Webes kereső

Feladat

A feladat egy webes keresőrendszer elkészítése, amely segítségével egy adott oldalon kereshetünk egy-egy kifejezésre. A Google-höz hasonló keresés a weboldalak indexelt adatbázisán történik, tehát szöveges alapú alkalmazás.
A keresés különböző típusú állományokban történik, ezek között vannak a HTML, az MS Word, az RTF, a PDF (pl. http://pdftohtml.sourceforge.net) formátumok. A kereséshez ezeket az állományokat fel kell dolgozni, a hozzájuk tartozó releváns információt tárolnunk kell. A keresésnél a beírt keresési szöveget előbb át kell alakítani olyan formába, hogy össze lehessen hasonlítani a tárolt dokumentumokkal. A keresési eredmény megjelenítésekor egy rangsoroló algoritmust is használunk, ez a relevánsabb találatokat jeleníti meg előbb; az algoritmus hasonló a Google PageRank algoritmusához [3]. A keresőrendszert alkalmazni is kell egy kisebb doménre. A Matematika és Informatika Kar honlapja például alkalmas ennek a rendszernek a tesztelésre.

A programozási nyelvet és a technikákat tekintve nincsenek megszorítások. Ez lehet PHP, Java, Perl, vagy bármilyen más programnyelv (pl. tetszőleges programnyelv ha CGI-programozási keretet használunk).

A feladat felbontása

Az alkalmazás főbb komponensei:

- indexelő
  - crawler
  - szövegfeldolgozó
  - invertált index elkészítése
  - rangsoroló
- kereső
- felhasználói felület

A legfontosabb és természetesen a legnagyobb komponens az indexelő. Először is egy crawler bejárja az oldalakat és lementi azokat (ez esetlegesen megoldható mentés nélkül is, ha minden bejárt oldal rögtön feldolgozásra kerül). A szövegfeldolgozó a különböző típusú dokumentumokból kinyeri a szöveges adatokat. Például HTML dokumentumok esetén ez a tag-ek, speciális szimbólumok elhagyását jelenti, és a szöveg helyes felbontását szavakra. A szavak a dokumentumok attribútumai, melyek alapján zajlik a keresés. Ezután a dokumentumokat (oldalakat) tárolnunk kell egy invertált index formájában, azaz a dokumentumokhoz tartozó szavakat illetve a szavaknak a gyakoriságát kell, hogy elmentsük. Az invertált indexet úgy kell elképzelni, mint egy szólista, ahol minden szóhoz tartozik egy lista, amelyben tároljuk, hogy az illető szó melyik dokumentumban hányszor fordult elő. Például:

```
computer      ->    {(0, 2), (5, 8), (5, 56)}
science       ->    {(0, 3), (4,78)}
math          ->    {(0,13)}
program       ->    {(3, 4), (3,109), (5, 57)}
read          ->    {(0,45), (1, 22), (2,104), (4, 2), (4, 54)}
```

A szavak listájában szereplő indexek „tartalmát” egy másik listában tároljuk, amely tartalmazza a dokumentum címét, és ezt a címet fogjuk megjeleníteni.
A releváns dokumentumok visszatérítése után egy rangsoroló (ranking) módszert is kell használnunk, amely megmondja, hogy az oldalakat milyen sorrendben kell megjeleníteni. Ez az

algoritmus súlyokat térít vissza, melyek alapján a visszatérített oldalakat megjelenítjük. A kereső tehát összehasonlítja a keresési szöveget (query) az indexelt dokumentumokkal, és a rangsoroló kimenetét használva sorrendbe jeleníti meg a találatokat.

A felhasználói felület egy olyan weboldal, ahol végrehajthatjuk a keresést.

Könyv/cikk/dokumentáció

[1] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008.
(http://www-csli.stanford.edu/~hinrich/information-retrieval-book.html)
[2] http://en.wikipedia.org/wiki/Index_(search_engine)
[3] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.