

A Wikipédia felhasználása az absztrakt címkézési feladatban

Berend Gábor¹, Farkas Richárd²

¹ Szegedi Tudományegyetem Informatikai Tanszékcsoport,
6720 Szeged, Árpád tér 2.
berendg@inf.u-szeged.hu

² MTA – SZTE Mesterséges Intelligencia Kutatócsoport,
6720 Szeged, Tisza Lajos krt. 103. III. lépcsőház
rfarkas@inf.u-szeged.hu

Kivonat: Az elektronikus, azon belül is az online tartalmak méretének robbanása újszerű megközelítést tesz szükségessé kategorizálásukra. Egy ilyen újszerű és elterjedt módszer az ún. címkézés, amely során dokumentumainkat azokat tömören és jól leíró kulcskifejezésekkel látjuk el. Ezek egy része egzaktul a szövegben is megtalálható, de kulcskifejezések lehetnek absztrakt címkék is, amik a dokumentumban nem fordulnak elő, mégis szemantikus kapcsolatba hozhatók a leírtakkal. Az [origo] hírportál archívumának automatikus felcímkézése során egyik részfeladatunknak a cikkekhez való absztrakt címkék hozzárendelését tekintettük, melyhez napjaink legnagyobb egységes formátumú, szabadon hozzáférhető tudásbázisát, a Wikipédiát használtuk föl.

1 Bevezetés

Az online tartalmak mennyiségének rohamos növekedésével egyre nehezebbé válik azok használata, katalogizálása. [4] szerint a 2007-ben 281 exabájtosra (281 milliárd gigabájtosra) becsült digitális univerzum mérete 2010-re várhatóan eléri az 1 zettabájtos határt, így nem is lehet kérdéses, hogy újszerű megközelítések szükségesek az online adatok rendszerezésére. Noha az egyszerű szöveges dokumentumok teljes digitális univerzumbeli részesedése csökkenő tendenciát mutat a multimédiás tartalmak térhódításának köszönhetően, fontosságukról így sem szabad megfeledkeznünk, hiszen mennyiségük így is változatlanul exponenciálisan nő. Ezt a növekedést támasztja alá [5] is, mely szerint a blogszféra mérete 5 havonta megduplázódik, naponta pedig átlagosan 30-40 ezer új blog kerül létrehozásra.

Éppen ezért a tartalmak kategorizálásának megkönnyítésére és a szövegekben történő könnyebb navigálás, keresés érdekében az utóbbi években – eleinte éppen a blogokon – bevezették az ún. címkézési (tagging) eljárást. Ezen Web2.0-ás eljárás során minden dokumentum szerzője az általa leírt tartalmat legtömörebben összegezni képes, néhány elemből álló kifejezéshalmazzal látja el írásait, amely alapján aztán könnyebben találhatjuk meg a minket érdeklő információkat. A módszer eredményességének láttán az eljárást időközben szinte minden tartalomszolgáltató bevezette, így a hírportálok is, mint például az [origo], amely szerkesztői 2009 eleje óta friss cikkei-

ket a bennük leírtakat legjobban megragadó kulcsszavakkal látják el. Egy ilyen megoldás hasznos szolgálatot nyújt mind a keresőoptimalizálás, mind pedig a weboldalon megjelenő hirdetések egyes célcsoportokhoz való eljuttatása terén is.

A címkézés automatizálására – felhasználói megerősítés mellett – több megoldási kísérlet [6, 9, 12] született a korábbiakban, hiszen segítségükkel kiküszöbölhető lenne a korábban föl nem címkézett, nagy mennyiségű adathalmazok emberi erővel történő fölcímkézése mindamellett, hogy ezzel az egyes, tipikusan emberi címkézésre jellemző hibáktól [12] is mentesíteni lehetne a jelölést. A korábbi megoldások jellemzően kézi címkékkel ellátott dokumentumok alapján ajánlottak címkejelölteket a címkézetlen dokumentumoknak.

A dokumentumokhoz elvárhatóan rendelendő címkék egy része a szövegben is fellelhető – még ha esetleg nem is egységes formátumban (pl. a rövidítések vagy éppen toldalékolás miatt), vagy csupán implicit módon (*foci – labdarúgás*) –, más részük egyáltalán nem: hiszen például egy motorsportról szóló cikk esetében nem feltétlenül kell szerepeljen maga a *motorsport* kifejezés is a szövegben. Utóbbi kifejezéseket absztrakt címkéknek nevezzük. Az absztrakt címkék esetenként alkalmasabbnak bizonyulnak nem absztrakt társaikhoz képest, hiszen jóval informatívabbnak találjuk egy adalékanyagokkal foglalkozó dokumentum esetében az *élelmiszer-adalékanyagok* címke használatát (még ha az konkrétan nem is került megemlítésre a dokumentumban), mint a ténylegesen megemlített adalékanyagok listáját (pl. *tartrazin, gellángumi, nátrium-tartarát, csontfoszfát*).

Az előzőekben leírt okok miatt cikkünk az ilyen, ún. absztrakt címkék problémájára ad megoldási javaslatot, felhasználva napjaink legnagyobb egységes formátumban fellelhető, szabadon felhasználható elektronikus tudásbázisát, a Wikipédiát. Eljárásunkkal, amely a cikkekben előforduló releváns kifejezések Wikipédia-szócikkeire támaszkodik, tovább javítható a címkézés minősége: a fedésen, valamint a pontosságon túl a címkefelhő kohéziója egyaránt.

Munkánk során a cikkek szövegeiben előforduló potenciális címkék Wikipédia-szócikkeinek tartalmát éppúgy fölhasználtuk, mint a szócikkek közt hiperlinkek formájában megtestesülő kvázi-szemantikus viszonyokat. Az egyes szócikkekkel gyakran együtt előforduló egyéb fogalmak (szócikkek), valamint az egyes oldalakra mutató és belőlük kifelé irányuló relációk (linkek) vizsgálata éppúgy hasznosnak bizonyult, akár csak a szócikkek közötti átirányítások (redirect) figyelembevétele.

2 Kapcsolódó munkák

A számítógépes nyelvészeti munkák közül leginkább az automatikus címkézéssel, valamint a tervek közötti szemantikus relációk Wikipédia segítségével történő automatikus föltérképezésével foglalkozó irodalomra támaszkodtunk.

2.1 Automatikus címkézés

Az eddigi automatikus címkézésről szóló munkák két fő irányvonalba sorolhatók. Az egyik megoldási módozat, az ún. címke- vagy kulcsszókinyerés (*tag / keyphrase*

extraction) során a fölcímkezendő cikkek szövegéből nyerik ki a címkejelölteket, akárcsak [3]-ban. Egy hátulütője az efféle kulcsszókinyerő rendszereknek, hogy ezek csak a dokumentumokban ténylegesen is előforduló címkék szövegből történő kiemelésére alkalmasak.

Absztrakt címkézési megközelítésünkhöz legközelebb álló megoldások a [9]-hez hasonló, ún. címke-hozzárendelő (*tag assignment*) rendszerek. Ezek a megoldások a fölcímkezendő dokumentumokhoz hasonló, kézi jelöléssel már ellátott dokumentumok címkéinek hozzárendelésével oldják meg a címkézési feladatot, így ezek a megoldások is absztrakt címkézésként foghatók föl, ugyanis egy dokumentumhoz olyan címkék is hozzárendelhetők, melyek annak szövegében nem fordulnak elő. Az ilyen módszerek hátránya azon túl, hogy a hozzárendelt címkék megőrzik az emberi címkézés esetlegességeit, hogy a dokumentumokhoz rendelt címkék egy zárt halmazból kerülhetnek csupán ki, vagyis a tárgyalt témákban az időben végbe menő változásokat nem tudják naprakész, friss címkékkel követni. Ezzel szemben az általunk javasolt rendszernek nincs szüksége kézi címkékkel ellátott dokumentumokra, az absztrakt címkék meghatározása során pedig a hasonló dokumentumok keresésén túlmutató, szemantikus kapcsolódó címkéket javasol.

2.2 Szemantikus viszonyok vizsgálata

Az automatikus címkézés során hasznos, ha képesek vagyunk meghatározni kifejezések között fennálló szemantikus viszonyokat: segítségével ki lehet szűrni egy dokumentum kulcsszójelöltjei közül azokat, melyek nem koherensek a többivel, vagy épp ellenkezőleg, a jelöltek közötti kohézió megtartása mellett újakkal lehet kiegészíteni azokat. A szemantikus relációk vizsgálata során az utóbbi években többen is a legnagyobb, részben strukturált online tudásbázist, a Wikipédiát használták föl szemben a korábbi megközelítésekkel [10], amelyek ontológiákra vagy különféle korpuszokon mért kifejezések együttes előfordulásának kiszámítására támaszkodtak.

[11] a szövegekben előforduló többértelmű tulajdonnevek (pl. *Kennedy (repülőtér)* – *Kennedy (személy)*) egyértelműsítésére használta föl a Wikipédiát. [1, 7] egyaránt termék között fennálló szemantikus viszony erősségét meghatározó rendszert mutatnak be, melyek a szócikkek által kifeszített vektortérben vett hasonlósági mértékek alapján hoznak döntést.

Munkánkhöz legközelebb az előbbi munkákra is támaszkodó [6] áll, mely egy dokumentum szavaihoz egyértelműsítés után rendelt Wikipédia-szócikkek közül gráfanalízist használva választja ki azokat, amelyek leginkább képesek lehetnek az eredeti dokumentum tartalmának megragadására.

3 Módszerek

Absztrakt címkéző eljárásunk az egyes cikkek szövegeiből kinyert, abban egzaktul előforduló kifejezések halmazát várja bemenetül, majd ezekhez rendeli hozzá a velük vélhetően szemantikus relációban álló Wikipédia-szócikkek halmazát. A bementként szolgáló címkejelölteket a cikkekből a [2]-ben leírtak szerint nyertük ki. Ezután a

szövegből kinyert címkeaspiránsokhoz meghatároztuk azon Wikipédia-szócikket, amelyek egy az egyben megfeleltethetők a címkejelöltek halmazának legalább egy elemével. Olyan szócikkek esetében, amelyek egyértelműsítő lappal rendelkeztek, nem választottuk ki a szócikk egyik egyértelműsítő lapját sem, elkerülendő ez által esetleges rossz választásokból adódó zajt a továbbiak során.

Az absztrakt címkék megtalálására alkalmazott módszereink egyaránt támaszkodnak a hírportál cikkeiből kinyert címkejelöltek Wikipédia-szócikkeinek szöveges tartalmára, valamint a közöttük meglévő gazdag linkstruktúrára. A következő fejezetek ezeket az eljárásokat mutatják be részletesen.

3.1 Átirányítások figyelembevétele

A Wikipédia felépítéséből adódóan azonos tartalmak több szócikk alól is elérhetők. Így például akár az *USA*, akár pedig az *Amerikai Egyesült Államok* szócikkekre keressünk rá, egyazon oldalt kapjuk találatul. Ezen ún. átirányító (*redirect*) Wikipédia-oldalak szinonimák, illetve asszociációk meghatározására, rövidítések feloldásai valamint korlátozott mértékig elíráskezelésre egyaránt alkalmazhatók (például 1. táblázat). Segítségükkel kanonikus alakra tudunk hozni eltérő formában előforduló, de azonos jelentéssel bíró címkejelölteket, amivel a teljes címkézés kohézióját javíthatjuk (mivel azonos jelentésű címkék nem fordulnak elő több formában, mint nyereség – profit).

1. táblázat: A Wikipédiában szereplő Amerikai Egyesült Államok szócikkre irányuló átirányítások listája.

Amerikai	Amerikai Egyesült Államok
Amerikaiak	Amerikai Egyesült Államok
Amerikai egyesült államok	Amerikai Egyesült Államok
Egyesült államok	Amerikai Egyesült Államok
Egyesült Államok	Amerikai Egyesült Államok
United States	Amerikai Egyesült Államok
United States of America	Amerikai Egyesült Államok
US	Amerikai Egyesült Államok
USA	Amerikai Egyesült Államok

Absztrakt címkéző módszerünk a címkeaspiránsokhoz rendelt Wikipédia-szócikkek közül lecseréltük mindazokat, amelyek más szócikkre voltak irányítva. Ezen a ponton az automatikus címkézés eredményeképp előálló címkefelhő kohézió növelése volt a cél, mivel így elkerülhető volt az eltérő alakban álló, de ugyanazzal a szemantikus jelentéssel bíró címkék alkalmazása.

3.2 Definíciók kinyerése

Ebben a lépésben a Wikipédia oldalnak megfeleltethető címkejelöltekhez rendeltünk definíciókat, amelyek aggregálása után újabb címkejelöltet voltunk képesek javasolni

a már meglévők mellé. Az ilyen módon nyert definíciók jól megragadják az egyes szócikkekben leírt fogalmak hiponim relációit: a *krizoin*-ról például megállapítható, hogy az egy *adalékanyag*.

Megfigyelhető, hogy a Wikipédia enciklopédikus jellegéből adódóan az egyes oldalak elején megtalálható a bennük tárgyalt fogalom definiálása. Úgy jártunk el, hogy minden egyes címkejelölthöz meghatároztuk annak Wikipédiáról automatikusan kinyert definícióját, és amennyiben egy definíció címkejelöltek egy adott halmazán több esetben is alkalmasnak bizonyult, úgy azt absztrakt címkeként javasoltuk.

Egy szócikk által leírt fogalom potenciális definícióinak kinyeréséhez elsőként meg kellett határozni azt a mondatot, amelyből az kinyerhető lehet. Megközelítésünkben ez a mondat minden esetben az volt, amelyik elsőként megemlítette a szócikket magát, vagy amennyiben nem szerepelt ilyen az egész oldalon, úgy a szócikk első bekezdésének első mondatát tekintettük ilyennek. Az ily módon kinyert szócikk-mondat megfeleltetésekre példákat a 2. táblázat hoz.

2. táblázat: Wikipédia-szócikkekből kinyert definíciót tartalmazó mondatok.

Erdős Pál	Erdős Pál , a 20. század egyik legkiemelkedőbb <i>matematikusa</i> , az <i>MTA tagja</i> .
Gottlob Frege	Friedrich Ludwig Gottlob Frege, <i>német matematikus, logikátudós, filozófus</i> , a modern matematikai logika és analitikus filozófia megalapítója, művelője.
Maffiózók	A Maffiózók egy <i>amerikai TV-sorozat</i> , amelynek David Chase a kitalálója és producere.

Az előzőek szerint generált potenciálisan definíciót tartalmazó mondatokból következő lépésként magukat a lehetséges definíciókat nyertük ki. Ezen lépés során a mondaton belüli szöveggörnyezetet figyelembe véve, továbbá morfológiai és szintaktikai megfontolásokat alkalmazva határoztuk meg az adott szócikkhez tartozó definíciókat, melyeknek vagy önmaguknak is vagy pedig tagonként önálló Wikipédia-szócikk-címeknek kellett lenniük. (Így lett alkalmas definíció az *amerikai TV-sorozat*, ahol az *amerikai* és a *TV-sorozat* külön szócikként szerepel a Wikipédiában.) A leírtak alapján nyert szócikk-definíció párosokra a 3. táblázatban láthatók példák.

3. táblázat: Példa definíciógenerálásra.

Erdős Pál	<i>matematika</i>
Gottlob Frege	<i>matematika, német, filozófia</i>
Maffiózók	<i>producer, amerikai TV-sorozat, TV-sorozat</i>

Átfedő definíciójelöltek esetén (pl. *amerikai, TV-sorozat* és *amerikai TV-sorozat*) a leghosszabb szupersztringet választottuk (*amerikai TV-sorozat*). Végül egy dokumentum címkejelöltjeihez akkor rendeltünk hozzá definíciókat is absztrakt címkeként, ha az több címkejelölt esetében is relevánsnak lett minősítve, vagyis például egy olyan esetben, ahol egy dokumentum címkejelöltjei között szerepelt *Erdős Pál* és *Gottlob Frege* is, ott fölveztük a *matematika* szót is mint címkejelöltet, hiszen az mindkettő esetében értelmes definíciónak lett titulálva.

3.3 A linkstruktúra kiaknázása

Adott dokumentumból kinyert címkejelöltekhez rendelhető absztrakt fogalmakat a Wikipédia linkstruktúrája szempontjából is vizsgáltuk: megkerestük azokat a további szócikkeket, amelyek jellemzően együtt fordulnak elő egy potenciális címkéhez rendelt szócikkkel, vizsgáltuk azokat a szócikkeket, amelyekre egy hírdokumentumhoz rendelt szócikkek közül több is hivatkozott, illetve megkerestük azokat a szócikkeket, amely egy dokumentum címkejelöltjeihez generált szócikkek halmazát a leginformatívabban tartalmazzák.

Együtt-előfordulás vizsgálata

Ebben az esetben minden egyes címkejelölthöz, melyhez hozzárendeltünk Wikipédia-szócikket, megkerestük azon egyéb szócikkeket, amellyel együtt az gyakran előfordul. A vizsgálat elvégzését csak olyan szócikkek esetében végeztük el, amely legalább 10 és legfeljebb 150 oldalon lett hivatkozva. Ennek oka az volt, hogy a 10 esetenél kevesebbet hivatkozott szócikkek nem tűntek eléggé relevánsnak, a 150-nél többször előfordulók pedig túl általános gyűjtőoldalnak bizonyultak.

Az olyan szócikkekre, amelyekre a hivatkozások száma az előbb említett két korlát között volt megkerestük azokat a szócikkeket, amelyek legalább az esetek felében ugyanúgy megfigyelhetők voltak a hivatkozó oldalakon linkek formájában. Így például, mivel *Sébastien Loeb* raliversenyző *rali-világbajnokság* szócikkkel való együttes előfordulása 0.7073 volt, a Sébastien Loeb nevet tartalmazó cikkhez a rali-világbajnokság címke is fölvételre került.

A kimenő linkek vizsgálata

A kimenő linkek esetében azokat a szócikkeket kerestük, amelyek relevánsnak tekinthetők szócikkek egy adott halmazára nézve. Ehhez vettük a bemeneti szócikk-halmaz egyes elemeiből kifelé irányuló megbízható linkekhez tartozó szócikkeket. Megbízhatónak tituláltunk egy linket, ha az általa hivatkozott oldal tartalmazott visszaértelt a hivatkozó dokumentum irányába, vagy a hivatkozó oldal linkjeinek legalább 25%-át a másik oldalra való hivatkozás tette ki, és ezen linkek száma legalább 3 volt (kivéve a portál – és kategória gyűjtőoldalakra mutató linkeket, mivel azok a szerkesztési konvenciókból adódóan az oldalak alján egy példányban szerepelnek többnyire).

Az előbbieket szerint minden egyes Wikipédia-szócikkkel rendelkező címkejelölthöz az általuk hivatkozott szócikkek közül azokat tartottuk ténylegesen is relevánsnak a teljes hírcikkre nézve, melyekre nem csupán egy szócikkből mutatott relevánsnak titulált link. Például egy cikk esetében, amely címkejelöltjei között szerepelt a *BUX* és a *Budapesti Értéktőzsde* is, egyúttal implikálta a *Magyarország gazdasága* címke fölvételét is, mivel arra mindkét oldalhoz tartozó Wikipédia-szócikk referál.

Tartalmazások vizsgálata

Az eddigieken túl szemantikus kapcsolatok tárhatók föl szócikkek egy halmaza és egy további szócikk között, ha megvizsgáljuk, hogy egy potenciális absztrakt címkének megfeleltethető szócikk az inputként kapott szócikkhalmaz elemeit milyen mértékben tartalmazza.

A termhalmazok és az absztrakt címkejelöltként funkcionáló szócikkek közötti tartalmazás mértékének számszerűsítésére a tf-idf metrikát adaptáltuk. A bemenetként szolgáló címkeaspiráns-halmaz alapján meghatároztuk azokat a szócikkeket, amelyek legalább egyet is tartalmaznak közülük link formájában. Ezek után az összes szócikk előző feltételnek eleget tevő részhalmának minden elemére kiszámítottuk az adott bemeneti szócikk halmazra vett átlagos tf-idf értéküket, amely ha adott küszöbérték feletti volt, akkor absztrakt címkeként kezeltük a továbbiakban az adott szócikket.

4 Eredmények

Absztrakt címkézési eljárásunk kiértékelésére az [origo] hírportál dokumentumainak kézi címkézésének megkezdése óta keletkezett, január és február hónapokból választott 600-600 dokumentumát választottuk ki. A kiértékelést két annotátorra bíztuk, a 600-600 dokumentumból pedig 100 mindkét annotátor esetében azonos volt, így összesen 1100 különböző cikk került kiválasztásra. Az 1100 dokumentumból azonban csak 1073 esetben állt rendelkezésünkre az absztrakt címkéző eljárásunk inputjaként szolgáló, a cikkek szövegéből kinyert címkejelöltek halmaza, aminek az oka az, hogy az [origo] specifikációja alapján a film-blog csatornájukba tartozó dokumentumaik címkézését nem kellett elvégezzük (a kérdéses 27 dokumentum pedig ebbe a csatornába esett). Így legvégül 584, illetve 589 dokumentum automatikus absztrakt címkézésének kiértékelése történt meg.

Az annotátorok feladata az volt, hogy minden dokumentum esetében a Wikipédia 2009. szeptember 14-i tartalma és struktúrája alapján az egyes hírcikkekhez rendelt absztrakt címkékről döntsék el, hogy azok az adott cikk esetében elfogadhatók-e, valamint hogy határozzák meg, hogy az automatikusan generált absztrakt címkék megfeleltethetők-e a manuális címkézés egy vagy több cikkben ténylegesen elő nem forduló elemével. A végső pontosságot az alkalmasnak talált absztrakt címkézési eljárással nyert címkék arányának (pontosság) és a manuális címkékhez viszonyított fedés értékekének kombinált értékeiből számított F-mértékkel határoztuk meg.

A vizsgált dokumentumokhoz az [origo] munkatársai összesen 1192 alkalommal rendelték a szövegben elő nem forduló kifejezéseket címkeként, ami dokumentumonként átlagosan 1,11 absztrakt címkét jelent. Az 1192 alkalommal összesen 554 különböző absztrakt címkét használtak. Az annotálás során azt tapasztaltuk, hogy egyes esetekben a cikkek szövegben elő nem forduló címkéként használt termék szinonimája (pl. *gazdasági válság – recesszió*) már megtalálható volt, és ezt az absztrakt címkézést megelőző lépésekben eredményesen ki is nyertük. Más esetekben pedig csupán az absztrakt címke kézi hozzárendelése során történő elírások (pl. Sony Ericsson – Sony Ericsson) tettek absztrakttá (vagyis a cikk szövegében elő nem fordulóvá) egyes kifejezéseket, így az automatikus absztrakt címkék fedésének vizsgálata során az ezekkel való pontos egyezést nem követeltük meg. Ezen „kvázi-absztrakt” címkék figyelmen kívül hagyásával összesen 1114 ténylegesen is absztrakt címke található az 1073 dokumentumból álló teszhalmazon (dokumentumonként átlagosan 1,038), melyek dokumentumok szerinti eloszlását a 4. táblázat tartalmazza.

4. táblázat: Hírdokumentumok és a manuálisan meghatározott absztrakt címkék eloszlása.

Absztrakt címkék száma	Dokumentumok száma	Címkék mennyisége
0	339	0
1	465	465
2	184	368
3	65	195
4	18	72
5	1	5
9	1	9
Összesen	1073	1114

Az 1073 vizsgált dokumentum esetében összesen 13689 címkeaspiránst nyertünk ki az absztrakt címkézést megelőző lépésekben, amelyekhez 5239 esetben voltunk képesek Wikipédia-szócikket rendelni. Az egyedi címkeaspiránsok száma 6578 volt, közülük 1766-hoz (26,85%) határoztunk meg Wikipédia-szócikket, melyek segítségével 5014 alkalommal rendeltünk hozzá összesen 2028 különböző automatikus absztrakt címkét cikkekből kinyert címkeaspiránsok halmazaihoz. A dokumentumok eddigiek alapján vett eloszlásai az 5. táblázatban szerepelnek, melyből az is kitűnik, hogy 32 dokumentum egyetlen címkeaspiránsához sem tudtunk Wikipédia-szócikket kötni.

5. táblázat: Dokumentumok eloszlása a hozzájuk rendelt kezdeti címkeaspiránsok/ Wikipédia-szócikkek/ absztrakt címkék száma szerint.

	Dokumentumok száma n darab		
	szövegből származó címkeaspiránssal	Wikipédia-szócikk-hozzárendeléssel	automatikus absztrakt címkével
n = 0	0	32	157
0 <n <=5	72	669	639
5 <n <=10	388	320	174
10 <n <=20	509	51	73
n > 20	104	1	30
Összesen	1073	1073	1073

Az 5014 absztrakt címke 5733 címke-hozzárendelésnek volt köszönhető, mely azal magyarázható, hogy bizonyos absztraktcímke-jelöléseket egyszerre több módszer is javasolt, az egyes módszerek közötti eloszlás pedig a 6. táblázatban látható.

6. táblázat: Az absztrakt címkézõ eljárások közötti eloszlás.

Módszerek	Címke-hozzárendelések száma
Átírányítás	1155 darab (20.146%)
Definíciók	1471 darab (25.658%)
Együttes előfordulás	1998 darab (34.676%)
Kimenő linkek	558 darab (9.733%)
Tartalmazó szócikkek	551 darab (9.611%)
Összesen	5733 darab (100%)

Mind az 5733 hozzárendelést külön módszerenként vizsgálva, a pontosság értékére a 7. táblázatban lévő adatokat kaptuk.

7. táblázat: Az egyes módszerek által bevont absztrakt címkék pontossága.

Módszerek	Címke-hozzárendelések száma	Elfogadott hozzárendelések	Pontosság
Átírányítás	1155	836	0.7238
Definíciók	1471	414	0.2814
Együttes előfordulás	1998	697	0.3488
Kimenő linkek	558	227	0.4068
Tartalmazó szócikkek	551	90	0.1633
Összesen	5733	2264	0.3949

Az absztrakt címkézés kiértékelésének végső eredményét a két annotátor döntései alapján a 8. táblázat tartalmazza.

8. táblázat: A kézi kiértékelés végső eredménye.

	Pontosság	Fedés	F-mérték
1. annotátor	0.3933	0.1057	0.1666
2. annotátor	0.3848	0.1077	0.1683
Összesítve	0.3891	0.1067	0.1675

5 Konklúzió

Módszerünket az [origo] hírportál címkézetlen archívumán teszteltük, a Wikipédia segítségével bevont absztrakt címkék fölvételével pedig sikerült javítanunk a legvégül előálló címkefelhő minőségén.

Az eredmények figyelembevételénél fontos szem előtt tartani, hogy az automatikus absztrakt címkézés fedésének értéke a cikkekhez ténylegesen hozzárendelt címkékhez lett mérve, ami pedig olyan fogalmakat is tartalmazott, amelyekre a magyar Wikipédiában egyáltalán nem létezik szócikk (pl. *gyárbezárás*), vagy pedig helyességük megkérdőjelezhető ("*Hearts, FTC*" vagy a "*fogászat, árak*" [mindkettő egybe, egy címkéként]). Az ilyen címkék Wikipédia fölhasználásával történő cikkekhez rendelése pedig nemcsak, hogy nem lehetséges, de esetenként nem is lenne célszerű.

Módszerünkre jellemző, hogy eredményessége függ a bemenetként kapott címkeaspiránsok halmazától, így fontos, hogy azok minősége megfelelő legyen. Ezen túl, ahogy az az 5. táblázatban is látható, 32 dokumentum esetében egyáltalán nem tudunk Wikipédia-szócikket társítani a bemenetként kapott címkejelöltekhez, így ezekben az esetekben nem is volt lehetőség absztrakt címkék bevonására (a legtöbb módszer ugyanis legalább kettő, a cikk szövegéhez kapcsolódó szócikk címének meglétét igényli). Ezért úgy gondoljuk, hogy tovább javítható lenne módszerünk, amennyiben az eddigiekben figyelmen kívül hagyott (szócikkkel nem rendelkező) címkejelöltekhez is társítani tudnánk Wikipédia-oldalakat. További javítási lehetőség látunk még az egyes szócikkeken előforduló linkek alkalmas súlyozásában is, annak megfelelően, hogy azok mekkora mértékben kötődnek az adott szócikkben tárgyaltakhoz.

Ugyan a kézi címkézés során alkalmazott 554 különböző absztrakt címkének megközelítőleg 20%-a bír csak Wikipédia-szócikkkel, ezek közül 58-at sikerült pontosan, vagy legalább egy közeli szinonimájával meghatározni módszereink valamelyikével. Az esetlegesen tévesen kiválasztott absztrakt címkéket pedig a későbbi címkeszűrés lépések során igyekeztünk eredményesen eltávolítani, amit a teljes címkéző rendszerünk eredeti várakozásainkat meghaladó végső 77.5%-os értékelése is alátámaszt.

Eljárásunkról az is elmondható, hogy a Wikipédia többnyelvűségéből fakadóan más nyelvekre is könnyűszerrel adaptálható, eredményessége pedig várhatóan az adott nyelven elérhető Wikipédia szócikkeinek számától, valamint az oldalak szerkesztésének (a köztük lévő linkstruktúra) minőségétől is függ.

6 Köszönetnyilvánítás

A kutatást – részben – a TEXTREND projekt (Jedlik Ányos program) keretében az NKTH támogatta.

Hivatkozások

1. Gabrilovich, E., Markovitch, S.: Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis (2007)
2. Farkas R.: Az [origo] automatikus címkézési projekt tapasztalatai. In: Tanács A., Szauder D., Vincze V. (szerk.): VI. Magyar Számítógépes Nyelvészeti Konferencia (2009) 84-92
3. Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., Nevill-Manning, C. G.: KEA: Practical Automatic Keyphrase Extraction
4. Gantz, J. F. et al.: The Diverse and Exploding Digital Universe - An Updated Forecast of Worldwide Information Growth Through 2011. <http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf> (2008)
5. Kim, J. W., Selçuk Candan, K., Tatemura, J.: CDIP: Collection-Driven, yet Individuality-Preserving Automated Blog Tagging (2008)
6. Grineva, M., Grinev, M., Lizorkin, D.: Extracting Key Terms From Noisy and Multi-theme Documents. (2009)
7. Strube, M., Ponzetto, S. P.: WikiRelate! Computing Semantic Relatedness Using Wikipedia. American Association for Artificial Intelligence (2006) 1419-1424

8. Mihalcea, R.: Using Wikipedia for Automatic Word Sense Disambiguation. Proceedings of NAACL HLT 2007 (2007) 196-203
9. Sood, S. C., Owsley, S. H., Hammond, K. J., Birnbaum, L.: TagAssist: Automatic Tag Suggestion for Blog Posts. 1th International Conference on Weblogs and Social Media (ICWSM'2007)
10. Patwardhan, S., Banrjee, S., Pedersen, T.: Using Measures of Semantic Relatedness for Word Sense Disambiguation. CICLing 2003, LNCS 2588 (2003) 241-257
11. Cucerzan, S.: Large-Scale Named Entity Disambiguation Based on Wikipedia Data. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (2007) 708-716
12. Waltinger, U., Mehler, A., Heyer, G.: Towards Automatic Content Tagging: Enhanced Web Services in Digital Libraries Using Lexical Chaining. 4th Int. Conf. on Web Information Systems and Technologies (WEBIST '08) (2008) 231-236