Sistema híbrido para clasificación de documentos aplicado al problema bioinformático de inferencia de interacción proteína-proteína

Rocío L. Cecchini[1] **Carlos M. Lorenzetti**[1] **Ana G. Maguitman**[1] **Ignacio Ponzoni**[1,2]

[1]Departamento de Ciencias e Ingeniería de la Computación,
Universidad Nacional del Sur, Av. Alem 1253, (8000) Bahía Blanca, Argentina
[2]Planta Piloto de Ingeniería Química, CCT-CONICET-Bahía Blanca,
Cno. la Carrindanga km 7, (8000), Bahía Blanca, Argentina
e-mail: {rlc,cml,agm,ip}@cs.uns.edu.ar

1. INTRODUCCIÓN

En la actual era postgenómica, el estudio de las interacciones existentes entre proteínas resulta una pieza clave en la compresión de los complejos mecanismos moleculares presentes en los procesos biológicos. Es sabido que la información acerca de interacciones proteína-proteína mejora el entendimiento sobre ciertas enfermedades, así como también puede ofrecer nuevas perspectivas para el desarrollo de tratamientos específicos. Sin embargo, gran parte del conocimiento sobre estas interacciones muchas veces permanece oculto en la inmensa cantidad de artículos científicos, reportes técnicos y demás trabajos científicos que día a día son publicados. Asimismo, el número creciente de revistas especializadas en Biología Computacional, más su naturaleza interdisciplinaria que hace viable la publicación de resultados en foros muy diversos (tales como medicina, biología, física o informática), tornan aún más difícil esta tarea de extracción y curación de nuevo conocimiento biológico. Para tomar una dimensión de este problema, basta con mencionar que tan solo la base de datos MEDLINE [16], fuertemente empleada en Biología Computacional, contiene alrededor de 19 millones de citas. Por otra parte, las publicaciones científicas son una de las principales fuentes que los curadores de bases de datos utilizan para sus anotaciones manuales, por lo que gran parte de la información funcional contenida en las bases de datos biológicas ha sido extraída directa o indirectamente de estas publicaciones.

Estas razones motivaron a que durante los últimos años hayan surgido distintos enfoques de minería de texto para extraer de forma sistematizada este conocimiento subyacente en la literatura científica [8, 5, 19]. La relevancia de la minería de texto en los dominios de la biología y la medicina se ve reflejada en varias competencias, tales como la "Knowledge Discovery and Data Mining (KDD) Challenge Cup" [18], la "Text Retrieval Conference (TREC) Genomics" [10] o las diferentes ediciones de "Critical Assessment of Information Extraction systems in Biology (BioCreative)" [11, 13]. La tarea de automatizar la interpretación y exploración de esta clase de publicaciones científicas no es computacionalmente sencilla. A la inherente complejidad del lenguaje natural, se suman otros problemas específicos del ámbito de Bioinformática, tales como la ambigüedad en los nombres de los genes de las eucariotas [4], y la existencia de distintos sistemas de notación [15]. Debido a estas cuestiones, se ha tornado necesario descomponer esta tarea en varios subproblemas [20], a fin de lograr una metodología de minería de texto que sea tan eficaz como eficiente.

En este contexto, un primer paso hacia la definición de un método exitoso para la reconstrucción de redes de interacción entre proteínas estaría dado por la correcta identificación de los artículos científicos que son relevantes para este tópico. De este modo, se reduciría el universo de artículos a explorar en las fases siguientes. Una alternativa para abordar esta tarea es la propuesta por Abi-Haidar *et al*. [2008], quienes adaptaron una técnica de detección de correo basura al problema de identificar pu-

blicaciones relativas a la interacción de proteínas. Este clasificador utiliza un número reducido de características y una superficie de decisión lineal. Para el problema en cuestión, esta técnica ha demostrado ser tanto o más efectiva que otras técnicas conocidas, tales como Máquina de Vectores de Soporte (SVM) [6] y métodos basados en la aplicación de Análisis de Semántica Latente (LSA) [7]. Otras técnicas que han tenido un buen desempeño en estas tareas han utilizado SVM como clasificador, considerando además la identificación de características especiales a través del uso de ontologías o nomenclaturas [14].
En esta línea de investigación proponemos diseñar, implementar y evaluar una infraestructura inteligente, combinando distintas herramientas computacionales para clasificación de documentos. A partir de un conjunto de documentos previamente clasificados, el sistema inteligente utiliza: un *módulo de clustering* para identificar los posibles subtópicos dentro del conjunto de documentos, un *módulo evolutivo* para determinar la importancia global de las palabras y un *módulo de clasificación* que utiliza dicha información para inferir la clase a la que pertenece un nuevo documento. Particularmente, se propone el uso del sistema para un problema importante en el área de bioinformática: *detectar si un documento trata sobre interacción entre proteínas*, el cual puede verse como un caso particular del problema de clasificación.

2. LÍNEA DE INVESTIGACIÓN PROPUESTA

Nuestro objetivo es identificar las palabras más relevantes para los documentos que tratan de interacciones entre proteínas, una vez identificadas estas palabras proponemos entrenar un clasificador para que sea capaz de inferir si un nuevo documento corresponde al tópico de interacción entre proteínas. Con el fin de alcanzar esta meta, comenzamos dividiendo en subtópicos un conjunto inicial de documentos previamente identificados como documentos pertenecientes al tópico de interacción entre proteínas. Cada subtópico identificado en esta división es utilizado por el módulo evolutivo para determinar la importancia de las palabras. Finalmente, un mecanismo clasificador es entrenado utilizando: el corpus inicial de documentos y la información de relevancia inferida por el módulo evolutivo.

2.1. Esquema General del Sistema

La figura 1 muestra un esquema general del sistema propuesto, donde se destacan los siguientes componentes:

- **Modelo inicial del corpus:** contiene una representación de todos los documentos contenidos en el corpus destinado al entrenamiento del sistema. Estos documentos han sido previamente clasificados en las clases de interés.
- **Módulo de clustering:** analiza el modelo inicial del conjunto de documentos previamente clasificados con el fin de identificar subtópicos.
- **Módulo evolutivo:** para cada subtópico identificado se genera una población inicial de consultas, la cual evoluciona guiada por una función de aptitud utilizando los operadores genéticos de selección, recombinación y mutación. Durante esta evolución, se mantiene un banco de mutación que es enriquecido con las nuevas palabras que se van encontrando en los documentos recuperados por las consultas. Las consultas de la población final son utilizadas para determinar la importancia de las palabras.
- **Modelo de ponderación ajustado:** es un nuevo modelo del corpus de documentos utilizado para entrenamiento, ajustado según la importancia inferida por el módulo evolutivo.
- **Módulo clasificador:** es un sistema de clasificación de documentos que utiliza el modelo de ponderación ajustado para la fase de entrenamiento. La porción del corpus reservada para testeo en la etapa inicial es usada para analizar el desempeño de la arquitectura.

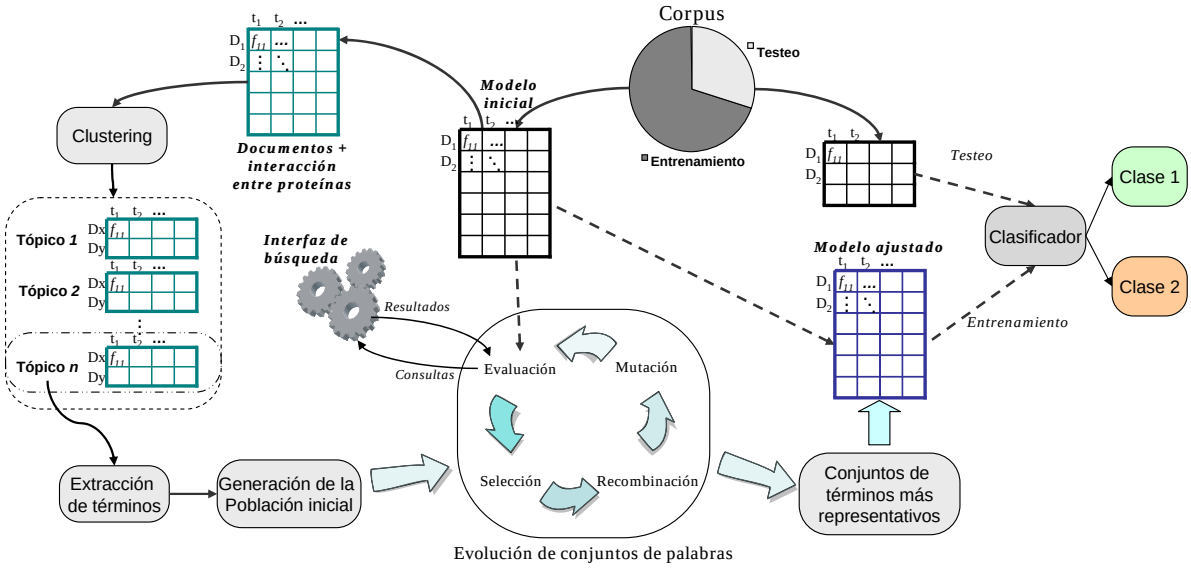

Figura 1: Arquitectura del sistema

2.2. Módulo de clustering

Dada la posibilidad de encontrarnos ante una gran cantidad de material en el corpus inicial, la arquitectura incorpora el módulo de clustering, el cual es capaz de realizar un refinamiento previo sobre los documentos con el fin de mejorar la precisión del módulo evolutivo. Es natural pensar que dentro de un conjunto grande de documentos relacionados con determinado tópico convivan a su vez subtópicos más pequeños. La finalidad del módulo de clustering es identificar posibles subtópicos y generar una nueva estructura para que el algoritmo evolutivo (EA) pueda ser ejecutado sobre cada subtópico.

Existen diferentes formas de denominar a los grupos identificados al realizar una tarea de clustering, por ejemplo, se llaman *exclusivos* aquellos en los que cada ítem se corresponde con un único grupo, *solapados* son aquellos en los que una instancia puede pertenecer a varios grupos, los *probabilísticos* son aquellos en los que cada elemento pertenece a cada grupo con cierta probabilidad y jerárquicos son grupos organizados bajo una estructura de jerarquías. La elección entre las posibilidades mencionadas depende de los mecanismos que subyacen bajo el fenómeno de clustering y por razones prácticas usualmente es impuesta por las herramientas de clustering disponibles. En esta etapa de la arquitectura se proyecta aprovechar las facilidades provistas en el entorno Weka [9]. Este paquete contiene una colección de herramientas para visualización y análisis de datos, entre las cuales se encuentran algunas herramientas para clustering que utilizan algoritmos basados en distancia (para clasificación exclusiva y clasificación jerárquica) y algoritmos basados en modelos estadísticos (para clustering probabilístico).

2.3. Módulo evolutivo

Utiliza algoritmos evolutivos mono- o multiobjetivo dependiendo de las medidas de aptitud requeridas para evolucionar una población consultas. El ciclo evolutivo comienza con una población de consultas compuestas por palabras extraídas de uno o más documentos pertenecientes al tópico actual. Las consultas se evalúan de acuerdo a la calidad de los resultados recuperados a partir de cada una de ellas. A medida que las generaciones avancen, 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. Este ciclo evolutivo se ejecuta para cada uno de los subtópicos encontrados por el mecanismo de clustering, permitiendo que las posibles diferentes regiones del corpus

sean representadas por diferentes conjuntos de palabras.

Población y Representación de Cromosomas. 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 con un número fijo de consultas generadas a partir de palabras seleccionadas aleatoriamente del tópico actual.

Evaluación de los individuos. Se pueden utilizar distintas funciones de aptitud y distintos esquemas evolutivos. Por ejemplo, se podrían utilizar las métricas de precisión y cobertura dentro de un esquema Pareto o dentro de un esquema agregativo que combine ambas funciones en una única fórmula. En el segundo caso se pueden utilizar distintas métricas conocidas en el área de recuperación de información, tal como la media armónica ponderada de precisión y cobertura F_1 [17].

Operadores Genéticos. Cada nueva generación se obtiene luego de aplicar probabilísticamente los operadores de selección, recombinación y mutación sobre las consultas de la población actual:

- *Selección.* Se genera una nueva población seleccionando probabilísticamente las consultas de mayor calidad. La probabilidad de que una consulta $\mathbf{q}$ 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. La recombinación de un par de consultas se lleva a cabo copiando palabras de cada uno de los padres en los descendientes.
- *Mutación.* Son pequeños cambios aleatorios en las consultas que consisten en reemplazar una palabra t_i^q seleccionada al azar por otra palabra t_j^p. Esta última se obtiene del *banco de mutación* que describimos a continuación.

Banco de Mutación. Para cada subtópico, el banco de mutación es un conjunto de palabras compuesto inicialmente por palabras provenientes de uno o más documentos pertenecientes al subtópico. A medida que el sistema recupera resultados relevantes para dicho subtópico, las palabras que aparecen en los documentos devueltos por el buscador se irán agregando al banco de mutación. Este procedimiento da al EA la posibilidad de generar consultas con palabras nuevas, permitiendo así una exploración más amplia del espacio de búsqueda.

2.4. Módulo clasificador

Finalmente, en base a la información de relevancia brindada por el modelo ajustado del corpus, un sistema de clasificación identificará si cada documento trata de interacción proteína-proteína o no. En esta última fase, se espera utilizar el *modelo de ponderación ajustado* para entrenar al clasificador, mientras que la porción del corpus reservada inicialmente nos permitirá evaluar el desempeño global del sistema y compararlo con otros métodos. Durante esta etapa se proyecta utilizar el paquete de clasificación brindado por el entorno Weka, el cual contiene una variedad de algoritmos de clasificación entre los cuales se encuentran un grupo de algoritmos basados en estadística bayesiana, un grupo de algoritmos basados en árboles de decisión, un grupo de algoritmos basados en reglas de decisión y otros clasificadores basados en regresión lineal por cuadrados mínimos, regresión logística y regresión por *k*-vecinos.

3. COLECCIONES DE DATOS Y EVALUACIONES

Como conjunto de entrenamiento planeamos utilizar una colección de resúmenes de artículos científicos indexados por PubMed. Cada artículo que forma parte del conjunto de entrenamiento se encuentra etiquetado como relevante o irrelevante para el tópico interacción proteína-proteína. Estos artículos han sido curados de acuerdo a los estándares IntAct [12] y MINT [3]. Por otra parte utilizaremos un conjunto de testeo consistente de resúmenes de

artículos no etiquetados. Este conjunto de datos fue utilizado en la competencia BioCreative II [13]. Para determinar la eficacia de los métodos propuestos planeamos utilizar métricas clásicas de evaluación de clasificadores tales como precisión, cobertura, F-score y área bajo la curva ROC.

4. CONCLUSIONES

Las técnicas propuestas aquí pueden ser utilizadas en la implementación de diferentes aplicaciones para clasificación de documentos. Técnicas basadas en algoritmos evolutivos multi-objetivo han sido utilizadas con éxito en problemas de recuperación de información [2], anticipando que las tareas de identificación de subtópicos propuestas en estas líneas de investigación contribuirán a mejorar la calidad de los clasificadores resultantes.

REFERENCIAS

[1] A. Abi-Haidar, J. Kaur, A. Maguitman, P. Radivojac, A. Rechtsteiner, K. Verspoor, Z. Wang, and L. Rocha. Uncovering protein interaction in abstracts and text using a novel linear model and word proximity networks. *Genome Biology*, 9(Suppl 2):S11, 2008.

[2] R. L. Cecchini, C. M. Lorenzetti, A. G. Maguitman, and N. B. Brignole. Multi-objective evolutionary algorithms for context-based search. *Journal of the American Society for Information Science and Technology*, -:In press, 2010.

[3] A. Chatr-Aryamontri, A. Ceol, L. M. Palazzi, G. Nardelli, M. V. Schneider, L. Castagnoli, and G. Cesareni. MINT: the Molecular INTeraction database. *Nucleic Acids Research*, 35:D572–D574, January 2007.

[4] L. Chen, H. Liu, and C. Friedman. Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics*, 21(2):248–256, 2005.

[5] J.-H. Chiang, H.-C. Yu, and H.-J. Hsu. GIS: a biomedical text-mining system for gene information discovery. *Bioinformatics*, 20(1):120–121, 2004.

[6] N. Cristianini and J. Shawe-Taylor. *An introduction to support Vector Machines: and other kernel-based learning methods*. Cambridge University Press, New York, NY, USA, 2000.

[7] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[8] I. Donaldson, J. Martin, B. de Bruijn, C. Wolting, V. Lay, B. Tuekam, S. Zhang, B. Baskin, *et al*. Prebind and textomy - mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*, 4(1):11, 2003.

[9] E. Frank, M. A. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I. H. Witten, and L. Trigg. WEKA - a machine learning workbench for data mining. In *The Data Mining and Knowledge Discovery Handbook*, pages 1305–1314. Springer, 2005.

[10] W. Hersh, A. M. Cohen, P. Roberts, and H. K. Rekapalli. TREC 2006 genomics track overview. In *TREC Notebook*. NIST, 2006.

[11] L. Hirschman, A. S. Yeh, C. Blaschke, and A. Valencia. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6(S1):S1, May 2005.

[12] S. Kerrien, Y. Alam-faruque, B. Ar, I. Bancarz, A. Bridge, C. Derow, E. Dimmer, M. Feuermann, A. Friedrichsen, R. Huntley, *et al*. Intact–open source resource for molecular interaction data. *Nucleic Acids Research*, 35:561–565, 2007.

[13] M. Krallinger, F. Leitner, C. Rodriguez-Penagos, and A. Valencia. Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biology*, 9(2):S4, 2008.

[14] M. Krallinger, R. Malik, and A. Valencia. Text mining and protein annotations: the construction and use of protein description sentences. *Genome Informatics*, 17:121–130, 2006.

[15] U. Leser and J. Hakenberg. What makes a gene name? Named entity recognition in the biomedical literature. *Briefings in Bioinformatics*, 6(4):357–369, January 2005.

[16] Pubmed. http://www.ncbi.nlm.nih.gov/pubmed/.

[17] C. J. v. Rijsbergen. *Information Retrieval*. Butterworth - Heinemann, Newton, MA, USA, 1979.

[18] A. S. Yeh, L. Hirschman, and A. A. Morgan. Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. *Bioinformatics*, 19(S1):i331–i339, 2003.

[19] D. Zhou, Y. He, and C. K. Kwoh. Extracting Protein-Protein Interactions from MEDLINE using the Hidden Vector State model. *International Journal of Bioinformatics Research and Applications*, 4(1):64–80, 2008.

[20] D. Zhou, Y. He, and C. K. Kwoh. From Biomedical Literature to Knowledge: Mining Protein-Protein Interactions. In *Computational Intelligence in Biomedicine and Bioinformatics*, volume 151 of *Studies in Computational Intelligence*, pages 397–421. Springer, Berlin / Heidelberg, 2008.