Implementation of the Plagiarism detection software used in universities

Xây dựng phần mềm phát hiện đạo văn dùng trong các trường đại học

Tran Anh Vu [1], Nguyen Phan Kien [1], Nguyen Tuan Anh [1], Pham Thi Viet Huong [2]

[1] School of Electronics and Telecommunications, Hanoi University of Science and Technology, Hanoi, Viet Nam

[2] University of Engineering and Technology, Vietnam National University Hanoi, Hanoi, Vietnam

Received: July 20, 2017; Accepted: May 25, 2018

Abstract

Plagiarism is a serious problem not only in Vietnam but also in almost all other countries. Many software packages have been proposed to detect plagiarism on the content of the text. However, most of these packages are from foreign countries, it is difficult to check the documents using Vietnamese language. The plagiarism detection software named BKCheck is designed and developed to solve this problem. This software has the ability to provide the similarity among documents based on user's database. The results are reported as the percentage of similarity among documents. Besides, the software also shows the position of similar sentences or paragraphs, which allows users to easily verify the results. It has been tested and provides reliable results in School of Electronics and Telecommunications (SET), Hanoi University of Science and Technology (HUST).

Keywords: Software, Plagiarism, Database, Checking, Thesis

1. Introduction

In order to graduate from universities, students are required to submit their disssertation to the commitee. Students are expected to put their motivation, knowledge and effort on it. Hence, the dissertation is said to be students' achievement after they are done with years of studying hard in universities. Although it is not very difficult to write a dissertation, it is not easy to produce a good quality, interesting and practical dissertation. It requires students have serious attitude, dilligence and excellent writing skill. Students nowadays are very lucky with the help of the booming Internet. They can easily find some online outstanding dissertation/papers for reference. Good students take advantage of the Internet by browsing and understanding what the world is doing and find some interesting ideas for their dissertations. Some lazy students are not doing that way. They may browse the Internet, find some published papers and copy the content into their dissertation. It saved their time and finally they still have a qualified dissertation in hand. As the resourse are huge and diversed, professors may not realize their sutdents was copying from somewhere else into their dissertation. Hence, the clarification and evaluation of students may not be exact. The misclarification happens not only in universites but also in the whole education system, which can cause serious problems. For example, bad students may be misclarified as good, then they are assigned a good position at work. When facing the real work, bad students of course not do well, which can causes damage or harm to the companies. Hence, we need to prevent such a phenomena of copying without crediting the source. The plagiarism software are proposed in this paper to serve this purpose. According to the Merriam-Webster online dictionary [1] , plagiarism is:

- to steal and pass off the ideas or words of another as one's own
- use another's production without crediting the source
- present as new and original an idea or product derived from an existing source

In general, plagiarism is using ideas, sentences or products of another as one's own as default. Or in other way, use another's ideas and content without crediting the source.

Nowadays, when the Internet and computers become popular in all over the world, plagiarism has increased at fast speed and become a big challenge for the education. We should say that, plagiarism helps students meet requirements of the courses or thesis while they do not put their effort on it. Meanwhile, we cannot clarify between students who study hard and those who copy another's product. They are evaluated equally. Hence, it forms the bad habit of stealing and passing off another's content as one's own. We should have solution to solve this problem completely; otherwise, it may cause bad effect to all other students.

So what should we do? Here comes the main content of this paper. As said before, in the booming of technology, plagiarism has been so easy, we need to utilize new technology to get rid of it. Plagiarism software is proposed with the main function of

[1] Corresponding author: Tel.: (+84) 944.639.471
Email: vu.trananh@hust.edu.vn

comparing documents and finding how many percent they look alike. This paper has the following sections:

- The first part introduces some existing plagiarism software in the world as well as their advantages and disadvantages.
- The second part talks about the reality and solutions against plagiarism in Vietnam.
- The third part gives requirements for the nowadays plagiarism detection software.
- The forth part implements the plagiarism detection software which has been tested in the School of Electronics and Telecommunications, Hanoi University of Science and Technology.

2. Requirements of the Plagiarism detection software

2.1. The existing Plagiarism detection software evaluation

There are a lot Plagiarism detection software packages which are being used in the world. Following are some popular ones.

Plagiarism Checker [3]: This software allow users to compare documents by providing websites which have exactly the same content with documents. However, it is difficult to compare with Vietnamese documents, because it can not handle the Vietnamese font display. Hence, it is not effective to use this software in Vietnam

Plagiarism.net [3]: It is an online tool to check the source of documents. However, users have difficulty in account registration and search. Otherwise, it is not a free website, users have to pay in order to use it. Moreover, the result is not clear for users to have the proof of plagiarism and to evaluate the level of plagiarism.

Turnitin [4]: It is the popular plagiarism detection software package in the world. It can search and compare 300 million essays and 110,000 documents over the Internet. The Lotus University has applied this software into the plagiarism checking in the university. This software can be integrated with the Cloud services like Google Drive and Dropbox. Students can submit their work to the Tunirin Original Check. However, due to the high cost and the complicated use, it is not been widely used.

There are some other similar software packages such as Plagiarism Detector [4], Writecheck [5], Plagium [6], Dupli Checker [7], etc. They are all required account registration and costly so users are hesitate to use them.

From the above examples, we can see the main disadvantage of those software packages is that they can only be applied to English documents; it is difficult to check and test in Vietnamese. Besides, the evaluation of the percentage of similarity among documents is not presented; the validity is not exact; the method is not easy to use especially for those who are not good at English. Moreover, some software requires users to pay, which is not effective to use.

2.2. The reality and solutions against plagiarism in Vietnam universities

The conference on "academic integrity" was held in Lotus University in May 29th, 2015. More than 20 universities have taken part in this conference. They share their research experiences on plagiarism prevention. Fig. 1 shows the Duy Tan University's survey result on students' plagiarism.

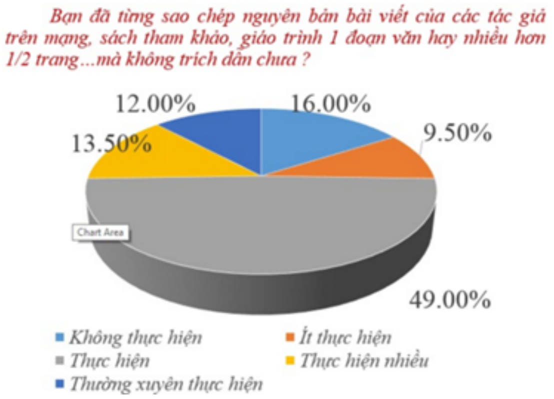


Fig. 1. Survey on students' plagiarism [8]

After the conference, most of universities have agreed to build a plagiarism detection software package. However, until now, two years from the conference, no plagiarism detection software has been officially proposed.

This work is expected to be the solution to the Vietnamese documents plagiarism detection. The software should handle with Vietnamese documents. It should not be costly to use, so a large number of people could use it. Moreover, the output should be clear, easy to read, so users can be able to evaluate the level of plagiarism in each document.

3. Introduction of BKCheck Plagiarism detection software

With the urgent demand of a Plagiarism detection software package, the BKCheck has been researched and developed, which can be satisfied all the above requirements. The software package works as follow: all the theses/dissertations are collected into a library. This work is done by an administrator, who has the right to add or remove files from library. Users are not allowed to access to the library. When a user inputs a document to check, the software will compare it with all files in the library and return the percentage of similarity between the input and each of the documents in the library.

The main program can check file by file in the library, but it takes time to open and close each file.

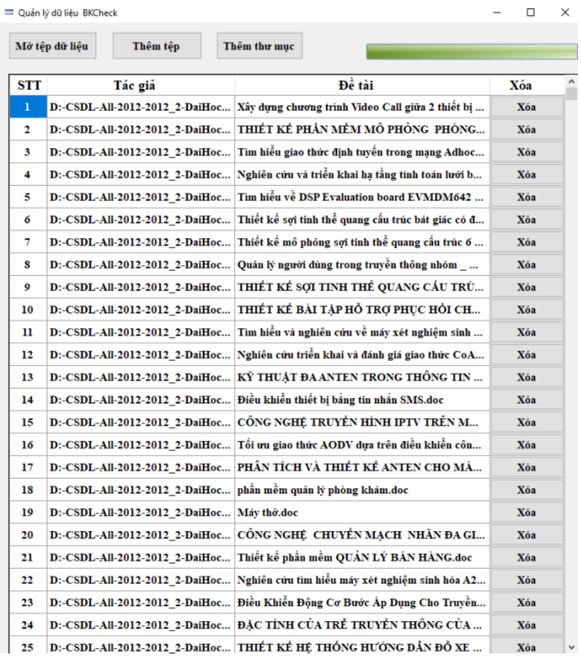
Thus, the total time for checking is increased due to the time of reopening and reclosing the files, especially when the number of files in the database is large. To prevent multiple file opening and closing, all separate files in the library are simply recorded in one database text file. The locations of each word in each file are recorded to reduce the searching time for the main program. This work is done by the database management program.

The database management program is used to manage the database text file. The interface of the database management program is as in figure 2, which represents the location and title of each file in the library.

To manage the database, administrators can add or delete files. The usable file formats are pdf, doc/docx and txt. Administrators can add each file separately, or the whole folder at the same time.

The main checking program finds and compares users' input file with each file in the library, then return the percentage of similarity.

The algorithm of checking is presented by the flow chart as in figure 3. In which:



STT	Tác giả	Đề tài	Xóa
1	D:-CSDL-All-2012-2012_2-DaiHoc...	Xây dựng chương trình Video Call giữa 2 thiết bị ...	Xóa
2	D:-CSDL-All-2012-2012_2-DaiHoc...	THIẾT KẾ PHẦN MỀM MÔ PHỎNG PHÒNG...	Xóa
3	D:-CSDL-All-2012-2012_2-DaiHoc...	Tìm hiểu giao thức định tuyến trong mạng Adhoc...	Xóa
4	D:-CSDL-All-2012-2012_2-DaiHoc...	Nghiên cứu và triển khai hạ tầng tính toán lưới b...	Xóa
5	D:-CSDL-All-2012-2012_2-DaiHoc...	Tìm hiểu về DSP Evaluation board EVMDM642 ...	Xóa
6	D:-CSDL-All-2012-2012_2-DaiHoc...	Thiết kế sợi tinh thể quang cấu trúc bát giác có đ...	Xóa
7	D:-CSDL-All-2012-2012_2-DaiHoc...	Thiết kế mô phỏng sợi tinh thể quang cấu trúc 6 ...	Xóa
8	D:-CSDL-All-2012-2012_2-DaiHoc...	Quản lý người dùng trong truyền thông nhóm _ ...	Xóa
9	D:-CSDL-All-2012-2012_2-DaiHoc...	THIẾT KẾ SỢI TINH THỂ QUANG CẤU TRÚ...	Xóa
10	D:-CSDL-All-2012-2012_2-DaiHoc...	THIẾT KẾ BÀI TẬP HỖ TRỢ PHỤC HỒI CH...	Xóa
11	D:-CSDL-All-2012-2012_2-DaiHoc...	Tìm hiểu và nghiên cứu về máy xét nghiệm sinh ...	Xóa
12	D:-CSDL-All-2012-2012_2-DaiHoc...	Nghiên cứu triển khai và đánh giá giao thức CoA...	Xóa
13	D:-CSDL-All-2012-2012_2-DaiHoc...	KỸ THUẬT ĐA ANTEN TRONG THÔNG TIN ...	Xóa
14	D:-CSDL-All-2012-2012_2-DaiHoc...	Điều khiển thiết bị bằng tin nhắn SMS.doc	Xóa
15	D:-CSDL-All-2012-2012_2-DaiHoc...	CÔNG NGHỆ TRUYỀN HÌNH IPTV TRÊN M...	Xóa
16	D:-CSDL-All-2012-2012_2-DaiHoc...	Tối ưu giao thức AODV dựa trên điều khiển côn...	Xóa
17	D:-CSDL-All-2012-2012_2-DaiHoc...	PHÂN TÍCH VÀ THIẾT KẾ ANTEN CHO MÁ...	Xóa
18	D:-CSDL-All-2012-2012_2-DaiHoc...	phần mềm quản lý phòng khám.doc	Xóa
19	D:-CSDL-All-2012-2012_2-DaiHoc...	Máy thở.doc	Xóa
20	D:-CSDL-All-2012-2012_2-DaiHoc...	CÔNG NGHỆ CHUYỂN MẠCH NHÃN ĐA GI...	Xóa
21	D:-CSDL-All-2012-2012_2-DaiHoc...	Thiết kế phần mềm QUẢN LÝ BÁN HÀNG.doc	Xóa
22	D:-CSDL-All-2012-2012_2-DaiHoc...	Nghiên cứu tìm hiểu máy xét nghiệm sinh hóa A2...	Xóa
23	D:-CSDL-All-2012-2012_2-DaiHoc...	Điều Khiển Động Cơ Bước Áp Dụng Cho Truyền...	Xóa
24	D:-CSDL-All-2012-2012_2-DaiHoc...	ĐẶC TÍNH CỦA TRỄ TRUYỀN THÔNG CỦA ...	Xóa
25	D:-CSDL-All-2012-2012_2-DaiHoc...	THIẾT KẾ HỆ THỐNG HƯỚNG DẪN ĐỖ XE ...	Xóa

Fig. 2. Interface of the database management program

A, B is the word array of the tested document and database document, respectively. *A.Count*, *B.Count* is the number of words of the tested and the database document.

M is the total number of copied words.

g is the default number of copied words, it depends on users' selection

<u>Algorithm explanation</u>

The algorithm works as follow: Start from the first word of the tested document, search for it in the database. There are two cases:

<u>Case 1</u>. There is no such word in the database => move to second word of the tested document and continue to search for the similar word in the database. The algorithm works in this way until it reaches the last word of the tested document.

<u>Case 2</u>. Find the similar word in the database, then:

Step 1: Count the similar words in the two documents, in the place we find them

Step 2: Continue the search process and count until reach the end of tested document

Step 3: Compare the numbers and output the maximum

Step 4:

- ➢ If the maximum is greater than the default number of copied words (chosen to be 10), then the number of copied words is assigned to the maximum. The algorithm moves to the next word and continue.
- ➢ If the maximum is smaller or equal to the default number of copied words, the algorithm moves to the next word and continue.

When the search is done, the software will calculate the number of the copied words of the tested document. The percentage will be calculated as:

Percentage of similarity= (number of copied words)/(total words of the document)×100%

The document in the library which has the highest percentage of similarity will be updated continuously and presented at the top of the result table. Users can stop the program at any time to check more detail at any file they want.

In order to increase the chance to find the document which has similar content, the order of searching will prioritize documents which have the title's keywords similar to those of the input.

With the above algorithm, the result is optimized; the copied words will be found disregards of their places or formats.

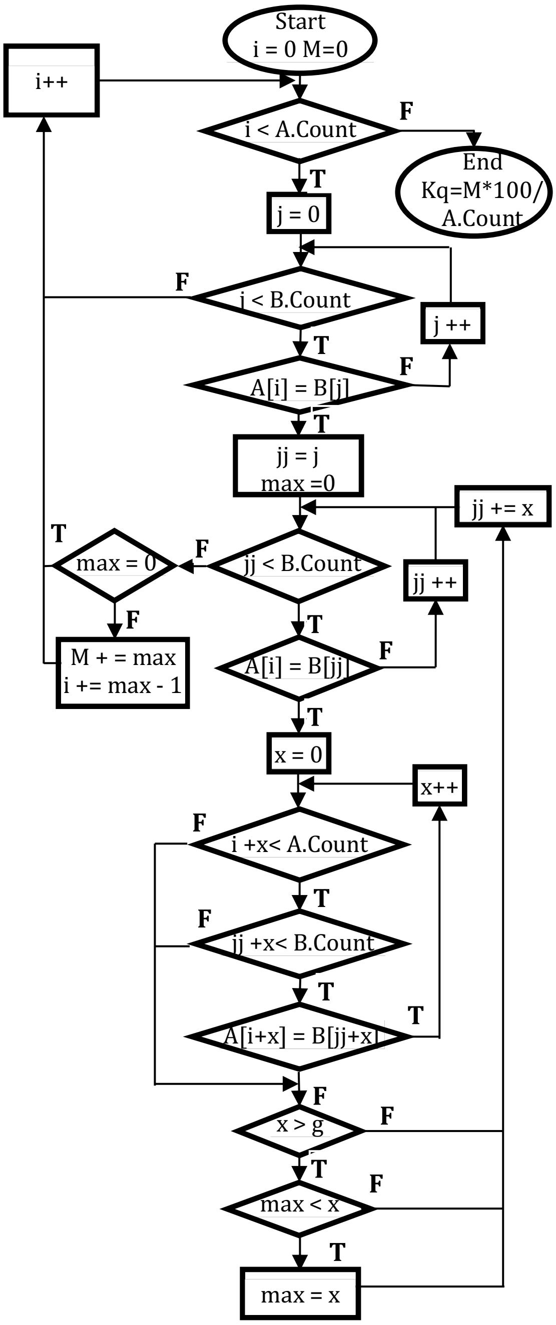
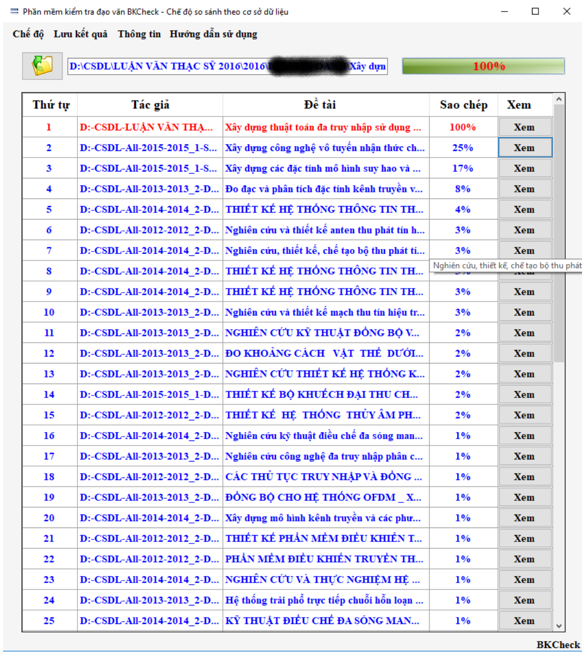


Fig. 3. The algorithm flowchart

4. Results

All documents in the library will be uploaded to the database of the software by administrators. The content of the database can be added or remove by the administrator. When a document needs to be checked, it is loaded to the program, the software will then automatically compare it with the database. It will take only about 30 second to check a document with a database of 1300 documents. The result appears in figure 4.

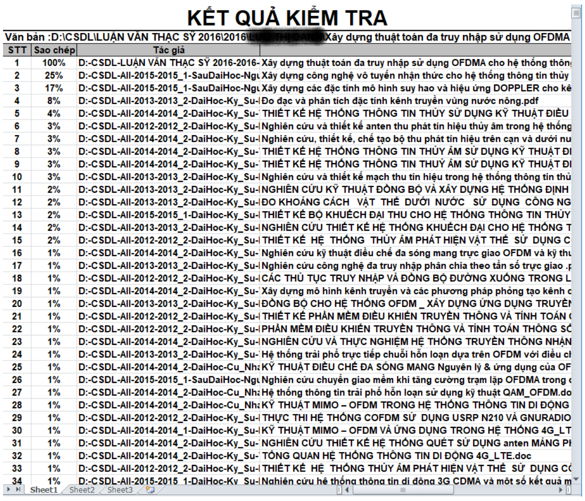

Phần mềm kiểm tra đạo văn BKCheck - Chế độ so sánh theo cơ sở dữ liệu
Chế độ
Lưu kết quả
Thông tin
Hướng dẫn sử dụng
100%
BKCheck

D:\CSDL\LUẬN VĂN THẠC SỸ 2016\2016\[redacted]Xây dựn

Thứ tự	Tác giả	Đề tài	Sao chép	Xem
1	D:-CSDL-LUẬN VĂN THẠ...	Xây dựng thuật toán đa truy nhập sử dụng ...	100%	Xem
2	D:-CSDL-All-2015-2015_1-S...	Xây dựng công nghệ vô tuyến nhận thức ch...	25%	Xem
3	D:-CSDL-All-2015-2015_1-S...	Xây dựng các đặc tính mô hình suy hao và ...	17%	Xem
4	D:-CSDL-All-2013-2013_2-D...	Đo đạc và phân tích đặc tính kênh truyền v...	8%	Xem
5	D:-CSDL-All-2014-2014_2-D...	THIẾT KẾ HỆ THỐNG THÔNG TIN TH...	4%	Xem
6	D:-CSDL-All-2012-2012_2-D...	Nghiên cứu và thiết kế anten thu phát tín h...	3%	Xem
7	D:-CSDL-All-2014-2014_2-D...	Nghiên cứu, thiết kế, chế tạo bộ thu phát tí...	3%	Xem
8	D:-CSDL-All-2014-2014_2-D...	THIẾT KẾ HỆ THỐNG THÔNG TIN TH...	3%	Xem
9	D:-CSDL-All-2014-2014_2-D...	THIẾT KẾ HỆ THỐNG THÔNG TIN TH...	3%	Xem
10	D:-CSDL-All-2013-2013_2-D...	Nghiên cứu và thiết kế mạch thu tín hiệu tr...	3%	Xem
11	D:-CSDL-All-2013-2013_2-D...	NGHIÊN CỨU KỸ THUẬT ĐỒNG BỘ V...	2%	Xem
12	D:-CSDL-All-2013-2013_2-D...	ĐO KHOẢNG CÁCH VẬT THỂ DƯỚI...	2%	Xem
13	D:-CSDL-All-2013-2013_2-D...	NGHIÊN CỨU THIẾT KẾ HỆ THỐNG K...	2%	Xem
14	D:-CSDL-All-2015-2015_1-D...	THIẾT KẾ BỘ KHUẾCH ĐẠI THU CH...	2%	Xem
15	D:-CSDL-All-2012-2012_2-D...	THIẾT KẾ HỆ THỐNG THỦY ÂM PH...	2%	Xem
16	D:-CSDL-All-2014-2014_2-D...	Nghiên cứu kỹ thuật điều chế đa sóng man...	1%	Xem
17	D:-CSDL-All-2013-2013_2-D...	Nghiên cứu công nghệ đa truy nhập phân c...	1%	Xem
18	D:-CSDL-All-2012-2012_2-D...	CÁC THỦ TỤC TRUY NHẬP VÀ ĐỒNG ...	1%	Xem
19	D:-CSDL-All-2013-2013_2-D...	ĐỒNG BỘ CHO HỆ THỐNG OFDM _ X...	1%	Xem
20	D:-CSDL-All-2014-2014_2-D...	Xây dựng mô hình kênh truyền và các phư...	1%	Xem
21	D:-CSDL-All-2012-2012_2-D...	THIẾT KẾ PHẦN MỀM ĐIỀU KHIỂN T...	1%	Xem
22	D:-CSDL-All-2012-2012_2-D...	PHẦN MỀM ĐIỀU KHIỂN TRUYỀN TH...	1%	Xem
23	D:-CSDL-All-2014-2014_2-D...	NGHIÊN CỨU VÀ THỰC NGHIỆM HỆ ...	1%	Xem
24	D:-CSDL-All-2013-2013_2-D...	Hệ thống trải phổ trực tiếp chuỗi hỗn loạn ...	1%	Xem
25	D:-CSDL-All-2014-2014_2-D...	KỸ THUẬT ĐIỀU CHẾ ĐA SÓNG MAN...	1%	Xem

Nghiên cứu, thiết kế, chế tạo bộ thu phát

Fig. 4. Display of the result



KẾT QUẢ KIỂM TRA

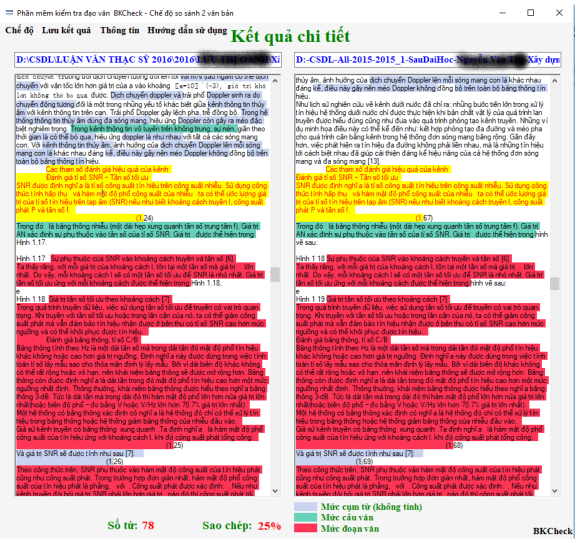
Văn bản :D:\CSDL\LUẬN VĂN THẠC SỸ 2016\2016\[redacted]\Xây dựng thuật toán đa truy nhập sử dụng OFDMA

STT	Sao chép	Tác giả	
1	100%	D:-CSDL-LUẬN VĂN THẠC SỸ 2016-2016-	Xây dựng thuật toán đa truy nhập sử dụng OFDMA cho hệ thống thôn
2	25%	D:-CSDL-All-2015-2015_1-SauDaiHoc-Ngu	Xây dựng công nghệ vô tuyến nhận thức cho hệ thống thông tin thủy
3	17%	D:-CSDL-All-2015-2015_1-SauDaiHoc-Ngu	Xây dựng các đặc tính mô hình suy hao và hiệu ứng DOPPLER cho kê
4	8%	D:-CSDL-All-2013-2013_2-DaiHoc-Ky_Su-I	Đo đạc và phân tích đặc tính kênh truyền vùng nước nông.pdf
5	4%	D:-CSDL-All-2014-2014_2-DaiHoc-Ky_Su-T	THIẾT KẾ HỆ THỐNG THÔNG TIN THỦY SỬ DỤNG KỸ THUẬT ĐIỀU
6	3%	D:-CSDL-All-2012-2012_2-DaiHoc-Ky_Su-I	Nghiên cứu và thiết kế anten thu phát tín hiệu thủy âm trong hệ thống
7	3%	D:-CSDL-All-2014-2014_2-DaiHoc-Ky_Su-I	Nghiên cứu, thiết kế, chế tạo bộ thu phát tín hiệu trên cạn và dưới nu
8	3%	D:-CSDL-All-2014-2014_2-DaiHoc-Ky_Su-T	THIẾT KẾ HỆ THỐNG THÔNG TIN THỦY ÂM SỬ DỤNG KỸ THUẬT ĐI
9	3%	D:-CSDL-All-2014-2014_2-DaiHoc-Ky_Su-T	THIẾT KẾ HỆ THỐNG THÔNG TIN THUỶ ÂM SỬ DỤNG KỸ THUẬT ĐI
10	3%	D:-CSDL-All-2013-2013_2-DaiHoc-Ky_Su-I	Nghiên cứu và thiết kế mạch thu tín hiệu trong hệ thống thông tin thủ
11	2%	D:-CSDL-All-2013-2013_2-DaiHoc-Ky_Su-I	NGHIÊN CỨU KỸ THUẬT ĐỒNG BỘ VÀ XÂY DỰNG HỆ THỐNG ĐỊNH
12	2%	D:-CSDL-All-2013-2013_2-DaiHoc-Ky_Su-I	ĐO KHOẢNG CÁCH VẬT THỂ DƯỚI NƯỚC SỬ DỤNG CÔNG NG
13	2%	D:-CSDL-All-2015-2015_1-DaiHoc-Ky_Su-I	THIẾT KẾ BỘ KHUẾCH ĐẠI THU CHO HỆ THỐNG THÔNG TIN THỦY
14	2%	D:-CSDL-All-2013-2013_2-DaiHoc-Ky_Su-I	NGHIÊN CỨU THIẾT KẾ HỆ THỐNG KHUẾCH ĐẠI CHO HỆ THỐNG T
15	2%	D:-CSDL-All-2012-2012_2-DaiHoc-Ky_Su-I	THIẾT KẾ HỆ THỐNG THỦY ÂM PHÁT HIỆN VẬT THỂ SỬ DỤNG C
16	1%	D:-CSDL-All-2014-2014_2-DaiHoc-Ky_Su-T	Nghiên cứu kỹ thuật điều chế đa sóng mang trực giao OFDM và kỹ thu
17	1%	D:-CSDL-All-2013-2013_2-DaiHoc-Ky_Su-I	Nghiên cứu công nghệ đa truy nhập phân chia theo tần số trực giao .p
18	1%	D:-CSDL-All-2012-2012_2-DaiHoc-Ky_Su-I	CÁC THỦ TỤC TRUY NHẬP VÀ ĐỒNG BỘ ĐƯỜNG XUỐNG TRONG L
19	1%	D:-CSDL-All-2014-2014_2-DaiHoc-Ky_Su-T	Xây dựng mô hình kênh truyền và các phương pháp phỏng tạo kênh c
20	1%	D:-CSDL-All-2013-2013_2-DaiHoc-Ky_Su-I	ĐỒNG BỘ CHO HỆ THỐNG OFDM _ XÂY DỰNG ỨNG DỤNG TRUYỀN
21	1%	D:-CSDL-All-2012-2012_2-DaiHoc-Ky_Su-I	THIẾT KẾ PHẦN MỀM ĐIỀU KHIỂN TRUYỀN THÔNG VÀ TÍNH TOÁN (
22	1%	D:-CSDL-All-2012-2012_2-DaiHoc-Ky_Su-I	PHẦN MỀM ĐIỀU KHIỂN TRUYỀN THÔNG VÀ TÍNH TOÁN THÔNG SỐ
23	1%	D:-CSDL-All-2014-2014_2-DaiHoc-Ky_Su-I	NGHIÊN CỨU VÀ THỰC NGHIỆM HỆ THỐNG TRUYỀN THÔNG NHẬN
24	1%	D:-CSDL-All-2013-2013_2-DaiHoc-Ky_Su-T	Hệ thống trải phổ trực tiếp chuỗi hỗn loạn dựa trên OFDM với điều ch
25	1%	D:-CSDL-All-2014-2014_2-DaiHoc-Cu_Nha	KỸ THUẬT ĐIỀU CHẾ ĐA SÓNG MANG Nguyên lý & ứng dụng của OF
26	1%	D:-CSDL-All-2015-2015_1-SauDaiHoc-Ngu	Nghiên cứu chuyển giao mềm khi tăng cường trạm lặp OFDMA trong (
27	1%	D:-CSDL-All-2014-2014_2-DaiHoc-Cu_Nha	Hệ thống thông tin trải phổ hỗn loạn sử dụng kỹ thuật QAM_OFDM.do
28	1%	D:-CSDL-All-2014-2014_2-DaiHoc-Cu_Nha	KỸ THUẬT MIMO – OFDM TRONG HỆ THỐNG THÔNG TIN DI ĐỘNG
29	1%	D:-CSDL-All-2012-2012_2-DaiHoc-Ky_Su-I	THỰC THI HỆ THỐNG COFDM SỬ DỤNG USRP N210 VÀ GNURADIO
30	1%	D:-CSDL-All-2014-2014_1-DaiHoc-Ky_Su-I	KỸ THUẬT MIMO – OFDM VÀ ỨNG DỤNG TRONG HỆ THỐNG 4G_LT
31	1%	D:-CSDL-All-2014-2014_2-DaiHoc-Ky_Su-T	NGHIÊN CỨU THIẾT KẾ HỆ THỐNG QUÉT SỬ DỤNG anten MẢNG PH
32	1%	D:-CSDL-All-2014-2014_2-DaiHoc-Ky_Su-T	TỔNG QUAN HỆ THỐNG THÔNG TIN DI ĐỘNG 4G_LTE.doc
33	1%	D:-CSDL-All-2012-2012_2-DaiHoc-Ky_Su-I	THIẾT KẾ HỆ THỐNG THỦY ÂM PHÁT HIỆN VẬT THỂ SỬ DỤNG C(
34	1%	D:-CSDL-All-2015-2015_1-DaiHoc-Ky_Su-I	Nghiên cứu hệ thống thông tin di động 3G CDMA và một số kết quả m

Fig. 5. Result extraction to Excel file

The fast checking brings users the similarity between the tested documents and those in the library. The result after checking can be exported in Excel format (figure 5). Users will know how many percent the tested document looks alike with the document in the database with the order of highest similarly percentage at top position.

In some cases, users want to have the exact information about positions of the similar paragraphs or sentences between tested documents and those in the database; the program will allow users to change to the direct comparison mode. With this mode, the position of each sentence or paragraph will be marked and colored for a convenient checking. Results are illustrated in figure 6.



D:\CSDL\LUẬN VĂN THẠC SỸ 2016\2016\[redacted]\Xá

dịch chuyển trường đối dịch chuyển tương đối lên tới vài m/s (tàu ngầm có thể dịch chuyển với vận tốc lớn hơn giá trị của a vào khoảng [a-10] ¯(-3), giá trị khó lơn không thể bỏ qua được. Dịch chuyển doppler và trải phổ Doppler sinh ra do chuyển động tương đối là một trong những yếu tố khác biệt giữa kênh thông tin thủy âm với kênh thông tin trên cạn. Trải phổ Doppler gây lệch pha, trễ đồng bộ. Trong hệ thống thông tin thủy âm dùng đa sóng mang, hiệu ứng Doppler còn gây ra méo đặc biệt nghiêm trọng. Trong kênh thông tin vô tuyến trên không trung, sự nén, giãn theo thời gian là có thể bỏ qua, hiệu ứng doppler là như nhau với tất cả các sóng mang con. Với kênh thông tin thủy âm, ảnh hưởng của dịch chuyển Doppler lên mỗi sóng mang con là khác nhau đáng kể, điều này gây nên méo Doppler không đồng bộ trên toàn bộ băng thông tín hiệu.

Các tham số đánh giá hiệu quả của kênh

Đánh giá tỉ số SNR – Tần số tối ưu

SNR được định nghĩa là tỉ số công suất tín hiệu trên công suất nhiễu. Sử dụng công thức tính hấp thụ và hàm mật độ phổ công suất của nhiễu ta có thể ước lượng giá trị của tỉ số tín hiệu trên tạp âm (SNR) nếu như biết khoảng cách truyền l, công suất phát P và tần số f.

(1.24)

Trong đó là băng thông nhiễu (một dải hẹp xung quanh tần số trung tâm f). Giá trị AN xác định sự phụ thuộc vào tần số của tỉ số SNR. Giá trị được thể hiện trong Hình 1.17.

Hình 1.17 Sự phụ thuộc của SNR vào khoảng cách truyền và tần số [6]

Ta thấy rằng, với mỗi giá trị của khoảng cách l, tồn tại một tần số mà giá trị lớn nhất. Do vậy, mỗi khoảng cách l sẽ có một tần số tối ưu để SNR là nhỏ nhất. Giá trị tần số tối ưu ứng với mỗi khoảng cách được thể hiện trong Hình 1.18.

e

Hình 1.18 Giá trị tần số tối ưu theo khoảng cách [7]

Trong quá trình truyền dữ liệu, việc sử dụng tần số tối ưu để truyền có vai trò quan trọng. Khi truyền với tần số tối ưu hoặc trong lân cận của nó, ta có thể giảm công suất phát mà vẫn đảm bảo tín hiệu nhận được ở bên thu có tỉ số SNR cao hơn mức ngưỡng và có thể khôi phục được tín hiệu.

Đánh giá băng thông, tỉ số C/B

Băng thông tính theo Hz là một dải tần số mà trong dải tần đó mật độ phổ tín hiệu khác không hoặc cao hơn giá trị ngưỡng. Định nghĩa này được dùng trong việc tính toán tỉ số lấy mẫu sao cho thỏa mãn định lý lấy mẫu. Bởi vì dải biên độ khác không có thể rất rộng hoặc vô hạn, nên khái niệm băng thông sẽ được mở rộng hơn. Băng thông còn được định nghĩa là dải tần trong đó mật độ phổ tín hiệu cao hơn một mức ngưỡng nhất định. Thông thường, khái niệm băng thông được hiểu theo nghĩa băng thông 3-dB. Tức là dải tần mà trong dải đó thì hàm mật độ phổ lớn hơn nửa giá trị lớn nhất(hoặc biên độ phổ – đo bằng V hoặc V/Hz lớn hơn 70.7% giá trị lớn nhất).

Một hệ thống có băng thông xác định có nghĩa là hệ thống đó chỉ có thể xử lý tín hiệu trong băng thông hoặc hệ thống giảm băng thông của nhiễu đầu vào.

Giả sử kênh truyền có băng thông xung quanh .Ta định nghĩa là hàm mật độ phổ công suất của tín hiệu ứng với khoảng cách l, khi đó công suất phát tổng cộng:

(1.25)

Và giá trị SNR sẽ được tính như sau [7]:

(1.26)

Theo công thức trên, SNR phụ thuộc vào hàm mật độ công suất của tín hiệu phát cũng như công suất phát. Trong trường hợp đơn giản nhất, hàm mật độ phổ công suất của tín hiệu phát là phẳng, với . Công suất phát được xác định: . Nếu như kênh truyền đòi hỏi giá trị SNR phải lớn hơn giá trị nào đó thì công suất phát tối

D:-CSDL-All-2015-2015_1-SauDaiHoc-[redacted]-Xây dựn

thủy âm, ảnh hưởng của dịch chuyển Doppler lên mỗi sóng mang con là khác nhau đáng kể, điều này gây nên méo Doppler không đồng bộ trên toàn bộ băng thông tín hiệu.

Như lịch sử nghiên cứu về kênh dưới nước đã chỉ ra: những bước tiến lớn trong xử lý tín hiệu hệ thống dưới nước chỉ được thực hiện khi bản chất vật lý của quá trình lan truyền được hiểu đúng cũng như đưa vào quá trình phỏng tạo kênh truyền. Những ví dụ minh họa điều này có thể kể đến như: kết hợp phỏng tạo đa đường và méo pha cho quá trình cân bằng kênh trong hệ thống đơn sóng mang băng rộng. Gần đây hơn, việc phát hiện ra tín hiệu đa đường không phải liền nhau, mà là những tín hiệu tới cách biệt nhau đã giúp cải thiện đáng kể hiệu năng của cả hệ thống đơn sóng mang và đa sóng mang [13].

Các tham số đánh giá hiệu quả của kênh

Đánh giá tỉ số SNR – Tần số tối ưu

SNR được định nghĩa là tỉ số công suất tín hiệu trên công suất nhiễu. Sử dụng công thức tính hấp thụ và hàm mật độ phổ công suất của nhiễu ta có thể ước lượng giá trị của tỉ số tín hiệu trên tạp âm (SNR) nếu như biết khoảng cách truyền l, công suất phát P và tần số f.

(1.67)

Trong đó là băng thông nhiễu (một dải hẹp xung quanh tần số trung tâm f). Giá trị AN xác định sự phụ thuộc vào tần số của tỉ số SNR. Giá trị được thể hiện trong hình vẽ sau:

Hình 1 18 Sự phụ thuộc của SNR vào khoảng cách truyền và tần số [6]

Ta thấy rằng, với mỗi giá trị của khoảng cách l, tồn tại một tần số mà giá trị lớn nhất. Do vậy, mỗi khoảng cách l sẽ có một tần số tối ưu để SNR là nhỏ nhất. Giá trị tần số tối ưu ứng với mỗi khoảng cách được thể hiện trong hình vẽ sau:

e

Hình 1 19 Giá trị tần số tối ưu theo khoảng cách [7]

Trong quá trình truyền dữ liệu, việc sử dụng tần số tối ưu để truyền có vai trò quan trọng. Khi truyền với tần số tối ưu hoặc trong lân cận của nó, ta có thể giảm công suất phát mà vẫn đảm bảo tín hiệu nhận được ở bên thu có tỉ số SNR cao hơn mức ngưỡng và có thể khôi phục được tín hiệu.

Đánh giá băng thông, tỉ số C/B

Băng thông tính theo Hz là một dải tần số mà trong dải tần đó mật độ phổ tín hiệu khác không hoặc cao hơn giá trị ngưỡng. Định nghĩa này được dùng trong việc tính toán tỉ số lấy mẫu sao cho thỏa mãn định lý lấy mẫu. Bởi vì dải biên độ khác không có thể rất rộng hoặc vô hạn, nên khái niệm băng thông sẽ được mở rộng hơn. Băng thông còn được định nghĩa là dải tần trong đó mật độ phổ tín hiệu cao hơn một mức ngưỡng nhất định. Thông thường, khái niệm băng thông được hiểu theo nghĩa băng thông 3-dB. Tức là dải tần mà trong dải đó thì hàm mật độ phổ lớn hơn nửa giá trị lớn nhất(hoặc biên độ phổ – đo bằng V hoặc V/Hz lớn hơn 70.7% giá trị lớn nhất).

Một hệ thống có băng thông xác định có nghĩa là hệ thống đó chỉ có thể xử lý tín hiệu trong băng thông hoặc hệ thống giảm băng thông của nhiễu đầu vào.

Giả sử kênh truyền có băng thông xung quanh .Ta định nghĩa là hàm mật độ phổ công suất của tín hiệu ứng với khoảng cách l, khi đó công suất phát tổng cộng:

(1.68)

Và giá trị SNR sẽ được tính như sau [7]:

(1.69)

Theo công thức trên, SNR phụ thuộc vào hàm mật độ công suất của tín hiệu phát cũng như công suất phát. Trong trường hợp đơn giản nhất, hàm mật độ phổ công suất của tín hiệu phát là phẳng, với . Công suất phát được xác định: . Nếu như kênh truyền đòi hỏi giá trị SNR phải lớn hơn giá trị nào đó thì công suất phát tối

Số từ: 78

Sao chép: 25%

- ▨ Mức cụm từ (không tính)
- 🟩 Mức câu văn
- 🟥 Mức đoạn văn

Fig. 6. Diplay file comparison in detail

The software also allow users to define parameters for the similarity, for example, paragraph-level similarity is marked as red (the default number of words > 50), sentence-level similarity is marked as blue (the default number of words > 10) and word-level similarity is marked as gray (this part will not be counted toward the similarity percentage).

In order for users to check conveniently, the software will mark the same position of the two files in which users are observing. If users click on any position in one file, the same position in the other file will be colored accordingly. This position is illustrated in yellow.

An improvement of this software compared to the existing ones is the result validation. By visualizing the position and defining the length of similar words, sentences or paragraphs, users are able to validate the results.

Hence, the BKCheck has following advantages:

- ✓ The interface is easy to use (in Vietnamese). Results are designed visually and conveniently.
- ✓ The speed of testing is fast (approximately 30 seconds to check a document for a database over 1300 documents)
- ✓ The accuracy is high. The results are displayed in detail which is easy to evaluate tested documents. It has the ability of searching similar paragraphs even when their formats are changed (words adding, sentences stop, line stop, marks changing, etc)
- ✓ The software is able to handle with Vietnamese documents (and English, of course) effectively.
- ✓ The software is able to handle with different document formats: .doc, .pdf, .txt
- ✓ The database is built by users; hence it is easy to manage.
- ✓ However, there are some disadvantages we need to get over:
- It is not able to compare images or graphs
- Due to the fact that the database is built by users, so if we want to use the software in a large scale, we need the integration of the database from different parties.

5. Conclusion

The BKCheck plagiarism detection software package has been researched and tested in School of Telecommunications and Electronics, Hanoi University of Science and Technology. The software satisfies the requirements such as the ability to test Vietnamese documents, friendly interface, easy database management, especially the ability to change the default number of copied words depending on users. The software has been tested with the database of more than 1300 documents, and the result is validated. In the coming days, authors will put the software online to make it easier for the testing process.

Acknowledgments

This research is funded by the Hanoi University of Science and Technology (HUST) under project number T2016-PC-120.

The authors thank you a lot for the help of the administration as well as all the lecturers in the School of Telecommunications and Electronics, Hanoi University of Science and Technology in the BKCheck testing process.

References

[1] https://www.merriam-webster.com/dictionary/plagiarize

[2] www.grammarly.com

[3] Plagiarisma.net

[4] Turnitin.com

[5] plagiarism-detector.com

[6] WriteCheck.com

[7] www.plagium.com

[8] www.duplichecker.com

[9] http://dantri.com.vn/giao-duc-khuyen-hoc/nan-dao-van-ngay-cang-gia-tang-1433666610.htm