# FIFTY WAYS TO DETECT A GHOSTWRITER

*Katerina Zdravkova*
Faculty of Computer Science and Engineering
University Ss. Cyril and Methodius in Skopje
Rudjer Boskovikj bb, 1000 Skopje, Macedonia
e-mail: katerina.zdravkova@finki.ukim.mk

**ABSTRACT**

Ghostwriting became students' most popular way to avoid writing of boring essays, or the best way to easily earn by writing on behalf of another student.

This paper presents several markers indicating a presence of potential ghostwriters. Proposed methodology suggests various inspection techniques, which do not prove anything in isolation. Whenever they are jointly implemented, they successfully cluster the essays, suggesting plausible absence, potential and almost certain presence of one or few ghostwriters. After the initial clustering, all the papers go through subtle linguistic check. In our sample, later approach discovered some unexpected phrases which confirmed the presence of the same ghostwriters not only in the current, but also in the previous generations.

## 1 INTRODUCTION

Massive storage technologies and search engines immensely provide and facilitate information access, and at the same time they enable a smooth and undoubted detection of most plagiarism sins. It is no longer a problem to discover students' naïve 'copy and paste' activities using search engines. When the original source is written in another language, Google Translate provides help for both, the students, and the teachers who easily capture the machine translated and usually unedited parts of essays. Even fairly translated essays are easily noticed, because students usually translate articles using an extremely professional writing style. Teachers can always use plagiarism detecting tools such as the famous iThenticate, Turnitin, or WriteCheck, (recommended by [1]) capable of comparing essays with the databases of stored texts.

In the recent years, writing essays, papers, and even theses has become a very popular and frequent activity. Magnificent article [2] reveals the presence of extremely well-paid professional writers such as Ed Dante, who "completed 12 graduated theses of 50 pages or more". Many ghostwriters work on their own, while others are organized by specialized agencies called essay or paper mills [3]. Chinese estimate "that university students spend up to half a billion yuan ($73 million) a year to have other people write their essays" [4]. Most of their works are still academically very inexperienced, and the presence of classical 'cut and paste' plagiarism is usually abundant. Unlike them, ghostwriters in America, Australia and Europe seem to be highly professional. They generate impeccable works with no evident plagiarism, and as aka El Dante claims in The Chronicle Review [2], they are all "based on specific instructions provided by cheating students". It seems that essay mills are on great demand in the academic world, and scholarly mercenaries daily finish many extremely ambitious tasks. Although not intended to support human writing, there are some authoring tools, such as GhostWriter [5, 6], which can facilitate the preparation of different contents. This case-based reasoning system can effectively support content authors suggesting them feature values.

If computer facilitated preparation of written products becomes reality, the opposite direction is not so prosperous. In spite of the presence of many plagiarism detecting tools, ghostwriting is rarely detected and almost impossible to prove. In the recent years it has become a lucrative business and young students or academicians willing to apply it are 'sprouting up like mushrooms after the rain'. As a consequence, procedural concerns grow, and one proposed solution in USA is to expand the federal rules to diminish its side effects [7]. Wherever the legislation is not prepared to handle with the problem, concerned faculties minimize the contribution of individual essays in the final grade [8].

Experienced teachers usually intuitively feel the cheat. Unfortunately, they have no means to prove it with indisputable certainty. Students always have an accurate and very rational excuse for all teachers' accusations. The only prove that the essay was not individually prepared is student's inability to tell what is written in it. But, very few teachers have the courage to find material evidence of the cheat, and time to personally enquire the student.

This paper presents several markers indicating the presence of potential ghostwriters who have prepared many essays on related topics for the same course over years. Second section is dedicated to the most obvious indicators derived from document properties and student activities in the learning management and storage system. Third sections presents the techniques connected with IP addresses through which students accessed the desired activity. The approach can be even more effective if the system has a track of all the previous accesses of all the students. After finishing the external inspection, the text itself becomes a target of a subtitle text and linguistic mining. These mining techniques are described in the fourth section.

Before presenting the final conclusion and further work, the results of the estimation of proposed approach are submitted. They are based on a sample of 185 short essays distributed into 5 groups of related topics. Although the correlation of proposed indicators is negligible, there are many exact matches between some of them. They confirmed the existence of several works done by another student, and suggest that there could be an anonymous ghostwriter. Before more reliable indicators are invented, the claim that a work was not done by the student are still a speculation.

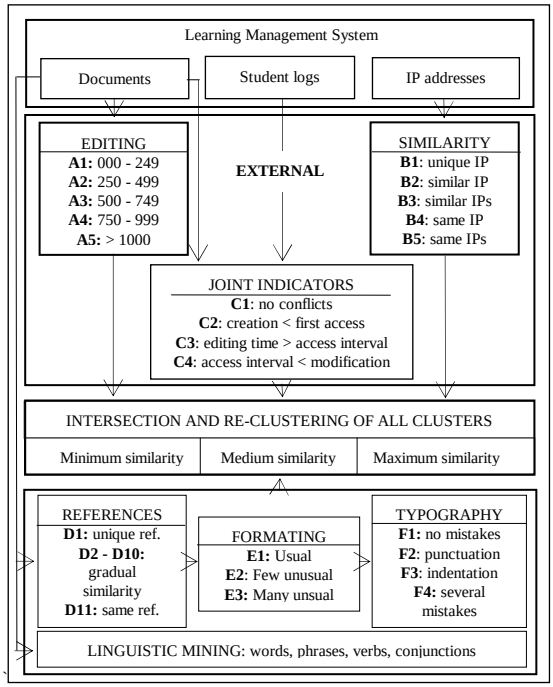


Figure 1: *System architecture*

## 2 EXTERNAL INDICATORS

External indicators are the information that can be extracted without looking into the contents of the essay. In order to be able to extract them, let's presume that:

- all the essays are stored and delivered as documents
- the course is maintained with a learning management system (or LMS) capable of reporting all the activities.

The document itself contains information such as: title of the paper, name of the first author, name of the user who last saved it, time of paper creation, revision number and total editing time. Only very naïve or extremely imprudent students deliver a document revealing initial creation in the past, no editing time, or a name of another colleague. But, they always have an acceptable excuse ("I used the template from last year.", "I use OpenOffice and saved it as Word.", "My computer is out of order, I went to my colleague".)

Traditional ghostwriters deliver the final essay without influencing the introductory activities prior to final upload of the essay. It seems that they deliver the essay as part of a mail message [9], which students simply copy into a newly opened document and deliver it with no editing time. More careful students use older documents or convert into pdf.

On the other hand, LMS reports offer information concerning the moment of first access of observed activity, how many times and when was it accessed, when was the document uploaded, how many times and how long has the student viewed the activity after final uploading of the essay, how many times and how long was the activity visited, and when was it last viewed. Using these times, time between first upload and first access and time between last upload and last access can also be calculated.

Having a long teacher's experience, the author of this paper can state that students who submit other's outcome usually access the definition of the task at most once, but afterwards regularly check it, sometimes upload it again, and eagerly wait for the final grade. Again, whenever someone is accused of uploading an essay 'borrowed' by another colleague, there are many excuses. The most frequent is: "We jointly prepared the essay".

There are at leats these joint indicators which can be derived from document properties and activity reports:

- the difference between time when the document with the essay was created and the first access of the definition of its topic,
- the difference between document total editing time and the difference between first uploading and first access,
- the difference between final uploading and document last modification.

They should all be positive. Surprisingly, we always discover few cases with negative differences. Predictably, students offer an excuse that they edited an earlier document rather than creating a new one. And, their excuse is very plausible.

All the external indicators are very useful to catch a student who delivers an essay prepared by a colleague from own generation, or from another student of previous generations. They can discover the presence of a ghostwriter who only delivers the final version of the essay. However, there are even more sophisticated ghostwriters. They take student's ID and password in their possession, and behave on student's behalf. ID fraudsters never make any mistake measurable with external indicators. Whenever they decide to commit an ID exchange, there is absolutely no proof that they exist. In order to catch them, additional techniques must be implemented.

## 3 INDICATORS BASED ON IP ADDRESSES

Most learning management systems are able to keep track of all users' IP addresses throughout all the courses. The first indicator is a comparison of all IP addresses participants use during the activities concerned with the essay. They are downloaded in a separate worksheet.

IP indicators are defined as follows:

- all the records are numbered consecutively according to the time of their access
- the whole worksheet is sorted according to IP address
- IP addresses during essay upload are separated in an individual set

Each IP address is labeled using following equations:

$$Label(IP_i) = n_i$$

where $i$ is the consecutive index of IP address, and $n_i$ is the number of students approaching to the activity from same IP address.

Faculty IP addresses are excluded from labeling, because all the students can access from any computer in the students' laboratories.

Particular attention is paid to IP addresses during upload. If a student uploaded the essay from an IP address with a label greater than 1, its IP value is doubled. This correction weakens the influence of joint work to further strict clustering.

After IP labeling, each student ID is also labeled

$$Label(ID_j) = \sum_{k=1}^{m_j} \frac{\ln(Label(IP_j))}{m_j}$$

where $j$ is student's order in the course, and $m_j$ is the number of IP addresses assigned to $j^{th}$ student with $IP_j > 1$.

At the end, each student ID is included into one of five clusters.

$$Cluster(ID_j) = \begin{cases} 0{,}00 & 0 \le ID_j < M/5 \\ 0{,}25 & M/5 \le ID_j < 2M/5 \\ 0{,}50 & 2M/5 \le ID_j < 3M/5 \\ 0{,}75 & 3M/5 \le ID_j < 4M/5 \\ 1{,}00 & 4M/5 \le ID_j \le M \end{cases}$$

where M is the maximum average value of IP addresses labels assigned to all students. Students belonging to the first clusters are those who seem that prepare their essays individually.

Students belonging to the last cluster are checked most thoroughly. Whenever they belong to the same cluster in other course activities, they are also checked at other courses. The match is perfect. Furthermore, the access to same IP address is simultaneous. But, controversial IP addresses usually belong to student dormitories. Students usually insist on collaborative work, which is stimulated, rather than punished.

## 4 INTERNAL EXAMINATION

After opening the document itself, the exactness of the defined assignment topic with the prepared essay is checked. There are very seldom mistakes, but they always reveal deliberate swap of the topic enabling a fake or 'collaboration'.

### 4.1 References

It has been noticed that some very special references appear in several student essays. Therefore, references are considered a valuable ghostwriter indicators.

References are labeled using the same strategy as IP addresses. Before the reference labeling, all the references are subtracted from the texts and each reference is assigned to the student. They are sorted and labeled using exactly the same formulas presented in the left column of this page. Very popular sites, such as English Wikipedia, or popular aggregators of ICT news are excluded from labeling.

At the end, each student ID is again included into a corresponding reference cluster. As far as the number of references is usually very high, clusters can be more refined. We propose 11 clusters. Students who used very special references belong to the cluster with value 0, while those who used the same references as their colleagues belong to the cluster with value 1. The most suspicious in the light of ghostwriters are those students who do not belong to extreme clusters. Namely, the person preparing several essays on the same topic collects a limited number of references and carefully divides them into almost disjunctive sets. But, they have many things in common, such as the language of the original reference, news aggregators.

### 4.2 Formatting styles

It has been noticed that most essays are written using normal formatting style. However, unusual styles such as short_text, long_text, long_text + Arial, apple_style_span, apple-converted-text, or yellowfadeinnerspan appear in several essays. At the moment, we do not have an application capable of rearranging texts according to their styles. Therefore, we have manually distributed students into groups according to the most frequent styles. And again, those students who belong to same clusters were joined together. It can be a coincidence, but also an indicator of dishonest student behavior.

### 4.3 Typographic similarities

Recent trial about Facebook ownership includes several checks with techniques belonging to linguistic forensics [10]. The first two (apostrophes and suspension points) are typographic. Students usually make many typographic mistakes. They:

- forget to put a space after the punctuation,
- indent the line by adding several spaces,
- add a point after reference bracket, although teacher example excludes it

In absence of an application dealing with typographic mistakes and similarities, simple replacement with highlighted text is very useful. Essays are again divided into clusters according to the type of highlighted replacement.

Believe it or not, some students have already been united in several of these clusters. They had too many similarities, including the way of signing the paper, so it was evident that their effort was either joint, or done by few of them. However, these student cheatings are not as severe as the presence of an unwanted author who is gaining profit.

### 4.4 Linguistic similarities

The best way to catch a ghostwriter is to compare writing styles in all essays. One interesting approach is offered by Rong Zheng et al. who are dealing with authorship identification of online messages [11]. Apart from proposing their own framework, they also offer a comparison of previous studies in authorship identification.

The main reason to start chasing the ghostwriter was an essay with 20% identical titles. In all of them, the subtitle 'used sources' or 'literature' preceded the references.

All the essays were processed separately. The crucial elements were:

- frequency of the words and short phrases consisting of at most five words
- frequency of the most frequent verbs
- frequency of conjunctions

After examining these markers, it appeared that many students used the verbs: exist, create, and select. Some of them used either blessing or some religious phrases. Furthermore, the frequency of conjunctions of these students was higher than regularly. And, at the end, all of them were either with no editing time, or in pdf. The most interesting is the fact that all these students belonged to some of medium clusters.

## 5 EXPERIMENTAL RESULTS

The effectiveness of proposed approach was tested over a pool of 185 short student essays dealing with assistive technologies. They were first tested on plagiarism. Only three essays contained literal copies of texts found on the Internet. Another four essays contained unedited Google translations. They were excluded from further inspection.

Major external indicators were calculated and correlated mutually and with the essay grade and course final grade. There was a small correlation between editing time and essay final grade, very high correlation between number of views and views before upload. Most factors were close to zero. All of these proved nothing in particular.

Joint indicators derived from document properties and activity reports were much more sensitive to potential uploading of other's intellectual property. They located several students who knew nothing about the contents of the essay. At least one goal was accomplished, few cheaters were uncovered.

The best results were obtained using the clustering. As mentioned before, all the internal clusters were consisting of the same students. Students suspected of using ghostwriter services were always stuck together. This fact might be a proof that such a person existed, but he or she was not discovered.

In order to verify the cheat, typographic and linguistic check were performed over essays from previous generations. The arguments didn't exist few years ago, but their presence was noticed for the first time two years ago, first moderately, nowadays easily noticed.

## 6 CONCLUSION AND FURTHER WORK

This paper presented the attempt to uncover the dilemma, is there a ghostwriter among students. Almost fifty indicators were established to prove that such a person, or may be several of them exist. These indicators were very successful to catch students committing harmless fakes and cheats. They have also revealed that students are sometimes grouped together and jointly prepare their assignments. Whenever they are capable of presenting the contents of the essay, their fault or sin again remains unpunished.

All the indicators were sensitive to small student deficiencies. But, the professional outsourcer was never caught in the net. The crucial evidence of his or her presence were linguistic similarities.

We have already started the creation of a plagiarism tool intended to integrate student essays with search engines, Google Translate, and the pool of previous essays. The tool will be soon enlarged with ghostwriter detector. We do hope that it will discover the cheat and reduce it to the level of previous years.

## References


[1] Plagiarism in a Digital Age: Voices from the Front Lines: What's Happening in High Schools Today? http://www.plagiarism.org/plag_webinar_high_schools.html

[2] P. Davis. The Ghowstwriter Behind Student Papers. Society for Scholarly Publishing, The Scholarly Kitchen, http://scholarlykitchen.sspnet.org/2010/11/18/the-ghostwriter-behind-student-papers/#comments

[3] Essay Mill, http://en.wikipedia.org/wiki/Essay_mill

[4] Medeiros, I., "Education in China: Ghostwriting hits frightening levels at universities", Thoughts on Design, Technology and Culture, http://designative.info/2010/03/26/education-in-china-ghostwriting-hits-frightening-levels-at-universities/

[5] D. Bridge, A. Waugh. Using experience on the read/write web: The GhostWriter System. Proc. of WebCBR, Reasoning from Experiences on the Web, Working Prorgamme at the 8th International Conference on Case Based Reasoning. 2009.

[6] D. Bridge, A. Waugh. An Evaluation of the GhostWriter System for Case-Based Content Suggestions. http://www.cs.ucc.ie/~dgb/papers/Waugh-Bridge-2010.pdf

[7] J. P. Justman. Capturing the Ghost: Expanding Federal Rule of Civil Procedure 11 to Solve Procedural Concerns with Ghostwriting. *Social Science Research Network. Minnesota Law Review*, Vol. 92, p. 1246, 2008

[8] I. Trajkovska. "I prepare seminar papers" http://www.vecer.com.mk/?ItemID=2CEA0A6FC904DB4B8E0664391C49478B. In Macedonian

[9] K. Zdravkova. Can Web 2.0 Reduce Plagiarism and Cheating. *The 8th International Conference for Informatics and Information Technology*. 2011.

[10] M. Liberman. High-stakes forensic linguistics. *Language Log*, http://languagelog.ldc.upenn.edu/nll/ ?p=3309

[11] R. Zheng, J. Li, H. Chen, Z. Huang. A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques. *Journal of the American Society tor Information Science and Technology*, 57(3):378–393. 2006