Análisis de textos con estructura

Marina Cardenas, Julio Castillo

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

{ing.marinacardenas, jotacastillo}@gmail.com

RESUMEN

Este artículo describe un proyecto de investigación relacionado con el análisis de textos que presentan cierta estructura, en particular se centra en el procesamiento de archivos de código fuente en un lenguaje de programación determinado. El proyecto aborda el problema de la detección de reutilización de código y presenta otras aplicaciones derivadas, como la detección de plagio en el código utilizado.

Se utilizan técnicas adaptadas del área lingüística computacional, y de una subtarea que se denomina implicación textual que es definido originalmente para textos sin estructura subyacente.

El proyecto se encuadra dentro de una línea de investigación en aprendizaje automático por computadora.

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

CONTEXTO

En este artículo se presenta el proyecto denominado *Modelado para el procesamiento de textos estructurados*, el cual se trata de un proyecto acreditado por la Secretaria de Ciencia y Tecnología de la UTN con código: UTN4518.

El mismo aborda temáticas específicas del área de lingüística computacional, inteligencia artificial y compiladores. Su desarrollo se lleva 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.

Este proyecto se encuentra dentro del grupo formal UTN denominado GA2LA: Grupo de Aprendizaje Automático, Lenguajes y Autómatas y del LIS[1].

El grupo GA2LA [1] reúne a diversos proyectos de investigación relacionados con redes neuronales, gramáticas y lenguajes de programación, como así también de calidad y trazabilidad en el desarrollo de aplicaciones informáticas. A su vez, trabaja desde un abordaje práctico en colaboración con profesionales de ciencias sociales, a los efectos de poder resolver problemáticas de salud comunitaria desde un enfoque de aprendizaje de computadoras y de la construcción de modelos teóricos computacionales.

Este grupo está compuesto por los integrantes de los diferentes proyectos que nuclea, y que incluye a docentes investigadores, doctores, pasantes y becarios.

1. INTRODUCCIÓN

El proyecto *Modelado para el procesamiento de textos estructurados* tiene como objetivo construir modelos y aplicaciones software que permitan determinar las similitudes a nivel de archivos de códigos fuentes, utilizando para ello diversas técnicas de procesamiento del lenguaje natural y de minería de textos en

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

general [2] . Debido a ciertas similitudes de los lenguajes de programación con los lenguajes naturales, y al ser objeto de estudio dentro de la clasificación general de lenguajes y gramáticas atribuidas a Noam Chomsky, es tomado como punto de partida para integrar conocimientos de áreas de compiladores con otros de lingüística computacional [3] [4].

La utilidad de la detección de similitudes en códigos fuentes es basta [5], y entre algunas aplicaciones podemos mencionar: la evolución de un archivo de código fuente, y de un proyecto de desarrollo de software en general, la detección de código reutilizado en un mismo proyecto (útiles al momento de la "refactorización" del código y para el seguimiento de defectos), la detección de prácticas de plagio, entre otras [6] [7].

En el proyecto abordamos la problemática de detección de similitudes de código fuente con fines de reutilización, que de ser hecha manualmente requeriría de altos costos debido a lo laborioso de dicha actividad, en particular con proyectos de escala mediana - grande.

En la bibliografía pueden encontrarse tres tipos común de abordajes a esta tarea, que las describimos brevemente 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. Por ejemplo, se puede utilizar el tamaño del código fuente (número de caracteres, palabras y líneas) como atributo comparable de tamaño [6], o también el número de variables, el número de funciones, el número de clases, entre otros atributos.
- Aproximaciones basadas en Tokens, en las cuales se convierte el código fuente en una secuencia de "fragmentos", para una evaluación posterior y luego la selección de estas secuencias de tokens según ciertas métricas [7] [8] [9].
- Aproximaciones basadas en la estructura, en este caso, el código fuente es convertido a una representación intermedia interna (IR), que luego es la que se utiliza para la comparación [10] [11].

Dentro de este contexto, en este proyecto se utiliza una combinación de los dos primeros enfoques, que permitan la detección de similitudes de código fuente en los lenguajes de programación Java y Python en base a un corpus elaborado específicamente 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 lingüística computacional.

En particular, nos centramos en enfoques basados en el aprendizaje automático. Los desarrollos en esta línea de investigación están constituidos, por un lado, por las herramientas desarrolladas para facilitar el análisis y procesamiento de archivos de texto, en este caso código fuente, y por otro, los sistemas de reconocimiento de similitudes entre dos archivos de código fuente.

La innovación del proyecto radica en los nuevos métodos propuestos para el análisis y procesamiento de textos, así como a los algoritmos creados para abordar los problemas antes mencionados. Los algoritmos diseñados aprovechan las diferentes características que se pueden aprender de los textos y que se recopilan y crean a partir de las herramientas de procesamiento de textos.

En cuanto a las posibles aplicaciones de los resultados, los mismos pueden ser de utilidad en tareas de extracción de información [12] [13], evaluación de las traducciones automáticas, evaluación de la calidad de las traducciones [14], reconocimiento de paráfrasis [15] [16] e implicación de textos [17][18][19][20].

3. RESULTADOS OBTENIDOS/ESPERADOS

En esta sección mencionaremos los resultados obtenidos hasta el momento, en el proyecto que ha concluido su primera etapa, pero que continúa en desarrollo.

Se ha desarrollado una aplicación web que permite realizar la comparación entre archivos de códigos fuentes (en múltiples lenguajes de programación), brindando una métrica de similitud entre dos archivos de código fuente en un lenguaje específico.

La aplicación web brinda la posibilidad de la ejecución "en lotes" de un conjunto grande de archivos. Esta actividad es muy útil cuando se necesita comparar un archivo contra un conjunto a los efectos de identificar una copia (plagio) o bien archivos que representen reutilización de código fuente.

Los archivos de códigos fuentes permitidos en la comparación son por el momento:

- .java, del lenguaje JAVA
- .cpp, del lenguaje de Programación C++
- .c, del lenguaje de Programación C.

Adicionalmente esta aplicación genera reportes con las salidas de las ejecuciones realizadas, que permitirán al decisor identificar rápidamente un subconjunto de archivos que presenten gran similitud entre sí, dentro de un conjunto grande de archivos.

Se está trabajando en el tiempo de comparación, el cual se torna demasiado costoso cuando se quieren comparar decenas de archivos, dependiendo de su tamaño en kilobytes y de las características con las que se lo desee comparar.

Para lograr las clasificaciones de los archivos se trabaja principalmente (por el momento) a nivel léxico con una exploración sintáctica superficial.

Otro de los resultados obtenidos hasta el momento consiste en un corpus de pares de archivos de código fuente. Se trata de un corpus de texto estructurado, realizado manualmente y requirió un considerable tiempo de clasificación, de dos etiquetadores por el lapso de dos años.

Este corpus fue generado con el Programa Asistente de Creación de Corpus v3 [21] que es producto de otro proyecto de investigación de procesamiento de textos [22].

Si bien inicialmente se pensaba en la construcción de un Corpus de 1000 pares, se pudo construir un corpus con 300 pares de códigos fuentes en lenguaje Java en los que se los ha etiquetado en base a las características presentes en los mismos.

Este corpus sirve de material de entrenamiento para el sistema. Actualmente se encuentra en un trabajo en progreso y se está entrenando el sistema con este conjunto de entrenamiento.

Como trabajos futuros mencionamos la necesidad de modelar el fenómeno de similitud textual en base a diferentes niveles de abstracción, que involucre características a nivel léxico, sintáctico y semántico, cuidando al mismo tiempo de no incrementar su complejidad computacional para que pueda ser utilizado en la práctica y no como un mero desarrollo teórico.

4. FORMACIÓN DE RECURSOS HUMANOS

El proyecto físicamente se lleva adelante en Laboratorio de Investigación de Software LIS[2] del Dpto. de Ingeniería en Sistemas de Información de la Universidad Tecnológica Nacional Facultad Regional Córdoba, en el cual realizan su labor diaria docentes-investigadores, becarios y pasantes de diferentes proyectos de investigación.

En cuanto a las personas que constituyen este proyecto mencionamos:

- Un doctor en ciencias de la computación formado en el área de lingüística computacional y que tiene experiencia en el área de detección y generación de paráfrasis, y en la

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

detección de implicaciones textuales. Entre las actividades que realiza se encuentra la formación de becarios, pasantes y la dirección de investigadores.

- Un doctorando en ingeniería con mención en sistemas de información cuyo tema de tesis se relaciona con el proyecto actual. A su vez, realiza la dirección de becarios, pasantes, y otros miembros del equipo.

- Participan del proyecto alumnos que necesitan realizar su práctica supervisada que es uno de los requisitos para la obtención del grado de Ingeniero. Su asignación depende de las actividades del proyecto. Siguen un plan de trabajo y colaboran con los demás integrantes del proyecto.

- Participa un becario de posgrado como parte de un programa de formación inicial de investigadores en proyectos homologados en UTN.
- Participan anualmente, uno o dos becarios alumnos a los que se les asigna tareas específicas del proyecto, se les enseña a trabajar en el contexto de un proyecto de investigación, y se les enseña la metodología de investigación científica.

5. BIBLIOGRAFÍA

[1] Juan C Vázquez, Julio J Castillo, Leticia Constable, Marina E Cardenas. (2018). XX Workshop de Investigadores en Ciencias de la Computación (WICC 2018, Universidad Nacional del Nordeste).

[2] Cardenas, Marina E., Castillo, Julio

J. Procesamiento de textos estructurados. (2018). XX Workshop de Investigadores en Ciencias de la Computación (WICC 2018).

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

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

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

[6] 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.

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

[8] 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.

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

[10] 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.

[11] 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.

[12] Feldman, R. y Hirsh, H.. (1996). Exploiting Background Information in Knowledge Discovery from Text. Journal of Intelligent Information Systems.

[13] Lewis, D. (1995). Evaluating and optimizing 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.
[14] Julio Castillo and Paula Estrella. (2012). Semantic textual similarity for MT evaluation. In Proceedings of the Seventh Workshop on Statistical Machine Translation, pages 52–58, Montr´eal, Canada.
[15] Castillo J., Cardenas M. (2010). Using Sentence Semantic Similarity Based on WordNet in Recognizing Textual Entailment. Iberamia, LNCS, vol. 6433, pp. 366-375, 2010.
[16] Castillo J. (2010). An approach to Recognizing Textual Entailment and TE Search Task using SVM. Procesamiento del Lenguaje Natural 44, 139-145, 2010. 4.
[17] Castillo J. (2010). Using Machine Translation Systems to Expand a Corpus in Textual Entailment. Proceedings of the Icetal 2010, LNCS, vol. 6233, pp.97-102.

[18] Julio J. Castillo. (2010). Recognizing textual entailment: experiments with machine learning algorithms and RTE corpora. Special issue: Natural Language Processing and its Applications, Research in Computing Science, 46:155–164.

[19] Castillo, J. J. (2010). Textual entailment search task: An initial approach based on coreference resolution. Intelligent Computing and Cognitive Informatics, International Conference on: 388-391.
[20] Castillo, J. J. (2010). Sagan in TAC2010: A machine learning approach to RTE within a corpus. In Proceedings of the Text Analysis Conference (TAC'10).

[21] Julio J Castillo, Marina E Cardenas, Martin Navarro, Nicolás A Hernández, Melisa Velazco. (2018). Sistemas de análisis textual en formato no estructurado. XX Workshop de Investigadores en Ciencias de la Computación (WICC 2018, Universidad Nacional del Nordeste).

[22] Julio J Castillo, Marina E Cardenas, Adrián Curti, Osvaldo Casco, Martin Navarro, Nicolás A Hernández, Melisa Velazco. (2017), "Desarrollo de sistemas de análisis de texto," in XIX Workshop de Investigadores en Ciencias de la Computación, 2017, pp. 58–62.