Esettanulmány: böngészés támogatása kivonatolással kézi számítógépeken

A kivonatolás egyik speciális és kézenfekvő felhasználási területe a kisképernyős (kézi számítógép, PDA, mobiltelefon) tartalomszolgáltatás támogatása. A vezeték nélküli internet használata manapság egyre elterjedtebbé válik. A távolkeleten (Japán, Korea) az internethasználat jelentős része kisképernyős eszközökön keresztül történik, ám a kisméretű kijelzők gyakran akadályt jelentenek az internet kényelmes használatában (Jones , 1999), ugyanis a honlapok a kijelző méretéből adódóan többnyire nehezen áttekinthetőek. További problémát jelent az adatbevitel nehézkessége, valamint az a tény, hogy rádióhullámokkal a letöltési sebesség még mindig sokkal kisebb, mint vezetékes kapcsolat esetén.

Ezen problémák egy részére az egyik lehetséges megoldás az internetes tartalomszolgáltatás több lépésben, kivonatolással történő megvalósítása. A felhasználók ugyanis általában nem teljes internetoldalakra kíváncsiak, különösen PDA-n vagy mobiltelefonon böngészve, hanem az oldal azon részére, ahol a releváns információ megtalálható. Ezek többnyire tényszerű adatok vagy linkek.

A továbbiakban a Buyukkokten és munkatársai által javasolt megoldást ismertetjük (Buyukkokten et al, 2001), amely a weboldalakat fokozatosan, a felhasználó igényétől függően jeleníti meg. Ezzel a módszerrel jelentősen csökkenthető mind a letöltött adatmennyiség — s ezzel párhuzamosan a letöltési idő is —, mind pedig a keresett információ megtalálásához szükséges navigálási műveletek száma, valamint a böngészésre fordított idő.

Weboldalak kivonatolásának speciális kérdései

Első lépés az eredeti weboldal tartalmának feldarabolása ún. *szemantikus szövegegység*ekre. A feldarabolás az oldal szerkezetét követi, amely az oldal forrását (HTML, XML stb.) feldolgozva a tartalomból szövegegységek hierarchikus struktúráját állítja elő. A szövegegységek a weboldalt alkotó részegységek, pl. bekezdések, listák és elemeik, táblázatok, képek stb. Ezekből a szöveges módon megjeleníthető egységeket dolgozzuk fel a továbbiakban, a képeket, illetve a túl nagy méretű táblázatokat elhagyjuk.

A szövegegységek kivonatolása felvet néhány problémát. Mivel itt nem teljes dokumentumokra, hanem azok kisebb egységeire kívánunk kivonatolót alkalmazni, ezért nehezebb feladatot jelenthet a kulcsszavak, ill. -mondatok meghatározása, mivel a szövegegységek terjedelme jellemzően rövid. Másik különbség az, hogy a hagyományos kivonatoló módszerek statikusak — a kivonatot egy lé-

pésben állítják elő —, így nem támogatják a fokozatos megjelenítést. Szintén megfontolást igényel a hiperlinkek ábrázolása is (megjelenítés, aktivitás, hossz, fontosság a tartalmazó mondatra vonatkozóan).

Végül problémát okoz a kivonatolásnál használt statisztikák elkészítése, hiszen a legtöbb módszer szóelőfordulások és -gyakoriságok alapján határozza meg egy adott mondat elentőségét a szövegegységen belül. Mivel jelen esetben a dokumentumgyűjtemény az egész világháló tartalma, így ezekre a statisztikai adatokra csak becslések adhatók.

Szövegegységek fokozatos megjelenítésének alternatívái

A szövegegységek fokozatos megjelenítésére az alábbi megoldásokat tesztelték:

- **inkrementális:** három lépésben: egy sor, három sor, egész szövegegység;
- **összes:** rögtön az egész szövegegység megjelenik, nincs fokozatosság;
- **kulcsszó:** első lépésben a szövegegységben azonosított kulcsszavak jelennek meg, a következő fokozatban az első három sor, majd végül az egész szöveg látható lesz;
- **összegzés:** itt csak két lépcső van: a legfontosabb mondat, majd a teljes szöveg megjelenítése;
- **kulcsszó/összegzés:** ez az előző két módszer kombinációja, ahol először a kulcsszavak, majd a kiemelt mondat, végül az egész szöveg jelenik meg.

A hiperlinkek minden esetben aktívan megjelennek, kivéve a kulcsszavak fázist. Amennyiben egy link nem fejeződik be a sor végén, a látható fragmense akkor is aktív.

A kulcsszavak és az összegzés meghatározása

A kulcsszavak a szövegegységben szereplő egyes szavak kiértékelése alapján határozhatók meg. A tf-idf súlyozás kiszámolásához szükséges a korpuszban előforduló összes szó ismerete, ami itt nem áll rendelkezésre, ezért közelítő becslésekre van szükség. Ehhez internetes gyakorisági szótárat webrobot segítségével készíthetünk.

Egy szövegrészlet feldolgozása során minden szóra szótövesítést alkalmazunk, majd a szótár, illetve az adott weboldalon való előfordulási értékek alapján meghatározzuk a szóhoz tartozó tf-idf értéket. A szótárban nem szereplő szavak esetén a szótárban szereplő legkisebb gyakorisági értékkel számolhatnuk. A megadott kü-

szöbérték elérése esetén a szó a kulcsszavak közé kerül. Lehetőség van a speciális szedésű (félkövér, dőlt stb.) szavak erősebb súlyozására.

A kivonat meghatározására a könyv 7.4. szakaszában ismertetett bármelyik módszer alkalmazható. Buyukkokten és munkatársai egy nagyon egyszerű és könnyen implementálható, Luhn nevéhez fűződő módszer (Luhn, 1958) módosított verzióját használták.

A megjelenítő módszerek összehasonlítása

A fent ismertetett fokozatosan megjelenítő heurisztikákat egy 15 főből álló, internetes böngészésben jártas csapat segítségével tesztelték. Tíz tipikusan vezeték nélküli internetezés közben felmerülő feladatot tűztek ki a tesztelésre, pl. link megkeresése adott oldalon, nyitvatartási idő megkeresése, filmmel, tudományos konferenciával, ill. tanulmánnyal kapcsolatos adat, valamilyen termék árának és egyéb paraméterének meghatározása stb. — úgy, hogy a kiinduló oldalak adottak voltak. A teszt eredményei azt mutatták, hogy böngészési időt tekintve az összegzés, ill. kulcsszó/összegzés fokozatokat használó megjelenítési forma a legkézenfekvőbb a felhasználóknak, míg az inkrementális és az összes módszer a legkevésbé hatékony. A navigálási műveletek számát tekintve még erőteljesebb az említett két módszer dominanciája, esetenként akár 97%-kal csökkent az egér, ill. billentyűzethasználat mértéke. Itt egyértelműen a kombinált kulcsszó/összegzés módszer bizonyult a legjobbnak.

Vizsgálták még a letöltött adat mennyiségének csökkenési arányát. Az összegzés, kulcsszó és a kombinált módszerek esetén az alapértékként tekintett (HTML-elemektől, képektől és táblázatoktól mentes) adatmennyiséghez képest némi pluszt jelent, hogy a kulcsszavak, illetve az összegzés elejét és végét jelző indexértéket is továbbítani kell a rendszernek a protokollban az átvitel során. Ez azonban mindössze rendre 4%, 24%, ill. 28% volt. A letöltött adatmennyiség a „legdrágább” esetben is átlagosan 87%-kal kevesebbnek bizonyult, ami alátámasztja a kivonatoláson alapuló módszer hatékonyságát a kisképernyős böngészés támogatására.

Forrás:

O. Buyukkokten, H. Garcia-Molina, and A. Paepcke. Accordion summarization for end-game browsing on pdas and cellular phones. In *Proc. of the SIGCHI conference on Human factors in computing systems*, pages 213–220, Seattle, Washington, 2001.

O. Buyukkokten, H. Garcia-Molina, and A. Paepcke. Seeing the whole in parts: Text summarization for web browsing on handheld devices. In *The 10th Int. WWW Conf. (WWW10)*, pages 652–662, Hong Kong, China, 2001.

M. Jones, G. Marsden, N. Mohd-Nasir, K. Boone, and G. Buchanan. Improving web interaction on small displays. In *Proc. of 8th Int. WWW Conf. (WWW8)*, pages 51–59, Toronto, Canada, 1999. ACM, Bővített verzió

H. P. Luhn. The automatic creation of literature abstracts. *IBM J. of Research & Development*, 2(2):159–165, 1958.