Detección de plagio con adversarios

Taquías Daniel, Villagra Andrea

Laboratorio de Tecnologías Emergentes (LabTEm)
Unidad Académica Caleta Olivia – Universidad Nacional de la Patagonia Austral
Ruta 3 Acceso Norte s/n. (9011) – Caleta Olivia – Santa Cruz – Argentina
danieltaquias@hotmail.com
avillagra@uaco.unpa.edu.ar

Errecalde Marcelo
Laboratorio de Investigación y Desarrollo en Inteligencia Computacional (LIDIC)
Departamento de Informática – Universidad Nacional de San Luis
Ejercito de los Andes 950 – Local 106 (5700) – San Luis - Argentina
merreca@unsl.edu.ar

Resumen

La categorización automática de textos (CAT) es una tarea de gran relevancia debido a la necesidad de procesar y analizar volúmenes cada vez mayores de información textual disponible en la Web y en las empresas. La CAT, usualmente implementada mediante técnicas de aprendizaje automático (AA), se ha convertido en una de las pocas alternativas realistas al análisis manual de documentos, con resultados de alto impacto en áreas como la detección de *spams*, filtrado de noticias, detección de plagios, análisis de opinión y tendencias, organización de páginas web, determinación del perfil de autores, etc. Sin embargo, en muchos dominios como la detección de spam, fraudes, plagio, intrusos, etc., la CAT debe enfrentar un problema cada vez más frecuente: la presencia de adversarios, personas o sistemas automáticos, que deliberada y maliciosamente intentan detectar y explotar las vulnerabilidades propias de los sistemas de CAT en beneficio propio. Esta área de investigación, denominada categorización con adversarios ha sido aplicada en enfoques tales como la detección de spams, fraude, intrusos en una red, virus, etc. pero, hasta donde sabemos, no ha sido analizada en el contexto de los sistemas de detección de plagio. Esta nueva área, que surge de la intersección de la categorización con adversarios y la detección de plagios, la denominaremos detección de plagio con adversarios y constituye el eje central de esta línea de investigación.

Palabras clave: Plagio, Detección de plagio, Categorización de textos,

Contexto

Esta línea de investigación se desarrolla en el Laboratorio de Tecnologías Emergentes (LabTEm) en el marco del programa de Investigación en Ciencia y Tecnología de la Universidad Nacional de la Patagonia Austral.

Introducción

En el presente trabajo se describe una línea de investigación en el campo de la categorización automática de textos (CAT), más precisamente en el área de detección de plagio con adversarios. En este sentido, se define como 'adversario' a la persona o herramienta informática que, en forma deliberada o maliciosa, intenta detectar o explotar las fallas o vulnerabilidades que estos sistemas de detección poseen para realizar plagio sin ser descubiertos. Para esto, se estudian los sistemas de detección de plagio que existen actualmente y se trata de descubrir las fallas o debilidades que poseen para analizar la presencia de adversarios.

La CAT cumple un rol fundamental dentro de la minería de textos y de la web, y adquiere cada día mayor relevancia debido a la necesidad de procesar y analizar, enormes volúmenes de información textual disponible en la web y en las empresas. En este contexto, la CAT y el aprendizaje automático (AA), se han convertido en una de las pocas alternativas realistas al análisis manual de documentos.

Esto ha quedado de manifiesto en un número significativo de trabajos donde se ha aplicado la CAT para la detección de spams [14], filtrado de noticias [19], detección de plagios e identificación de autores [10, 15], análisis de opinión y sentimientos [9, 10, 13], organización de patentes en categorías [11], clasificación y organización de páginas web [6], detección de 'bullying' y pedófilos en la web [5, 16], determinación del perfil del autor (sexo, grupo etario, nacionalidad, etc.) [3, 12], identificación de información de calidad [20], etc.

Los puntos fuertes de los AA son su adaptabilidad y su capacidad de inferir patrones que pueden ser utilizados en predicciones o en la toma de decisiones [17]. Sin embargo, este AA pueden ser alterados por la manipulación contradictoria del entorno y quedarían expuestas sus técnicas y aplicaciones a una nueva clase de vulnerabilidades. El alcance de estas amenazas tienen que ser evaluadas, las técnicas de aprendizaje deben ser seleccionadas, para minimizar el impacto del adversario, y evitar las fallas en estos sistemas.

La posibilidad de realizar plagio, denominando a este término como el uso no reconocido de la obra original de otro autor [1], es cada vez más tentadora en estos tiempos. La facilidad que se tiene actualmente para acceder a la información a través de las nuevas tecnologías que existen, la gran cantidad de información que se puede encontrar en las bases de datos de la web, hace que los estudiantes recurran a estas malas costumbres para poder obtener mejores calificaciones o incorporar material superior en sus documentos. Las razones más comunes que tienen los estudiantes para inclinarse a cometer plagio van desde la ignorancia del uso de citas y referencias hasta la seguridad de que no van a ser descubiertos.

Las formas de plagio [2] son varias: copiar en forma textual palabras, frases o pasajes de un texto publicado sin citarlo correctamente, utilizar un concepto, idea u opinión que no es de conocimiento común, cambiar la gramática o reordenar las frases de una obra original, reiterar el mismo contenido cambiando el orden de las palabras, reemplazar las palabras por sinónimos, sustituir oraciones largas por cortas, añadir citas o referencias sin proporcionar información de la fecha de los enlaces a las fuentes originales o añadir referencias incorrectas, etc.

El método más común utilizado por los sistemas de detección automática de plagio [4], es el uso de algoritmos de correspondencia de cadenas que permiten

la comparación del contenido de un documento. A su vez, estos se pueden clasificar en sistemas externos e intrínsecos [3, 5, 7]. Los sistemas externos tratan de identificar casos de plagio fuera del conjunto de ensayos que se está analizando, y basándose en el uso de un servicio de búsqueda en la Web, y los sistemas intrínsecos son los que examinan únicamente las colecciones locales de documentos.

Usualmente se asume que los patrones de detección de plagio aprendidos por los sistemas de AA, mantendrán su validez y efectividad a lo largo del tiempo. La realidad en cambio muestra que las personas que realizan plagio, al igual que en otros dominios 'con adversarios' [17], adaptan sus estrategias para burlar a los sistemas de CAT. Esto queda de manifiesto en artículos como [18], donde se muestra las técnicas específicamente desarrolladas por los estudiantes que, conociendo el funcionamiento de los sistemas de detección de plagio disponibles a la fecha, permiten realizar plagio sin ser detectados.

Nuestra hipótesis por lo tanto, es que los métodos de detección de plagio son vulnerables a los ataques de 'adversarios' que pueden perjudicar su funcionamiento y que se verían beneficiados por un enfoque 'basado en adversarios'. Esta nueva área, que surge de la intersección de la categorización con adversarios y la detección de plagios, la denominaremos detección de plagio con adversarios y constituye el eje central de esta línea de investigación.

Líneas de Investigación, Desarrollo e Innovación

Con este trabajo se quiere demostrar que los métodos de detección de plagio son vulnerables a los ataques de 'adversarios', que se puede perjudicar su funcionamiento, y que los sistemas de detección de plagio actuales se verían beneficiados con este enfoque 'basado en adversarios'. Se trata de evaluar el grado en que los sistemas de detección de plagio existentes son capaces de detectar el plagio. Un sistema de detección debe ser capaz de detectar todos los casos y estilos de plagio [18], algunos de estos métodos de plagio son más fáciles de detectar que otros.

Los estudiantes que conocen la forma en que funcionan los sistemas de detección de plagio, usan una serie de trucos técnicos, como se explica en [18], deliberadamente con el fin de engañar los sistemas. Los sistemas existentes se limitan a la detección de estos tres tipos de plagio, dejando a un lado la detección del uso incorrecto o intencional de referencias, porque esto requeriría el uso de herramientas automáticas sofisticadas de seguimiento de citas y de referencias, algo que no está todavía disponible. El tipo de plagio entre idiomas [8] es extremadamente difícil de detectar y encontrar de forma automática y es muy probable que lo siga siendo en el futuro.

Además de ser capaz de detectar con precisión los tipos de plagio enumerados anteriormente, un sistema de detección debe evitar falsas detecciones (es decir, no devolver falsos positivos). Dado que el volumen de textos en Internet es tan grande, es posible que partes del texto de un estudiante casualmente pueda parecerse al de una página web existente (a pesar de que el estudiante nunca pudo haber visto la página web en cuestión). Este tipo de semejanza se refiere a menudo como 'similitud casual'.

El conocimiento de los trucos técnicos [18], para evadir a los principales sistemas de detección de plagio como insertar caracteres de aspecto similar al de alfabetos extranjeros, insertar letras

invisibles de color blanco o la inserción de páginas escaneadas como imágenes en un documento son algunos ejemplos de estos engaños que se le hacen a los sistemas de detección de plagio.

Por esta razón, se estudian los sistemas de detección de plagio que existen actualmente y se trata de descubrir fallas o debilidades que poseen para analizar la presencia de adversarios.

Resultados y Objetivos

El objetivo principal de esta línea de investigación es poder analizar el comportamiento de los sistemas de detección de plagio en un contexto con adversarios. Como parte de este objetivo general, podemos identificar los siguientes objetivos específicos:

1. Lograr un entendimiento más profundo de los aspectos principales vinculados a esta línea de trabajo:
   a) clasificación con adversarios,
   b) detección de plagio,
   c) enfoques que reflejen el estado del arte en la clasificación automática de documentos en general y de documentos con plagio en particular.

2. Analizar críticamente potenciales amenazas de seguridad en la detección de plagio, y el grado en que se pueden evaluar estas amenazas.

3. Evaluar al menos un sistema del mundo real utilizado en la detección de plagio, analizar sus vulnerabilidades, mostrar ataques del mundo real contra sus mecanismos de AA y CAT y proponer defensas que puedan mitigar exitosamente la efectividad de estos ataques.

Como resultado final, se espera lograr una mayor comprensión teórica y práctica de las limitaciones y vulnerabilidades que los sistemas de detección de plagio exhiben en la actualidad, y de las formas en que esas falencias podrían ser solucionadas o aliviadas. Es importante remarcar que la detección de plagio con adversarios no ha sido un área desarrollada aún en el ámbito científico nacional ni internacional por lo que la contribución de este trabajo puede ser significativa, y abrir nuevas líneas de investigación.

Formación de Recursos Humanos

Un integrante ha comenzado su Maestría orientando sus cursos y trabajos a esta línea de investigación.

Referencias

[1], Hermann Maurer, Frank Kappe, and Bilal Zaka Plagiarism. A Survey. Journal of Universal Computer Science, vol. 12, no. 8, 1050-1084, 2006.

[2] Paul Clough. Old and new challenges in automatic plagiarism detection. National UK Plagarism Advisory Service, 2003.

[3] Benno Stein, Nedim Lipka, and Peter Prettenhofer. Intrinsic Plagiarism Analysis. Language Resources and Evaluation (LRE), 45 (1): 63-82, 2010.

[4] Enrique Vallés Balaguer. Empresa 2.0: Detección de plagio y análisis de opiniones. Master's thesis, Universidad Politécnica de Valencia, 2011.

[5] Potthast M., Stein A., Eiselt A., Barrón-Cedeño A., Rosso P. Overview of the 1st International Competition on Plagiarism Detection, 2009.

[6] Xiaoguang Qi and Brian D. Davison. Web page classification: Features and algorithms. ACM Comput. Surv., 41(2):12:1–12:31, February 2009.

[7] Benno Stein, Nedim Lipka, and Peter Prettenhofer. Intrinsic plagiarism analysis. Language Resources and Evaluation, 45(1):63–82, 2011.

[8] Potthast M., Barrón-Cedeño A., Stein B., Rosso P. Cross-Language Plagiarism Detection. Languages Resources and Evaluation, 2010.

[9] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2):1–135, January 2008.

[10] Gabriel Oberreuter, Gaston LHuillier, Sebastián A. Ríos, and Juan D. Velásquez. Approaches for intrinsic and external plagiarism. In CLEF (Notebook Papers/Labs/Workshop), 2011.

[11] C. J. Fall, A. Törcsvári, K. Benzineb, and G. Karetka. Automated categorization in the international patent classification. SIGIR Forum, 37(1):10–25, April 2003.

[12] Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. "How old do you think I am?" a study of language and age in twitter. In Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, 2013.

[13] Pollyanna Gonçalves, Matheus Araújo, Fabrício Benevenuto, and Meeyoung Cha. Comparing and combining sentiment analysis methods. In Proceedings of the First ACM Conference on Online Social Networks, COSN '13, pages 27–38, New York, NY, USA, 2013.

[14] Thiago S. Guzella and Walmir M. Caminhas. A review of machine learning approaches to spam filtering. Expert Systems with Applications, 36 (7):10206–10222, September 2009.

[15] Alberto Barrón-Cedeño Benno Stein Martin Potthast, Andreas Eiselt and Paolo Rosso. Overview of the 3rd international competition on plagiarism detection. Notebook Papers of CLEF 2011 Labs and Workshops, 2011.

[16] India Mcghee, Jennifer Bayzick, April Kontostathis, Lynne Edwards, and Emma Jakubowski. Learning to identify internet sexual predation. International Journal on Electronic Commerce, 15(3):103–122, 2011.

[17] Blaine Nelson. Behavior of Machine Learning Algorithms in Adversarial Environments. PhD thesis, EECS Department, University of California, Berkeley, Nov 2010.

[18] Tuomo Kakkonen and Maxim Mozgovoy. Hermetic and web plagiarism detection systems for student essays–an evaluation of the state-of-the-art. Journal of Educational Computing Research, 42(2):135–159, 2010.

[19] Ken Lang. Newsweeder: Learning to filter netnews. In Proceedings of the 12th International Conference on Machine Learning, pages 331–339, 1995.

[20] Elisabeth Lex, Michael Voelske, Marcelo Errecalde, Edgardo Ferretti, Leticia Cagnina, Christopher Horn, Benno Stein, and Michael Granitzer. Measuring the quality of web content using factual information. In Proceedings of the 2Nd Joint WICOW/AIRWeb Workshop on Web Quality, WebQuality '12, pages 7–10, New York, NY, USA, 2012. ACM.