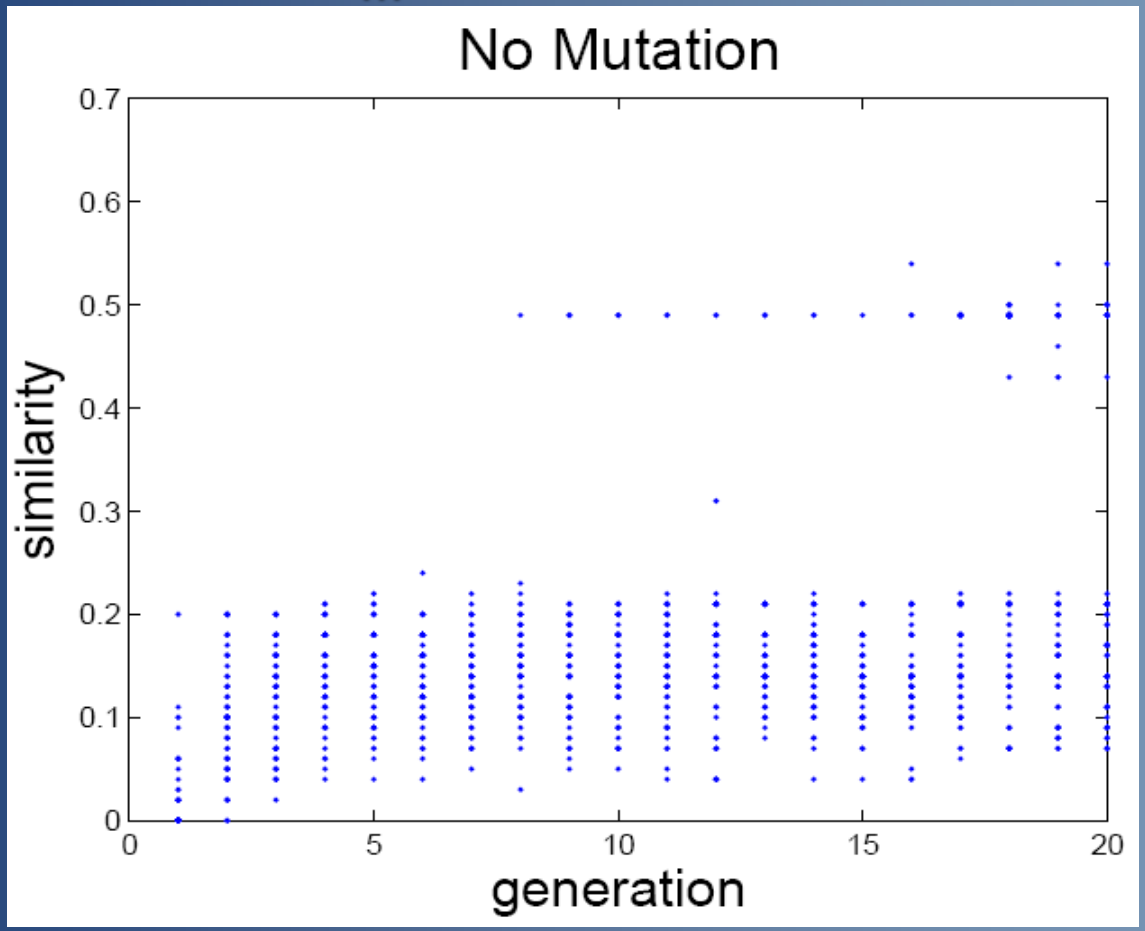



Para cada uno de los testeos se realizaron 5 corridas con los siguientes parámetros genéticos:

- n = 60 , g = 20, $P_c$ = 0,7
- Las diferentes tasas de mutación fueron:
  - $P_{m2}$ = 0 (Nula)
  - $P_{m1}$ = 0,03 (Normal),
  - $P_{m3}$ = 0,3 (Hipermitación)
- Con una longitud máxima inicial de las consultas de 32 palabras

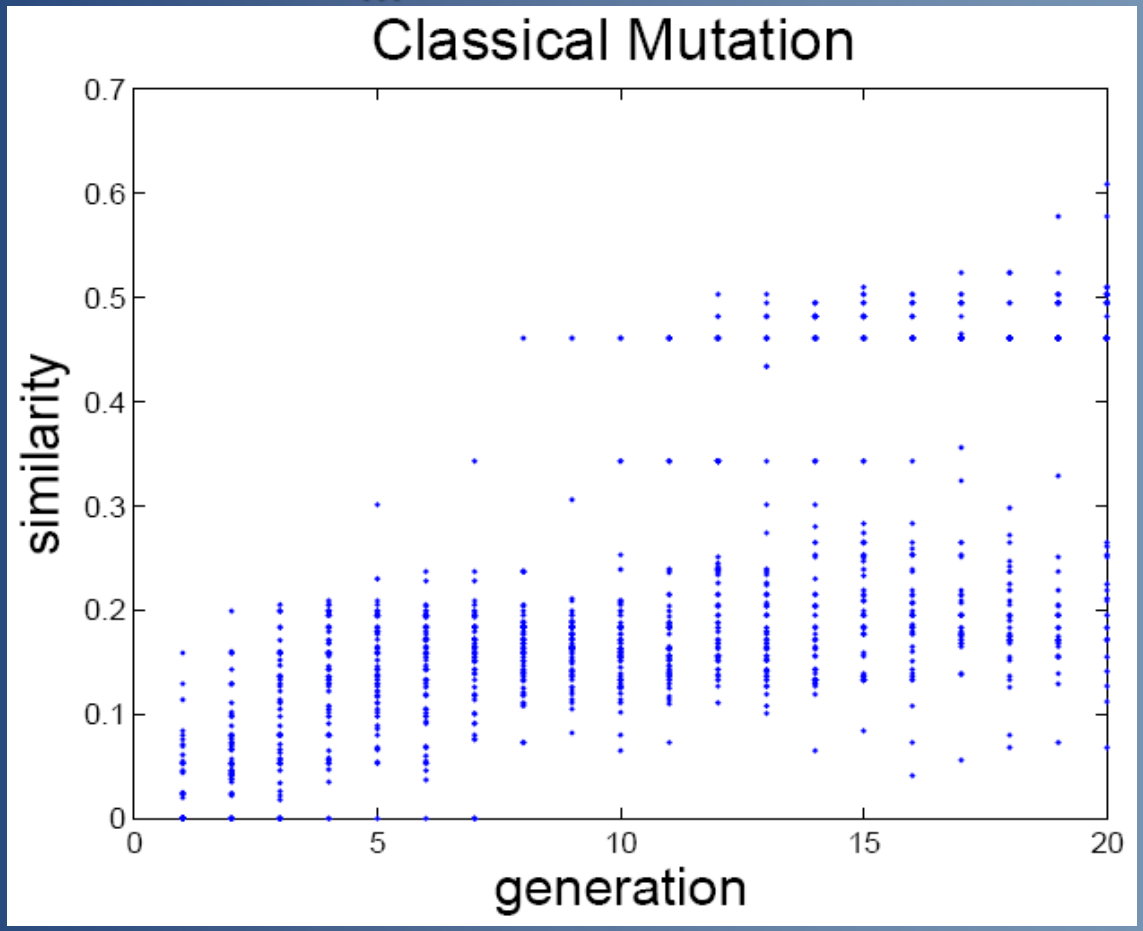
# Efecto de las diferentes tasas de mutación

## Resultados para $P_m = 0$



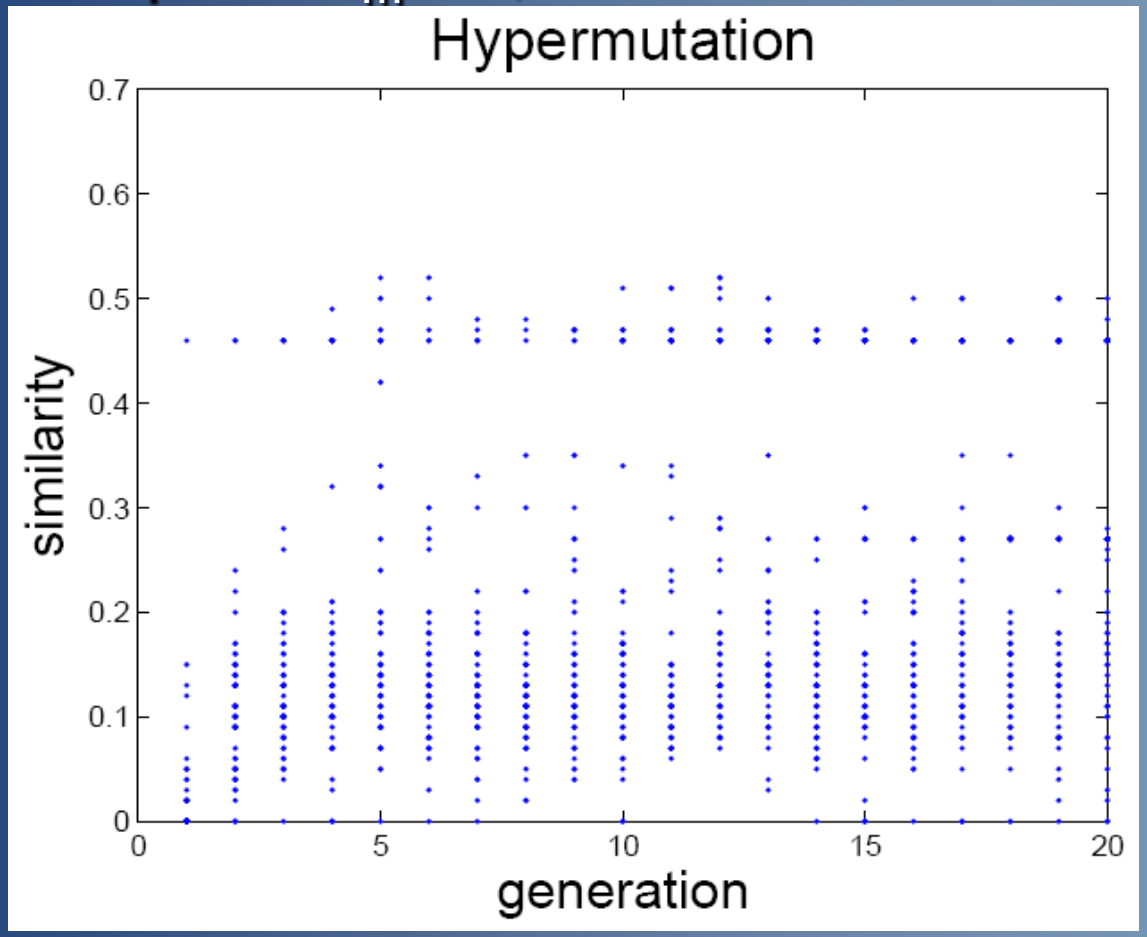
# Efecto de las diferentes tasas de mutación

## Resultados para $P_m$= 0,03

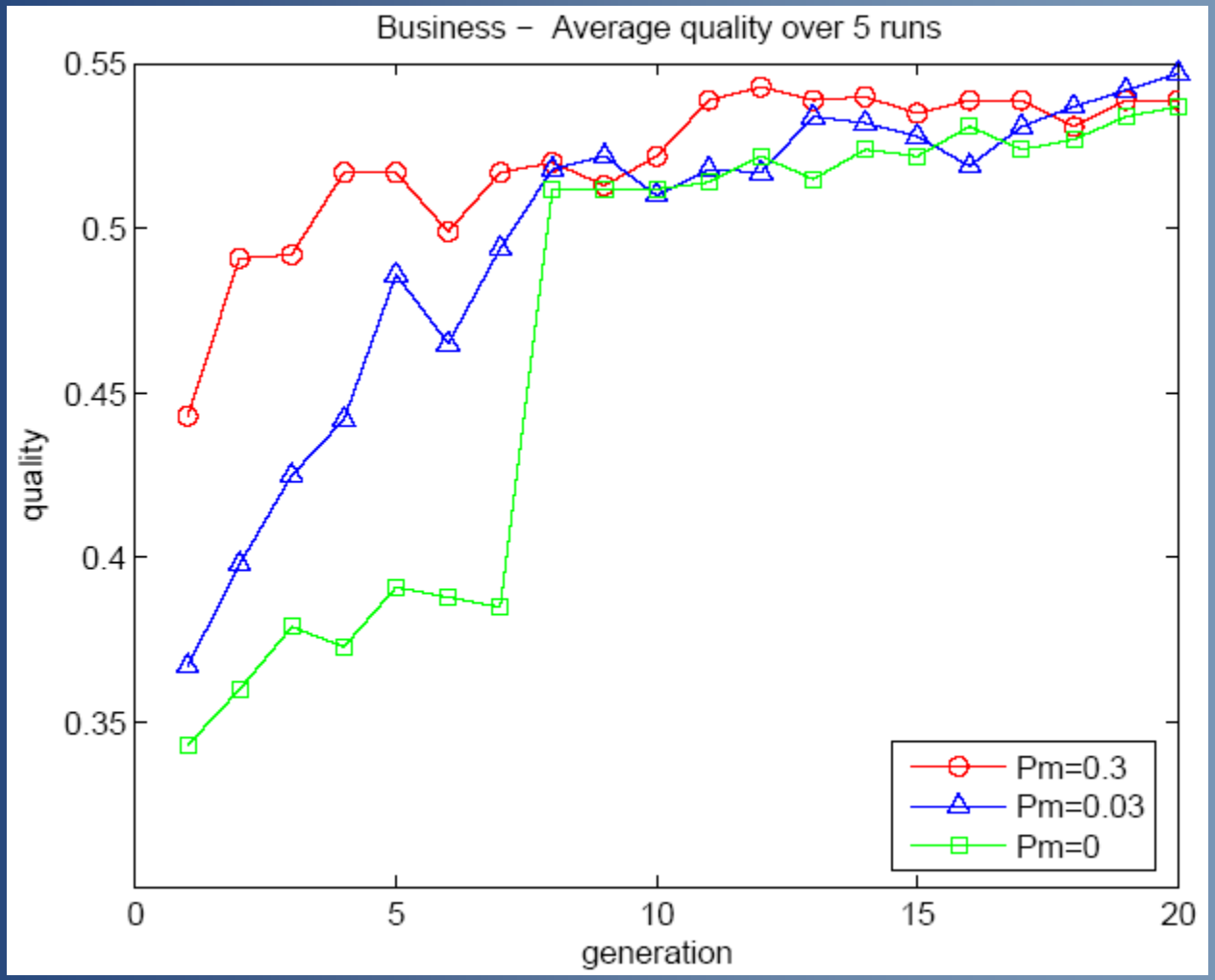


# Efecto de las diferentes tasas de mutación

## Resultados para $P_m$= 0,3

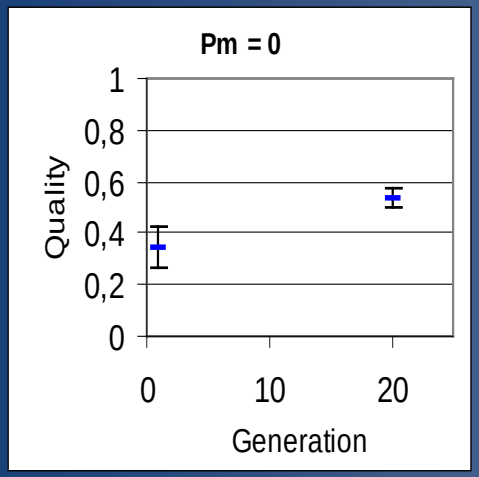


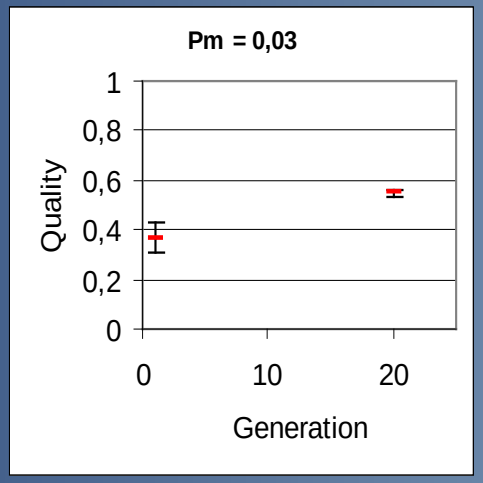
# Evaluación de la calidad de las consultas

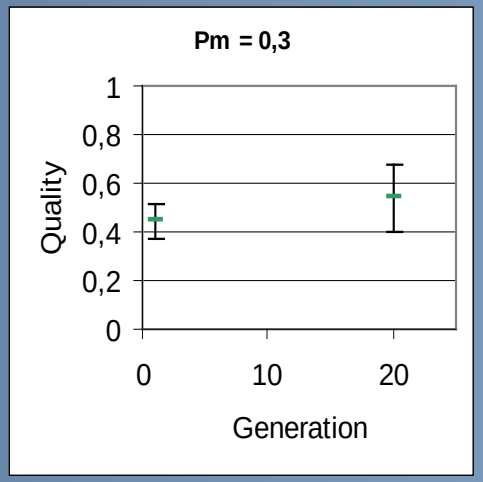


# Análisis estadístico

Intervalos de confianza para 95% de certeza para la media de la calidad para la primer y última generación




| MEAN 95% C.I. | MEAN 95% C.I | MEAN 95% C.I. |
|---|---|---|
| g=1 0.343 (0.264,0.421)<br>g=20 0.537 (0.500,0.574) | g=1 0.367 (0.305,0.429)<br>g=20 0.547 (0.530,0.564) | g=1 0.443 (0.375,0.511)<br>g=20 0.539 (0.404,0.673) |
| Pm = 0 | Pm = 0.03 | Pm = 0.3 |

# Conclusiones y trabajo futuro

## Conclusiones

- Hemos estudiado el efecto de aplicar diferentes tasas de mutación y se ha comprobado que a mayores tasas se amplia notablemente el rango de exploración del espacio de búsqueda.
- Esta herramienta es aplicable a diferentes dominios en los cuales sea posible construir un modelo de términos a partir de un tópico.
- 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.
- Aplicar otros métodos de selección.
- Implementar un esquema elitista.
- 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