Procesamiento de textos estructurados

Marina Cardenas, Julio Castillo

Laboratorio de Investigación de Software LIS, Dpto. Ingeniería en Sistemas de Información
Facultad Regional Córdoba, Universidad Tecnológica Nacional

{ing.marinacardenas, jotacastillo}@gmail.com

RESUMEN

En este artículo se describe un proyecto de investigación relacionado al procesamiento de textos estructurados, en particular a archivos de códigos fuentes en algún lenguaje de programación.

Se propone abordar este problema obteniendo información a diferentes niveles de abstracción, adaptando técnicas que son específicas de implicación textual.

El proyecto se inserta en una línea de investigación de aprendizaje automático y de lingüística computacional.

Se describe el proyecto, los resultados obtenidos hasta el momento, como así también, los resultados que se esperan de la utilización del sistema. Finalmente, se presenta la línea de investigación en la que se enmarca este proyecto.

Palabras clave: análisis de texto, extracción de información, corpus.

CONTEXTO

El proyecto acreditado por la Universidad Tecnológica Nacional denominado *Modelado para el procesamiento de textos estructurados* (código UTN4518) se encuentra consolidado dentro de la línea de investigación relacionada con lingüística computacional y es llevado a cabo en el Laboratorio de Investigación de Software LIS[1] del Departamento de Ingeniería en Sistemas de Información de la Universidad Tecnológica Nacional Facultad Regional Córdoba.

A su vez, este proyecto se encuentra dentro del grupo de investigación denominado Grupo de Inteligencia Artificial (o GIA) de la UTN-FRC.

El grupo GIA reúne a proyectos de investigación que se hayan todos en temáticas concernientes a la inteligencia artificial entre las que podemos destacar análisis de imágenes, algoritmos evolutivos, y su aplicabilidad en problemas de la ingeniería, de las ciencias naturales, y de las ciencias sociales.

Este grupo está compuesto de becarios, pasantes, docentes investigadores y doctores.

1. INTRODUCCIÓN

En el proyecto de procesamiento de textos estructurados se plantea el desarrollo de un modelo para detección de similitudes de código fuente para poder determinar la existencia de prácticas de reutilización aplicando técnicas vinculadas a la lingüística computacional, tales como minería de datos sobre texto y procesamiento del lenguaje natural, dado que según [1] "los lenguajes de programación se parecen a los lenguajes naturales en tanto que ambos, códigos fuente y textos escritos en lenguaje natural, se pueden representar como cadenas de símbolos (caracteres, palabras, etc.)".

La identificación de similitudes de código puede servir para varios propósitos [2], entre los que se puede mencionar el

[1] www.investigacion.frc.utn.edu.ar/mslabs/

estudio de la evolución del código fuente de un proyecto, detección de prácticas de plagio, detección de prácticas de reutilización, extracción de un fragmento de código para "refactorización" del mismo y seguimiento de defectos para su corrección.

La detección de similitudes de código fuente con fines de reutilización es una tarea laboriosa que puede demandar altos costos de tiempo y dinero, dependiendo del impacto de la tarea de detección según el ámbito de aplicación.

Actualmente existen diferentes aproximaciones para abordar la problemática planteada en el presente proyecto, alguna de ellas se mencionan a continuación:

- Aproximaciones basadas en atributos, donde las métricas se calculan a partir del código fuente y se utilizan para la comparación de los distintos archivos. Un ejemplo sencillo sería: utilizar el tamaño del código fuente (número de caracteres, palabras y líneas) como atributo comparable de tamaño [3], o también el número de variables, el número de funciones, el número de clases, entre otros.
- Aproximaciones basadas en Tokens, en las cuales se convierte el código fuente en una secuencia de "tokens" o marcas, para una posterior evaluación y selección de estas secuencias de tokens según ciertas métricas [4] [5].
- Aproximaciones basadas en la estructura, donde el código fuente es convertido a una representación intermedia interna (IR), la cual es utilizada posteriormente para la comparación [6] [7].

Dentro de este contexto, se propone la construcción de un sistema, basado en un modelo que utiliza técnicas de Aprendizaje Automático Supervisado, comúnmente utilizadas para minería de datos [8], que permita la detección de similitudes de código fuente en los lenguajes de programación Java y Python en base a un corpus que será elaborado especialmente para tal fin.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

La línea de investigación en las que se enmarca el proyecto de modelado para el procesamiento de textos estructurados es el área de inteligencia artificial, más concretamente una sub-especialidad que se denomina computación lingüística. En particular, nos enfocamos en los enfoques basados en aprendizaje automático.

Los desarrollos de esta línea de investigación, lo constituyen, por un lado, las herramientas elaboradas para facilitar el análisis y procesamiento de archivos de textos, en este caso de código fuente, y por el otro, los sistemas de reconocimiento de similitudes entre dos archivos de código fuente.

La innovación del proyecto concierne a los nuevos métodos propuestos para el análisis y procesamiento de textos, como así también a los algoritmos creados para abordar las problemáticas anteriormente mencionadas. Los algoritmos diseñados aprovechan las diferentes características que se pueden aprender de los textos y que son recolectados y creados a partir de las herramientas de procesamiento de textos.

Son múltiples las posibles sub-disciplinas que podrían valerse de los resultados de este proyecto, entre las que podemos destacar a las tareas de recuperación de información, evaluación de las traducciones automáticas [9], evaluación de la calidad de las traducciones, reconocimiento de paráfrasis [10] e implicación de textos [11][12][13][14][15]. Adicionalmente, la creación de corpus sobre texto estructurado es una actividad de relevancia y que puede impactar en otras tareas relacionadas al procesamiento del lenguaje.

La confluencia del trabajo de varias líneas de investigación que se desarrollan en el Laboratorio de investigación de software LIS, lugar donde se lleva a cabo este proyecto, ha llevado a la necesidad de plantear el surgimiento de un nuevo grupo

UTN que nuclee a las siguientes líneas de investigación: teoría de autómatas y gramáticas formales, modelos de predicciones de fenómenos climatológicos, y el modelado de problemas del área de ciencias sociales.

En este contexto, el presente proyecto se haya incluido en la presentación de este nuevo grupo de investigación en UTN. Actualmente, la aprobación de dicho grupo se encuentra en trámite dentro de la UTN.

3. RESULTADOS OBTENIDOS/ESPERADOS

Este proyecto se encuentra en sus primeras etapas y, hasta el momento, se han conseguido los siguientes resultados.

Se ha desarrollado un prototipo que permite seleccionar dos archivos de códigos fuentes de Java o Python, sobre los que se puede aplicar un conjunto de medidas de similitud léxica. A su vez, se está integrando una medida de similitud sintáctica.

Hasta el momento, el sistema puede identificar dos archivos con alto grado de superposición léxica, pero no es capaz de identificar archivos similares que utilizan estructuras sintácticas equivalentes pero que se escriben utilizando sintaxis diferente. Claramente, se debe a la imposibilidad de capturar la información presente a un nivel más profundo que el nivel léxico.

Uno de los inconvenientes que se presentan actualmente es el tiempo necesario para la ejecución de las métricas sobre un conjunto de archivos. Supongamos que se desee obtener los dos archivos más similares de un conjunto. Para ello es necesaria una comparación entre cada par de elementos del conjunto, la cual es una operación muy costosa desde el punto de vista computacional, y es por esto que los algoritmos propuestos como medidas de similitud no solo tienen que ser efectivos, sino también escalables y paralelizables.

Un sistema de detección de similitudes que contemple a todos los elementos de un conjunto de N-elementos requerirá (N*N)/2 comparaciones entre los archivos de código fuente.

Por consiguiente, se están desarrollando algoritmos que sean capaces de modelar de manera precisa el fenómeno de similitud textual a diferentes niveles léxico, sintáctico y semántico, pero al mismo tiempo cumpliendo con el requerimiento no funcional de que debe tratarse de un algoritmo de baja complejidad computacional.

4. FORMACIÓN DE RECURSOS HUMANOS

El equipo de investigación está formado por docentes investigadores del Laboratorio de Investigación de Software LIS[2] del Dpto. de Ingeniería en Sistemas de Información de la Universidad Tecnológica Nacional Facultad Regional Córdoba, se detallan a continuación los responsables del proyecto:

- Una magister en ingeniería en sistemas de información que está evaluando la posibilidad de desarrollar su tema de tesis de doctorado en la esta temática con una variación del enfoque desde el punto de vista de los sistemas de Generación del Lenguaje Natural (NLG), y dirige a los integrantes miembros del equipo.
- Un doctor en ciencias de la computación que desarrolló su tesis de doctoral en la temática de implicación de textos y paráfrasis, y colaborara en la coordinación de becarios.
- Participan del proyecto alumnos que necesitan realizar su práctica supervisada que es uno de los requisitos para la obtención del grado de Ingeniero. Los alumnos que intervienen aprenden a realizar actividades de investigación, y cómo integrarse en un equipo existente.

[2] www.investigacion.frc.utn.edu.ar/mslabs/

- Por año participan uno o dos becarios alumnos a los que se les enseña como trabajar en un proyecto de investigación, y como llevar adelante actividades de investigación.
- Un becario de investigación de posgrado se ha incorporado recientemente, por lo cual el proyecto complementará su formación profesional.

5. BIBLIOGRAFÍA

[1] Frantzeskou, G., MacDonell, S., Stamatatos, E., Gritzalis S. (2008). Examining the significance of high-level programming features in source code author classfication. The Journal of Systems and Software, 81(3):447–460.

[2] Smith, R. y Horwitz, S. (2009). Detecting and Measuring Similarity in Code Clones. International Workshop on Software Clones (IWSC'09), pp. 28-34.

[3] Wise M. (1992). Detection of similarities in student programs: YAP'ing may be preferable to plaguing. In ACM SIGCSE Bulletin, volume 24, pp. 268–271.

[4] Wise M. (1993) Running Karp-Rabin matching and greedy string tiling. Basser Dept. of Computer Science, University of Sydney, Sydney.

[5] Schleimer, S., Wilkerson, D., y Aiken, A. (2003). Winnowing: Local Algorithms for Document Fingerprinting. En: Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, pp. 76-85.

[6] Li, X., y Zhong, X. (2010). The source code plagiarism detection using AST. In International Symposium IPTC, pp. 406–408.

[7] Baxter I., Yahin, A., Moura, L., Sant'Anna, M., y Bier, L. (1998). Clone detection using abstract syntax trees. En Proceedings de IEEE ICSM 1998, pp. 368–377.

[8] Jadon, S. (2016). Code clones detection using machine learning technique: Support vector machine. 2016 International Conference on Computing, Communication and Automation (ICCCA). IEEE. Noida, India.

[9] FeldmanR. y Hirsh H.. Exploiting Background Information in Knowledge Discovery from Text. Journal of Intelligent Information Systems, 1996.

[10] Lewis, D.. Evaluating and optmizing autonomous text classification systems. In Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval. Seattle, US, págs. 246-254, 1995.

[11] M. Craven y J. Shavlik. Using Neural Networks for Data Mining. Future Generation Computer Systems, 13, págs. 211-229, 1997.

[12] Castillo J. An approach to Recognizing Textual Entailment and TE Search Task using SVM. Procesamiento del Lenguaje Natural 44, 139-145, 2010. 4, 2010.

[13] I. Goodfellow, Y. Bengio y A. Courville. Deep Learning. MIT Press. 2016.

[14] Castillo J. Using Machine Translation Systems to Expand a Corpus in Textual Entailment. Proceedings of the Icetal 2010, LNCS, vol. 6233, pp.97-102, 2010.

[15] Castillo J., Cardenas M. Using Sentence Semantic Similarity Based on WordNet in Recognizing Textual Entailment. Iberamia 2010, LNCS, vol. 6433, pp. 366-375, 2010.