Hibatűrő keresés digitalizált magyar nyelvű szövegekben

Pataki Máté, Füzessy Tamás, Kovács László, Tóth Zoltán

{mate.pataki, laszlo.kovacs, zoltan.toth}@sztaki.hu tfuzessy@freesoft.hu

MTA SZTAKI FreeSoft Nyrt.

Magyar Tudományos Akadémia

Számítástechnikai és Automatizálási Kutató Intézet

Elosztott Rendszerek Osztály

Kivonat

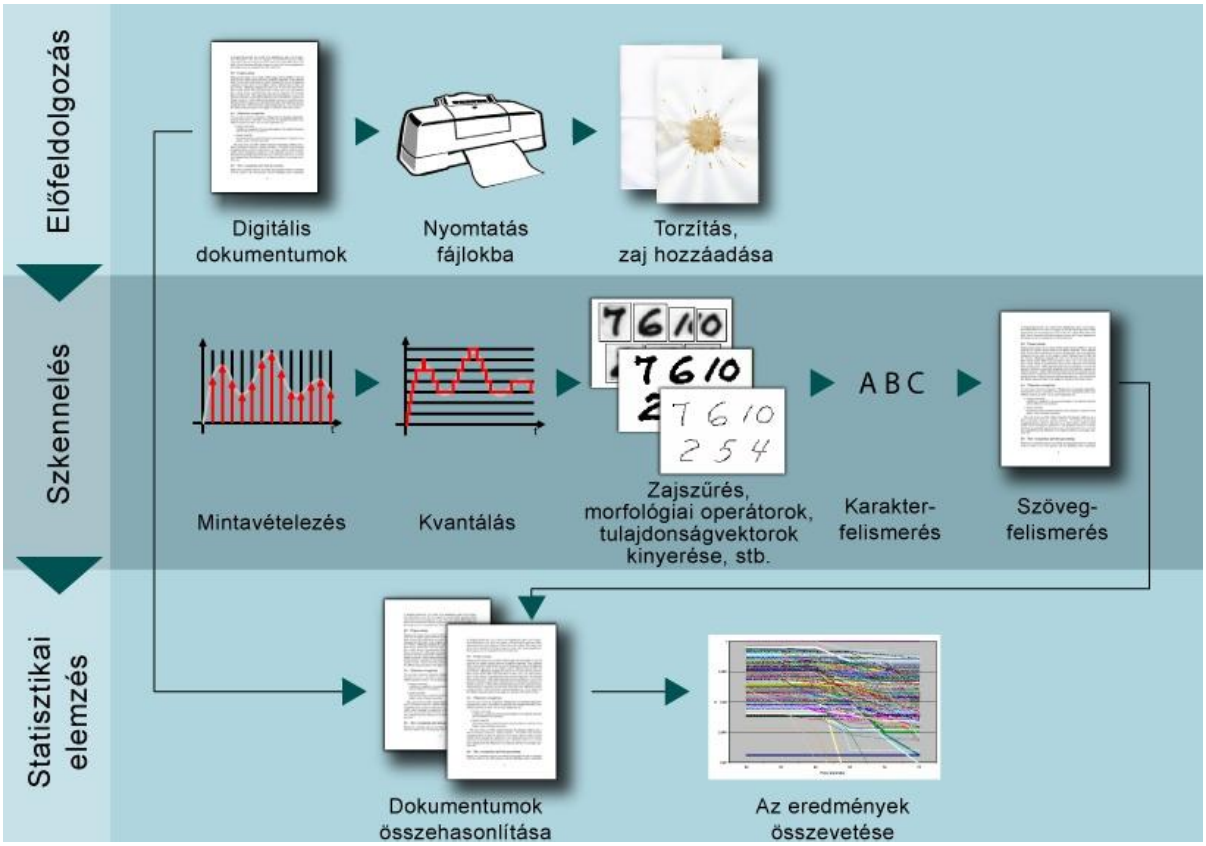
Az írásos emlékek megőrzése a jövő számára egy fontos feladata korunknak. A ma készülő művek jelentős része már digitális formában is tárolásra kerül, ugyanakkor fontos, hogy a régebben, csak nyomtatásban megjelent könyvekhez, írott anyagokhoz is hozzáférést biztosítsunk korunk digitális nemzedékének. Az így előállt adatbázis azonban használhatatlan a tárolt anyagokban történő kereshetőség megteremtése nélkül.

Az MTA SZTAKI Elosztott Rendszerek Osztálya egy GVOP pályázat keretében azt vizsgálta, hogy magyar nyelvű szkennelt szövegekben milyen hibák keletkeznek az eredeti dokumentumhoz képest. Ezeket az eredményeket felhasználva egy olyan kereső prototípusát építettük meg, amely egy digitalizálásból eredő hibákat tartalmazó adatbázisban is megbízhatóan tud keresni.

Szöveges dokumentumok digitalizálása

A szöveges dokumentumok digitalizálásának az alábbi öt fő lépése van:

- Mintavételezés (szkennelés)
- Kvantálás
- Előfeldolgozás
- Karakterfelismerés
- Szófelismerés, szövegfeldolgozás



A tesztkörnyezet felépítése

Digitalizálás során azt feltételezzük, hogy maga a digitalizált anyag zajjal terhelt. Ez a zaj eredhet a digitalizálandó dokumentumot ért sérülésekből, a papír típusából, a nyomtatás elégtelen minőségéből, stb. Sajnos a digitalizálás mind az 5 lépése – amellett, hogy alapvető céljuk ennek a zajnak a csökkentése – további „zajt", digitalizálási hibát visz a rendszerbe (alulmintavételezési hiba, kvantálási hiba, a zajszűrés során kiszűrt nem zaj jellegű információ, karakterfelismerési hiba, hibás szóelemzésből adódó hiba). Azzal együtt, hogy ez a hiba az egyre jobb digitalizáló eszközöknek és szoftvereknek hála egyre kevesebb, még mindig számolni kell vele, ha a kérdés a dokumentum kereshetősége.

Tesztkörnyezet a digitalizálás során fellépő hibák elemzéséhez

A karakterfelismerés tesztelése egy olyan tesztkörnyezetben történt, amely a lehető legjobban próbálta modellezni a dokumentumok valós életútját a digitalizálás során.

A tesztkörnyezet a következő elemekből állt:

1. Tesztadatok (nagymennyiségű magyar nyelven íródott, különféle formátumú: *rtf, txt, pdf, doc* dokumentum);

2. Digitalizáló szoftver, mely alkalmas különféle digitális képformátumok feldolgozására is;
3. Saját fejlesztésű eszközök a tesztek futtatására, az eredmények kiértékelésére.

Első lépésben a dokumentumok kinyomtatásra kerültek, annyi különbséggel a valós élethez képest, hogy digitális formátumba, veszteségmentesen tömörített TIFF képekbe történt a nyomtatás. Ezek után mesterséges zaj került a képekre, szimulálandó, hogy a dokumentumok meggyűrődnek, elkoszolódnak, a szkenner üvege koszos lehet stb. A hibákat a kinyomtatott képekhez véletlenszerűen adta hozzá a rendszer. Ezek a zajjal terhelt képek mentek a digitalizáló alkalmazásnak. A digitalizálás eredményéül kapott dokumentumokat összevetettük a bemenetként használt dokumentumokkal, és különféle statisztikai elemzéseket végeztünk rajtuk.

Először az összevetést viszonylag kis mennyiségű adaton kézzel végeztük, hogy különféle hibakategóriákat, hibatípusokat határozhassunk meg. Ezek után következhetett az automatikus elemzés a teljes adatbázison. A manuális összehasonlítás eredményeit felhasználva az automatikus elemzés során besoroltuk a digitalizálás során kapott digitalizálási hibákat, illetve különféle statisztikákat generáltunk.

Javasolt algoritmus

Egy intelligens keresőrendszer a szkennelt, hibával terhelt szövegben is tud keresni és megtalálja a szkennelés során keletkezett hibák ellenére is a keresett szöveget. Tapasztalataink alapján a következő pszeudo-kóddal leírt algoritmus egy ilyen intelligens keresőrendszert valósít meg.

```
min_results = 10
query = input
result = search(query, "phrase")
if count(result) < min_results {
  if count_words(query) > 1 {
    new_result = search(query, "all")
    result = append_results(result, new_result)
  }
  if count(result) < min_results {
    new_result = search_fuzzy(query, 70, "all")
    result = append_results(result, new_result)
    if count(result) < min_results {
      new_result = search_stemmed(query, "all")
      result = append_results(result, new_result)
      parameter = 60
      while count(result) < min_results AND parameter > 30 {
        new_result = search_fuzzy(query, parameter, "all")
        result = append_results(result, new_result)
        parameter = parameter - 10
      }
    }
  }
}
```

Ez a rendszer kihasználja a jelenlegi adatbáziskezelő rendszerek nyújtotta szolgáltatásokat, megpróbálja minimalizálni a keresés idejét és a lehető legvalószínűbb találatokat adja vissza.

Első lépésben egy teljes szavas keresés történik, majd, ha ez nem ad megfelelő eredményt a rendszer automatikusan teljes szöveges keresést végez. Amennyiben még így sincs elég találat, akkor egy enyhe fuzzy keresés történik, amely csak a nagyon hasonló szavakat adja vissza. Csak ezek után jön a sokkal lassabb szótöves keresés, illetve a sok hibás találatot eredményező, nagyobb különbségeket is megengedő fuzzy keresés.

Összefoglalás

A kutatás eredményeként pontos képet kaptunk a szövegfelismerés során fellépő hibák típusairól és előfordulási gyakoriságairól magyar nyelvű szövegek feldolgozása esetén. A kialakított univerzális tesztkörnyezet alkalmas lehet más szoftverek illetve nyelvek esetén is a hatékonyság és a hibák tesztelésére, működő rendszerek optimalizálására.

A kutatás eredményei beépítésre kerülnek a FreeSoft Rt. Contentum tartalomkezelési alkalmazáscsomagjába.

Hivatkozások

Meta-Contentum R&D project (FreeSoft, “A meta-contentum k+f projekt.”)

http://www.contentum.hu/hu/news/meta-contentum-kf

MTA SZTAKI Elosztott Rendszerek Osztály

http://dsd.sztaki.hu

Kapcsolat

Kovács László

MTA SZTAKI Elosztott Rendszerek Osztály

Tel: +36 1 279 6212

E-mail: laszlo.kovacs@sztaki.hu