RUSC Revista de Universidad y Sociedad del Conocimiento
Universities and Knowledge Society Journal

http://rusc.uoc.edu

ARTICLE

Determination of writing styles to detect similarities in digital documents

Yohandri Ril Gil
rilltt@uci.cu
Technology and Information Technology Security Advisor, FORTES Educational Technology Centre,
University of Information Sciences, Havana

Yuniet del Carmen Toll Palma
ytoll@uci.cu
Quality Assessor, FORTES Educational Technology Centre,
University of Information Sciences, Havana

Eddy Fonseca Lahens
elahens@uci.cu
Specialist, IdeoInformatics Centre,
University of Information Sciences, Havana

Submitted in: March 2013
Accepted in: April 2013
Published in: January 2014

Recommended citation

Ril, Y., Toll, Y.C. & Fonseca, E. (2014). Determination of writing styles to detect similarities in digital documents. *Revista de Universidad y Sociedad del Conocimiento (RUSC), 11*(1). doi http://dx.doi.org/10.7238/rusc.v11i1.1783

Original title: *Determinación de estilos de escritura para la detección de similitudes entre documentos digitales*

Abstract
Anything involving human intellect is at risk of being plagiarised. This includes scientific and literary works such as articles, theses, audiovisual works, plans, projects and computer programs. However, this article pays special attention to the existence of this phenomenon in written works in general, and in digital documents in natural or programming languages in particular. The objective of the research is to develop and apply a mathematical model that allows the writing style used in the drafting of texts to be determined. The results obtained from the application of the procedure are intended to serve as the basis for reducing the number of documents that need to be compared in order to analyse and detect similarities in them. The procedure was experimentally applied to a set of articles classified by topic and author, where the writing styles used to draft them differed.

Keywords
writing style, digital documents, plagiarism, procedure

Determinación de estilos de escritura para la detección de similitudes entre documentos digitales

Resumen
Todo lo inherente al intelecto humano es susceptible de actos de plagio: obras científicas y literarias tales como artículos, tesis, obras audiovisuales, planos y proyectos, códigos fuentes de programas, entre otros. Sin embargo, el presente trabajo dedica especial atención a la existencia de este fenómeno en obras escritas, en concreto documentos digitales provenientes de lenguajes naturales o de programación, y centra su objetivo en el desarrollo y aplicación de un modelo matemático que permite determinar el estilo de escritura empleado en la redacción de los textos. Los resultados que se esperan obtener a partir de la aplicación del procedimiento servirán de base para la reducción en el número de documentos que se deben comparar en el análisis y detección de similitudes entre estos documentos. De forma experimental se aplica el procedimiento a un grupo de artículos clasificados por temáticas y autores y que difieren entre ellos en el estilo de escritura utilizado para su redacción.

Palabras clave
estilo de escritura, documentos digitales, plagio, procedimiento

1. Introduction

The advantages offered by information and communication technologies (ICTs) are irrefutable and borne out by numerous examples. However, along with the positives come the negatives like plagiarism, which is undoubtedly one of the most illustrative examples. The *Real Academia Española*, the institution responsible for regulating the Spanish language, defines the Spanish term '*plagiar*' (Real Academia Española, 2001) – 'to plagiarise' in English – in a simple, categorical manner: "*Copiar en lo sustancial obras ajenas, dándolas como propias*". A definition in English of the term is "to appropriate (ideas, passages, etc.) from (another work or author)" (Plagiarise, n.d.).

Anything involving human intellect is at risk of being plagiarised. This includes scientific and literary works such as articles, theses, films, sheet music, audiovisual works, plans, projects, computer programs and websites. However, this article pays special attention to the existence of this phenomenon in written works in general, and in digital documents in natural or programming languages in particular.

Using elements from other works is a common practice. The way it is done determines the intentionality of respecting – or otherwise – the authorship of the sources. Among the various methods used are the following:

- **Correct:** the provenance of the referenced textual material is accurately shown, adding information about the author of the work cited.
- **Paraphrasing:** based on an interpretation, certain ideas from other documents are expressed in an author's own words.
- **Multiple sources:** a new document is created by the textual or modified use of fragments of text from other documents.
- **Mosaic:** segments of an original document are disarranged so as to be textually used in a new document.
- **Odd modifications:** segments of an original document are used while words or phrases are inserted to modify them.
- **Copying and pasting:** the document is copied in part or in full without any modifications. This is relatively easy to detect by doing a comparison to determine whether or not there are any differences between the documents.

Technological advances mean that systems are now available to detect plagiarism in documents; however, human supervision of the process is still essential. An accusation of plagiarism is serious and should not be made lightly. The use of automated tools reduces the number of documents that require human intervention by differentiating between them, sometimes using heuristics, as is the case for determining that the original of several similar websites is the one with the highest ranking. For comparisons between works done by students, the one by the student with the highest performance could be considered the original. Unfortunately, these heuristics are not always valid, and it is not even possible to be sure that they are so in the majority of instances. As is often the case, the proposed analysis depends on a subjective assessment. Thus, from this point forward, the term 'determination of similarities' will be used instead of 'plagiarism'.

Many articles have discussed topics relating to the determination of similarity in documents. Of particular note among these are:

"Detecting similar documents using salient terms" (Cooper et al., 2002)

The comparison between two documents is done by pre-processing both and searching for tokens[1] defined in advance, such as proper nouns, acronyms, locations, abbreviations, etc. An Information

1. A token, also called a 'lexical component', is a string of characters that has a coherent meaning in a particular natural or programming language.

Quotient (IQ) (Cooper et al., 2002) is assigned to each term, which is equivalent to the amount of information that its appearance in the text provides. The score for the similarity in two texts reflects proportionality with the number of terms appearing in one document but not in the other.

"Tool support for plagiarism detection in text documents" (Gruner & Naven, 2005)

A method is shown for the analysis of writing styles on the basis of statistical behaviour. It seeks to determine how language is used by the author and to compare it with other documents. The search for styles can be performed on paragraphs or on fragments of text.

"Check: a document plagiarism detection system" (Si et al., 1997)

It determines similarity on the basis of how often the terms appear. A weightings vector for each compared documents is generated, and the cosine between these vectors is used as a parameter in a function that returns an estimate of similarity. The procedure is repeated for each section of the documents until the existence or otherwise of copying between them is determined.

"Sim: a utility for detecting similarity in computer programs" (Gitchell & Tran, 1999)

The article is about detecting similarity in programs. The programs are separated into tokens in order to then calculate the alignment score between two token streams. Each gap or mismatch in the alignment is assigned a weighting. The similarity is computed using the expression:

$$s = \frac{2 \times score(p1,p2)}{score(p1,p2) + score(p1,p2)} \qquad [I]$$

Where **score (p1, p2)** is the alignment score between program 1 and program 2.

"Plagiarism in natural and programming languages: an overview of current tools and technologies" (Clough, 2000)

It summarises a set of tools available for detecting similarities in texts in natural or programming languages. It concisely describes the distinctive elements of some algorithms used by those tools to perform comparisons, as is the case for determining the Longest Common Subsequence.

The objective of the research is to develop and apply a mathematical model that allows the writing style used in the drafting of texts to be determined.

Research method and theory used

In order to conduct the research, various scientific methods were used. For example, a literature review was conducted to look for sources of information to theoretically underpin the study. Analysis and synthesis methods were generally used throughout the research process, and particularly for specifying the theoretical fundamentals relating to the detection of similarities in digital documents, and to the analysis and interpretation of the results obtained from the application of the tool to detect plagiarism in digital documents. In addition, descriptive statistics were used to process the

data obtained from the application of the tools to detect similarities in digital documents, and inferential statistics were used to make decisions about whether or not to reject the data obtained from the process for detecting similarities in digital documents.

2. Development

In the case of systems for detecting similarities in documents, the biggest challenge is the enormous volume of data that has to be processed; that is why a pre-classification of the whole sample or the original documents available for the comparison is required.

The criteria taken into account for the classification often include document type, language, category, subcategory, keywords, authors and dates. Our proposal is to incorporate writing style as a comparative criterion when it comes to determining whether or not a document is original. Thus, the number of documents to be compared in the search for similarities would be reduced to those that have a direct relationship – in terms of classification – with the compared document, which in turn have a similar writing style.

2.1. Procedure for writing style extraction

Any text can be represented by a statistical model that identifies its intrinsic characteristics, which relate to the author's writing style. Among these we would mention:

Stop words: this refers to the use of articles, adverbs, prepositions and conjunctions. The frequency with which these words are used denotes the author's writing style. Consequently, stop word mean is defined as:

$$Pp = \frac{Cpp}{Tp} * 100\ \% \qquad [II]$$

Where *Pp* is stop word mean, *Cpp* is the number of stop words used and *Tp* is the total number of words in the text.

Level of difficulty: determines the level of education that someone needs to have in order to understand the text. There are several indices available to calculate this level. Gunning (Wikipedia, 2011), Dale and Chall (1948) and Flesch-Kincaid (DuBay, 2004) are some of them, although the latter is the most commonly documented and cited.

The expression to determine the Flesch-Kincaid index is as follows:

$$I_{FK} = 1.599\Lambda - 1.015\beta - 31.517 \qquad [III]$$

Where λ is the mean of one-syllable words per 100 words, and β is the mean sentence length measured by the number of words.

Richness of vocabulary: proposed by Honoré (1979), this tries to determine the richness of the author's vocabulary on the basis of the total unrepeated words used in the text. The following expression is used for that purpose:

$$R = \frac{100\ log\ (M)}{M^2} \qquad [IV]$$

Where *M* is the number of different words in the text.

Depending on the type of document being analysed, the calculation of *R* has more or less validity. Certain specialist articles and computer programs are good examples of this, as their very nature requires the constant repetition of words.

As a consequence of the above, an approach proposed by Yule (1944) is introduced as a calculation alternative, defining:

$$K = \frac{10^4 (\sum_i^{\infty} = 1\ i^2\ V_i - M)}{M^2} \qquad [V]$$

Where *Vi* is the number of words that appear *i* times in the text. *M* has the same meaning as in the previous calculation.

Mean sentence length: is a reliable measure of grammatical knowledge that the author uses in the composition of sentences.

$$Lp_0 = \frac{\sum_i^{No} = 1\ lo_i}{No} = \frac{Tp}{No} \qquad [VI]$$

Where lo_i is the length in words of the sentence occuring in position *i*, *No* is the total number of sentences and *Tp* is the total number of words in the text.

Mean word length: this term is directly connnected with the richness of the author's vocabulary and measures his or her ability to use complex words[2].

$$Pp = \frac{Cpp}{Tp} * 100\ \% \qquad [II]$$

2. Complex words are considered to be those formed by three or more syllables that do not represent proper nouns, prefixes, suffixes or compound words.

Where *Lpp* is mean word length, *Tc* is the total characters used (excluding spaces), and *Tp* is the total number of words used.

The final definition is that of the writing style vector (E), whose components are described above:

$$E < Pp, I_{FK}, K, Lp_0, Lp_p > \quad [VIII]$$

While it is true to say the determination of an author's writing style implies some degree of uncertainty, knowing what that degree is enables a suitable margin of deviation to be established when it comes to determining who the creator of a document is. Among the main parameters that have an influence on uncertainty are the topic covered, the document length (in terms of the number of paragraphs, sentences or words used), the author's experience and the document end user.

Having a statistical estimate of the writing style is important to ensure that there is a comparative criterion for selecting and classifying documents, and for determining their authorship. One of the applications is the determination of similarities in documents, where processing time needs to be optimised. The materials to be compared are stored on vast databases, so processing everything is not an option, nor is processing only those documents belonging to one classification category. Other arguments are required to differentiate between and considerably reduce the number of documents to be processed, like a writing style identifier, for example.

3. Research analysis and results

In order to apply the proposal for determining writing styles, 37 documents by 5 authors in 4 thematic areas were selected: education, history, medicine and children's stories. Two of the authors belong to the thematic area of medicine. Chart 1 shows the document distribution by thematic area.

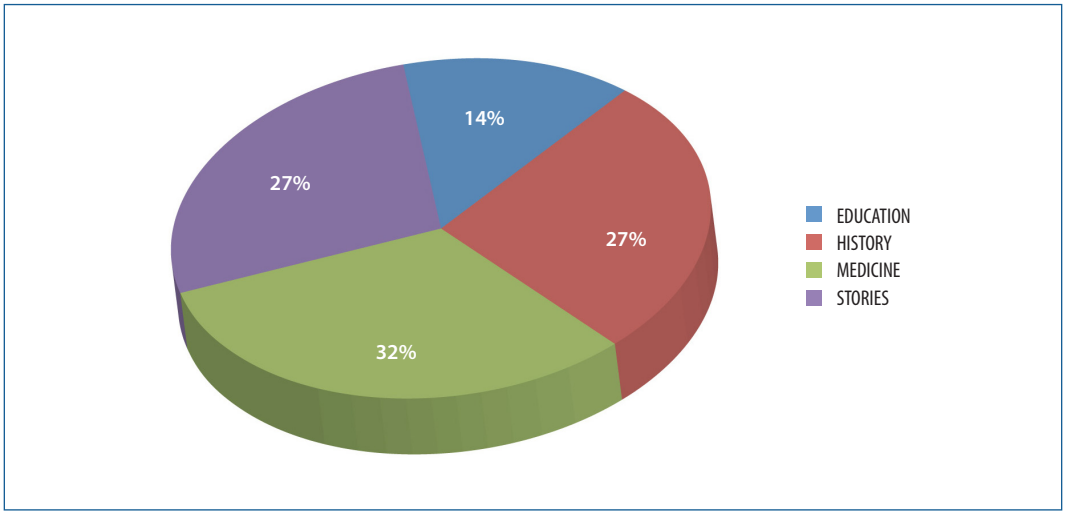

Chart 1. Document distribution by thematic area

Original title: *Determinación de estilos de escritura para la detección de similitudes entre documentos digitales*

In order to automatically determine the writing style vector, a tool was developed (Figure 1) to enable the extraction of a set of statistics from the texts analysed as the step prior to calculating the vector.

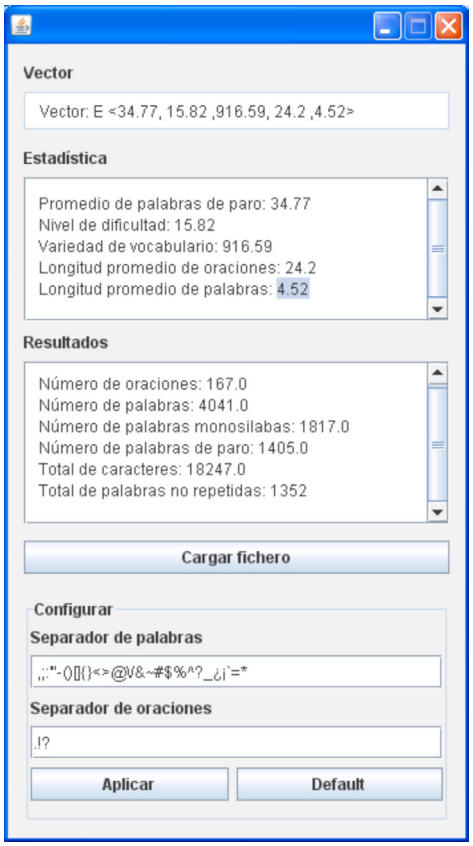

Figure 2. Tool for calculating the style vector

The characteristics extracted from the documents are listed below (to aid understanding of the content of Figure 1 for non-Spanish speakers, the English translation of each characteristic is given in brackets):

- *Promedio de palabras de paro* (Stop word mean)
- *Nivel de dificultad* (Level of difficulty)
- *Variedad de vocabulario* (Richness of vocabulary)
- *Longitud promedio de oraciones* (Mean sentence length, measured in words)
- *Longitud promedio de palabras* (Mean word length, measured in characters)
- *Número de oraciones* (Number of sentences)
- *Número de palabras* (Number of words)
- *Número de palabras monosílabas* (Number of one-syllable words)
- *Número de palabras de paro* (Number of stop words)
- *Total de caracteres* (Total characters)
- *Total de palabras no repetidas* (Total unrepeated words)

In order to determine trends and others statistics, a Microsoft Office 2007 spreadsheet was used. Table 1 shows the mean values obtained for the style vector components for each author. As we can see, the parameter that makes the biggest difference is richness of vocabulary. Authors 1 and 2 stand out in this case, who covered education and history topics, respectively. In particular, author 2 has the lowest level of difficulty owing to the use of easily understandable narrative documents.

Table 1. Style vector means by author

	Stop word mean	Level of difficulty	Richness of vocabulary	Mean sentence lenght	Mean word length
Author 1	37.49	21.89	2482.43	20.46	5.314
Author 2	35.88	11.40	2357.63	30.39	4.742
Author 3	30.52	26.88	1011.01	12.41	5.162
Author 4	24.21	28.96	1121.34	10.57	5.356
Author 5	33.34	26.70	665.76	12.54	4.399

For each author, it is essential to determine how the writing style parameters vary, that is to say, whether not a particular author is able to maintain his or her writing style across a set of documents on the same topic. An analysis was performed on each vector parameter to calculate its mean deviation, as shown in Table 2. As can be anticipated from the information shown in the previous table, it is precisely the richness of vocabulary parameter that varies the most, and the mean word length parameter that varies the least – so little so that it was necessary to increase the number of decimal places to three.

Table 2. Mean deviations by vector parameter

	Stop word mean	Level of difficulty	Richness of vocabulary	Mean sentence lenght	Mean word length
Author 1	0.65	0.82	107.03	1.26	0.079
Author 2	0.53	0.74	82.57	1.10	0.075
Author 3	1.96	1.05	69.78	1.05	0.066
Author 4	1.25	1.53	229.08	0.99	0.092
Author 5	1.10	5.43	75.88	1.87	0.079

Chart 2 shows the distribution of deviations through a graphic representation of areas. As we can see, author 2 has the most consistent writing style because the area representing it (red) is more uniform. The opposite is the case for authors 4 and 5, who have several peaks and troughs in their respective deviation distribution areas.

Despite the fact that authors 3 and 4 both belong to the thematic area of medicine, there is a considerable difference in their writing styles.

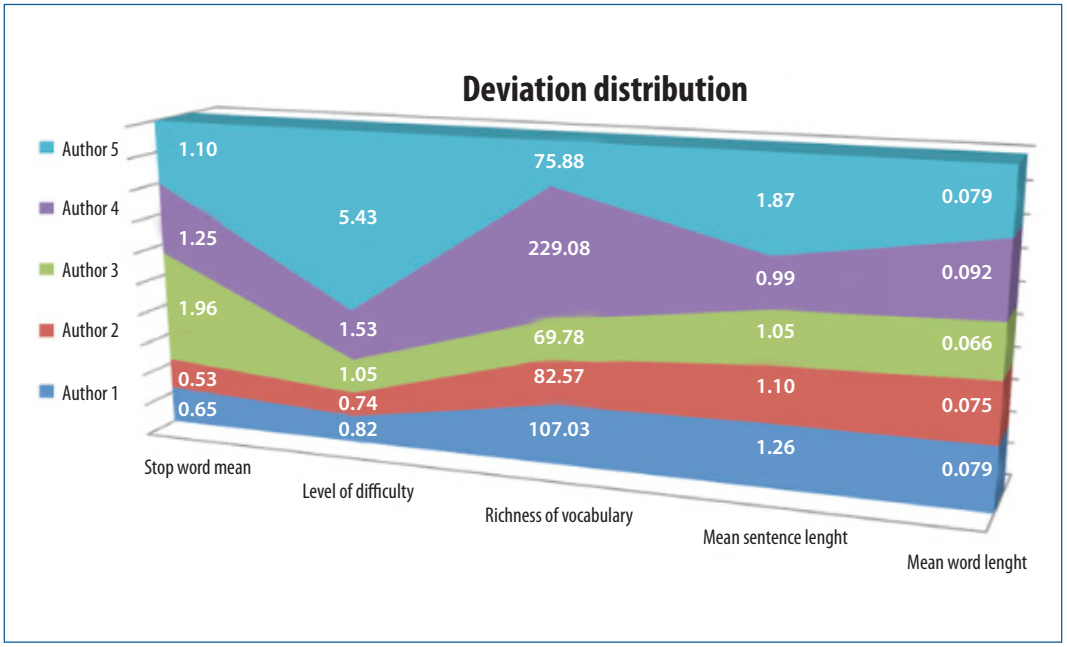

Figure 3. Deviation distribution areas

The proposed method for determining writing styles can be used in a scenario where it is necessary to describe documents whose authorship has been validated. Some descriptors are commonly used, such as author, title, keywords, category, subcategory and document type, yet having a descriptor like writing style will enable a reduction in the number of documents that need to be compared in the search for possible plagiarism. In this respect, when the aim is to analyse the presence of plagiarism in a new document, it will only need to be compared with a sub-set made up of those with a similar writing style vector.

Although it is possible to start with the idea that every author has a unique writing style, it is nevertheless important to consider that the main threats to the validity of this method reside in the selection of documents to determine an author's writing style. They should be documents whose authorship has been authenticated, while bearing in mind that more than one author may have participated in the drafting of a document. When the document is not about one of the topics customarily covered by the author, it is acceptable for the deviation indices to vary within a reasonable range. Thus, it is advisable to use documents on topics that the author usually covers. Finally, there are other elements that are equally as important, such as the learning process and the changing reality surrounding the individual, which gradually give rise to variations in every author's writing style. In this respect, the determination of writing styles will be more reliable when, from the viewpoint of writing style, the individual is more mature and experienced.

4. Discussion and conclusions

The proposed procedure focuses on the determination of writing styles and enables the academic community to observe similarities in documents. However, it does not constitute a definitive solution to the serious problem of academic plagiarism. Mindful, systematic training in values such as responsibility and honesty is required.

Why insist on academic honesty? If academic honesty is non-existent, then an impression of knowledge will be given that does not match the reality of what is genuinely in the cognitive structure. Dishonest behaviour in the academic sphere is a means of deceit and, above all, of self-deceit. It erodes the core of the educational aim of our teaching activity. Owing to its ethical nature, we might assume that the concept of honesty is implicit, but this assumption is not absolutely certain. Acts of plagiarism may be committed deliberately, accidentally or unknowingly.

What can be done to strengthen these values among our students? When we show professional coherence and commitment, creating a high-level pedagogical and intellectual atmosphere, we manage to eliminate or reduce academic dishonesty and its negative effects. In higher education, there are clearly major opportunities to learn and, at the same time, major opportunities to defraud. In some ways, we trust that most of the students are aware of their obligations to learn, influenced by the need to get a degree. Our education function cannot be limited to the use of tools to oversee plagiarism; this does not guarantee that the students will not act fraudulently. We must insist on the fact that ethical behaviour is based on free acceptance of the rules of conduct that cannot be imposed by force of authority.

From a very young age, human beings gradually enrich their vocabulary, gain more experience and training, and develop their writing styles as they learn. That is why the creation of a unique writing style is perceived as a slow process. In future studies, the intention is to explore (a) the procedures in order to obtain the development curves for writing styles, and (b) the characterisation of these styles on the basis of the parameters describing them, all of which have formed a fundamental part of this study.

The appropriation of third-party works as one's own will continue and evolve at the same pace as technology, so being ready to counteract this phenomenon is of vital importance. Most of the studies reviewed for this research will form the basis for the development of future works on the detection of similarities in documents, and the proposal presented here will serve as the starting point for determining which documents to compare. The extraction of the style vector marks the difference between authors, whether or not they cover the same topic. By applying the proposed mathematical model to a considerable set of documents, it was found that trends really do exist when it comes to drafting, and that such trends put a stamp of authenticity onto a document.

References

Clough, P. (2000). Plagiarism in natural and programming languages: an overview of current tools and technologies. *Research Memoranda: CS-00-05, Department of Computer Science, University of Sheffield, UK,* 1-31. Retrieved from http://ir.shef.ac.uk/cloughie/papers/plagiarism2000.pdf

Cooper, J. W., Coden, A. R., & Brown, E. W. (2002). Detecting similar documents using salient terms. In *Proceedings of the 11th international conference on Information and Knowledge Management*. New York, NY: ACM. Retrieved from http://www.labsoftware.com/flahdo/Papers/CIKMDuplicates.pdf

Dale, E., & Chall, J. S. (1948). A formula for predicting readability. *Educational Research Bulletin*, *27*(1), 11-20. Retrieved from http://www.ecy.wa.gov/quality/plaintalk/resources/classics.pdf

Dubay, W. H. (2004). *The principles of readability*. Costa Mesa, CA: Impact Information. Retrieved from http://files.eric.ed.gov/fulltext/ED490073.pdf

Gitchell, D., & Tran, N. (1999). Sim: a utility for detecting similarity in computer programs. In *The proceedings of the 30th SIGCSE technical symposium on Computer Science Education*. New York, NY: ACM. Retrieved from http://www.eng.uwi.tt/depts/elec/staff/feisal/ee302/sim-gitchell.pdf

Gruner, S. & Naven, S. (2005). Tool support for plagiarism detection in text documents. In *Proceedings of the 2005 ACM symposium on Applied Computing*. New York, NY: ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1066677.1066854. doi http://dx.doi.org/10.1145/1066677.1066854

Honoré, A. (1979). Some simple measures of richness of vocabulary. *Association for Literary and Linguistic Computing Bulletin*, *7*(2).

Plagiarise (n.d.). In *The Collins English Dictionary*. Retrieved from http://www.collinsdictionary.com/dictionary/english/plagiarise

Real Academia Española (Ed.) (2001). *Diccionario de la Real Academia Española*. Madrid, Spain: Real Academia Española.

Si, A., Leong, H. V., & Lau, R. W. H. (1997). Check: a document plagiarism detection system. In *Proceedings of the 1997 ACM symposium on Applied Computing*. New York, NY: ACM. Retrieved from http://www.cs.cityu.edu.hk/~rynson/papers/sac97.pdf. doi http://dx.doi.org/10.1145/331697.335176

Wikipedia (2011). Gunning fog index. *Wikipedia*. Online: Wikipedia.org. Retrieved from http://en.wikipedia.org/wiki/Gunning_fog_index

Yule, G. U. (1944).The statistical study of literary vocabulary. *Journal of the Royal Statistical Society*, *107*(2), 129-131. Retrieved from http://www.jstor.org/discover/10.2307/2981280?uid=3737824&uid=2129&uid=2&uid=70&uid=4&sid=21102626763567. doi http://dx.doi.org/10.2307/2981280

About the Authors

Yohandri Ril Gil
rilltt@uci.cu
Technology and Information Technology Security Advisor, FORTES Educational Technology Centre, University of Information Sciences, Havana

He holds a bachelor's degree in Telecommunications and Electronic Engineering and is an assistant lecturer at the University of Information Sciences (UCI), Havana, specifically for Teleinformatics I and II on the Informatics bachelor's degree course. He is a technology and information technology (IT) security advisor in the FORTES Educational Technology Centre He is currently taking a master's degree in Distance Education at the UCI. His main lines of research are telecommunications networks, IT security, virtual learning environments, the quality of learning objects and software architecture. He has had various articles published in renowned journals such as *No Solo Usabilidad* and *RUSC. Universities and Knowledge Society Journal*, and in the proceedings of national and international events.

Yuniet del Carmen Toll Palma
ytoll@uci.cu
Quality Assessor, FORTES Educational Technology Centre, University of Information Sciences, Havana

She holds a master's degree in Software Quality and is an assistant lecturer at the University of Information Sciences (UCI), Havana. As an information expert, she collaborates with the Information and Knowledge Group at the FORTES Educational Technology Centre. She is responsible for editing the *Producción de recursos didácticos* [Production of didactic resources] and *Archivo* [File] sections of the FORTES centre *TeduScopio* newsletter, and is also the general editor of the newsletter. She is currently a quality assessor in the FORTES centre. Her main lines of research are the quality assessment of learning objects and other software products, information and knowledge management, and the architecture of information for FORTES centre projects. She has had various articles published in renowned journals such as *EDUTEC*, *No Solo Usabilidad, REDC* and *RUSC. Universities and Knowledge Society Journal*, and in the proceedings of national and international events, in which she has participated widely.

Eddy Fonseca Lahens
elahens@uci.cu
Specialist, IdeoInformatics Centre, University of Information Sciences, Havana

He is a specialist in the IdeoInformatics Centre at the University of Information Sciences (UCI), Havana. He leads the Motor project for the intelligent categorisation of content and is an architect on the Motor project for the intelligent categorisation of content for e-mail. His main lines of research are the development of information technology (IT) solutions for the Internet, the development of concurrent and distributed applications, the automatic categorisation of content and the automation of learning object assessment. He has had various articles published in national journals and in the proceedings of national and international events.

Universidad de las Ciencias Informáticas
Carretera a San Antonio de los Baños, km 2 ½
Torrens, municipio de La Lisa
La Habana, Cuba