A szöveg mint skálafüggetlen hálózat

Makrai Márton és Sass Bálint

MTA Nyelvtudományi Intézet
{makrai.marton,sass.balint}@nytud.mta.hu

Kivonat Cikkünkben a szöveget egy egyszerű konstrukció segítségével skálafüggetlen hálózatként ábrázoljuk, és megvizsgáljuk, hogy a hálózatelmélet sztenderd eszközei mit mondanak az ilyen módon reprezentált szövegről.
Kulcsszavak: hálózatok, skálafüggetlen hálózatok, köztiség, Zipf, gyakoriság, szöveg

1. Bevezetés

Azokat a hálózatokat, ahol a befokszámok Zipf-eloszlást követnek, *skálafüggetlen hálózat*oknak nevezzük [1] [2, 703. oldal]. Régóta tudjuk [3], hogy a szöveg szavainak gyakorisági eloszlása Zipf-eloszlású. Tanulmányunkban azt az élsúlyozott irányított gráfot vizsgáljuk, melynek csúcsai a szóalakok, a $\langle w_1, w_2 \rangle$ él súlya pedig a $w_1 w_2$ bigram gyakorisága. Az előbbiek miatt ez a gráf skálafüggetlen.

Ebben a bevezető szakaszban egy játékpéldán szemléltetjük a konstrukciót, majd bemutatjuk a kapcsolódó irodalmat és a HITS linkelemzési algoritmust, amivel a leglátványosabb eredményeket kaptuk. A következő szakaszban mutatjuk be a kísérleti eredményeket.

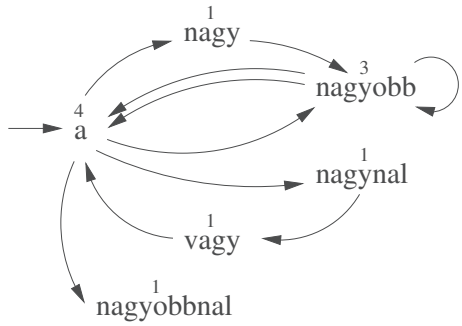
1.1. Gráf bigramokból

Az a feladat, hogy találjunk egy konstrukciót, melynek segítségével a szövegbeli szógyakoriság éppen a befokszám lesz egy alkalmasan megalkotott gráfban. Így skálafüggetlen hálózatot kapunk.

Az ötlet nagyon egyszerű: a hálózat csomópontjai a korpusz szavai lesznek, menjünk végig a korpuszon és minden szó esetén rajzoljunk be egy (1 súlyú) nyilat a hálózatba, ami az adott szóból indul ki és őt közvetlenül követő szóhoz vezet. Ha egy szópár többedszerre fordul akkor rajzoljunk be még egy nyilat, vagy ami ezzel ekvivalens, a nyíl súlyához adjunk hozzá 1-et. Az 1. ábrán látható hálózatot kapjuk. Ez egy olyan gráf lesz, melyben a csomópontok szavak, és az $A \to B$ él akkor létezik, ha van AB bigram a korpuszban, és az él súlya AB bigram gyakorisága.

A fenti reprezentáció tehát skálafüggetlen hálózatot eredményez. Ez azért nagyon jó, mert a skálafüggetlen hálózatok elméletének az elmúlt évek során kidolgozott összes eszközét, módszerét, mérőszámát [4] alkalmazni tudjuk rá.

Bízunk benne, hogy ezzel az eszköztárral valami újat tudunk mondani a szövegről, új módon tudjuk megragadni a szöveg bizonyos jellegzetességeit.



1. ábra: „A nagyobb nagyobb a nagynál vagy a nagy nagyobb a nagyobbnál.” példamondat ábrázolása. A dupla nyilat ábrázolhatjuk egy 2-es súllyal bíró szimpla nyíllal is.

1.2. Kapcsolódó irodalom

A mienkhez leghasonlóbb kutatás alighanem a TextRank [5]. Mihalcea és Tarau az irányított, 2-széles ablakkal végzett kutatásaikról azt írják, hogy rosszabb eredményeket hoztak, mint a az irányítatlan eset. Azt a meglepő következtetést vonják le, hogy a szövegnek nincs természetes iránya. Ha nyelvi adatból készült skálafüggetlen gráfról beszélünk, nem kerülhető meg [6] sem, ők azonban szemantikus hálókat vizsgálnak, míg mi magából a korpuszból vonunk le első sorban szintaktikai tanulságokat.

A mi konstrukciónk lényegesen eltér attól a bevett megközelítéstől, mely esetében akkor húzunk be egy (irányítatlan!) élt két szó között, ha egy közös trigramban megtalálhatók, más szóval, ha az egyik a másiktól (jobbra vagy balra) 1 vagy 2 szó távolságra van [7]. Ennél a modellnél a szöveg természetes balról jobbra rákövetkezése nincs reprezentálva, szemben a mi modellünkkel, ahol viszont lényegi elem. Emiatt a most bemutatott modell várhatóan kevésbé szemantikai, inkább szintaktikai jellemzőket tud majd megragadni. A kutatás feltáró alapkutatás jellegéből adódóan az eredmények esetleges majdani alkalmazása nem témája jelen cikknek.

A következő alszakaszban bemutatandó HITS algoritmus, amivel a leglátványosabb eredményeket kaptuk. nagyjából egyidős a skálafüggetlen gráfok ma népszerű fogalmával. Az utóbbi évtizedekben természetesen számos kutatás vizsgált skálafüggetlen gráfokat a HITS segítségével [8].

1.3. HITS

A HITS (Hyperlinkindukált témakeresés, *Hyperlink-Induced Topic Search* vagy hubok és tekintélyek, *hubs and authorities*) egy linkelemzési algoritmus [9], körülbelül egyidős a PageRankkel [10], csak persze sokkal kevésbé elterjedt. Az az alapötlete, hogy a fontos internetes oldalak kétfélék. A *hubok*, mint az index.hu

vagy a vajdasag.lap.hu, nagy linkgyűjteményként működnek: a rajtuk magukon megjelenő információnak nincs tekintélye, viszont más, hiteles oldalakra, a *tekintélyekre* irányítják a felhasználókat. A hubok és a tekintélyek definíciója kölcsönös: jó hub egy olyan oldal, amely sok tekintélyes oldalra mutat, nagy tekintélyük pedig azoknak az oldalaknak van ebben a modellben, melyekre számos jó hub mutat.

A hub- és tekintélyérték számítása iteríatíve történik. Kezdetben a számokat tetszés szerint választjuk (például minden oldalnak ugyanazt), majd minden iterációban egy oldal mértékadósága a rá mutató oldalak hubértékének összege lesz, a hubérték pedig a lap által mutatott oldalak tekintélyének összege. Az iterációk között a hub- illetve tekintélyértékek négyzetösszegét normalizáljuk. Ezzel az algoritmussal kaptuk a leginkább szembeötlő eredményünket.

2. Eredmények

Vizsgálatainkat[1] az MNSZ2 [11] véletlenül választott 1000 illetve 10000 mondatán végeztük. Az elemzéseinkhez a NetworkX python csomagot használtuk [12].

2.1. Erős összefüggőség

Az első nagyon egyszerű kérdés, hogy erősen összefüggő-e a gráf, azaz minden szóból elérhető-e irányított úton az összes többi szó. Lényegében erősen összefüggő lesz a gráf, esetleg a korpusz elején és végén lévő hapax szavakból álló „farok" fordulhat elő, ami megbontja az erős összefüggőséget (ld. az 1. ábrán a *nagyobbnál* szót).

2.2. Kisvilág tulajdonság

A kisvilág-szerkezet szemléltetésére idézzük Karinthy 1929-es *Láncszemek* című novelláját:

> Tessék egy akármilyen meghatározható egyént kijelölni a Föld másfél milliárd lakója közül, bármelyik pontján a Földnek – [a társaság egyik tagja fogadást ajánlott], hogy legföljebb öt más egyénen keresztül, kik közül az egyik neki személyes ismerőse, kapcsolatot tud létesíteni az illetővel, csupa közvetlen – ismeretség – alapon

A meglepően kis távolság, melyet úgy formalizálhatunk, hogy az L átlagos távolság csak logaritmikusan növekszik a csúcsok számában, nem a skálafüggetlen gráfok sajátja: Erdős–Rényi-féle véletlen gráfoknál is fennáll [13], ha a élek beválasztását kontrolláló p elég nagy ahhoz, hogy az egész hálózat összekapcsolódjon. (Vegyük észre a definícióban, hogy önmagában egy gráf kisvilág tulajdonságáról nem beszélhetünk, csak egy olyan gráfsorozat esetében, ahol a csúcsszám

[1] `https://github.com/makrai/TextBetweenness.git`

végtelenhez tart.) A kisvilág-szerkezet jellemzésében az alacsony C klaszterezési együtthatót is meg szokták követelni [14], de ennek az irányított gráfokra való általánosítása nem tűnik triviálisnak, ezért tanumányunkban a legrövidebb utak hosszából számított statisztikákra szorítkozunk. A mi 7 K-csúcsú, 13 K-élű gráfunkban az átlagos távolság $L = 4.89$. Az aszimptotikus tulajdonságot a poszteren elemezzük, amit a projekt repójában talál meg az olvasó.

2.3. Skálafüggetlenség

A skálafüggetlen gráfok definiáló tulajdonsága, hogy a fokszámok hatványeloszlást követnek [1]. Ezt nálunk Zipf törvénye miatt biztosítja az, hogy (az első és az utolsó szó kivételével) a be- és a kiélek súlyösszege egyaránt megegyezik a szónak a korpuszban való gyakoriságával. Az elméletileg garantált tulajdonságot statisztikailag is ellenőriztük. A fokszámeloszlást exponenciális eloszlással összehasonlítva 133-as likelyhood-ratiót és $5.91{\cdot}10^{-6}$-os szignifikanciaszintet kaptunk. A Zipf-együttható 2.20-nak adódott, ami összhangban van azzal, hogy az angolban 1.25 körülre teszik [15], a magyar pedig gazdagabb morfológiájú, így a Zipf-együttható is magasabbnak várható.

2.4. Távolságok

Egy másik gráfmérték is hasznosnak tűnik az együtt-előfordulási gráf elemzésében, a csúcsok különcsége (*eccentricity*), vagyis az adott csúcsból az összes többi csúcsba vezető legrövidebb utak hosszának maximuma. Míg a többi vizsgálatot tízezer mondatos mintán végeztük, az ebben a pontban említetteket csak ezer mondatoson, mert az összes (rendezett) csúcspárra ki kell hozzá számolni a legrövidebb út hosszát, aminek nagy az időkomplexitása. A *sugár* (a legkisebb különcség) nálunk 9, az *átmérő* (a maximális különcség) 19. Szemléletesen *közép*nek *(center)* hívják azokat a pontokat, amelyeknek a különcsége megegyezik a sugárral. Esetünkben egy ilyen csúcs van, a sok funkcióban használatos vessző (,) token.

2.5. Closeness centrality

A közelségi központiság (*closeness centrality*) mértéke (mely az adott csomóponttól az összes többi csomópontba vezető legrövidebb utak hosszának átlagaként adódik) egy érdekes jelenséget mutat (2. ábra). A szavak egységes eloszlásban helyezkednek el. Az eloszlásból néhány olyan elem lóg ki, amely „nem illeszkedik a magyar szövegbe”: ilyen az egyenlőségjel és egy HTML entity (`&verbar;`), illetve két angol szó (a *the* és az *of*), melyek előfordulnak a korpuszban. Ezeknek a tokeneknek tehát kisebb a közelségi központiság értékük annál, mint amit gyakoriságuk alapján várnánk. A kilógó elemek pontos karakterizálásához további vizsgálat szükséges.

2.6. HITS

A bigramgyakorliságokból épített gráfban egy szó hubértéke az őt követő szavak tekintélyének összege, egy szó tekintélyét pedig az őt megelőző szavak együttes hubértéke adja. A 3. ábrán első körben azt látjuk, hogy a gyakoribb (nagyobb fokszámú) szavaknak nagyobb a tekintélye – ez nem meglepő. Érdekesebb, hogy a kötőszavak és az igék jól elkülönülnek: az egész szókincsre jellemző bal lent – jobb fent átlótól inkább följebb vannak a kötőszavak, és ettől az átlótól inkább lejjebb vannak az igék. Azaz a kötőszavak tekintélye nagyobb annál, ami a gyakoriságuk alapján várható, az igéké pedig kisebb. (Tehát nem egyszerűen arról van szó, hogy a kötőszavak gyakorisága kiemelkedően magas, hiszen az azonos gyakoriságú igék és kötőszavak is jól elválnak egymástól.)

Az eredményre a következő intuitív magyarázatot javasoljuk: a kötőszavak nagy tekintélye azt jelzi, hogy az őket megelőző tokenek együttes hubértéke nagy. Mivel a szavak hubértéke az őket követő szavak tekintélyének összege, az ördögi körből úgy léphetünk ki, ha a kötőszavak előtti tokenek után más tekintélyes szavak is megjelennek. A gráf tehát tükrözi azt, hogy a kötőszavak olyan helyen állnak, ahol a balkontextus alapján sok más szó is állhat. A tipikus példa ilyen balkontextusra a már a center kapcsán is említett vessző token. Az igék esetében épp ellenkező a helyzet: az igék előtt megjelenő szavak (igemódosítók és bővitményi frázisok utolsó szavai) jobban determinálják, hogy igének kell következnie, mint más szavak balkontextusa az adott szót.

3. Összefoglalás

Az irányított skálafüggetlen hálózattá alakított szöveget érintő első vizsgálatainkban feltérképeztük, hogy az egyes hálózatelméleti eszközök mit mondanak erről a hálózatról, milyen értékek a jellemzőek. Legérdekesebb kezdeti eredményünk az, hogy úgy tűnik, hogy a HITS algoritmus képes egymástól elválasztani bizonyos szócsoportokat, és ezek a csoportok összefüggésben vannak a szófajokkal.

Köszönetnyilvánítás

Sass Bálint kutatásait az MTA Bolyai János Kutatási Ösztöndíja támogatta (ügyszám: BO/00064/17/1; időtartam: 2017-2020).

Hivatkozások

1. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. science **286**(5439) (1999) 509–512
2. Kovács, L., Orosz, K., Pollner, P.: Magyar szóasszociációk hálózata. Magyar Tudomány **173**(6) (2012) 699–705
3. Zipf, G.K.: The Psycho-Biology of Language; an Introduction to Dynamic Philology. Houghton Mifflin, Boston (1935)

4. Barabási, A.L.: Scale-free networks: A decade and beyond. Science **325** (2009) 412–413
5. Mihalcea, R., Tarau, P.: Textrank: Bringing order into text. In: Proceedings of the 2004 conference on empirical methods in natural language processing. (2004)
6. Steyvers, M., Tenenbaum, J.B.: The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. Cognitive science **29**(1) (2005) 41–78
7. i Cancho, R.F., Solé, R.V.: The small world of human language. Proceedings of The Royal Society of London. Series B, Biological Sciences **268** (2001) 2261–2266
8. Zhang, J., Ackerman, M.S., Adamic, L.: Expertise networks in online communities: structure and algorithms. In: Proceedings of the 16th international conference on World Wide Web, ACM (2007) 221–230
9. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM **46** (1999) 604–632
10. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab (1999)
11. Oravecz, Cs., Váradi, T., Sass, B.: The Hungarian Gigaword Corpus. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC2014), Reykjavík (2014)
12. Hagberg, A.A., Schult, D.A., Swart, P.J.: Exploring network structure, dynamics, and function using NetworkX. In: Proceedings of the 7th Python in Science Conference (SciPy2008), Pasadena, CA USA (2008) 11–15
13. Erdos, P., Rényi, A.: On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci **5**(1) (1960) 17–60
14. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world'networks. nature **393**(6684) (1998) 440–442
15. Kornai, A.: Mathematical Linguistics. Advanced Information and Knowledge Processing. Springer (2008)