Systém pro odhalování plagiátů na českých vysokých školách

Miroslav Křipač, Michal Brandejs, Jan Kasprzak, Jitka Brandejsová, FI MU

1 Vyhledávání plagiátů na MU

S rozvojem Internetu a počítačů obecně je stále jednodušší nejen možnost vyhledání informací, ale také jejich zneužití. Studenti, kteří připravují pomocí počítačů své domácí úkoly, seminární či závěrečné práce, mohou snadněji opisovat či přímo přebírat cizí texty a tím podvádět.

Na Masarykově univerzitě byla otázka systematického boje proti plagiátům ve studentských pracích poprvé ve větší míře diskutována před rokem 2004, kdy došlo k plošnému zveřejnění plných textů závěrečných prací v rámci Informačního systému Masarykovy univerzity (IS MU) všem studentům a učitelům univerzity. Volání řady členů akademické obce po programu, který by neoprávněné opisování z takto přístupných prací detekoval, tak začalo nabírat konkrétních podob.

Vzhledem k tomu, že IS MU jakožto zdroj plných textů prací je primárně studijní systém, rozhodli jsme se nejprve využít nabídky na vytvoření externího řešení, do kterého by IS MU své práce dodával a pouze interpretoval jeho výsledky. Ukázalo se však, že tento způsob implementace programu pro vyhledávání podobností není možné v prostředí IS MU použít, neboť je nutné začlenit celou řadu dalších mechanismů, které fakticky umožňují produkční provoz v prostředí rozsáhlé univerzity a její existující infrastruktury.

Proto došlo v roce 2006, kdy byl již v platnosti nový zákon o vysokých školách, který ukládá školám přímo povinnost všechny práce zveřejňovat celému světu, k implementaci vlastní verze softwaru na odhalování plagiátů přímo uvnitř Informačního systému Masarykovy univerzity. Tento program, který mohou využít nejen všichni učitelé, ale i studenti, absolventi či administrativní pracovníci, se setkal hned od počátku s velkým zájmem nejen členů akademické obce, ale také široké veřejnosti a médií.

Princip fungování tohoto systému je jednoduchý: po vložení do IS MU je každý dokument nejprve zpracován a porovnán s ostatními dříve uloženými dokumenty. Hledají se skutečné podobnosti, to znamená, že jako podobný se zobrazí i takový dokument, který autor okopíroval a částečně pozměnil (například použil jiné termíny, prohodil části textu apod.). Účinnost takového porovnání a ochrana před možností vyhledání podobnosti obejít vhodnou úpravou textu byla postupně zvyšována a nyní je na velmi vysoké úrovni. Je zřejmé, že opsanou myšlenku, kterou autor popíše svými vlastními slovy, žádný program neodhalí. To ovšem často není cílem, protože právě přidání vlastní invence do existujících znalostí může být cílem práce. Systém pro odhalování plagiátů tak má za cíl zejména zvýšit náročnost opsání dané práce oproti náročnosti na vypracování práce zcela nové. Princip je tedy jednoduchý – než by se student snažil odhalení plagiátu zabránit, raději práci napíše sám, neboť je to pro něj efektivnější.

Samotný software zobrazí po zvolení funkce na nalezení podobných dokumentů (ta je v systému symbolizována ikonou dvou vajíček, což představují úsloví „podobné jako vejce vejci“) seznam souborů, se kterými jsou vyhledávané dokumenty podobné. Takto nalezené soubory se pak zobrazují včetně čísla, které udává přibližné procento nalezení podobností a odkaz na zobrazení částí textu, které jsou podobné. Samotné rozhodnutí o tom, zda se jedná o plagiát, ale stále zůstává na učiteli nebo odborníkovi na dané téma.

V rámci IS MU je pak aplikace dostupná nejen pro závěrečné práce, ale pro jakékoliv soubory, tedy včetně domácích úloh, esejí a seminárních prací. Potvrdilo se také, že jen samotná skutečnost, že studenti vědí o probíhající automatické kontrole, odradí celou řadu z nich od pokusů o plagiátorství. Preventivně také působí výsledky disciplinárních řízení, kdy děkan může udělit (a v minulosti již udělil) i trest vyloučení ze studia za zneužití cizí práce v práci vlastní.

Postupem času byla do systému doplněna také funkce pro zobrazení všech dokumentů, u kterých je podezření na plagiátorství, které spadají pod danou fakultu. Správce nebo člen vedení fakulty tak nemusí postupně procházet a kontro-

lovat všechny nově vložené dokumenty, ale aplikace sama upozorní na ty, které jsou podobné.

V květnu 2008 systém vyhledával podobnosti již ve více než milionu dokumentů, což je asi čtyřnásobek oproti počtu dokumentů, které byly srovnávány při prvním spuštění.

2 Rozšíření na ostatní školy

Takřka ihned po zveřejnění tohoto programu v rámci IS MU se ozvalo několik škol, které by měly zájem systém také využít. Vhledem k tak velké integraci s IS MU však bylo možné zpočátku těmto prosbám vyhovět jen u těch škol, které používají stejný informační systém.

Během loňského roku byl však sestaven nový projekt, který měl za cíl sjednotit snahy jednotlivých českých (a později i některých slovenských) vysokých škol do jednoho projektu, který by umožnil vyhledávat plagiáty centrálně, a to i u těch dokumentů, které si daná škola nepřeje úplně zveřejnit (například starší diplomové práce apod.).

V rámci tohoto projektu byl vytvořen nový systém, který je dostupný na adrese `http://theses.cz/` a do kterého již nyní (v květnu 2008) mohou první školy vkládat své závěrečné práce. Cílem tohoto projektu však není pouze vyhledávání podobností, ale, tak jak je tomu na MU, umožnit veřejnosti nahlédnout do plného textu práce (fakticky tak škola může dostát své zákonné povinnosti na zveřejňování prací díky tomuto softwaru bez dalších nákladů na udržování vlastní veřejně přístupné databáze). Jako další funkci lze zmínit také funkčnost tzv. *Národního registru vysokoškolských kvalifikačních prací*, což je knihovnický registr, který je určen pro evidencí údajů o diplomových a bakalářských pracích.

Pro tak rozsáhlý projekt a zejména pro velký nárůst dokumentů i v rámci samotné MU muselo dojít k přepracování klíčových programů pro vyhledávání podobností. Nově je tak v systému theses.cz přístupná verze, která využívá plně distribuované uložení a zpracování dat uvnitř IS MU (vyhledávání probíhá v současné době na clusteru několika desítek počítačů). Tím došlo k výraznému zvýšení propustnosti systému při zachování velmi vysoké přesnosti. Zároveň byla přidána podpora pro všechny jazyky, které jsou v dokumentech MU běžně používány.

Přestože ne všechny školy jsou se sběrem elektronických verzí dokumentů tak daleko jako MU, ukazuje se, že další rozvoj této spolupráce může přinést užitečné výsledky. V budoucnu tak bude možné použít tento systém i pro seminární práce a domácí úkoly, publikace nebo další soubory, u kterých je potřebné zkoumat původ. □