Algoritmusok egynyelvű és különböző nyelvek közötti fordítások és plágiumok megtalálására

doktori (Ph.D.) disszertáció tézisei

Pataki Máté

Témavezető:
Prószéky Gábor, az MTA doktora

Pázmány Péter Katolikus Egyetem,
Információs Technológiai Kar,
Multidiszciplináris Műszaki Tudományok Doktori Iskola

Firenze, 2011.

Budapest, 2012.

Bevezetés

A plágium nemcsak a felsőoktatásban, hanem számos más szakterületen is komoly problémákat okoz. Ahogy terjednek a számítógéppel beadható dolgozatok és a diákok egyre fiatalabb korban ismerkednek meg a számítógéppel, internettel, úgy szivárog be a plagizálás a középiskolákba is. A tudományos életben is sajnos egyre gyakrabban lehet találkozni plagizált cikkekkel, gondolatokkal. A digitális könyvtárak terjedését is lassítják az illegális másolatok, mert a szerzők – nem teljesen alaptalanul – tartanak a bevételkieséstől. A könyvkiadóknál is gyakran azért ragaszkodnak a papír alapú kiadványokhoz, mert ott sokkal könnyebb az illegális másolást normál keretek közé szorítani. A cégek honlapján található tartalmakat vagy akár teljes honlapokat is egyre gyakrabban másolják le konkurens cégek. A legnagyobb internetes lexikon, a Wikipédia is küzd a plágiumokkal. A Wikipédiára felkerülő anyagok bárki számára ingyenesen elérhetőek és bárki fel is tölthet tartalmat, emiatt viszont rendszeresen ellenőriznie kell az adminisztrátoroknak a tartalmakat, mert nem engedhetik meg, hogy valaki (akár jószándékból), engedély nélküli, jogvédett tartalmat tegyen fel az oldalaikra.

A plágiumkeresés ma már elképzelhetetlen számítógépes segítség nélkül. Senki sem ismerheti az összes, az adott témában megjelenő művet, cikket, diplomát, honlapot. Egy szakdolgozat esetében nem elég érezni, hogy az adott mű plágium, azt be is kell bizonyítani. Ehhez elengedhetetlen egy olyan eszköz, amely hatalmas mennyiségű anyagot rövid idő alatt át tud nézni, és meg tudja nevezni az adott dolgozathoz felhasznált forrásokat és az egyezés mértékét.

A plágiumok elleni védekezés műszaki megoldásait alapvetően két csoportba oszthatjuk, a másolás megakadályozását elősegítő eszközök (másolásvédelem), és a másolás felderítését lehetővé tevő eszközök (plágiumkeresők). Nehéz megóvni 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, az általa használt szoftverkörnyezettől függetlenül. A legtöbb másolásvédelmi rendszer könnyen megkerülhető, így csak névleges védelmet biztosít; más rendszerek sokkal jobban védenek, körülményes a megkerülésük, de csak kiegészítő szoftverekkel, esetenként dedikált hardverrel együtt használhatóak, amit csak akkor fog installálni, megvenni a felhasználó, ha számára igazán értékes a tartalom, amelyet véd. A hátrányos helyzetűek (vakok, gyengénlátók, siketek, elavult gépet használók…) gyakran nem is képesek elérni ezeket a védett tartalmakat, így

ezen eljárások bizonyos esetekben még akár jogsértőek is lehetnek (1998. évi XXVI. törvény 6.§).

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 tűntesse fel azt. Ez a védelem kettős: egyrészt másolatot találva a rendszer rögtön meg is nevezi az eredeti forrást és az átfedés mértékét; másrészt, ha az ilyen rendszer létezése széles körben ismert és használata elterjedt, akkor a legtöbben nem fogják felvállalni a plagizálás kockázatát, kitéve magukat a lebukás veszélyének.

Módszertan

A plágiumkereső rendszereknek igen sok fajtája létezik, és legtöbbjük jól használható bizonyos területeken, ugyanakkor jelentős részükre vonatkoznak olyan megkötések, melyek miatt például digitális könyvtárak vagy egyetemi diplomák esetében nem használhatóak. A kutatásaim során két alapvető irányelvet vettem alapul. Az eljárások, algoritmusok, amelyeket továbbfejlesztek, illetve létrehozok, tegyék lehetővé, hogy nagy mennyiségű szöveget, automatikusan, emberi beavatkozás nélkül lehessen feldolgozni és rajta hasonlóság, illetve plágiumkeresést végezni. A másik fontos szempont a nyelvfüggetlenség volt, illetve, hogy az algoritmus működjön magyar nyelvre. Ez utóbbi azért is lényeges, mert Magyarországon ilyen rendszerek a kutatás kezdetén egyáltalán nem is léteztek, nem voltak elérhetőek.

A kutatásaim eredményeképp létrejövő algoritmusokat minden esetben igyekeztem működő, és mások által is kipróbálható, tesztelhető rendszerbe beépíteni, ezzel nem csak bizonyítva azok működőképességét, de elősegítve a téma hazai elterjedését is.

Új tudományos eredmények

A ma használatos plágiumkereső algoritmusok, amelyek kisebb egyezést is ki tudnak mutatni – azaz nem csak teljes dokumentumokat, több oldalas egyezéseket keresnek – valamilyen daraboló eljáráson nyugszanak. A daraboló eljárások vizsgálata során sikerült kettőnek is egy-egy negatív tulajdonságát az algoritmus megváltoztatásával kijavítanom.

1. tézis

Létrehoztam a félig átlapolódó szavas darabolást, mely a szavas darabolás és az átlapolódó darabolás egyesítése plágiumkeresési célokra, amiről bebizonyítottam, hogy ugyanolyan hatékony a hasonlóságok felismerésében, mint az átlapolódó szavas darabolás, ugyanakkor

implementációtól függően vagy n-ed akkora adatbázist igényel, vagy a lekérdezési idő csökken n-ed akkorára (ahol n a szavas darabolás paramétere). [45]

Jelöljük *W*-vel a dokumentumban található szavak számát, valamint *n*-nel a daraboló eljárás paraméterét, *Ch*-val a töredékek halmazát. A kutatás során elkészítettem egy új darabolási, illetve lekérdezési eljárást is, amelyet félig átlapolódó darabolásnak neveztem el. Ennek lényege, hogy az egyik dokumentumot (*q*) átlapolódó szavas darabolással, a másikat (*db*) szavas darabolással dolgozzuk fel, majd ezeket hasonlítjuk össze egymással. Ez a megoldás kiküszöböli a szavas darabolásnál tapasztalt fázisproblémát, hiszen az egyik dokumentumból az összes lehetséges darabot előállítjuk. A másik dokumentumot viszont csak szavas darabolással daraboljuk, így lehetőségünk van vagy az adatbázis méretét ($\sim |Ch_{db}|$) csökkenteni *n*-ed részre: $|Ch_{db}| = W/n$, $|Ch_q| \approx W$ vagy a keresési időt ($\sim |Ch_q|$), azaz a lekérdezések számát: $|Ch_{db}| \approx W$, $|Ch_q| = W/n$

Ezen darabolási eljárás használata esetén egy szó beszúrása, törlése, illetve átírása mind-mind maximum egy-egy hibát okoznak, azaz egy töredék fog csak módosulni, a többit a rendszer továbbra is azonosnak fogja értékelni. Ez teljes mértékben egyezik az átlapolódó szavas darabolás esetében tapasztalttal, ahol *n*-szer ennyi töredék van, de ezek a hibák mind *n* darab töredéket érintenek, azaz mindkét eljárás esetében, ahhoz, hogy egy egyezőséget ne találjon meg a rendszer minimum minden *n* szavanként egy különbségnek kell lennie.

2. tézis

Bebizonyítottam az átlapolódó hash-kódon alapuló darabolásról, hogy segítségével kiküszöbölhető a hash-kódon alapuló darabolás szövegfüggősége, és így a gyakorlatban is alkalmassá válik akár ismeretlen szövegek darabolására is. [19] [44] [45]

Bebizonyítottam, hogy a hash kódon alapuló darabolási eljárás teljesítménye nagyban függ a választott hash kódtól és a szöveg stílusától, témájától. Emiatt ez az algoritmus nem hatékony ismeretlen eredetű szövegek vagy nem homogén, nem azonos témájú és stílusú szövegeket tartalmazó adatbázisok hasonlóságainak kimutatására. Megmutattam, hogy több hash érték párhuzamos használatával ki lehet küszöbölni ezt az szövegfüggőséget. Ez lehetőséget biztosít arra, hogy a felhasználási terület függvényében kompromisszumot kössünk a magasabb és egyenletesebb fedési érték és egy nagyobb adatbázisméret között.

3. tézis

Létrehoztam a jelenleg nyelvfelismerésre széles körben használt n-gram algoritmusnak egy új változatát, amely a korábbi algoritmus végeredményét megtisztítja a nyelvek hasonlóságából adódó hamis pozitív találatoktól. Az új algoritmus a dokumentum szövegében 30%-nál nagyobb arányban jelen lévő nyelveket képes felismerni, akkor is, ha ezek a részek nem egyben, összefüggő szövegként, hanem elszórva találhatóak meg. Az új algoritmus alkalmas webes korpuszokban található, a plágiumkeresést már negatívan befolyásoló mennyiségben más nyelvet tartalmazó dokumentumok kiszűrésére. [24]

Nyelvfelismerésre az egyik leggyakrabban használt algoritmus az n-gram algoritmus, melyet használva csak egyszer kell végigolvasni a dokumentumot és az n-gram statisztikákból meg lehet állapítani, hogy a dokumentum milyen nyelven íródott, valamint – ha vannak megfelelő mintáink – még a kódolását is meg tudja határozni. Ez az algoritmus a szöveg legvalószínűbb nyelvét meg tudja állapítani, ugyanakkor többnyelvű szövegek esetén a második legvalószínűbb nyelv nem a második legalacsonyabb pontszámot kapja. Ennek az oka a nyelvek hasonlósága. Ezt az n-gram szinten lévő, a nyelvrokonsággal nem mindig egyező, hasonlóságot használtam fel arra, hogy a hasonlósági listából kiszűrjem a hamis pozitív találatokat.

Egy D dokumentumra kapott százalékos n-gram hasonlóság (h) a százalékos hasonlóság mértékének csökkenő sorrendjében legyen: h_1, h_2, h_3 stb., a nyelveket jelölje L_1, L_2, L_3, azaz a h_1 a D dokumentum hasonlóságát mutatja az L_1 nyelvű mintánkkal, százalékban. A nyelvek közötti százalékos hasonlóságot pedig jelöljük h^{LiLk}-val. h_i' az új algoritmus által az L_i nyelvre adott érték.

$$h_i' = h_i \quad \text{ha } i=1$$

$$h_i' = h_i - \frac{\sum_{k=1}^{i-1} h_k \times h^{LiLk}}{\sum_{k=1}^{i-1} h_i} \quad \text{ha } i>1$$

Az algoritmus tulajdonképpen minden nyelv valószínűségét csökkenti az előtte megtalált nyelvek valószínűségével, így kompenzálva a nyelvek közötti hasonlóságból adódó torzulást.

4. tézis

Létrehoztam egy fordítási plágiumok megtalálására képes új algoritmust, amely az n-szavas darabolás helyett mondatokra bontja a szöveget, és egy a fordítás menetét utánzó hasonlósági metrika segítségével hasonlítja össze a mondatokat egymással, hogy megállapítsa, mekkora eséllyel fordításai azok egymásnak. Az új algoritmus lényeges tulajdonsága, hogy nem kell hozzá gépi fordító, elég csupán egy szótár is, amit sokkal könnyebb beszerezni és folyamatosan fejleszteni. [1][11][23]

Az algoritmus alapja a szózsák modell: egy *n* szóból álló mondatot (S_n) képviseljenek a benne lévő szavak ($w_1, w_2, ..., w_n$).

$$S_n^x = \left\{ w_1^x, w_2^x, ..., w_n^x \right\}$$

Két mondat hasonlósági mértékét (*Sim*) az alábbiak szerint definiáljuk:

$$\mathrm{Sim}\left(S_n^x, S_m^y\right) = \min\left(\alpha \cdot \left|S_n^x \cap S_m^y\right| - \beta \cdot \left|S_n^x \setminus S_m^y\right|, \alpha \cdot \left|S_m^y \cap S_n^x\right| - \beta \cdot \left|S_m^y \setminus S_n^x\right|\right)$$

Ahol α és β a közös illetve a hiányzó szavak súlyozására használt konstansok. Jelölje Sim_a és Sim_b a hasonlósági metrika által adott hasonlóság mértékét az *a*-adik és *b*-edik mondatok esetében. Definiáljunk két állandót, két hasonlósági mértéket SIM_1 és SIM_2, ahol $SIM_1 < SIM_2$ valamint egy távolság-értéket, *d*-t. Ezek alapján leírhatjuk, hogy akkor tekintjük az adott dokumentumot találatnak, ha az alábbi igaz:

- $\mathrm{Sim}_a \geq \mathrm{SIM}_2$ vagy
- $\mathrm{Sim}_a \geq \mathrm{SIM}_1$ és $\mathrm{Sim}_b \geq \mathrm{SIM}_1$ és $|a - b| < d$

Az új hasonlósági metrika használhatóságát a következő tézis mondja ki.

5. tézis

Megmutattam, hogy a fordítási plágiumok megtalálására képes új algoritmus a gyakorlatban is használható, más algoritmusokkal összemérhető eredményt ér el, valamint hogy a fedés értéke nem függ az adott nyelvpártól: a magyar-angol nyelvpár esetén a fedése 83%, míg a pontossága 40% volt, a német-angol nyelvpárnál a fedése 83%, míg a pontossága 77% a 12 Wikipédia-cikket tartalmazó tesztkorpuszon. [11][23]

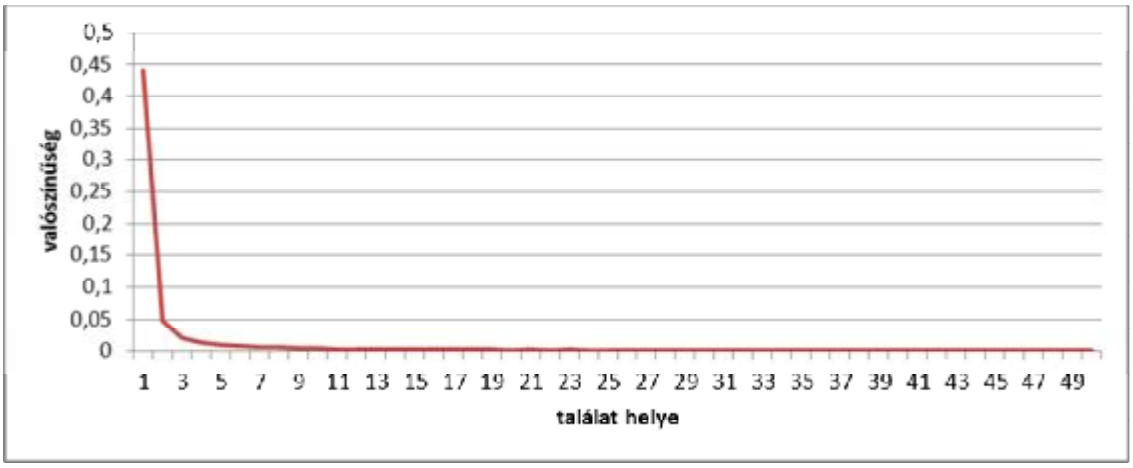
A tesztek elvégzéséhez a teljes, körülbelül négymillió oldalt tartalmazó, angol Wikipédiát dolgoztam fel. Ebből választottam ki véletlenszerűen 12 szócikket, melyeket fordítók magyarra és németre fordítottak le. Megmutattam, hogy az új algoritmus képes ezeket megtalálni, és mindkét szövegen, mindkét nyelven magas, 80% feletti fedést ér el. Ez a

mondat szintű fedés (*r*). Annak a valószínűsége, hogy minimum egy lefordított mondatot megtalál a rendszer *k* darab mondatból: $1-(1-r)^k$ ami könnyen belátható, hogy a mondatok számának a növekedésével tart az 1-hez, azaz nagyobb egyező részeket sokkal magasabb valószínűséggel talál meg.

6. tézis

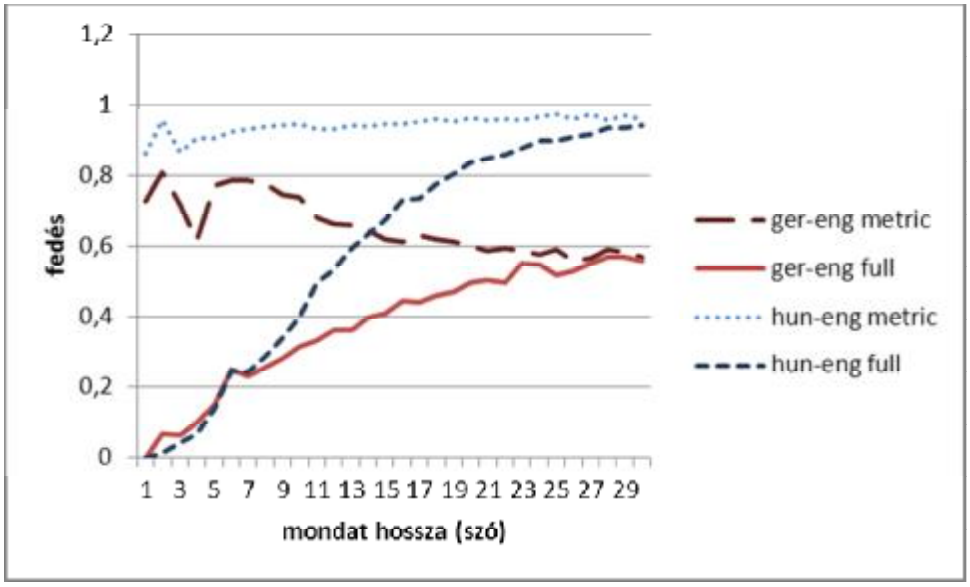
Megmutattam, hogy a fordítási plágiumok megtalálására képes új algoritmusnak az adatbázis méretéhez viszonyított lineáris futási ideje konstansra csökkenthető információ-visszakereső algoritmus használatával. [1][21][23]

Bebizonyítottam, hogy az általam használt információ-visszakereső algoritmus alkalmas arra, hogy egy Wikipédia méretű adatbázisból kiszűrje azokat a mondatokat, amelyek a legnagyobb valószínűséggel hasonlítanak a keresett mondathoz. A kísérlet eredménye géppel fordított tesztkorpuszra: a magyar mondatok esetében a jó találat 44% eséllyel az első találat, 13%-kal a 2. és a 10. helyezett között van, és csak 6% annak az esélye, hogy a 11. és 50. között található, valamint 37%, hogy nincs az első 50 között (lásd. 6.1. ábra).



6.1. ábra:Az indexált keresés által visszaadott jó találatok helyezése (angol-magyar)

Megmutattam, hogy a fedés mértéke az információ-visszakereső algoritmus miatt erősen függ a mondat hosszától: míg a rövid mondatokat alacsony fedési érték jellemzi, addig a hosszabb, tartalmasabb mondatokat nagyobb valószínűséggel adja vissza (lásd. 6.2. ábra).



6.2. ábra: A hasonlósági metrika és a teljes rendszer fedés értéke a mondat hosszának függvényében (angol-német és angol-magyar nyelvpárokra)

Az információ-visszakereső algoritmus segítségével az algoritmus adatbázis méretéhez viszonyított lineáris futási idejét tudtam konstansra csökkenteni, ami lehetővé teszi az algoritmus gyakorlatban történő alkalmazhatóságát.

7. tézis

Bebizonyítottam, hogy a fordítási plágiumok megtalálására képes új, hasonlósági metrikán alapuló algoritmusom nemcsak fordítások, hanem azonos nyelven írt szövegek összehasonlítására is alkalmas. Egy kétszeres fordításon átesett szöveg esetén 92% illetve 83% volt a mondat szintű fedés. A mondaton belüli szórendre ez az algoritmus teljesen érzéketlen, ellentétben az n-gramon alapuló algoritmussal, ahol a szavak sorrendjének a változása a fedés csökkenését eredményezi. A találatok sorrendezése a mondatok száma és a találatok értéke szerint a 12 Wikipédia-cikket tartalmazó tesztkorpuszon a 11 jó találatot az első 15 hely közé sorolta be, a hamis pozitív találatok egy kivétellel mind a lista végén szerepeltek.

Az új, fordítási plágiumok megkeresésére használt algoritmust akár egynyelvű szövegek összehasonlítására is alkalmazhatjuk. Ebben az esetben a szótári azonosság helyett szinonima-, antonima-, hiponima- és hipernima-azonosságokat vezetünk be, és ezek alapján értékelhetjük két szöveg azonosságát: két szózsák metszetét illetve különbségét a hasonlósági metrikában. Az algoritmus minden más eleme változatlan maradt, csupán a kétnyelvű szótárat cseréltem le az angol WordNetre. A plagizálást legegyszerűbben a szavak szinonimákra cserélésével lehet

szimulálni, ugyanakkor ez az algoritmus felépítéséből adódóan túl könnyű feladat lett volna, így a plagizálást úgy szimuláltam, hogy a kézzel, fordító által angolról magyarra lefordított 12 Wikipédia-cikket tartalmazó korpuszt két különböző gépi fordítóval is visszafordítottam angolra. Ezen a két, komoly változtatásokat, hibákat tartalmazó korpuszon teszteltem az algoritmust. Az egyik gépi fordítóval fordított szöveg esetén a 12 szócikkből 11-et megtalált és az első 15 helyen szerepel a jó találat, a másik fordítóval is nagyon hasonló eredményt értem el, a 10 helyes találat az első 17 találat között van.

Az eredmények gyakorlati alkalmazása

Az átlapolódó szavas darabolás a gyakorlatban is alkalmazásra került az SZTAKI KOPI Plágiumkeresőben, annak 2004-es indulása óta ez az algoritmus biztosítja az egynyelvű keresés alapját. Az adatbázisba a dokumentumok szavas darabolással kerülnek be és az összehasonlítandó dokumentumokat átlapolódó szavas darabolással dolgozza fel a rendszer, így a félig átlapolódó szavas darabolás segítségével egy kisebb adatbázist tudtunk létrehozni.

A fordítási plágiumok keresésére irányuló kutatásnak az volt a célja, hogy kiderítsem, lehetséges-e, és ha igen, milyen hatásfokkal, angol és magyar nyelvek között fordítási plágiumokat felismerni. Mivel az eredmények nagyon bíztatóak, és a hasonlósági metrikán és információ-visszakereső algoritmuson alapuló új algoritmusom a gyakorlatban is használhatónak bizonyult, lehetővé vált, hogy ez az algoritmus is beépítésre kerüljön a KOPI Plágiumkeresőbe. 2011. év végén a világon elsőként nyújtott fordítási plágiumkereső szolgáltatást a KOPI Portál.

A KOPI Portálnak elsődleges célja, hogy visszaszorítsa a plagizálást a felsőoktatásban. Ezt úgy éri el, hogy egy jól működő eszközt ad az oktatók kezébe, ami kockázatossá teszi a plagizálást. A KOPI rendszer működésének ismeretében belátható, hogy sokkal több energiabefektetés a másolás nyomait eltűntetni, mint becsülettel megírni az adott házi feladatot, diplomadolgozatot.

Köszönetnyilvánítás

Isaac Newton szavaival élve „If I have seen further, it is by standing on the shoulders of giants.”, aki amikor ezt mondta, éppen Bernard of Chartres vállán állt.

Köszönöm szépen szüleimnek a támogatást, és hogy olyan magasra tették a lécet. Feleségemnek és Édesapámnak a folyamatos és fáradhatatlan ösztönzést.

Köszönöm Monostori Krisztiánnak, hogy több mint 10 éve bevezetett a plágiumkeresés témakörébe; Hodász Gábornak a közös munkát, az első lépéseket ezen a területen; Arkady Zaslavskynak az ausztrál ösztöndíjat, ahol még jobban beleáshattam magam ebbe a témába.

Köszönöm szépen Kovács Lászlónak, hogy meghívott, hogy dolgozzak a SZTAKI-ban; Micsik Andrásnak a szakmai támogatást; Tóth Zoltánnak a KOPI rendszer elkészítésében, Vajna Miklósnak a nyelvfelismerő algoritmus tesztelésében és implementálásában, Pataki Balázsnak pedig a fordításiplágium-kereső algoritmus implementációja során nyújtott segítséget; Pallinger Péternek a rendszergazdai támogatását a kísérletekhez; Zsivnovszki Magdolnának és Virág Évának az angol cikkek angolosítását és a magyar cikkek magyarosítását; Inzelt Péternek a belső pályázati lehetőséget, aminek a segítségével be tudtam fejezni a kutatásomat.

Köszönöm szépen Prószéky Gábornak támogatását a Pázmányon töltött két év alatt és a részletes, mindenre kiterjedő lektorálásokat.

A szerző publikációi

Folyóiratcikk

[1] **Máté Pataki** and Attila Csaba Marosi, "Searching for Translated Plagiarism with the Help of Desktop Grids", *Journal of Grid Computing*, Volume 11, Issue 1, pp 149-166, Springer, ISSN: 1570-7873, 2013.

[2] **Pataki Máté**, „Digitális könyvtárak védelme a KOPI plágiumkereső rendszerrel", *Tudományos és Műszaki Tájékoztatás* 54/3., 2007.

[3] Kovács László és **Pataki Máté**, „E-ügyintézés bevezetése Kaposvárott", *Jegyző és közigazgatás 8/1.*, ISSN: 1589-3383, 2006.

[4] **Pataki Máté**, „W3C ajánlások magyarul", *Tudományos és Műszaki Tájékoztatás 52/9.*, 2005.

Könyv

[5] **Pataki Máté** és Abonyi-Tóth Andor, Szerkesztő: **Pataki Máté** „Bevezetés az info-kommunikációs akadálymentesítés világába II.", ISBN: 978-615-5043-62-8, 2011.

[6] Abonyi-Tóth Andor, **Pataki Máté** és Mátételki Péter, Szerkesztő: Abonyi-Tóth Andor, „Bevezetés az info-kommunikációs akadálymentesítés világába I.", ISBN: 978-615-5043-18-5, 2011.

[7] **Pataki Máté**, „Infokommunikációs akadályok", ISBN: 978-615-5043-66-6, 2010.

Könyvfejezet

[8] Jókai Erika, Koloszár Kata, Mogánné Tölgyesy Szilvia és **Pataki Máté**, „Rehabilitációs támogató technológiák", ISBN: 978-963-2790-97-8, 2010.

[9] Deákné Orosz Zsuzsa, Dr. Kecskeméti Éva, Zalabai Péterné, Abonyi-Tóth Andor, Fehérné Kovács Zsuzsa, Helfenbein Henrik és **Pataki Máté** „Fogyatékos személyek szociális segítése – Szociális ellátás", ISBN: 987-615-5043-08-6, 2009.

[10] Dr. Kecskeméti Éva, dr. Nagy Janka Teodóra, Abonyi-Tóth Andor, Fehérné Kovács Zsuzsa, Földiné Angyalossy Zsuzsa, Helfenbein Henrik, dr. Márkus Eszter, **Pataki Máté**, dr. Perlusz Andrea és dr. Szabó Ákosné, „Esélyegyenlőség a joggyakorlatban - Szociális jogi szabályozás", ISBN: 987-615-5043-10-9, 2009.

Külföldi konferenciakötet

[11] **Máté Pataki**, „A new approach for searching translated plagiarism”, *5th International Plagiarism Conference*, Newcastle-Upon-Tyne, 2012.

[12] Hannes Eichner, András Micsik, **Máté Pataki** and Robert Woitsch, „A use case of service-based knowledge management for software development”, *IFIP international conference on research and practical issues of enterprise information systems*, Győr, 2009.

[13] László Kovács and **Máté Pataki**, „Copy Protection via Plagiarism Search”, *3rd International Plagiarism Conference*, Newcastle-Upon-Tyne, 2008.

[14] László Kovács, Zoltán Szentirmay, **Máté Pataki** and Péter Pallinger, „Development of the new National Cancer Registry”, *microCAD 2007, International Scientific Conference*, ISBN: 978-963-661-759-2, Miskolc, 2007.

[15] László Kovács and **Máté Pataki**, „KOPI protection instead of copy protection”, *Axmedis 2006,. 2nd International Conference on Automated Production of Cross Media Content for Multi-channel Distribution*, ISBN: 88-8453-526-3, Leeds, 2006.

[16] Roland Alton-Scheidl, András Micsik, **Máté Pataki**, Wolfgang Reutz, Jürgen Schmidt and Thomas Thurner, „StreamOnTheFly: a Peer-to-peer network for radio stations and podCasters”, *Proceedings of the First International Conference on Automated Production of Cross Media Content for Multi-channel Distribution, AXMEDIS'05*, Florence, 2005.

[17] László Kovács, András Micsik, **Máté Pataki** and Robert Stachel, „StreamOnTheFly: a network for radio content dissemination”, *Lecture Notes in Computer Science 3664*, 2005.

[18] **Máté Pataki**, „Plagiarism detection and document chunking methods”, *Proceedings of the Twelfth International Conference on World Wide Web, WWW2003*, Budapest, 2003.

[19] Krisztián Monostori, Raphael Finkel, Arkady Zaslavsky, Gábor Hodász and **Máté Pataki**, „Comparison of Overlap Detection Techniques”, *The 2002 International Conference on Computational Science, Lecture Notes in Computer Science 2329*, Amsterdam, 2002.

Hazai konferenciakötet

[20] **Pataki Máté**, Pataki Balázs, Tóth Zoltán, Pallinger Péter, Kovács László, „DRM megoldások áttekintése”, *Networkshop*, Sopron, 2013.

[21] Micsik András, **Pataki Máté**, Garzó András, „A KOPI Plágiumkereső terhelésének elosztása cloud környezetben”, *Networkshop*, Sopron, 2013.

[22] **Pataki Máté**, „Algoritmus fordítások keresésére”, *BJMT Alkalmazott Matematikai Konferencia*, Győr, 2012.

[23] **Pataki Máté**, „Fordítási plágiumok keresése”, *MSZNY 2011. VIII. Magyar Számítógépes Nyelvészeti Konferencia*, ISBN: 978-963-306-121-3, Szeged, 2011.

[24] **Pataki Máté** és Vajna Miklós, „Többnyelvű dokumentum nyelvének megállapítása”, *MSZNY 2011. VIII. Magyar Számítógépes Nyelvészeti Konferencia*, ISBN: 978-963-306-121-3, Szeged, 2011.

[25] **Pataki Máté**, „Plágiumkeresés különböző nyelvek között”, *Networkshop*, Kaposvár, 2011.

[26] **Pataki Máté**, Abonyi-Tóth Andor és Helfenbein Henrik, „A "Bevezetés az esélyegyenlőséget szolgáló info-kommunikációs technológiákba" kurzus tapasztalatai az ELTE Informatikai Karán”, *III. Oktatás-Informatikai konferencia*, Tanulmánykötet, ISBN: 978-963-312-037-8, Budapest, 2011.

[27] **Pataki Máté**, „Plagizálás a felsőoktatásban”, *Magyar Tudomány Napja: Kövek, szavak, gondolatok – A kultúrák találkozása*, Szombathely, 2010.

[28] **Pataki Máté**, „Webes Akadálymentesítési Útmutató 2.0 - W3C WCAG 2.0”, *Akadálymentes web-tervezés workshop*, Veszprém, 2010.

[29] **Pataki Máté**, „Rámpát a honlapokra – úton az akadálymentes honlapok felé”, *Networkshop*, Szeged, 2009.

[30] **Pataki Máté** és Micsik András, „Üzleti modellen alapuló webes tudásprezentáció”, *Networkshop*, Szeged, 2009.

[31] **Pataki Máté**, Füzessy Tamás, Kovács László és Tóth, Zoltán, „Hibatűrő keresés digitalizált magyar nyelvű szövegekben”, *Networkshop*, Dunaújváros, 2008.

[32] **Pataki Máté**, Richter Viktor, „A W3C szabványosítási törekvései”, *Networkshop*, Dunaújváros, 2008.

[33] **Máté Pataki** and Tamás Füzessy, „Digitization errors in Hungarian documents”, *AACS '07 Automation and Applied Computer Science Workshop*, ISBN: 978-963-420-909-6, Budapest, 2007.

[34] **Pataki Máté**, Kovács László és Pataki Balázs, „Egy országos méretű orvosi adatbázissal kapcsolatos informatikai kihívások”, *Networkshop*, Eger, 2007.

[35] **Pataki Máté** és Tóth Zoltán, „Szkennelt szövegek digitalizálása során keletkező hibák elemzése magyar szövegek esetében”, *Networkshop*, Eger, 2007.

[36] **Pataki Máté**, „A W3C és a Mobilweb”, *Magyarországi Web Konferencia*, Budapest, 2007.

[37] **Pataki Máté**, „KOPI Plágiumkereső a digitális tartalmak védelmében”, *DAT 2006 - A digitális kreatív iparágak szerepe Magyarországon*, Budapest, 2006.

[38] **Máté Pataki**, „Distributed similarity and plagiarism search”, *AACS 2006, Proceedings of the Automation and Applied Computer Science Workshop*, ISBN: 963-420-865-7, Budapest, 2006.

[39] **Máté Pataki**, „Plagiarism search within one document”, *AACS 2006, Proceedings of the Automation and Applied Computer Science Workshop*, ISBN: 963-420-865-7, Budapest, 2006.

[40] **Pataki Máté**, „W3C WAI, avagy weblapok akadálymentesítése”, *Magyarországi Web Konferencia 2006*, Budapest, 2006.

[41] **Pataki Máté** és Kovács László, „W3C WAI - weblapok akadálymentesítése”, *Networkshop*, Szeged, 2005.

[42] **Pataki Máté**, „KOPI online plágiumkereső és információs portál”, *Networkshop*, Győr, 2004.

[43] Kézdi Tamás, Kovács László, Micsik András és **Pataki Máté**, „Elosztott digitális hangtárak a közösségi rádiózásért (SotF)”, *Networkshop*, Pécs, 2003.

Egyéb publikációk

[44] **Pataki Máté**, „Szöveges dokumentumok darabolása és tömörítése hash-kódolással - darabolási technikák és másolatkeresés”, *Budapesti Műszaki és Gazdaságtudományi Egyetem, diplomadolgozat*, Budapest, 2002.

[45] Hodász Gábor, **Pataki Máté**, „Szöveges dokumentumok darabolása és tömörítése”, *Budapesti Műszaki és Gazdaságtudományi Egyetem, TDK dolgozat*, Budapest, 2001.