Aplicación de Reconocimiento de Patrones en Repositorios Digitales de Datos Biológicos

Joaquín R. Lima[1], Gustavo D. Samec[1,2], Romina Stickar[1], Renato Mazzanti[1,2]

[1]Dpto. de Informática, Fac. de Ingeniería, Universidad Nacional de la Patagonia San Juan Bosco.
+54-0280-4472885-int. 117 - Puerto Madryn, Argentina
[2]Centro Nacional Patagónico – CONICET – Servicio Centralizado de Computación
+54-0280-4450401-int. 1260 - Puerto Madryn, Agentina

limajoaquin@hotmail.com, gsamec@gmail.com romistickar@gmail.com renato.mazzanti@gmail.com

Resumen

El presente trabajo surge como una necesidad puntual de sistematizar un gran volumen de datos históricos relacionado con citas bibliográficas de especimenes invertebrados bentónicos marinos del sur de Sudamérica, con el objetivo de extraer información en forma eficiente teniendo presente las características propias de los datos registrados.

El trabajo, además de las metas de investigación y desarrollo, persigue como objetivo académico la formación de recursos humanos. Se desarrolla en el marco de pasantías de la carrera de grado Licenciatura en Informática.

Palabras clave: Reconocimiento de Patrones, Búsqueda Semántica, Repositorios Digitales

Contexto

El presente trabajo se realiza en el marco de pasantías de la carrera de grado de uno de sus participantes en el Centro Nacional Patagónico como Instancia Supervisada de Formación Práctica Profesional (ISFPP) de la materia Inteligencia Artificial de La Carrera Licenciatura en Informática y como aporte futuro al Proyecto de Investigación "Técnicas de recuperación de información en grandes volúmenes de datos heterogéneos con bases de datos No-Sql" desarrollado por docentes y alumnos de la Fac. de Ingeniería de la Universidad Nacional de La Patagonia San Juan Bosco (UNPSJB) Sede Puerto Madryn.

Introducción

El manejo de información en los repositorios digitales de datos biológicos con un volumen significativo de datos, demanda métodos de búsqueda eficientes y precisos. Se debe dotar a las búsquedas con la inteligencia necesaria para extraer lo que los usuarios realmente quieren encontrar en sus consultas.

La búsqueda semántica aplicando Inteligencia Artificial nos permiten desarrollar soluciones a esta problemática [1].

Contar con herramientas que faciliten la recuperación de datos históricos de registros biológicos es de gran ayuda para quienes trabajan en líneas de investigación relacionadas con la distribución y evolución de las especies,

desplazamientos, crecimiento, adaptación, etc.

Algunas de las características de los datos a manejar son:

- La referencia a un determinado taxón dentro del árbol taxonómico con un mayor o menor grado de determinación dentro del mismo.
- Nombre de autores, investigadores, colectores que lo descubrieron, clasificaron, reclasificaron.
- Referencias a expediciones o campañas donde se colectaron los especimenes. Georreferencias de los mismos.
- Repositorio donde se encuentran los especimenes, número de catálogo, estado de conservación, etc.

Si bien la referencia a un taxón pareciera facilitar su clasificación, en la recopilación de citas bibliográficas se han encontrado problemas tales como:

- Un mismo espécimen ha sido clasificado por distintos investigadores a lo largo de su historia con un taxón diferente.
- El árbol taxonómico no es estático va evolucionando con el tiempo y ello puede llevar a reexaminar y reclasificar nuevamente algunos especimenes.
- Los nombres de autores, investigadores, colectores, etc. no siguen un estándar determinado sobretodo en las citas bibliografiítas antiguas. Es frecuente encontrar referencias a un mismo autor escrito de manera distinta. También se encuentran casos donde el mismo nombre puede pertenecer a personas distintas y donde el contexto donde se lo menciona permite diferenciarlo.
- Las georreferencias son ambiguas o incompletas sobre todo en los registros más antiguos. Hay referencias a nombres geográficos que son similares en muchas regiones. Las coordenadas geográficas son poco precisas, etc.

Los patrones son entidades con algún nombre y están representadas por un conjunto de propiedades medias y relacionadas entre ellas [2]. El reconocimiento de patrones (RP) estudia la descripción y clasificación de patrones a clases. Las clases son conjunto de entidades que comparten alguna característica que las diferencia de otras. Básicamente, el RP es un proceso que se encarga de capturar y pre-procesar datos de algún patrón, para luego extraer las características significativas y en base a algún algoritmo y dichas características, clasificar al patrón en alguna categoría.

Líneas de Investigación, Desarrollo e Innovación

Este proyecto se centra, en un principio, en la investigación de diversos modelos de RP, su análisis y comparación

para después aplicar un candidato en un desarrollo integral con un sistema existente de repositorios bibliográficos digitales. El objetivo general es dotar de búsquedas semánticas al sistema existente.

Resultados y Objetivos

El proyecto se encuentra en su etapa inicial. Se están evaluando las características que presentan lo datos y estudiando distintos modelos de reconocimiento de patrones.

Los objetivos de la investigación y desarrollo de este trabajo se pueden resumir en:

Investigar la aplicabilidad de los modelos de reconocimiento de patrones en textos de diferentes formatos [3] (texto plano, pdf, etc.)

Estudiar los diferentes modelos de clasificación (Clasificador Bayesiano y Redes Neuronales [4][5]). Analizar ventajas y desventajas de cada uno y decidir el mejor modelo a utilizar en este contexto.

Implementar un reconocedor de patrones para el conjunto de datos existente.

Como una segunda etapa del proyecto se pretende modificar el vector de características para lograr una nueva categorización en base a:

- Reconocimiento de Entidades Nominales
- Sinonimia
- Polisemia
- Expresiones Temporales

Formación de Recursos Humanos

Este trabajo se lleva a cabo en el contexto de la pasantía realizada por uno de sus autores dentro de La Carrera Licenciatura en Informática. Por otra parte, el alumno pretende ampliar su investigación durante el desarrollo de su tesina de grado.

Referencias

[1] Califf, Mary and Raymond J. Mooney (1997) CoNLL: Computational Natural Language Learning.

[2] Marie-Francine Moens. (2006) Information Extraction: Algorithms and Prospects in a Retrieval Context

[3] F. Cruz, J. A. Troyano and F. J. Ortega (2006): Procesamiento de Lenguaje Natural

[4] Haykin S. (1994) Neural Networks-a comprehensive foundation.

[5] Cazorla Quevedo, Miguel Angel (2000) Un enfoque bayesiano para la extracción de características y agrupamiento en visión artificial.