Spektrális dokumentummodell

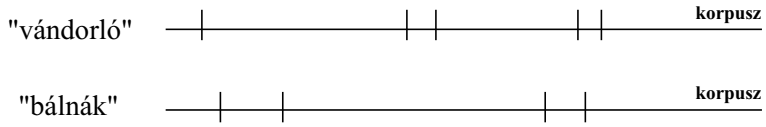
A *spektrális szövegbányászat* lényege abban rejlik, hogy a korpusz vektortérbeli reprezentációja helyett a teljes korpuszt áttranszformáljuk az ún. frekvencia tartományba, ahol a jelfeldolgozásból ismert diszkrét Fourier- (DFT) vagy diszkrét koszinusz- (DCT), illetve más hasonló transzformációval történik a szöveges dokumentumok további feldolgozása (Park et al, 2002). A szövegbányászat területén történő első megjelenése az ún. Fourier tartománybeli értékelés (Fourier domain scoring, FDS) módszerének ismertetésével kezdődött (Park, 2003). A vektortérmodell ismertetésénél kitértünk a modell hiányosságaira — ilyen pl. a dokumentumbeli pozícióra vonatkozó információ elvesztése. A vektortérmodell pontatlansága ellenére sikeresen és főleg hatékonyan alkalmazható a legtöbb szövegbányászati problémánál, folynak ugyanakkor kutatások olyan modellek megalkotására, amelyek kiküszöbölik a modell hátrányait, míg hatékonyságban összemérhető vagy jobb eredményt mutatnak annál. Ez a szakasz az egyik alternatív modellt, a spektrális szövegbányászatot ismerteti.

A jelenleg elterjedt információ-visszakereső algoritmusok számos hátrányával találjuk magunkat szemben, amikor szöveges információkeresési tevékenységünk során klasszikus keresőrendszerek (pl. Google) használatával olyan találatokat kapunk, amelyek nagyrésze számunkra teljesen irreleváns. Ennek legfőbb oka, hogy a jelenleg elterjedt keresőrendszerek kizárólag a keresőszavak dokumentumokban történő előfordulási gyakoriságát veszik figyelembe. A keresőszavak gyakorisági rangsorolását általában a tf-idf (ld. a könyv 36. oldalán) találati rangsoroló algoritmus segítségével végzik a klasszikus szövegbányászati és információkinyerési rendszerekben. Ezzel szemben az információ-visszakeresés egyik viszonylag új (2003) szövegbányászati megközelítése a spektrális elvű információkinyerés, amely képes felülmúlni a klasszikus módszerek hatékonyságát, ezáltal az információkeresést végző személy számára nagyobb relevanciahányadú találati listát állít előugyanazon dokumentumkorpusz felett.

A spektrális információkinyerés lényege abban rejlik, hogy a találati eredmények rangsorolásánál az algoritmus nem csak a keresőszavak gyakorisági értékeit veszi figyelembe, hanem azok dokumentumbeli pozícióit is. A spektrális szövegbányászat során a dokumentumkorpuszt nem a gyakorisági értékekre épülő szó-dokumentum mátrix formában reprezentált vektortérben elemezzük, hanem hullámtérben. A szövegmodellezés vektortérszerű ábrázolása a hagyományos esetben elveszíti a szövegek szintaktikai információit, hiszen nem veszi figyelembe a szavak egymáshoz képesti elhelyezkedését. A szavak szövegeken belüli pozícióját megragadni képes spektrális szövegbányászati modell már jóval összetettebb

matematikai formalizmust igényel, azonban megtartja a szavak pozíciójából eredő információtartalmat, amelynek hasznosításával esetenként jobb találati lista kapható.

A spektrális reprezentáció alapja egy tenzor, amely ebben az esetben az egyes térbeli helyekhez rendelt egyedi és egymástól független vektorterek összességeként képzelhető el. A kifejezések szövegen belüli térbeli ábrázolása új, statisztikai információkat rejt magában, ha elvégzünk egy frekvenciatartományba képező transzformációt. A dokumentumok spektrális térbe történő transzformációjának alapja az ún. *szószignál*ok (term signal) képzése. Egy szószignál azt mutatja meg, hogy az adott szó a teljes korpuszban mely pozíciókon fordul elő. Ez gyakorlatilag egy bináris vektor, amely ott tartalmaz 1-eseket, ahol az adott szó előfordul, egyébként 0 értéket vesz fel. A vektor hossza megegyezik a dokumentumkorpusz összes szavainak számával. A szószignált az 1. ábrán látható példa szemlélteti. A szignálok azt jelzik, hogy a kiterített korpusz mely pozíciójánál szerepel a vizsgált szó (Park, 2003).



1. ábra. Példa szószignálra

A dokumentumkorpusz reprezentációjának hullámtartományba történő transzformációja legtöbbször Fourier-, wavelet- vagy diszkrét koszinusz-transzformációval (vagy ezek módosított verzióival) történik. Ezen transzformációk alkalmazásával az egyes szavak pontos pozícióit tartalmazó információ megmarad. A transzformáció során az egyes szószignálok transzformációjára kerül sor az éppen alkalmazott transzformációs módszerrel, így alakítva azokat matematikai értelemben vett hullámmá.

Megvalósíthatósági szempontból a Fourier-transzformáció (ill. annak változatai, pl. gyors Fourier-transzformáció) alkalmazása a legcélszerűbb, mert megfelelő algoritmusokkal a módszer jól skálázható, nagy adatbázisokon is hatékonyan alkalmazható. A frekvenciatartományban a kifejezések és a dokumentumok közötti kapcsolatot már fázis- és amplitúdóértékek is jellemzik. Ezekre a jellemzőkre épül az információkeresés spektrális változata, a Fourier-tartománybeli értékelés (Park et al, 2002; Park, 2003).

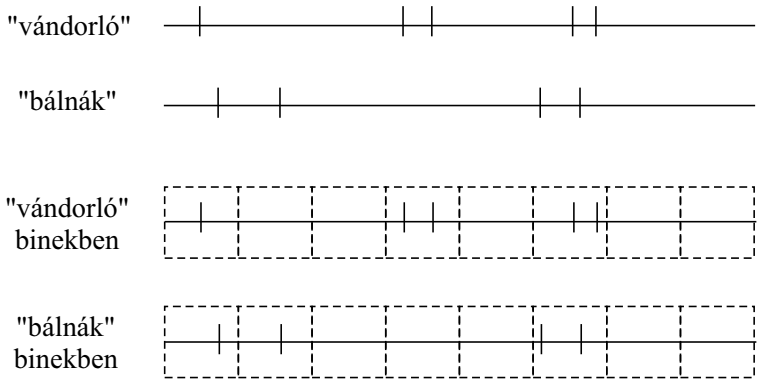
A spektrális transzformáció során kapott hullámtér-reprezentáció felett ezt követően sor kerülhet az információkinyerési feladatok elvégzésére. Adott keresőszavak esetén a megfelelő szószignálok uniójaként képzett ún. *lekérdezésszignál*t (query signal) kapunk, amely szintén egy bináris vektor lesz csakúgy, mint az egyszerű szószignálok. A lekérdezési vektor mentén (ahol a lekérdezési vektor 1-et vesz fel értékül) a hullámtérbeli dokumentummátrix oszlopainak összegzése után jutunk el az egyes dokumentumok adott lekérdezésre vonatkozó számszerű ranglistájához. Ezzel a módszerrel tulajdonképpen hullámok interferenciáját vizsgáljuk. Ahol a lekérdezési vektor hulláma és a korpusz hullámai azonos fázisban vannak, ott találjuk a lekérdezés szempontjából legrelevánsabb információt (Vázsonyi, 2005).

A spektrális információkinyerés pontos folyamatát az alábbi lépések szemléltetik.

- szószignálok generálása;
- a korpusz tisztítása (pl. stopszószűrés) után megmaradt szavak ún. binekbe sorolása (egy *bin* egy meghatározott számú, egymás után következő szó tárolására szolgáló csoportosító egység);
- invertált index készítése;
- dokumentum (és lekérdező sztring) súlyozásának elvégzése;
- transzformáció a frekvenciatartományba;
- amplitúdószámítás a lekérdező sztring esetében a lekérdezés súlyozása mellett;
- a lekérdezés fázis pontosságának számítása;
- a lekérdezés és korpuszhullámok kombinálása a dokumentumok relevancia értékeinek meghatározására (Park et al, 2001; Park, 2003).

A szószignálok generálása minden szó esetében egy bináris egydimenziós vektort eredményez. A bináris szószignálvektor hossza annyi lesz, ahány szót tartalmaz a teljes korpusz egészében, tehát egy szó mindannyiszor számít, ahányszor előfordul. Tegyük fel, hogy a teljes korpusz mindössze ennyi: *A vándorló bálna a bálnák egy speciális fajtája, amely a többi bálnával ellentétben nagyobb távolságokat tesz meg az év során a főbb óceáni áramlatok mentén.* Ez a korpusz a stopszószűrés és a szótövezés után az alábbiak szerint módosul: *vándorol bálna bálna speciális fajta bálna ellentét nagy távolság tesz év fő óceán áramlat mentén.* Ekkor a *bálna* szó szignálja az alábbi lesz: [011001000000000].

A következő lépés a szószignálok binekbe rendezése (ld. a 2. ábrát), amelynek elsődleges célja a komplexitás csökkentése. A felső két szignál a *vándorló* és a *bálna* szavak előfordulását jelzi a teljes korpuszban, amelynek kiterített formáját a vízszintes vonal szemlélteti.



2. ábra. Binekbe rendezés

A binekbe rendezés során nem gyakorisági értékek alapján képzett vektorokként reprezentáljuk az adott dokumentumokat, hanem binek összességeként, ahol a binek adott számú szót tartalmaznak. Ha egy binben B darab szót tárolunk, és egy adott dokumentum W számú szót tartalmaz, akkor ez a dokumentum $\frac{W}{B}$ darab bint fog tartalmazni. A 2. ábra példáján a *vándorló* szó a binekbe sorolás után a következő vektort eredményezi: $[10020200]$, a *bálna* pedig ennek mintájára a $[11000200]$ vektort $B = 8$ paraméter mellett. Ezeket a vektorokat szóvektoroknak nevezzük (word vectors).

Ezt követően készítjük el az invertált indexet. A vektortérmodell mintájára jelen esetben is az invertált index a szóvektorok gyors visszakeresésére szolgál. Mivel a spektrális modellben nem az egyes szavaknak a dokumentumokban való előfordulását tároljuk, hanem a dokumentumok számánál nagyobb számú binekben való előfordulásait, így az invertált index tárolása a vektortérmodellhez képest nagyobb tárhelyigénnyel jár. A tárolás struktúrája az $n, \langle b_1, f_1 \rangle \langle b_2, f_2 \rangle \ldots \langle b_n, f_n \rangle$ formát ölti, ahol n a binek száma, b jelzi az aktuális bin számát, f pedig az adott binben az adott szó előfordulási gyakoriságát. Ez a fajta tárolás az ún. helyzeti teret reprezentálja.

A már ismertetett hasonlósági metrikák (pl. koszinusz) hatékonysága és reprezentációs képessége nagymértékben növelhető, ha előtte valamilyen alkalmas súlyozást hajtunk végre a reprezentációs mátrixon, differenciálva ezzel a szöveg

1. táblázat. Invertált index

	bin_1	bin_2	$\cdots$	bin_n
t_1	$\langle b_1, f_1 \rangle$	$\langle b_2, f_1 \rangle$	$\ldots$	$\langle b_n, f_1 \rangle$
t_2	$\langle b_1, f_2 \rangle$	$\langle b_2, f_2 \rangle$	$\ldots$	$\langle b_n, f_2 \rangle$
$\ldots$	$\ldots$	$\ldots$	$\vdots$	$\ldots$
t_N	$\langle b_1, f_N \rangle$	$\langle b_2, f_N \rangle$	$\ldots$	$\langle b_n, f_N \rangle$

szavait aszerint, hogy milyen módon oszlanak el a korpuszban. A spektrális információvisszakeresés szakirodalmában az alábbi formulákkal is találkozhatunk, ahol a felső képlet a dokumentumok, míg az alsó a lekérdező sztring súlyozására használható hatékonyan.

$$w_d(f_{d,t}) = \frac{1+\ln f_{d,t}}{(1-s)} + \frac{sW_d}{av_{d\in DW_d}}$$
$$w_q(d_{d,t}) = (1+\ln f_{q,t}) \cdot \ln(1+\frac{f^m}{f_t}),$$

ahol w_d és w_q rendre a dokumentum, illetve a lekérdező sztring súlya, $f_{d,t}$ és $f_{q,t}$ a t szó előfordulási gyakorisága a dokumentumban, illetve a keresőkifejezésben, $s \in [0,1]$ a *meredekségi faktor* (slope factor), W_d és $av_d \in DW_d$ a dokumentum, illetve az átlagos dokumentum vektornorma, f_t a t szót tartalmazó dokumentumok száma, f_m pedig az f_t értékek maximuma. A dokumentumok súlyozása azt a célt szolgálja, hogy csökkentsük egy adott szó egy adott dokumentumban történő többszöri előfordulásának, illetve a dokumentumméretből adódó különbözőségek torzító hatásait. A lekérdezővektor súlyozásával pedig egyrészt el tudjuk érni, hogy egy adott szó lekérdező sztringben történő többszöri előfordulásából, illetve a lekérdező sztringben lévő stopszó jellegű, gyakori szavak dokumentum rangsorolási hatásaiból eredő torzító tényezőket kiiktassuk. A spektrális szövegbányászatban ennek kiemelt szerepe van, mert ezt alkalmazva a későbbi Fourier-frekvenciatartománybeli értékeléssel erre építve még jobban növelhető a találati relevancia értéke. Amennyiben a spektrális reprezentációban binekre osztást is alkalmazunk a komplexitás csökkentésére, akkor a súlyozást ezekre a binekre végezzük el a dokumentumvektor helyett az alábbi módon, ahol $f_{d,t,b}$ a t szó

előfordulási gyakorisága a d dokumentum b binjében.

$$w_d(f_{d,t,b}) = \frac{1+\ln(f_{d,t,b})}{(1-s)} + \frac{sW_d}{av_{d\in DW_d}}$$

A következő lépés a frekvenciatartományba történő transzformáció. Ezt a fajta transzformálást a mérnöki tudományokban széleskörben használják. Célja, hogy egy jel összetevőit egymástól lineárisan független szinuszoid összetevőkre bontsa. A Fourier-transzformáció alkalmazásával (diszkrét esetben a diszkrét Fourier-transzformáció) képlete szerint történik a frekvenciatartományba történő transzformáció az alábbi képlet szerint.

$$\nu_{d,t,\beta} = \sum_{b=0}^{\beta-1} \omega_{d,t,b} e^{\frac{-i2\Pi\beta b}{B}}$$

Spektrális szövegbányászati alkalmazások esetében a súlyozott szószignálok lesznek a transzformálandó szignálok, ahol a szignál elemei az $\omega_{d,t,b}$ értékek, ahol $b \in \{0,1,\ldots,B-1\}$. Mivel minden $\nu_{d,t,b}$ érték az $\omega_{d,t}$ szószignál projekciója egy β frekvenciájú szinuszoid hullámmá, így $\nu_{d,t}$ az adott szószignál spektruma. A spektrális komponens $\beta \in \{0,1,\ldots,B-1\}$. A fentiekből adódóan a diszkrét Fourier traszformáció az eredeti idő vagy helyzeti tartományból a jeleket frekvenciatartományba transzformálja át.

Amint láthatóvá válik egy jel spektruma, képesek vagyunk azonosítani a legfőbb frekvenciakomponenseket, amik a legjobban meghatározzák a jel alakját. A DCT tehát a következő megfeleltetést hajtja végre szövegbányászati alkalmazásokban.

$$\{\omega_{d,t,b}\} \Longrightarrow \{\nu_{d,t,b}\} = \{H_{d,t,b}\exp(i\varphi_{d,t,b})\},$$

ahol $\omega_{d,t,b}$ a t szó súlya a d dokumentum b binjében, $\nu_{d,t,b}$ a t szó b-edik frekvenciakomponese a d dokumentumban, $H_{d,t,b}$ és $\varphi_{d,t,b}$ az amplitúdó és fázis valós értékkészletű tényezője a $\nu_{d,t,b}$ frekvenciatényezőnek, i pedig a komplex képzetes egység.

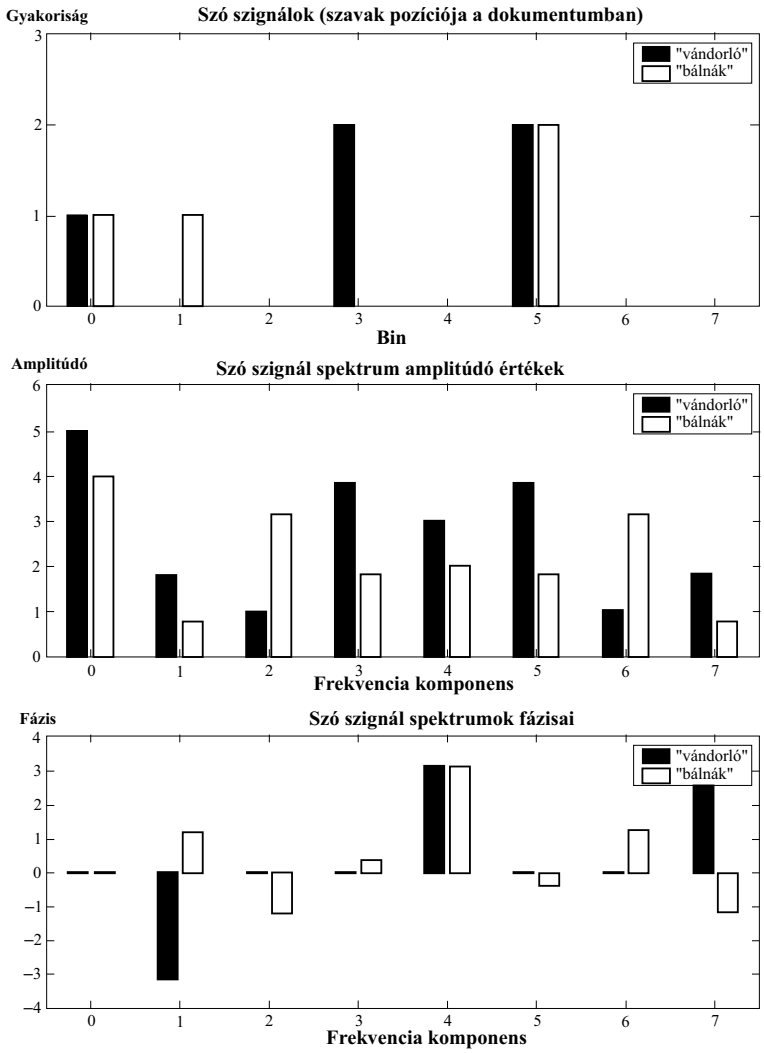
A Nyquist–Shannon mintavételi elmélet szerint a legnagyobb fellelhető frekvenciakomponens egy valós értékkészletű jelben a mintavételezési szint fele. Ebből következik, hogy ha B számú bint választottunk a szószignál esetében, akkor elegendő a frekvenciakomponenseket csak 0 és $\frac{B}{2}$ között vizsgálni. A helyzeti információt tároló szó bineken végrehajtott DCT vagy egyéb hullámtérbe irányuló transzformáció segítségével képesek vagyunk érzékelni, hogy az adott szó milyen formában van jelen a dokumentumkorpusz egészében, azon — mint jel — hogyan vonul keresztül.

Minden frekvenciakomponens amplitúdó- és fázisértékeket tartalmaz, amelyek úgy értelmezhetőek, mint a komponens hatása és az eltolódása. A hatás (amplitúdó) a szószignál alakjáról informál bennünket. Amennyiben egy kisebb frekvenciakomponens amplitúdója nagy a többi komponenshez képest, akkor a szó csoportosan fordul elő a dokumentumkorpusz kevés számú helyén. Amennyiben egy nagy frekvenciakomponens amplitúdója nagy a többi komponenshez képest, akkor a szó által alkotott klaszterek elszórva végig előfordulnak a dokumentumban.

Az eltolódás (fázis) a szó pozícióját jelzi a dokumentumkorpusz egészén, és radiánban adja meg annak értékét. Az eltolódás akkor válik igazán hasznossá, amikor két szószignált hasonlítunk össze. Ekkor, ha a két szó egy fázisban van, arra következtethetünk, hogy a két szó általában egymással együtt fordul elő. Ha több szó fázisa esik egybe, akkor az arra utal, hogy a szavak előfordulásai egymással összefüggenek. Amennyiben két szó fázisa általában nem esik egybe, akkor előfordulnak ugyan a dokumentumban, de jellemzően nem egymás közelében.

Amennyiben a fenti *vándorló bálnák* példával élünk, akkor látható, hogy a szavak a nulladik és az ötödik binben együtt fordulnak elő. Ha a transzformáló algoritmust a szószignálokra alkalmazzuk, akkor a spektrális komponensek amplitúdó- és fázisvizsgálatával kimutathatjuk, hogy hol fordulnak elő együtt. Amennyiben az első szó egy c komponense hasonló fázist mutat a másik szó c komponensével, akkor ezek a komponensek azonos fázisban vannak, és ez arra utal, hogy a két szó együttesen fordul elő a spektrum azon régiójában. Ez az eset áll fent a nulladik, harmadik és negyedik binek esetében, ha a fenti példában összehasonlítjuk a *vándorló* és *bálnák* szavak szószignáljait. Ha az említett komponensek amplitúdója nagy a többi komponenshez képest, akkor ez azt jelenti, hogy ezek a frekvenciakomponensek a szószignálok fő komponensei. Ebben az esetben ez az adott dokumentum relevánsnak tekinthető a lekérdezésre nézve. A leírtakat szemlélteti a 3. ábra.

A spektrális modell megalkotásában a következő lépés a szóspektrumok kombinálása. Amint végrehajtottuk a frekvenciatartományba (hullámtérbe) történő transzformációt a B hosszúságú vektorokon, $\lfloor \frac{B}{2} \rfloor + 1$ darab független komplex számot kapunk minden szóra. Ezeket a vektorokat szóspektrumoknak nevezzük. A spektrális modell szerint ahhoz, hogy egy dokumentum releváns legyen egy lekérdezésre nézve, a dokumentumnak nagy amplitútóértékkel kell rendelkeznie, és a vonatkozó fázisok egy fázisban kell lenniüka lekérdező sztring minden szavával. Ennek értelmében az amplitúdó és a fázis vizsgálatával külön kell foglalkoznunk.



3. ábra. Frekvenciatartománybeli összehasonlítás

A spektrumok kombinálásával megkaphatjuk az amplitúdó- $(H_{d,b})$ és fázis- $(\varphi_{d,b})$ pontossági értékeket a *d* dokumentum minden egyes *b* binjére (Park, 2003).

A spektrális szövegbányászat számos kihívás előtt áll, elsősorban a módszer számításigényét illetően. Alapvetően két lehetőség van, amelyek során a tárhely-szükséglet, illetve a keresési idő között kell kompromisszumot kötnünk. Lehető-

ség van a teljes dokumentumkorpusz minden term-szignáljának előzetes transzformációjára, ezzel elkerülhetjük, hogy a számításigényes hullámtranszformációt futásidőben végezzük el. Ekkor a módszer tárhelyigénye lesz nagyobb. Amennyiben a hullámtérbe történő transzformációkat futásidőben végezzük, akkor a módszer tárhelyigénye kisebb lesz, futáside azonban meghosszabbodik. A legígéretesebbnek a kompakt dokumentum reprezentálási módszerek alkalmazása tűnik, amely során a dokumentumkorpuszt olyan módon reprezentáljuk (pl. véges állapotú automatában), amely hatékonyan támogatja a futásidejű transzformációt (Vázsonyi, 2005).

Forrás:

L. A. F. Park. *Spectral Based Information Retrieval*. PhD thesis, The University of Melbourne, 2003.

L. A. F. Park, M. Palaniswami, and R. Kotagiri. Internet document filtering using fourier domain scoring. In *Proc. of PKDD-01, 5th European Conf. on Principles of Data Mining and Knowledge Discovery*, pages 362–373, 2001.

L. A. F. Park, M. Palaniswami, and K. Ramamohanarao. A novel web text mining method using the discrete cosine transform. In *Proc. of PKDD-02, 6th European Conf. on Principles of Data Mining and Knowledge Discovery*, pages 385–396. Springer, 2002.

M. Vázsonyi. A szövegbányászat új irányzata: spektrális szövegbányászat. In *Adatbányászati alkalmazások perspektívái konferencia*, Veszprém, 2005.