Técnicas Evolutivas para la Extracción Automática de Conocimiento

Cecilia Baggio[†] **Rocío L. Cecchini**[‡] **Ana G. Maguitman**[†]

[†] LIDIA - Laboratorio de Investigación y Desarrollo en Inteligencia Artificial

[‡] LIDeCC - Laboratorio de Investigación y Desarrollo en Computación Científica

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: ceciliabgg@gmail.com,{rlc,agm}@cs.uns.edu.ar

1. RESUMEN

Esta línea de investigación propone el diseño, desarrollo y evaluación de técnicas automáticas para extracción de conocimiento, de tal forma que sean capaces de sobrellevar la búsqueda dentro de grandes espacios de información. Para ello se propone, en primera instancia, la resolución de un problema de interés general: el de reformulación automática de consultas. Una resolución automática para este problema podría ser utilizada en diversas aplicaciones, tales como monitorear un tópico de interés, especificar trackers temáticos sobre redes sociales, identificar entidades y relaciones entre entidades en grandes corpus de documentos o recolectar material para portales temáticos. Por sus características (alta dimensionalidad del espacio de búsqueda, carencia de subestructura óptima, posibilidad de aprovechamiento de múltiples soluciones) el uso de computación evolutiva parece adecuado para abordar su resolución. Un primer aporte de esta línea dentro del área radica en la consideración de la incorporación de operadores booleanos y otro tipo de modificadores a las consultas reformuladas y el control de la diversidad, ambos pensados como un mecanismo para lograr mayor expresión en las consultas y, por lo tanto, mayor poder para expresar los conceptos de interás involucrados. El segundo aporte consiste en proponer un marco de evaluación adecuado para la metodología desarrollada y el estudio y comparación con otras técnicas. Por útimo, el aporte final aborda la aplicación de los métodos desarrollados en dominios específicos tales como bioinformática (e.g. para identificación de interacciones entre entidades biológicas) o redes sociales (e.g. para realizar minería de opiniones mediante trackers temáticos).

Palabras clave: Extracción de Conocimiento, Computación Evolutiva, Minería de Datos y Texto.

2. CONTEXTO

Este trabajo de investigación será financiado por la Universidad Nacional del Sur en el marco del proyecto *Diseño y Evaluación de Mecanismos de Búsqueda Contextualizada en Sistemas Centralizados y Distribuidos* (Código: 24/N029) y por el Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) en el marco del proyecto *Diseño y Evaluación de Herramientas Inteligentes para el Acceso Contextualizado a Recursos Digitales en Entornos Distribuídos.*

3. INTRODUCCIÓN

En los últimos años ha habido un crecimiento enorme en la cantidad de información disponible, ya sea en forma de documentos, opiniones o datos numéricos. Por ejemplo, con respecto a los documentos en [13] se hizo un estudio en el que se muestra el crecimiento especialmente veloz de la información científica. Por otro lado, las redes sociales reportan estadísticas de crecimiento de información abrumadoras: 460000 usuarios nuevos en promedio por día en Twitter [1] o 4.75 millones de items de información compartidos diariamente en Facebook [2]. Estos y otros ejemplos conducen al interés y la necesidad de proponer métodos automáticos que sean capaces de manipular y

analizar dicha información para poder extraer conocimiento. En este sentido, los buscadores web, son una herramienta muy útil ya que permiten recuperar documentos relevantes basándose únicamente en un conjunto limitado de palabras brindado por el usuario. Dado que los búscadores son el único punto de acceso a la mayoría de la información distribuída por la red, una tarea interesante consiste en aprender de manera automática, cuáles son los términos más relevantes que logran ser buenos descriptores o buenos discriminadores para recursos pertenecientes a cierto tópico o contexto de interés. No sólo es de gran importancia identificar dichos términos, sino también encontrar una sintaxis apropiada para combinarlos. Otra tarea interesante consiste en inferir entidades participantes dentro de cierto tópico de interés así como las relaciones existentes entre dichas entidades. La cantidad de datos y artículos científicos disponibles en diversas áreas, como por ejemplo en bioinformática, permite suponer que gran cantidad de conocimiento se encuentra oculto esperando a ser descubierto.

Si fuese posible especificar las necesidades de información mediante consultas de alta calidad o mediante frases que involucren los términos y operadores apropiados, dichas especificaciones podrían ser aprovechadas en una amplia variedad de aplicaciones, tales como definir alertas para monitorear un tópico de interés, especificar trackers temáticos sobre redes sociales, generar consultas para diversos tipos de motores de búsqueda (tales como Google, Wikipedia o PubMed), recolectar material para la construcción de portales temáticos, explorar la web profunda (deep web), o facilitar la captura y construcción de ontologías y modelos de conocimiento, entre otras.

El problema de generar automáticamente consultas o frases temáticas para especificar necesidades de información es de difícil resolución, tanto desde un punto de vista teórico como computacional. Entre las características y desafíos que presenta dicho problema podemos mencionar los siguientes:

- El problema de generar consultas temáticas puede ser abordado como un problema de optimización, donde el espacio de búsqueda se define como el conjunto de posibles consultas que pueden ser construidas. La función objetivo a optimizar deberá incorporar criterios que reflejen la efectividad de una consulta a la hora de recuperar material relevante (tales como precisión, cobertura, novedad).
- El problema de optimización de consultas no tiene subestructura óptima, lo que significa que una solución óptima no puede ser construida eficientemente a partir de soluciones subópitmas [8].
- El espacio de búsqueda está formado por un gran número de dimensiones, donde cada término o atributo incorpora una nueva dimensión. La complejidad se ve aumentada aún más si tomamos en cuenta la posibilidad de construir consultas con sintaxis complejas.
- La construcción de consultas múltiples y diversas, aún cuando las mismas no sean las más óptimas, resulta de gran utilidad a la hora de especificar los requerimientos de información.

En vista a las características del problema en cuestión y en base a resultados preliminares exitosos [5, 6], anticipamos que una propuesta basada en el uso de algoritmos evolutivos brindará una solución apropiada para abordar los desafíos planteados. En particular, anticipamos que la aplicación de algoritmos evolutivos multiobjetivo, con especial énfasis en la incorporación de técnicas de programación genética para generar soluciones representadas con estructura de árboles, resultará en una solución efectiva y novedosa al problema en cuestión. Por otra parte, la incorporación de técnicas para diversificar el conjunto de soluciones redundará en el desarrollo de métodos capaces de capturar mayor número de resultados relevantes y novedosos, sin perder el foco del contexto bajo análisis.

4. ANTECEDENTES Y LÍNEA DE INVESTIGACIÓN ESPECÍFICA

La reformulación y extensión de consultas ha sido reconocida como una tarea importante en los

sistemas de recuperación de información [20]. Se sabe que los usuarios concretan sus necesidades de información por medio de un lenguaje simple y acotado. Por un lado, se ha encontrado una tendencia a utilizar un número muy reducido de términos en la formulación de consultas [21]. Por otro lado, no siempre las palabras que caracterizan a cierto tópico de interés serán obvias y cotidianas. Por estas razones la reformulación o extensión de consultas juega un rol fundamental a la hora de recuperar información.

En varias propuestas, se han usado técnicas de reformulación de consultas para aumentar automáticamente las consultas del usuario, incorporando términos seleccionados del contexto [3, 15, 11]. Dicho contexto podría consistir en páginas web que el usuario está visitando o ha visitado, en emails que está editando o en documentos que está creando.

Los primeros intentos de aplicar computación evolutiva en el área de la recuperación de información se remontan a finales de la década de los 80. El foco en ese entonces era usar técnicas de computación genética para derivar mejores descriptores de documentos, con el fin de facilitar el indexado o el agrupamiento (clustering) de información [12, 19]. Algunas técnicas basadas en algoritmos genéticos fueron también aplicadas a reforzar el peso de los términos durante la optimización de consultas [22], a construir conceptos representados por gran número de términos [17] y a extender el conjunto inicial de resultados mediante consultas mejoradas [14].

Existen algunas propuestas más recientes que son más cercanas a nuestra propuesta [16, 5, 6]. En estos trabajos el objetivo es proponer una metodología para evolucionar consultas mediante técnicas basadas algoritmos genéticos multiobjetivos. Sin embargo, en ellos no se abordan dos temas centrales que planteamos en esta línea: la evolución de consultas con sintaxis complejas y la incorporación de técnicas orientadas a promover la diversidad de resultados.

La utilización de operadores booleanos (tales como AND, OR y NOT), como así también otros modificadores de búsqueda (tales como el manejo de sinónimos o búsqueda por proximidad) han demostrado ventajas en la reformulación de consultas cuando los usuarios presentan dificultades de búsqueda [7, 10]. Por lo tanto, creemos que su incorporación será también beneficiosa en la formulación automática de consultas. Por ejemplo, en el caso de la búsqueda basada en un tópico de interés, el uso de consultas con sintaxis más complejas permitirá expresar los conceptos involucrados en las necesidades de información con un mayor grado de precisión. Cabe aclarar que a diferencia de estos dos trabajos ([7, 10]), los métodos desarrollados en esta línea de investigación estarán destinados tanto a aprender de forma completamente automática (sin la asistencia del usuario) los términos y los operadores (booleanos y sintácticos) que los relacionan, como a identificar subtópicos dentro del tópico de interés dado.

Otro aspecto importante del problema a abordar consiste en la identificación de múltiples soluciones óptimas o cuasi-óptimas. En el campo de la computación evolutiva multi-objetivo se han propuesto diferentes estrategias para promover la diversidad de la población. Entre dichas estrategias podemos mencionar los métodos de *niching* tales como *crowding* [9] y *fitness sharing* [9]. Sin embargo, estas técnicas apuntan a que los valores obtenidos por las funciones de aptitud (e.g., precisión y cobertura) aplicada a diversos individuos (e.g., consultas) cubran de manera uniforme el frente de Pareto. Es decir, se intenta diversificar en el el espacio de objetivos. Esto es diferente a nuestra propuesta, la cual se orienta a obtener diversidad en el conjunto de resultados recuperados por las consultas generadas automáticamente.

5. RESULTADOS PREVIOS Y OBJETIVOS

En una primera etapa, una tarea concreta será trabajar en la utilización de algoritmos evolutivos como mecanismo heurístico para lograr la exploración y explotación de los grandes espacios de búsqueda presentes actualmente. Más específicamente, se proyecta trabajar en el estudio y mejora de técnicas para reformulación y extensión de consultas. En este punto las primeras tareas estarán orientadas al diseo de métodos evolutivos que logren la recuperación de material novedoso

al mismo tiempo que relevante para un tópico de interés dado. Para lo cual se trabajará, en principio, en dos aspectos puntuales: *proponer e implementar mecanismos que permitan controlar la diversidad poblacional* y *estudiar el impacto de la incorporación de operadores booleanos a la sintaxis de las consultas evolucionadas.*

En últimos experimentos realizados con la arquitectura propuesta en [4], se observó que las últimas poblaciones tienden a aprender conjuntos de términos con un alto nivel de solapamiento. Creemos que una diversidad equilibrada aportará un doble beneficio a nivel de obtención de resultados. Por un lado, contribuirá a una mejor exploración del espacio de búsqueda, dado que consultas que sean diferentes entre sí tenderán a explorar diferentes sectores del espacio de búsqueda. Por otro lado, si cada consulta puede ser representante de un sector diferente de un conjunto de documentos relevantes, entonces se podrían identificar subtópicos de tal manera que cada consulta sea responsable de determinada sugerencia de interés. Con este objetivo, una de las tareas principales consistirá en la formulación, implementación y análisis de nuevas métricas de evaluación de aptitud y nuevas métricas de evaluación de desempeo para este tipo de arquitecturas. Hasta aquí, los algoritmos que hemos desarrollado utilizan funciones de aptitud que evalúan a cada individuo de forma particular, es decir sin tener en cuenta a la población en forma colectiva. Como parte de estas investigaciones se estudiarán diferentes alternativas que permitan incorporar conocimiento comunitario en la evaluación de cada individuo, para ello se estudiará cómo puede llevarse a cabo la adaptación de diferentes métricas clásicas (tales como precisión o cobertura) de forma que puedan incorporar esta mirada más globalizada y también se estudiarán otras métricas que ya lo hacen por definición, tales como la entropía.

Paralelamente, en una tarea a mayor plazo, se disearán, estudiarán y evaluarán diferentes alternativas evolutivas para la reformulación de consultas con sintaxis más complejas para estudiar los resultados que surgen al incorporar operadores booleanos. También se estudiará la incorporación de otro tipo de operadores o modificadores de búsqueda, como “”, ‘+’, ‘-’, ’∼’, u otras sintaxis especializadas, dependiendo del buscador subyacente a ser utilizado.

Para los experimentos se utilizarán corpus de documentos pre-clasificados por tópico, tales como ODP (http://dmoz.org) y motores de búsqueda configurables tales como la infraestructura Terrier [18]. También anticipamos utilizar buscadores web como Yacy (http://yacy.net) o Faroo (http://www.faroo.com). En estos últimos casos las evaluaciones deberán basarse en métricas que no incorporen la noción de relevancia, por tratarse de buscadores a gran escala donde la relevancia de cada documento no se conoce de antemano .

6. FORMACIÓN DE RECURSOS HUMANOS

Este trabajo será parte de la formación de posgrado de Cecilia Baggio, actual Tesista de doctorado de la UNS y Becaria de CONICET desde abril de 2014. Cecilia realizará actividades de investigación como parte del *Knowledge Management and Information Retrieval Research Group*, en el cual participan actualmente los investigadores: Ana Maguitman (directora del grupo), Carlos Chesñevar, Carlos Lorenzetti, Rocío Cecchini e Ignacio Ponzoni, y los Becarios y estudiantes de Posgrado: Cristian Briguez, Ana Nicolini, Eduardo Xamena y Cecilia Baggio entre otros.

REFERENCIAS

[1] https://blog.twitter.com/2011/numbers.

[2] http://techcrunch.com/2013/05/17/facebook-growth/.

[3] Jay Budzik, Kristian J. Hammond, and Larry Birnbaum. Information Access in Context. *Knowledge-Based Systems*, 14:37–53, 2001.

[4] Rocío L. Cecchini. *Computación Evolutiva como Soporte en Minería de Datos y Texto*. AV Akademikerverlag GmbH & Co. KG (EAE), 1 edition, 2012.

[5] Rocío L. Cecchini, Carlos M. Lorenzetti, Ana G. Maguitman, and Nélida B. Brignole. Using genetic algorithms to evolve a population of topical queries. *Information Processing and Management*, 44(6):1863–1878, 2008.

[6] Rocío L. Cecchini, Carlos M. Lorenzetti, Ana G. Maguitman, and Nélida B. Brignole. Multi-objective Evolutionary Algorithms for Context-based Search. *Journal of the American Society for Information Science and Technology*, 61(6):1258–1274, June 2010.

[7] O. Cordón, E. Herrera-Viedma, and M. Luque. Improving the learning of boolean queries by means of a multiobjective {IQBE} evolutionary algorithm. *Information Processing & Management*, 42(3):615 – 632, 2006.

[8] Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest. *Introduction to Algorithms*. MIT Press, Cambridge, MA, USA, 1st edition, August 1990.

[9] Kenneth Alan De Jong. *An analysis of the behavior of a class of genetic adaptive systems.* PhD thesis, Ann Arbor, MI, USA, 1975.

[10] Huizhong Duan, Rui Li, and ChengXiang Zhai. Automatic query reformulation with syntactic operators to alleviate search difficulty. In Craig Macdonald, Iadh Ounis, and Ian Ruthven, editors, *CIKM*, pages 2037–2040. ACM, 2011.

[11] Ariel Fuxman, Patrick Pantel, Yuanhua Lv, Ashok Chandra, Pradeep Chilakamarri, Michael Gamon, David Hamilton, Bernhard Kohlmeier, Dhyanesh Narayanan, Evangelos Papalexakis, and Bo Zhao. Contextual insights. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, pages 265–266, Republic and Canton of Geneva, Switzerland, 2014.

[12] Michael Gordon. Probabilistic and genetic algorithms in document retrieval. *Communications of the ACM*, 31(10):1208–1218, 1988.

[13] Peder Olesen Larsen and Markus von Ins. The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics*, 84(3):575–603, 2010.

[14] Gondy Leroy, Ann M. Lally, and Hsinchun Chen. The use of dynamic contexts to improve casual internet searching. *ACM Trans. Inf. Syst.*, 21(3):229–253, 2003.

[15] Carlos M. Lorenzetti and Ana G. Maguitman. A semi-supervised incremental algorithm to automatically formulate topical queries. *Information Sciences*, 179(12):1881–1892, 2009. Including Special Issue on Web Search.

[16] Antonio Gabriel López-Herrera, Enrique Herrera-Viedma, and Francisco Herrera. A study of the use of multi-objective evolutionary algorithms to learn boolean queries: A comparative study. *JASIST*, 60(6):1192–1207, 2009.

[17] Zacharis Z. Nick and Panayiotopoulos Themis. Web Search Using a Genetic Algorithm. *IEEE Internet Computing*, 5(2):18–26, 2001.

[18] Iadh Ounis, Christina Lioma, Craig Macdonald, and Vassilis Plachouras. Research Directions in Terrier: a Search Engine for Advanced Retrieval on the web. *Novatica/UPGRADE Special Issue on Web Information Access, Ricardo Baeza-Yates et al. (Eds), Invited Paper*, VIII(1):49–56, February 2007.

[19] Vijay Raghavan and Brijesh Agarwal. Optimal determination of user-oriented clusters: an application for the reproductive plan. In *Proceedings of the Second International Conference on Genetic Algorithms on Genetic algorithms and their application*, pages 241–246, Mahwah, NJ, USA, 1987. Lawrence Erlbaum Associates, Inc.

[20] Joseph John Rocchio. Relevance feedback in information retrieval. In Gerard Salton, editor, *The Smart retrieval system - experiments in automatic document processing*, chapter 14, pages 313–323. Prentice-Hall, Englewood Cliffs, NJ, USA, 1971.

[21] Haocheng Wu, Wei Wu, Ming Zhou, Enhong Chen, Lei Duan, and Heung-Yeung Shum. Improving search relevance for short queries in community question answering. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 43–52, New York, NY, USA, 2014. ACM.

[22] Jing-Jye Yang and Robert Korfhage. Query optimization in information retrieval using genetic algorithms. In *ICGA*, pages 603–613, 1993.