

KOPI-Fotó: Plágiumkeresés egy lefotózott oldal alapján

KOPI-Photo: Searching for plagiarism via a photo

Pataki Máté^a, Micsik András^b, Kovács László^c, Szabó Mihály

MTA SZTAKI

^amate.pataki@sztaki.hu

^bandras.micsik@sztaki.hu

^claszlo.kovacs@sztaki.hu

Absztrakt: Az MTA SZTAKI közismert, számos felsőoktatási intézményben használt KOPI plágiumkeresője új szolgáltatással bővült: az iPhone alkalmazás képes egy lefotózott oldalról megmondani, hogy annak ki az eredeti szerzője. A KOPI Online Plágiumkereső Portál egy egyedülálló, nyílt szolgáltatás az internetező közönség számára, amely lehetővé teszi, hogy a felhasználók saját dokumentumaik valamint mások által feltöltött dokumentumok között azonos részeket, hasonlóságot, esetleg teljes egyezést keressenek. A nyelvfüggetlen algoritmusnak köszönhetően ez bármely európai nyelven megtehető. Ehhez a szolgáltatáshoz kapcsolódik a KOPI-Fotó alkalmazás, amely egy szöveges oldalról mobiltelefonnal készült kép alapján tud keresni a KOPI Plágiumkereső adatbázisában, a magyar weben és az angol Wikipédiában. Használata egyszerű, csak le kell fotózni egy jól olvasható részt egy szövegből, és a program megmondja, hogy az megtalálható-e valamelyik a KOPI által korábban eltárolt dokumentumban, cikkben, oldalon. Amennyiben van találat, a KOPI-Fotó az egyező részeket kijelzi, és lehetővé teszi, hogy a felhasználó a megfelelő oldalra navigáljon. Ha a szöveg vagy annak egyes részei több oldalon is megtalálhatóak, akkor mindegyiket felsorolja, a legtöbb egyezést tartalmazó találattól kezdve. Előadásunkban egy demó keretében is bemutatjuk az új szolgáltatás működését.

Kulcsszavak: plágiumkereső, mobil alkalmazás, fotó, OCR, KOPI

Abstract: KOPI, the well-known Plagiarism Search Portal in Hungary, developed by the Institute for Computer Science and Control of the Hungarian Academy of Sciences (MTA SZTAKI) came out with a new service called KOPI-Photo: the iPhone app is capable to determine the original author of a text using a photo taken from a page of the document. The KOPI Online Plagiarism Search Portal is an open service for web users that enables them to check for identical or similar contents between their own documents and the files uploaded by other authors. This check function works in any European language, due to the language-independent algorithm. The new KOPI-Photo application enables the users of the service to search for the source of a text document by taking a picture of a page. The search is performed in the KOPI database, the English and the Hungarian Wikipedia and on the Hungarian Web. When a match is found the similar texts are displayed to the user so they can navigate to the source found by KOPI. If there are more sources, than all of them are listed, starting with the one that includes the most identical match. During the presentation we demonstrate how KOPI-Photo works.

Keywords: plagiarism search, mobile application, OCR, KOPI

1. Bevezetés

A KOPI Portál kifejlesztését a volt Informatikai és Hírközlési Minisztérium támogatásával az MTA SZTAKI Elosztott Rendszerek Osztálya (DSD), a Melbourne-i Monash Egyetemmel együtt, annak eredményeit felhasználva végezte. 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 felkutatásában, a digitális könyvtárak védelme az illegális másolatoktól, a diákok tájékoztatása a plagizálásról és az idézés helyes módjáról, valamint a cikkek, dolgozatok, diplomamunkák értékének növelése az eredetiségük igazolásával. Érdemes kiemelni – ez az összes hasonló szolgáltatásra is igaz –, hogy a rendszer maga nem tudja megállapítani, hogy egy adott szövegrész idézet-e vagy plágium; az ilyen rendszer csak arra képes, hogy jelzi a felhasználónak, az adott dokumentumban mely más dokumentumból talált meg részeket, mekkora közöttük az átfedés vagy a hasonlóság. Annak a megállapítása, hogy ez szabályos módon történt idézés-e, és helyesen szerepel a forrás megjelölése, már a felhasználóra van bízva.

A KOPI Portál legfőbb célja a plágiumok illetve a plagizálás visszaszorítása, melynek elősegítésére az oldalon több szolgáltatás is található. A legfontosabb közülük az az információgyűjtemény, amely a plágiummal kapcsolatos tudnivalókat gyűjti össze. Mivel sokan nem tudják pontosan, mi számít plágiumnak vagy nem ismerik az idézés pontos szabályait, ezért a KOPI Portálon megtalálható a plágium definícióján kívül egy részletes leírás is arról, mi a plágium és milyen fokozatai vannak, valamint egy útmutató a helyes idézés módjáról. Az ide vonatkozó jogszabályok mellett az egyetemi szabályzatok is helyet kaptak a linkgyűjteményben, nem csak azért, hogy lássák a hallgatók, hogy milyen következménnyel jár a plagizálás, hanem azért is, mert még az idézésnek is pontos szabályai vannak. Egy diplomadolgozatban például nem lehet meghatározott mennyiségnél több idézet, hiába jelöljük meg a szerzőt, hiszen valami újat, valami sajátot is hozzá kell tenni a korábbi ismeretekhez, eredményekhez ahhoz, hogy a diplomamunka önálló, teljes munkának számítson. Hasonló módon, ha két diák közös témában ír diplomát, akkor is a diploma teljes terjedelmének csak megadott százaléka lehet közös, a többinek teljesen egyéni munkának kell lennie.

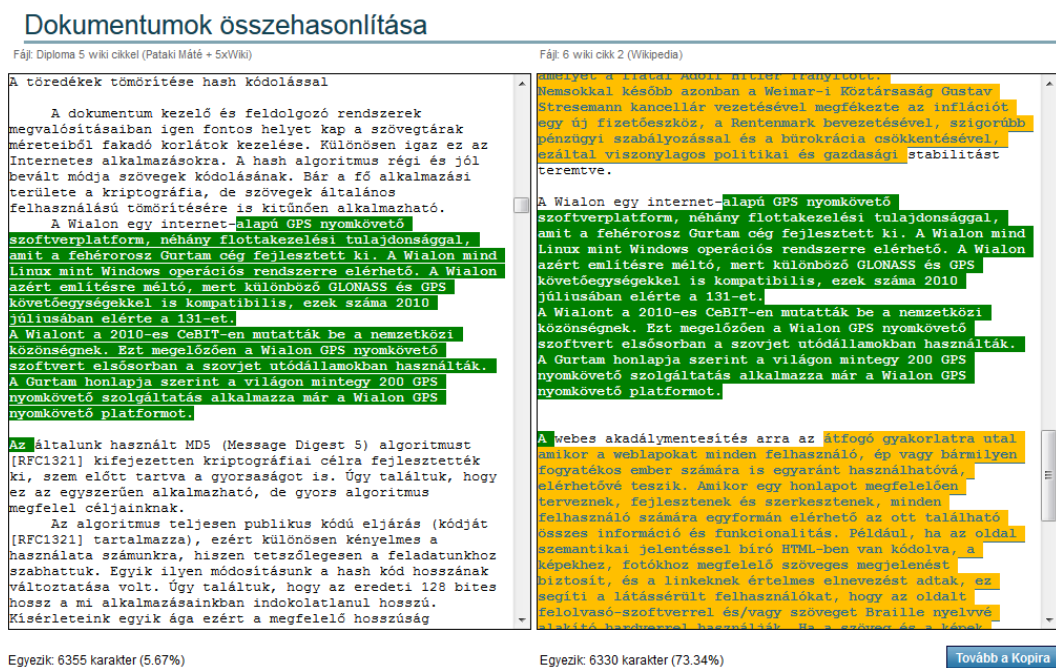
A rendszer lehetővé teszi azt is, hogy ha hasonlóságot vagy esetleg plágiumot talál egy felhasználó, akkor felvegye a kapcsolatot azzal, aki a másik művet feltöltötte, így meg lehet beszélni, hogy pontosan mi is az a mű, melyik az eredeti, ki kiről másolt. A portálnak magyar és angol felhasználói felülete is van, ezzel is elősegítve mind a felhasználói kör, mind az adatbázis bővülését. A KOPI Portál lényegét természetesen a plágiumkereső szolgáltatások adják. Azért, hogy később az adott felhasználó és a többiek által is azonosíthatóak legyenek a dokumentumok, a portálba feltöltött műveknek érdemes megadni a címét és a szerzőjét. Egyéb, részletesebb információk – mint például: kiadó, kiadás éve, kulcsszavak, vagy személyes megjegyzés – bevitelét is lehetővé teszi a rendszer.

2. Folyamatos fejlődés

Az informatika rohamos fejlődésével a KOPI Portálnak is lépést kell tartania, egyfelől hogy minél jobban kielégítse a felhasználók igényeit, emellett kellő visszatartó erő legyen a plágiumok visszaszorításában. Az algoritmusok fejlesztése 2001-ben kezdődött, a KOPI szolgáltatás 2004-ben indult. 2007-ben az első három év üzemeltetési tapasztalatai alapján továbbfejlesztettük a rendszert. 2011-ben a világon elsőként beépítettünk a KOPI Plágiumkeresőbe egy fordítási plágiumok megtalálására képes, saját fejlesztésű algoritmust, amely a teljes angol Wikipédiában keres.

2014-ben újabb szolgáltatásokkal bővült a közkedvelt plágiumkereső. Az eddig elfogadott fájlformátumokon (doc, docx, pdf, rtf, html) felül már a nyílt odt formátumot is támogatja a rendszer, valamint megbirkózik a szkennelt dokumentumot tartalmazó pdf fájlokkal is, amelyre erős igény mutatkozott a régebbi, csak papír formátumban meglévő diplomák, dokumentumok nagy száma miatt.

A hasonlóságok megtekintését, a másolások és idézetek jobb elkülönítését segíti az ugyancsak idén debütált új szolgáltatásunk, a dokumentum-összehasonlító. Ennek segítségével bármely két, egymással közös részt tartalmazó dokumentum megjeleníthető egymás mellett, az egyező részeket a rendszer színes háttérrel kiemelve mutatja.

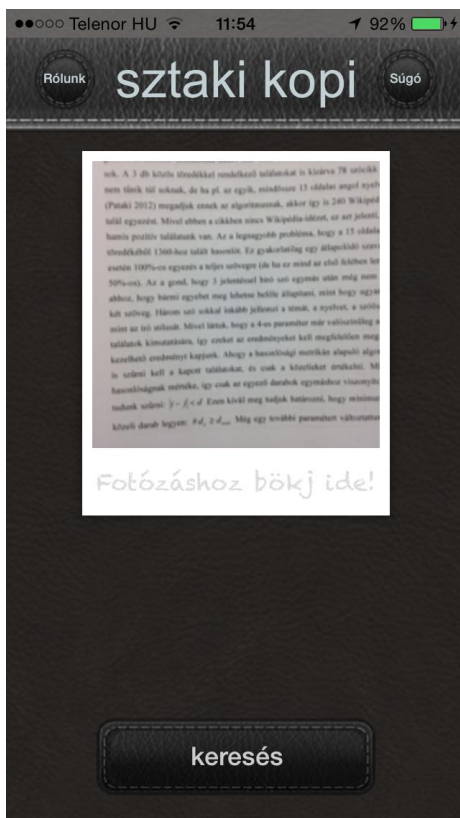


1. ábra. Dokumentumok összehasonlítása a KOPI-ban

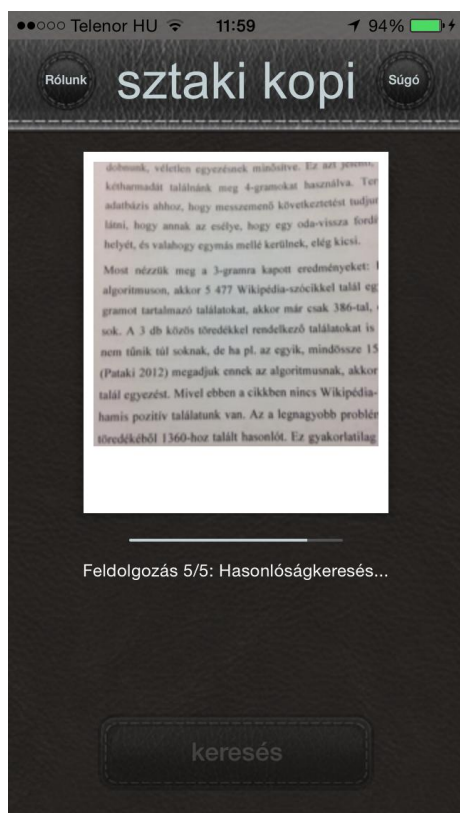
A színes kiemelés segítségével könnyen azonosíthatók a dokumentumok közötti egyezések, és látható az is, megfelelően jelölve van-e az idézet.

3. KOPI-Fotó

A KOPI Portál legújabb szolgáltatása az iPhone telefonra letölthető applikáció, a KOPI-Fotó, amely egy lefotózott oldal alapján annak eredeti forrását képes megállapítani. Az alkalmazás használata nagyon egyszerű: elindítás után egy fényképet készítünk az adott szövegről (2. ábra), majd a keresés gombbal elindíthatjuk a plágiumkeresést. A telefon előbb átalakítja a képet a megfelelő formátumba és méretre, majd elküldi azt a szervernek. Itt egy előfeldolgozás után egy OCR program segítségével a képet szöveggé alakítja a rendszer, majd ezen a szövegen végez plágiumkeresést.

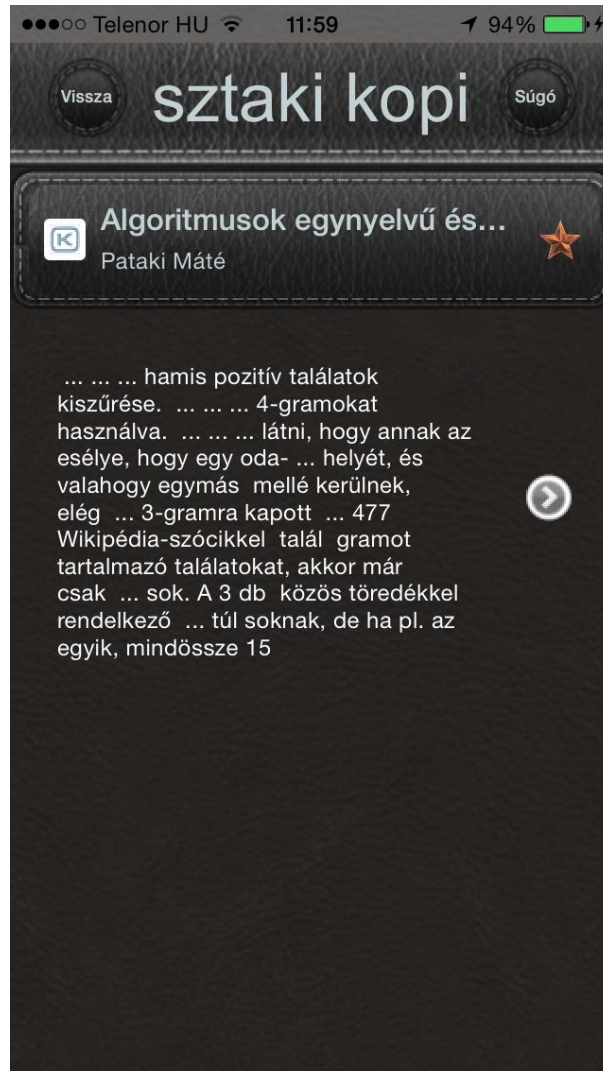


2. ábra. Fénykép készítés



3. ábra. Fut a keresés

Az eredmény a szöveg mennyiségétől függően pár másodpercen belül megjelenik. A találatokra kattintva meg lehet nézni az egyező részeket, valamint a Wikipédia esetén lehetőség van a megfelelő Wikipédia oldalra navigálni.



3. ábra. Keresés eredménye

Az applikáció a KOPI-ba feltöltött fájlok között, az angol és magyar Wikipédiában, valamint egy internetes adatbázisban keres. Utóbbiban több tíz millió magyar oldal található, főleg olyan honlapokról, amelyekben nagyobb mennyiségű szöveges tartalom van, illetve amelyek témában illeszkednek a KOPI-ba jelenleg feltöltött dolgozatokhoz.

4. Összefoglaló

A KOPI-Fotó rendszer jól kiegészíti az online plágiumkeresőt, és alkalmas arra, hogy ha valaki kétes eredetű szöveget találkozik, akár bírálatra beadott diplomadolgozatban, akár egy

újságcikkben, akkor megnézhesse, hogy a KOPI adatbázisában megtalálható-e, vagy nem a Wikipédiából, esetleg az internetről származik a kérdéses szövegrész.

Terveink között szerepel, hogy a másik két közkedvelt mobil platformra, Android-ra és Windows-ra is elkészítsük az applikációt, minden felhasználó számára elérhetővé téve a KOPI-Fotó szolgáltatást.

Irodalomjegyzék

- [1] Micsik András, Pataki Máté és Garzó András: A KOPI Plágiumkereső terhelésének elosztása cloud környezetben, *Networkshop 2013 Sopron*,
<http://nws.niif.hu/ncd2013/docs/ehu/061.pdf>
- [2] Pataki Máté és Marosi Attila: Searching for Translated Plagiarism with the Help of Desktop Grids, *Journal of Grid Computing* Vol 11, Issue 1, pp. 149-166, DOI: 10.1007/s10723-012-9224-5
- [3] Pataki Máté: Algoritmusok egynyelvű és különböző nyelvek közötti fordítások és plágiumok megtalálására, *Pázmány Péter Katolikus Egyetem, PhD disszertáció*, 2013
- [4] Pataki Máté, Vajna Miklós és Marosi Attila: Wikipedia as Text, *ERCIM News* 89, April 2012, <http://ercim-news.ercim.eu/en89/ri/wikipedia-as-text>