

Primeras Experiencias en Detección de Plagio en el Ambiente Educativo

Bordignon, Fernando R.A., Tolosa, Gabriel H., Rodriguez, Carlos G. y Peri, Jorge A.
{bordi, tolosoft, crodriguez, peri}@unlu.edu.ar

Universidad Nacional de Luján
Departamento de Ciencias Básicas
Laboratorio de Redes de Datos
+54-2323-423064

Resumen

Desde los últimos años es significativa la proliferación de sitios web dedicados a la provisión gratuita o paga de trabajos estudiantiles – a medida o no –, de exámenes, de trucos para copiarse en evaluaciones y demás. En Internet, en particular en el espacio web, la posibilidad de obtener grandes volúmenes de información provenientes de páginas personales, bibliotecas, enciclopedias, bases de datos, etc. para casi cualquier tema escolar ha acercado a nuestros alumnos una riqueza informativa incommensurable. Esta situación la perciben como una “solución” en la confección de sus tareas estudiantiles. Sin embargo, el uso inapropiado de los recursos digitales no solamente no constituye solución alguna sino que – en algunos casos – se transforma en una actividad de deshonestidad estudiantil.

En este documento se presentan los primeros resultados de un método propio de detección de pasajes similares de texto. Esta línea de trabajo persigue como objetivo a largo plazo la construcción de una plataforma de software institucional que permita a los profesores el control de plagio sobre los trabajos recibidos de sus alumnos.

Introducción

A lo largo de la vida académica del alumno es común que los profesores les encomienden la confección de trabajos prácticos y monografías sobre la más diversa temática. Tales actividades tienden a que el estudiante reflexione, ejercite y desarrolle habilidades que le permitan apropiarse de los conocimientos impartidos.

Un tema relacionado con lo planteado anteriormente, que se ha incrementado de acuerdo a la alta disponibilidad de contenidos digitales es el engaño estudiantil. Algunas de las formas de engaño más comunes son la copia o parafraseo de trabajos de terceros de forma parcial o total sin citar la fuente, la copia de trabajos entre pares, la adquisición de trabajos hechos y la descarga y copia de trabajos desde Internet sin citar la fuente. En estas prácticas de deshonestidad académica el alumno solo logra engañarse a sí mismo, saltando pasos necesarios para su normal formación.

Algunos antecedentes de tales prácticas se citan a continuación. En Colombia [1], una universidad detectó que tres de sus alumnos de postgrado cometieron plagio en un trabajo de una asignatura. Lo destacado de la noticia es que los estudiantes en cuestión son políticos del orden de diputados y alcaldes. En el año 2001, en la Universidad de Virginia el profesor Bloomfield [2] expuso una situación de fraude académico de significativa importancia: alrededor de 160 alumnos fueron acusados de plagio en sus trabajos.

Un destacado investigador sobre honestidad estudiantil, el profesor Donald McCabe – fundador del CAI, Center for Academic Integrity – a través de distintos trabajos de campo [3] reveló los

siguientes datos: En el año académico 2000-2001 se encuestaron 2,294 estudiantes norteamericanos de *High School Juniors* de 25 instituciones educativas públicas y privadas. Algunos de los indicadores más alarmantes son que – en promedio – el 63% reportaron que una o más veces se copiaron en exámenes, el 76% declaró no haber recibido ayuda en la confección un trabajo práctico habiéndola tenido y el 52% reconoció haber copiado oraciones de algún sitio web sin la debida cita.

En otro trabajo de McCabe [4], se pidió la opinión sobre Internet como fuente de información a 2,200 estudiantes de *High School* y *College* de 21 campus diferentes en el año académico 1999-2000. Sus resultados son los siguientes:

	Estudiantes que informaron tal comportamiento		Estudiantes que piensan que este comportamiento es serio	
	High School	College	High School	College
Plagio de fuentes escritas				
Copia textual y envío como trabajo propio	34%	16%	70%	70%
Copia de algunas oraciones sin cita	60%	40%	39%	35%
Plagio en Internet				
Envío de un trabajo obtenido de una “fábrica de monografías” o un sitio web	16%	5%	74%	72%
Copia de algunas oraciones de un sitio web sin cita	52%	10%	46%	68%

El plagio ha sido un problema clásico en las instituciones académicas. Antes de Internet estuvo contenido por la disponibilidad de fuentes al alcance de los alumnos. Típicamente, se obtenía información de bibliotecas, archivos de diarios, publicaciones periódicas y documentos realizados por compañeros de estudio. Con la llegada de los soportes de almacenamiento masivo portables y la disponibilidad de las comunicaciones se logró un acceso instantáneo a grandes espacios de información. Por ende, la actividad denominada “copie y pegue” se incrementó notablemente y con ésta la posibilidad de plagio.

Lo expresado en el párrafo anterior puede resumirse en la frase de Posnick [5] *‘La tecnología informática ha hecho del engaño una actividad tan fácil la cual es una tentación para estudiantes que de no tenerla hubieran sido honestos’*. Bajo otro punto de vista, la profesora Ryan [7] declara que *“Internet es un recurso útil para los que realizan plagio, pero también una excelente herramienta a utilizar contra ellos”*.

Fábricas de monografías

A partir de mediados de la década de 1990, en Internet han empezado a aparecer empresas que asisten al plagio. En tales sitios (denominados *paper mills*) se ofrecen – en forma gratuita o paga – catálogos de trabajos de la más amplia temática. Algunos ejemplos de estos sitios son www.schoolsucks.com, www.bignerds.com y www.lazystudents.com. Para recursos en español existen www.elrincondelvago.com y www.monografias.com.

Los costos de provisión de documentos son variados. Desde alrededor de diez dólares se puede obtener una monografía con bibliografía actualizada de cuatro páginas con cinco citas. Por otro lado, cuando se trata de documentos especialmente redactados para un cliente específico, se pagan alrededor de diez dólares por página.

La informática al auxilio de los docentes

En la búsqueda de plagio en textos escolares los docentes pueden utilizar tres tipos de recursos para su detección:

- Motores de consulta: A partir de sospecha de plagio o parafraseo un docente puede extraer oraciones o n-gramas y enviarlos a un motor de consulta de gran cobertura – como son Google, Yahoo o Altavista – y analizar las respuestas pertinentes.
- Aplicaciones remotas: El sistema de detección de plagio denominado *Plagiarism Advisory Service*¹ asiste a universidades inglesas en la tarea de detección de plagio. Cada trabajo enviado por los docentes es cotejado con documentos de una base de datos que contiene páginas web, trabajos monográficos, trabajos de investigación, diccionarios y enciclopedias, entre otros recursos. También existe la base de datos privada denominada Turnitin² donde las entidades educativas pueden suscribirse a un servicio externo de recepción y control de plagio de trabajos de alumnos
- Aplicaciones de escritorio: El software Glatt³ intenta detectar estilos de escritura semejantes basándose en la premisa que, analizando la escritura de una persona, se puede lograr una firma que la identifique. Por otro lado, CopyCatch⁴ es un programa que localiza semejanzas sintácticas entre pares de documentos, utilizando técnicas de análisis de frecuencia de términos. Wcopyfind⁵ es otro software de uso personal destinado a la detección de plagio desarrollado por el profesor Bloomfield en la Universidad de Virginia.

Independientemente de la herramienta utilizada, la decisión de plagio es tomada por el docente o directivo de la institución y no por el software de detección. Las aplicaciones se limitan – únicamente – a buscar pasajes de texto similares entre el documento que se está analizando y los documentos existentes en una base de datos, o bien, a determinar si el estilo de escritura del autor coincide con el del texto entregado.

Método de detección de plagio basado en bigramas

A los efectos de atacar la problemática de la detección de plagio en trabajos presentados por alumnos se propone un método que permite encontrar similitudes entre pasajes de texto. La base del mismo se encuentra en la descomposición de las oraciones de un texto en bigramas de palabras.

El método consta de dos fases. En la primera, se procesan todos los documentos que constituyen la base de datos textual o repositorio formada por diferentes fuentes: trabajos anteriores de estudiantes, tesis, monografías, capítulos de libros, entre otros. El resultado es un índice invertido donde cada entrada representa un bigrama con una lista asociada de documentos y oraciones donde aparece. La

¹ <http://jisc.northumbrialearning.co.uk/>

² <http://www.plagiarism.org/>

³ <http://www.plagiarism.com/>

⁴ <http://www2.warwick.ac.uk/elearning/tools/plagiarism/copycatch/>

⁵ <http://plagiarism.phys.virginia.edu/Wsoftware.html>

segunda fase consiste en la búsqueda de pasajes similares entre un documento dado la base de datos. Para ello, se computa una métrica de semejanza a partir de obtener – mediante una técnica de filtrado basada en bigramas – aquellas oraciones sospechosas por ser similares a algunas contenidas en la base de datos. A continuación se presentan los pasos principales que componen el algoritmo de armado del índice de la base de datos:

```

Para cada documento en la base de datos
  Eliminar palabras vacías
  Normalizar caracteres
  Llevar texto a minúsculas
  Para cada oración o del texto
    Extraer sus bigramas de palabras y almacenarlos en un
    índice indicando archivo y número o de oración
    (posting list).
  Fin-para
Fin-para

```

Luego, por cada documento que se quiera contrastar se realiza el siguiente procedimiento:

```

Eliminar palabras vacías
Normalizar caracteres
Llevar texto a minúsculas
Para cada oración o del texto
  Extraer sus bigramas de palabras
  Para cada bigrama
    Acceder al índice y recuperar sus entradas
  Fin-Para

  Computar la semejanza de la oración o con respecto a
  cada oración candidata extraída de la base de datos

  Si se supera un umbral u1 de semejanza entonces
    Marcar la oración como sospechosa
  Fin-para

  Si la cantidad de oraciones sospechosas superan un
  cierto umbral u2 entonces
    Indicar el documento analizado como sospechoso

```

Para validar el método se realizó una prueba experimental consistente en tomar un conjunto de 100 documentos provenientes de un sitio de provisión de monografías en castellano. En especial, se tomaron aquellas del tema Internet y Redes. A los efectos de realizar una prueba global se compararon los documentos todos contra todos. El primer resultado obtenido es que el 27% de los documentos posee 5 o más oraciones semejantes a otros con un umbral de un 80%.

En una segunda prueba, con los mismos valores de umbral, se encontraron 23 pares de documentos semejantes. Este corpus se evaluó – además – con el software Wcopyfind [2] el cual arrojó que 22 de los pares anteriormente encontrados son semejantes. Esta prueba confirma la validez y eficacia del método propuesto.

Consideraciones y trabajos futuros

La alta disponibilidad de contenidos digitales ha favorecido ciertas prácticas deshonestas en alumnos de distintos niveles. La bibliografía muestra que existe cierta conciencia sobre la gravedad del problema entre profesores y autoridades educativas de norteamérica y europa. Sin embargo, no se han encontrado documentos ni académicos ni de divulgación sobre esta cuestión en latinoamérica.

A partir de esto resulta necesario estudiar este fenómeno y su impacto en países de la región, en diferentes niveles educativos. Estos resultados permitirán establecer políticas tendientes a manejar situaciones de deshonestidad.

Tomando la base del método propuesto se plantea la construcción de una plataforma de software institucional que permita a los profesores el control de plagio sobre los trabajos recibidos de sus alumnos. Además, se deberán establecer políticas y estrategias técnicas de cooperación horizontal entre las instituciones académicas para compartir las bases de datos de referencia.

El engaño estudiantil impide a los alumnos pensar por sí mismos. Esta pérdida de libertad trae aparejado perjuicios en su educación, ya sea para adquirir nuevos conocimientos o desarrollar habilidades. Los docentes tienen como misión ineludible controlarlos y guiarlos – de forma continua – como parte de su normal formación. Según Stockman [6], *“El engaño no controlado puede ser peligroso y quizás puede pavimentar el camino de los ejecutivos de otras Enrons y WordComs del mundo”*.

Bibliografía

- [1] Artículo, Agencia EFE. “Universidad expulsa a políticos por copia en examen en Bogotá”, diario El Colombiano, 20/8/2003. <http://www.elcolombiano.com/historicod/200308/20030830/nun011.htm>
- [2] Bloomfield, L. Plagiarism Resource Site. Universidad de Virginia, USA. <http://www.plagiarism.phys.virginia.edu/>
- [3] McCabe, D. L. “Cheating - Why Students Do It and How We Can Help Them Stop”. American Educator, pp. 38:43. Winter, 2001.
- [4] McCabe, D. L., Trevino, L.K. y Butterfield, K.D. "Honor Codes and Other Contextual Influences on Academic Integrity". Research in Higher Education, 43, No. 3, pp. 357-378. 2002.
- [5] Posnick-Goodwin, S. "Point, Click, Cheat". California Educator, Volume 4, Issue 2, pp. 14, 1999.
- [6] Stockman, R. “Study Says 74 Percent Admit To Cheating”. TeenSpeakNews - Today for the Leaders of Tomorrow, v.3, n.4, pp 8-9. 2002.
- [7] Ryan, Julie J.C.H. "Student Plagiarism in an OnLine World". Prism Magazine, Nota de Tapa, Diciembre, 1998.