Algoritmos Genéticos para la Búsqueda Web basada en Contextos Temáticos*

Rocío L. Cecchini† **Carlos M. Lorenzetti**‡ **Ana G. Maguitman**‡

† LIDeCC - Laboratorio de Investigación y Desarrollo en Computación Científica
‡ LIDIA - Laboratorio de Investigación y Desarrollo en Inteligencia Artificial
Departamento de Ciencias e Ingeniería de la Computación
Universidad Nacional del Sur, Av. Alem 1253, (8000) Bahía Blanca, Argentina
phone: 54-291-4595135 fax: 54-291-4595136
e-mail: {cr,cml,agm}@cs.uns.edu.ar

1. INTRODUCCIÓN

El uso de contextos temáticos para seleccionar y filtrar información juega un papel fundamental en los sistemas de recuperación de información basados en la tarea del usuario (e.g., [3, 8]). Desafortunadamente, aprovechar la información del contexto durante la búsqueda en la Web es una tarea difícil. Los buscadores actuales imponen un límite a la longitud de las consultas, y aún si se permitieran consultas largas las mismas podrían volverse demasiado específicas, devolviendo muy pocos o ningún resultado. Esto dificulta la tarea de formular consultas adecuadas para describir contextos temáticos. Una alternativa para evitar este problema es el uso de ciertas sintaxis especiales provistas por algunos buscadores para la formulación de consultas. Sin embargo, aún con la flexibilidad provista por estos mecanismos de formulación de consultas, es posible que el vocabulario utilizado para describir el contexto difiera del usado para indexar los recursos relevantes. La meta de nuestro trabajo de investigación es desarrollar técnicas para refinar las consultas automáticamente y recolectar recursos relevantes para el contexto temático del usuario.

En este trabajo proponemos utilizar Algoritmos Genéticos (AGs) para abordar el problema de reflejar contextos temáticos en las consultas formuladas a un buscador Web. Nuestra propuesta se basa en nuevas técnicas incrementales que permiten evolucionar consultas útiles ligadas a un contexto temático bajo análisis.

1.1. Algoritmos Genéticos

Los AGs [7] son técnicas de optimización robustas basadas en el principio de *selección natural y supervivencia del más apto* que establece que "en cada generación los individuos más fuertes tienden a sobrevivir y los más débiles suelen perecer". Por lo tanto, cada nueva generación dará lugar a individuos más fuertes en comparación a sus ancestros.

Para utilizar AGs en problemas de optimización es necesario codificar la información de las soluciones posibles mediante cromosomas (compuestos por genes) y definir una función de aptitud a ser maximizada. El algoritmo mantiene una población de soluciones candidatas. Esta población evoluciona a través de iteraciones, llamadas generaciones. Las nuevas generaciones se forman usando los operadores genéticos de selección, cruzamiento y mutación. Los cromosomas padres son seleccionados para producir descendientes, favoreciendo a aquellos padres con mejor aptitud. El cruzamiento

*Este trabajo de investigación es financiado por la Agencia Nacional de Promoción Científica y Tecnológica (PICT 2005 Nro. 32373), la Universidad Nacional del Sur y el Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET).

consiste en combinar dos cromosomas padres dando como resultado dos cromosomas descendientes. Por otra parte, la mutación origina perturbaciones aleatorias de los cromosomas, por ejemplo, reemplazando un gen por otro. El siguiente algoritmo muestra los pasos generales de un AG:

Paso 1: Comenzar con una población de soluciones generadas aleatoriamente.
Paso 2: Evaluar la aptitud de cada individuo en la población.
Paso 3: Seleccionar probabilísticamente individuos a reproducirse favoreciendo a los más aptos.
Paso 4: Aplicar cruzamiento con probabilidad *Pc*.
Paso 5: Aplicar mutación con probabilidad *Pm*.
Paso 6: Reemplazar la población por la nueva generación de individuos.
Paso 7: Ir al paso 2.

Si bien los operadores de selección, cruzamiento y mutación pueden ser implementados de diferentes maneras, su propósito fundamental es explorar el espacio de soluciones candidatas, perfeccionando la población en cada generación mediante el agregado de descendientes mejorados y la eliminación de los peores individuos.

1.2. AGs para la Búsqueda Temática en la Web

Los sistemas de recuperación de información basados en contextos temáticos monitorean al usuario, infieren sus necesidades de información y buscan recursos relevantes en la Web o en otras fuentes electrónicas. Tradicionalmente, tales sistemas encuentran documentos relevantes extendiendo las consultas del usuario con palabras adicionales extraídas del contexto o formulando consultas automáticamente. Existen varios sistemas que siguen este enfoque y que han alcanzado resultados prometedores (e.g. [3, 9]). Por otra parte, los AGs han sido parte de diversas propuestas dentro del área de recuperación de la información. Entre ellas cabe destacarse la aplicación de técnicas de AGs para derivar mejores descripciones de documentos [6] y la optimización de consultas utilizando AGs para ponderar palabras [12, 2].

Podemos mencionar varias razones por las cuales los AGs son apropiados para abordar el problema de la búsqueda Web basada en contextos temáticos:

- **Búsqueda temática como un problema de optimización.** El problema de formular buenas consultas para la búsqueda Web basada en contextos temáticos puede ser planteado como un problema de optimización. El conjunto de búsqueda de este problema está dado por el conjunto de posibles consultas que pueden ser presentadas a un buscador. La función objetivo a ser optimizada toma como argumento una consulta y se define como la efectividad que posee dicha consulta para la recuperación de material relevante cuando la misma es presentada a un buscador.
- **Espacio de búsqueda con un gran número de dimensiones.** El espacio de consultas candidatas posee un gran número de dimensiones, donde cada palabra posible da lugar a una nueva dimensión. Esta clase de problemas no son fáciles de resolver mediante métodos analíticos pero pueden ser abordados con éxito mediante AGs.
- **Soluciones subóptimas aceptables.** En la búsqueda Web existen numerosos resultados subóptimos que vale la pena explorar y por tal motivo este tipo de búsqueda admite la formulación de consultas que no son óptimas. Generalmente los AGs no garantizan la identificación de soluciones óptimas pero son altamente efectivos cuando se trata de encontrar aquellas soluciones cercanas a las óptimas.

- **Soluciones múltiples.** Múltiples conjuntos de páginas Web pueden considerarse resultados satisfactorios para una búsqueda basada en un contexto temático. Por tal motivo, podría interesarnos formular varias en lugar de un única consulta. Los AGs pueden ser utilizados naturalmente para problemas de optimización multimodal, devolviendo múltiples soluciones globales.

2. NUESTRA PROPUESTA

Nuestro objetivo es evolucionar poblaciones de consultas que tengan la capacidad de recuperar material relevante para el contexto temático del usuario. Con el fin de alcanzar esta meta, comenzamos con una población de consultas compuestas por palabras extraídas directamente del contexto temático inicial y evaluamos estas consultas de acuerdo a la calidad de los resultados recuperados a partir de cada una de ellas. A medida que las generaciones avanzan, predominarán las consultas asociadas a los mejores resultados. Además, los operadores genéticos combinan y alteran continuamente estas consultas de maneras novedosas, generando soluciones cada vez más refinadas. La figura 1 esquematiza nuestra propuesta.

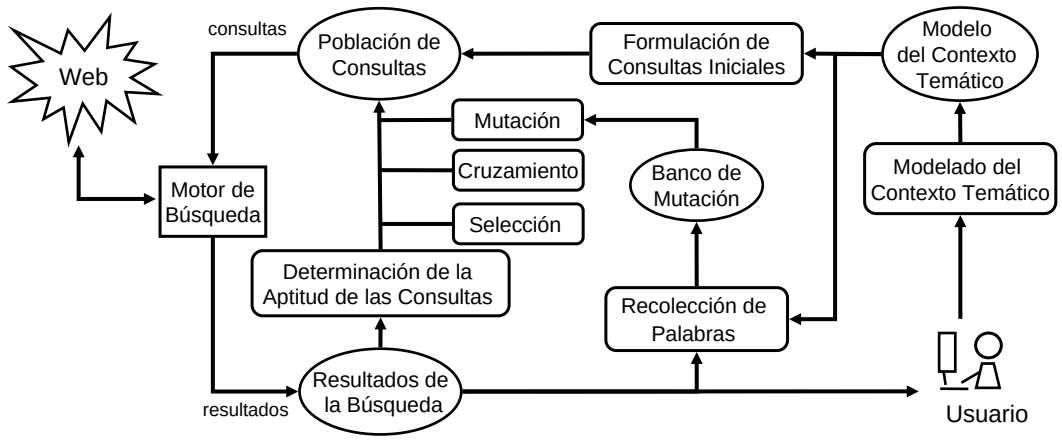

Figura 1: Arquitectura de un sistema basado en AGs para la búsqueda temática en la Web.

2.1. Población y Representación de Cromosomas

El espacio de búsqueda está constituido por todas las posibles consultas que un usuario puede ingresar en un buscador. La poblacón, por lo tanto, será un subconjunto de dichas consultas. En consecuencia, cada cromosoma se representa como una secuencia de palabras, donde cada palabra corresponde a un gen que puede ser manipulado por los operadores genéticos. La población es inicializada utilizando un número fijo de consultas generadas a partir de palabras seleccionadas aleatoriamente del contexto temático original. Si bien todas las palabras que conforman la población inicial de consultas provienen del contexto temático original, varias palabras novedosas serán incluidas en lo sucesivo como resultado de la mutación.

2.2. Función de Aptitud

Asociamos al espacio de búsqueda Q una función de aptitud F : $Q \rightarrow [0 \ldots 1]$ que puede evaluar numéricamente a las consultas individuales. La función de aptitud define el criterio con el cual se determina la calidad de una consulta. Nuestro concepto de consulta de alta calidad está basado en la capacidad de la consulta para devolver material similar al contexto temático c cuando se la envía al

buscador. La función de aptitud propuesta es

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

donde $\mathbf{A_q}$ es el conjunto de respuesta para la consulta **q** (el conjunto de documentos devueltos por el buscador cuando **q** se usa como consulta) y $\sigma : D \times D \rightarrow [0 \ldots 1]$ es la medida de similitud entre un par de documentos (nótese que el contexto c puede ser considerado uno de los documentos en D).

Medidas de similitud distintas, tales como la similitud por el coseno o la similitud de Jaccard [1], pueden ser usadas en la implementación de la función de aptitud. Una dificultad pragmática que encontramos al intentar implementar la función de aptitud es el uso del conjunto completo $\mathbf{A_q}$. Utilizar la totalidad de las páginas devueltas por un buscador es demasiado costoso para fines prácticos. Por tal motivo, nos limitamos a los diez primeros resultados y sólo consideramos los "snippets" devueltos por Google para calcular la medida de similitud (el snippet devuelto por Google es un extracto de la página resumiendo el contexto en el cual aparecen las palabras de la consulta).

2.3. Operadores Genéticos

Una nueva generación en nuestro AG se obtiene tras aplicar probabilísticamente los operadores de selección, cruzamiento y mutación sobre las consultas de la población actual:

- **Selección.** Una nueva población es generada al seleccionar probabilísticamente las consultas de mayor calidad. La probabilidad de que una consulta **q** sea seleccionada es proporcional a su propia su aptitud $\mathrm{F}(\mathbf{q})$ e inversamente proporcional a la aptitud de las otras consultas en la población actual.
- **Cruzamiento.** Algunas de las consultas seleccionadas son incluidas en la siguiente generación tal como son, mientras que otras son cruzadas para crear nuevas consultas. El cruzamiento de un par de consultas se lleva a cabo copiando palabras de cada uno de los padres en los descendientes. En nuestra propuesta utilizamos el operador de cruzamiento "en un punto".
- **Mutación.** Los pequeños cambios aleatorios resultantes de aplicar el operador de mutación sobre las consultas consisten en reemplazar una palabra t_i^q seleccionada al azar por otra palabra t_j^p . La palabra t_j^p se obtiene del *banco de mutación* que describimos a continuación.

2.4. Banco de Mutación

El banco de mutación es un conjunto de palabras compuesto inicialmente por palabras provenientes del contexto temático original. A medida que el sistema recupera resultados relevantes de la Web, las palabras que aparecen en los snippets devueltos por el buscador se irán agregando al banco de mutación. Este procedimiento da al AG la posibilidad de generar consultas con palabras nuevas, permitiendo así una exploración más amplia del espacio de búsqueda.

3. CONCLUSIONES Y LÍNEAS DE INVESTIGACIÓN FUTURAS

Las técnicas aquí propuestas pueden ser utilizadas en la implementación de diferentes aplicaciones para la recolección de material basado en contextos temáticos. Por ejemplo, las técnicas de refinamiento de consultas aquí propuestas pueden utilizarse para facilitar el acceso a material temático generado dinámicamente proveniente de lo que se conoce como Web invisible [11]. Otra aplicación con gran potencial es la creación de portales verticales o índices temáticos. Los métodos para evolucionar consultas presentados en este trabajo pueden considerarse una técnica alternativa o complementaria a los crawlers temáticos [5, 10] para la recolección de material sobre un tópico específico.

Tras la implementación y análisis inicial de los métodos aquí descriptos hemos observado que nuestra propuesta es muy promisoria. En una serie de pruebas preliminares pudimos notar que la calidad de las consultas evoluciona considerablemente a través de las generaciones sucesivas. Como parte de nuestra tarea de investigación futura esperamos estudiar variantes de las técnicas propuestas. Entre estas, nos interesa implementar diferentes operaciones de selección, cruzamiento y mutación, como así también analizar el impacto de diversos parámetros (probabilidad de cruzamiento y mutación, tamaño de la población, etc.) sobre la eficacia de las técnicas bajo estudio. Otra línea de investigación futura consistirá en el uso de la *programación genética* para evolucionar consultas con sintaxis especiales [4]. En tales métodos no sólo nos concentraremos en seleccionar buenas palabras para formular consultas sino que también se intentará evolucionar consultas que incluyan operadores booleanos y otros comandos especiales.

REFERENCIAS

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

[2] M. Boughanem, C. Chrisment, and L. Tamine. On using genetic algorithms for multimodal relevance optimization in information retrieval. *J. Am. Soc. Inf. Sci. Technol.*, 53(11):934–942, 2002.

[3] Jay Budzik, Kristian J. Hammond, and Larry Birnbaum. Information access in context. *Knowledge based systems*, 14(1–2):37–53, 2001.

[4] Tara Calishain and Rael Dornfest. *Google Hacks. 100 Industrial-Strengths Tips and Tools*. O'Reilly, 2003.

[5] Soumen Chakrabarti, Martin van den Berg, and Byron Dom. Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks (Amsterdam, Netherlands: 1999)*, 31(11–16):1623–1640, 1999. 1999a.

[6] M. Gordon. Probabilistic and genetic algorithms in document retrieval. *Commun. ACM*, 31(10):1208–1218, 1988.

[7] John H. Holland. *Adaptation in natural and artificial systems*. Ann Arbor: The University of Michigan Press, 1975.

[8] David B. Leake, Travis Bauer, Ana Maguitman, and David C. Wilson. Capture, storage and reuse of lessons about information resources: Supporting task-based information search. In *Proceedings of the AAAI-00 Workshop on Intelligent Lessons Learned Systems. Austin, Texas*, pages 33–37. AAAI Press, 2000.

[9] Ana Maguitman, David Leake, and Thomas Reichherzer. Suggesting novel but related topics: towards context-based support for knowledge model extension. In *Proceedings of the 10th international conference on Intelligent user interfaces*, pages 207–214, New York, NY, USA, 2005. ACM Press.

[10] Filippo Menczer, Gautam Pant, and Padmini Srinivasan. Topical web crawlers: Evaluating adaptive algorithms. *ACM Trans. Inter. Tech.*, 4(4):378–419, 2004.

[11] Alexandros Ntoulas, Petros Zerfos, and Junghoo Cho. Downloading textual hidden web content through keyword queries. In *JCDL '05: Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pages 100–109, New York, NY, USA, 2005. ACM Press.

[12] Jing-Jye Yang and Robert Korfhage. Query optimization in information retrieval using genetic algorithms. In *Proceedings of the 5th International Conference on Genetic Algorithms*, pages 603–613, San Francisco, CA, USA, 1993. Morgan Kaufmann Publishers Inc.