TMT 54. évf. 2007. 3. sz.

Pataki Máté

Digitális könyvtárak védelme a KOPI plágiumkereső rendszerrel

Az egyetemi és a digitális könyvtári világban a dokumentumok védelme fontos kérdés, ugyanakkor pusztán a másolásvédelmi eljárások nem igazán alkalmasak ennek a feladatnak az ellátására. A legtöbb védelem könnyen megkerülhető, mások jobban védenek, de bonyolult a használatuk, adott platformhoz kötöttek, így erősen leszűkítik a felhasználók körét. A plagizálás elleni védelemben segít a KOPI plágiumkereső rendszer, amely gyorsan megtalálja a másolt dokumentumokat, megjelöli az eredeti forrásokat és a szerzőket. Ezáltal kockázatossá válik a másolás a plágiumkereső védelme alatt álló dokumentumokból. Ha széles körben elterjed a plágiumkereső használata, a védett dokumentumokat szabadon lehet majd terjeszteni, és nem kell attól tartani, hogy valaki saját neve alatt fogja őket publikálni.

Bevezetés

A digitális tartalmak védelmét szolgáló megoldásokat alapvetően két csoportba lehet osztani. Az egyikbe azok tartoznak, amelyek valamilyen módon megakadályozzák az illegális másolást, felhasználást, a másikba azok, amelyek felfedik a másolás tényét. Nehéz megóvni a digitális tartalmat az illegális másolástól úgy, hogy közben a legális felhasználást ne nehezítse meg a rendszer, sőt egyes esetekben még azt is nehéz megoldani, hogy mindenki hozzáférhessen a tartalomhoz, függetlenül a használt szoftverkörnyezettől. A legtöbb másolásvédelmi rendszer könnyen feltörhető, így csak névleges védelmet ad. Vannak jobban védő rendszerek, amelyek megkerülése körülményes, és csak kiegészítő szoftverekkel együtt használhatók; telepítésük csak akkor kifizetődő, ha a felhasználónak igazán értékes a tartalom, amelyet véd. A hátrányos helyzetűek – akik speciális eszközökkel használják az internetet – gyakran nem is képesek elérni ezeket a védett tartalmakat.

A plágiumkeresés nem védi meg a tartalmat az illegális másolástól, de ha széles körben használják, követhetővé teszi a mű útját, és megakadályozhatja, hogy valaki a sajátjaként tüntesse azt fel. Ez a védelem kettős: egyrészt másolatot találva a rendszer azonnal megnevezi a forrást és az átfedés mértékét, másrészt, ha az ilyen rendszer létezése széles körben ismert, és használata elterjedt, akkor a legtöbben nem fogják megkockáztatni, hogy plagizáljanak, kitéve magukat a lebukás veszélyének.

Plágium és plagizálás

Definíció

A plágiumot a *Magyar Értelmező Szótár* (*MESZ*) így határozza meg:

> „plágium: szellemi tolvajlás, más művének közlése saját név alatt, a mű alapgondolatának vagy részleteinek felhasználása a szerzőre való hivatkozás nélkül. Perbe fogták plágiumért. Bebizonyosodott, hogy novellája az első betűtől az utolsóig plágium” [1].

Két fontos mondanivaló van a fenti idézetben: az egyik, hogy a szerzőre való hivatkozás elmulasztása miatt válik az idézet plágiummá, a másik, hogy elég egy részletet átvenni – azaz nem kell valaki másnak a teljes művét lemásolni és sajátként prezentálni –, egy rövid idézetnél is meg kell jelölni az eredeti szerzőt. Ezt akkor is meg kell tenni, ha a szerző nem tart rá igényt, esetleg lemondott a műről, nincsenek már hozzá fűződő jogai, vagy ismeretlen. Egy diplomamunkában, vagy házi feladatban nem az a fontos, hogy az elkészült munka ne sértse meg más szerzői jogait, hanem az, hogy készítőjének saját, önálló alkotása legyen. Teljesen lényegtelen, hogy kiről másolt, egyértelműen meg kell jelölnie, hogy mely részeket honnan és milyen forrásból vett át.

Plágium a felsőoktatásban

A plágium talán a felsőoktatásban okozza a legnagyobb gondot, ahol a legtöbb feladat, dolgozat és

diplomamunka digitálisan készül, és az ismerősökön, közösen használt gépeken, szervereken, honlapokon keresztül terjed a diákok között. Már a középiskolákban is ismertek az előre elkészített házi feladatok, olvasónaplók, érettségi tételek, sőt külön honlapok készülnek ezek megosztására, de itt sokkal nehezebb a diákok dolga, mivel a tanár (jobb esetben) ismeri őket, a korábbi teljesítményüket és stílusukat, így egy bárhonnan lemásolt dolgozatnál nagy a lebukás veszélye. Ezzel szemben a felsőoktatásban több ezer diák is felveheti ugyanazt a tárgyat, a beadott munkák javítását minden évben változó, nagy létszámú csoport végzi, ezért a lebukás veszélye is elenyésző.

Ha elképzeljük, hogy adott szakterületen és évben hány diploma születik az országban, akkor láthatjuk, hogy nincs az a professzor, aki ezeket mind ismerhetné, és észrevehetné, hogy másolás történt. Anélkül, hogy valakit is megsértenénk, kijelenthetjük, hogy a diplomamunkák jelentős része szakmai szempontból sajnos teljesen érdektelen, és erről nem feltétlenül a diák tehet. Mivel az egyetemek és a főiskolák tartanak a plágiumtól, nem teszik mindenkinek elérhetővé a korábbi években született dolgozatokat, így ezek évről évre ugyanazon témákban születnek anélkül, hogy egymás eredményeire építenének, azaz újból és újból „feltalálják a spanyolviaszt". Nem valószínű, hogy egy tanszéken belül ez így lenne, de egy egyetemen belül már biztosan előfordul, nem beszélve az ország különböző egyetemeinek és főiskoláinak tanszékeiről, ahol számos, egymást témájában majdnem teljesen átfedő diplomamunkát nyújtanak be.

Magyarországon a legnagyobb gondot valószínűleg az egymásról történő másolás okozza, de az angol és német nyelvterületeken – ahol nagyságrendekkel több tartalom található meg az interneten – a legfőbb gondot az internetes oldalakról, például a *Wikipédiá*ból másolt szövegek okozzák, és a trendek alapján hazánk is erre halad.

Plágium a tudomány világában

A plagizálás sajnos a tudományos területeken sem olyan ritka dolog, mint azt hinni szeretnénk. A jelenség valószínűleg az egyetemi diplomamunkáknál kezdődik, majd folytatódik a tudományos publikációknál, és az is előfordult, hogy valaki a doktori disszertációjában plagizált, ami már felettébb kellemetlen, nemcsak az illetőnek, hanem elsősorban annak az oktatási intézménynek, amelyben a doktori címét szerezte. Minden ilyen napvilágra került ügy után megkérdőjeleződik annak az intézménynek a színvonala, amelyben átengedték a plágiumot, és diplomával jutalmazták a plagizálót, holott az intézményeknek kevés eszközük van ennek megakadályozására. A diplomát értékelő szakembertől elvárható, hogy az összes fontosabb művet és szereplőt ismerje az adott szakterületen, de az nem, hogy minden egyes diplomamunkát és házi feladatot elolvasva rájuk is ismerjen, mivel fizikailag sem fér hozzá az eredetik jelentős részéhez.

A tudományos publikációknál a másolásnak egy másik formája is ismert, ez az önmagáról való másolás. Mivel sokakat érint a publikálási kényszer, vagy azért, hogy megkapják a tudományos fokozatukat, vagy mert olyan intézményben dolgoznak, ahol ennek alapján (is) mérik a teljesítményt, saját korábbi publikációikat próbálják meg minél többször megjelentetni, természetesen mindig egy pici változtatással. Ez utóbbi a kiadóknak okozhat gondot, mivel arra törekednek, hogy minél több tudományos újdonságot jelentessenek meg, és ha ezt nem tudják teljesíteni, illetve ha rendszeresen olyan tudományos értekezéseket jelentetnek meg, amelyek már máshol megjelentek, akkor nem lesz olyan értékes az adott kiadvány, kevesebben fogják olvasni, idézni, és ezért kevesebben is kívánnak majd ott publikálni. Mindez gondot okoz a tudományos közösségnek is, mivel a cikkek száma a sokszorosa lesz a tényleges tudásmennyiségnek, túlterhelik a szakma képviselőit, akik nehezebben jutnak hozzá az új információkhoz.

Plágium a digitális könyvtáraknál

A digitális könyváraknál a plagizálás kétféleképpen is történhet. A legegyszerűbb, hogy valaki talál valamilyen szép gondolatot az egyik műben, és azt beépíti a sajátjába, anélkül, hogy megnevezte volna az eredeti szerzőt és a forrást. Ez végül is megegyezik az előzőleg tárgyaltakkal, csak az ellenkező oldalról tekintünk rá. A másik az – és valószínűleg ez a legkárosabb a digitális könyvtárakra –, hogy mások átveszik a teljes művet, és saját gyűjteményükben helyezik el. Ennek különböző módjai léteznek, és megítélésük is attól függ, hogy például az eredeti művet milyen forrásból digitalizálták, milyen szerzői jogok vonatkoznak rá, amikor eladták, vagy milyen feltételeket szabtak annak, hogy letöltsék. A digitális könyvtárnak mindenesetre ez forgalom-, illetve bevételkiesést jelent, és még az ingyenesen hozzáférhető gyűjteménynél is rossz lehet, hogy nem ismernek pontos statisztikákat arról, hogy melyik műre hányan kíváncsiak, és mely műveket kellene még digitalizál-

niuk, mert nem tőlük töltik le az érdeklődők a tartalmat, hanem harmadik szolgáltató oldaláról. Sok digitális könyvtár az oldalán elhelyezett reklámokból is bevételhez jut; ilyenkor is komoly hátrány éri őket, ha más kereskedik a művükkel, függetlenül attól, hogy az illető ezt pénzért teszi, vagy ingyen bocsátja mások rendelkezésére.

Plágiumkereső rendszerek

A plágiumkereső rendszereknek sok fajtája létezik, és legtöbbjük jól használható bizonyos területeken. Jelentős részükre azonban olyan megkötések vonatkoznak, amelyek miatt például digitális könyvtáraknál vagy egyetemi diplomamunkagyűjteménynél nem használhatók. Ebben a fejezetben rövid ismertetés található a fontosabb típusokról, előnyeikről és hátrányaikról.

Vízjel és ellenőrző összeg

Sok rendszer használ vízjelet vagy valamilyen ellenőrző összeget a művek eredetiségének vagy származásának a megállapítására. Az ellenőrző összegek jól használhatók annak az ellenőrzésére, hogy a művet, vagy annak részeit megváltoztatták-e, illetve a mű „útját" követik nyomon a segítségével. A vízjel képeknél és videóknál a legelterjedtebb, de szöveges dokumentumoknál is gyakran használják. Utóbbinál legtöbbször a szóközök méretének szemmel észrevehetetlen megváltoztatásával érik el a hatást, és így adott körülmények között még egy fénymásolatról is megállapítható, hogy honnan vették át. Mindkét megoldásnál az jelenti a legnagyobb gondot, hogy már egy kisebb változtatás is könnyen a védelem elvesztésével jár, és ha valaki tud arról, hogy a dokumentum ilyen védelem alatt áll, akkor könnyedén és automatizálva eltávolíthatja azt. További hátrány, hogy kisebb idézetek, részletek átvételénél egyik megoldás sem használható.

A szerző azonosítása

A szerző azonosítása (authorship attribution) erősen kutatott számítógépes nyelvészeti terület. Ennél a megoldásnál a szöveg nyelvi, nyelvtani elemzésével, a használt szavak alapján próbálják megállapítani, hogy egy művet ki írt, vagy két művet ugyanaz a személy írta-e. Irodalmi elemzésekben is használtak már ehhez hasonló eszközöket egy író különböző korban írt műveinek az elemzésére, vagy adott műben a stílusok változásának, a nyomon követésére [2]. A megoldásnak vannak hátrányai; az algoritmusok például – mivel legtöbb esetben nyelvtani elemzést használnak – nyelvfüggők, ezért minden nyelvre külön kell őket kifejleszteni. Ahhoz, hogy a rendszer meg tudja állapítani, hogy ki a szerző, rendelkeznie kell már megfelelő mintákkal a szerzőtől, ez ritkán oldható meg. A módszer jelenleg még nem elég megbízható [3] ahhoz, hogy több ezer szerző dokumentumai között megfelelő biztonsággal különbséget tegyen, ugyanakkor egy művön belül ki lehet mutatni vele a stílusváltozásokat. Érdemes lehet esetleg ezekre a változásokra, vagyis az ezt okozó pár mondatra mint kulcsmondatokra rákeresni egy keresőben, hogy máshol nincsenek-e meg.

Nyílt keresőszolgáltatások

Léteznek olyan plágiumkereső rendszerek, amelyek nyílt keresőrendszerekre – mint amilyen a Google – épülnek. Ilyen rendszer volt a *PSearch* [4]. A *Copyscape* [5] rendszerrel honlapok tartalmát lehet megvédeni a plagizálástól, azaz egy honlapot megadva, ahhoz hasonlókat, vagy azzal egyezőeket keres az interneten. Belső működésére nem térnek ki részletesen az oldalon, de annyi azért kiderül, hogy metakeresőről van szó, amely a Google-ra épül. Hasonló elven működik a *PCheck* [6] is, amely a feltöltött szöveges dokumentumból mondatot emel ki véletlenszerűen, és azt felhasználva keres a Google segítségével. Ezek a megoldások hasznosak lehetnek interneten megtalálható tartalmak megkereséséhez, de sajnos az igazán jól használhatónak tűnő megoldások fizetősek. Az ingyenesen elérhetők, mint az utóbb említett is, erőforrás híján nem végeznek teljes keresést, így ha nem talál egyezést, az még nem bizonyítja azt, hogy a mű teljesen eredeti. Ezt a programot ugyanakkor kombinálni lehetne az előző fejezet végén említett megoldással, és akkor nem véletlen mondatokra keresne, hanem a valami miatt oda nem illőkre, vagy más stílusban írottakra, ami feltehetően valamivel növelné a megbízhatóságát.

Az internetről plagizált művek megtalálásában valószínűleg az ilyen, nyílt keresőrendszerre épülő, online szolgáltatás bizonyulhat a leghatékonyabbnak, viszont az interneten közvetlenül meg nem található tartalmakban ezek a rendszerek nem képesek keresni. A diplomamunkájukat kevesen teszik fel az internetre, a könyv- és újságkiadók ritkán teszik elérhetővé a teljes tartalmakat a honlapjukon, sőt némely digitális könyvtár is csak regisztráció után érhető el, azaz automata kereső már nem találja meg az ott lévő tartalmakat.

Szöveges összehasonlítás

Két dokumentum egymással való összehasonlítása a hasonlóságkeresés legegyszerűbb módja. A legismertebb szövegszerkesztő, a *Microsoft Word* is tartalmazza ezt a funkciót, és a *TotalCommander* nevű, széles körben használt fájlkezelő program is használható két szöveges formátumú dokumentum összehasonlítására. Kevés dokumentum esetén ez az eljárás a leghatékonyabb, és ez adja a legpontosabb eredményt, ugyanakkor nagyobb dokumentumhalmaz elemeinek egymással való összehasonlítása nem oldható meg hatékonyan ezzel a módszerrel. Már tíz dokumentumnál is 45 összehasonlítási műveletet kell elvégezni, ha párosával szeretnénk össze hasonlítani a műveket. Több ezer dokumentumnál ez a módszer már egyáltalán nem használható. Ugyanakkor, ha a felhasználó egy másik, akár sokkal pontatlanabb módszerrel ki tudja szűrni nagy adatbázisából azt a húsz-harminc dokumentumot, amely egyáltalán szóba jön, második lépésben érdemes egy ilyen összehasonlító és vizualizáló programot használnia a hasonlóság mértékének pontosabb megállapítása, és az eredmények megmutatása céljából.

Kérdőív

Az előbbitől eltérő megoldást használ a *Glatt Plagiarism Screening Program* (*GPSP*) [7], amely afféle kérdőívet állít elő a műből olyan módon, hogy bizonyos szavakat kitöröl, és utána a szerzőnek ki kell töltenie a hiányzó részeket. A program készítői azzal a jogos feltételezéssel éltek, hogy az eredeti szerző valószínűleg a legtöbb helyen ugyanazokat a szavakat használná másodszor is, míg mások nagyobb százalékban illesztenének be eltérő, rokon értelmű szavakat a hiányzók helyére. Ennek a megoldásnak az a hátránya, hogy a teszt elvégeztetésével már meggyanúsítottuk a diákot plagizálással, ráadásul ez a módszer sok időt igényel mind a tanártól, mind a diákoktól. Egyetemi környezetben, ha kevés a diák, esetleg használható ez a módszer, de például egy digitális könyvtárban található dokumentumról történő másolást nem fedez fel, ha azt nem diák követi el, hanem például tudományos cikk szerzője.

Ismeretlen működésű keresők

Sok olyan rendszer található az interneten, amelynek belső működése teljesen ismeretlen, legtöbbször még olyan alapvető információkra sem derül fény, hogy milyen nyelvű dokumentumokhoz használható, illetve hogy milyen algoritmust használ, és mennyire megbízható. Mind a *Plagiarism Finder* (*PFind*) [8], mind az *EVE Plagiarism Detection System* [9] fizetős rendszerek, de a honlapjukon alig van információ arról, hogy hogyan működnek. Utóbbi például valószínűleg a korábban már említett internetes keresők egy változata saját adatbázissal. Sajnos ezeknél a rendszereknél nem lehet tudni, hogy milyen mértékű másolást találnak meg, vagy hogy mennyire lehet megbízni a készítőiben. Míg ez utóbb említett rendszer már régóta üzemel, és több mint valószínű, hogy megbízható, pár éve egy orosz plágiumkereső szolgáltatásról kiderült, hogy a plágiumkeresésre beérkezett dokumentumokat egy másik honlapon éppen plagizálás céljával árusítani kezdték. Egyetem, vagy nagyobb intézmény ezért valószínűleg nem engedheti meg magának, hogy a nála készült diplomamunkákat és egyéb dokumentumokat tömegesen kétes megbízhatóságú oldalra töltse fel.

A KOPI portál

A KOPI portált a volt *Informatikai és Hírközlési Minisztérium* támogatásával az *MTA SZTAKI Elosztott rendszerek osztálya* (*DSD*) [11] a melbourne-i *Monash Egyetem*mel együtt, annak eredményeit felhasználva fejlesztette ki. A Portál 2004-ben készült el, és azóta is szabadon hozzáférhető az érdeklődők számára.

A KOPI projekt célja elsősorban a tanárok, professzorok, konferenciaszervezők segítése a másolt művek eredetijének a felkutatásában, a digitális könyvtárak védelme az illegális másolatoktól, a diákok tájékoztatása a plagizálásról és az idézés helyes módjáról, valamint a cikkek, dolgozatok, diplomamunkák értékének a növelése az eredetiségük igazolásával.

Érdemes kiemelni – és ez az összes korábban említett szolgáltatásra is igaz –, hogy ezek a rendszerek nem tudják megállapítani, hogy valami idézet-e vagy plágium; az ilyen rendszer csak arra képes, hogy jelezze a felhasználónak, hogy az adott dokumentumban mely más dokumentumból talált meg részeket, mekkora az átfedés vagy a hasonlóság. Annak a megállapítása, hogy ez szabályos módon történt idézés-e, és helyesen meg van-e jelölve a forrás, már a felhasználóra van bízva.

Mielőtt kitérnénk arra, hogy a *KOPI Online Plágiumkereső és Információs Portál* (KOPI) [10] által is használt algoritmuson alapuló plágiumkereső szolgáltatás miként is védi meg a dokumentumokat a plagizálás ellen, és miként oldja meg az előző fejezetben felvetett problémákat, nézzük meg, hogy milyen szolgáltatásokat is nyújt.

Portálszolgáltatások

A KOPI portál legfőbb célja a plágiumok, illetve a plagizálás visszaszorítása, ezért az oldalon több szolgáltatás is található ennek elősegítésére. A legfontosabb ezek közül az az információgyűjtemény, amely a plágiummal kapcsolatos tudnivalókat gyűjti össze.

Információk
Mivel sokan nem is tudják pontosan, hogy mi a plágium, és nem ismerik az idézés pontos szabályait, a KOPI portálon a plágium definícióján kívül részletes leírás is található arról, hogy mi a plágium, és milyen fokozatai vannak, valamint egy útmutató a helyes idézés módjáról. A vonatkozó jogszabályok mellett az egyetemi szabályzatok is helyet kaptak az ugrópontgyűjteményben. Nemcsak azért, hogy lássák a hallgatók, milyen következménnyel jár a plagizálás, hanem azért is, mert az idézésnek is pontos szabályai vannak. Egy diplomamunkában például nem lehet meghatározott mennyiségnél több idézet, hiába jelöljük meg a szerzőt, hiszen valami újat, valami sajátot is hozzá kell tenni az eddigiekhez, hogy elfogadják. Hasonló módon, ha két diák közös témában ír diplomamunkát, akkor is csak megadott fejezetek lehetnek közösek, mondjuk a munka teljes terjedelmének 30%-a, a többinek teljesen egyéninek kell lennie.

Fórum
A fórumszolgáltatás is hozzájárulhat a plagizálás visszaszorításához, ha erről a problémakörről nyílt beszélgetések alakulnak ki az érintett felek között. A fórumszolgáltatás eléréséhez regisztrálni kell, de a felhasználók személyes adatai nem láthatók, így névtelenül beszélhetik meg például a diákok és az oktatók a problémáikat, és írhatják le tapasztalataikat, javaslataikat.

Egyéb szolgáltatások
A rendszer lehetővé teszi, hogy ha valaki plágiumot vagy hasonlóságot talált, felvegye a kapcsolatot azzal, aki a másik művet feltöltötte, így meg lehet beszélni, hogy melyik az eredeti mű, ki kiről másolt. A portálnak magyar és angol felhasználói felülete van, ez is hozzájárul ahhoz, hogy minél többen használják, és minél gyorsabban bővüljön az adatbázisa.

Hasonlóságkereső szolgáltatások

A KOPI portál lényegét természetesen a plágiumkereső szolgáltatások adják. Érdemes megadni a portálban feltöltött művek címét és szerzőjét, hogy később az adott felhasználó és a többiek által is azonosíthatók legyenek a dokumentumok. A rendszer egyéb, részletesebb információk megadását is lehetővé teszi, mint például: kiadó, kiadás éve, kulcsszavak, személyes megjegyzés. Jelenleg az alábbi dokumentumformátumokat támogatja: *doc*, *rtf*, *pdf*, *html*, *txt*, és ezekből álló tömörített *zip* állományokat, több dokumentum gyors feltöltése érdekében. A feltöltött dokumentumokkal ezek után plágiumkereséseket lehet indítani.

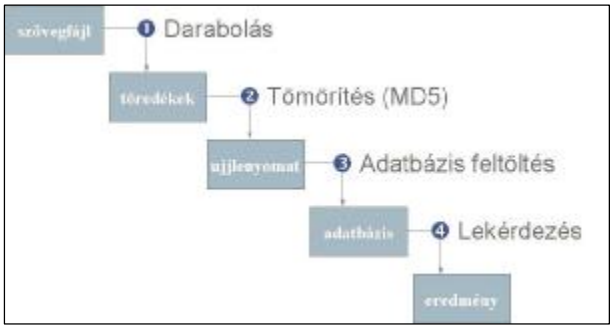
Adott dokumentumokhoz hasonlók keresése a rendszer adatbázisában
A legegyszerűbb keresés, amikor a felhasználó egy vagy több dokumentumot választ ki, és a rendszerben lévő összes többivel – köztük a saját maga által feltöltöttekkel is – összehasonlítja. Ennek a keresésnek az eredménye két helyen is elérhető lesz, és választástól függően e-mailes értesítést is küld róla a rendszer. Az e-mailben, és a keresés eredményét tartalmazó belső üzenetben rövid összefoglaló található a keresés eredményéről. Ebben a rendszer megjelöli azokat a dokumentumokat, amelyekhez hasonlót talált, valamint a hasonlóság mértékét %-ban, a másik dokumentum címét, szerzőjét és feltöltőjének a nevét. A dokumentumok listájában kis színes csík is jelöli, hogy milyen mértékben egyezik az adott dokumentum más, a rendszerben talált dokumentumokkal (*1. ábra*).

Keresés internetes és egyéb adatbázisokban
A rendszer támogatja teljesen különálló dokumentumhalmazok, adatbázisok bekapcsolását is a plágiumkeresésbe, és akkor ezek is megjelennek a rendszer jelenlegi adatbázisa mellett, mint kiválasztható lehetőségek, hogy azokban keressen a felhasználó dokumentumaihoz hasonlót. Erőforráshiány miatt még nem állították fel az internetről letöltött dokumentumokat tartalmazó adatbázist, amelyben szintén tud keresni a rendszer, és a távlati tervek között szerepel digitális könyvtárak (pl. a MEK) adatállományának a feldolgozása, valamint egyetemek diplomamunkáinak a begyűjtése és kereshetővé tétele.

1. ábra **A keresés eredményét tartalmazó üzenet**

Dokumentumok összehasonlítása egymással

A felhasználónak lehetősége van arra, hogy több kiválasztott dokumentumot összehasonlítson egymással. Ez akkor lehet kényelmes, amikor adott házi feladatra beérkezett műveket kell egyediség szempontjából ellenőrizni, vagy – hogy ne csak plagizálással kapcsolatos példákat említsünk – a diplomamunkához használt irodalomkészletet is feltölthetjük, és a rendszer megállapítja az idézetek mennyiségét.



2. ábra **A KOPI összehasonlítási folyamata**

A rendszer működése

A plágiumkereső rendszereknek az a csoportja, amelybe a KOPI is tartozik, adatbázist alakít ki a dokumentumokból, és aztán ebben az adatbázisban keres hasonló dokumentumokat. Ezen belül is vannak olyan rendszerek, amelyek például gráfot építenek a dokumentumokból, ahol a gráf élei a szavak közötti kapcsolatok (a szavak egymásutánisága), majd ezek között az élek között keresnek minél hosszabb egyezéseket. A KOPI ezzel szemben kisebb részekre darabolja a dokumentumot, azután ezeket a kisebb darabokat tömöríti, a tömörített darabokat adatbázisba tölti, majd ebben az adatbázisban keres azonos darabokat a különböző dokumentumok között. A teljes folyamatot a *2. ábra* szemlélteti.

Darabolás

A darabolás az eljárás lelke, ezen múlik, hogy mekkora és milyen egyezéseket lesz képes kimutatni a rendszer [12]. A darabolás történhet például mondathatárnál, adott gyakori szavaknál, vagy *n* szavanként. A KOPI ez utóbbit használja, mert ez bizonyult a legmegbízhatóbbnak; túlnyomórészt megtalálja az egyezéseket, és kevés hamis, nem jelentős egyezést ad. A könnyebb érthetőség kedvéért álljon itt egy példa ötszavas darabolásra.

Az eredeti szöveg:

```
Ezen project célja, hogy a Monash University-
vel együttműködve egy olyan rendszert hozzunk
létre, amely hatékony a dokumentum-másolatok
felderítésében.
```

Szavas daraboláskor ötös paraméterrel az alábbi négy töredéket kapjuk:

```
ezen project célja hogy a
monash university vel együttműködve egy
olyan rendszert hozzunk létre amely
hatékony a dokumentum másolatok felderí-
tésében
```

Ezek után olyan dokumentumokat keresünk majd az adatbázisban, amelyekben e töredékek közül valamelyik megtalálható. Az eljárással csak az a gond, hogy érzékeny a szavak beszúrására, illetve törlésére. Egy „az” szó beszúrásával a következő töredékeket kapjuk:

```
ezen project célja az hogy
a monash university vel együttműködve
egy olyan rendszert hozzunk létre
amely hatékony a dokumentum másolatok
```

Mint látható, a két dokumentumban már nincsenek azonos töredékek. Ezen segít az átlapolódó szavas darabolás, amely minden szónál elkezd egy töredéket, így teszi lehetővé, hogy elcsúszáskor is megtalálja a hasonlóságot.

Átlapolódó szavas darabolással az eredeti szövegből ezt kapjuk:

```
1. ezen project célja hogy a
2. project célja hogy a monash
3. célja hogy a monash university
4. hogy a monash university vel
5. a monash university vel együttműködve
6. monash university vel együttműködve egy
7. university vel együttműködve egy olyan
...
```

Az első példának két töredéke is megtalálható ezek között (ha végigírtuk volna, mind a négy megtalálható lenne: 1., 6., 11. és 16.), de a második dokumentummal is csak egyel kevesebb közös töredéke van (az 5., és később a 10. és 15.). A KOPI rendszer az adatbázisában lévő dokumentumokat szavas darabolással tagolja, míg azokat, amelyeket össze szeretnénk hasonlítani vele, átlapolódó szavas darabolással, így teszi lehetővé, hogy minden átírás, beszúrás, törlés maximum egy hibát okoz a keresésben, azaz ahhoz, hogy például egy hetes paramétert használó rendszert megtévesszen valamelyik felhasználó, legalább minden 7. szót át kell írnia. Ez már néhány oldalas dokumentum átírása esetén is nagy teljesítmény lenne, és még az is előfordulhat, hogy a rendszer kisebb paramétert használ, és az átírás ellenére lebukik a másoló.

Tömörítés

A tömörítés legfontosabb haszna az adatbázis méretében jelentkezik. Mivel minden szöveges darabot a veszteséges hash-kódolást használó MD5 kriptográfiai algoritmussal számmá alakít a rendszer (*3. ábra*), az adatbázis sokkal kisebb lesz, és a keresés is gyorsabb, hiszen számokkal sokkal gyorsabban tud dolgozni a számítógép, mint szövegekkel.

3. ábra **Szöveg tömörítése számmá**

MD5-ös algoritmusnál a bemenet bármilyen hosszú lehet, a kimenet mindig egy adott hosszúságú szám. Ez az algoritmus igen gyors, és annak az esélye, hogy két különböző bemeneti szöveghez ugyanazt a számot adja kimenetként, kicsi. Az irreverzibilis, veszteséges kódolásnak az az előnye, hogy nem állítható vissza a számból az eredeti szöveg, a dokumentumból ezzel az eljárással generált számok alapján nem állítható vissza emberi időn belül az eredeti dokumentum. Ennek ellenére lehetséges az azonos módon készített adatbázisban egyező számokat keresni, és így hasonló dokumentumokat találni. Ez az eljárás ezért kényes, értékes tartalmak védelmére is alkalmas. A korábban említett esetben például, amikor egy internetes plágiumkereső szolgáltatás elkezdte árulni a feltöltött műveket, jó védelem lehetett volna: csak a számokat tölti fel a felhasználó a saját rendszerébe, és annak ellenére, hogy nem jutottak hozzá a dokumentum tartalmához, tudnak keresni hasonló dokumentumokat.

Másolásvédelem

Most, hogy megismertük a plágiumkereső rendszerek működését, nézzük meg, miként viszonyulnak ezek a megoldások a másolásvédelemhez, és miként képesek megvédeni a digitális tartalmakat a plagizálástól, továbbá mikor érdemes ezeket használni. Semmiképpen sem állíthatjuk, hogy a másolásvédelem rossz lenne, és a plágiumkeresés feleslegessé tenné a használatát, sőt esetenként együtt hozhatják létre a leghatékonyabb védelmet. Célunk csak az, hogy alternatívát mutassunk, amely bizonyos esetekben jobb megoldás lehet, mint a másolásvédelem.

A másolásvédelmi eljárások előnyei

Először nézzük meg, milyen előnyökkel rendelkeznek a másolásvédelmi rendszerek. Mint az a nevében is benne van, megvédi a tartalmakat a má-

solástól. Nem állíthatjuk, hogy 100%-os védelmet nyújt, de még a gyengébb eljárásoknál is *megnehezíti, és körülményessé teszi a másolást.* Nem szorosan másolásvédelmi eljárás, de a *Digital Rights Management* (*DRM*) lehetővé teszi, hogy a védelem mellett *a mű útját és felhasználását is nyomon kövessék*. A kiadóknak ez pontos információt ad arról, mire is használták fel a művet, és lehetőséget arra, hogy mindenféle kiegészítő szolgáltatásokkal lássák el a dokumentumokat. Megoldható például, hogy a mű nyomtatását az eredeti licenc nem engedélyezi, és amikor a felhasználó ezt mégis megpróbálja, akkor felajánlja, hogy adott összeg befizetésével egy percen belül már ki is nyomtathatja a művet.

Ha minden mű korlátlanul és ingyen hozzáférhető lenne az interneten, a legtöbben onnan töltenék le, aminek következtében a szerzők, a kiadók és a forgalmazók hatalmas bevételtől esnének el. A másolásvédelemmel megnehezíthető azoknak a dolga, akik le szeretnék másolni, vagy közzé szeretnék tenni a műveket, és ezzel többen „kényszerülnek" megvenni a műveket, azaz legális csatornákon keresztül beszerezni őket, így *a szerzők több bevételhez jutnak*.

A másolásvédelmi eljárások hátrányai

Sajnos még a legegyszerűbb másolásvédelmi eljárásról is elmondható, hogy *megnehezíti a legális felhasználást is*. Ha csak a legegyszerűbb, például PDF fájlokban található védelemre gondolunk, már önmagában az, hogy nem sima szövegként, vagy html-formátumban tesszük közzé a művünket, gondot okozhat egyeseknek. A legtöbb számítógép alapfelszereltségében nincs PDF olvasására képes program; aki modemmel kapcsolódik az internetre, annak például az új 7-es verziójú Acrobat Reader program letöltése, amely 18 Mb, közel egy órát vesz igénybe. Ezt nem mindenki vállalja. A mobiltelefonos böngészés is kezd terjedni, aminél néha még lehetőség sincs ilyen kiegészítő programokat installálni. A hátrányos helyzetűeknek is gondot okozhat mindenféle kiegészítő programok installálása, ha azokat nem támogatja a böngészésüket segítő alkalmazás.

A *másolásvédelem* sajnos *nem tudja megakadályozni az illegális másolást*, és ha éppen azok, akik ennek a dokumentumnak a felhasználói csoportjába tartoznak, könnyedén megkerülik a védelmet, akkor teljesen értelmetlen a használata, csak terhet jelent a szolgáltatónak.

Előfordulhat, hogy egy *jogosult személy kénytelen megkerülni a másolásvédelmet*. Ilyen lehet például, amikor valaki a saját dokumentumát PDF-formába teszi át, és a program, amelyet használ, alapértelmezésben bekapcsolja a másolásvédelmet. Később, ha valamiért nincs már meg az eredeti dokumentum, a felhasználó fel fogja törni ezt a védelmet, hogy hozzájusson a dokumentum tartalmához.

Az *1999. évi LXXVI. törvény a szerzői jogról 95/A §*-a kimondja:

„… a szabad felhasználás kedvezményezettje követelheti, hogy a jogosult a műszaki intézkedések megkerülésével szemben a 95. § alapján biztosított védelem ellenére tegye lehetővé számára a szabad felhasználást…".

Itt a 95. § a műszaki intézkedések megkerüléséről szól, azaz a másolásvédelem megkerülésének a tiltásáról. Ez a szakasz tehát azt mondja ki, hogy annak ellenére, hogy másolásvédelem van a művön, *adott feltételek teljesülése esetén a felhasználók kérhetik a védelem eltávolítását* (pl. szabad felhasználás bizonyos eseteiben, fogyatékos személyek jogos igényei esetén).

Nem mindig jogszerű a másolásvédelem használata; erre legjobb példa a szoftver, amellyel kapcsolatban az eladó nem akadályozhatja meg, hogy a termékről a vevő biztonsági másolatot készítsen saját céljára. Ha valaki például tanulmányokat árul az interneten, akkor használhat másolásvédelmet, de erre fel kell hívnia a vevő figyelmét, hogy az tisztában vele, vásárlás után mire tudja majd használni a dokumentumot, különösen, ha a másolásvédelem megakadályozza, hogy idézeteket emeljen át a műből a sajátjába, ami többnyire jogos elvárás.

A korábban említett DRM *felvet néhány személyiségi jogi problémát*, hiszen az eladó a legtöbb rendszerről pontosan tudja, hogy ki, mikor, melyik művet nézi meg, nyomtatja ki stb. Nem biztos, hogy minden felhasználó szívesen ad ki magáról ilyen információkat, kivált teljesen idegen cégeknek, ahol nincs is lehetősége befolyásolni, hogy ezeket az információkat ki és mire fogja felhasználni. A kéretlen reklámlevelek korában az olyan információnak, hogy melyik felhasználónak mi az érdeklődési területe, mit olvas és milyen gyakran, felbecsülhetetlen értéke van, így még ha az adott cég nem is használná fel, akkor is lehet, hogy betörnek a rendszerébe, és ehhez az információhoz hozzájutva visszaélnek vele. Tudományos terüle-

ten fontos cél, hogy egy adott kutatás híre minél több másik kutatóhoz eljusson, és minél többen hivatkozzanak az adott cikkre vagy eredményre. Ilyenkor *a másolásvédelem csak megakadályozza, hogy mindenki hozzáférjen* a műhöz, és esetenként még azt is, hogy *a webes keresők indexeljék*. Ez azért kellemetlen, mert még ha keresi is valaki a cikkünket, akkor sem fogja megtalálnia például a Google-ban, mert az a másolásvédelem miatt nem fér hozzá a tartalmához.

Megoldások szöveges dokumentumoknál

A teljesség igénye nélkül érdemes néhány elterjedtebb másolásvédelmi eljárást közelebbről is megvizsgálnunk.

A *pdf* és *doc formátumú fájloknál* az Adobe, illetve a Microsoft beépített valamilyen másolásvédelmet. Ezek könnyen használhatók, és legtöbbször nem is okoznak gondot a másik félnek megnyitáskor, ugyanakkor mind a két megoldás könnyen és automatizálva megkerülhető. Ilyen gyenge védelmet egyébként azért is szoktak használni, hogy felhívják a felhasználó figyelmét: ezt a dokumentumot nem szabad másolni, így később – mivel a felhasználó szándékosan megkerülte a védelmet – nem hivatkozhat arra, hogy nem tudta, milyen feltételekkel használhatja az adott művet.

Léteznek olyan megoldások, amelyek *csak az online megjelenítést engedélyezik*. A szöveges változatok nem olyan ismertek, de hang- és videoanyagoknál már sokkal elterjedtebbek azok a műsorok, amelyeket nem lehet menteni, csak meghallgatni, illetve megnézni. Szöveges változataik is azonos elven működnek, és legtöbbször valamilyen kis programot kell installálni a gépre a megjelenítéshez. Ezek a megoldások erősen korlátozzák a felhasználást, és ha nem is olyan egyszerűen, mint az előzőleg említett védelmek, de kis utánjárással megkerülhetők.

Gyakori megoldás, hogy a gyártók olyan, *nem szabványos fájlformátumot alkalmaznak*, amelyet kizárólag az ő megjelenítőjük képes feldolgozni. Hazánkban még nem olyan népszerűek az elektronikus könyvek, mint külföldön, ahol ezek valószínűleg az e-papír elterjedésével válnak tömegessé. Az emberek többsége nem szeret képernyőn olvasni, ezért készítettek olyan eszközöket, amelyek jobban pihentetik a szemet olvasáskor, és ezekre az internetről letöltött könyveket tölthetünk fel. A legtöbb ilyen hardver ismeri a legelterjedtebb formátumú szöveges fájlokat, de a hozzá vásárolt könyvek – csak ez által a hardver által támogatott – zárt formátumban vannak. A megoldás legnagyobb hátránya az, hogy a tartalomhoz való hozzáféréshez rendelkeznünk kell ilyen hardverrel. Egy digitális könyvtár például nem engedheti meg magának, hogy ilyen formátumban adja közre az anyagait, mert ezek a hardverek ehhez nem elég elterjedtek, ráadásul gyártóspecifikus a formátumuk. Ha az ilyen hardver mégis elterjedne, hamarosan meg is jelenne hozzá egy olyan program, amely képes feltörni.

Sokszor használják a védelemnek azt a módját, hogy *korlátozzák a műhöz hozzáférők körét*, és ezzel próbálják megakadályozni, hogy illetéktelenek kezébe kerüljön. Jó ez a megoldás, mivel azok, akiknek szánjuk, nemcsak hozzáférnek, de valahogy meg is találják ezeket a műveket. Ezeknek a rendszereknek általában éppen az a hátrányuk, hogy a mű használatára jogosultak nem is tudnak arról, mihez is férhetnének hozzá. További hátrány, hogy ha egy ilyen rendszerből dokumentum szivárog ki, akkor attól kezdve már nem áll védelem alatt.

A legbiztonságosabb megoldás a *fizikai védelem.* Ha senki sem fér hozzá a dokumentumhoz, senki sem fogja lemásolni. Ez a megoldás kicsit túlzottnak tűnik, de sajnos gyakori. A legszomorúbb példa erre az egyetemi és főiskolai diplomamunkák sorsa, amelyek ugyan elvileg hozzáférhetők a könyvtárban, de nem lehet bennük keresni, ezért lehetetlen megtalálni a több ezer diplomadolgozat között a számunkra érdekeseket. Ezek a munkák a plágiumtól való félelem miatt kerültek erre a sorsra, pedig éppen az lenne a szakmai cél, hogy a műveket egy digitális könyvtárba rendezzék, és azon keresztül minél többen olvassák. Eszményi környezetben a diplomázónak át kellene futnia az összes releváns, és az adott a témában született korábbi dolgozatot, és azokhoz kellene hozzáadnia valami újat, azokból kellene ötleteket meríteni, bírálni az ott felvetett gondolatokat, megerősíteni a mérési eredményeket, kiegészíteni új módszerekkel stb. Ha a diplomamunkák szabadon hozzáférhetők lennének közös, jól kereshető és használható rendszerben, és az újak is ugyanebbe a rendszerbe kerülnének be, akkor a plagizálás könnyen visszaszorítható lenne, ráadásul gyanú esetén a bírálók is könnyedén hozzáférnének az adott művekhez, és kézzel is összehasonlíthatnák őket. Ezzel el is értünk a plágiumkeresők által nyújtott védelem kérdésköréhez.

KOPI-védelem

A másolásvédelem után most nézzük meg, hogy mi az a KOPI-védelem, azaz a plágiumkereső hogyan védheti meg az oktatási intézmények, könyvkiadók, digitális könyvtárak, konferenciaszervezők és más intézmények dokumentumait az illegális másolástól.

A KOPI-védelem előnyei

Nézzük meg, hogyan működik a KOPI-védelem, milyen előnyökkel jár a használata. Ha valaki másol a KOPI rendszerbe feltöltött dokumentumról, akkor *a plagizálás pillanatok alatt kideríthető*. Házi feladatoknál, diplomadolgozatoknál, szakmai cikkeknél a keresést automatikusan el is lehet végezni, és ahhoz lehet kötni a munka elfogadását, hogy a rendszer igazolást ad-e arról, hogy nem talált bizonyos számúnál több egyezést egyik korábbi munkával sem.

Adott egyetemi dolgozatnál például nem elég az, ha a tanár *érzi*, hogy a mű, amelyet a diák beadott, nem az ő munkája, ezt valahogy igazolnia is kell. A plágiumkereső rendszer azonnal *meg is jelöli a forrásokat*, így ezek felkutatására az oktatónak nem kell felesleges időt pazarolnia, sőt a rendszer olyan dokumentumokban is kereshet, amelyekhez neki nincs is hozzáférése, így meg sem találhatná az egyezést.

Az előbbiek miatt *a lebukás kockázata jelentősen megnő*, ami komoly visszatartó erő lehet azoknak, akik maguk is meg tudnák oldani a feladatot, csak egyszerűbb, gyorsabb utat kerestek a munka elvégzéséhez. Sajnos az is előfordul, hogy a diák valaki mással íratja meg a házi feladatát, de ezzel is nagy kockázatot vállal. Külföldön valaki így *bukott le* – nem plágiumkereső használatával, hanem egy figyelmes oktatónak köszönhetően –, mert pénzért vállalt dolgozatírást, amit azután többeknek eladott, mindig csak picit módosítva rajta. A plágiumkereső felfedheti ezeket az eseteket még akkor is, ha különböző oktatási intézményekbe került egy-egy példány a műből.

Mivel nem létezik tökéletes védelem, mindig fontos szempont az, hogy a védelem megkerülése nehezebb legyen, vagy több energiába, pénzbe kerüljön, mint annak az értéke, amit véd. Mint az algoritmus leírásakor kiderült, *ez a védelem nem kerülhető meg automatikusan*, mert legalább minden *n*-edik szót át kell írni a műben ahhoz, hogy ne ismerje fel, természetesen úgy, hogy utána is értelmes maradjon a szöveg, és ne hangozzanak erőltetettnek a mondatok. Ráadásul *n* értéke rendszerről rendszerre változhat, és az is lehet, hogy további finomításokat vezetnek be a rendszer üzemeltetői, azaz elképzelhető, hogy a leggyakoribb szavakat (stopword) törlik a dokumentumból darabolás előtt, a szinonimával rendelkezőket pedig a leggyakrabban használt párjukkal helyettesítik. A plágiumkereső legnagyobb előnye a másolásvédelemmel szemben talán éppen az, hogy *a mű szabadon terjeszthetővé válik*. Nem kell a védelem kérdésével foglalkozni, mindenki el tudja olvasni, még a speciális hardvert vagy szoftvert használók is, valamint a web keresőivel is megtalálhatók. Mindennek eredménye, hogy többen olvassák a művet, ismertebb lesz mind a mű, mind a szerzője, illetve kiadója, és természetesen többen hivatkoznak rá, ami tudományos körökben fontos szempont.

Az egyetemek és főiskolák – a diákszám csökkenésének és a fejkvóták bevezetésének köszönhetően – elkezdtek versenyezni a diákok kegyeiért. Nemcsak az oktatási intézménynek fontos, hogy az általa kibocsátott diplomának mekkora a presztízse, hanem az oda jelentkezőknek is, hogy amikor végeznek, minél jobb esélyeik legyenek a munkaerőpiacon, Többen fognak jelentkezni azokba az oktatási intézményekbe, amelyek diplomái többet érnek. *A plágiumkereső használatával több módon is növelni lehet az oktatási intézményekben a diplomák és dolgozatok értékét*. Az első szempont az lehet, hogy elkerülhetik az olyan kínos eseteket, amikor utólag, már a diploma kiosztása, vagy a dolgozat értékelése után derül fény a csalásra. További előny, hogy a diákok – éppen a lebukás veszélye miatt – sokkal ritkábban fognak plagizálni, és több energiát fektetnek a diplomadolgozatba, ezzel gyarapodik a tudásuk, munkájuk színvonala emelkedik. A legnagyobb hasznot feltehetően az jelenti, hogy forrásként tudják kiadni a korábbi évek munkáit a diákoknak a tömeges plagizálás kockázata nélkül. Így több olyan diplomamunka születhet, amely hozzátesz valamit az előző évek munkáihoz, valami újat nyújt a szakmának, és nem csak megismétli, amit már sokan leírtak az előző évben is. Lehet, hogy mindez utópisztikusnak tűnik, de az olyan digitális könyvtár használata, ahol kereshető formában, esetleg tematikusan rendezve megtalálhatók a szakdolgozatok, igen egyszerű formája lehet annak, hogy adott cégek adott területen jártas, új munkaerőre tegyenek szert, hiszen láthatnák, hogy a kérdéses témában milyen minőségű munkát tett le a valaki az asztalra. Ha valaki kiváló diplomamunkát írna, az

sem lenne kizárt, hogy mire kézbe kapja a diplomáját, már két-három állásajánlatot is kapna különböző cégektől.

A KOPI-védelem hátrányai

A plágiumkereső rendszereknek az előnyök mellett hátrányai, korlátai is vannak. Ahhoz, hogy a védelem érvényesüljön, *egy nagy rendszert érdemes használnia mindenkinek*, vagy pár nagyobbat, mert különben az összes rendszerben keresnie kell a felhasználónak ahhoz, hogy biztos legyen a kezébe került mű egyediségében. Ha pedig valaki biztos akar lenni abban, hogy a művét nem másolják, az összes plágiumkeresőbe be kell töltenie. Egyetemi diplomamunkáknál már az is elegendő, ha az összes, vagy a legtöbb egyetem ugyanazt a rendszert használja.

A másolásvédelem önmagában védi a dokumentumot. Ahhoz, hogy egy plágiumkereső rendszer is védje, *be kell tölteni a védeni kívánt dokumentumokat a rendszerbe*. Ez nagy mennyiségű, rendezetlen, illetve rendszerezetlen dokumentumnál komoly feladat lehet.

Továbbfejlesztési lehetőségek

A KOPI Online Plágiumkereső és Információs Portál többéves működtetése alatt rengeteg tapasztalatot gyűjtöttünk össze, és számos visszajelzést, javaslatot kaptunk felhasználóinktól. Terveink között szerepel ezeknek a megvalósítása, hogy új, még könnyebben használható, és már létező rendszerekbe is könnyen beépíthető plágiumkereső szolgáltatást alakítsunk ki.

A megoldandó feladatok közül a legfontosabb: pontosan azért, hogy minden egyetem, főiskola, digitális könyvtár, kiadó, kutatóintézet, cég stb. saját rendszert üzemeltethessen, valamilyen elosztott rendszert kell kialakítani, ahol minden intézményben önálló KOPI rendszer van, de ezek képesek egymás adatbázisaiban keresni. Ez megoldaná a közös rendszer használatának a problémáját, ráadásul a cégek többsége sokkal jobban megbízik a maga által üzemeltetett rendszerben, mint egy külső fél által fenntartottban. A korábban említett egyirányú tömörítési eljárás segítségével úgy tudnak keresni egymás rendszerében, hogy csak az ujjlenyomatokat (számokat) viszik át. Ez a megoldás nemcsak a dokumentumok biztonságát szavatolja, hanem a hálózati forgalmat is jelentősen csökkenti.

A portál felhasználói jelezték, hogy kényelmes lenne, ha valamilyen szabványos interfészen keresztül (pl. SOAP), programból érhetnék el a KOPI szolgáltatásait, így könnyen beépíthetővé válna ez a plágiumkereső funkció akármilyen külső rendszerbe. Tervezzük ennek az interfésznek a megvalósítását, hogy olyan helyen, ahol már valamilyen rendszer bevált, ne kelljen lecserélni, hanem könnyen kiegészíthető legyen egy ilyen plágiumkereső funkcióval.

A jelenlegi rendszer nem alkalmas a programkódok összehasonlítására, mert ott túl könnyű szisztematikusan kicserélni „szavakat”. Érdekes jövőbeni kutatási téma, hogy ezt miként lehetne megoldani, vagy egyáltalán megoldható-e. A KOPI portál jelenleg nem végez vizualizációt; ha talál egyező dokumentumokat, megnevezi azokat, és a felhasználóra bízza, hogy ha letölti őket, milyen eszközt használ az egyező részek megjelenítésére. Sokkal kényelmesebb lenne a rendszer használata, ha – természetesen a jelenlegi lehetőséget is megtartva – maga is el tudná végezni az egyező részek kiemelését.

Következtetések

Az ilyen rendszert használva a tartalomszolgáltatók – digitális könyvtárak, oktatási intézmények, kiadók – sokkal szabadabban hozzáférhetővé tehetnék a (KOPI-védelem alatt álló) dokumentumaikat, ami előnyös lenne számukra, mivel nagyobb lenne a forgalmuk, többen olvasnák a műveiket, és természetesen többen is hivatkoznának rájuk. A magyar internethasználó közönség is sokat nyerne azzal, ha a jelenleg teljesen elzárt, vagy nehézkesen hozzáférhető dokumentumok elérhetővé válnának, és könnyen használható, kereshető formában megjelennének a gyűjtemények tulajdonosainak a honlapján.

Kifejezések

Darabolás: az az eljárás, amelynél a dokumentumot töredékekre osztjuk fel.l

DRM (*Digital Rights Management):* olyan technológia, amelynek segítségével a jogtulajdonosok a digitális tartalomhoz vagy hardverhez való hozzáférést és használatot ellenőrizhetik, szabályozhatják.

Finomhangolás: a rendszer paramétereinek „kismértékű” változtatása, amelynek célja, hogy az adott felhasználási környezetben a lehető legjobb eredményt adja; esetünkben a darabolási eljárások paramétereinek mó-

dosításával lehet elérni, hogy a rendszer különböző alkalmazási területeken az optimumot nyújtsa.

Hamis pozitív eset: általánosságban olyan eset, amely megfelelőnek tűnik egy bizonyos kritériumnak, azonban valamilyen hiba folytán mégsem az; esetünkben azt a hash-kódolt töredéket hívjuk hamis pozitív esetnek, amely a kódolásnál egyező kódot kapott egy vele nem egyező töredékkel, így a másolatkereső lekérdezés egyezést fog találni ott, ahol ténylegesen nincs egyezés a két dokumentumban.

Hash-kódolás: olyan veszteséges kódolás, amely karakterláncot alakít át fix hosszúságú kóddá; felhasználási területe egyrészről a szöveges adatbázisok, másrészről a kriptográfia.

MD5 (*Message Digest 5*): kriptográfiai algoritmus, amelynek kódja publikus (rfc1321.txt); tetszőleges hosszúságú szöveget 128 bit hosszú kódra képez le, ezáltal veszteséges kódolását adja a bemenetnek.

RFC (*Request For Comments*): szabad terjesztésű ajánlások gyűjteménye, amelyek tényleges szabványnak tekinthetők; leírásuk egyszerű szöveges fájlokban rendelkezésre áll, többek között a *http://www.rfc-editor.org* címen.

Stopword: olyan szavak, amelyek gyakran előfordulnak, a szöveg jelentéstartalmával nem állnak összefüggésben, ezért eltávolításuk a szövegből nem okoz információcsökkenést; pl. névmások, létigék, névelők.

Töredék: egy dokumentum kisebb darabja; két töredék nem feltétlenül független egymástól (átlapolódó eset).

Irodalom

[1] Magyar Értelmező Szótár v1.1. *http://pistvan.extra.hu/mesz.htm*

[2] CSERNOCH Mária: A szavak véletlenszerű megjelenésén alapuló modellek és az irodalmi művek közötti eltérések magyarázata. II. Magyar Számítógépes Nyelvészeti Konferencia. Szeged, 2004. dec. 9−10.

[3] JUOLA, Patrick−SOFKO, John−BRENNAN, Patrick: A prototype for authorship attribution studies. = Literary and Linguistic Computing, 21. köt. 2. sz. 2006. p. 169−178.

[4] Plagiarism Search V 1.0.0. *http://baltic.cse.msu.edu/ heynige1/Search/*

[5] Copyscape by Indigo Stream Technologies. *http://www.copyscape.com/*

[6] Plagiarism Check using Google's Search API. *http://hip2b2.yutivo.org/2006/03/25/plagiarism-check-using-googles-search-api*

[7] Glatt Plagiarism Screening Program. *http://www.plagiarism.com/*

[8] Plagiarism Finder. *http://www.m4-software.de/en-index.htm*

[9] EVE Plagiarism Detection System. *http://www.canexus.com*

[10] KOPI Online Plágiumkereső és Információs Portál. *http://kopi.sztaki.hu*

[11] MTA SZTAKI Elosztott rendszerek osztály, *http://dsd.sztaki.hu*

[12] PATAKI Máté: Szöveges dokumentumok darabolása és tömörítése hash-kódolással − darabolási technikák és másolatkeresés. Budapesti Műszaki és Gazdaságtudományi Egyetem, diplomadolgozat. *http://dsd.sztaki.hu/people/mate_pataki/200201_DiplomaM25.pdf*

Beérkezett 2007. I. 4-én.

Pataki Máté
az MTA SZTAKI Elosztott rendszerek osztályán tudományos főmunkatárs.
E-mail: *Pataki.Mate@sztaki.hu*