DSD
DEPARTMENT OF DISTRIBUTED SYSTEMS

MTA SZTAKI Elosztott Rendszerek Osztály
1111 Budapest, Lágymányosi u. 11
tel: +36-1-279-6212, fax: +36-1-209-5288
web: http://dsd.sztaki.hu, email: laszlo.kovacs@sztaki.hu

KOPI

Online Plágiumkereső és Információs Portál

Pataki Máté

DSD
DEPARTMENT OF DISTRIBUTED SYSTEMS

MTA SZTAKI Elosztott Rendszerek Osztály
1111 Budapest, Lágymányosi u. 11
tel: +36-1-279-6212, fax: +36-1-209-5288
web: http://dsd.sztaki.hu, email: laszlo.kovacs@sztaki.hu

A projektről

A KOPI Online Plágiumkereső és Információs Portál fejlesztése 2003-ban kezdődött az MTA SZTAKI Elosztott Rendszerek Osztálya, valamint a Monash University közreműködésével. A fejlesztés célja egy hálózati plágiumkereső portál létrehozása, amelynek segítségével elektronikus formában rendelkezésre álló dokumentumok közötti hasonlóság felderítésével plágium vagy idézet, de akár ugyanannak a dokumentumnak egy régebbi vagy újabb változata is megtalálható.

A plagizálás

A plagizálás problémájára számos megoldás született, amelyek alapvetően két csoportba sorolhatóak:

- másolás megelőzése
- másolatok felderítése

A másolás megelőzése fontos dolog, de sok esetben a művek egyszerű közzétételét akadályozza, illetve a legtöbb védelmi megoldást idővel kijátsszák, feltörik és a dokumentumok védtelenekké válnak. Az oktatás területén az is elképzelhető, hogy az eredeti dokumentum nem áll védelem alatt, viszont felhasználása és saját munkaként való prezentálása már csalásnak, plágiumnak számít. Mindezen okokból kifolyólag nagy hangsúlyt kell fektetni a másolatok felderítésére, amely a KOPI projekt elsődleges célja.

Plagizálás a digitális könyvtárak területén

A számítástechnika fejlődésével az írott művek előállítási folyamata egyszerűsödött, publikációjuk gyors és könnyű lett, a hagyományos könyvtárak mellett megjelentek a digitális könyvtárak, amelyek megszüntették a földrajzi kötöttségeket, új dimenziót nyitottak a tudás tárházainak. A rengeteg előny mellett azonban a digitális adattárolás a végletekig egyszerűsíti a művek másolását, azok egészének vagy részeinek átvételét, így nagymértékben megkönnyíti a plágiumok létrehozását is. Természetes tehát, hogy a szellemi termékek (képek, irodalmi és zenei művek, tudományos publikációk) eredeti szerzői megfontoltak a digitális könyvtárakkal és az internetes publikálással kapcsolatban, hiszen ezzel jelentősen megkönnyítik a műveikhez való hozzáférést, valamint ezzel párhuzamosan azok illegális másolását, plagizálását is.

Sajnálatos, de az informatikai fejlődéssel képtelen a jog lépést tartani, a visszaéléseket hatékonyan visszaszorítani, így mindenképpen fontos, hogy olyan informatikai megoldások szülessenek, amelyek a digitális szellemi javak előállítóit műveik publikálására ösztönözzék. A KOPI plágiumkereső rendszer használatának egyik legfontosabb eredménye lehet tehát a digitális könyvtárakban tárolt anyagok plagizálóinak felderítése, valamint ezáltal a digitális könyvtárakkal szembeni bizalom növelése.

DSD
DEPARTMENT OF DISTRIBUTED SYSTEMS

MTA SZTAKI Elosztott Rendszerek Osztály
1111 Budapest, Lágymányosi u. 11
tel: +36-1-279-6212, fax: +36-1-209-5288
web: http://dsd.sztaki.hu, email: laszlo.kovacs@sztaki.hu

Plagizálás az oktatás területén

A plágiumkereső rendszer másik fontos területe a digitális könyvtárak mellett az oktatás, azon belül is kiemelten a felsőoktatás. Ezen a területen egyre nagyobb számban jelennek meg a digitális forrásokból összeállított (összeollózott) dolgozatok, diplomamunkák, cikkek, illetve publikációk. Az ilyen típusú visszaélések erősen rontják az intézmények hírnevét és az onnét kikerült diplomások értékét a piacon. A KOPI plágiumkereső rendszer bevezetésével erősen visszaszorítható lenne a másolt, plagizált munkák száma ezen a területen is, valamint jelentősen javulna az oktatási intézmények megítélése is.

A KOPI felhasználási módjai a plagizálás visszaszorításának érdekében

A KOPI rendszer képes dokumentumokat egymással összehasonlítani, ezáltal felderíteni az esetleges illegális másolást, plagizálást. A felhasználó által feltöltött dokumentumokat a rendszer egymással vagy az adatbázisban lévő, mások által feltöltött dokumentumokkal is képes összehasonlítani, alkalmas lesz akár a csak egy bekezdés hosszú átlapolás észrevételére is.

Fontos tulajdonsága a plágiumkeresésnek, hogy a rendszerbe feltöltött anyagok csak kódoltan tárolódnak, a kódolt anyagokból pedig nem nyerhető vissza az eredeti dokumentum. A plágiumkeresés tehát úgy valósítható meg, hogy nem az eredeti szöveghez, hanem annak kódolt változatához hasonlítja a rendszer a vizsgálandó dokumentumot.

A KOPI rendszer a plágiumkeresésen kívül képes kiszűrni a duplikátumokat, valamint meg tudja határozni, hogy a feltöltött szöveg milyen nyelven íródott, ezáltal automatikus kategorizálást tesz lehetővé.

A plágiumkereső szolgáltatásokon kívül a KOPI portál megismerteti az oldalra látogatókat az ide vonatkozó jogszabályokkal és rendeletekkel, valamint lehetőséget biztosít különböző témákban való eszmecserére is a fórumokon keresztül.

Miért van rá szükség?

Joggal kérdezhetné valaki, hogy miért volt szükség erre a plágiumkeresőre, hiszen az Interneten biztos található számtalan hasonló megoldás. Igen, egyre több plágiumkereséssel foglakozó program jelenik meg, viszont a megoldások igen eltérőek. A teljesség igénye nélkül nézzünk meg párat.

Sok olyan, úgynevezett vízjeles megoldás létezik, amely a teljes dokumentumot vagy annak nagyobb részeit védi. Ezek gyakran már formázásnál elvesztik ezt a beleágyazott információt, vagy ahogy valami apróbb változtatás történt a szövegben. Hasonló problémát észlelünk az ellenőrző összeges (checksum) megoldásoknál is: egy szó átírása már elég ahhoz, hogy ne találjuk meg az eredeti dokumentumot.

Sok nyelvfüggő megoldás is létezik. Ezeknek az a jellemzője, hogy csak olyan nyelvű dokumentumokat képesek feldolgozni, amelyekre előtte felkészítették a rendszert. Ide tartoznak az

MTA SZTAKI Elosztott Rendszerek Osztály
1111 Budapest, Lágymányosi u. 11
tel: +36-1-279-6212, fax: +36-1-209-5288
web: http://dsd.sztaki.hu, email: laszlo.kovacs@sztaki.hu

írás stílusát elemző megoldások is. Ezekkel azt lehet megállapítani, hogy két dokumentumot mekkora valószínűséggel írta ugyanaz a személy. Egy teljesen eltérő megközelítés, amikor tesztet generál a program a dolgozatból, ilyenkor a tanár leülteti a diákot maga mellé, és kitölteti vele a tesztet; ha gyenge eredményt ért el valószínűleg másolta a szöveget. Ennek a nyelvfüggőségen kívül még az is a hátránya, hogy időt kell szánni rá, illetve meg kell gyanúsítani a diákot.

Olyan plágiumkereső megoldás is létezik, amely valamelyik internetes keresőt használja fel hasonló dokumentumok után kutatva, így egy óriási méretű, szabadon hozzáférhető adatbázisban keres. Ez előny és hátrány egyben. Előny, mert nagy a rendelkezésre álló dokumentumok mennyisége, hátrány, mert csak a szabadon hozzáférhető oldalakon keres. Azaz a tavalyi diplomadolgozatok nem képezik a keresés részét, hiszen ezeket csak a legritkább esetben publikálják ily módon. Ugyanakkor ez egy igen gyakori probléma ma a felsőoktatásban.

Utoljára, de nem utolsó sorban, sok olyan fizetős szolgáltatást kínál az Internet, amely teljes egészében elfedi belső működését, és csak a hirdetéseiből lehet arra következtetni, hogy pontosan mit tud.

A KOPI ezzel szemben egy nyelvfüggetlen algoritmusra épülő rendszer, amely természetesen a magyar igényekre külön odafigyel. Ez alatt egy magyar felhasználói felületetet és magyar tartalmat értünk. Utóbbi azt jelenti, hogy a rendszer rendelkezni fog több, magyar dokumentumokat tartalmazó adatbázissal is, melyekhez a felhasználók saját dokumentumaikat hasonlíthatják. Ilyen adatbázis lesz például a magyar digitális könyvtárak gyűjteménye, illetve egy Internetes adatbázis is, amely a magyar oldalakat tartalmazza. A KOPI egy portálszolgáltatás, amely annyit jelent, hogy semmilyen installációt nem igényel, és bárhonnét elérhető, ahol van Internet hozzáférés. Ráadásul, akárcsak a SZTAKI Szótár (http://szotar.sztaki.hu), ez is ingyen áll majd a felhasználók rendelkezésére.

A KOPI Portál szolgáltatásai

A szokásos portálszolgáltatások, mint a beszélgetőfórumok, üzenetküldés és FAQ mellett számos más specifikus szolgáltatást is nyújt a rendszer. A fejlesztés teljes menete során megpróbáltunk nagy hangsúlyt fektetni arra, hogy az oldalak mindenki számára hozzáférhetőek legyenek. Ezt a W3C WAI irányelvei alapján tettük meg, így biztosítva, hogy a vakok, gyengén látók, régi böngészővel, gyenge géppel vagy lassú Internet-kapcsolattal rendelkezők is könnyen kezelhessék a portált.

Az oldalra látogatók megismerkedhetnek a plágiumokra vonatkozó jogszabályokkal, egyetemi szabályzatokkal is. Erre nagy szükség van, hiszen nem minden plágium, ami egyezik, például egy idézet nem plágium, ha egyértelműen jelölve van a forrás. Ugyanakkor hiába jelöli meg valaki, hogy a diplomája felét honnét idézi, kevés egyetem fogad el ekkora idézeteket.

Két eltérő hasonlóságkereső szolgáltatást is nyújt a rendszer. Az egyik a feltöltött dokumentumok összehasonlítása egymással. Ezt akkor célszerű használni, ha valaki például feltölti a diplomáját és az irodalomjegyzékben szereplő műveket, és akkor a rendszer megállapítja, hogy mekkora az egyezés az egyes forrásokkal. A másik szolgáltatás a felhasználó dokumentumait egy adatbázisban szereplő többi dokumentumhoz hasonlítja. Ilyen adatbázisok lesznek például:

- A felhasználó vagy mások által feltöltött dokumentumok

MTA SZTAKI Elosztott Rendszerek Osztály
1111 Budapest, Lágymányosi u. 11
tel: +36-1-279-6212, fax: +36-1-209-5288
web: http://dsd.sztaki.hu, email: laszlo.kovacs@sztaki.hu

- Internetről letöltött oldalak
- Digitális könyvtárak (MEK)
- Egyetemi diplomák

A másolatkereső rendszer működése

A legelső lépés egy ilyen programban a dokumentumok beszerzése. Mivel ehhez a felhasználáshoz a formázási paraméterekre nincs szükség, a legegyszerűbb egy sima szövegfájl használata. Minden olyan dokumentum, amelyik nem ilyen formában található, egy ezt megelőző lépésben konvertálásra kerülhet.

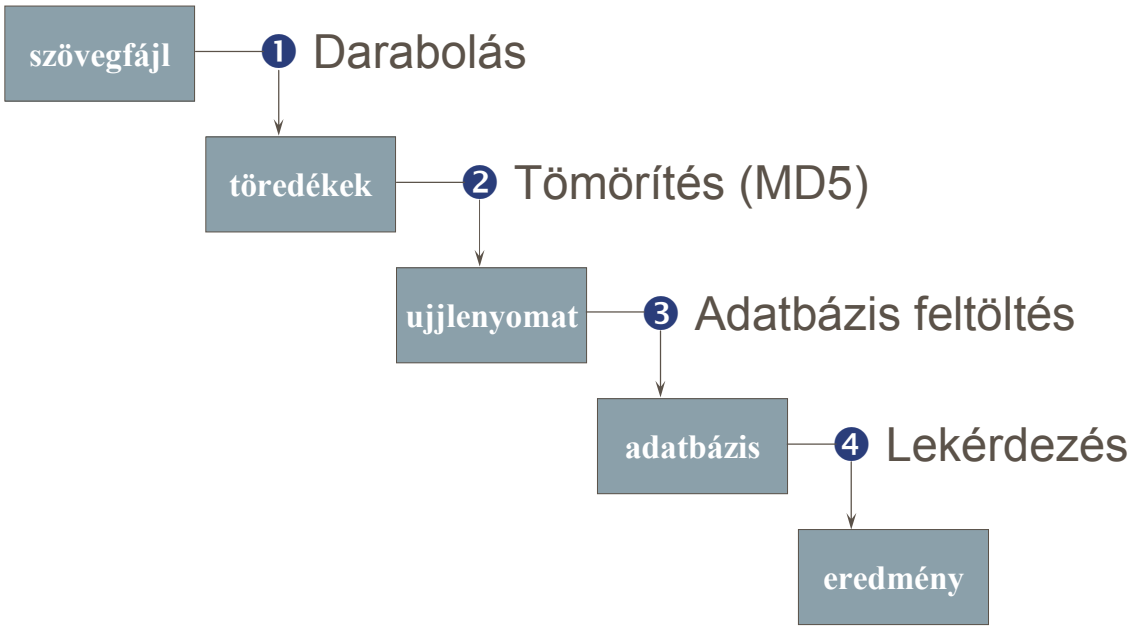
A szövegfájlokat fel kell darabolni kisebb részekre, úgynevezett töredékekre, majd az ezt követő lépésben a töredékek eltárolódnak egy adatbázisban. Mivel ezek a töredékek sok helyet foglalnak el, nem az eredeti töredék kerül eltárolásra, hanem annak egy úgynevezett „ujjlenyomata”. Ezt egy megfelelő tömörítő eljárással kapjuk az eredeti töredékből.

Az adatbázis feltöltése tetszőleges számú lépésben történhet, ehhez minden új dokumentumot fel kell darabolni, majd a töredékek ujjlenyomatát el kell tárolni. A lekérdezést is akármikor elvégezhetjük, akár minden újonnan beérkezett dokumentum eltárolása után is.

Ha később kíváncsiak vagyunk arra, hogy két dokumentum között van-e egyezés, csak le kell kérdeznünk az adatbázisból, hogy hány közös töredéke van ezen két dokumentumnak.

Amennyiben rendelkezésünkre állnak az eredeti dokumentumok, a felhasználó dolgát megkönnyítve – például a hasonlónak ítélt fájlokat egymás mellé téve – vizualizálhatjuk is eredményünket.

Az alábbi ábra a teljes folyamatot ábrázolja:



MTA SZTAKI Elosztott Rendszerek Osztály
1111 Budapest, Lágymányosi u. 11
tel: +36-1-279-6212, fax: +36-1-209-5288
web: http://dsd.sztaki.hu, email: laszlo.kovacs@sztaki.hu

A darabolási eljárás egy nagyon lényeges pontja a rendszernek, hiszen az egész rendszer működése az azonos darabok megtalálásán alapul. Amennyiben ezek túl nagyok, a rendszer nem fogja észrevenni a kisebb egyezéseket. Túl kicsi darabok esetén meg gyakran fog teljesen különböző dokumentumokban is fog azonos szófordulatokat, kifejezéseket találni.

A KOPI rendszer egy új eljárást használ, amely a szavas és az átlapolódó szavas darabolások egyesítéséből áll. Ennek ismertetésére most helyhiány miatt nincs módunk, viszont az alább felsorolt weboldalakról kiindulva részletes információkhoz lehet jutni:

- http://www2003.org/cdrom/papers/poster/p186/p186-Pataki.html
- http://www.csse.monash.edu.au/projects/MDR/papers/ICCS2002-monostori.pdf

A KOPI projekt helyzete

A főbb funkciói már elkészültek a rendszernek, és a fejlesztés is pár héten belül befejeződik. A tesztelések után, előre láthatóan

2004. május végétől

lesz hozzáférhető az alábbi címen:

http://kopi.sztaki.hu