Interfaz de usuario en sistemas de detección de plagio

Diego Campo Millán
School of Computer Science & Information Technology
The University of Nottingham - Jubilee Campus
NG7 1BB Nottingham
Reino Unido
dxc04o@cs.nott.ac.uk

Pedro M. Latorre Andrés
Grupo de Informática Gráfica Avanzada (GIGA)|
Departamento de Informática e Ingeniería de Sistemas
Centro Politécnico Superior
María de Luna, 1
50018 Zaragoza
platorre@unizar.es

Colin Higgins
School of Computer Science & Information Technology
The University of Nottingham - Jubilee Campus
NG7 1BB Nottingham
Reino Unido
cah@cs.nott.ac.uk

Resumen

Este trabajo se centra en la revisión y evaluación de la usabilidad de las interfaces de usuario de los sistemas de detección de plagio para documentos en lenguaje natural y, a partir de ellas, efectúa propuestas para la implementación de la interfaz de un nuevo sistema de detección de plagio desarrollado en la Universidad de Nottingham.

1. Introducción

El plagio es uno de los principales problemas que afectan al reconocimiento y explotación de los derechos de autor. También en el entorno educativo existe una dificultad de distinción de la originalidad de los trabajos, ya que el plagio resulta más fácil hoy día debido al uso extensivo de las tecnologías de la información.

Para la detección del posible plagio es crucial que el evaluador obtenga de manera precisa la información relevante a la originalidad del trabajo analizado en comparación con aquellos otros con los que presenta una proximidad sospechosa.

En la última época se han desarrollado diferentes sistemas y herramientas de aplicación general para detectar el plagio. Quedan fuera de este análisis, por considerarlos de escasa relevancia, los siguientes sistemas: MatchDetectReveal [12], WCopyFind [2], Ferret [8], y Sherlock [13]; asimismo este estudio no considera sistemas en entornos específicos, por ejemplo la detección de plagio de código software MOSS [1], o JPlag [10].

En este documento se estudian los distintos sistemas desde un punto de vista directamente ligado a la interfaz de usuario.

El sistema Verbatim Engine, que está siendo desarrollado en la Universidad de Nottingham, basará su interfaz de usuario en este estudio.

2. Clasificación de los sistemas de detección de plagio

En los últimos años se han realizado diferentes clasificaciones de sistemas de detección de plagio, basadas en diferentes criterios. En este texto se adoptará la propuesta por Fintan Culwin [5], que ofrece las siguientes características diferenciales de los sistemas:

Servidor – Escritorio. Según el sistema requiera infraestructura Cliente – Servidor o no.

Local – Remoto. Según el sistema requiera una estructura en red o no.

Basado en documentos – Basado en corpus. Según el sistema se aplique a un documento o a un conjunto de ellos.

Intracorpus – Intercorpus. Aplicado a un sistema *basado en corpus*, determina si el sistema sólo utiliza información interna al corpus o no.

Gratuito – Comercial. Según el sistema es gratuito o no.

Orientado a base de datos – No orientado a base de datos. Según el sistema utilice bases de datos o no.

Documentos con estilo – Texto plano. Según el documento tenga algún tipo de estilo o sea texto plano.

	TurnItIn	*URKUND*	*CopyCatch Gold*	*PRAISE*	*VAST*	*OrCheck*	*Plagiarism Finder*
Servidor/ Escritorio	Servidor	Servidor	Escrit.	Escrit.	Escrit.	Escrit.	Escrit.
Remoto/ Local	Remoto	Remoto	Local	Local	Local	Remoto	Remoto
Documental / Corpus	Doc.	Doc.	Inter corpus	Intra corpus	2 doc.	Doc.	Doc.
Comercial/ Gratuito	Comercial	Comercial	Comercial	Gratuito	Gratuito	Gratuito	Comercial
Base de datos	Si	Si	No	No	No	Si	Si
Estilos	Si	Si	Si	No	No	No	Si
Código fuente abierto	No	No	No	Sí	Sí	Sí	No
Multilingüe	?	?	Si	No	No	No	No
Despliegue	Web	Mail + web	Local	Applet	Applet	Applet	Local
Código de colores	Si	No	Limitado	Si	No	Si	No
Índice de similitud	Real	No	Real	Real	No	Real	Estimado
Visualización de formato de texto	Se mantiene	No se mantiene	No se mantiene	No hay	Es texto plano	Es texto plano	No se mantiene
Gráficos de apoyo	No	No	No	Sí	Sí	Sí	No
Uso buscador externo	No	No	No	No	No	Sí	Sí (caché)
Distribución en pasos	No	No	No	No	No	No	Sí

Tabla 1. Comparación de las características de los distintos sistemas

Código fuente abierto – Código fuente reservado. Según el código fuente sea libre o no.

Se ha añadido, además, la siguiente característica:

Multilingüe – Monolingüe. Según el sistema use técnicas específicas de varios lenguajes o no.

Este estudio comprende los sistemas: TurnItIn [7], URKUND [14], CopyCatchGold [4], PRAISE [3], VAST [3], OrCheck [3] y Plagiarism Finder [9]. La tabla 1 recoge su clasificación y características de interfaz, que se comentan en la siguiente sección.

3. Interfaces de los sistemas de detección de plagio

En esta sección se examinan las interfaces de los distintos sistemas. Éstas se han clasificado entre gráficas y textuales, según ofrezcan soluciones gráficas de representación de datos o no.

3.1. Interfaces textuales

Se comentan a continuación los tipos de distribución y visualización que no utilizan gráficos para representación de datos.

Distribución en pasos del proceso. Facilita el seguimiento por parte del usuario, y da una mayor impresión de control y de progreso. Sólo Plagiarism Finder lo implementa; CopyCatchGold se aproxima mediante un sistema de solapas, aunque nueve quizá son demasiadas, aplicando la teoría de memoria de corto plazo de Miller [11].

Investigación de varios documentos referidos a un solo texto crítico. TurnItIn ofrece un informe dividido en dos partes (Figura 1).

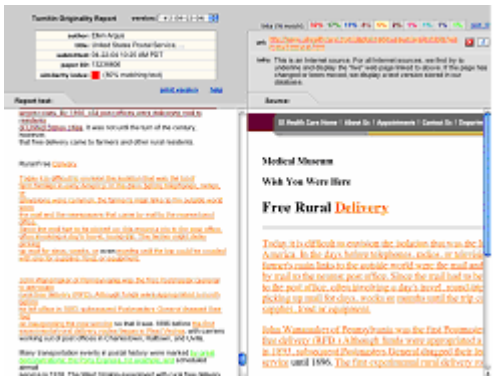

Figura 1. TurnItIn – Informe de Originalidad

En la parte izquierda aparece el texto del documento que se está sometiendo (texto crítico). En la parte derecha aparece un conjunto de posibles fuentes de origen de las partes sospechosas, ordenadas con un sistema de solapas. Las partes del texto crítico con las que se han encontrado similitudes aparecen relacionadas mediante un código de color con el documento fuente, y un índice de similitud entre los documentos, relativo al texto crítico. Se ofrece asimismo un porcentaje del total de texto coincidente. Asimismo, es el único sistema de visualización que mantiene el formato del texto.

En contraposición, en Plagiarism Finder este informe consiste en una página HTML monomarco donde aparece el texto original, y las partes coincidentes aparecen resaltadas en hipertexto enlazadas a la página de caché de Google [6] donde han sido encontradas. URKUND ofrece los resultados en una página HTML algo más sofisticada, con un sistema de cuatro cuadrantes, donde dos de ellos presentan el texto crítico y uno de los documentos sospechosos respectivamente, y los otros dos presentan datos propios de los ficheros. URKUND no ofrece índice de similitud. Ninguno de los dos últimos implementa un código de colores para las coincidencias, y la visualización y navegación entre documentos no es tan ágil como en TurnItIn.

Investigación intracorpus. CopyCatchGold es el único sistema intracorpus que ofrece una interfaz textual. La vista se divide en dos listas (arriba) y dos paneles de texto (abajo). La primera lista a la izquierda consiste en pares (texto analizado – documento original) ordenados por índice de similitud. Al seleccionar un elemento de esta lista, en la de su derecha aparece una nueva lista con las coincidencias existentes entre esos dos documentos. Al seleccionar un elemento en esta segunda lista, en los paneles de texto de la parte inferior aparecen las ocurrencias de la coincidencia en cada uno de los textos. Se utiliza un código de colores limitado (rojo para coincidencias y azul para indicar la posición del texto).

Investigación de dos documentos en detalle. Para este tipo de investigación, en interfaces textuales se utiliza un doble panel de texto. En diversos sistemas (TurnItIn, URKUND, PlagiarismFinder) no existe de manera separada ya que se engloba en el informe de originalidad. CopyCatchGold sí la ofrece, dando opción de marcado HTML de frases coincidentes.

3.2. Interfaces gráficas

OrCheck, VAST y PRAISE ofrecen interesantes soluciones gráficas de representación de resultados.

Investigación de varios documentos referidos a un solo texto crítico. OrCheck muestra un rectángulo de fondo blanco, donde el texto crítico se hace corresponder con el eje horizontal, y cada fuente conforma una línea en el eje vertical. La similitud se indica usando un código de colores.

Investigación de dos documentos en detalle. VAST ofrece una interfaz dividida verticalmente en dos partes (Figura 2). La parte derecha a su vez se divide horizontalmente, conteniendo en cada parte un panel de texto mostrando cada documento.

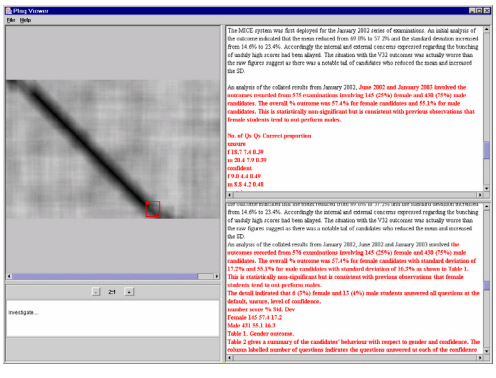

Figura 2. VAST – Análisis de resultados

La parte izquierda muestra el sistema de visualización, que consiste en una cuadrícula. Cada eje se corresponde con uno de los documentos. El índice de similitud se representa mediante niveles de gris. Al seleccionar un área en la cuadrícula, los textos correspondientes aparecen en la parte derecha, con el texto seleccionado en color rojo, para un examen detallado.

Investigación intracorpus. PRAISE ofrece una interfaz muy original (Figura 3). Los documentos se disponen a lo largo de un anillo. Las líneas entre dos puntos representan la relación entre dos documentos, y su color indica nivel de similitud.

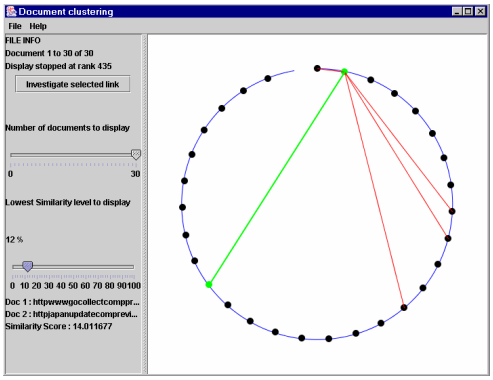

Figura 3. PRAISE – Anillo de similitud

4. Decisiones de diseño de la interfaz de Verbatim Engine

En esta sección se extraen las consecuencias más importantes del estudio y que se van a aplicar a la interfaz de Verbatim Engine.

Distribución en fases.

Código de colores.

Idiomas. Se adopta la opción multilenguaje.

Investigación de varios documentos referidos a un solo texto crítico. Adaptación del sistema usado en PRAISE. Consiste en un anillo en el que los documentos no críticos se sitúan a lo largo del mismo relacionando el índice de similitud con el ángulo de apertura. De esta manera se pueden apreciar a primera vista los pares de interés.

Investigación de dos documentos en detalle. Se propone el sistema de cuadrícula usado en VAST, añadiendo la posibilidad de definir un umbral de similitud para las zonas.

El diseño de Verbatim Engine se validará mediante una serie de sesiones de recorrido cognitivo con evaluadores y usuarios potenciales de la herramienta, y las conclusiones se incorporarán a la solución presentada.

5. Conclusiones

Se han presentado las decisiones de diseño de la nueva herramienta Verbatim Engine, basadas en el estudio mostrado en el punto 2.

Verbatim Engine pretende paliar la dificultad de utilización de algunos sistemas de detección de plagio mediante la aplicación de técnicas de usabilidad para el diseño de su interfaz.

Referencias

[1] Aiken, A. MOSS. University of California Berkeley, USA. http://www.cs.berkeley.edu/~aiken/moss.html.

[2] Bloomfield, L. WCopyFind. University of Virginia, USA. http://plagiarism.phys.virginia.edu/

[3] Centre for Interactive Systems Engineering. South Bank University, UK. http://cise.lsbu.ac.uk/tools.html

[4] CFL Software Development. http://www.copycatchgold.com

[5] Culwin, F. Practical Free-text Plagiarism Investigation. Proceedings of Seminars in University of Lincoln, UK, 4 June 2004.

[6] Google Search Engine. http:// www.google.com.

[7] IParadigms. http://www.iparadigms.com/

[8] Lyon, C. Ferret. University of Hertfordshire, UK. http://homepages.feis.herts.ac.uk/~pdgroup/

[9] M4 Software. http://www.m4-software.de

[10] MalPohl, G. JPlag. Universität Karlsruhe, Germany. http://www.jplag.de

[11] Miller, G. "The Magical Number Seven, Plus or Minus Two", The Psychological Review, vol. 63 pp. 81-97.

[12] Monostori, K., Schmidt, H., Zaslavsky, A. MatchDetectReveal. Monash University, Australia http://www.csse.monash.edu.au/projects/MDR/

[13] Pike, R. Sherlock. University of Sydney, Australia. http://www.cs.usyd.edu.au/~scilect/

[14] PrioInfo. http://www-prioinfo.se