ŠTUDENTSKÉ ZADANIA – PLAGIÁTY?

Henrieta Telepovská, František Gajdoš, Marek Szanyi
Technická univerzita v Košiciach

Abstrakt.
Príspevok sa zaoberá problémom plagiátorstva študentských zadaní na akademickej pôde. Záverečné práce bakalárskeho ako aj inžinierskeho resp. magisterského štúdia boli v predošlom akademickom roku podrobené kontrole plagiátorstva prostredníctvom centrálneho registra záverečných prác. S problémom kopírovania, či obchodovania so študentskými prácami sa stretávame nielen pri záverečných prácach, ale pri tvorbe študentských zadaní v rámci jednotlivých predmetov. Na Katedre počítačov a informatiky FEI Technickej univerzity v Košiciach bol v rámci diplomových prác vytvorený prototyp systému na detekciu plagiátov v študentských zadaniach. Je zložený z dvoch základných modulov – modul konverzie vybraných dokumentových formátov na text a modul odhaľovania podobnosti v textoch. Systém bol otestovaný na vybranej vzorke zadaní predmetov, ktoré zabezpečuje Katedra počítačov a informatiky.
Kľúčové slová: plagiát, konverzia dokumentov, podobnosť v textoch

ÚVOD

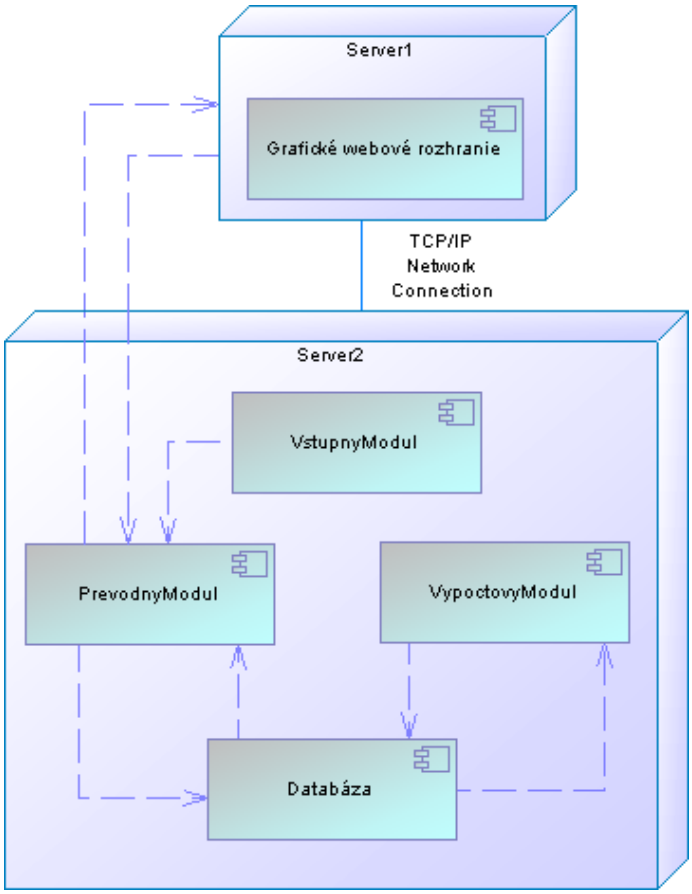
Problém plagiátov sa v posledných rokoch dostal do popredia aj vďaka značnej medializácii v rôznych komunikačných prostriedkoch. Problém plagiátov sa netýka iba záverečných prác, na ktoré bola zameraná pozornosť, ale stretávame sa s ním už pri študentských zadaniach. Dnešný svet počítačov a internetu, kde má k dispozícii ľubovoľný používateľ rýchlo a veľké množstvo informácií z rôznych zdrojov, sa stáva lákadlom pre vytvorenie novej práce. Pohodlná metóda Skopíruj a Vlož to umožňuje za relatívne krátky čas. Podľa [1,2] plagiátorstvo je úmyselné, ale aj neúmyselné používanie cudzej práce bez uvedenia citácie a vyhlasovanie tejto práce za svoju vlastnú.
V súčasnosti existujú komerčné aj nekomerčné riešenia a systémy na odhaľovanie plagiátov [5,7]. Systém popisovaný v príspevku je zameraný na kontrolu dokumentácie študentských zadaní a v súčasnosti sa pripravuje sa modul na kontrolu zdrojových textov.

ARCHITEKTÚRA SYSTÉMU

Na Obr. 1 je uvedená architektúra prototypu systému pre odovzdávanie zadaní – Elektronická kontrola zadaní [8,11,12]. Vstupný, prevodný a výpočtový modul a databáza sa nachádzajú na jednom serveri.

Grafické webové rozhranie slúži na komunikáciu používateľov so systémom. Medzi používateľov patria študenti, vyučujúci a administrátor. Študenti prostredníctvom webového rozhrania odosielajú do systému archívny súbor, ktorý obsahuje dokumenty podľa predpísanej štruktúry. Ďalšou funkciou rozhrania pre študentov je spätná väzba, kde si môžu pozrieť hodnotenie svojho zadania a percentuálnu zhodu svojho konkrétneho zadania s ostanými prácami študentov.
Vyučujúci si prostredníctvom rozhrania prezerajú výsledky kontroly plagiátorstva a hodnotenie jednotlivých zadaní. Vyučujúci má k dispozícii všetky potrebné informácie o študentských prácach, a v prípade podobnosti dvoch zadaní má možnosť pozrieť si príslušnú dvojicu, kde sú zvýraznené podobné časti. K ďalším funkciám, ktoré sú dostupné pre

pedagógov, patria definovanie adresárovej štruktúry zadania a priradenie študentov k predmetu.
Funkcie administrátora sú rozdelené do troch základných okruhov – správa zadaní, správa predmetov a vyučujúcich a správa systému.



Obr. 1 Architektúra systému [11,12]

Vstupný modul je určený na príjem zadaní z grafického webového rozhrania. Zadania sú v archívnom súbore formátu zip. Modul archívny súbor rozbalí a zároveň kontroluje, či súbory tvoriace zadanie majú predpísanú štruktúru archívu. Štruktúru archívu definuje pedagóg, prip. administrátor systému. Ak štruktúra súborov je správna, zadanie je odoslané na ďalšie spracovanie do prevodného modulu. V prípade chyby je študent upozornený na nedostatky.

Prevodný modul realizuje prevod dokumentových formátov do čistého textu. Po prevode a uložení všetkých dokumentov tvoriacich zadanie do databázy je aktivovaný výpočtový modul, ktorý porovná nové zadanie s ostatnými zadaniami systému. Ďalšou úlohou prevodného modulu je zvýraznenie podobných častí dvoch dokumentov, ktoré sa zobrazia v grafickom webovom rozhraní.

Výpočtový modul porovnáva navzájom zadania študentov a vypočíta percentuálnu zhodu medzi zadaniami na základe algoritmu zvoleného administrátorom systému. Vstupom výpočtového modulu sú čisté texty jednotlivých zadaní. Výstupom je percentuálna zhoda medzi zadaniami a indexy zhodných symbolov, pomocou ktorých je možné zvýrazniť podobné časti dvoch dokumentov. Výsledok porovnávania je uložený do databázy.

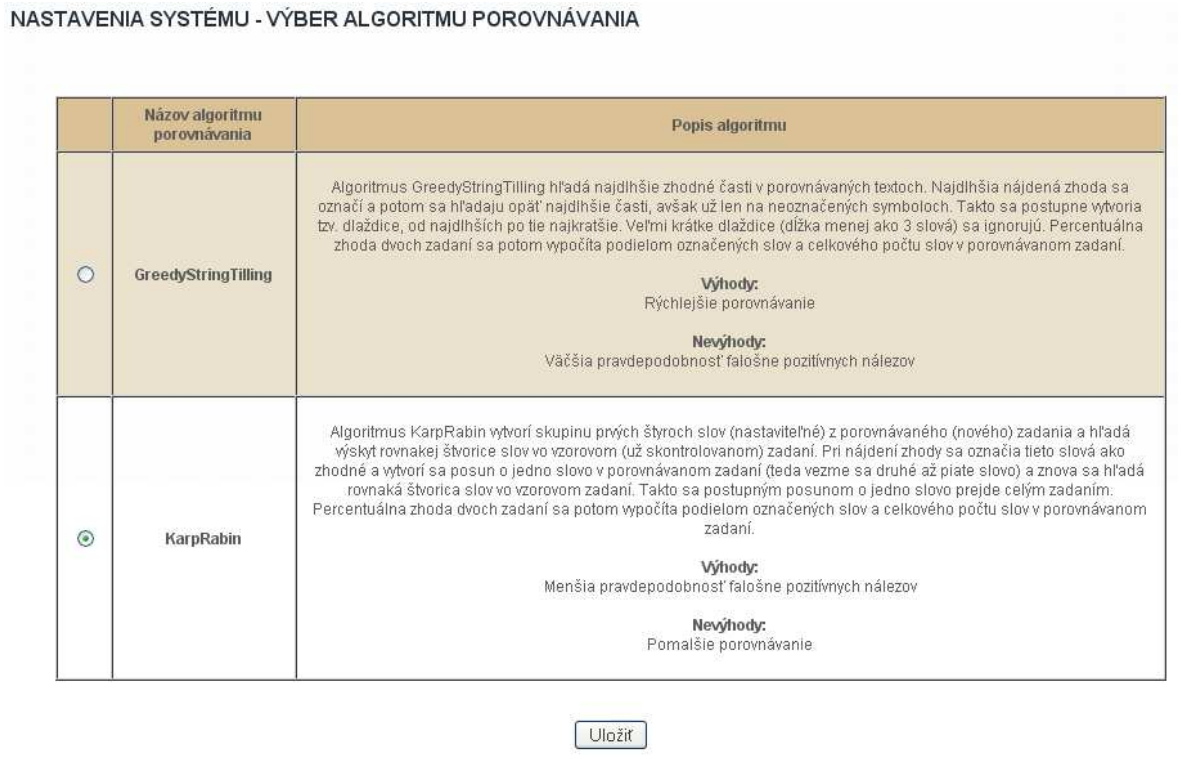
Databáza uchováva všetky potrebné informácie, aby systém fungoval správne. V databázových tabuľkách sú uložené informácie o predmetoch, prácach študentov a zhode medzi príslušnými prácami. Vzhľadom k tomu, že identifikácia a autentifikácia používateľov systému je založená na Centrálnom autentifikačnom systéme Technickej univerzity, v databáze sa nachádzajú iba jednoznačné identifikátory pedagógov a študentov. Ostatné potrebné informácie – napr. meno, priezvisko – sa získajú z autentifikačného systému.

MODUL KONVERZIE DOKUMENTOV

Modul konverzie dokumentov na Obr. 1 uvedený ako Prevodný modul realizuje konverziu rôznych dokumentových formátov do dvoch základných formátov – do textového tvaru a HTML formátu. Modul podporuje konverziu najrozšírenejších formátov - Portable Document Format (PDF), Microsoft Office Document 97-2003 (doc), Microsoft Office Document 2007 (docx), Open Document Format (odt) a LaTeX. Pri implementácii modulu boli použité aj knižnice tretej strany, kde každá z nich poskytuje špecifickú funkcionalitu. [12]

MODUL ODHAĽOVANIA PODOBNOSTI V TEXTOCH

Modul odhaľovania podobnosti v textoch na Obr. 1 uvedený ako Výpočtový modul realizuje porovnávanie študentských prác, ktoré sú konvertované z nejakého dokumentového formátu do formátu textového a vypočíta percentuálnu zhodu medzi dvojicami zadaní v systéme.



Obr. 2 Výber algoritmu porovnávania [11,12]

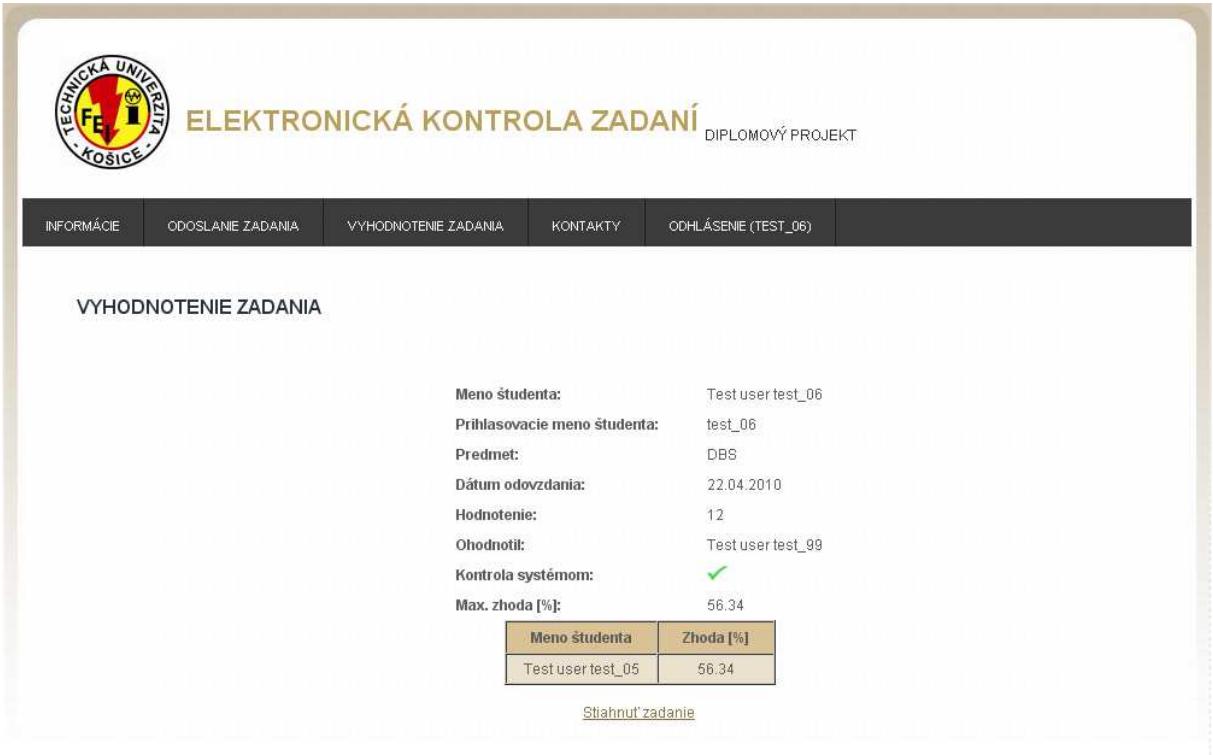
Existuje viacero algoritmov, ktoré je možné použiť na odhalenie podobnosti v textoch [9, 10]. V module sú implementované dva algoritmy [11] – Karp-Rabinov a Greedy String Tilling algoritmy. Na Obr. 2 je uvedený výber algoritmov, ktoré sú realizované v systéme.

Karp-Rabin Algoritmus
Hlavnou črtou tohto algoritmu je použitie hash funkcie [3, 4]. Hash funkcia transformuje vstupné dáta na iné, zvyčajne do numerickej formy nazývanej hash kód alebo odtlačok (fingerprint). Hash funkcia pre rovnaké vstupné hodnoty dáva tú istú výstupnú hodnotu. Algoritmus analyzuje text zľava doprava a pre nejaký reťazec vypočíta hash hodnotu. Pomocou hash funkcie sa vypočítajú odtlačky reťazcov celého textu. Ak sa dva odtlačky zhodujú, potom dva texty sú zhodné.

Greedy String Tilling algoritmus
Algoritmus hľadá najdlhší rovnaký reťazec v dvoch textoch (nazývaný „tile“) [6]. Najdlhšie zhodné reťazce sú označené a tieto označené slová už nie sú znovu zahrnuté do iného rovnakého reťazca. Algoritmus potom znovu hľadá najdlhší rovnaký reťazec v dvoch textoch, ale iba na neoznačených symboloch. Takto sú postupne vytvárané „dlaždice“ (tile) od najdlhšej po najkratšiu. Veľmi krátke reťazce-dlaždice sú ignorované. Percentuálna zhoda dvoch textov je vypočítaná ako pomer označených slov a celkového počtu slov v porovnávanom texte.

ZÁVER

Prototyp systému pre elektronickú kontrolu zadaní bol testovaný na vybranej vzorke zadaní predmetov, ktoré zabezpečuje Katedra počítačov a informatiky [Obr. 3].



Obr. 3 Vyhodnotenie zadania [11,12]

Systém je navrhnutý modulárne a je otvorený ďalším rozšíreniam – pridanie podpory pre konverziu ďalších dokumentových formátov, modul pre detekciu podobnosti v zdrojových kódoch.

POĎAKOVANIE
Tento príspevok je výsledkom projektu: Knowledge-Based Software Life Cycle and Architectures (Project VEGA No. 1/0350/08)

LITERATÚRA

1. What is plagiarism?, http://www.plagiarism.org/plag_article_what_is_plagiarism.html
2. Plagiarism.org, Learning center: Plagiarism Definitions, Tips on avoiding Plagiarism, Guidelines for proper citation, & Help Indentifying Plagiarism, http://www.plagiarism.org/plag_article_types_of_plagiarism.html
3. MUTIARA, A. B., AGUSTINA, S.: Anti plagiarism application with algorithm Karp-Rabin at thesis in Gunadarma University. In: Graduate Program in Information System. Gunadarma University, Depok, Indonesia
4. Karp-Rabin algorithm, http://www-igm.univ-mlv.fr/~lecroq/string/node5.html
5. MOZGOVOY, M.: Enhancing Computer-aided Plagiarism Detection. In: Department of Computer Science and Statistics. University of Joensuu, ftp://cs.joensuu.fi/pub/Dissertations/mozgovoy.pdf
6. WISE, M. J.: String Similarity via Greedy String Tiling and Running Karp−Rabin Matching. In: Department of Computer Science, University of Sydney, Australia. http://www.pam1.bcs.uwa.edu.au/~michaelw/ftp/doc/RKR_GST.ps
7. Genči, J., Maťašovská, M., Pločica, O., Telepovská, H.: Plagiátorstvo a spôsoby jeho obmedzenia, UNINFOS 2009, Zborník príspevkov z medzinárodnej konferencie, Nitra 2009
8. Havlice , Z.: Modelovanie a prototypovanie pri projektovaní informačných systémov, Elfa 1999
9. Professor Clifford Stein: Analysis of Algorithms, CSOR 4231. In: Department of Computer Science, Columbia University, Fall 2007. [citované 8.4.2009]. Dostupné na internete: <http://www.columbia.edu/~cs2035/courses/csor4231.F07/lcs.pdf>
10. EPPSTEIN, D.: Design and Analysis of Algorithms. In: Computer Science Department, Donald Bren School of Information and Computer Sciences, University of California, Irvine. [citované 8.4.2009]. Dostupné na internete: http://www.ics.uci.edu/~eppstein/161/960229.html
11. Gajdoš, F.: Plagiáty – vyhľadávanie podobnosti v textoch, diplomová práca, Technická univerzita v Košiciach, 2010
12. Szanyi, M.: Plagiáty – konverzia dokumentov, diplomová práca, Technická univerzita v Košiciach, 2010

Abstract.
The article deals with plagiarism problem and the ways of its detection in academic environment. This paper describes a prototype of the system for plagiarism detection at Faculty of Electrical Engineering and Informatics, Technical University of Košice. This system focuses to student projects of various subjects, especially on project documentation. The system consists of two basic modules – conversion module and computing module.
Keywords: plagiarism, detection of plagiarism, document conversion