UDK 004.04 DOI: 10.31673/2412-4338.2018.044752

Otrokh S. I. *State University of Telecommunications, Kyiv*
Kuzminykh V. O., Osipenko M. V. *Ihor Sikorskyi Kyiv Polytechnic Institute, Kyiv*
Hryshchenko O. O. *State University of Telecommunications, Kyiv*

METHOD FOR ASSESSING IDENTICAL TEXT

The article discusses the issues and possibilities of using anti-plagiarism tools, types of plagiarism and ways to circumvent the means of verification. Also proposed a method for analyzing the assessment of the identity of the texts. Detailed description of the algorithm for evaluating the text on a specific example is described. The advantage of this method is ability to check texts in different languages and compare them by context.

Keywords: *system plagiarism, borrowing fragments, plagiarism detection, unlawful borrowing, analysis of identity evaluation, contextual comparison*

Отрох С. І. *Державний університет телекомунікацій, Київ*
Кузьміних В. О., Осипенко М. В. *НТУУ «КПІ ім. Ігоря Сікорського», Київ*
Грищенко О. О. *Державний університет телекомунікацій, Київ*

МЕТОД ОЦІНЮВАННЯ ІДЕНТИЧНОСТІ ТЕКСТУ

Розглянуто сутність плагіату та основні його прояви. Зосереджено увагу на причинах виникнення плагіату. З'ясовано, що появі плагіату сприяють швидкий розвиток Інтернету, збільшення ресурсів та полегшення доступу до цих ресурсів. Наведено основні види плагіату, представлено класифікацію його за категоріями та способами прояву. Акцентовано увагу на необхідності розробки ефективних засобів виявлення та запобігання плагіату.

Зроблено формалізовану постановку задачі стосовно аналізу оцінювання ідентичності багатомовного тексту під час виявлення плагіату. Окреслено низку проблем, що виникають під час оцінювання ідентичності тексту та виявлення плагіату. Визначено, що основні труднощі в оцінюванні ідентичності тексту виникають через різній структурі речень і неоднозначність виявлення перекладу в текстах на різних мовах. Охарактеризовано існуючі методи оцінювання подібності між двома документами та зосереджено увагу на їх недоліках та обмеженнях. Визначено, що основний недолік існуючих методів оцінювання подібності між двома документами полягає в неможливості порівняння рядків тексту різної довжини. З метою уникнення цього недоліку запропоновано метод оцінювання ідентичності двох рядків довільної довжини шляхом вирівнювання тексту до "єдиної" мови. Докладно описано алгоритм запропонованого методу на конкретному прикладі.

Доведено, що запропонований метод є ефективним при аналізі тестових текстів. Зазначено, що перевагою такого методу є здатність до перевірки текстів на різних мовах шляхом їх порівняння за контекстом. Запропонований метод може бути використаний при побудові програмних систем анти плагіату загального використання.

Ключові слова: *системний плагіат, запозичення фрагментів; виявлення плагіату; незаконне запозичення, аналіз оцінювання ідентичності, порівняння за контекстом*

Отрох С. И. *Государственный университет телекоммуникаций, Киев*
Кузьминых В. А., Осипенко М. В. *НТУУ «КПИ им. Игоря Сикорского», Киев*
Грищенко Е.А. *Государственный университет телекоммуникаций, Киев*

МЕТОД ОЦЕНИВАНИЯ ИДЕНТИЧНОСТИ ТЕКСТА

Рассмотрены вопросы и возможности использования средств антиплагиата, виды плагиата и способы обхода средств проверки. Предложен метод для анализа оценки идентичности текста.

Подробно описан алгоритм оценки идентичности текста на конкретном примере. Преимуществом такого метода является способность проверки идентичности текстов на разных языках и их сравнение по контексту.

Ключевые слова: *системный плагиат; заимствования фрагментов; выявления плагиата, анализ оценки идентичности, сравнение по контексту*

1. Introduction

To determine plagiarism of any form is necessary to have knowledge of its possible forms and shapes, as well as the various instruments and systems of detection. Plagiarism can be manifested in any article or text in various ways, damaging one way or another. The rapid development of the Internet, increasing resources and improving access to them has given rise to the phenomenon of plagiarism. This is especially true in higher education and Science [1]. To prevent plagiarism insufficient legal, moral and ethical standards.

According to the Law of Ukraine "About copyright and related rights" on 01.13.16" plagiarism – the disclosure (publication), in whole or in part, of another's work under the name of a person who is not the author of this work."

Solving the problem of combating plagiarism in all modern forms of its manifestation, it requires the development of effective means to detect and prevent plagiarism. Currently, there are quite a number of methods, systems, and services for plagiarism detection.

For a long time created a large number of tools and techniques for the detection of plagiarism, however, there are many ways and opportunities to cheat tool for plagiarism detection. The last decades, the number of digital resources is increasing every year with great pace. And with the growth of such resources increases the ability to copy someone else's information and plagiarism. Gradually it becomes more difficult to identify the original author. It introduces readers astray, is detrimental to the author, and gives undeserved good plagiarists. Despite a large number of tools and techniques for its detection, they are still far from ideal.

Plagiarism can be:

- Full and partial;
- Simple (single source) and complex (many sources);
- Fragmented and continuous;
- Monolingual and multilingual.

Plagiarism can appear in different ways:

- Issuance of someone else's work of his own;
- Using someone else's work without permission of the author;
- Restructuring of another's work and her own.

Text plagiarism can be divided into several categories:

- Deliberate copying: copying someone else's work without proper attribution;
- Paraphrasing, as the use of someone else's work with text by changing the order of words, sentences and grammatical structures style;
- Metaphorical plagiarism - used solution idea or from another source and are issued for its own for more accurate and understandable manner;
- Self-plagiarism, that is, the author uses its own previous articles in the new work.

2. Formulation of the problem.

Multilingual word plagiarism is one of the most complex and least solved problems. Most of the means of detecting plagiarism work for one language, as due to different sentence structure and ambiguity of translation plagiarism detection in tests in different languages the problem of detecting the identity of the text more complicated.

To determine plagiarism, need to measure the similarity between two documents. Most of the funds anti-plagiarism use two types of similarity assessment:

– evaluation string similarity;

– stock assessment.

Hamming distance [2] it is one example of a similarity evaluation line. In the method, it considered the number of positions in which the respective symbols of the two different rows of the same length. The Hamming distance is the number of positions in which elements of two vectors do not match. There have been attempts to apply the Hamming distance to determine the difference between the two character strings. In this case, the determination of the Hamming distance is defined as the number of positions in which the characters in the strings being compared are different. Note that the Hamming distance is limited only suitable for determining the difference between lines since it allows comparing a string of the same length. For this reason, comparisons of rows more frequently used Levenshtein distance [3] as a minimum number of insertions of one symbol, one symbol deletion and replacement of one symbol to the other, necessary for the conversion of one line to the other [7].

Stock valuation has become the most widely used in recent years. a large number of methods for evaluating the comparison vector [4] has been presented in recent decades. One of these methods is based on the respective odds. It calculates the similarity between two vectors of equal length. Also, common methods constructed on similarity theory [5], in particular, based on the Jaccard coefficient [6], the essence of which consists in calculating the ratio of the number of common terms in compares the text to the total number of terms of the texts.

3. Description of the proposed method

The proposed algorithm solves the problem of assessing the identity of two strings of arbitrary length. In order to implement the comparison tests in different languages need to align the text to the "uniform" language [4]. This solves two problems:

1. The algorithm can be used for texts in different languages. To do this, you must have a dictionary that includes the root word in the nominative case, including synonyms.

2. Unification of the text to determine whether the author of the bypass text comparison algorithm by substituting characters trying to be visually identical, but for the word of the program will be identified as different.

To do this, replace the relevant labels. Dictionary label is a table where the key is the label and value are the words that have a similar meaning to the one or more languages. Thus, the text is not tied to one specific language and can detect similarities in the rehashed proposals.

Next, the text should be cleared of frequently used words that may affect the assessment result. It also requires a dictionary of words. The resulting standardized text should be divided into proposals that can be done with regular expressions.

Consider this example:

1. Dictionary of pointers is a table 1, where the key is the pointer and value – words having a similar meaning to the one or more languages.

2. The table contains dictionary pointers, in which the key – a pointer and the words having the same sense in one or several languages are value.

Thus we get the following expression:

[0] [1] [2] [3] [4] [1] [5] [6] [7] [8] [9] [10] [11],

[3] [0] [1] [4] [1] [6] [7] [8] [9] [10] [11] [9].

Table 1. Example pointers table

Pointer	Words
0	dictionary
1	pointer
2	is
3	table
4	key
5	value
6	word
7	have
8	identical, similar
9	sense, meaning
10	several, many
11	language

Analyzing the line, you must create a table 2 relationship marks. Thus each label will correspond to several other labels with different estimation depending on how often occur in the same sentence corresponding tags and how far apart they are in the sentence.

Table 2. First sentence analyzing

Pointer	Distances to next pointers in sentence 1	Distances to previous pointers in sentence 1
0	[1]: 1, [2] 2, [3] 3, [4], 4 [5]: 5 [6]: 6 [7] 7, [8] 8, [9] 9, [10] 10 [11] 11	-
1	[2]: 1, [3] 2, [4] 3, [5] 4, [6] 5, [7] 6, [8] 7, [9] 8 [10] 9, [11]: 10	[0]: 1
2		[0] 2, [1]: 1
3		[0] 3, [1]: 2
4		[0] 4, [1]: 3
5		[0] 5, [1]: 4
6		[0]: 6 [1]: 5
7		[0] 7 [1]: 6
8		[0] 8 [1]: 7
9		[0] 9 [1]: 8
10		[0] 10, [1]: 9
11		[0] 11, [1]: 10

Thus, analyzing the string of two. It is possible to estimate successively:

1. For the label [3] tag [0], they have met in one sentence at a distance of 3. We add to estimate ⅓

2. For label [0] is the label [1], in which the distance – 1, therefore, in the evaluation of 1/1 add.

Similarly performed subsequent steps of the method.

The maximum estimation value for each label will be 1. Dividing the sum of the estimates to obtain the value of the number of words identity factor in the range [0, 1] and that the evaluation value will rows identity.

4. Conclusions

Multilingual texts comparison analysis revealed using nonlinear dependence of the sensitivity of the method of the percentage of the volume of identical text in text fragments and fragment length, which are compared.

The proposed method is shown to be effective in the analysis of the test texts. Which suggests the possibility of its use as in the construction of software anti-plagiarism systems general use, and in particular for the construction of the university anti-plagiarism systems for the analysis of projects and dissertations of students of higher educational institutions [8].

References

1. Bolilyi V.O. (2011). Checking the Uniqueness of the Text When Evaluating Student Works of Creative Or Research Character. Science. Notes NDU im. M. Hohol. Seriya: Psycho-pedagogic science: ST. Sciences. pr. / Nizhin. keep. Univercity im. M. Gogol. Nizhin, 7. 134-145.

2. Bleikhut R. (1986). Theory and Practice of Error Control Codes Codes. Moscow: Mir, 1986. 576.

3. Levenstein V. I. (1965). Binary Codes with Correction for Deletions, Insertions and Substitutions of Characters. Reports of the USSR Academy of Sciences. 163.4:845-848.

4. Manning K.D., Raghavan, P., Schutze X. (2011). Introduction to Information Retrieval. Moscow: OOO "ID Williams", 528p.

5. Voronin Y. A. (1991). Start Similarity Theory. Novosibirsk: Nauka, Sibir Department, 128.

6. Jaccard P. (1991). Distribution de la Flore Alpine Dans le Bassin des Dranses et Dans Quelques Regions Voisines // Bull. Soc. Vaudoise sci. Natur. V. 37. Bd. 140, 241-272.

7. Komarnickaia O. I. (2014). Improvement of Algorithm Latent Semantic Analysis Fuzzy Text Information // The Modern Scientific Bulletin. 29(225). Series: Philology. Belgorod: Rusnauchkniga, 58-62.

8. Shostak I. V, Gruzdo I. V. (2013). The Computerization of the Process of Identifying Plagiarism in Student Papers // Collection of scientific works of the Military Institute of Kyiv National Taras Shevchenko University. Vol. 41, 99-109.

Список використаної літератури

1. Болілий В. О. Перевірка унікальності тексту при оцінюванні студентських робіт творчого або дослідницького характеру / В. О. Болілий, В. В. Копотій // Наукові записки

НДУ ім. М. Гоголя. Серія: Психолого-педагогічні науки : зб. наук. пр. / Ніжин. Держ. Ун-т ім. М. Гоголя. – Ніжин, 2011. – № 7. – С. 134–145.

2. Блейхут Р. Теория и практика кодов, контролирующих ошибки / Р. Блейхут // Москва: Мир, 1986. – 576 с.

3. Левенштейн В. И. Двоичные коды с исправлением выпадений, вставок и замещений символов / В. И. Левенштейн // Доклады Академии Наук СССР, 1965. 163.4:845-848.

4. Маннинг К. Д. Введение в информационный поиск / К. Д. Маннинг, П. Рагхаван, Х. Шютце. – Москва: ООО «И.Д. Вильямс», 2011. – 528 с.

5. Воронин Ю. А. Начала теории сходства / Ю. А Воронин // Новосибирск: Наука, Сибирское отделение, 1991. – 128 с.

6. Jaccard P. Distribution de la flore alpine dans le Bassin des Dranses et dans quelques regions voisines/ P. Jaccard // Bull. Soc. Vaudoise sci. Natur. 1991. V. 37. Bd. 140. P. 241-272.

7. Комарницкая О. И. Совершенствование алгоритма латентно–семантического анализа нечеткой текстовой информации / О. И. Комарницкая // Современный научный вестник. 2014. № 29(225). Серия: Фил. Науки. Белгород: Руснаучкнига. – С. 58–62.

8. Шостак И. В. Компьютеризация процесса выявления плагиата в студенческих работах / И. В. Шостак, И. В. Груздо // Сборник научных трудов Военного института Киевского национального университета имени Тараса Шевченко. Киев, 2013. Вып. 41. – С. 99–109.

Автори статті (Authors of the article)

Отрох Сергій Іванович – д.т.н., завідувач кафедри мобільних та відеоінформаційних технологій (Otrokh Serhii Ivanovych – Dr.Sci. in technic, head of Mobile and Video Information Technologies Department) Phone: +380 66 981 2439. E-mail: 2411197@ukr.net.

Кузьміних Валерій Олександрович – к.т.н., доцент кафедри АПЕПС (Kuzminykh Valeriy Oleksandrovych – PhD in technic, associate professor of APEPS Department). Phone: +380 67 466 4733. E-mail: vakuz0202@gmail.com.

Осипенко Марк Валерійович – аспірант кафедри АПЕПС (Osipenko Mark Valeriiovych – postgraduate student, APEPS Department). Phone: +380 99 048 3461. E-mail: mark.osypenko@gmail.com.

Грищенко Олена Олександрівна – аспірант, кафедра мобільних та відеоінформаційних технологій (postgraduate student, Mobile and Video Information Technologies Department). Phone: +380 63 616 9972. E-mail: elena.grischenko1@gmail.com.