Népi Hiedelem Gyűjtemény Analízise Fuzzy Pseudo-tezaurusszal

Szaszkó Sándor [I]. Kóczy T László [I, II] Gedeon Tamás [III]

[I] Budapesti Műszaki és Gazdaságtudományi Egyetem
Távközlési és Médiainformatikai Tanszék
1117 Budapest, Magyar tudósok krt. 2.
{szaszko, koczy}@tmit.bme.hu

[II] Széchenyi István Egyetem
Villamosmérnöki és Informatikai Intézet
9026 Győr, Egyetem tér 1.
koczy@sze.hu [II]

[III] Ausztráliai Nemzeti Egyetem
Informatikai Tanszék
ACT 0200 Canberra
Ausztrália
tom.gedeon@anu.edu.au

Kivonat: Az ideális tezaurusz szócsoportokból áll. A csoport szavai egy olyan fogalomhoz tartoznak, amely absztrakt is lehet, vagyis a valóságban nem megtalálható. Kutatásunk célja automatikus tezauruszgenerátor módszer kifejlesztése és hangolása szavak dokumentumokban való közös-előfordulása alapján. A pszeudó-tezaurusz szócsoportjai egy-egy témát írnak le. E témák köré felépíthető a dokumentumok tartalmi csoportosítása. 2704 magyar népi hiedelem szöveg feldolgozását végeztük el. A lehető legnagyobb számú fogalom megtalálása céljából egy tapasztalati súlyt vezettünk be, melynek kiválasztását részletesen indokoljuk. Már a kutatás eddigi részeredményei is segítettek néprajzkutatóknak elemezni és megérteni a korpusz rejtett struktúráját.

Bevezetés

A természetes nyelvek sok hasonló szót használnak egy vagy több hasonló fogalom kifejezésére. Speciális szótárak használata válik szükségessé, ha szeretnénk az összes olyan dokumentumot visszakeresni, amelyek egy adott témához tartoznak. Tezaurusz kifejezések gyűjteménye, amelyben az azonos csoportba besorolt szavak egy adott fogalmat írnak le. A tezaurusz használatával felfedhetjük a kapcsolatot olyan dokumentumok között is, amelyek nem tartalmaznak azonos kifejezéseket, bár azonos témáról szólnak.

Az automatikus kulcsszó keresés a legelterjedtebb megoldás a dokumentumkeresésre, habár könnyen belátható, hogy kulcsszót nem tartalmazó dokumentum is lehet releváns a keresésben. Vegyük például azt az esetet, amikor a „puha számítástudo-

mány" (soft computing) kulcsszóra keresünk; a Fuzzy rendszerekről vagy neurális hálózatokról szóló szövegek nem lesznek benne a keresés eredményhalmazában, habár relevánsak a keresett témához. Ugyancsak nem találjuk ezzel a módszerrel azon közösségek dokumentumait, akik „számítástudományi intelligencia" (Computational Intelligence) kifejezést preferálják azonos kontextusban.

Az [1][4] tanulmányokban szavak hierarchikus közös-előfordulási mértékét javasoltuk az egyes szavak, illetve szócsoportok fontosságának jelzésére. A dokumentumok címében, alcímében, kivonatában szereplő szavak kiemelt fontosságot kaptak és hozzákapcsolódtak a dokumentumtörzs minden szavához. Általában is elmondható, hogy fuzzy logika alkalmazása automatikus dokumentumok keresésben nem új keletű, a legfontosabb eredmények a [5]-ban vannak összegezve.

A hiedelem gyűjtemény fuzzy élőfeldolgozása

Ha adott területről származó szöveget analizálunk, akkor a szavakat négy fő halmazba csoportosíthatjuk, ahogy ez az 1. ábrán is látszik.

A stop szavak a nyelv olyan eszközei melyeket szinte minden szövegben megtalálhatunk, a szöveg tartalmáról nem hordoznak információt. A relatív stop szavaknak hasonló szerepe van a stop szavakhoz, de csak adott típusú szövegek esetén, mint például a jogi dokumentumok esetén a „törvény" szó. Az általunk feldolgozott hiedelem gyűjteményben nem azonosítottunk relatív stop szavakat.

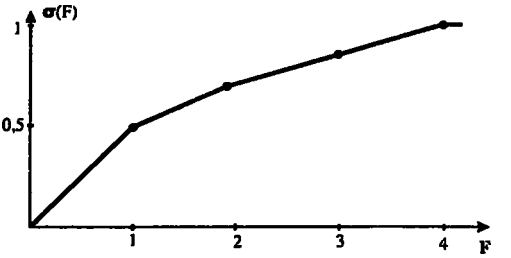
1. ábra Szókategóriák a dokumentumokban

A stop szavak eltávolítása után fennmaradó szavak a fontos szavak, melyek felhasználhatóak további analízisre. A fontos szavak egy része pontosabb információt ad a dokumentum tartalmáról, ezeket a szavalat nevezzük kulcsszavaknak.

A gazdag magyar néphagyományból 2704 db népi hiedelemszöveget bocsátott digitális formában rendelkezésünkre a a Néprajzi Múzeum. Ezen szövegek átlagos hossza 2–5 sor, vagyis igen rövid. Igen sok a régies, vagy egy-egy nyelvjárás szerinti szó, ezért szükség volt egy előfeldolgozási szótárra, amelyben kb. 13 000 szó (ragozott formákat is beleértve) 1704 fő bejegyzés alá lett besorolva. Ez után a korpuszt a K' mátrixban ábrázoltuk; a K' mátrix felső élén a szavakat soroltuk fel, míg oldalán a dokumentumokat.

Kritikus pont a K' mátrix előfordulási értékeinek fuzzy tagsági értékekké transzformálása. A tagsági érték fejezi ki, milyen mértékben jellemző egy szó az adott dokumentumra.

A traszformációra $\sigma(F)$ sigmoid függvényt választottuk [2]. A „S" formájú $\sigma(F)$ függvény konkrét alakját az adott korpuszhoz kell illeszteni. A dokumetumok rövidsége miatt a mi estünkben már az egyszeri előfordulásnak is nagy jelentősége van, az esetek 99,9%-ban nem fordul elő egy szó 4-nél többször egy dokumentumban. Ezek alapján alakítottuk ki a 2. ábrán látható függvény.



2. ábra A korpuszon használt sigmoid függvény

2.1 Szógyakorisági mérték

Későbbi használatra definiáljuk:

$$I_w = \sum_{i=1}^{N} \sigma_{d_i W} \tag{1}$$

Az l_w szógyakorisági mérték mutatja meg egy a w szó fontosságát a dokumentumgyűjteményben.

Szavak együttes-előfordulása alapján Fuzzy Pseudo-tezaurusz

Az ideális tezaurusz úgy definiálható, hogy minden szót hozzá rendelünk egy vagy több fogalomhoz. A fogalmak lehetnek absztraktak is, a való világban nem megtalálhatók. Két szót akkor tekintünk egymás szinonimájának, ha azonos fogalmakhoz tartoznak. Fuzzy tezaurusz esetén a fogalomhoz tartozás és így a szinonimaság mértéke (azaz az összetartozás foka a tezauruszban) fuzzy mérték, nulla és egy közötti szám.

A javasolt automatikus tezauruszgenerálás estén a fogalmak szerepét a dokumentumok töltik be, tehát akkor mondjuk két szóról, hogy szinonimák, ha azonos dokumentumokban szerepelnek. Természetesen ez a kapcsolat is leírható fuzzy mértékkel.

Sajnos az általános fogalmak és a dokumentumok között koránt sem létezik egy-egy leképezés. Egy dokumentum általában több absztrakt fogalmat is tartalmaz és sokat érintőlegesen taglal, így az automata módon generált tezaurusz nagy eltéréseket mutathat az ideálistól, ezért e módszerrel kapott tezauruszt pseudo- vagyis áltezaurusznak nevezzük.

Fuzzy pseudo-tezaurusz létrehozása:

1. lépés: Közös előfordulási mérték számítása

A számítások alapját a szöveg előfeldolgozásakor $\sigma(F)$ értékek adják, amelyek azt mutatják meg, hogy egy adott szó (W_i) milyen mértékben jellemző egy dokumentumra (D).

$$\mu'_{ij}(D) = \min(\sigma_{W_iD}, \sigma_{W_jD}) \tag{2}$$

$$\mu_{ij} = \frac{1}{C}\frac{1}{s}\sum_{z=1}^{N}\mu'_{ij}(D_z)$$

ahol C konstans feladata μ_{ij} értékét a [0,1] intervallumban tartani. C állandó az egész korpuszon, míg az s súly paraméter értéke változhat a szavak (i,j) függvényében is. A legegyszerűbb választás C értékére N, a dokumentumok száma, de ekkor az összes μ_{ij} érték nagyon kicsi. μ_{ij} jobban kihasználja a [0,1] intervallumot ha

$$C = \max_{i,j}\left(\frac{1}{s}\sum_{z=1}^{N}\mu'_{ij}(D_z)\right) \tag{3}$$

2. lépés: α-vágat

Ha a fontos szavak száma M, akkor a közöselőfordulási mértékek (μ_{ij}) egy MxM mátrixot alkotnak, hívjuk ezt W-nek. W mátrix mindkét oldalán a fontos szavak szerepelnek.

Mivel $\mu_{ij} = \mu_{ji}$ W ábrázolható egy irányítatlan gráffal. Válasszunk egy olyan α-t, amellyel elkészítve az α-vágatot – azaz kinullázva az összes α-nál kisebb értéket W-ben – csak 30-40 sor tartalmaz 0-tól különböző értéket. Ábrázoljuk ezt a redukált gráfot, ez reprezentálja a speudo-tezauruszt.

3. lépés: Maximum kikkek keresése

Ha a gráfban két pont (szó) éllel van összekapcsolva, akkor ők egymás szinonímái (kiterjesztett értelemben). Ha szavak egy részhalmazában mindenki mindenkinek kapcsolatban áll, akkor ők egy egy általánosabb értelemben vett fogalomhoz tartoznak.

A legnagyobb teljesen összekötött szócsoportok, maximum klikkek megkeresésével a korpusz főbb fogalmait azonosítjuk.

4. lépés: Fuzzy klikk

Sok esetben a maximum klikkeknek sok közös pontjuk van és csak egy-két szóban térnek el egymástól. Mivel az α-t önkényesen választottuk, ezért értelmes lehet megvizsgálni, hogy ezek az egymáshoz közeli klikkek azonos fogalmakat írnak-e le.
Válasszunk ki két klikket, melyek csak egy pontban különböznek és vizsgáljuk meg, hogy a két pont össze van-e kötve $\alpha'=0.7\alpha$ vágat szintjén. Ha igen, akkor a két klikket egyesítjük.

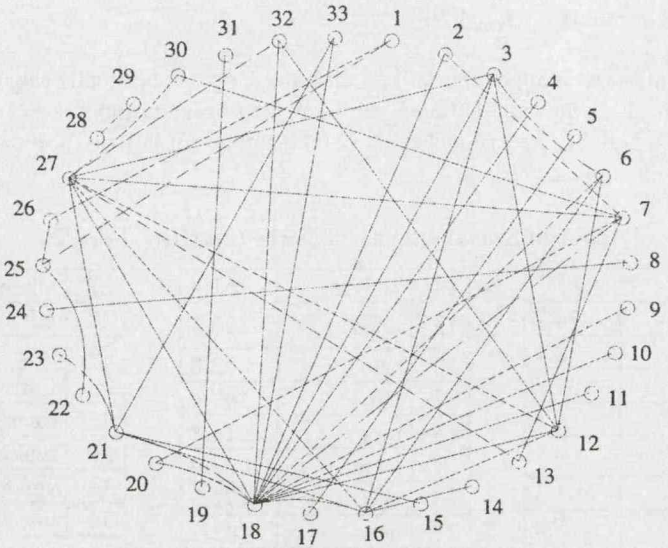
3.1 Súly s=1

A legtöbb fogalom felderítése céljából különböző s súly értékeket próbálunk ki.
Itt a közös előfordulás mértéke egyenesen arányos a szópárok együtt előfordulásának számával. Az I_w oszlop tanulsága szerint csak nagy gyakoriságú szavak maradtak benne az α-vágatban. A leggyakoribb szavaknak (mint pl. megy, tesz) van a legtöbb élük.

3. Táblázat Szavak listája súly s=1

NR	HUN.	I_w
1	ad	74.7
2	asszony	92.1
3	este	103.3
4	fent	69.1
5	fog	123.5
6	férj	66.1
7	gyermek	160.8
8	György	31.8
9	haza	42.8
10	3	83.7
11	ház	173.6
12	karácsony	90.4
13	kicsi	110.9
14	legény	41.7
15	Luca	52.8
16	lány	142.4
17	meghal	90.8
18	megy	218.4
19	mise	27.6
20	mond	108.3
21	Nap	172.3
22	ront	48.7
23	sok	86.0
24	szent	37.8
25	tehén	97.1
26	tej	61.3
27	tesz	177.8
28	tojik	32.3
29	tyúk	77.4
30	víz	98.9
31	éjfél	40.5
32	éjjel	87.2
33	év	94.7

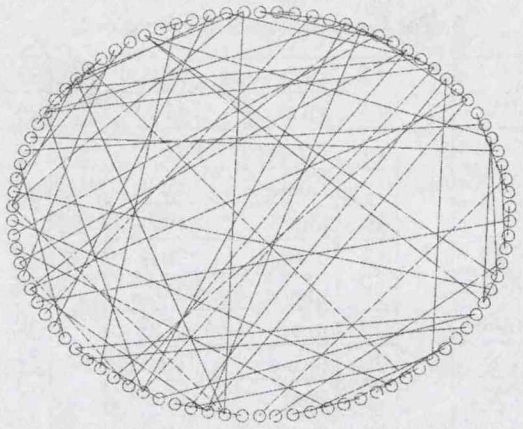
Habár nem is az összes, de sok szópár mutat jelentésbeli kapcsolatot. Találhatunk mögöttes logikát a maximum klikkekben, de általában egy-egy nagyon gyakori szó megjelenik, mint kakukktojás a csoportban. (Lásd 3.ábra)



3. ábra Pseudo-tezaurusz gráfja, súly s=1

3.2 Súly s=max(I_{WA} ,I_{WB})

A gyakori szavak dominanciájának elkerülésére osszuk el a közöselőfordulási mértéket, μ_{ij}-t a gyakoribb szó szógyakorisági mértékével (I_w). Ebben az esetben C=1 mivel I_w sohasem kisebb, mint μ'_{ij}.



4. ábra Pseudo-tezaurusz gráfja, súly s= max(I_{WA} ,I_{WB})

A legkisebb nem üres α-vágat 90 pontot tartalmaz. Minden bent maradó szó esetén I_w=0.5, tehát ezek a szavak csak egyetlen egyszer fordulnak elő az egész dokumentum gyűjteményben. Emiatt a feltárt kapcsolatoknak nem lehet értelmük, és ha megvizsgáljuk nincs is.

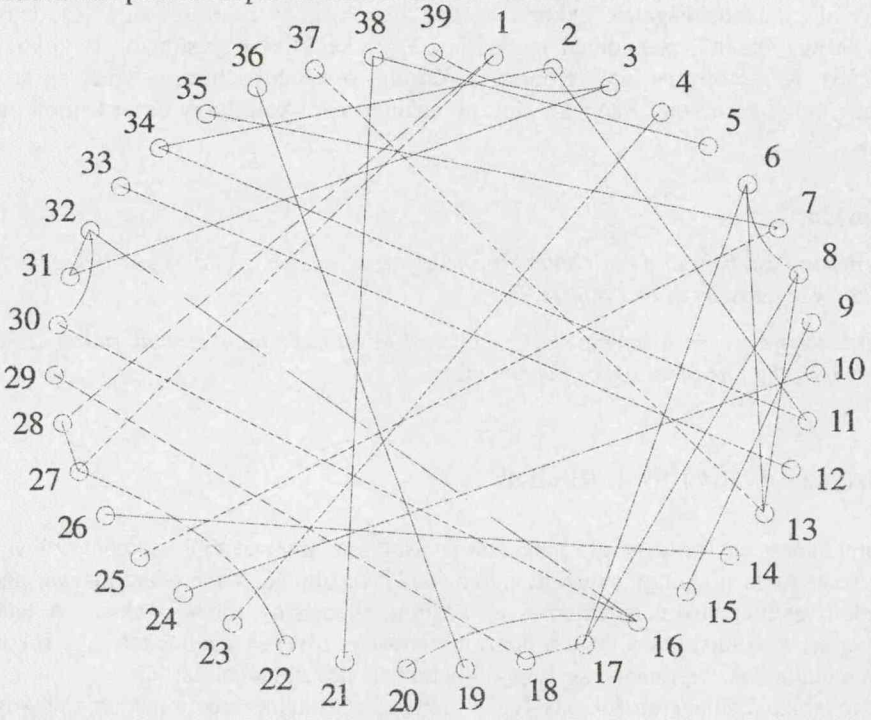
3.3 Súly s=1+ (max(I_{WA} ,I_{WB}))/20

Az előző két alfejezet alapján arra kell gondoljunk, hogy az optimális súly paraméter 1 és max(I_{WA}, I_{WB}) között van. Több súlytényező megvizsgálta után az s=1+ (max(I_{WA}, I_{WB}))/20 adta a legjobb eredményeket. A 2. táblázatban jól látható, hogy az I_w értékek széles skálán mozognak.

4. táblázat Szavak listája, súly s=1+ (max(I_{WA}, I_{WB}))/20

NR	HUN.	I_w
1	ad	74.7
2	bal	28.4
3	csibe	35.2
4	csörög	5.5
5	cédula	9.5
6	este	103.3
7	fecske	16.8
8	férj	66.1
9	gyermek	160.8
10	György	31.8
11	jobb	30.9
12	jön	91.1
13	karácsony	90.4
14	kenyér	51.4
15	kicsi	110.9
16	Luca	52.8
17	lány	142.4
18	megy	218.4
19	mise	27.6
20	nap	172.3
21	név	43.3
22	ront	48.7
23	szarka	6.7
24	szent	37.8
25	szeplő	10.3
26	süt	33.6
27	tehén	97.1
28	tej	61.3
29	tesz	177.8
30	tojik	32.3
31	tojás	43.8
32	tyúk	77.4
33	vendég	29.8
34	viszket	25.5
35	vér	17.6
36	éjfél	40.5
37	éjjel	87.2
38	ír	20.7
39	ültet	21.3

Az 5. ábrán látható gráfon az egyik pontnak sincs túl domináns, az eredményeket torzító szerepe, még az olyan kifejezetten gyakori szavaknak is, mint tesz, vagy megy nincs több, mint kettő éle, tehát sikerült elkerülni, hogy legyen néhány szó amely szinte minden más ponthoz kapcsolódik a 3. ábrához hasonlóan.



5. ábra Pseudo-tezaurusz gráfja, súly s=1+ (max(I_{WA}, I_{WB}))/20

5. táblázat Maximum klikkek

ad (1)	tehén (27)	tej (28)
bal (2)	jobb (11)	viszket (34)
este *(6)*	férj *(8)*	karácsony *(13)*
este *(6)*	férj *(8)*	lány *(17)*
Luca (16)	tojik (30)	tyúk (32)

csibe	ültet
csörög	szarka
cédula	ír
Fecske	szeplô
Fecske	vér
gyermek	kicsi
György	szent
jön	vendég
karácsony	éjjel

kenyér	süt
luca	Nap
lány	megy
megy	tesz
mise	éjfél
név	ír
ront	tehén
tojás	tyúk

A 3. táblázat az 5. ábra gráfjának maximum klikkjeit sorolja fel. Ezek a szócsoportok egy-egy általános értelemben vett, a korpusz számára fontos fogalmat határoznak meg. A 3. táblázat harmadik és negyedik sora csak egy szóban tér el. Ha megvizsgáljuk W mátrix alacsonyabb vágatait, akkor megállapíthatjuk, hogy a „Karácsony" és „lány" $\alpha'=0.8\alpha$ mértékben kapcsolódnak egymáshoz, így a két klikk egyesíthető. Az új klikk: este, férj, Karácsony és lány. Könnyű elképzelni olyan hiedelmeket, amelyek arról szólnak, hogy karácsony estéjén a lánynak valamit tenni kell, hogy férjet kapjon magának.

Két példa:

"Karácsony estélyén doboskát sütnek s a leány az elsővel kiszalad és amely legénnyel találkozik legelőször az lesz a férje."

"Karácsony estélyén a lánynak egy öl fát kell felvenni és ha a számuk páros, akkor férjhez meg, ha páratlan, akkor nem megy."

Összegzés és további kutatások

Automatikus tezauruszgeneráló metódust javasoltunk, amelyet több súlytényezővel is megvizsgáltunk, majd egy hatékony paramétert javasoltunk. A metódust magyar népi hiedelem gyűjteményen alkalmazva azonosítottunk néhány főbb fogalmat. A talált szócsoportok értelmesek voltak, a hozzájuk tartozó szövegek egymással nagy rokonságot mutattak. Az eredményeket népraj kutatók alkalmazhatónak tartják.

A továbbiakban szeretnénk egy fuzzy mértéket definiálni, amely megadja egy-egy talált szócsoport, vagyis „fogalom" és dokumentum közelségét, majd ennek segítségével klaszetrezzük a dokumentumokat.

Köszönetnyilvánítás

Köszönjük Darányi Sándornak és Kiss Ferencnek a segítségét, és hogy lehetővé tették a korpuszhoz, illetve az előfeldolgozó szótárhoz való hozzáférést.

Referenciák

[1] K. Chakrabarty, L.T. Kóczy, T.D. Gedeon, Analysis of fuzzy relational charts in information retrieval, IETR99-01, School of Computer Science and Engineering, University of New South Wales, Sydney, 1999.

[2] L.T. Kóczy, Interactive σ-algebras and fuzzy objects of type N, J. Cybernet. 8 (1978) 273–290

[3] L T. Kóczy, T. D. Gedeon and J. A. Kóczy, Fuzzy tolerance relations and relational maps applied to information retrieval, Fuzzy Sets and Systems 126 (2002) 49–61

[4] L.T. Kóczy, T.D. Gedeon, Information retrieval by fuzzy relations and hierarchical co-occurrence, Part II, TR97-03, Department of Information Engineering, School of Computer Science and Engineering, University of New South Wales, Sydney, 1997.

[5] S. Miyamoto, Fuzzy Sets in Information Retrieval and Cluster Analysis, Kluwer, Dordrecht, 1990, 259p