



# Információkeresés a weben és tanítása

Doktori (PhD) értekezés

**Krauszné Princz Mária**

Debreceni Egyetem  
Természettudományi Kar  
Debrecen, 2006



Ezen értekezést a Debreceni Egyetem TTK Matematika Doktori Iskola Informatika programja keretében készítettem a Debreceni Egyetem TTK doktori (Ph.D.) fokozatának elnyerése céljából.

Debrecen, 2006. szeptember 12.

Krauszné Princz Mária  
jelölt

Tanúsítom, hogy Krauszné Princz Mária doktorjelölt a fent megnevezett Doktori Iskola Informatika programja keretében irányításommal végezte munkáját. Az értekezésben foglalt eredményekhez a jelölt önálló alkotó tevékenységével meghatározóan hozzájárult. Az értekezés elfogadását javaslom.

Debrecen, 2006. szeptember 12.

Dr. Fazekas Gábor  
témavezető



# Tartalomjegyzék

|                                                                  |    |
|------------------------------------------------------------------|----|
| INFORMÁCIÓKERESÉS A WEBEN ÉS TANÍTÁSA .....                      | I  |
| 1. A TÉMAVÁLASZTÁS INDOKLÁSA .....                               | 1  |
| 2. A KUTATÁS CÉLJA .....                                         | 2  |
| 2.1. AZ INFORMÁCIÓ-VISSZAKERESÉS ELMÉLETÉNEK ÁTTEKINTÉSE .....   | 2  |
| 2.2. A HALLGATÓK INFORMÁCIÓ-VISSZAKERESÉSÉNEK FELMÉRÉSE .....    | 4  |
| 2.3. EGY PEDAGÓGIAI KÍSÉRLET A KERESÉSI KÉSZSÉG JAVÍTÁSÁRA ..... | 5  |
| 3. KUTATÁSI MÓDSZEREK.....                                       | 5  |
| 4. INFORMÁCIÓ-VISSZAKERESÉS .....                                | 6  |
| 4.1. BEVEZETÉS .....                                             | 6  |
| 4.2. ALAPFOGALMAK .....                                          | 6  |
| 4.2.1. Böngészés, keresés .....                                  | 6  |
| 4.2.2. Adat, információ .....                                    | 7  |
| 4.2.3. Srukturált és nem srukturált információ .....             | 8  |
| 4.2.4. Hasonlóság, mint a relevancia jellemzője .....            | 8  |
| 4.2.5. Teljesség és pontosság .....                              | 9  |
| 4.2.6. Pontos egyezés és legjobb egyezés.....                    | 9  |
| 4.3. INFORMÁCIÓ-VISSZAKERESŐ RENDSZEREK MŰKÖDÉSE .....           | 10 |
| 4.4. AZ INFORMÁCIÓ-VISSZAKERESŐ MODELL .....                     | 11 |
| 4.4.1. Boole modell.....                                         | 11 |
| 4.4.2. Vektortér modell .....                                    | 12 |
| 4.4.3. Valószínűségi modell.....                                 | 15 |
| 4.5. LEKÉRDEZÉSEK .....                                          | 17 |
| 4.5.1. Kulcsszón alapuló lekérdezések.....                       | 18 |
| 4.5.1.1. Boole-lekérdezések .....                                | 18 |
| 4.5.1.2. Szövegkörnyezet lekérdezése .....                       | 19 |
| 4.5.1.3. Természetes nyelvű lekérdezés.....                      | 19 |
| 4.5.2. Minta egyezés.....                                        | 20 |
| 4.5.3. Szerkezeti lekérdezések.....                              | 21 |
| 4.5.3.1. Rögzített struktúra .....                               | 21 |

|                                                        |           |
|--------------------------------------------------------|-----------|
| 4.5.3.2. Hipertext szerkezet.....                      | 21        |
| 4.5.3.3. Hierarchikus struktúra .....                  | 22        |
| 4.5.4. <i>A lekérdezésnél használt technikák</i> ..... | 22        |
| 4.5.4.1. A relevancia visszacsatolása .....            | 22        |
| 4.5.4.2. Klaszterezés .....                            | 23        |
| 4.6. DOKUMENTUMOK JELLEMZŐI .....                      | 23        |
| 4.6.1. <i>Leíró nyelvek</i> .....                      | 24        |
| 4.6.1.1. SGML.....                                     | 25        |
| 4.6.1.2. HTML .....                                    | 25        |
| 4.6.1.3. XML.....                                      | 27        |
| 4.6.1.4. Szemantikus web.....                          | 30        |
| 4.7. SZÖVEG MŰVELETEK .....                            | 31        |
| 4.8. INDEXELÉS ÉS KERESÉS .....                        | 32        |
| 4.8.1. <i>Invertált fájlok</i> .....                   | 33        |
| 4.8.2. <i>Szuffixek tömbje</i> .....                   | 34        |
| 4.8.3. <i>Szignatúra fájlok</i> .....                  | 36        |
| 4.8.4. <i>Szekvenciális keresés</i> .....              | 37        |
| 4.9. MULTIMÉDIA VISSZAKERESÉS.....                     | 38        |
| 4.9.1. <i>Az adatmodell</i> .....                      | 39        |
| 4.9.2. <i>Kulcsszón alapuló visszakeresés</i> .....    | 39        |
| 4.9.3. <i>Tartalom alapú visszakeresés</i> .....       | 40        |
| <b>5. INFORMÁCIÓ-VISSZAKERESÉS A WEBEN.....</b>        | <b>42</b> |
| 5.1. KIHÍVÁSOK .....                                   | 43        |
| 5.1.1. <i>Problémák az adattal</i> .....               | 43        |
| 5.1.1.1. Osztott adatok.....                           | 43        |
| 5.1.1.2. Változó adatok .....                          | 43        |
| 5.1.1.3. Nagy mennyiségű adat .....                    | 43        |
| 5.1.1.4. Strukturálatlan és redundáns adatok .....     | 45        |
| 5.1.1.5. Az adatok minősége .....                      | 46        |
| 5.1.1.6. Heterogén adatok .....                        | 46        |
| 5.1.2. <i>Problémák a felhasználókkal</i> .....        | 48        |
| 5.2. A WEB MEGOSZLÁSA .....                            | 48        |
| 5.2.1. <i>A látható web</i> .....                      | 50        |

|                                                                |           |
|----------------------------------------------------------------|-----------|
| 5.2.2. <i>A láthatatlan web</i> .....                          | 50        |
| 5.2.2.1. Nem átlátszó web .....                                | 50        |
| 5.2.2.2. Magán web .....                                       | 51        |
| 5.2.2.3. Szabadalmazott web .....                              | 52        |
| 5.2.2.4. Mély web .....                                        | 52        |
| 5.2.2.5. Az igazán láthatatlan web.....                        | 52        |
| 5.3. SZABADSZAVAS KERESŐK .....                                | 53        |
| 5.3.1. <i>Begyűjtés</i> .....                                  | 54        |
| 5.3.2. <i>Indexelés</i> .....                                  | 57        |
| 5.3.3. <i>Keresés</i> .....                                    | 58        |
| 5.3.3.1. Lekérdezések megadása .....                           | 58        |
| 5.3.3.2. Rangsorolás .....                                     | 60        |
| 5.3.3.2.1. A keresés irányultságának fogalmi felismerése ..... | 60        |
| 5.3.3.2.2. Hivatkozások analízise .....                        | 61        |
| 5.3.3.2.2.1. HITS algoritmus .....                             | 61        |
| 5.3.3.2.2.2. PageRank .....                                    | 62        |
| 5.3.3.2.2.3. HITS és PageRank összehasonlítása:.....           | 63        |
| 5.3.3.2.3. Az oldal népszerűségének elismerése .....           | 63        |
| 5.3.3.3. Eredmények megjelenítése .....                        | 64        |
| 5.3.4. <i>Népszerű szabadszavas keresők</i> .....              | 64        |
| 5.4. TEMATIKUS KERESŐK .....                                   | 65        |
| 5.5. METAKERESŐK .....                                         | 66        |
| 5.6. KERESÉS A HIPERHIVATKOZÁSOK ALAPJÁN.....                  | 67        |
| 5.6.1. <i>Web lekérdező nyelvek</i> .....                      | 67        |
| 5.6.2. <i>Dinamikus keresés és ügynök programok</i> .....      | 69        |
| 5.6.3. <i>Webbányászat</i> .....                               | 69        |
| <b>6. WEBOLDALAK OPTIMALIZÁLÁSA.....</b>                       | <b>71</b> |
| 6.1. A RANGSORT MEGHATÁROZÓ TÉNYEZŐK .....                     | 72        |
| 6.1.1. <i>Kulcsszavak</i> .....                                | 72        |
| 6.1.1.1. A weboldal szövege .....                              | 73        |
| 6.1.1.2. A dokumentum neve .....                               | 73        |
| 6.1.1.3. Meta jelölő elemek.....                               | 74        |
| 6.1.1.3.1. A dokumentum leírása .....                          | 74        |

|                                            |    |
|--------------------------------------------|----|
| 6.1.1.3.2. Meta kulcsszavak.....           | 74 |
| 6.1.1.3.3. Az ALT elem .....               | 75 |
| 6.1.1.4. Kulcsszó az URL-ben.....          | 75 |
| 6.1.2. Téma és kulcsszavak összhangja..... | 76 |
| 6.1.3. Hivatkozási népszerűség .....       | 76 |
| 6.1.4. Webhelyek népszerűsége.....         | 77 |
| 6.2. WEBHELYEK ELŐTERJESZTÉSE .....        | 78 |
| 6.3. A RANGSOR BEFOLYÁSOLÁSA .....         | 78 |

## **7. A HALLGATÓK INFORMÁCIÓ-VISSZAKERESÉSÉNEK FELMÉRÉSE . 80**

|                                                                                                                               |    |
|-------------------------------------------------------------------------------------------------------------------------------|----|
| 7.1. EREDMÉNYEK.....                                                                                                          | 82 |
| 7.1.1. A hallgatók mely Internet szolgáltatásokat és milyen sűrűn vesznek igénybe? .....                                      | 83 |
| 7.1.2. Tanulmányaikhoz tankönyveiken kívül mely információs forrásokat részesítik előnyben?<br>.....                          | 84 |
| 7.1.3. Hogyan vélekednek az Interneten található információk közötti keresési képességükről?.....                             | 85 |
| 7.1.4. Milyen Internet hozzáférési lehetőségek állnak a hallgatók rendelkezésére?.....                                        | 86 |
| 7.1.5. Honnan szerzik ismereteiket az Interneten található, a tanulmányaikhoz felhasználható<br>információk kereséséhez?..... | 86 |
| 7.1.6. A hallgatók milyen ismeretekkel rendelkeznek a különböző Internetes kereső<br>szolgáltatásokról? .....                 | 87 |
| 7.1.7. Milyen stratégiákat és technikákat használnak a webes információ-visszakeresésnél? ...                                 | 90 |
| 7.1.8. Mi befolyásolja a hallgatók keresési készségét? .....                                                                  | 93 |
| 7.1.8.1. Az Internethez való hozzáférés és a keresési készség kapcsolata.....                                                 | 94 |
| 7.1.8.2. Az informatika iránti érdeklődés és a keresési készség kapcsolata .....                                              | 95 |
| 7.1.8.3. A nemek és a keresési készség kapcsolata.....                                                                        | 95 |
| 7.1.8.4. Milyen eltérések vannak az egyes karokon tanulók keresési készségében? .....                                         | 95 |
| 7.1.9. Milyen felhasználói szokások jellemzik a hallgatókat a kereséseknél? .....                                             | 96 |
| 7.1.10. Hogyan vélekednek a keresők hatékonyságáról? .....                                                                    | 97 |

## **8. EGY PEDAGÓGIAI KÍSÉRLET A KERESÉSI KÉSZSÉG JAVÍTÁSÁRA..... 99**

|                                                 |     |
|-------------------------------------------------|-----|
| 8.1. AZ INFORMÁCIÓ-VISSZAKERESÉS TANÍTÁSA ..... | 99  |
| 8.2. A KÍSÉRLET .....                           | 100 |
| 8.3. DIDAKTIKAI ALAPOK .....                    | 101 |
| 8.4. KERESÉSI STRATÉGIÁK.....                   | 101 |

|                                                                                            |            |
|--------------------------------------------------------------------------------------------|------------|
| 8.4.1. <i>Tematikus keresők</i> .....                                                      | 102        |
| 8.4.1.1. A tematikus keresők használatának előnyei, hátrányai.....                         | 103        |
| 8.4.1.2. A jónak vélt URL cím beírása .....                                                | 104        |
| 8.4.2. <i>Szabadszavas keresők</i> .....                                                   | 104        |
| 8.4.2.1. Általános hibák.....                                                              | 106        |
| 8.4.2.2. A szabadszavas keresők használatának előnyei, hátrányai.....                      | 107        |
| 8.4.3. <i>Webes adatbázisok</i> .....                                                      | 108        |
| 8.5. A KÍSÉRLET LEÍRÁSA ÉS EREDMÉNYE .....                                                 | 109        |
| <b>9. ÖSSZEGZÉS .....</b>                                                                  | <b>111</b> |
| <b>10. IRODALOMJEGYZÉK.....</b>                                                            | <b>113</b> |
| <b>11. PUBLIKÁCIÓS JEGYZÉK.....</b>                                                        | <b>118</b> |
| <b>12. SUMMARY .....</b>                                                                   | <b>122</b> |
| 12.1. INTRODUCTION, VERIFICATION OF THE CHOICE OF THE TOPIC .....                          | 122        |
| 12.2. RESEARCH GOALS .....                                                                 | 123        |
| 12.2.1. <i>Review of the Theory of Information Retrieval</i> .....                         | 123        |
| 12.2.2. <i>A Survey on How Students Seek Information on the Internet</i> .....             | 125        |
| 12.2.3. <i>A Pedagogical Experiment to Improvement Searching Skill</i> .....               | 128        |
| 12.3. CONCLUSION .....                                                                     | 129        |
| <b>13. FÜGGELÉK .....</b>                                                                  | <b>I</b>   |
| 13.1. KÉRDŐÍV .....                                                                        | I          |
| 13.2. A HALLGATÓK KERESÉSI KÉSZSÉGÉT BEFOLYÁSOLÓ TÉNYEZŐK STATISZTIKAI<br>VIZSGÁLATA ..... | VI         |
| 13.2.1. <i>Faktoranalízis</i> .....                                                        | vii        |
| 13.2.2. <i>Homogenitás-vizsgálatok</i> .....                                               | xi         |
| 13.2.2.1. Az Internethez való hozzáférés és a keresési készség kapcsolata .....            | xii        |
| 13.2.2.2. Az informatika iránti érdeklődés és a keresési készség kapcsolata.....           | xv         |
| 13.2.2.3. A nemek és a keresési készség kapcsolata .....                                   | xvii       |
| 13.2.2.4. Milyen eltérések vannak az egyes karokon tanulók keresési készségében?.....      | xx         |



# 1. A témaválasztás indoklása

Napjaink tudásközpontú társadalma feltételezi, igényli és ösztönzi tagjainak folyamatos tanulását. A globális gazdaság hosszú távú, dinamikus fejlődésének feltétele, hogy az emberek mind nagyobb többsége hozzájusson a szellemi tőkéhez és felhasználja azt. A már elsajátított szellemi tőke hamar leértékelődik, a szaktudás, az ismeretek elavulnak. Ezért az egyén sikeressége nem a konkrét ismeretektől, a tudás mennyiségétől, hanem a gyarapítására való képességtől, vagyis mindennél inkább a tanulástól függ.

Az Európai Bizottság 2000 októberében **Egész életen át tartó tanulás** (Lifelong learning) [EC, 2000] címmel kiadott memoranduma hat kulcsüzenetben foglalta össze a probléma megközelítésével kapcsolatos legfontosabb tennivalókat. Az előtérbe helyezett új alapkészségek az információs technológiában, az idegen nyelvekben, a technikai kultúrában, a vállalkozásban és a szociális készségekben (úgy mint magabiztosság, önállóság és kockázatvállalás) való jártasság. További alapvető és kiemelt fontosságú készség a tanulás képessége, ennek elsajátítása pedig a változáshoz való alkalmazkodást, és az egyre kiterjedtebbé váló **információáramlásban való eligazodás** biztosítását jelenti.

Új módszerek, technikák segítik az egyént a tanulásban, s a legjelentősebb ezek közül a számítógép felhasználása, illetve az **Internet s a webes felületek biztosította lehetőség**. Az Internet használata egyre inkább tért hódít a mindennapi életben, így az oktatás területén is. Az e-learning rendszerek megtervezett, összeválogatott tudástartalmat közvetítenek a hallgatók felé formális, illetve nem formális oktatási keretek között. Természetesen megjelenhetnek a tananyaggal összefüggő kérdések, amelyekre a válaszokat nem tartalmazza a rendszer tudásbázisa, de a mindennapi életben is felbukkanhatnak olyan kérdések, amelyekre a választ önszántából, érdeklődési körének, előképzettségének megfelelően keresi az egyén. A válaszkeresésben a hozzáértő személy (tanár, könyvtáros, szakértő) mellett egyre gyakrabban fordulunk a weben át elérhető információk felé. Ezen óriási mennyiségű információban való eligazodás szükséges feltétele annak, hogy a web alapú tanulás sikeres lehessen.

## 2. A kutatás célja

A web hatalmas méretű információs tengernek tekinthető, amelyen való eligazodás korunk elvárása. A weben lévő információ keresésekor a leggyakrabban használt eszköz - a szabadszavas kereső - az információ-visszakereső rendszerek közül a legismertebb.

Kutatási tevékenységünk a weben található információk keresésével kapcsolatban három részből áll:

### 2.1. Az információ-visszakeresés elméletének áttekintése

Az információ-visszakeresés nagyon széles és tág tárgyterület, amely magában foglalja az információ tárolását, megjelenítését és visszakeresését is. Számos kutatási terület kapcsolódik hozzá: pl. algoritmusok, metaadatok használata, multimédia visszakeresés, mesterséges intelligencia, természetes nyelvi feldolgozás, stb.

Az információ-visszakeresés elméletéhez tartozó kutatási területek eredményeinek nemzetközi irodalma gazdag [Blackwell, Amazon], de a téma összefoglaló, rendszerező áttekintésére – a téma nagyságára való tekintettel – angol nyelven is csak néhány könyv áll rendelkezésre [Baeza-Yates&Ribeiro-Neto,1999], magyarul pedig nem érhető el összefoglaló mű [OSZK]. Értekezésünk első részében arra vállalkozunk, hogy e széles tárgyterületet felölelő, gazdag témát saját szempontjainknak megfelelően, a webről történő információ-visszakeresésre fókuszálva áttekintsük, és magyar nyelven is hozzáférhetővé tegyük.

A webes keresések legfőbb eszköze, a szabadszavas keresők kifejlesztése, üzemeltetése napjainkban egyre inkább üzleti jellegű vállalkozás, amely szabadalmazott technológiát is jelent egy erős versenypiacon. Ebből következik, hogy pontos működésükről annyit tudunk, amennyit a fejlesztők biztonságosnak tartanak közzétenni, s ez általában a technikai részletekről kevés publikációt

jelent. A tájékozódás legfőbb eszközei a weben át elérhető cikkek, konferencia előadások anyaga, a szakmai webportálokon át elérhető információ, illetve a próbálkozás és a következtetés.

- Dolgozatunk első részében az információ-visszakeresés alapfogalmait, az információ-visszakereső rendszerek működését, az információ-visszakereső modelleket, a dokumentumok jellemzőit ismertetjük. Bemutatjuk a dokumentumok formai és tartalmi leírására szolgáló leíró nyelveket. Áttekintjük a lekérdezések fajtáit, valamint bemutatunk néhány, a lekérdezésnél használt technikát. Példákon keresztül bemutatjuk a szöveges dokumentumok indexelésére és a bennük való keresésre kifejlesztett módszereket. Ismertetjük a multimédia objektumokban való keresés lehetőségeit.
- *Az információ-visszakeresés a weben* című fejezet elején a webről történő információ-visszakeresés problémáit, a web megoszlását tekintjük át. Ismeretes, hogy a szabadszavas keresők a webnek csak a kisebb részét térképezik fel, ezért szólni kell a láthatatlan web létezéséről, valamint e területen való információkeresés lehetőségeiről. Ismertetjük a webes keresések három alaptípusát: a kulcsszó, a téma és a hiperhivatkozások alapján végrehajtható kereséseket és eszközeiket. Áttekintjük a szabadszavas keresők, mint a webes keresések leggyakrabban használt eszközének felépítését, az általános és egyedi tulajdonságokat a begyűjtő, indexelő, lekérdező és rangsoroló részben. Ismertetjük az információs igény minél pontosabb megfogalmazását lehetővé tevő összetett és részletes kereséseket. Bemutatjuk a rangsorolások jobbítására létrehozott megoldásokat. Példát adunk a hiperhivatkozások alapján történő keresésekre, ismertetjük a webbányászat taxonómiáját.
- *A weboldalak optimalizálása* című fejezetben megvizsgáljuk, hogy a lekérdezéseknél mi befolyásolja az eredményül kapott dokumentumoknak a rangsorban elfoglalt helyét. Ha tudjuk, hogy a szabadszavas keresők mely tulajdonságok alapján szerepeltetik a keresés eredménylistáján előkelőbb helyen az egyes oldalakat, akkor a webre szánt oldalak optimalizálásával javíthatjuk az eredménylistán weboldalunk helyezését, s a jobb eredménnyel valószínűleg több emberhez jutnak el közlésre szánt

információnk. Ekképpen a szabadszavas keresők tulajdonságainak ismerete hasznosítható a webre szánt dokumentumok elkészítésénél is.

## **2.2. A hallgatók információ-visszakeresésének felmérése**

A weben át elérhető információk közötti eligazodás az információs társadalom tagjai számára egyre inkább elvárás, s különösen érvényes ez a megállapítás a ma felsőfokú tanulmányokat folytató hallgatókra, a jövő értelmiségére.

A keresési készség fogalma alatt a kereséskor alkalmazott ismeretek összességét értjük. A hallgatók weben történő információ keresési készségével kapcsolatban a következő kérdésekre keressük a választ:

1. A hallgatók mely Internet szolgáltatásokat és milyen sűrűn vesznek igénybe?
2. Mely információs forrásokat részesítenek előnyben?
3. Hogyan vélekednek az Interneten található információk közötti keresési készségükről?
4. Milyen Internet hozzáférési lehetőségek állnak a hallgatók rendelkezésére?
5. Honnan szerzik ismereteiket az Interneten található, a tanulmányaikhoz felhasználható információk kereséséhez?
6. Milyen ismeretekkel rendelkeznek a különböző Internetes kereső szolgáltatókról?
7. Milyen stratégiákat és technikákat használnak a weben át elérhető információ-visszakeresésnél?
8. Mi befolyásolja a hallgatók keresési készségét?
9. Milyen felhasználói szokások jellemzik a hallgatókat a kereséseknél?
10. Hogyan vélekednek a szabadszavas keresők hatékonyságáról?

## 2.3. Egy pedagógiai kísérlet a keresési készség javítására

Tapasztalataink szerint, az e tárgyban megjelent tanulmányokkal egyezően, a hallgatók weben történő információ-visszakeresésére az egy-két keresési kulcsszó megadása jellemző, s nem ismerik és használják a szükséges információ pontosítását lehetővé tevő, az eredményhalmazt szűkítő különböző lehetőségeket, illetve nem használnak alternatív keresési eszközöket. Felmérésünk egyik eredménye szerint a hallgatók keresési készségének kialakításánál, azaz a keresések tanulásánál a tanárok szerepe kicsi, a hallgatók döntően maguktól, illetve társaiktól szerzett információk alapján tanulják a webről az információ-visszakeresést.

Feltételezésünk szerint tanári irányítás mellett, a didaktikai elvárásoknak megfelelő keresési feladatok elvégzése, az alkalmazott stratégiák és technikák megbeszélése által a hallgatók keresési készsége jelentősen növelhető. Kísérletünk ezen állításunk igazolására szolgál.

## 3. Kutatási módszerek

Főbb kutatási módszereink a következők voltak:

- A szakirodalom követése, a publikált eredmények értelmezése.
- Pedagógiai megfigyelés és adatgyűjtés.
- Kérdőívek alkalmazása és értékelése. Statisztikai kiértékelések és elemzések.
- Részvétel és előadások tartása nemzetközi konferenciákon. A nemzetközi tapasztalatok egybevetése.

## 4. Információ-visszakeresés

### 4.1. Bevezetés

Az információ-visszakeresés az e-mail mellett a legáltalánosabb hálózati tevékenység, amelyet az Internet, az elektronikus katalógusok, adatbázisok és bármely típusú számítógépes gyűjtemény használója végrehajt.

Az információ-visszakeresésnek a tudományos és kulturális örökségben nagy jelentősége van, s különböző katalógusok használatával az ember már évszázadok óta gyakorolja. Az adat és információ-visszakeresésben igazi nagy változást az információs forradalom beköszöntekor a számítógépes feldolgozás teljesítménye eredményezett. Az információkeresésben a számítógép használatának két főbb körülhatárolható területe van: a visszakereséshez az *információ elemzése* (pl. mesterséges intelligencia használata az azonosításra) valamint *statisztikai feldolgozás* (pl. teljesen automatizált visszakeresési technikák).

A web robbanásszerű elterjedése óriási hatással volt az alkalmas információ-visszakeresési technikák keresletére és kínálatára. A szabadszavas keresők szolgáltatásait szinte mindenki használja, aki navigálni szeretne a weben.

Az információ-visszakereséshez tartozó tárgyterületek közül dolgozatunkban a visszakeresésre, mégpedig a szöveges, nem bibliográfiai tipológiájú visszakeresésre helyezzük a hangsúlyt, hiszen a webre – még multimédiás dokumentumokból való visszakeresés esetén is – napjainkban ez a jellemző.

### 4.2. Alapfogalmak

#### 4.2.1. Böngészés, keresés

A web tartalmának feltérképezésekor két különböző tevékenységet hajthatunk végre: *böngészhetünk* illetve *kereshetünk*. Mindkét tevékenység különböző módszert igényel. Böngészésnél a felhasználó manuálisan fedez fel bizonyos területeket (elektronikus könyvtárakat, hivatkozások gyűjteményét, web

oldalakat), miközben navigál köztük a remélt információ fellelése érdekében. A felhasználó dönt arról, mely utat járjon be, alkalmazkodva a könyvtárak struktúrájához, elnevezési rendszeréhez. A megtalálási folyamat lassúbb az automatikus kereséshez képest, de mivel a forrásanyagok tematikusan vannak összekapcsolva, ez a felhasználó számára lehetővé teszi, hogy egyszerre több jelentős anyagot is megtaláljon.

Keresésnél szükséges, hogy a felhasználó az információ igényét egy kérdésben fogalmazza meg, ahol a kérdés általában az információs igényt összegző kulcsszavak halmaza. A kérdés megadása után a számítógép a kérdésnek leginkább megfelelő dokumentumot előkeresi egy bizonyos információs halmazból.

Az információ-visszakeresés célja – szemben a böngészéssel – az információ azonnali elérése. Mind a keresés, mind a böngészés célja az információ megtalálása, de csak a keresési tevékenységnél használatosak az információ-visszakeresési technikák.

#### **4.2.2. Adat, információ**

Bár a laikusok gyakran használják e két fogalmat egymás szinonimájaként, fontos különbséget tenni köztük. Az adat egy objektum egy meghatározott változójának (attribútumának, jellemzőjének) értéke (karakterállapota, megvalósult formája). Az adat strukturálatlan tények halmaza, amely önmagában nem képvisel elválaszthatatlan jelentést. Amikor ezeket az adatokat összegyűjtjük, strukturáljuk, elemezzük és értelmezzük, akkor válik információvá. Az információ az adat felhasználható formája, amely valamilyen bizonytalanságot szüntet meg, új ismeretet jelent, valamilyen döntés, szabályozás alapját képezheti. Az adaton és az információn különböző módon lehet műveleteket végrehajtani, ennélfogva az adat-visszakeresés és az információ-visszakeresés is különbözik: más módszereket és technikákat használ, amint azt az alábbi táblázat tartalmazza:

|                         | Adat-visszakeresés | Információ-visszakeresés                                 |
|-------------------------|--------------------|----------------------------------------------------------|
| Egyezés                 | Pontos egyezés     | Részleges vagy legjobb egyezés                           |
| Modell                  | Determinisztikus   | Valószínűségeen alapuló                                  |
| Lekérdező nyelv         | Mesterséges        | Természetes (vagy a természeteshez a leginkább közelítő) |
| A kérdés megadása       | Teljes             | Hiányos                                                  |
| A visszakapott eredmény | Egyező             | Releváns                                                 |
| Hibára reagálás         | Érzékeny           | Nem érzékeny                                             |

1. táblázat. Adat és információ-visszakeresés összehasonlítása

#### 4.2.3. Szstrukturált és nem szstrukturált információ

Az adat és az információ is létezhet szstrukturált (pl. szabványos adatbázisok) és nem szstrukturált (pl. szabad szöveg) formákban, ami kihatással van a visszakeresés módjára. Adatbázisok lekérdezésénél felhasználjuk, hogy az adatbázis-kezelő ismeri az adatok szstrukturáját, s ez hatékonnyá teszi a visszakeresést. Azonban valamennyi információt szabványosan szstrukturált formába konvertálni időigényes, nehéz, és olyan szabványok alkalmazását igényli, mint a leíró nyelvek vagy a mező definíciók. Karaktorsorozatok összehasonlításának technikájával egyszerű, formázatlan szövegekből is lehetséges a visszakeresés, de más szstrukturálatlan média információkból (pl. kép, hang) a visszakeresés már összetettebb visszakeresési módszerek kifejlesztését igényli.

#### 4.2.4. Hasonlóság, mint a relevancia jellemzője

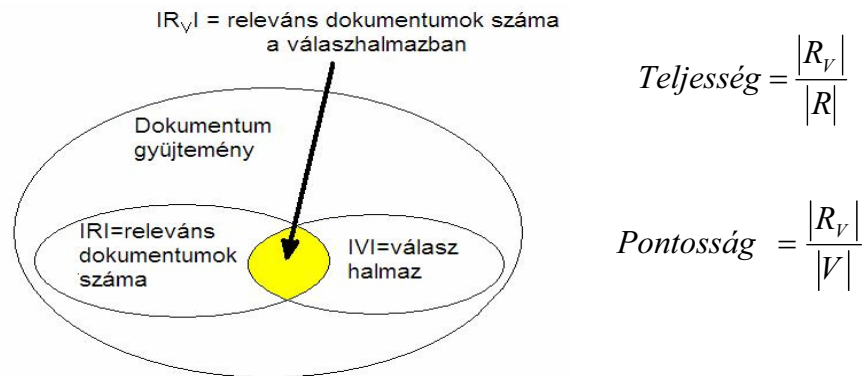
Az információ-visszakereső rendszerekben általános, hogy a hasonlóságot használják a tárgyhoz tartozás jellemzésére. Egy dokumentum relevanciáját a lekérdezéshez képest az adja meg, hogy mennyire hasonlít a lekérdezéshez: minél

nagyobb a hasonlóság, annál relevánsabb a dokumentum. Az ember általában könnyen el tudja dönteni, hogy egy dokumentum hasonlít-e a lekérdezésben megfogalmazotthoz, de a számítógép számára modellt kell megadni a hasonlóság definiálására és mérésére.

A hasonlóság megadására kidolgozott modellek attól függenek, hogy milyen típusú információ objektumot hasonlítanak össze. (Például szövegek hasonlóságát karaktersorozatok összehasonlításával lehet meghatározni, képek hasonlóságának vizsgálatánál használják a szín hisztogramot, stb.)

#### 4.2.5. Teljesség és pontosság

A teljesség (recall) fogalma alatt a releváns dokumentumok arányát értjük, amelyet a lekérdezés visszaad, míg a pontosság (precision) a lekérdezés eredményeként kapott dokumentumok azon részére utal, amelyek relevánsak a lekérdezéshez.



1. ábra. Teljesség, pontosság

A webes keresőszolgáltatások általában a teljességre törekszenek, s a pontosság háttérbe szorul.

#### 4.2.6. Pontos egyezés és legjobb egyezés

A pontos egyezés olyan visszakeresési modell, ahol a lekérdezést és a lekérdezett alaphalmaz elemeit összehasonlítva az eredmény igaz vagy hamis. A lekérdezés eredményeként az összehasonlításban igaz értéket kapó adatok, információk jelennek meg.

A legjobb vagy részleges egyezés azon a feltevésen alapul, hogy az információ-visszakeresés alapján bizonytalan rendszer. A bizonytalanságnak két oka is van:

- A felhasználók nem tudják pontosan megadni, mit is keresnek valójában.
- Az információ különböző típusainak tulajdonságait nehéz leírni (pl. kép, hang, videó), így visszakeresésük is bizonytalan.

A legjobb egyezés elve alapján visszakapott eredmény a lekérdezésnek nem feltétlen pontosan felel meg. Az információ-visszakereső rendszerek az eredményhez egy fontossági jelzést is kapcsolnak, s ez alapján rendezik az eredményül kapott dokumentumokat, mielőtt az eredménylistát a felhasználó számára megjelenítik. Minél jobban egyezik egy dokumentum a lekérdezéssel, annál előrébb található az eredménylistán.

Míg a pontos egyezés elve az adat visszakeresésekre jellemző, addig a legjobb egyezés elvét az információ-visszakereső rendszerek használják. A valóságban azonban a modellek kombinálhatók, így rövid lekérdezések esetén a legjobb egyezés elvét használó rendszereknél is a pontos egyezés az alapértelmezett.

### 4.3. Információ-visszakereső rendszerek működése

A hagyományos információ-visszakereső rendszerek indexkifejezéseket használnak arra, hogy névmutatóval ellássanak és visszakeressenek dokumentumokat. Az **indexkifejezések** olyan kulcsszavak (vagy szókapcsolatok), amelyek jelentéssel bírnak. Az információ-visszakeresés alapja az az elgondolás, hogy a dokumentumok jelentése és a felhasználók információ szükséglete is kifejezhető indexkifejezések halmazaként. A visszakeresésben a problémát az okozza, hogy a dokumentum vagy a felhasználó kérésének jelentéséből is sok elvész, amikor a dokumentum szövegét vagy a felhasználó információ igényét szavak halmazával helyettesítjük. Az információ-visszakereső rendszerek egyik feladata annak az eldöntése, hogy mely dokumentumok relevánsak egy lekérdezéshez, s melyek nem. Ennek megítélése a rangsoroló

algoritmustól függ, amely a visszakapott dokumentumok között megkísérel egy rangsort felállítani, s amely a visszakereső rendszerek lényeges része.

## 4.4. Az információ-visszakereső modell

Egy dokumentum indexkifejezéseinek halmazában megfigyelhető, hogy nem minden kifejezés bizonyul egyformán hasznosnak a dokumentum tartalmának összegzésére. A dokumentumok tartalmának leírására használt minden egyes indexkifejezéshez egy számértéket kapcsolunk, amely értéket a kifejezés súlyának nevezünk, s amely a kifejezés relevanciáját határozza meg.

**Definíció:** Jelölje  $k_i$  indexkifejezéshez és tetszőleges  $d_j$  dokumentumhoz tartozó súlyértéket  $w_{ij}$ . Ekkor  $w_{ij} \geq 0$ .  $w_{ij}=0$  azon indexkifejezések esetében igaz, amelyek nem fordulnak elő a dokumentum szövegében.

Az információ-visszakeresésben három klasszikus modellt különböztetünk meg: A Boole, a vektortér, és a valószínűségi modellt. Az *elméleti* modellnek is nevezett Boole modellben a dokumentumokat és a lekérdezéseket is indexkifejezések halmazaként tekintjük. A vektortér modellben a dokumentumokat és a lekérdezéseket is úgy vesszük, mint egy  $t$  dimenziós vektortér vektorait, s a modellt *algebrainak* mondjuk. A valószínűségi modellben a dokumentumok és a lekérdezések modellezése is a valószínűségi tételeken alapul, innen kapta a modell is a nevét.

### 4.4.1. Boole modell

A Boole modell a pontos egyezés elvét megvalósító legáltalánosabb modell. Egy  $q$  lekérdezés a megszokott Boole kifejezés, amelynek értéke vagy igaz, vagy hamis, így a  $q$  lekérdezés és  $d_j$  dokumentum hasonlósága

$$\text{sim}(d_j, q) = \begin{cases} 1, & \text{ha a lekérdezés minden kifejezése szerepel } d_j \text{ - ben} \\ 0 & \text{egyébként} \end{cases}$$

A Boole modell esetében az indexkifejezések súlya bináris, azaz  $w_{i,j} \in \{0,1\}$ .

A lekérdezés Boole kifejezéssel való megadása hatékony keresést eredményez. A Boole lekérdezések világos, egyértelmű nyelvét jól használhatja

minden olyan gyakorlott kereső, aki pontosan tudja, hogy mit keres. A gyakorlatlanok számára azonban problémát jelenthet a lekérdezés megfogalmazása, hiszen az alapoperátorok egyszerűsége ellenére bonyolult és komplex megfogalmazások is megadhatók.

A Boole kifejezéseknél fordított összefüggés van a teljesség és a pontosság között, s ez korlátozza a lekérdezésnek megfelelő valamennyi releváns dokumentum visszaadását, nagy dokumentum alaphalmaz esetén pedig megnehezíti egy alkalmas méretű eredményhalmaz előállítását. Boole lekérdezések eredménye mindig pontosan egyezik a lekérdezésben megfogalmazottakkal, bár valószínűtlen az eredményhalmaz 100%-os teljessége. Azonban kiszámíthatósága, jól kezelhetősége és eredményessége miatt a Boole lekérdezés népszerű és jól használható számos alkalmazásnál.

Noha egy Boole kereső alapján véve nem rangsorolja az eredményül kapott dokumentumokat relevanciájuk szerint (ehelyett egyszerűen kilistáz minden olyan dokumentumot, amely megfelel a lekérdezésben megfogalmazottaknak), de az eredmény rangsorolható és rendezhető, felhasználva egyéb tulajdonságokat. A leggyakrabban a lekérdezésben található kifejezések dokumentumbeli előfordulási gyakoriságát veszik figyelembe, az előfordulás helyét a dokumentumon belül (pl. a TITLE tagban, címsorokban, a dokumentum első bekezdésében, stb.), a kifejezések előfordulásának viszonylagos közelségét, a kifejezés szemantikai jelentőségét.

#### **4.4.2. Vektortér modell**

Míg a Boole lekérdezések a lekérdezésnek pontosan megfelelő dokumentumokat adnak vissza, addig a vektortér modell a lekérdezéshez legjobban hasonlító dokumentumokat eredményez, és egyben megadja a hasonlóság fokát a lekérdezés és az eredményhalmazként kapott dokumentumok között.

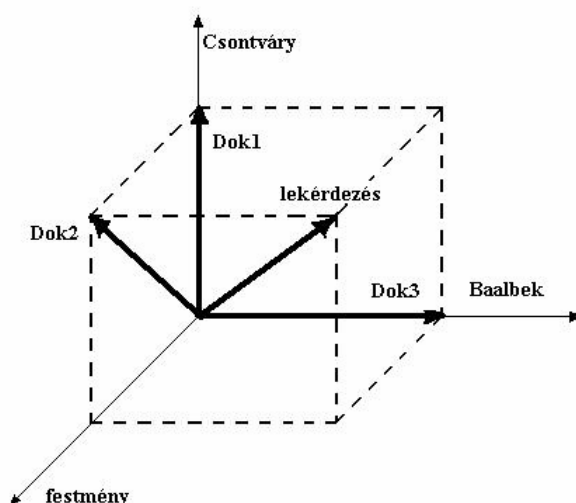
A vektortér modell a lekérdezés és a keresés eredményének hasonlóságát geometriai perspektívából közelíti. Ez azon a feltételezésen alapul, hogy mind a lekérdezés, mind egy tetszőleges dokumentum vektorként ábrázolható egy sokdimenziós térben, ahol a tér dimenziójának száma a lekérdezésben szereplő kifejezések számával egyenlő, s a lekérdezésben szereplő minden egyes kifejezés egy bázisvektornak (iránynak) felel meg.

A dokumentum vektora a bázisvektorok lineáris kombinációjaként áll elő, ahol a bázisvektor együtthatója 1, ha a kifejezés szerepel a lekérdezésben, és 0, ha nem. A dokumentum és a lekérdezés hasonlóságát leggyakrabban a két vektor közbezárt szögének koszinuszával adják meg, amely a két vektor skaláris szorzatából fejezhető ki.

Az alábbiakban 3 kifejezést tartalmazó lekérdezésre mutatunk példát:

Csontvárynak a Baalbek című festményével kapcsolatban keresünk dokumentumokat.

| Tartalmazza? | Dokumentum1 | Dokumentum2 | Dokumentum3 | Lekérdezés |
|--------------|-------------|-------------|-------------|------------|
| Csontváry    | igen        | igen        | nem         | igen       |
| Baalbek      | nem         | nem         | igen        | igen       |
| festmény     | nem         | igen        | nem         | igen       |



2. ábra. A 3 dimenziós vektortér modell

A háromdimenziós térben a Dokumentum2 és a Lekérdezés közti szög a legkisebb, így a lekérdezéshez leginkább a Dokumentum2 hasonlít. Ha egy dokumentum lekérdezésben szereplő összes kifejezést tartalmazza, akkor a dokumentum és a lekérdezés vektora egybeesik, a  $0^\circ$ -os közbe zárt szög pontos egyezést eredményez.

Lehetséges a kifejezések gyakoriságát és jelentőségét is bevezetni a vektortér modellbe, módosítva az értékeket valamely tengely mentén. A nem bináris súlyozással az eredményhalmaz minősége javítható. Tekintsük a  $q$  lekérdezést és egy tetszőleges  $d_j$  dokumentumot:

$$\vec{d}_j = (w_{1,j}, w_{2,j}, \dots, w_{t,j})$$

$$\vec{q} = (w_{1,q}, w_{2,q}, \dots, w_{t,q})$$

A dokumentum és a lekérdezés hasonlósága a dokumentum és a lekérdezés vektora által bezárt szög alapján adható meg.

$$\text{sim}(d_j, q) = \frac{\vec{d}_j \bullet \vec{q}}{|\vec{d}_j| \times |\vec{q}|} = \frac{\sqrt{\sum_{i=1}^t w_{ij} \times w_{iq}}}{\sqrt{\sum_{i=1}^t w_{ij}^2} \times \sqrt{\sum_{i=1}^t w_{iq}^2}}$$

$W_{ij}$  súlyok kiszámításánál az egyik legnépszerűbb módszer két faktor kombinációján alapul a következő módon:

Megadják a dokumentumban előforduló valamennyi index kifejezés fontosságát. Ez azon számmal mérhető, hogy a kifejezés hány alkalommal fordul elő a dokumentumban. Ezt nevezik a kifejezés gyakoriságának (*term frequency*), és a  $tf$  szimbólummal jelölik.

Megadják az index kifejezés fontosságát a gyűjteményben. Egy index kifejezés, amely a gyűjtemény valamennyi dokumentumában előfordul, nem túl hasznos, azonban ha csak néhány dokumentumban található meg, az azt jelezheti, hogy az a néhány dokumentum releváns lehet olyan lekérdezéseknél, amelyek tartalmazzák az adott kifejezést. A gyűjteményben az index kifejezés fontossága a gyűjtemény dokumentumai közötti előfordulás gyakoriságának az inverzével fejezhető ki. Ezt a faktort inverz dokumentum gyakoriságnak (*inverse document frequency*) nevezik és jelölésére az  $idf$  szimbólumot használják. Matematikailag a következő módon fejezhető ki:

$$idf_i = \log \frac{N}{n_i}$$

, ahol

$N$  = a gyűjteményben található dokumentumok száma,

$n_i$  = azon dokumentumok száma, amelyek tartalmazzák a lekérdezés  $i$ -edik kifejezését.

A lekérdezés  $i$ -edik index kifejezésének súlya a két faktor segítségével a

$$w_{i,j} = tf \times idf = tf \times \log \frac{N}{n_i}$$

módon adható meg. Ezt a súlyozást a ***tf-idf*** súlysémának nevezik.

### 4.4.3. Valószínűségi modell

A valószínűségi alapuló visszakeresés bizonyos paraméterek valószínűségi becslésén alapul. A valószínűségi modellek feltételezik, hogy a keresési szavak, kifejezések releváns és nem releváns dokumentumokban fordulnak elő, és megpróbálják azt megbecsülni, hogy egy dokumentum egy lekérdezéshez viszonyítva releváns-e vagy sem, valamint a relevancia visszajelzése révén megpróbálják a releváns dokumentumok számát megnövelni.

A lekérdezés indexkifejezések részhalmazaként tekinthető. E modellnél a dokumentumok és a lekérdezések súlyai is binárisak:  $w_{i,j} \in \{0,1\}$ ,  $w_{i,q} \in \{0,1\}$ .

Egy  $d_j$  dokumentumnak a  $q$  lekérdezéshez való hasonlóságát a következő valószínűségek hányadosa adja meg:

$$sim(d_j, q) = P(d_j \text{ releváns } q\text{-hoz}) / P(d_j \text{ nem releváns } q\text{-hoz}).$$

A dokumentum és a lekérdezés hasonlósága a súlyokkal kifejezve [Baeza-Yates&Ribeiro-Neto,1999]:

$$sim(d_j, q) \approx \sum_{i=1}^t w_{i,q} \times w_{i,j} \times \left( \log \frac{P(k_i|R)}{1 - P(k_i|R)} + \log \frac{1 - P(k_i|\bar{R})}{P(k_i|\bar{R})} \right)$$

,ahol  $R$  jelöli azon dokumentumok halmazát, amelyek relevánsak  $q$  lekérdezéshez,  $\bar{R}$  megegyezik  $R$  komplementerével (a nem releváns dokumentumok halmaza),  $P(k_i|R)$  feltételes valószínűség pedig annak a valószínűsége, hogy  $k_i$  indexkifejezést tartalmazza egy  $R$ -ből véletlenszerűen kiválasztott dokumentum.

|                             | <b>Előny</b>                                                                                                                                                                                                                         | <b>Hátrány</b>                                                                                                                                                                                                                                                                                                    |
|-----------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Boole modell</b>         | <ul style="list-style-type: none"> <li>▪ Tiszta formalizmus;</li> <li>▪ Könnyű megvalósítani.</li> </ul>                                                                                                                             | <ul style="list-style-type: none"> <li>▪ A pontos egyezés túl sok vagy túl kevés dokumentumot eredményez;</li> <li>▪ Az információs igény Boole kifejezéskénti megfogalmazása nehézséget okozhat;</li> <li>▪ Nem szolgáltatja a kifejezések súlyozását;</li> <li>▪ Nem rangsorolja az eredményhalmazt.</li> </ul> |
| <b>Vektortér modell</b>     | <ul style="list-style-type: none"> <li>▪ A szavak súlyozása javíthatja az eredmény minőségét;</li> <li>▪ Megengedi a részleges egyezéseket;</li> <li>▪ Az eredményül kapott dokumentumokat relevanciájuk szerint rendezi.</li> </ul> | <ul style="list-style-type: none"> <li>▪ Feltételezi a szavak egymástól való függetlenségét. Ez néhány esetben nehézséget okoz a végrehajtásnál.</li> </ul>                                                                                                                                                       |
| <b>Valószínűségi modell</b> | <ul style="list-style-type: none"> <li>▪ Az eredményhalmazt a dokumentumok relevanciájának valószínűsége szerint rangsorolja.</li> </ul>                                                                                             | <ul style="list-style-type: none"> <li>▪ A lekérdezést megelőzően szükséges a dokumentumok szétválasztása releváns és nem releváns halmazra;</li> <li>▪ A módszer nem veszi figyelembe a keresési kulcsszavak előfordulásának gyakoriságát;</li> <li>▪ Függetlenséget feltételez.</li> </ul>                      |

**2. táblázat. A három klasszikus modell összehasonlítása**

A három modell közül a Boole modellt tartják a leggyengébbnek. Különböznek a vélemények abban, hogy a vektortér modell a valószínűségi modellnek egy speciális esete, vagy sem, illetve a kettő közül melyik szolgáltat

jobb eredményt. [Baeza-Yates&Ribeiro-Neto,1999, Salton&Buckley,1990, Croft&Harper,1979]

Mindhárom klasszikus modellben különböző alternatívákat dolgoztak ki az alapmodell erősítésére. A Boole modell alternatívái a fuzzy halmazok modellje és a kiterjesztett Boole modell. A vektortér modell alternatívái az általánosított vektortér modell, a szemantikus indexelés és a neurális hálók modellje. A valószínűségi modell alternatívája a Bayesian hálók.

## 4.5. Lekérdezések

A felhasználó információs igényét lekérdezésben fogalmazza meg. Az lekérdezés alapvetően öt különböző módon adható meg [Schneiderman,1997]: parancs nyelven, adatbeviteli mezők kitöltésével, menü választásával, közvetlen manipulációval és természetes nyelven. A felhasználó által megadható lekérdezés típusa nagymértékben az alkalmazott visszakeresési modelltől függ.

A legtöbb lekérdező nyelv megpróbálja felhasználni a szöveg tartalmát (szemantika) és struktúráját (a szöveg szintakszisa). Számos technikai elemet használnak a lekérdezés eredményességének növelésére: *szinonimaszótár*, *szóelemzés*, *stop-szavak*.

A szinonimaszótár használatával nemcsak az indexek mérete csökkenthető, de a lekérdezések is kibővíthetők a különböző formájú, de tartalmilag azonos jelentésű szavakkal, s ez a keresés eredményeként relevánsabb dokumentumokat eredményezhet.

A szóelemzés alkalmazásával a szó különböző ragozott és képzett alakjai visszavezethetők az eredeti szótőre. A szótövek megkeresésével (stemming) szintén csökkenthető a szükséges index mérete, és növelhető a keresés teljessége.

Egy gyűjteményen belül a túl gyakran előforduló szavak általában nem relevánsak a gyűjteményre nézve (pl. névelők, kötőszavak). Ezen szavak indexelése felesleges, ezért az index nyelvből legtöbbször kizárják őket. Ennek módja az ún. stop-szavak (stopwords) megadása, amely szavakat a keresők nem indexelik.

### 4.5.1. Kulcsszón alapuló lekérdezések

A lekérdezések legegyszerűbb formája egy vagy több kulcsszó felsorolásából áll, s azon dokumentumokat adja eredményül, amelyek a felsorolt kulcsszavakat tartalmazzák. A szavak keresési mintán belüli elhelyezkedése indifferens. A kulcsszón alapuló lekérdezések népszerűek, mert ösztönösen alkalmazhatóak, könnyen megadhatók, és lehetővé teszik az eredmények gyors rangsorolását.

#### 4.5.1.1. Boole-lekérdezések

A lekérdezések megfogalmazása a Boole-algebra szerint logikai kifejezések megadásával történik, ahol az operandusok a keresett dokumentumot leíró kulcsszavak vagy kifejezések, amelyeket az AND (és), az OR (vagy) és a BUT (kivéve) logikai operátorokkal (Boole-operátorokkal) kapcsolhatunk össze logikai kifejezésekké.

Tekintsük  $k_1$  és  $k_2$  kulcsszavakat vagy kifejezéseket:

- OR  
( $k_1$  OR  $k_2$ ) lekérdezés kiválasztja valamennyi dokumentumot, amely  $k_1$  vagy  $k_2$  lekérdezésnek megfelel. A mindkét lekérdezésnek megfelelő dokumentumokat csak egyszer szerepelteti az eredmények között.
- AND  
( $k_1$  AND  $k_2$ ) lekérdezés valamennyi dokumentumot kiválaszt, amely  $k_1$  és  $k_2$  lekérdezésnek megfelel.
- BUT  
( $k_1$  BUT  $k_2$ ) lekérdezés valamennyi dokumentumot kiválaszt, amely a  $k_1$  lekérdezést kielégíti, de  $k_2$  lekérdezésnek nem felel meg. A klasszikus Boole algebra a NOT operátort használja, ahol (NOT  $k_2$ ) akkor igaz, amikor  $k_2$  nem teljesül. Ebben az esetben a  $k_2$  lekérdezésnek nem eleget tevő dokumentumok nagy tömegét kapná vissza a felhasználó. E helyett a BUT operátor eredményül a  $k_1$  kérdésnek megfelelő dokumentumokra szorítkozik, amelyek közül kizárja a  $k_2$  lekérdezésnek is megfelelő dokumentumokat.

A Boole lekérdezés legnagyobb hátránya, hogy az emberek többsége számára nehezen érthető, nehezen alkalmazható.

#### 4.5.1.2. Szövegkörnyezet lekérdezése

Számos rendszernél az egyszerű kulcsszó megadáson kívül lehetőség van a szó szövegkörnyezetben történő keresésére is, azaz megadható, hogy a keresett szó mely szó (szavak) mellett tűnjön fel a dokumentumban. Két formáját különböztetjük meg: a *kifejezések* keresését, valamint a szó keresését más *szó szomszédságában*.

A szavak távolsága rendszertől függően karakterekben illetve szavakban mérhető. A szavak elhelyezkedésének leírására az egyes rendszerek a NEAR (közeli), FAR (távoli), vagy ADJ (rákövetkező) operátorokat használják.

#### 4.5.1.3. Természetes nyelvű lekérdezés

A természetes nyelvű lekérdezések a természetes nyelv szabályainak megfelelő keresési mintát alkalmaznak (például kérdéseket lehet feltenni). A rendszer a lekérdezést nemcsak a lekérdezésben megadott szavak felsorolásaként tekinti, hanem előredefiniált szabályok alapján megpróbálja a szavak közötti kapcsolatot is felismerni. A számítógép a mondatokat szemantikai részekre bontja (pl. főnév, ige, melléknév), és meghatározza a részek közötti kapcsolatot. Miután a nyelv lehet kétértelmű, jelképes, a rendszer megpróbálja a szó jelentésének bizonytalanságát megszüntetni oly módon, hogy egy-egy szót a környezetében előforduló szavakkal egybevetve vizsgál. A lekérdezést szabállyá alakítva, a számítógép feltevéseket gyárt, amelyek a logikán alapulnak, és eredményül visszaad minden olyan dokumentumot, amely részben egyezik a lekérdezésben megadottakkal. A rangsorolásnál előbb szerepelnek azok a dokumentumok, amelyek a lekérdezés nagyobb részével egyeznek meg.

Szöveges állományok természetes nyelvi analízise még nem alkalmazott széles körben. Ennek oka, hogy ez a technika még nem elég hatékony, viszont elég költséges, különösen nagy adatállományok esetén. A legtöbb esetben csak jól strukturált szövegnél, tezaurusz alkalmazása mellett hatékony.

A természetes nyelvű feldolgozás (Natural Language Processing, NLP) a mesterséges intelligencia alkalmazásának egyik területe.

## 4.5.2. Minta egyezés

A kulcsszón alapuló lekérdezésekhez képest összetettebb lekérdezések adhatók meg a minta egyezésen alapuló lekérdezések használatával, amelyek nagyon hasznosak a nyelvészet, a szöveg statisztikák, s a különböző adat kivonatolás (extrakció) területén. A minta mondattani tulajdonságokkal bíró szövegrész, amelynek előfordulását keressük a szövegben. A visszakereső rendszerek – rendszerenként változóan - a minták néhány típusának megadását engedélyezik.

A leggyakrabban használt minta típusok:

**Szavak:** Karakterek sorozata, a legáltalánosabb minta.

**Prefixek:** Egy szó elejétől számított karaktersorozat. Pl. a 'szám' prefix minden olyan dokumentumot visszaad, amelyben szerepel a 'számítógép', 'számol', 'számegyenes' szó.

**Suffixek:** Egy szó végéig tartó karaktersorozat. Pl. az 'ász' suffix minden olyan dokumentumot visszaad, amelyben szerepel a 'vadász', 'kalász', 'madarász' szó.

**Részszttring:** Egy karaktersorozat a szöveg szavain, vagy a szövegen belül. Pl. az 'int' részszttring előfordul az 'integet', 'hint', 'hintaló' szavakban.

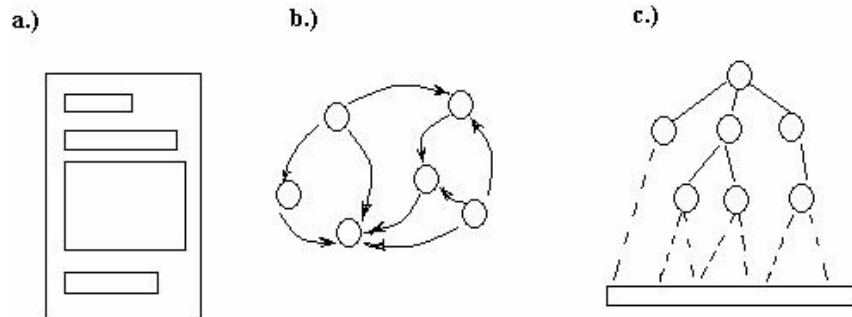
**Terjedelem:** Egy sztringpárral megadható szavak összessége, amelybe minden olyan szó beletartozik, amely lexikografikusan az adott két sztring közé esik. Pl. a 'hat' és 'hét' szavak által adott terjedelemben esnek a 'három', 'havas', 'hegy' szavak is.

**Megengedett hibák:** Az elv azon alapul, hogy mind a minta, mind a szöveg tartalmazhat elgépelésből, helyesírásból adódó hibákat. Ez a fajta keresés visszaad minden olyan szót, amely hasonló az eredeti szóhoz. A hasonlóságot a *Levenshtein távolsággal* vagy *szerkesztési távolsággal* adjuk meg. Két karaktersorozat szerkesztési távolsága azon karakterek minimális számával egyezik meg, amelyeket beszúrva, törölve vagy helyettesítve a két karaktersorozat egyenlővé alakítható.

**Reguláris kifejezések:** Néhány szöveges visszakereső rendszer megengedi karaktersorozatok unióját, konkatenációját, vagy egy karaktersorozat adott számú ismétlését.

### 4.5.3. Szerkezeti lekérdezések

Ez a fajta visszakeresés lehetővé teszi, hogy egy szövegben nemcsak a tartalom, hanem a szövegben lévő struktúra alapján is kereshessünk, s ezzel a lekérdezés hatékonyságát növeljük. A szövegben lévő struktúrákat szemlélteti a 3. ábra.



3. ábra. A szövegben lévő szerkezetek: a.) rögzített struktúra, b.) hipertext szerkezet, c.) hierarchikus struktúra

#### 4.5.3.1. Rögzített struktúra

A legtöbb forgalomban lévő rendszer e modell szerint működik. A dokumentum mezők rögzített halmazát tartalmazza, hasonlóan egy kitöltött űrlaphoz. Minden mező tartalmaz szöveget, de nem minden mező van jelen minden dokumentumban. A visszakeresésnél a mezők tartalmára kérdezzük rá.

Rögzített struktúrát használnak például a mail archívumok, ahol minden egyes mail tartalmazza a feladót, címzettet, dátumot, tárgyat és a szövegtörzs mezőt. Szöveges adatállományok form-szerű rögzített struktúrája relációs adatmodellhez hasonlítható, ahol a lekérdezés az SQL nyelv alkalmazásával történik.

#### 4.5.3.2. Hipertext szerkezet

A hipertext egy irányított gráfnak tekinthető, ahol a csomópontok szöveget tartalmaznak, és a hivatkozások fejezik ki a kapcsolatot a csomópontok vagy a csomóponton belüli pozíciók között. Az információ visszakeresése a hipertextből böngészés útján, a hipertext csomópontjainak áttekintésével, manuálisan történik.

### 4.5.3.3. Hierarchikus struktúra

A rögzített struktúra és a hipertext struktúra közötti köztes modell a hierarchikus struktúra. Ez a modell a szöveg rekurzív felbontásának tekinthető, és számos szöveggyűjteménynek (pl. könyvek, cikkek, strukturált programok) természetes modellje. A hierarchikus modellben a struktúra jelölése a szövegben tag-ekkel történik (hasonlóan a HTML nyelvhez). A hierarchikus struktúra alkalmazására példa a PAT szöveges adatbázis-kezelő rendszer.

### 4.5.4. A lekérdezésnél használt technikák

Az információ-visszakereső rendszerek a felhasználó információs igényének minél jobb megértése érdekében alkalmazzák az alábbi technikákat:

#### 4.5.4.1. A relevancia visszacsatolása

A relevancia visszajelzése számos információ-visszakeresési modell fontos része. Az eredményhalmaz emberi értékelése és a modell számára történő visszajelzése az információ-visszakereső rendszerek egyik alapproblémája, hiszen a relevancia szubjektív, jelentés, feladat és felhasználó függő. Ez gátolja, hogy egy automatizált folyamattal el lehessen dönteni, vajon a visszakapott eredmény releváns-e.

A relevancia visszajelzésénél alapvetően két technika használatos: egyrészt *a lekérdezés módosítása*, másrészt *a kulcsszavak újrasúlyozása*.

A lekérdezés módosítása a leggyakrabban a lekérdezés kibővítését jelenti, amely a releváns dokumentumokból vett újabb kulcsszavak hozzáadásával valósul meg. A lekérdezés kibővítésére a klaszterezés egy alapvető technika.

A relevancia visszajelzésére használatos másik technika a kulcsszavak újrasúlyozása a felhasználó ítélete alapján. A változtatás a felhasználó értékelése alapján megy végbe: mennyire elégedett a rendszerrel, a szolgáltatott eredmények mennyire relevánsak az információs igényéhez.

A rendszer működésének javítása végett az egyes modelleken belül is számos paraméter változtatása lehetséges. A visszacsatolás a teljesítmény növelése érdekében a felhasználó megelégedéséig folytatható ugyanarra a keresési mintára és dokumentum gyűjteményre.

A relevancia visszajelzése különösen a tartalom alapú visszakereső rendszereknél fontos, ahol a relevanciát gyakran sokkal nehezebb definiálni, mint a karaktersorozatok egyszerű egyezését.

#### **4.5.4.2. Klaszterezés**

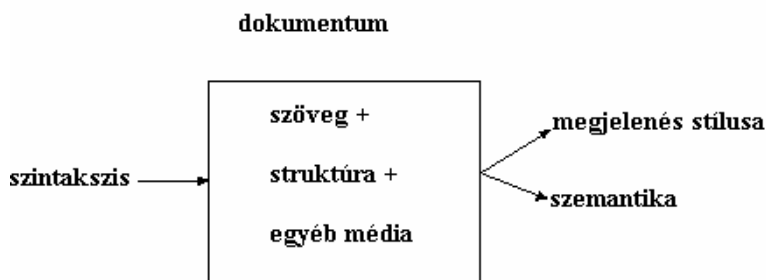
A klaszterezés egy olyan művelet, amely a hasonló, vagy az egymással összefüggő dokumentumokat egy osztályba sorolja. Célja a relevancia visszajelzésének elősegítése. A klaszterezésnek két fő típusa van: globális és lokális.

Globális klaszterezésnél a dokumentumokat a gyűjtemény egészében elfoglalt helyük szerint osztályozzuk egy tezaurusz-szerű struktúrába, amely mutatja a fogalmak kapcsolatát. Ezt a struktúrát lehet megmutatni a felhasználónak, amikor a lekérdezés kibővítésekor újabb keresési kulcsszót választ.

Lokális klaszterezésnél a rendszer automatikusan, az aktuális lekérdezésnek megfelelően csoportosítja a dokumentumokat (pl. szavak jelentése alapján tömörít). A klaszterezés célja olyan általános struktúrákat létrehozni (pl. asszociációs mátrixok), amelyekkel mérhetőek a kifejezések korrelációja, s a korrelált kifejezéseket felhasználni a lekérdezés kibővítésére. E megközelítéssel az a fő probléma az, hogy nem következetesen bizonyított, az általános struktúrák ténylegesen javítják-e a visszakeresés végrehajtását egy általános gyűjteményben [Attar&Fraenkel,1977].

### **4.6. Dokumentumok jellemzői**

Az információ-visszakeresés folyamán különböző típusú, formátumú dokumentum között kereshetjük a szükséges információt, de a leggyakrabban használt formátum a szöveg.



4. ábra. A dokumentum jellemzői

Egy dokumentum adott stílussal, azaz szintaktikával és struktúrával rendelkezik, amely a dokumentumot generáló program vagy a dokumentumot létrehozó személy által meghatározott. A dokumentumok adatokat tartalmazhatnak a bennük fellelhető adatok szervezéséről, típusairól, a köztük lévő kapcsolatokról, megkönnyítve a katalogizálást, az információ-visszakeresést. A *metaadatoknak* (adatok az adatokról) alapján két formáját különböztethetjük meg: a leíró jellegű, valamint a szemantikus jelentéssel bíró metaadatokat.

A legáltalánosabb leíró jellegű metaadat, amikor a szöveghez hozzákapsoljuk a szerző nevét, a kiadás dátumát, a dokumentum forrását (könyv, cikk) stb. Erre példa a Dublin Core Metaadat Kezdeményezés, amelyet a könyvtárak igényeinek megfelelően hoztak létre 1995-ben. A Dublin Core 15 meta leíróelemet tartalmazó készlettel (pl. kulcsszó, leírás, szerző, kiadó, forrás, azonosító, szerzői jogok, stb.) adja meg valamely dokumentum bibliografikus leírását, és ezáltal kísérli meg feljavítani a weben való kereshetőségét.

Számos folyóirat (pl. MEDLINE) a megjelentetett cikkekhez a témakör leírását tartalmazó kulcskifejezéseket kapcsol hozzá. A szemantikus kifejezések szabványosítása végett számos területen használnak speciális *ontológiákat*, amelyek egy adott témakört leíró kifejezések hierarchikus osztályozását tartalmazzák.

#### 4.6.1. Leíró nyelvek

A dokumentumban előforduló jelölést (tag) egy többlet szöveges szintakszisnak lehet tekinteni, amelyet a megjelenítés leírására, az attribútumok,

az információ struktúrájának, a szöveg szemantikai jelentésének megadására használhatunk.

#### 4.6.1.1. SGML

Az SGML (Standard Generalized Markup Language) szöveges dokumentumok és multimédiás anyagok számítógépes jellemzésére, tárolására bevezetett szabványos, kiterjesztett leíró nyelv. Kialakulása a hetvenes-nyolcvanas évekre nyúlik vissza. Kidolgozásának mozgatórugója az elektronikus formában érkező dokumentumok egységes gépesített feldolgozásának elősegítése volt. 1986 óta nemzetközi ISO szabvány. A szabvány két fő részből áll: az információ szerkezetének megadására szolgáló metanyelvből és az SGML dokumentumok tárolási formátumának leírásából. Az SGML szabvány nem rendelkezik az információ megjelenítéséről.

Az SGML technológia alkalmazásakor az egyik leglényegesebb feladat a dokumentumok típusokra osztása és ezen típusok szerkezetének leírása, azaz a DTD - dokumentumtípus-definíció - elkészítése. Erre szolgál az SGML szabványban rögzített metanyelv. A DTD ún. *tartalmazási modell*je megadja, hogy az adott típusú dokumentum milyen kötelező vagy opcionális elemekből áll. Az elemekre ismétlődési és sorrendi megkötések szabhatók. A DTD leírja továbbá az elemekhez rendelhető attribútumokat, a dokumentumon belüli és kívüli objektumokra mutató hivatkozásokat. Ugyancsak a DTD adja meg a dokumentumban használható úgynevezett entitásokat. Ezek speciális karakterek, hosszabb sztringek, és az állományon kívüli szövegrészek, elemcsoportok, képek, táblázatok, stb. jelölésére szolgálnak.

Az SGML szabvány második része az SGML-ben készült dokumentumok tárolási formátumát rögzíti. A formátumban az elemek kezdetét `<leírónév>`-vel, a végét pedig `</leírónév>`-vel jelöljük. A *tag*-eket mi definiálhatjuk tetszés szerint, olyan nevet adhatunk nekik, amit logikusnak tartunk. Ezeket a *tag* definíciókat egy DTD dokumentumleíró fájlban kell megadni.

#### 4.6.1.2. HTML

A HTML (HyperText Markup Language) egy leíró nyelv, melyet web oldalak készítéséhez fejlesztettek ki. Az SGML leegyszerűsített változata, amely a

W3C támogatásával ma már internetes szabvánnyá vált. Fejlesztése 1990-ben kezdődött.

A HTML állományok olyan szimbólumokat tartalmaznak, amelyek a megjelenítő programnak leírják, hogyan is kell megjeleníteni, illetve feldolgozni az adott állomány tartalmát. A megjelenítő program lehet web böngésző, aural böngésző (a felhasználónak felolvasa a megjelenítendő szöveget), braille olvasó (konvertálja a szöveget braille formátumba), levelező program (pl. MS Outlook, Eudora), valamint egyéb eszközök, például mobiltelefon, palm pilot, WebTV.

Négyfajta leíró elem található meg a HTML-ben:

- *Strukturális* elemek, amelyek leírják az adott szöveg célját. Például `<h1>Téma 1</h1>` mint első szintű címsor (alcím).
- *Prezentációs* szimbólumok, melyek leírják, hogy az adott szöveg hogyan nézzen ki. Pl. `<b>vastag</b>` **vastag** kinézetet eredményez. Ez a forma azonban ma már elavultnak számít, helyette egymásba ágyazott stíluslapok (Cascade Style Sheets, CSS) használata javasolt. Ezáltal elérhető a tartalom (amit a HTML kódol) és a forma (CSS) szétválasztása, így minden megjelenítő program számára elérhetővé tehetőek a weblapok.
- *Hípertext* elemek, melyek segítségével kapcsolat létesíthető a dokumentum egyes elemei és más dokumentumok között.  
(Pl. `<a href="http://www.unideb.hu/">Debreceni Egyetem</a>`)
- *Eszköz* elemek, melyek segítségével gombok, listák, beviteli mezők hozhatók létre. (Pl. `<input type="checkbox">`)

Egy HTML állomány három nagy részre bontható:

- A Dokumentum Típus Definíció az állomány legelején, ami a használni kívánt DTD-t adja meg, pl.: „HTML 4.01”;
- a HTML fejléc, ami technikai és dokumentációs adatokat tartalmaz. (tartalmát az internet böngésző nem jelenít meg);
- a HTML törzs, amely a megjelenítendő információkat tartalmazza.

Tehát egy internetes oldal alapszerkezete a következőképpen nézhet ki:

```

<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN"
"http://www.w3.org/TR/html4/strict.dtd">
<html>
<head>
<title>Az oldal megnevezése</title>
<meta http-equiv="Content-type" content="text/html; charset=az
oldal kódolása">
esetlegesen további fejlécinformációk
</head>
<body>
Az oldal tartalma
</body>
</html>

```

A DHTML (Dynamic HTML ) a HTML nyelv azon kiegészítéseinek az összessége, melyek weboldalak tartalmának dinamikus megjelenítését szolgálják. A DHTML nem W3C szabvány, hanem technológiák (leggyakrabban HTML 4.0, stíluslapok, JavaScript) kombinációja.

#### 4.6.1.3. XML

Az interneten található, vizuális megjelenítésre szánt dokumentumok kiválóan leírhatóak a HTML jelölőnyelv segítségével, de az automatizált, gépi feldolgozáshoz más leírónyelvre van szükség, amelynek segítségével a dokumentum logikai struktúrája kialakítható. Ennek a megoldására fejlesztették ki az XML nyelvet (eXtensible Markup Language), amely az SGML egyszerűsített részhalmazaként tekinthető.

Az XML fejlesztése 1996-ban kezdődött, és 1998 februárja óta már hivatalos W3C ajánlás. A nyelv kifejlesztésével az volt a cél, hogy lehetővé tegyék a strukturált szöveg és információ megosztását az Interneten keresztül, feloldják a viszonylag egyszerű struktúrájú HTML korlátait, optimálisan kihasználják az SGML adta lehetőségeket, továbbá minimalizálják a támogatásához szükséges programozói erőforrásokat.

**Az XML technológia egyik jelentős előnye a tartalom és forma szétválasztása.** Az XML szabályok gyűjteménye, melyek segítségével olyan szöveges formátumokat állíthatunk elő, amelyek alkalmasak adatok strukturált, a számítógép által is értelmezhetően feldolgozható leírására, viszont az XML dokumentumok általában nem tartalmaznak információt arról, hogyan jelenjék

meg a dokumentum a képernyőn vagy nyomtatásban. Amíg a HTML előre definiálja, hogy az egyes leírók és attribútumok mit jelentenek, s hogy a leírók közötti szöveg a böngészőben miképp jelenjen meg, addig az XML a leírókat csak az egyes adatszoportok elválasztására használja, az adatok értelmezését meghagyja az alkalmazások számára, amelyek az XML-t olvassák.

A leíró elemek egyértelmű olvasatához az XML *névtérket* használ. Területfüggő tag-ek hozhatók létre különböző témakörökben, így téve lehetővé az egyszerű, szektorokon belüli kommunikációt.

Az XML dokumentumoknál az adatmegjelenítés módja független az adatstruktúrától, a különböző platformokon stíluslapok határozzák meg az adatok reprezentációját. A CSS nyelv az XML-hez ugyanúgy használható, mint a HTML-hez. Az XSL egy továbbfejlesztett nyelv stílusnyelvek leírására. Ez az XSLT-n alapul, ami egy transzformációs nyelv, mely XML fájlok átrendezésére, tag-ek és attribútumok beszúrására, törlésére szolgál. A DOM funkcióhívások egy halmazát szabványosítja, melyek segítségével XML (és HTML) fájlokat manipulálhatunk különböző programozási nyelvekből. Az XSL három részből áll: a fent említett XSLT transzformációs nyelvből; az XML dokumentum részeinek bejárására, címezésére szolgáló XPath nyelvből; és az XML dokumentumok megjelenítését lehetővé tevő XSL FO formázó nyelvből. További lehetőség XML dokumentumok megjelenítésére a DSSSL (Document Style Semantics and Specification Language) nyelv használata.

**Az XML technológiák egy egész családja.** A már említett XSL stílusleíró nyelven kívül az XLink egy olyan szabvány, mely leírja, hogy miként kell XML fájlokba linkeket elhelyezni. Az XPointer és XFragments szintaxissal az XML dokumentumon belül található részekre lehet hivatkozni.

Az XML-en alapuló nyelvek formális módon vannak leírva, ami programok számára is lehetővé teszi a dokumentumok módosítását és validálását.

Példák a különböző szakterületeken alkalmazott XML alapú leíró nyelvekre:

- MathML (Mathematical Markup Language) a matematikai problémák leírására. Kétféle tag-et használ: az egyik a formulák megjelenítésének leírására, a másik a kifejezések jelentésének megadására szolgál.

- SMIL (Synchronized Multimedia Integration Language) különböző multimédiás források egy dokumentumba ágyazására alkalmas, ahol a különböző objektumok helye és aktiválásának ideje is beállítható.
- PGML (Precision Graphics Markup Language), 2D vektorgrafika PDF és PostScript alapokon.
- MusicML (Music Markup Language) kották leírása alkalmas.
- CML (Chemical Markup Language) vegyipari dokumentumok leírására.
- RDF- (Resource Description Framework) Mint a neve is jelzi, keretet ad a web-alapú információforrások és dokumentumok különböző metaadat szabványainak egységes kezelésére. Az RDF egy olyan deklaratív nyelv, amelyet állításoknak precíz, formális szókészletek segítségével történő kifejezésére, weben keresztüli elérésére és alkalmazására fejlesztettek ki.

Az XML platform, szoftver és hardver független eszköz az információ továbbítására. Rendkívül fontos szerepet tölt be az információ- és tudásmenedzsment területén belül. Az XML támogatja a Unicode-ot, ami lehetővé teszi bármely információ bármely emberi nyelven történő közlését.

| HTML                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | XML                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <ul style="list-style-type: none"> <li>▪ Leíró nyelv szövegek leírására</li> <li>▪ Megjelenésre koncentrál</li> <li>▪ Az adat szöveggé válik</li> <li>▪ Előre meghatározott címke készlet</li> <li>▪ “Szabados” szintaxis <ul style="list-style-type: none"> <li>▫ <code>&lt;B&gt;&lt;I&gt;...&lt;/B&gt;&lt;/I&gt;</code> megengedett</li> <li>▫ Záró címke elhagyható</li> <li>▫ <code>...bgcolor=blue...</code> megengedett</li> <li>▫ Nem különbözteti meg a kis és nagybetűt</li> </ul> </li> </ul> | <ul style="list-style-type: none"> <li>▪ Leíró nyelv adatok leírására</li> <li>▪ Tartalomra koncentrál</li> <li>▪ Az adatok megtartják jelentésüket</li> <li>▪ Szabadon definiálható címke készlet</li> <li>▪ Rögzített szintaxis <ul style="list-style-type: none"> <li>▫ Szabályos zárójelezés</li> <li>▫ Záró címke nem hagyható el</li> <li>▫ Attribútumok értéke idézőjelben</li> <li>▫ Különbség kis és nagybetű között</li> </ul> </li> </ul> |

3. táblázat. A HTML és az XML összehasonlítása

A HTML és az XML tulajdonságainak összehasonlítását tartalmazza a 3. táblázat [Rutkovszky, Princz, 2001].

A HTML-t XML-ben újradefiniálták, ez a nyelv az XHTML, ahol az X utal arra, hogy ez a HTML nyelv az XML szintakszist követi

#### 4.6.1.4. Szemantikus web

A web ontológiák nyelve (Web Ontology Language, OWL) arra szolgál, hogy leírjuk vele a web dokumentumaiban és alkalmazásaiban előforduló lényeges és jellemző osztályokat, és a közöttük lévő kapcsolatokat.

Az OWL nyelv arra használható, hogy

- formalizáljunk egy tématerületet osztályok és tulajdonságaik definiálása útján;
- egyedeket határozzunk meg, és a tulajdonságaikról kijelentéseket tegyünk;
- többlet információkat következtessünk ki ezekről az osztályokról és egyedekről (amilyen mértékben ezt az OWL nyelv formális szemantikája lehetővé teszi).

Az RDF, az ontológiák és a jelentések számítógép számára értelmezhető ábrázolása áll a szemantikus web kutatási terület középpontjában. A szemantikus web következtetéseket lehetővé tevő infrastruktúra a weben, amely kiterjeszti a web jelenlegi lehetőségeit. Keretrendszert biztosít az adatok megosztására és újrafelhasználására, gyakorlatilag globálisan összefüggő adatbázisként lehet felfogni.

Manapság a szemantikus webet gyakran emlegetik úgy, mintha az a webet felváltó technológia lenne. Eric Miller, a Szemantikus Web Fejlesztési Terület vezetője szerint - "a szemantikus web inkább továbbfejlesztést, mint fordulatot jelent a web életében. Fokozatosan a számítógép által is megérthető leírásokat kapcsolnak azokhoz az adatokhoz és dokumentumokhoz, amelyek korábban már a weben voltak, és így alakult ki a szemantikus web. Így az XML, az RDF és az OWL segítségével a web a dokumentumok és adatok olyan globális infrastruktúrája lett, ahol könnyebb és megbízhatóbb az információk keresése és felhasználása."

## 4.7. Szöveg műveletek

Szöveges dokumentumokban való keresés egyik módja, hogy a keresett szó, kifejezés, szövegrészlet előfordulását szekvenciálisan, karakterről karakterre ellenőrizzük az állományokban (pl. grep). A szekvenciális vagy online keresés nem igényli a szöveg előkészítését a keresésre, azonban csak kisebb, vagy gyorsan változó szöveges állományok esetén megfelelő, sok időt követel, és a dokumentumoknak is elérhetőnek kell lenniük a keresés végrehajtásakor. Az esetek többségében azonban a dokumentum-gyűjtemény nem érhető el a kereséskor, így a keresés csak a dokumentumokat leíró információk között történhet.

Szöveges állományokban való keresésnek egy másik lehetséges módja, ha a szöveg fölé adatstruktúrát – pl. indexet - építünk a keresés meggyorsítása érdekében. Az indexek létrehozása különösen nagy és nem túl gyakran változó szöveges gyűjtemények esetén célszerű - mint amilyen a web -, de természetesen a kétféle keresés kombinálható is.

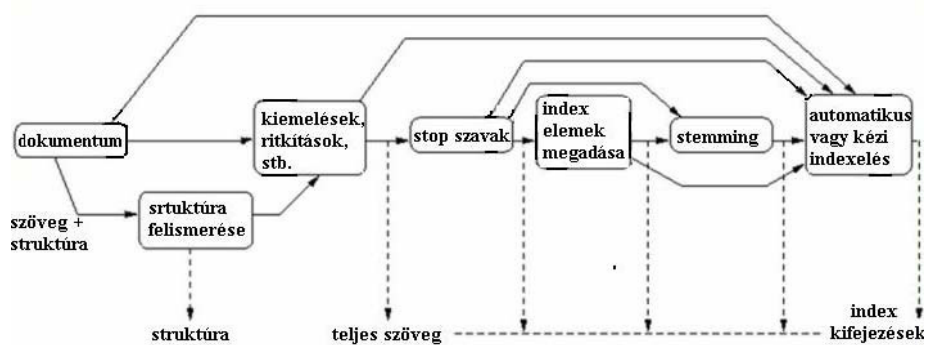
A dokumentumok jól reprezentálhatók indexkifejezések halmazával (pl. könyvtári katalógusok), ahol az indexkifejezések alatt szavakat vagy szókapcsolatokat értünk. A szavak különböző mértékben képviselik egy-egy dokumentum szemantikai tartalmát. Általában a főnevek reprezentálják leginkább egy dokumentum tartalmát, éppen ezért egy dokumentum-gyűjtemény indexkifejezései elsősorban főneveket tartalmaznak. Vannak azonban olyan keresőrendszerek (pl. a legnépszerűbb szabadszavas keresők), amelyek a stop-szavakon kívül a dokumentumok valamennyi szavát indexelik.

Az információ-visszakeresésére a dokumentumot elő kell készíteni. A dokumentum előfeldolgozása egy olyan folyamat, amely főként öt szöveg műveletre bontható:

- A szöveg lexikai elemzése a számjegyek, kötőjelek, írásjelek, a kis- és nagybetűk figyelembe vételével.
- A stop szavak eltávolítása. E lépéssel az index struktúra nagyságát akár 40%-kal is csökkenthetjük. [Baeza-Yates, Ribeiro-Neto, 1999]
- A maradék szó szótöveinek megkeresése a prefixek és a szuffixek eltávolításával (stemming), megengedve a visszakeresésnél a lekérdező kifejezés mondattani változatait (pl. kapcsol, kapcsolni, kapcsolt). E

művelet is csökkenti az index struktúra nagyságát, bár a stemming előnyeiről megoszlanak a vélemények. Talán éppen ezért számos szabadszavas kereső nem alkalmazza ezt az eljárást.

- Annak eldöntése, mely szavakat (szótöveket) használjuk indexelő elemként. Előfordulhat, hogy az egymás mellett, vagy egymás közelségében gyakran előforduló két vagy három szót is egy indexelő kifejezéssé kapcsolunk össze (pl. Arany János, számítógép-tudomány).
- Dokumentumok osztályozására használt struktúra (pl. tezaurusz) létrehozása, megengedve az eredeti lekérdezés kibővítését a szinonimákkal.



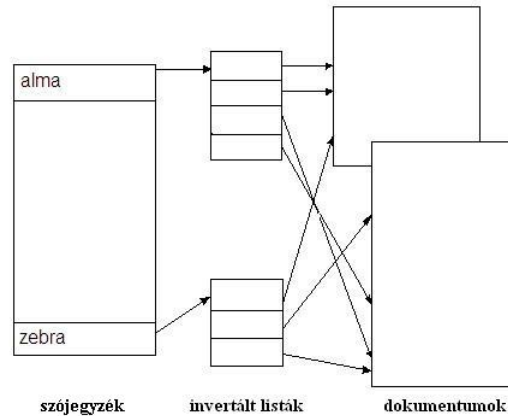
5. ábra. A szöveg előfeldolgozási folyamatának fázisai

## 4.8. Indexelés és keresés

Több különböző indexelési technikát fejlesztettek ki, hogy hatékonyan támogassák a szöveges dokumentumok között az információ-visszakeresést. Ezek közé tartoznak az invertált fájlok, a suffixek tömbje és a szignatúra fájlok. A három technika közül az invertált fájlok módszere a leggyakrabban alkalmazott technika, ezt használják a legnépszerűbb webes keresőrendszerek is. A suffixek tömbje gyorsabb a kifejezések keresésénél, de fáradságos megépíteni és fenntartani. A szignatúra fájlok népszerűek voltak az 1980-as években, de napjainkban az invertált fájlok módszere kiszorította ezt a technikát.

### 4.8.1. Invertált fájlok

A web tartalmában való keresésre az invertált fájlok alkalmazása vált általánossá, amely a kulcsszó alapú keresést teszi lehetővé. Az invertált fájlok struktúrájának két összetevője van: a szójegyzék vagy lexikon, valamint az előfordulások listája (invertált listák).



6. ábra. Az invertált fájlok felépítése

A szójegyzék tartalmazza a szövegyűjteményben előforduló szavakat. A szójegyzék minden egyes szóhoz készül egy lista, amely rendezett sorrendben tartalmazza a szó minden előfordulását a dokumentumokban. Az előfordulás megadja a tartalmazó oldal azonosítóját és az oldalon belül az előfordulás helyének pozícióját (postings). Amennyiben nem szükséges feltüntetni minden egyes előfordulást az oldalon belül, a hely megadás csak az oldal azonosítóját tartalmazza, valamint opcionálisan az előfordulások számát az oldalon belül. Így egy adott szót tartalmazó dokumentum előkeresésére csak a szót kell megkeresni a lexikonban, és az előfordulások listája alapján a szót tartalmazó dokumentumok előkereshetők. A Boole lekérdezéseknél a keresési szavaknak megfelelő előfordulások metszetét vagy unióját kell venni, az értelmezésnek megfelelően.

Az indexelő technikák tökéletesítésével (pl változó hosszúságú azonosítók) az invertált fájlok méretét a szöveg méretének 30%-áig lehet csökkenteni, de további tömörítő technikákkal az index mérete a szöveg méretének 10%-ig is csökkenthető [Witten,Moffat,Bell,1999]. E megközelítés hátránya, hogy az

invertált listákat dekódolni kell a visszakereséseknél, de az ilyen kitömörítések gyorsan végrehajthatók. [Zobel et al. 1998].

Az előfordulások listájának helyigénye csökkenthető a blokk címzéses technika használatával. A szövegeket ekkor blokkokra osztják, és az előfordulások listája a pontos előfordulás felsorolása helyett csupán azon blokkok azonosítóját tartalmazza, amelyekben a szöveg feltűnik. E technika alkalmazásánál azonban a szövegnek a keresés idejében elérhetőnek kell lennie, ezért a webes keresések esetében nem használatos. Gyakran alkalmazzák azt a technikát, amikor a szavak előfordulásának feltüntetésére egy dokumentumon belül a pointerek helyett a szó sorszámát tartalmazza az invertált lista. Ez a módszer elősegíti a kifejezések, és a közelségi, szomszédsági keresések megvalósítását is.

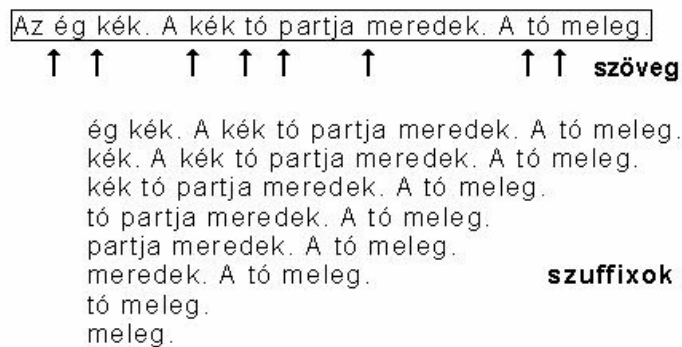
A szójegyzék tárolására hatékony struktúra a B-fa, de tömbök és hash táblák is alkalmasak rá. A szójegyzék általában kevesebb helyet igényel az invertált listáknál: Például az 1Gb méretű TREC-2 gyűjtemény esetében mindössze 5 Mb. [Baeza-Yates&Ribeiro-Neto,1999].

#### **4.8.2. Szuffixek tömbje**

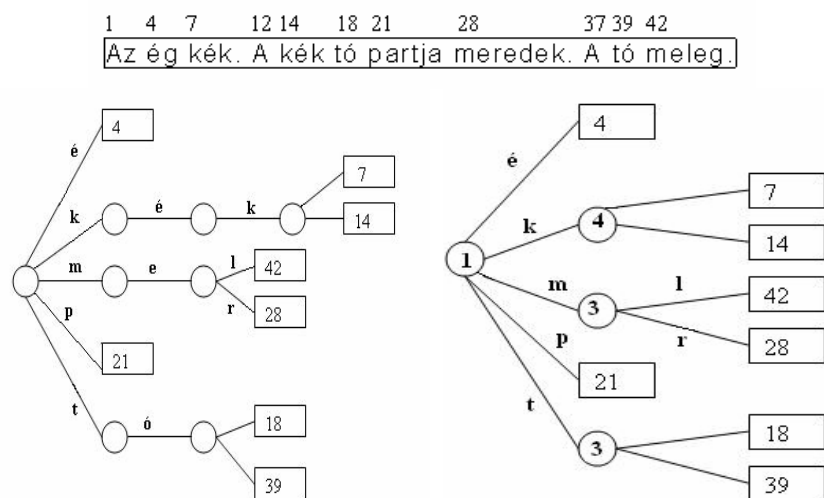
Az indexelésnek ez a típusa lehetővé teszi bonyolultabb lekérdezések (pl. kifejezések keresése) hatékony megvalósítását is, de hátránya a költséges megvalósíthatósága (különösen hosszú szöveges állományok esetén), s az a tény, hogy a szövegnek a lekérdezés idején olvashatónak kell lennie.

Ez a fajta indexelés a szöveget egy hosszú karaktersorozatnak tekinti. A szöveg minden pozíciója egy szuffixnek (a pozíciótól a szöveg végéig tartó karaktersorozatnak) tekinthető, s minden szuffix egyértelműen meghatározható a pozíciójával.

A szöveg nem minden pozícióját szükséges indexelni: elegendő csupán azon pozíciók indexelése, amelyek a visszakereshető szavak vagy kifejezések kezdő pozíciójával egyeznek meg.



7. ábra. Mintaszöveg az indexelési pontok jelölésével



8. ábra. Szuffixfa és a tömörített szuffixfa a mintaszövegre

A szuffixfa a szöveg valamennyi szuffixa fölé épített adatstruktúra, ahol a szuffixokra mutató pointerok a fa végpontjaiban található. A jobb térkihasználás végett a fa tömöríthető: azon csomópontokon átvezető utak, amelyekből csak egy további csomópontba lehet jutni, a 8. ábra szerint tömöríthetők.

A szuffixfával kapcsolatos probléma, hogy nagyon sok helyet igényel: Megvalósítástól függően egy-egy csomópont tárigénye 12 és 24 byte közötti érték, s ha csak a szavak kezdőpontját indexeljük, akkor is az eredeti szöveg többszöröse lehet az index nagysága (120% és 240% közötti érték [Baeza-Yates&Ribeiro-Neto,1999]).

A szuffixtömb a szuffixfához hasonlóan működik, de jóval kevesebb a tárigénye. Ha a szuffixfa végpontjait (felülről lefelé haladva) egy tömbbe írjuk, akkor a szöveg szuffixai között abc szerint rendezve kereshetünk.

|                                                 |   |    |    |    |    |    |             |    |    |    |
|-------------------------------------------------|---|----|----|----|----|----|-------------|----|----|----|
| 1                                               | 4 | 7  | 12 | 14 | 18 | 21 | 28          | 37 | 39 | 42 |
| Az ég kék. A kék tó partja meredek. A tó meleg. |   |    |    |    |    |    |             |    |    |    |
| 4                                               | 7 | 14 | 42 | 28 | 18 | 39 | szuffixtömb |    |    |    |

9. ábra. Szuffixtömb a mintaszövegre

A szuffixek tömbje tehát egy olyan tömb, amely lexikografikus sorrendben tartalmazza a szöveg szuffixaira mutató valamennyi pointert. A szuffixek tömbjében bináris kereséseket lehet végrehajtani, összehasonlítva a pointerek tartalmát a keresett kifejezéssel.

### 4.8.3. Szignatúra fájlok

Szignatúra fájlok egy valószínűségi módszer a dokumentumok indexelésére. A dokumentum minden egyes szavához egy szignatúrát kapcsolunk, ami azonos számú bitekből álló vektor. Ezeknek a hozzákapcsolt vektoroknak a tartalma véletlenszerű, előállításuk hash függvény alkalmazásával történik.

A dokumentum deszkriptorát úgy kapjuk, hogy a dokumentumban lévő szavak szignatúráinak bitenként a logikai VAGY kapcsolatát vesszük. Ha a dokumentum elég hosszú, akkor a szöveget blokkokra osztjuk, és a blokkban található szavak szignatúráit bitenként összeadva elkészítjük a blokkok szignatúráit is.

Ha ellenőrizni akarjuk, hogy egy  $S$  szó egy adott  $D$  dokumentumban (blokkban) megtalálható-e, akkor meg kell néznünk, hogy valamennyi bit, amely az  $S$  szignatúrájában be van állítva, be van-e állítva a  $D$  deszkriptorában is, azaz igaz-e, hogy  $S \& D = S$ , ahol  $\&$  a bitenkénti AND (ÉS) kapcsolatot jelöli  $S$  és  $D$  szignatúrái között. Ha nem igaz az egyenlőség, akkor  $S$ -t nem tartalmazza a dokumentum. Ha igen, akkor  $S$  valószínűleg megtalálható a dokumentumban, de előfordulhat más kifejezések szignatúrájának kombinációjaként is, ezért nem lehet megmondani biztosan, hogy egy adott szó ténylegesen előfordul-e az adott dokumentumban.

| Szó   | Szignatúra          |
|-------|---------------------|
| boci  | 1000 0000 0000 0100 |
| tarka | 0100 0010 0010 0000 |
| lepke | 1000 1000 0000 0001 |
| kis   | 0000 0000 0001 0001 |
| mese  | 0000 0001 1000 0010 |

| Dokumentum | Szöveg                   | Deszkriptor         |
|------------|--------------------------|---------------------|
| 1          | Boci, boci tarka         | 1100 0010 0010 0100 |
| 2          | Tarka lepke, kis mese... | 1100 1011 1011 0011 |

4. táblázat. Példa szignatúra fájlokra

A szignatúra fájlok lekérdezése néha azt eredményezi, hogy a kifejezés a dokumentumban található, noha a dokumentum ténylegesen nem tartalmazza azt. A szignatúra fájlok tervezésének legkényesebb része a hamis egyezések valószínűségét minél alacsonyabban tartani a szignatúrák hosszának a lehető legrövidebb megválasztása mellett.

Ha  $S$  a szignatúrában szereplő bitek száma, és  $X$  a szignatúrákban beállított bitek száma (értékük =1), akkor általában

$1000 \leq S \leq 10\,000$  és  $6 \leq X \leq 20$  javasolt [Witten et al, 1999; Blelloch, 1997].

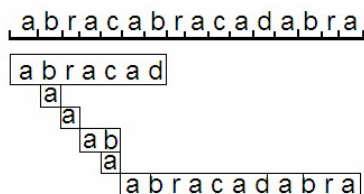
A szignatúra fájlok legfőbb előnye, hogy – az invertált fájlokkal szemben – nem igényelnek lexikont, így a lekérdezés alatt nem igénylik a lexikon memóriában való jelenlétét sem. Szignatúra fájlok alkalmazására példa a TREC (Text REtrieval Conference) adatbázis.

#### 4.8.4. Szekvenciális keresés

A szekvenciális keresés a szöveges dokumentumokban való kereséseknek egy olyan módja, amelynél semmiféle adatstruktúra nem épül a szövegre. Ez a módszer számos indexelő technika része.

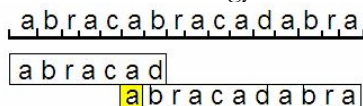
Karaktersorozatok pontos egyezésének vizsgálatára szolgál, alkalmazásával megadható egy hosszú szövegben az összes olyan pozíció, amelytől kezdve a minta előfordul. Klasszikus keresési probléma, amelyre számos elméleti

megoldást kifejlesztettek már. Az algoritmusok közül a legegyszerűbb megoldás a nyers erő alkalmazása (Brut-force algorithm), azaz végig kell próbálni az összes lehetséges esetet, amikor a szöveg megegyezhet a mintával. Ez a szövegben az összes olyan pozíció vizsgálatát jelenti, amely a minta kezdőkarakterével megegyezik. Az alábbi példákban az abracadabra szó előfordulását keressük a szövegben:



10. ábra. Keresés a “Brut-force” algoritmus használatával

Átlagban egy kissé gyorsabb keresést tesz lehetővé a Knuth-Morris-Pratt algoritmus. Ez annyiban különbözik a “Brut-force” algoritmustól, hogy a megelőző egyezési vizsgálatok eredményét felhasználva kizárja eleve azokat a pozíciókat a szövegben, ahol a minta nem egyezhet.



11. ábra. Keresés a Knuth-Morris-Pratt algoritmus használatával

A szekvenciális keresés megvalósítására született számos modell közül az összehasonlításoknál az ismertett két modell nyújtja a legjobb eredményt [Baeza-Yates&Ribeiro-Neto,1999].

## 4.9. Multimédia visszakeresés

A multimédia objektumokból az információ-visszakeresésére két mód kínálkozik: Lehetséges ezen elemek között böngészni, navigálni (pl. CAD/CAM/CASE környezetben), illetve információ-visszakereső rendszert használni a remélt információ fellelése érdekében.

A multimédia információ-visszakereső rendszerek két szempontból különböznek a hagyományos információ-visszakereső rendszerektől: Egyrészt a multimédia objektumok struktúrája jóval összetettebb, mint a hagyományos szöveges adatok struktúrája, másrészt az objektumok visszakeresése főleg a hasonlóságon alapul. A multimédia információ-visszakereső rendszerek ötvözik az adatbázis-kezelő rendszerek és az információ-visszakereső rendszerek tulajdonságait, így képesek a tulajdonság-alapú és a tartalom-alapú lekérdezések megválaszolására is. A multimédia visszakereséseknél a legtöbb próbálkozás az információ kiemelésére és összegzésére történik.

#### 4.9.1. Az adatmodell

Az adatbázisoknál a tulajdonságokat az adatbázis sémája tartalmazza, amely strukturálja is az adatot. A multimédia adatok (kép, videó, hang) tipikusan strukturálatlanok, ezért szükségesek olyan módszerek, amelyek megadják ezen adatok tartalmát, tulajdonságait, azaz valamelyest strukturálják ezen adatokat (szemistrukturált adatok). A multimédia adatok kezelésére az objektumorientált adatbázis-kezelő rendszerek használhatóbbak, mint a relációs adatbázis-kezelő rendszerek, viszont az előbbiek nem szabványosítottak még. A két rendszert egyesítik az objektum-relációs adatbázis-kezelők, amelyek összetett adattípusokat képesek kezelni a relációs modellhez hasonlóan (SQL3 szabvány).

Problémát okoz a multimédia adatok reprezentálása is, azaz ezeket az adatokat olyan leíró tulajdonságokkal ellátni, amelyek a visszakeresést elősegítik. Az objektumokhoz a tulajdonságokat hozzá lehet kapcsolni a felhasználó által manuálisan (időigényes és drága), illetve a rendszer által automatikusan, bár a leggyakoribb a vegyes megvalósítás: A rendszer meghatároz az értékek közül néhányat, amit a felhasználó javít vagy gyarapít. Problémát okozhat az is, hogy a leírás lehet szubjektív és lehet részleges. Például egy képnek különbözhet a *fogalmi* jelentése (erdő, hegy, vidék, tájkép), és eltérhet *jelképes* jelentése (pl. a színek különböző érzelmi hatásokat válthatnak ki), így a kép mást-mást jelenthet különböző embereknek.

#### 4.9.2. Kulcsszón alapuló visszakeresés

Multimédia objektumok esetén az információ-visszakeresésre elsőként a szöveges adatállományoknál is alkalmazott módszert használták: kulcsszón

alapuló visszakeresés az objektum nevéből, az objektumot leíró szövegmagyarázatból vagy metaadatokból. A visszakeresést megelőzően azonban a multimédia objektumokat el kell látni leíró tulajdonságokkal. Az ilyen típusú visszakeresések a lekérdezéssel pontosan egyező dokumentumokat adnak eredményül.

### 4.9.3. Tartalom alapú visszakeresés

Az Internet népszerűsége és a multimédia objektumok számának óriási mértékű növekedése miatt a szöveges leírásokkal szemben egyre nagyobb az igény a **tartalom alapú**\* multimédia visszakeresési technikák kifejlesztésére.

A tartalomalapú megközelítések esetében a keresés olyan tulajdonságokon alapszik, melyek közvetlenül a multimédia objektumból származnak, mint például szín vagy textúra, elmozdulás, ritmus, hangszín.

A tartalom alapú információ-visszakeresés lekérdezései három fő csoportba oszthatók.

- **Lekérdezések példa megadás alapján**

A hasonló alapján történő lekérdezések esetében a felhasználó megad egy példa dokumentumot, majd a rendszer visszaad az adatbázisból minden olyan dokumentumot, amely hasonló a felhasználó által megadott példányhoz.

- **Struktúra alapján történő lekérdezések**

A struktúra alapú kereséseknél a felhasználó felvázolja az objektumot, amely a későbbiekben a keresés alapja lesz. Például képeknél megad egy alakzatot, audio visszakeresésnél megad egy dallamot.

- **Ikon alapú lekérdezések**

Az ikon alapú lekérdezésnél a felhasználó a keresett dokumentum tulajdonságát leíró fogalmakat ad meg. Például egy kép, amely vizet ábrázol, egy videó, ahol leesik egy tárgy, egy hang, amely hahotázik.

A tartalom-alapú visszakereső rendszerek mindig a lekérdezéshez leginkább hasonló objektumokat adnak eredményül. A tanulási funkcióval

---

\* Szöveges információ között tartalom alapú visszakeresésekkel a természetesnyelv-feldolgozás (Natural Language Processing) foglalkozik.

rendelkező rendszerek gyakran várnak egy visszacsatolást, melyben a felhasználó értékeli a kapott eredményeket relevanciájuk szerint. Az értékelést a rendszer felhasználja a következő hasonló jellegű kereséskor a minél pontosabb eredmény szolgáltatása érdekében.

## 5. Információ-visszakeresés a weben

A hagyományos információ-visszakereső szoftvereket tipikusan direkt elérésű dokumentumok statikus gyűjteményében való visszakeresésre tervezték, ahol az adatbázis viszonylag kevés és homogén adatot tartalmaz.

A weben hatalmas számú dokumentum található, amely igencsak különbözik tartalmában, formátumban, minőségben, a kihelyezés céljában. A webet egy hatalmas, strukturálatlan, széles körben használt adatbázisnak tekinthetjük, ahol a dokumentumok mennyisége, megoszlása és dinamikus jellege különösen nehezíti egy teljes és naprakész információ-visszakereső rendszer létrehozását. A webet irányított gráfként lehet modellezni, amelynek csúcsai az egyes dokumentumok, élei pedig egyik dokumentumról a másikra mutató hipervivatkozásoknak felelnek meg.

Alapjában véve három különböző formája van a weben történő keresésnek: kulcsszó, téma és hivatkozás szerint. Az első eset a szabadszavas keresők használatát igényli, amelyek a weben lévő dokumentumok egy részét indexelik és teszik visszakereshetővé. A második esetben tematikus keresők (webes könyvtárak, hivatkozásgyűjtemények) használhatók, amelyek kiválasztott dokumentumokat osztályoznak téma szerint. A harmadik forma kevésbé ismert és használt: ez a weben való keresést teszi lehetővé a hiperlinkek struktúrájának felhasználásával.

A dolgozatban a webről való információ-visszakeresésnél elsősorban a szövegre összpontosítunk, hiszen bár léteznek technikák a multimédia adatok között a tartalom-alapú keresésre, ezek egyelőre még nem használatosak általánosan a weben. A lekérdezéseknél a szintaktikus lekérdezéseket szeretnénk kihangsúlyozni. Szintaktikus lekérdezéseknél a felhasználó információs igényét szavakban vagy mintákban fogalmazza meg. Ennek alternatívája a lekérdezés természetes nyelvű megfogalmazása, ami a szövegek természetes nyelvi analízisét igényli. Ez a technika ma még nem elég hatékony, így nem is használatos széles körben.

## 5.1. Kihívások

A webbel kapcsolatban számos olyan probléma adódik, amelyek megnehezítik az információ-visszakeresést. Ezek oka alapján két csoportba sorolhatók: egyrészt az adatok, másrészt a felhasználók, valamint a visszakereső-rendszerrel való kapcsolatuk

### 5.1.1. Problémák az adattal

#### 5.1.1.1. Osztott adatok

A web természetéből adódóan az adatok számtalan számítógépről és platformon át érhetők el. A számítógépek nem definiált topológiában vannak összekapcsolva, és a hálózatok összekapcsolásánál a rendelkezésre álló sávszélesség és megbízhatóság is nagyon eltérő lehet.

#### 5.1.1.2. Változó adatok

A web dinamikájának velejárója, hogy új számítógépeket és adatokat lehet könnyen hozzáadni vagy eltávolítani. A tartalom jelentős százalékanak gyors változása, létrejötte vagy megszűnése különbözteti meg a webet a hagyományos információ-visszakereső rendszerektől.

- Brewington és Cybenko (2000) figyelte meg, hogy a weboldalak fele nem idősebb 100 napnál, és csak 25 %-a idősebb 1 évnél.
- Cho és Garcia-Molina (2000. a) megállapítása, hogy a *.com* domainen a weboldalak 40%-a változik minden nap. A weboldalak felezési ideje a *.gov* és *.edu* domaineken négy hónap.
- Koehler (1999) szerint a web tartalmának felezési ideje két év.
- Markwell és Brooks (2002) megállapítása, hogy a tudományos oldalak felezési ideje 25 hónap.
- Cockburn és McKenzie (2001) a könyvjelzők felezési idejét két hónapra értékeli.

#### 5.1.1.3. Nagy mennyiségű adat

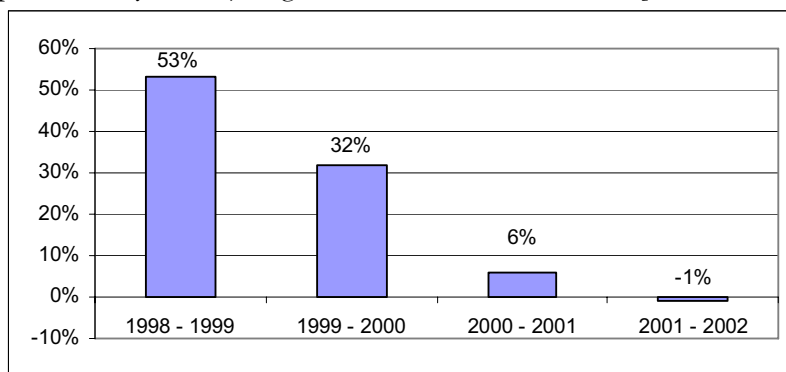
A web hatalmas mennyiségű információt tartalmaz, melynek nagyságára, a növekedés ütemére különböző becslések léteznek.

A web tartalmát két csoportra lehet osztani: a látható (surface) webre - ezen oldalakat képesek bejárni a szabadszavas keresők indexelő programjai -, valamint a láthatatlan (invisible) webre.

Egy 2000-ben publikált felmérés [How Much Information: Internet, 2000] a látható web akkori méretét 25 és 50 terabyte közötti értékre becsülte, amiből 10 és 20 terabyte között volt a szöveges állományok nagysága. E forrás szerint a látható weben át elérhető dokumentumok száma 2,5 milliárd körül volt, s akkoriban minden nap 7,3 millió új weblap jelent meg a weben. A növekedés nagysága 0,1 terabyte volt naponta. A láthatatlan web nagyságát (weben elérhető adatbázisok, intranet, stb.) jóval nagyobbbnak, akár a látható web méretének 400-550-szeresére becsülte. A láthatatlan web méretére vonatkozó hasonló becslés olvasható a BrightPlanet honlapján is.

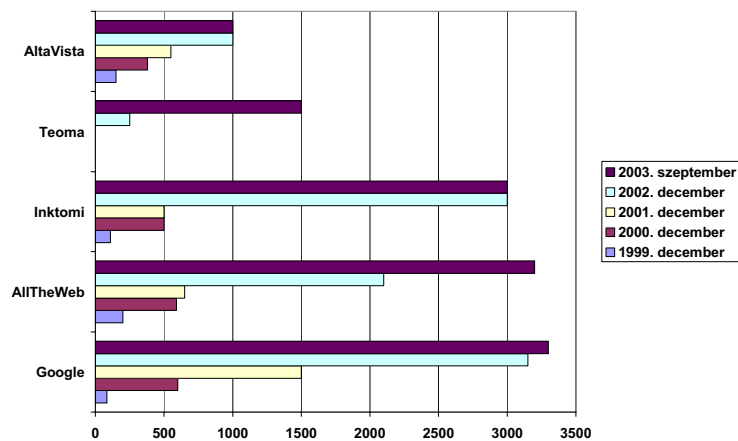
Az egyik legfrissebb adat szerint 2005. januárjában az indexelhető web mérete meghaladta a 11,5 milliárd oldalt [Gulli & Signorini,2005].

A web kezdetben robbanásszerű gyorsasággal növekedett, de mára a növekedés üteme lelassult. A növekedés adatai olvashatók az OCLC (Online Computer Library Center) megbízásából készült felmérésben [O'Neil at al,2003].



12. ábra. A web növekedésének üteme

A SearchEngineWatch által közölt adatok szerint a 2003. szeptemberében a legnagyobb indexállománya a következő keresőknek volt:



13. ábra. A legtöbb oldalt indexelő szabadszavas keresők mérete (millió weboldal)

A rendelkezésre álló adatok szerint 2005-ben a Google már 8,1 milliárd, az MSN 5,0 milliárd, a Yahoo 4,2 milliárd, az Ask Jeeves (Teoma) 2,5 milliárd oldalt indexelt [Gulli & Signorini,2005].

#### 5.1.1.4. Strukturálatlan és redundáns adatok

A weben található dokumentumok többségét emberi olvasásra szánták, s így csak vizuális megjelenítés céljára vannak strukturálva (szemistrukturált dokumentumok). Az oldalak általában HTML formátumúak, ahol a strukturálást, s ezzel a lekérdezhetőséget a különböző meta elemek, jelölő tag-ek alkalmazásával érjük el (pl. címsor, bekezdés, felsorolás).

A World Wide Web továbbfejlesztése a szemantikus web, ahol már a weben lévő források jelentésére kérdezhetünk rá, és nem csupán szavakra, kifejezésekre kereshetünk. A weben lévő dokumentumok gépi feldolgozásához szükséges szemantikai strukturáltságot (pl. "születési hely", "lakóhely") az XML alkalmazásával hozhatjuk létre, de az XML használata a weben még egyáltalán nem elterjedt. A szabadszavas keresők többsége napjainkban még nem indexeli az XML elemeket.

|                                                                  | 1998  | 1999  | 2000  | 2001  | 2002  |
|------------------------------------------------------------------|-------|-------|-------|-------|-------|
| A nyilvános webhelyek száma (ezer)                               | 1 457 | 2 229 | 2 942 | 3 119 | 3 080 |
| A webhelyek aránya, amelyeknek nyitólapja metaadatokat tartalmaz | 70%   | 79%   | 82%   | 85%   | 87%   |
| A jelölő elemek átlag száma                                      | 2,75  | 2,32  | 2,72  | 2,97  | 3,14  |
| Metaadatokat tartalmazó oldalak aránya                           | 45%   | 50%   | 59%   | 62%   | 70%   |
| A jelölő elemek átlag száma az oldalakon                         | 2,27  | 2,16  | 2,46  | 2,63  | 2,75  |

**5. táblázat. Metaadatok használata a nyilvános webhelyeken [O'Neil et al,2003]**

A weben lévő adatok gyakran ismétlődnek tükrözés vagy másolás eredményeként, és gyakran nagyon hasonlóak. Shivakumar és Garcia-Molina a weboldalak közel 30%-át másolatnak tartja [Shivakumar&Garcia-Molina,1998]. Az oldalak szemantikus redundanciája még nagyobb lehet.

#### **5.1.1.5. Az adatok minősége**

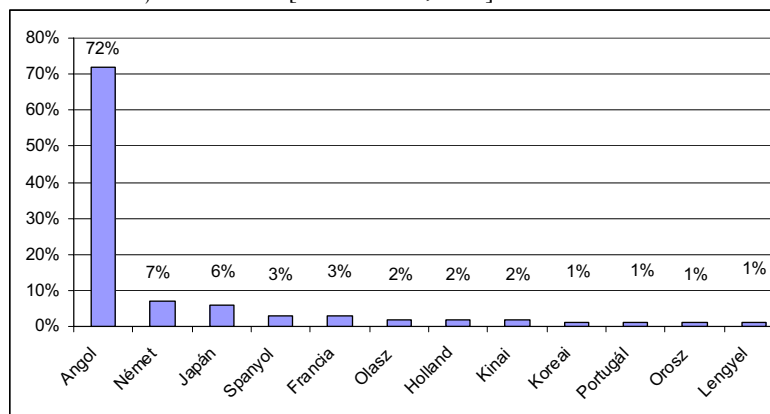
A webet úgy tekinthetjük, mint egy új kiadói formát. A legtöbb esetben azonban itt nincs meg a szerkesztőségi folyamat, így az adatok lehetnek tévesek, érvénytelenek (pl. időközben megváltoztak), a szöveg lehet gyengén megírt, vagy különböző típusú hibákat tartalmazhat (elgépelés, nyelvtani hibák, stb.) Navarro tanulmánya szerint gyakori szavaknál 200 esetből egyszer, míg idegen vezetéknévknél 3 esetből egyszer fordul elő elgépelés [Navarro, 1998].

#### **5.1.1.6. Heterogén adatok**

A weben többféle média található meg, és mindegyik típus számtalan formában lehet tárolva. A legnépszerűbb forma a HTML, ezt követi a GIF és JPG (mindkettő képek tárolására szolgál), az ASCII szövegek tárolására szolgáló TXT, majd a Postscript formátum.

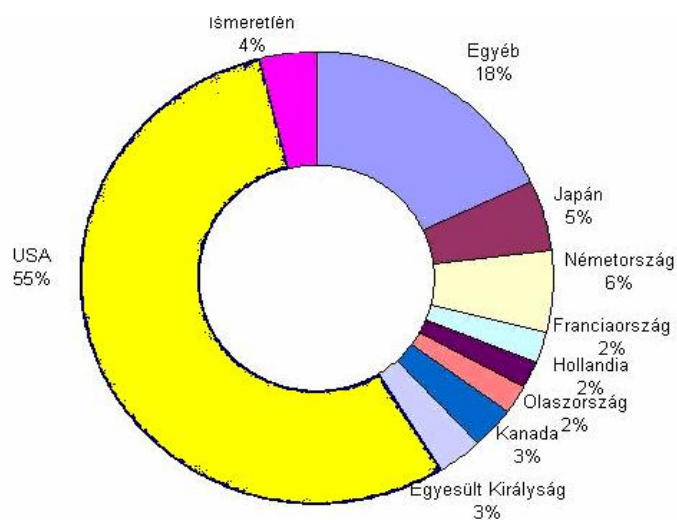
A dokumentumok különböző nyelveken íródnak, ráadásul különböző karakterkészletet használva. A kódolás sem egységes (pl. ASCII, Unicode). Mindez megnehezíti a dokumentumok megírását, megjelenítését, tárolását.

A webben domináns az angol nyelvű tartalom. A nyilvános oldalak nyelvi megoszlását mutatja a 14. ábra [O'Neil at al,2003].



14. ábra. A nyilvános web megoszlása nyelvek szerint. (2002-es adat)

Az angol nyelv dominanciájára magyarázat, hogy a szabadon elérhető webhelyek közel kétharmada angol anyanyelvű országokban található.



15. ábra. Nyilvános webhelyek ország szerinti megoszlása

### **5.1.2. Problémák a felhasználókkal**

A problémák másik csoportja a felhasználókkal kapcsolatos. Ezek egyrészt abból adódnak, hogy a felhasználók az információ igényüket hogyan fogalmazzák meg, másrészt hogyan értelmezik a rendszer szolgáltatása eredményeket. Sokszor a rosszul megfogalmazott kérdés miatt (elírás, kis és nagybetű érzékenység, rosszul használt operátorok) kiesnek releváns dokumentumok. A felhasználó nem feltétlenül veszi észre hibáját, de az így kapott gyenge minőségű oldalak nem elégítik ki információ igényét.

Nem könnyű precízen megfogalmazni a kérdést, tűnjék bár az egyszerűnek, s nem könnyű az eredményül kapott oldalak tömegében sem az eligazodás. Hasznos a felhasználókat a webről való információ-visszakeresésre tanítani, megismertetni velük a különböző lehetőségeket, felhívni figyelmüket az általános hibákra.

## **5.2. A web megoszlása**

A Világhálón át elérhető dokumentumok száma több milliárdra tehető. A felhasználók a szabadszavas keresők alkalmazásával szinte minden témáról kimerítő mennyiségű tudásanyagot találnak a weben, éppen ezért hajlamosak azt gondolni, hogy a szoftverek segítségével minden dokumentumot fellelnek, ami a weben át elérhető [Princz,2006].

1998-ban publikálták az első eredményeket arról, hogy az egyes keresők a web hány százalékát indexelik [Lawrence & Giles,1998]. Az adatok szerint az akkori 6 legnépszerűbb és legjobb eredményt felmutató keresőszoftver együttesen a webnek mintegy 60%-át fedte le, egyenkénti eredményük a következő volt:

| <i><b>Kereső</b></i> | <i><b>A látható web hány százalékát indexeli? (1998)</b></i> |
|----------------------|--------------------------------------------------------------|
| HotBot               | 34%                                                          |
| AltaVista            | 28%                                                          |
| Northern Light       | 20%                                                          |
| Excite               | 14%                                                          |
| Infoseek             | 10%                                                          |
| Lycos                | 3%                                                           |

**6. táblázat. A web lefedettsége a legjobb eredményt mutató szabadszavas keresők által (1998)**

A következő évben a legjobb eredményt a Northern Light mutatta fel, de a webnek már csak 16%-át indexelte, s az akkori első 10 keresőszoftver együttvéve is a webnek csupán 42%-át fedte le [Lawrence & Giles,1999]. A web robbanásszerű növekedésével a keresők sokáig nem tudtak lépést tartani. A 2000-es évek elejére mérséklődött a web növekedése, és a keresők a nyilvános web egyre nagyobb részét indexelik [Gulli & Signorini,2005]:

| <i><b>Kereső</b></i> | <i><b>A látható web hány százalékát indexeli? (2005)</b></i> |
|----------------------|--------------------------------------------------------------|
| Google               | 76%                                                          |
| Yahoo                | 69%                                                          |
| MSN Beta             | 62%                                                          |
| Ask (Teoma)          | 58%                                                          |

**7. táblázat. A web lefedettsége a legjobb eredményt mutató szabadszavas keresők által (2005)**

A nyilvános, indexelhető web lefedettségének aránya az egyes keresők esetén különböző, a teljes lefedettséget egyik kereső sem éri még el. A 7. táblázatban szereplő 4 kereső indexállományának közös részét Gulli és Signorini a teljes indexelhető web 28,85%-ra, mintegy 2,7 milliárd weboldalra becsülte, míg az indexek összegét 9,36 milliárd weboldalra tartják. Ez azt jelenti, hogy célszerű minél több keresőt használni, ha valamely témában alaposan át szeretnénk nézni a weben tárolt dokumentumok tömegét.

A weben számos olyan dokumentum, információ is elérhető, amelyek megtalálásában a szabadszavas keresők nem segítenek. Éppen ezért fontos tudatosítani a webes keresések tanításánál, hogy a web kétfelé osztható: A

szabadszavas keresők által indexelt, s így kereshető részre, valamint egy méretében jóval nagyobb, e szoftverek által nem kereshető részre.

### **5.2.1. A látható web**

A látható web (Surface Web vagy Visible Web) a web azon része, amely a szabadszavas keresők által begyűjtött és indexelt dokumentumokat tartalmazza.

### **5.2.2. A láthatatlan web**

A láthatatlan web (Invisible Web) részét képezik mindazon dokumentumok, amelyek a keresőszoftverek számára láthatatlanok. Ezen dokumentumokat a keresőszoftverek nem tudják (technikai korlátok), vagy nem akarják (elvi döntések miatt) adatbázisukba beemelni.

Különböző becslések találhatók a weben a láthatatlan web nagyságáról, de ezek leggyakrabban Bergman megállapításait idézik [Bergman,2001]. E szerint

- a láthatatlan web 500-szor nagyobb, mint a látható;
- az új információk leggyorsabban gyarapodó kategóriáját teszi ki,
- 95%-a nyilvánosan hozzáférhető,
- a láthatatlan web anyagának több mint a fele tematikus adatbázisokban található.

Bergman cikkének megjelenése óta a legnépszerűbb keresők addig nem indexelt fájl típusokat is indexelnek (pl. .pdf, .doc, .xls, .ppt, .ps, .rtf), s ez nyilván csökkenti a láthatatlan web méretét, de az még napjainkban is sokszorosa a látható web méretének.

A láthatatlan web részei:

#### **5.2.2.1. Nem átlátszó web**

Azon dokumentumok tartoznak ide, amelyek könnyen elérhetőek lennének a begyűjtő robotprogramok számára, de valamely okból mégis kimaradnak a keresőszoftverek adatbázisaiból (Opaque Web).

Ennek oka lehet a begyűjtés mélységének korlátja, az egy domainről begyűjthető dokumentumok számának korlátja, a frissítési periódus alatt megváltozott, megjelent oldalak, a dokumentumszigetek.

### 5.2.2.2. Magán web

Magán web (Private Web) részét képezik azok a dokumentumok, amelyek az Interneten át elérhetőek, de különböző megoldásokkal a nyilvánosan látható oldalak közül ki vannak zárva (pl. az intranet hálózatokon található dokumentumok).

A kizárás történhet a robot.txt fájl alkalmazásával, vagy meta-tag-ek használatával a *Standard for Robot Exclusion* előírásnak megfelelően [The Web Robots FAQ], de tűzfallal, jelszóval való védelem is lehetséges.

A robots.txt fájlt a HTTP szerver gyökérkönyvtárában kell elhelyezni, így tartalmát csak a rendszeradminisztrátorok tudják megadni. A robots.txt fájl használatával az összes robot kizárása a teljes webhelyről az alábbi két sor megadásával érhető el:

```
User-agent: *  
Disallow: /
```

A robots.txt használatával elérhető az is, hogy a kizárás csak bizonyos robotokra, vagy csak bizonyos könyvtárakra terjedjen ki:

```
# a http://nev.hu webhely /robots.txt fájlja  
User-agent: webcrawler  
Disallow: /  
User-agent: *  
Disallow: /tmp  
Disallow: /logs
```

A fenti példa első sora megjegyzés, majd a webcrawler nevű robotot utasítja ki az egész webhelyről. A következő utasítás a többi robotnak tiltja meg a /tmp és a /logs könyvtárak meglátogatását.

Az egyes oldalak tulajdonosainak az oldalaik fejlécében megadott metaadatok használatával lehetőségük van a robotok kizárására a webhely általános irányelveitől függetlenül is. Az egyes oldalokról a robotok a noindex beállításával zárhatók ki, míg a nofollows érték a robotot arra utasítja, hogy ne elemezze és kövesse az oldalon található hivatkozásokat.

```
<html>  
<head>  
<meta name="robots" content="noindex,nofollow">  
<meta name="description" content="Az oldal tartalma ....">  
<title>...</title>
```

```
</head>  
<body>  
...
```

#### **5.2.2.3. Szabadalmazott web**

A szabadalmazott web (Proprietary Web) kategóriája alá olyan adatbázisok, dokumentumok tartoznak, amelyeket tartalomszolgáltatók állítanak elő, és csak regisztráció után, előfizetéssel vagy külön díjért tekinthetők meg. A jelszót vagy login nevet igénylő bejelentkezés technikai korlát a robot számára.

#### **5.2.2.4. Mély web**

A mély web (Deep Web) a láthatatlan web legnagyobb részét teszi ki. A mély web a weben át szabadon elérhető adatbázisokat tartalmaz (pl. menetredek, sárga oldalak, tőzsdei információk, időjárás, stb.). Az adatbázisokban tárolt adatok lekérdezését lekérdező nyelv teszi lehetővé. A lekérdezés megadásakor valamely karaktersorozatot be kell gépelni, vagy tulajdonságok egy sorozatát ki kell választani, amire a begyűjtő robotok nincsenek felkészítve, így ezen adatbázisok tartalma technikai korlát miatt láthatatlan a szabadszavas keresők számára.

Lekérdezhető webes adatbázisok tartalma, azaz a dinamikusan generált web oldalak teljesen vagy részlegesen láthatók vagy láthatatlanok a begyűjtést végző robotprogramok számára attól függően, hogy mennyit tartalmaz abból egy hiperhivatkozás által elérhető statikus oldal. Ha nem mutat egy hiperhivatkozás a web oldalra, akkor az a robot számára láthatatlan.

#### **5.2.2.5. Az igazán láthatatlan web**

Az igazán láthatatlan webet (Truly Invisible Web) azon dokumentumok alkotják, amelyeket elsősorban formátumuk miatt képtelenek a szabadszavas keresők indexelni (pl. audió, videó, képfájlok). A szabadszavas keresők a dokumentumok, vagy a dokumentumra mutató hivatkozások szövegét indexelik, a nem szöveges állományok tartalma alapján nem tudnak visszakeresni. Ezért szükséges a nem szöveges dokumentumokhoz valamilyen szöveget, leírást kapcsolni, ami alapján ezen állományok visszakereshetők a weben. Erre példa a

képek leírására használható ALT elem, valamint az image mapben megadható leírások, bár ezen elemeket sem indexeli minden kereső.

Néhány szöveget tartalmazó fájl formátumra (pdf, flash, office fájlok, stb.) is igaz, hogy időigényes indexelni őket, ezért egyes szabadszavas keresők kizárják ezen típusú dokumentumokat.

Az egyes szolgáltatók döntése alapján kerülnek teljes vagy részleges kizárásra a scriptek által generált dinamikus tartalmak. Ezen oldalak URL címe tartalmazza a kérdőjelet. (pl. <http://www.forums.com/viewtopic.php?t=1130>). A kizárás oka, hogy a kereső robot nem tudhatja biztosan, nem egy végtelen ciklust eredményező csapda-e a script által generált oldal, amely a fenntartónak idő- és pénzvesztést eredményezne, így megelőzés végett kerül sor az oldal kizárására.

Néhány kereső nem támogatja a frame szerkezetet, ezért a framekben megadott dokumentumokat sem indexeli (pl. MSN, HotBot).

A láthatatlan web kereshetővé tételére különböző technikai megoldások születtek [Princz, Rutkovszky, 2004]. A láthatatlan web részét képező források, adatbázisok elérését segítik a különböző katalógusok, hivatkozások rendszerezett gyűjteményei. Ilyen található például a *www.invisible-web.net* címen.

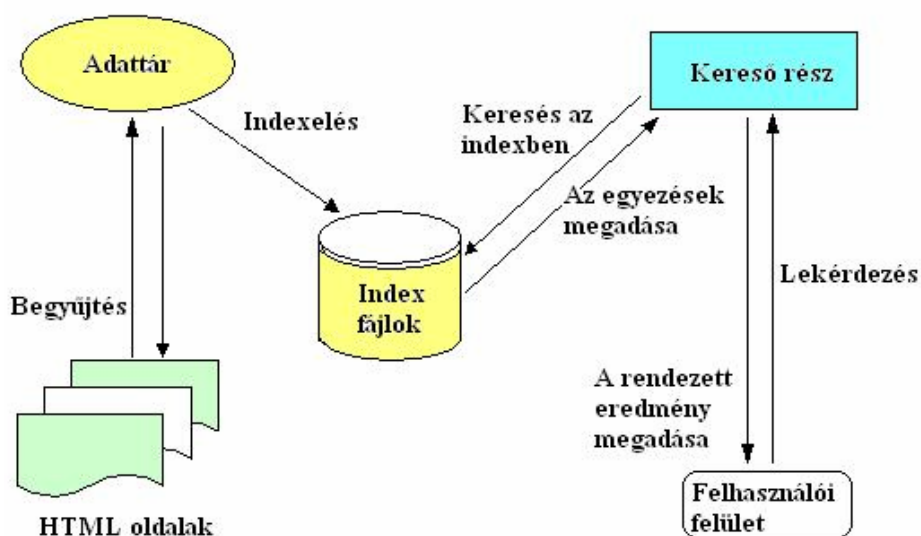
A láthatatlan web nagyon gyors mértékben növekszik, így tartalmának lekérdezhetővé tétele sürgető feladat.

### 5.3. Szabadszavas keresők

Az angol *search engine* kifejezés magyarra fordítására a *szabadszavas kereső* kifejezésen kívül előfordulnak még a *keresőgép*, *keresőmotor*, *keresőrendszer*, *keresőszoftver* kifejezések is. Ez az információ-visszakereső architektúra úgy modellezi a webet, mint egy teljes szövegű adatbázist. A hagyományos információ-visszakereső rendszertől az különbözteti meg, hogy visszakeresést a szöveg elérése nélkül, az adatok előfeldolgozásával valósítja meg. A weboldalak jelenlétéhez vagy az oldalakról kellene egy-egy másolatot helyben tárolni (ez igen hely és költségigényes lenne), vagy lekérdezéskor a hálózatot használva kellene elérni a távoli oldalakat (túl lassú).

Napjainkban a szabadszavas keresők többsége üzleti jellegű, amely szabadalmazott technológiát is jelent egy erős versenypiacon. Ebből következik, hogy pontos működésükről annyit tudunk, amennyit a fejlesztők biztonságosnak tartanak közzétenni, s ez általában a technikai részletekről kevés publikációt jelent. Az egyik legrészletesebb nyilvános leírás a Google-ról érhető el [Brin & Page,1998].

A legtöbb szabadszavas kereső működése a központosított robot-indexelő architektúrán alapul. A keresés a tárolt indexekben történik, amelyek a web robotok által begyűjtött dokumentumok alapján épülnek fel. A szabadszavas keresők működését szemlélteti a 16. ábra.



16. ábra. A szabadszavas keresők működésének általános folyamata

A szabadszavas keresők a következő főbb funkcionális részekből állnak: a begyűjtő robot, az indexelő eljárás valamint a kereső rész.

### 5.3.1. Begyűjtés

A begyűjtés feladata a web oldalainak letöltése és háttértárolóra írása. A folyamat közben két adatbázis készül. Az egyikben az oldalakon szereplő információkat, azaz a HTML fájlokat írja egymás után a rendszer. Erre az

adatbázisra az indexelés folyamán lesz szükség, tartalma a feldolgozás után törlődik. A másik adatbázisban a rendszer a hiperlinkek strukturális adatait (webgráf) tárolja, amely alapján a begyűjtés útvonalát szervezi, és amelyet később a rangsoroló és klaszterező eljárásoknál is felhasználhat. A begyűjtést robot programok végzik.

A web robotok (crawlers, spiders, wanderers, knowbots) olyan számítógépes programok, amelyek a helyi rendszeren futnak, és kéréseket küldenek távoli web szervereknek, majd a kapott válaszokat a kereső indexelő része felé továbbítják. Jellemzőik:

- Önállóság: bizonyos előre meghatározott korlátok között a működésüket emberi beavatkozás és irányítás nélkül végzik.
- Rekurzivitás: egy adott weboldaltól kiindulva képesek bejárni az összes olyan oldalt, amely a kiindulási ponttal közvetlen kapcsolatban áll, majd a kapcsolt oldalakat, mint kiindulási pontot tekinteni és a folyamatot egészen addig folytatni, amíg valamilyen kilépési feltétel érvényre nem jut.

A szabadszavas keresők robotprogramjai különböznek abban, hogy a begyűjtésnél mely szervereket tekintenek kiindulási pontnak, az információ kigyűjtésének mely módszerét alkalmazzák, milyen frissítési periódust használnak.

Míg kezdetben a szabadszavas keresők egy része csak HTML oldalakat gyűjtött indexelésre, addig ma már a legnépszerűbb keresők a Postscript, PDF, Word, Excel, PowerPoint, text formátumokat is begyűjtik és indexelik.

A robot begyűjtés közben egy prioritási sort kezel, amely a még letöltetlen oldalak URL címeit tartalmazza fontossági sorrendben. A sor elejéről kivett címekhez tartozó oldalakat letölti, és kigyűjti belőlük a hiperlinkeket. A kapott linkekről eldönti, melyeket kell követni, ezeket beteszi a sorba, a többit figyelmen kívül hagyja.

Az információ begyűjtésénél a keresők különböznek abban, hogy egy domainről kiindulva mélységi vagy szélességi bejárást alkalmaznak a domainről induló hivatkozások felkeresése során [Breadth first search and depth first search,1996]. Szélességi bejárást alkalmazva egyből látható az összes olyan oldal, amely az aktuális oldalhoz van kapcsolva, így gyorsan elérhetők a hasonló témakörű webhelyek. Ez a fajta bejárás széles terjedelmet, de keresésnél felszínes eredményt biztosít. A mélységi bejárásnál az oldal első hivatkozását követi végig a

robot, s amikor már nincs további hivatkozás, tér vissza rekurzívan a fentebbi szintre. Ez a bejárás keskeny, de mély haladást tesz lehetővé. Általában igaz, hogy mennél nagyobb egy szabadszavas kereső indexállománya, annál valószínűbb, hogy mélységi keresést hajt végre a robotja (pl. AllTheWeb, Google, Inktomi). Számos kereső felső korlátot ad meg, hogy egy domainról hány dokumentumot indexel (pl. Alta Vista, Teoma).

Fontos szempont, hogy milyen sűrűn írja felül saját adatbázisát egy kereső, végignézve újra azokat a webhelyeket, amelyeken egyszer már járt. Ez a paraméter jelentősen eltér a különböző keresőknél: egy héttől akár több hónapig is terjedhet, így ha közben módosulnak, megszűnnek az indexelt oldalak, a szabadszavas kereső adatbázisában még a régi adatok szerepelnek (dead links). A frissítésnél alapvetően két stratégiát követnek a robotok: Az egyik szerint minden weboldalt ugyanolyan gyakorisággal keresnek fel (egységes kiosztási eljárás), a másik módszer, amikor a gyakrabban módosuló, fontosabb oldalakat sűrűbben látogatják és indexelik (arányos kiosztási eljárás). Kimutatható, hogy az egységes kiosztási eljárás a hatékonyabb [Friedman,Uher,Windhagerl,2003].

A robotoknak a begyűjtés folyamán be kell tartaniuk bizonyos működési és etikai szabályokat. Ezek közé tartozik a webhely, valamint az egyes oldalak hozzáférhetőségének korlátozása a webhely robot.txt fájlja, illetve meta-tag-ek által, valamint figyelembe kell vennie azt a szabályt, hogy csak 60 másodperc elteltével küldhet újabb kérést ugyanannak a HTTP szervernek.

A legtöbb szabadszavas kereső lehetővé teszi, hogy a felhasználók saját nyitólapjukat regisztráltathassák, azaz web szerverük, nyitólapjaik URL címét átadják a kereső robotprogramjának, bár ez nem jelent automatikus bekerülést a kereső adatbázisába. A legtöbb keresőnél azonban különböző nagyságú regisztrációs összegek befizetése mellett biztosítható, illetve gyorsítható az adatbázisba való bekerülés [SearchEngineWatch].

Példa a robot feladatait megvalósító alkalmazásokra:

- Lynx: A Lynx a Unix operációs rendszer alatt futó szöveges böngésző.

Például a

```
lynx -dump -source http://www.w3c.com/
```

parancs letölti a <http://www.w3c.com/> weboldal forráskódját.

- Java.net: A java.net programcsomag számos hálózati segédprogramot biztosít. A programcsomag két osztályát, a *java.net.URL*-t és a *java.net.URLConnection*-t weboldalak letöltésére lehet használni.
- CPAN (Comprehensive Perl Archive Network): A CPAN weboldalain (<http://www.cpan.org>) számos script található, amely hasznos robotprogramok létrehozásakor.

### 5.3.2. Indexelés

A szabadszavas keresők a weben lévő dokumentumok indexelésére, és a dokumentumokban való keresésre az invertált fájlok módszerét alkalmazzák.

Az indexelő rész elemzi a robot által begyűjtött oldalakat: Kigyűjti a dokumentumban előforduló szavakat (szójegyzék), és feljegyzik a szó előfordulását (invertált lista). Az invertált listák rendezett sorrendben tartalmazzák azon dokumentumok azonosítóit, amelyekben a szó előfordul, valamint a dokumentumokon belül az előfordulás pozícióját. Egyes keresők eltérő mértékben további információkat is feljegyeznek a szó előfordulásával kapcsolatban. Ezen metaadatokat a rangsorolásnál felhasználják, illetve a felhasználó számára lehetővé teszik ezen helyeken is a keresést:

- A szabadszavas keresők feljegyzik a szó megnevezését (<TITLE> megnevezés </TITLE>), s rangsorolásnál kiemelten kezelik a szó előfordulását ezen a helyen. Sok népszerű kereső lehetővé teszi a keresést a dokumentum nevében (pl. AllTheWeb, Yahoo, Google, Ask Jeeves).
- Az URL címeket (host, könyvtár, file név) tartalmazzák a keresők adatbázisai, s néhány keresőnél keresési szempontként külön is megadható (pl. Yahoo, Google, Ask Jeeves).
- Egyes keresők egyéb HTML elemeket is indexelnek: például címsorokat, lapon belüli hivatkozást (<A> és </A> közötti rész), kiemelt szövegrészeket (pl. Google).
- A *keywords*, *description*, *ALT* metaadatokat a szabadszavas keresők különböző mértékben építik be az adatbázisukba. Erről a 6.1.1.3 Meta jelölő elemek fejezetben bővebben írunk.
- A stop szavakat (pl. névelők, elöljárók, számok) a szabadszavas keresők különböző mértékben zárják ki az indexelésből.

A keresők különböznek azon szavak gyűjteményének nagyságában, amelyeket indexelnek, azaz tárolják az előfordulási helyüket. A Google esetében ez több, mint 14 millió szót jelent. A legtöbb kereső ma már a szöveg egészét indexeli, a szöveg bizonyos részeit pedig kiemelten kezeli.

### 5.3.3. Keresés

Amikor a felhasználó elindít egy lekérdezést, a kereső az index segítségével kiválogatja a kritériumoknak megfelelő oldalakat, a hozzájuk társított metainformációk alapján sorrendbe állítja őket, és az összes találat közül az első néhányat visszaadja a felhasználónak.

A legtöbb kereső a “legjobb egyezés” elvét használja a visszakereséseknél, néhány csak a pontos egyezéseket tekinti eredménynek. Egy-két szolgáltató az eredmény javítására használja a relevancia visszajelzését, klaszterezést (hasonlók keresése, találatok csoportosítása).

#### 5.3.3.1. Lekérdezések megadása

A kereső résznél lényeges, hogy a szabadszavas kereső milyen lehetőségeket nyújt adatbázisának lekérdezéséhez. A legáltalánosabb lehetőség, hogy a lekérdezési interfész egy szövegmező, amelybe egy vagy több szót lehet beírni. A legtöbb szabadszavas kereső az egyszerű lekérdezéseken túl az *összetett lekérdezések* lehetőségét is biztosítja, ami egyfajta parancsnyelvnek tekinthető. Összetett keresésnek nevezzük, ha a keresőszavakat logikai műveletekkel kapcsoljuk össze, vagy ha szűkítő előtagokat írunk a keresőszavak elé, valamint ezek kombinációit.

Az összetett lekérdezéseknél használható operátorok, kiegészítő feltételek:

- Logikai operátorokkal (AND, OR, BUT) köthetjük össze a keresőszavakat. Az egyszerűen szóközzel egymás mellé írt szavakat a legtöbb kereső úgy értelmezi, mintha logikai ÉS-sel (AND) kapcsoltuk volna össze őket. A szó előtti + jel használatával előírhatjuk, a – jellel kizárhatjuk a szó jelenlétét az eredményhalmazban.
- A kulcsszó-egyezéses kereséseken kívül kifejezések keresésére is lehetőség van a legtöbb keresőnél. Ekkor a szavak sorrendje lényeges és a kifejezést idézőjelek közé kell zárni.

- Néhány keresőnél előírhatjuk, hogy egy adott szó környezetében kell előfordulnia a keresett szónak (NEAR operátor). A szavak közötti távolság keresőnként változhat: 2-től (WebCrawler) akár 25 szóig (Lycos) is terjedhet.
- Keresési megszorítások is megadhatók: Előírhatunk a dokumentumok létrehozására vonatkozó időkorlátot, megadhatjuk a dokumentum nyelvét, típusát, méretét, előfordulási helyét domainenként, site-onként, földrajzi helyenként.
- Számos keresőnél kereshetünk csak a dokumentum nevében, a hiperhivatkozások között, illetve az URL alapján.
- Egy-két keresőnél lehetőség van a szavak végének levágására. Ekkor a *rádió\** keresés eredménye egyaránt lehet a *rádió*, *rádiózás*, *rádióbullám* szavak valamelyikét tartalmazó dokumentum.
- Kereséskor érdemes arra is figyelni, hogy néhány kereső kis- és nagybetű érzékeny.

A leggyakoribb típusú összetett kereséseket a *részletes keresés* oldalon lehet könnyen megadni anélkül, hogy ismernünk kellene az összetett lekérdezés megfogalmazásának pontos szabályait. Napjainkban már számos szabadszavas kereső felkínálja ezt a lehetőséget (pl. Google, AllTheWeb, AltaVista, Yahoo, Vizsla24).

**Google** **Speciális keresés** [Keresési tippek](#) | [Minden a Google-ről](#)

**Találatok keresése** Tartalmazza-e ezen szavak **mindegyikét**:  10 találatok   
 Tartalmazza-e pontosan ezt a **kifejezést**:   
 Tartalmazza-e ezen szavak **legalább egyikét**:   
 Nem tartalmazza-e ezeket a szavakat:

**Nyelv** Ilyen nyelvű lapok közt keresés:  bármely nyelv

**Fájlformátum**  bármilyen formátum

**Dátum** Ekkor frissített weblapok:  bármikor

**Előfordulások** Olyan találatok megjelenítése, amikben a keresett kifejezések itt fordulnak elő:  a lapon bárhol

**Tartomány**   [További információ](#)

**Lapfüggő keresés**

**Hasonlóság** Olyan lapok keresése, melyek hasonlítanak erre a lapra:    
 Pl.: [www.google.co.hu/intl/hu/help.html](http://www.google.co.hu/intl/hu/help.html)

**Hivatkozások** Olyan lapok keresése, melyek erre a lapra hivatkoznak:

©2006 Google

17. ábra. A részletes keresés ablaka a Google-nél

Az operátorok explicit megadása helyett a megfelelő beviteli mezők kiválasztásával fogalmazhatók meg a kulcsszavak közötti logikai kapcsolatok, adhatók meg a szűkítések, így tehermentesítve a felhasználókat a szintaktikai szabályok ismerete alól. A részletes keresést biztosító oldal általában a nyitólapon található hivatkozáson keresztül érhető csak el. Ennek az az oka, hogy sok felhasználó első ránézésre elbizonytalanodik a felkínált, s számára bonyolult lehetőségektől, s inkább nem is használja az oldalt, így - a felmérések hatására - ezt a lehetőséget száműzték a nyitólapról.

A keresések megadásakor használható operátorok, kiegészítő feltételek szabadszavas keresőnként változnak, így egy-egy új szabadszavas kereső használata előtt célszerű megismerni annak lehetőségeit.

### **5.3.3.2. Rangsorolás**

A legtöbb szabadszavas kereső a Boole és/vagy a vektortér modell variációit használja a rangsorolásnál. Ahogy a keresés, úgy a rangsorolás is a szöveg elérése nélkül, csak az indexek felhasználásával valósul meg. Valamennyi kereső úgy rendezi a keresés eredményét, hogy az eredmény lista elejére az általa legfontosabbnak tartott dokumentumok kerüljenek. A rangsorolási algoritmusok keresőnként különböznek, s szigorúan titokban tartják őket.

A szabadszavas keresők rangsorolásnál alkalmazzák a klasszikus tf-idf sémát, azaz az eredményoldal relevanciájának meghatározásakor figyelembe veszik a kulcsszó előfordulásának helyét (a dokumentum nevében, címsorokban, az első bekezdésekben, stb.), valamint az előfordulás gyakoriságát. E feltételek meglétét különböző súlyozással tekintik a keresők. A valóságban azonban bizonyos kulcsszavak túlsúlya nincs mindig arányban az oldal jelentőségével. Éppen ezért egyre több kereső a keresési kulcsszavak számolgatása mellett új megoldásokat is alkalmaz egy-egy oldal fontosságának meghatározásakor (második generációs szabadszavas keresők). A klasszikus tf-idf séma rangsorolási eredményének megjavítására több változtatást is kifejlesztettek:

#### **5.3.3.2.1. A keresés irányultságának fogalmi felismerése**

Ha egy adott kulcsszó egyszerre több témát is felölel, akkor az eredménylista tetején a népszerűbb téma fog szerepelni, míg a keresett jelentésben esetleg csak a lista végén tűnnek fel oldalak. Ezt a jelenséget

témasodródásnak nevezzük (topic-drift). A problémára megoldást jelenthet, ha téma szerint csoportosítjuk, azaz klaszterezzük az oldalakat.

A klaszterezés a felhasználó számára több szempontból is előnyös:

- Áttekintést kap az elérhető témákról, témakörökről.
- A kisebb témához tartozó oldalak sem vesznek el a hosszú találati listákon.
- A mappában együtt látja a hasonló eredményeket, s nem kell a hosszú listán keresgélnie.

Klaszterező technikát használó keresőkre példa a Northern Light, amely webhely és/vagy tartalom alapján csoportosítja az oldalakat, valamint a Vivisimo Clusty nevű metakeresője, amely a különböző keresőkből származó találatokat csoportokba is rendezi.

A témasodródás elkerülésére egy másfajta megoldás, ha a kereső a feltett kérdés alapján megpróbálja kitalálni, hogy valójában mire is kíváncsi a felhasználó, s ennek megfelelően szolgáltatja az eredményoldalakat. Ez a megközelítés az alkalmazott jelentéstan, a természetes nyelvi feldolgozás alkalmazását igényli.

E technikát alkalmazó kereső az Ask Jeeves. Az Ask Jeeves adattárában számos, különböző tárgyerületek szakértői által előre megfogalmazott kérdés van, amelyekhez hozzárendeli a legrelevánsabb válaszoldalakat. A kereső elemzi a felhasználó által feltett kérdést, majd a hozzá legjobban hasonlító előredefiniált kérdés alapján szolgáltatja a válaszokat.

#### **5.3.3.2.2. Hivatkozások analízise**

Az új rangsoroló algoritmusok közül több a hivatkozások tartalmazta információt használja fel az eredményoldalak rangsorának kialakításához. Ezek közé tartozik Kleinberg HITS algoritmus [Kleinberg,1998] és a Google által használt PageRank [Brin & Page,1998].

##### **5.3.3.2.2.1. HITS algoritmus**

A HITS (Hyperlink-Induced Topic Search) algoritmus az egyes csomópontok fontosságáról a weboldalak közötti hivatkozásokból nyer ki információt. Az algoritmus minden egyes kérdéshez először összegyűjti a válaszoldalakat, majd rekurzívan kiszámítja minden oldal súlyát. Alapötlete a releváns oldalak felbontása két kategóriára: a keresett téma szempontjából

központi (hub), illetve tekintély (authoritive) oldalakra. Tekintély oldal például egy színvonalas szakmai portál, vagy egy elismert kutató saját publikációit tartalmazó személyes honlapja; központi oldal egy hírportál, linkgyűjtemény vagy katalógus. A csoportosítás egy rekurzív definíció alapján történik: a központi oldalak azok, amik sok tekintély oldalra mutatnak, a tekintély oldalak pedig azok, amikre sok központi oldal mutat. A két csoport meghatározása iteráció alkalmazásával megy végbe: minden oldalhoz rendelünk két számot, amik az oldal központiságát, illetve tekintélyét jelzik. Kezdetben a számokat tetszős szerint választjuk (például minden oldalnak ugyanazt), majd minden iterációban egy oldal tekintélye a rá mutató oldalak központiságának összege, és a központisága az általa mutatott oldalak tekintélyének összege lesz. Az eredményoldal rangsorolása az oldal hub és authoritative értéke alapján történik.

A HITS által visszaadott eredmények nagyon jól használhatóak közösségi szűrésre. Ha két oldal ugyanazokra, vagy majdnem ugyanazokra az oldalakra mutat, akkor a két oldal valószínűleg azonos témájú, illetve azonos közösséghez tartozik. Ugyanez igaz arra a két oldalra, amire ugyanazok az oldalak mutatnak.

#### **5.3.3.2.2. PageRank**

A PageRank a Google szabadszavas kereső legfontosabb eleme. A Google arra a feltételezésre épít, hogy a weboldalak készítői általában azokra az oldalakra linkelnek a saját lapjukról, amiket jónak tartanak, vagyis minden hiperlink felfogható egy-egy szavazatként a céloldalra. Minél több szavazatot kap egy oldal, annál fontosabb, s a népszerűbb oldalról történő külső hivatkozásokat súlyozottan veszi figyelembe.

A PageRank algoritmus a hiperlinkekkel összekötött dokumentumokhoz számokat rendel azoknak a hiperlink-hálózatban betöltött szerepe alapján. Ezt a számot szintén PageRank-nek nevezik, s egy weboldal fontosságának a számszerűsített értéke.

Az A oldalhoz tartozó PageRank kiszámítása a következő elv alapján történik: Tekintsük az A oldalra hivatkozó  $T_1, \dots, T_n$  weboldalakat. Legyen  $d$  egy 0 és 1 közé eső paraméter, aminek az értéke általában 0,85. Jelölje  $C(A)$  az A oldalról kimutató hivatkozások számát. Ekkor az A oldalhoz tartozó PageRank érték

$$PR(A) = (1 - d) + d \times \sum_{i=1}^n \frac{PR(T_i)}{C(T_i)}$$

formulával határozható meg. A PageRank értéke valójában a normalizált link-mátrix legnagyobb sajátértékéhez tartozó sajátvektornak felel meg, így viszonylag gyorsan számolható a megfelelő iteratív algoritmussal.

A PageRank egy valószínűségi eloszlást határoz meg a weboldalakon, az összes oldal PageRankjének összege 1. Annak a valószínűsége, hogy egy hosszú böngészés után a felhasználó éppen egy adott oldalt néz, az adott oldal PageRankjével egyenlő. Egy oldalnak akkor lesz a PageRankje magas, ha rá sokan mutatnak, vagy vannak olyan oldalak, amelyeknek nagy a PageRankje, és rámutatnak.

### 5.3.3.2.3. HITS és PageRank összehasonlítása:

A HITS első ránézésre nagyon hasonlónak tűnik a PageRank algoritmushoz, de van egy fontos különbség közöttük: a PageRank egy téma független, csak a linkstruktúra által meghatározott érték, a HITS viszont mindig egy konkrét témára nézve keres (tipikusan egy másik, vektortér-modell alapú kereső által visszaadott találatok között). Ez azt jelenti, hogy egyrészt a HITS eredménye sokkal pontosabb és relevánsabb lesz, másrészt viszont minden egyes kérdésnél újra ki kell számolni, tehát kevésbé hatékony. Mindkét algoritmus hibája, hogy nem szünteti meg a témasodródást.

### 5.3.3.2.3. Az oldal népszerűségének elismerése

Számos kereső az eredménylista sorrendjének meghatározásánál figyelembe veszi az eredményoldalak népszerűségét. (pl. Google, DirectHit).

A második generációs szabadszavas keresőkre jellemző, hogy egy oldal helyezését a rangsorban előnyösen befolyásolja az oldal népszerűsége. Ha ehhez hozzávesszük, hogy a felhasználók ritkán böngésznek a keresési eredménylista második oldalán túl, akkor az a várható trend, hogy az eddig is népszerű oldalak még népszerűbbek lesznek, az új oldalaknak pedig egyre nehezebb lesz jó helyezést elérni.

Az eredménylistán szereplő dokumentumok rangsorát meghatározó tényezőkről, a rangsor befolyásolásáról a Weboldalak optimalizálása fejezet foglalkozik.

### 5.3.3.3. Eredmények megjelenítése

A lekérdezésnek megfelelő dokumentumokat a szabadszavas keresők a rangsorolás után rendezve jelenítik meg. Rendszerint az eredménylisták tartalmazzák a dokumentum megnevezését (TITLE), címét (URL), valamint rövid kivonatot a dokumentumról, az állomány típusát, méretét, az utolsó módosítás dátumát.

Néhány szabadszavas kereső lehetővé teszi, hogy a felhasználók további információk megjelenítését is kérhessék (pl. a dokumentumnak a kereső adatbázisába való bekerülésének az időpontját), illetve beállíthassák, hogy egy oldalon hány eredményt jelenítsen meg a kereső.

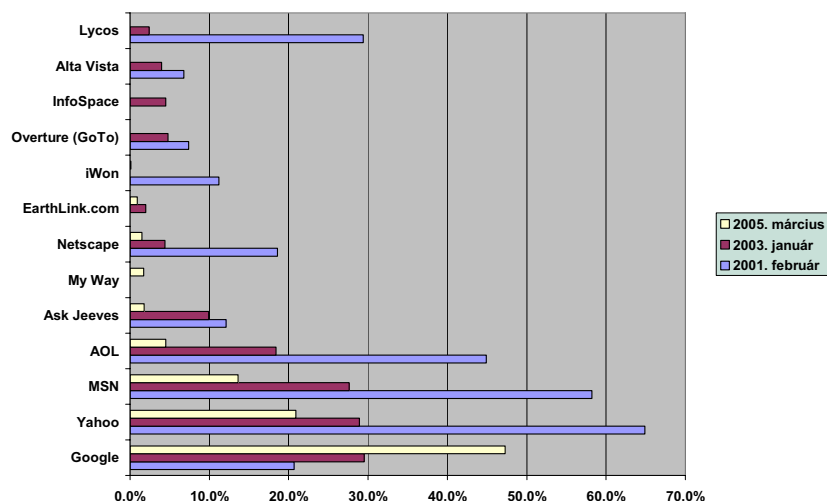
A dokumentumok összegzésénél különböző szemléletek figyelhetők meg: Számos szabadszavas kereső a szerzők által megadott metaadatokat használja az összegzésnél, vagy a dokumentum azon mondatát jeleníti meg, amely a lekérdezéssel egyezést mutat, esetleg az oldal szövegének első 100-200 karakterét írja ki.

### 5.3.4. Népszerű szabadszavas keresők

Több cég minősíti a kereső szolgáltatókat különböző szempontok szerint. Az egyik gyakori szempont, hogy mekkora az indexállományuk, azaz hány millió web oldalból építik fel adatbázisukat? Egy másik gyakori szempont, hogy mely kereső szolgáltatót használják a leggyakrabban egy-egy keresés végrehajtásakor?

Az USA-ban a legnépszerűbb szabadszavas keresők között a comScore Media Metrix 2004. decemberi, illetve a Nielsen NetRatings 2005. márciusi felmérése szerint [SearchEngineWatch] egyaránt a Google áll az első helyen, majd a sorrend Yahoo, MSN, AOL mindkét felmérés szerint. Európában ezen szolgáltatók mellett minden országban dobogós helyen áll valamely nemzeti szolgáltató [SearchEngineWatch]. Így például Németországban a T-Online, Olaszországban a Virgilio, a Libero, Franciaországban a Wanadoo, stb. Magyarországon az Origo által működtetett Vizsla24 a legnépszerűbb magyar kereső, de a Heuréka és a Góliát magyar nyelvű keresőt is sokan használják.

Egy-egy kereső szolgáltató népszerűsége időről időre változik, ahogy azt a következő ábra is mutatja.



18. ábra. A legnépszerűbb kereső szolgáltatók az USA-ban

## 5.4. Tematikus keresők

A fogalomra az angol nyelvben is több kifejezés található (subject directories, web directories, catalogs), s a magyar nyelvben is több szó szolgál a fogalom megnevezésére: tematikus keresők, hivatkozások gyűjteménye, webes könyvtárak, katalógusok, stb.

A tematikus keresők a weben található dokumentumokra mutató hiperhivatkozások hierarchikus gyűjteménye, ahol témakörök szerint felépülő könyvtárakban kereshetünk.

A tematikus keresők legnagyobb problémája a dokumentumok katalogizálása. A szövegek automatikus megértésére és osztályozására létrejött megoldások még nem elég hatékonyak, így a dokumentumok témakörönkénti besorolását ma még legtöbbször szakemberek végzik. Ezek a rendszerek nem vállalkoznak a teljes Web katalogizálására. A manuális katalogizálás ugyanakkor lehetőséget teremt a tartalom szerinti szűrésre, azaz a katalógus rendszerektől elvárható a relevánsabb tartalom egy adott témához.

A tematikus keresőknél a keresés történhet böngészéssel, de számos szolgáltató lehetővé teszi dokumentum gyűjteményének adatbázis-szerű

lekérdezését is. Ekkor a témakörökre, a taxonómia leíró elemeire kérdezhetünk rá.

Míg 10 évvel ezelőtt a különböző webes szolgáltatók élesen elkülönültek abban, hogy a weben, vagy saját válogatott dokumentum gyűjteményükben teszik lehetővé a keresést, mára a volt különbségek már nem érzékelhetők: A legtöbb szabadszavas kereső nyitólapjáról biztosítja valamely válogatott, rendszerezett hivatkozásgyűjtemény elérését (pl. Google, AltaVista, Vizsla24), illetve a tematikus keresőként induló szolgáltatók is lehetővé teszik nyitólapjukon a weben történő keresést (pl. Yahoo, LookSmart). A robotok által összegyűjtött és az ember által válogatott anyagokban való keresést egyaránt lehetővé tevő szolgáltatást hibrid keresőnek is nevezik (hybrid search engine).

Számos kereső szolgáltató társul egymással, és eredményeik javítására különböző technológiát, könyvtárakat, index állományokat vásárolnak egymástól: Például az AltaVista a Yahoo könyvtárát kínálja fel nyitólapján, míg a Yahoo régebben az AltaVista, napjainkban a Google indexállományát használja a weben történő kereséshez. Az Open Directory könyvtárat egyaránt használja a Google, AOL, HotBot, Netscape, stb.[SearchEngineWatch, SearchEngines]

A tematikus keresők a webnek szűkebb részét fedik le, mint amit a keresőszoftverek adatbázisai tartalmaznak. A tematikus keresőknél az összegyűjtött anyag jellege lehet akadémiai vagy egyéb szakgyűjtemény (pl. Magyar Elektronikus Könyvtár, InfoMine, CiteSeer), illetve a közönség széles rétegeinek szánt, minél több szolgáltatást nyújtó üzleti, közszolgálati portál (pl. Yahoo, MSN, AOL, Origo).

## 5.5. Metakeresők

A metakeresők olyan web szerverek, amelyek egy adott lekérdezést több szabadszavas keresőhöz, és/vagy adatbázishoz továbbítanak, összegyűjtik a válaszokat, és egyesítve jelentetik meg őket. A metakereső használata a felhasználó számára lehetővé teszi, hogy egyszer adja csak meg a lekérdezést, s utána egy időben több szolgáltató adatai között kereshessen. A metakeresőkre példa: MetaCrawler, Dogpile.

## 5.6. Keresés a hiperhivatkozások alapján

A szabadszavas keresők és a tematikus keresők használatával a weboldalak tartalmában kereshetünk, de lehetséges a keresés a weboldalakot összekötő hivatkozások struktúrájában is, ami a weboldalak közötti hivatkozásokat, az oldalon belüli hivatkozásokat és az URL által megadott struktúrát is tartalmazza.

### 5.6.1. Web lekérdező nyelvek

Néhány példa struktúra alapú lekérdezésekre:

- Egy adott oldalról kiindulva olyan oldalakat keresünk, amelyek legfeljebb három hivatkozást követve elérhetők, és legalább egy képet tartalmaznak.
- Tudni szeretnénk, mik a megnevezései egy adott honlapra mutató oldalaknak?
- Melyik a leginkább hivatkozott weboldal, amely egy adott kifejezést tartalmaz?
- Mely más szervereken található oldalakon tűnik fel ugyanaz a tartalom, mint egy adott honlapé?

Az ilyen típusú kérdések megválaszolásához az eddigiektől eltérő adatmodelleket kell használni, amelyek a teljes szövegben való keresést integrálják a topológián alapuló kereséssel. Ezek közül a legfontosabb a címkézett gráf, amely a weboldalakot tekinti csomópontoknak, az oldalak közötti hivatkozásokat éleknek, és a szemiszerkezturált adat modellt a weboldalak tartalmának. Az adat modell, azaz a séma nem mindig ismert, időnként változhat, nagyobb vagy leíróbb lehet.

A web lekérdező nyelvek első generációjának a célja az volt, hogy kombinálják a dokumentumon belül feltűnő minták keresését a dokumentumok közötti hivatkozások alkotta gráf lekérdezésével. A hivatkozások struktúrája útvonal megadó kifejezésekkel fejezhető ki. E nyelvek közé tartozik a W3QL, WebSQL, WebLog, WebQL.

| Útvonal meg-<br>adó kifejezés | Jelentése                                                           |
|-------------------------------|---------------------------------------------------------------------|
| -> -> =>                      | Három egymás utáni hivatkozás:két lokális linket egy globális követ |
| ->   =>                       | Egy hosszúságú hivatkozás, amely vagy lokális, vagy globális.       |
| ->*                           | Tetszőleges számú helyi hivatkozás egymás után.                     |
| => ->*                        | Egy globális hivatkozást tetszőleges számú helyi link követ.        |
| =   #>   ->                   | Nulla vagy egy hosszú helyi hivatkozás                              |

**8. táblázat. Példa útvonal megadó kifejezésekre WQL-ben**

Példa struktúra lekérdezésekre WQL használatával [Arocena et al,1997]:

A lekérdezés azonos nevű, de különböző címen található dokumentumokat keres:

```
SELECT d1.url, d2.url
FROM Document d1, Document d2
WHERE d1.title = d2.title
AND NOT (d1.url = d2.url)
```

A következő példa szabadszavas keresőkkel kapcsolatos oldalakat keres a Debreceni Egyetem honlapján:

```
SELECT d.url
FROM Document d SUCH THAT " http://www.unideb.hu/" ->* d,
WHERE d.text CONTAINS " szabadszavas keresők "
OR d.title CONTAINS " szabadszavas keresők "
```

A második generációs nyelvek, amelyeket webes adatkezelő nyelvnek hívnak, a hangsúlyt a szemisstrukturált adatra helyezik. E nyelvek az első generációs nyelvek nyújtotta lehetőségen túl lehetővé teszik a weboldalak struktúrájának (azaz a belső struktúrának) az elérését is. E kategóriába tartozó nyelvek a STRUQL, FLORID.

### 5.6.2. Dinamikus keresés és ügynök programok

A weben a dinamikus keresés ekvivalens a szekvenciális szövegkereséssel, a célja pedig releváns információk kinyerése az oldalak közötti hivatkozásokból. A fő előnye az online keresésnek, hogy a web aktuális struktúrájában keres, és nem valamelyik szabadszavas kereső indexében. E módszer a web egészére alkalmazva nagyon lassú, de a web kicsi és dinamikusan változó területén való keresésre már alkalmas. Különböző algoritmusokat dolgoztak ki a dinamikus keresésre (pl. fish search, shark search). Ezek azon megfigyelésen alapulnak, hogy sok esetben a releváns oldalakhoz közeli oldalak is relevánsak. A keresés egy releváns oldalról indul, s bizonyos prioritási szabályok szerint követi a hivatkozásokat a releváns dokumentumokban.

Dinamikus keresést hajtanak végre az ügynök (*agent*) programok, amelyek valamilyen speciális információ után keresnek a weben. Ezek a programok önállóan, a felhasználó beavatkozása nélkül, a felhasználó helyett hajtják végre a keresést.

### 5.6.3. Webbányászat

A web lekérdező nyelveknek a web struktúrájának, valamint a felhasználók web használati szokásainak feltérképezésében van szerepe. A dinamikus kereséssel, ügynökprogramok használatával a web tartalmában kereshetünk. Mindez része a webbányászatnak (web mining).

Az egyre több forrásból – szövegekből, webhelyekről és idősoros tranzakciókból – származó adatokból nem csak kimutatások, hanem prediktív modelleken alapuló előrejelzések is készíthetők, grafikai megjelenítéssel könnyen azonosíthatók a bonyolult összefüggések. A webbányászattal gyorsan meghatározhatók a webhely látogatóinak navigálási jellegzetességeit leíró leglényegesebb webes útvonalak, s előrejelző modellek készíthetők a webes forgalom viselkedésére, az idősoros tranzakciókra vonatkozóan. A webbányászat számos területet ölel fel [Princz, 2001], ezeket a következő ábrán láthatjuk:



19. ábra. A webbányászat taxonómiája

## 6. Weboldalak optimalizálása

A szabadszavas keresők tulajdonságainak ismerete nemcsak a lekérdezések minél hatékonyabb megfogalmazásánál, s ezáltal a találati eredmények leszűkítésénél segít, de hasznos a webre szánt dokumentumok elkészítésénél is. Ha tudjuk, hogy a keresők mely tulajdonságok alapján szerepeltetik az eredménylistán előkelőbb helyen az egyes oldalakat, akkor a webre szánt oldalaink optimalizálásával jobb eredményt érhetünk el a rangsorban. Felmérések igazolják [iProspect,2004, Greenspan,2002; Nielsen,2001], hogy a felhasználók többsége az egyes eredménylistáknak csak az elejét nézi át, ezért az üzleti vállalkozásoknak létérdeke, hogy a cég honlapja minél előrébb szerepeljen a cég profiljával kapcsolatos lekérdezések eredménylistáin. Kimutatható, hogy a jelentős szabadszavas keresők eredménylistájának első 20 helyén lenni rendkívül megnöveli a web site forgalmát, s ez jelentős hatással van az üzletmenetre is.

Webhelyek, weboldalak optimalizálása egy olyan eljárás, amelynek célja az oldal számára a legjobb helyezést biztosítani a szabadszavas keresők eredménylistáin a weboldal tartalmát leíró legrelevánsabb keresési szavak, kifejezések esetén. Ezek azok a szavak, amelyeket a felhasználók a legnagyobb valószínűséggel használnak az adott oldal tartalmával, kínálatával kapcsolatos kereséseknél.

Mivel van rá kereslet, ezért honlapok optimalizálásával számos cég foglalkozik, de a weben keresve olyan oldalak is találhatók, amelyek szabadon közölnek ötleteket az oldalak optimalizálásával kapcsolatban (pl.searchengines.com, searchenginewatch.com).

A folyamat első lépése a webhely, a weboldal számára legalkalmasabb keresési szavak, kifejezések meghatározása, majd ezen szavak elhelyezése a dokumentum legfontosabb részeiben.

Az oldalak közötti mozgásoknál lényeges összhangban lenni a kulcsszavak jelentőségével. Ajánlatos a webhelyet több oldalra bontani, és a legfontosabb oldalra a webhely minden oldaláról hivatkozni, valamint a webhely oldalaira hivatkozni a főoldalról és a webhely tartalomjegyzékéből.

Célszerű a webhelyet a legfőbb tematikus keresőknél kezdve előterjeszteni minél több kereső szolgáltatónál, majd a webhely tartalmával rokon oldalakon is hivatkozásért folyamodni.

A szabadszavas keresők pontos rangsorolási algoritmusai szigorúan üzleti titok, de azért megfogalmazhatók általános szabályok a weboldalak készítésénél:

- Ahhoz, hogy egy webhely igazán népszerű és sikeres legyen, legfontosabb a tartalom minősége. Fontos a közzétett információ időről-időre megvalósuló frissítése, az adatok aktualizálása.
- A vonzó külső növelheti a látogatottságot, a webhelyen eltöltött időt, s ez néhány keresőnél javítja a webhely helyezését. Azonban nem árt mérlegelni: A sávszélesség elegendő-e az alkalmazni kívánt animáció, Flash intro gyors letöltéséhez, vagy az oldal minden böngészőben megfelelően látható-e?
- Egy webhely elkészítése előtt érdemes a szabadszavas keresők működését tanulmányozni, hogy mely tényezők befolyásolják a legnépszerűbb szolgáltatók eredmény-rangsorolását?

## **6.1. A rangsort meghatározó tényezők**

### **6.1.1. Kulcsszavak**

Egy dokumentum az általa tartalmazott szavak, kifejezések alapján kereshető vissza a szabadszavas keresők indexéből, ezért lényeges, hogy milyen szavakat használunk egy-egy honlap kialakításakor.

- Fontos, hogy minél széleskörűbben írjuk le a képviselni kívánt tartalmat (pl. szinonimák használata, a hasonló témakörök megemlítése, stb.)
- Ne használjunk általános kulcsszavakat (pl. Internet), a specifikus szavak használatával könnyebben megtalálják a specifikus tartalmat a felhasználók.
- Tegyük könnyen kereshetővé az oldalainkat az idegen szavak, kis és nagybetű, többes szám átgondolt használatával! Figyeljünk a helyesírássra! Használjuk a kulcsszavak hosszabb és rövidebb formáját is (pl. TV, televízió)!

Kulcsszavakat a dokumentumok különböző részein helyezhetünk el:

### 6.1.1.1. A weboldal szövege

Napjainkban a legnépszerűbb keresők a dokumentumok teljes szövegét indexelik, de a szövegen belüli elhelyezésnél is vannak prioritások. Fontos, hogy a nyitólap minél több kulcsszót tartalmazzon. A kulcsszavakat lehetőleg emeljük ki, azaz helyezzük a mondatok, illetve a szöveg elejére! Alkalmazzuk a kulcsszavakat minél közelebb egymáshoz, hiszen számos kereső (pl. a Google) a rangsorolásnál ezt figyelembe veszi!

A kulcsszavak gyakorisága meghatározza egy oldal relevanciáját. Ajánlott, hogy 100 szavanként 3-7 alkalommal szerepeljen egy kulcsszó.

Az oldalak legalább 100 szó hosszúak legyenek, mert némely kereső a hosszabb oldalakat a rangsorolásnál előrébb veszi (a hosszabb oldal várhatóan nemcsak kulcsszavak felsorolásából áll).

Soha ne használjunk nagyon parányi vagy láthatatlan betűket, mert a keresők spamnek veszik, és kizárják az ilyen oldalakat az adatbázisukból!

### 6.1.1.2. A dokumentum neve

Érdemes figyelmet fordítani a webre szánt dokumentumunk nevére, hiszen tartalma a szabadszavas keresők és a felhasználók számára egyaránt olvasható.

A <TITLE> elemmel közbezárt részt a legtöbb szabadszavas kereső indexeli, és az eredménylista elkészítésénél előrébb rangsorolja azokat a dokumentumokat, ahol a keresési kifejezés a dokumentum nevében előfordul.

A dokumentum nevének fel kell kelteni az emberek érdeklődését, ezért célszerű a semmitmondó *Honlap* név helyett például *Cipők, csizmák, szandálok, táskák, övek és egyéb bőrárak forgalmazója: Kényelem Nagykereskedelmi Kft* nevet használni. Az eredménylistán számos kereső az azonos fontosságú dokumentumokat név szerint abc sorrendben jeleníti meg, érdemes tehát erre is figyelni.

A dokumentum nevében ne ismételgessük kettőnél többször a kulcsszavakat, mert néhány kereső spamnek veheti. Szintén bünteti néhány rangsoroló, ha a név csupa nagybetűvel van megadva.

### 6.1.1.3. Meta jelölő elemek

A dokumentumok tartalmának megadására használatos HTML meta jelölő elemek segítségével különböző információt adhatunk át az indexelő résznek. A legtöbb keresőgép különböző mértékben használja fel ezen elemeket [Shapiro, Lehoczky]. Noha vannak olyan szabadszavas keresők, amelyek figyelmen kívül hagyják ezen elemeket, de jelenlétüket egyetlen kereső sem bünteti, így használatuk javasolt. A két leggyakrabban használt meta elem a *description* és a *keywords*. Ezen elemeket a böngésző programok nem jelenítik meg, s megadásuk mindig a dokumentumok fejlécében történik

#### 6.1.1.3.1. A dokumentum leírása

A *description* elem segítségével a dokumentum tartalma írható le. Számos keresőgép az eredménylistában – ha a dokumentum tartalmaz ilyen meta elemet – ezt az összegzést adja vissza, így mi magunk adhatjuk meg dokumentumunk lényegi leírását, s ezzel kelthetjük fel a kereső felhasználó figyelmét.

Példa a *description* elem használatára:

```
<html>
<head>
<meta name="description" content="A weboldal tartalmának
összegzése">
</head>
</html>
```

Jó, ha a *description* elemben többször szerepel a dokumentum tartalmát leíró kulcsszó. A jobb rangsorolás érdekében a kulcsszavakat célszerű az összegzés elején, lehetőleg egymáshoz közel elhelyezni.

A *description* használatát leginkább az AltaVista, AllTheWeb és a Teoma támogatja. A szabadszavas keresők különböznek abban, hogy milyen hosszúságú meta elemeket preferálnak. A 150 karaktertől rövidebb szöveget mindegyik elfogadja a *description* megadásakor.

Kerüljük el, hogy egy kulcsszót 3-7 alkalommal többször használjunk az összegzés megadásakor, mert egynémely szabadszavas kereső spamnek veheti.

#### 6.1.1.3.2. Meta kulcsszavak

A *keywords* elem alkalmazásával a dokumentum tartalmára vonatkozó kulcsszavak adhatók meg az alábbi módon:

```
<html>
<head>
<meta name="keywords" content="kulcsszavak felsorolása">
</head>
</html>
```

A kulcsszavak megadásánál érdemes minél több, a dokumentumra illő kulcsszót feltüntetni, gondolva a szinonimákra, egyes-többes számra, általánosításokra, stb. A kulcsszavakat ajánlatos fontossági sorrendben megadni, legfeljebb három alkalommal megismételni, de akkor is külön sorban, vagy nem közvetlenül egymás után. A leíró elem hosszára 1024 karakter javasolt.

A *keywords* elem különböző visszaélésekre ad lehetőséget, így indexelését számos szolgáltató nem támogatja (pl. nem támogatja: Google, AllTheWeb, AltaVista, támogatja: HotBot, MSN, Inktomi, Teoma).

#### 6.1.1.3.3. Az ALT elem

A weboldalak kialakításakor ne feledjük el, hogy a szabadszavas keresők csak a szöveg alapján tudnak visszakeresni. Ezért a képeket - de egyéb multimédia fájlokat is - el kell látni szöveges leírásokkal, beszélő megnevezésekkel ahhoz, hogy a szabadszavas keresők által kereshetővé tegyük őket. Ha a szöveg egy kép részeként tűnik fel (pl. logo, jelmondat, stb.), az olvashatatlan a szabadszavas kereső számára.

A képek tartalmának leírására az ALT jelölő elem szolgál. Az ALT elemben megadott kulcsszavak növelik a kulcsszó gyakoriságát is, a szöveges böngészők pedig ezen elem tartalmát jelenítik meg. Az ALT elemet az AltaVista, Google, Teoma használja, az AllTheWeb és az Inktomi pedig nem.

#### 6.1.1.4. Kulcsszó az URL-ben

Napjainkban egyre több szabadszavas kereső (pl. Google, Inktomi) teszi lehetővé a dokumentum web címében, azaz az URL-ben való keresést, és rangsorolásnál is előnyben részesítik az ilyen oldalakat.

A kisebb keresők, valamint a tematikus keresők a rangsorolásnál az alfabetikus sorrendet követik, így a kulcsszónak az URL-ben való szerepeltetésénél érdemes erre odafigyelni. Ha lehetőség van rá, válasszunk olyan

URL nevet, amely speciális karakterrel vagy számmal kezdődik (pl. [www.7csoda.hu](http://www.7csoda.hu)).

Néhány szabadszavas kereső nem indexeli az olyan oldalakat, amelyeknek a címében szerepel a “?” karakter (script által generált oldalak). A legjelentősebb szabadszavas keresők azonban a számokat és pl. a “@” karaktert tartalmazó URL címeket indexelik.

### **6.1.2. Téma és kulcsszavak összhangja**

A legfontosabb kulcsszavak megválasztásánál gondoljunk a weboldalunk témájára! Egy weboldal témája az oldal által tartalmazott témakörök összessége. Egy dokumentum témáját a dokumentum nevében, a dokumentum meta elemeiben, szövegében és hivatkozásaiban előforduló kulcsszavak határozzák meg. Számos szabadszavas kereső kifejlesztett már olyan mesterséges technológiát, amellyel képes egy weboldal témáját kivonatolni.

Jelenleg a Google, az AltaVista és a Lycos végrehajtja és felhasználja a téma kivonatolását, az Inktomit alkalmazó AOL, MSN és HotBot pedig tervezi a bevezetését.

### **6.1.3. Hivatkozási népszerűség**

Számos szabadszavas kereső használja fel a hivatkozási népszerűség rangsorolási algoritmusában. A Google-nál ez a legfontosabb tényező, de felhasználja a HotBot, AltaVista, MSN, Inktomi is.

A hivatkozási népszerűséget lényegesen nehezebb manipulálni, mint a hagyományos, vektortér modellen alapuló módszereket, de nem lehetetlen. Az egyik ilyen módszer a hivatkozási népszerűség növelésére a linkfarmok létrehozása. Az ilyen címeken nagyszámú olyan oldal van, amelynek egyetlen feladata, hogy visszahivatkozzon a rá hivatkozó és nagyobb hivatkozási népszerűséget akaró oldalra. A Google a PageRank érték 0-ra csökkentésével bünteti azokat a linkfarmokat, amikről tudomást szerez, és bünteti az ilyen weboldalakra hivatkozó oldalakat is. Egy másik lehetőség, hogy a nagyobb hivatkozási népszerűséget akaró webmester belép egy csere programba, ahol időről időre közlik a programban részt vevők URL címét, amelyre minden résztvevőnek az oldaláról hivatkoznia kell.

A hivatkozási népszerűség növelésénél fontosak a minőségi oldalról történő hivatkozások, amelyek bővelkednek a hivatkozott oldal tartalmát leíró kulcsszavakban. Miután a link farmok tartalma nem nevezhető minőséginek, és a csereprogramok résztvevői is gyakran különböző területeket képviselnek, így egyik lehetőség sem biztosítja a minőségi hivatkozást, ráadásul számos szabadszavas kereső helyteleníti és bünteti az ilyen próbálkozást.

Egy elfogadott módszer a hivatkozási népszerűség növelésére a webhelyre mutató hivatkozások elhelyezése bárki által írható oldalakon, például vendégkönyvekben, blogokban vagy wikipedia oldalakon. Egy-egy ilyen link csak elenyésző mértékben növeli meg a céloldal PageRank-jét, de nagy mennyiségben alkalmazva már jelentős növekedést lehet elérni.

Követhető út még közvetlenül felvenni a kapcsolatot a vetélytársakra hivatkozó oldalak webmestereivel, ajánlani a saját oldalunkra való hivatkozást, s kérni a saját legfontosabb kulcsszavaink szerepeltetését az alábbi módon:

```
<a href="http://www.név.kiterjesztés">kulcsszavak felsorolása</a>
```

A hivatkozási népszerűség a jó helyezéssel együtt a weboldal forgalmát is megnöveli, ezért kialakítása és fenntartása fontos minden cég számára.

Ha arra vagyunk kíváncsiak, hogy hány hivatkozás van webhelyünkre a Google, Yahoo, MSN keresőktől, akkor azt a <http://www.linkpopularity.com/> címen lekérdezhetjük.

#### **6.1.4. Webhelyek népszerűsége**

A webhelyek népszerűségének mérésérét elsőként a DirectHits kereső alkalmazta, de azóta alkalmazza az MSN, HotBot és Lycos is. A DirectHit egy webhelyt azon elv alapján rangsorol, hogy egy egyszerű keresés eredménylistájából hányan választják az adott webhelyet (click popularity), illetve mennyi időt töltenek el az oldalain. A rangsorolás e módszerének kialakításával az volt a cél, hogy megakadályozzák a webmestereknek a rangsor manipulálására irányuló törekvéseit, és a rangsor kialakítását a felhasználókra bízzák.

Ez a módszer sem tökéletes azonban: Írtak már olyan programot, amely tetszőleges kulcsszóra keres, s ha az eredménylistában megtalál egy adott webhelyet, ráklikkel, és bizonyos ideig a nyitólapon marad, majd kezdi előlről a folyamatot.

## 6.2. Webhelyek előterjesztése

A webhelyet, optimalizálását követően, forgalmának növelése végett ajánlatos előterjeszteni a keresők felé, és nem elég arra várni, hogy majd indexeli valamely szabadszavas kereső. Az előterjesztés célja a keresők figyelmét egy adott webhelyre irányítani, hogy indexeljék azt, vagy jelenítsék meg a könyvtárukban. Ehhez a keresők nyitólapján át elérhető lapon egy kérést kell elküldeni a szolgáltatóhoz, amely tartalmazza a webhely nevét, URL címét, az előterjesztő e-mail címét, stb. Bár vannak olyan szoftver programok, webes alkalmazások, amelyek által keresők százai felé lehet automatikusan előterjeszteni, de a jelentősebb szabadszavas keresők, illetve tematikus keresők esetén nem érdemes az automatikus előterjesztésre hagyatkozni. Egyrészt vannak olyan szabadszavas keresők, amelyek nem fogadják el a robot általi előterjesztést, másrészt ha tematikus kereső felé történik az előterjesztés, meg kell adni azt is, hogy a kereső mely fogalomköréhez tartozik a webhely. A tematikus keresők felé könnyebb az előterjesztés, mint a szabadszavas keresők felé.

Vannak olyan kereső szolgáltatók, amelyek bizonyos összegek befizetése után jobb rangsorolást biztosítanak a webhelynek, és a rangsor a befizetett összeg nagyságától is függ. Egy másik megoldás az ilyen keresőknél, hogy egy társaság csak akkor fizet, ha a szolgáltató eredménylistájáról klikkek egy látogató a társaság honlapjára. ("Pay-Per-Click" keresők).

A webhelyek előterjesztéséről a [searchenginewatch.com](http://searchenginewatch.com), [searchengines.com](http://searchengines.com) oldalakon olvashatunk bővebben.

## 6.3. A rangsor befolyásolása

A webhelyek optimalizálásakor a rangsort emelő tényezők mellett elengedhetetlenül fontos tudni azt is, hogy mit nem szeretnek a szabadszavas keresők. Valamennyi szabadszavas kereső küzd azon taktikák ellen, amelyek jobb rangsorolást adnak egy kevésbé releváns webhelynek. Ezen taktikák összességét spamnek nevezzük.

Spam-nek minősülő technikák:

- A kulcsszavak túlzásba vitt ismétlése, ami a felhasználó számára láthatatlan, ha a háttérszín és a betűszín azonos, vagy ha a betűméret elég kicsinek van megválasztva, ugyanakkor a kereső számára a kulcsszavak láthatók.
- A TITLE elem többszörözése, amelyek közül csak az elsőt jelenítik meg a böngészők, de a robot valamennyit indexeli.
- A tartalomhoz nem releváns kulcsszó szerepeltetése a dokumentum megnevezésében, illetve a meta elemekben.
- A tartalom duplikálása ugyanazon, vagy közel ugyanazon oldalak más hoston való elhelyezése által.
- A nyitólapra mutató „mesterséges linkek” elhelyezése.
- A dokumentum frissítését kérő `http-equiv="refresh"` meta elem jelenléte a fejlécben, amely a felhasználó böngészőjét az adott oldal néhány másodperces időközökben történő újbóli letöltésére utasítja.
- Túl sok oldal előterjesztése egy rövid idő alatt, vagy tematikus keresők esetén a nem megfelelő fogalomkörbe történő előterjesztés.

Az említett próbálkozások eredménye számos keresőnél az, hogy az érintett webhelyet alacsonyabb rangsorolással büntetik, vagy automatikusan kizárják az adatbázisukból.

## 7. A hallgatók információ-visszakeresésének felmérése

A felhasználók keresési szokásait, ismereteit számos tanulmány elemzi. Néhány jellemző adat:

- Az Internetet használók 56,3%-a naponta használ szabadszavas keresőt, 33% heti rendszerességgel, s ettől ritkábban csak 9,4%. A felhasználók csupán 1,2%-a válaszolta, hogy soha sem használta még ezen szoftverek valamelyikét. [iProspect,2004].
- A felhasználók 52 %-a ugyanazt a szabadszavas keresőt használja az információk keresése esetén, és csak 35% használ alternatív kereső eszközt. 13% esetében igaz, hogy különböző típusú kereséseknél különböző keresőszoftvert használ [Greenspan,2002]. Hasonló arányokat mutat egy két évvel későbbi felmérés is: 56,7%; 30,5%; 12,8% [iProspect,2004].
- A keresőszoftvert használók 16%-a csak az első egy-két eredményt nézi meg, 32% végignézi az első eredményoldalt, 23 % eljut a második eredményoldalig, 10,3 % a harmadikig, 8,7% még a harmadik után néz, és 10 % végignézi az összes eredményt, ha az nem oldalak tucatja [Greenspan,2002]. Hasonló arányok jöttek ki egy két évvel későbbi felmérés eredményeként: A felhasználók 81,75%-a csak az első 3 eredményoldalt nézi át. Az idősebbekre, háztartásbeliekre, munkanélküliekre jellemző, hogy csak az első eredményoldalt nézik végig. A nők kevésbé ásnak mélyre az eredménylistán, mint a férfiak [iProspect,2004].
- Általában a felhasználók 45,9%-a érzi sikeresnek a kereséseit. Sikertelen keresés után csupán 7,5%-a finomítja tovább a kereső kérdést, 27,2% pedig más keresőszoftverrel próbálkozik [Greenspan,2002].
- A felhasználók átlag két szóból álló kérdéseket tesznek fel. A legtöbb felhasználó nem használja az összetett keresést, illetve a részletes keresést. Szintén nagyon kevesen alkalmazzák az eredményoldalak korlátozására,

szűrésére biztosított lehetőségeket [Griffiths&Brophy,2002, Nielsen,2001; Gunn&Hepburn,2003].

Különösen érdekes volt számunkra egy kanadai felmérés [Gunn&Hepburn,2003], amely Új-Skócia 12. évfolyamos középiskolás diákjainak Internetes keresési ismereteit vizsgálja. E tanulmány 198 középiskolás (átlag életkoruk 17,5 év) információ-visszakeresési stratégiáját és technikáját méri, valamint választ keres oly kérdésekre, hogy a diákok mennyire ismerik a különböző Internetes kereső szolgáltatókat, hogyan értékelik keresési képességüket, kitől tanulták az információ-visszakeresést?

A különböző tanulmányokat olvasva kíváncsiak lettünk arra, mi a helyzet a mi hallgatóinknál, hogyan, milyen hatékonysággal keresnek? A Debreceni Egyetem Műszaki Főiskolai Karán az első éves hallgatók körében vizsgáltuk, mennyire képesek tájékozódni a weben. A gyakorlati órákon feladott néhány keresési feladat során tapasztaltuk, hogy

- hallgatóink szinte kizárólagosan néhány keresési kulcsszó megadásával keresnek;
- a többség nem ismeri a szabadszavas keresők által biztosított lehetőségeket a szükséges információ pontosítására, valamint nem tudják az eredményhalmazt szűkíteni, szűrni;
- nem tudatosult bennük, hogy bizonyos információs igények megválaszolása más-más típusú keresőeszköz használatát igényli.

A hiányosságokat látva kíváncsiak lettünk arra is, hogy ez a megfigyelés mennyire általánosítható, azaz vajon hogyan keresnek az egyetem többi karán tanuló hallgatók? Ennek felmérése végett kérdőívet készítettünk, amelynek kérdései részben megegyeznek Gunn és Hepburn, illetve Greenspan tanulmányában vizsgáltakkal. A kérdőíveket előzetesen a Debreceni Egyetem két karán, a Műszaki Főiskolai Karon, illetve az Általános Orvostudományi Karon teszteltük. A módosított, bővített kérdőívek kitöltésére a Debreceni Egyetem különböző karain tanuló nappali tagozatos hallgatók közül, a karokon tanuló hallgatók arányának megfelelően 200 főt kértünk fel. A kérdőíveket különböző, de általában alsóbb évfolyamon tanuló hallgatók töltötték ki. A kérdőívet kitöltő hallgatók karonkénti megoszlását az alábbi táblázat tartalmazza:

| Kar  | Megnevezés                                | Fő     | Százalék |
|------|-------------------------------------------|--------|----------|
| ÁJK  | Állam- és Jogtudományi Kar                | 9 fő   | 4,5%     |
| ÁOK  | Általános Orvostudományi Kar              | 19 fő  | 9,5%     |
| AVK  | Agrárgazdasági és Vidékfejlesztési Kar    | 15 fő  | 7,5%     |
| BTK  | Bölcsészettudományi Kar                   | 36 fő  | 18,0%    |
| DK   | Debreceni Konzervatórium                  | 2 fő   | 1,0%     |
| EFK  | Egészségügyi Főiskolai Kar                | 16 fő  | 8,0%     |
| FOK  | Fogorvostudományi Kar                     | 3 fő   | 1,5%     |
| GYTK | Gyógyszerésztudományi Kar                 | 3 fő   | 1,5%     |
| HPFK | Hajdúböszörményi Pedagógiai Főiskolai Kar | 8 fő   | 4,0%     |
| IK   | Informatikai Kar                          | 11 fő  | 5,5%     |
| KTk  | Közgazdaságtudományi Kar                  | 10 fő  | 5,0%     |
| MFK  | Műszaki Főiskolai Kar                     | 19 fő  | 9,5%     |
| MTK  | Mezőgazdaságtudományi Kar                 | 15 fő  | 7,5%     |
| TTK  | Természettudományi Kar                    | 34 fő  | 17,0%    |
|      | Összesen                                  | 200 fő | 100%     |

**9. táblázat. A kérdőívet kitöltő hallgatók karonkénti megoszlása**

A kérdések kiértékelésekor Excel táblázatkezelőt, valamint SPSS statisztikai programcsomagot használtunk.

## 7.1. Eredmények

A kutatás kérdéseit e dolgozat elején, a kutatás célkitűzései között már megfogalmaztuk. Az eredmények ismertetésénél a vizsgált kérdéseket a fejezetek címeiként szerepeltetjük. A kérdések kiértékelésekor a megegyező kérdéseknél összevetjük a kapott eredményeket Gunn és Hepburn, illetve Greenspan tanulmányának eredményeivel.

Az alábbi táblázat a kérdőívben megfogalmazott 21 kérdés és a kutatás 10 kérdése közötti kapcsolatot tartalmazza. A teljes kérdőív a függelék 13.1 fejezetében olvasható.

| Kérdőív                | A kutatás kérdése | Témája                                                      |
|------------------------|-------------------|-------------------------------------------------------------|
| I.-III. kérdés         |                   | További információk az eredmények megoszlásához             |
| IV.-VI. kérdés         | 1                 | A hallgatók Internet használati szokásai                    |
| VII. kérdés            | 2                 | Mely információs forrásokat részesítenek előnyben?          |
| VIII. kérdés           | 3                 | A hallgatók vélekedése keresési képességükről               |
| IX.-X. kérdés          | 4                 | A hallgatók hozzáférése az Internethez                      |
| XI. kérdés             | 5                 | Kitől tanulnak keresni a hallgatók?                         |
| XII.-XIII. kérdés      | 6                 | Mely Internetes keresőket ismernek?                         |
| XIV.-XVIII. kérdés     | 7                 | Milyen stratégiákat, technikákat használnak a kereséseknél? |
| I.-III., IX.-X. kérdés | 8                 | Mi befolyásolja a hallgatók keresési készségét?             |
| XIX.-XX. kérdés        | 9                 | Milyen felhasználói szokásaik vannak?                       |
| XXI. kérdés            | 10                | Hogyan vélekednek a keresők hatékonyságáról?                |

**10. táblázat. A kérdőív illesztése a kutatás kérdéseire**

A kérdőív kiértékelését a kutatás kérdéseinek sorrendjében végezzük.

### **7.1.1. A hallgatók mely Internet szolgáltatásokat és milyen sűrűn vesznek igénybe?**

A felmérés kezdetén rákérdeztünk, hogy a hallgatók mely Internet szolgáltatásokat és milyen arányban vesznek igénybe. A 11. táblázat a hallgatók által használt szolgáltatások arányát mutatja:

| Karkód | Kar      | WWW  | E-mail | Ftp | Telnet | Chat |
|--------|----------|------|--------|-----|--------|------|
| 1      | ÁJK      | 100% | 100%   | 0%  | 0%     | 11%  |
| 2      | ÁOK      | 95%  | 74%    | 16% | 0%     | 32%  |
| 3      | AVK      | 100% | 73%    | 0%  | 0%     | 40%  |
| 4      | BTK      | 97%  | 94%    | 8%  | 0%     | 36%  |
| 5      | DK       | 100% | 50%    | 0%  | 0%     | 50%  |
| 6      | EFK      | 100% | 88%    | 13% | 0%     | 19%  |
| 7      | FOK      | 100% | 67%    | 0%  | 0%     | 33%  |
| 8      | GYTK     | 100% | 100%   | 0%  | 0%     | 33%  |
| 9      | HPFK     | 100% | 75%    | 0%  | 0%     | 13%  |
| 10     | IK       | 100% | 91%    | 82% | 36%    | 9%   |
| 11     | KTk      | 100% | 100%   | 10% | 0%     | 30%  |
| 12     | MFK      | 100% | 89%    | 32% | 0%     | 26%  |
| 13     | MTK      | 100% | 67%    | 7%  | 0%     | 33%  |
| 14     | TTK      | 100% | 91%    | 9%  | 3%     | 32%  |
|        | Összesen | 99%  | 86%    | 14% | 3%     | 29%  |

**11. táblázat. Internet szolgáltatások használatának aránya kari bontásban**

A hallgatók döntő többsége kisebb nagyobb rendszerességgel információ-visszakeresésre is használja a webet. A hallgatók 32%-a rendszeresen, 65%-a alkalmanként szokott tanulmányaihoz az Interneten információt, tananyagot gyűjteni, míg személyes célból rendszeresen 44%, alkalmanként 53% keres a weben.

### **7.1.2. Tanulmányaikhoz tankönyveiken kívül mely információs forrásokat részesítik előnyben?**

A kérdésre adott válaszokat a következő táblázat mutatja:

|                      |       |
|----------------------|-------|
| web:                 | 11,6% |
| könyvek:             | 34,2% |
| egyformán mindkettő: | 52,8% |
| egyik sem:           | 1,5%  |

**12. táblázat. Információs források használatának aránya**

Különböző felmérések támasztják alá, hogy az Internetnek egyre inkább meghatározó szerepe van az iskolai tananyaghoz kapcsolódó kereséseknél, s számos diák az Internetet részesíti előnyben a könyvekkel szemben [Bilal,2000; Environics Research Group,2001]. A kanadai iskolák az Internetre fordított összegek növelése mellett drámai módon csökkentették a nyomtatott források beszerzésére szánt összegeket [Gunn&Hepburn,2003]. Ezen adatok ismerete mellett tettük fel a kérdést, vajon a DE hallgatói mely információs forrásokat részesítenek előnyben? A válaszok szerint a többség egyformán preferálja a webet és a könyveket, de összességében mégis többen választják a tankönyveken kívül a nyomtatott forrásokat. Ez a tény a felsőoktatásban kissé meglepő, hiszen a tankönyveknek tartalmazniuk kell a törzsanyagot, míg a tananyaghoz kapcsolódó kutatások legfrissebb eredményei leginkább a webről érhetők el. A nyomtatott források létrejöttéhez sokkal hosszabb idő szükséges, mint a weben publikálni egy anyagot, s a beszerzési források szűkössége miatt számos nyomtatott forrás nem is jut el az intézmények könyvtáráig.

### 7.1.3. Hogyan vélekednek az Interneten található információk közötti keresési képességükről?

A keresési képesség önértékelését szándékosan a kérdőív elejére, a keresési stratégiákra és technikákra kérdező kérdések elé tettük.

|          |       |
|----------|-------|
| Gyenge:  | 10,6% |
| Közepes: | 56,8% |
| Jó:      | 32,7% |

13. táblázat. A hallgatók vélekedése keresési képességükről

Felmérésünk eredményei szerint kevés diák használja a szabadszavas keresők nyújtotta lehetőségeket a keresés eredményének szűkítésére és finomítására, ennek ellenére keresési képességét a legtöbbjük átlagosnak vagy jónak tartja. A pozitív önértékelés még inkább megfigyelhető a kanadai diákoknál, ahol 5,1% gyengének, 13,6% közepesnek és 81,3%-uk jónak vagy kifejezetten jónak tartotta magát, miközben eredményeik ezt nem igazolták. A felhasználóknak a keresési képességükről kialakított képét valószínűleg meghatározza, hogy a kereséseknél milyen gyakorisággal találják meg a keresett

információt, s mekkora erőfeszítésre van ehhez szükség. Ha feltételezzük, hogy a felhasználók többsége anyanyelvén keres, akkor igaz, hogy a magyar nyelvűtől lényegesen nagyobb számban található angol nyelvű információ a weben, a jóval többről pedig egyszerű egy-két szavas keresésekkel is nagyobb esély van a remélt információ megtalálásának – az eltérő elégedettségnek talán ez lehet a magyarázata.

#### **7.1.4. Milyen Internet hozzáférési lehetőségek állnak a hallgatók rendelkezésére?**

A Internet elérhetősége előfeltétele annak, hogy a weben fellelhető információk a tanulásban értékes forrásanyagként szolgáljanak.

Az Internet munkahelyi és otthoni elérhetőségére külön kérdeztünk rá. Az *Oktatási intézményében rendelkezésre állnak olyan számítógépek, amelyeken át az Internet elérhető?* kérdésre a hallgatók 79 %-a az *igen*, 19 %-a a *korlátozottan*, 2%-a a *nem* választ adta. Karonkénti bontásban a válaszok között lényeges eltérés nem volt. Ezen válaszok alapján elmondható, hogy a hallgatók többségének lehetősége nyílik az Interneten át információkat beszerezni.

Az *Otthonában rendelkezik Internet eléréssel?* kérdésre a hallgatók 49,5%-a válaszolt igennel. A kanadai felmérésben a diákok 72,7% válaszolta, hogy otthonában is hozzáfér az Internethez.

#### **7.1.5. Honnan szerzik ismereteiket az Interneten található, a tanulmányaikhoz felhasználható információk kereséséhez?**

A *Kitől tanult meg az Interneten keresni?* kérdésre több válasz megadása is lehetséges volt.

|                     | Százalék |
|---------------------|----------|
| Tanár               | 21%      |
| Szülő / családtag   | 12%      |
| Barát / osztálytárs | 44%      |
| Magától tanulta meg | 60%      |
| Összesen:           | 137%     |

**14. táblázat. Kitől tanulnak keresni a hallgatók?\***

A hallgatók közül a többség - a felmérésben résztvevőknek 60%-a - arról számolt be, hogy önmagától tanult meg keresni, hasonlóan a kanadai diákokhoz, ahol ez az arány 72%. Második helyen a DE hallgató között a barát vagy osztálytárs áll, kb. kétszer annyi jelöléssel, mint a 3. helyen álló tanár. A kanadai felmérésben hasonló a rangsor, de ott a 2. és 3. hely szinte ugyanannyi jelölést kapott (39,8% illetve 39,3%). Mindkét felmérésnél a családnak volt a legkevésbé szerepe a keresések tanulásánál. A felmérések eredményei azt mutatják, hogy a diákok eléggé magukra, illetve társaikra hagyatkozva tanulják a webről az információ-visszakeresést, s a tanároknak ebben a tevékenységben kisebb szerep jut.

#### **7.1.6. A hallgatók milyen ismeretekkel rendelkeznek a különböző Internetes kereső szolgáltatókról?**

A felmérésünkben szereplő szolgáltatókat volt és jelenlegi népszerűségük, illetve indexállományuk nagysága alapján válogattuk ki. A felmérésben szerepeltettük a három magyar nyelvű szabadszavas kereső szolgáltatót is.

A Debreceni Egyetem nappali tagozatos hallgatói között a keresők ismertségi adatai a következők:

---

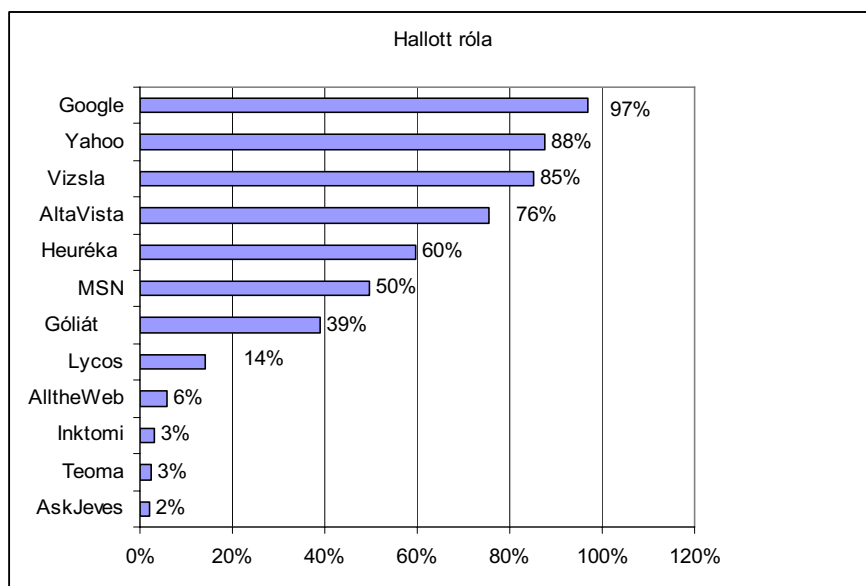
\* Ez a kérdés többszörös választást engedélyezett, ezért a százalékos arányok összege nem egyenlő 100%-kal.

| Kereső<br>szolgáltató | Nem<br>hallott róla | Alkalmanként<br>használja | Rendszeresen<br>használja |
|-----------------------|---------------------|---------------------------|---------------------------|
| AlltheWeb             | 94%                 | 3%                        | 0,50%                     |
| AltaVista             | 24,50%              | 38%                       | 7%                        |
| AskJeeves             | 98%                 | 0%                        | 0%                        |
| Google                | 3%                  | 35%                       | 62%                       |
| Lycos                 | 86%                 | 4%                        | 0%                        |
| MSN                   | 50,50%              | 18%                       | 8,50%                     |
| Teoma                 | 97,50%              | 0%                        | 0%                        |
| Yahoo                 | 12,50%              | 50,50%                    | 16%                       |
| Góliát                | 61%                 | 15,50%                    | 3%                        |
| Heuréka               | 40,50%              | 22%                       | 4%                        |
| Vizsla24              | 15%                 | 49%                       | 11,50%                    |

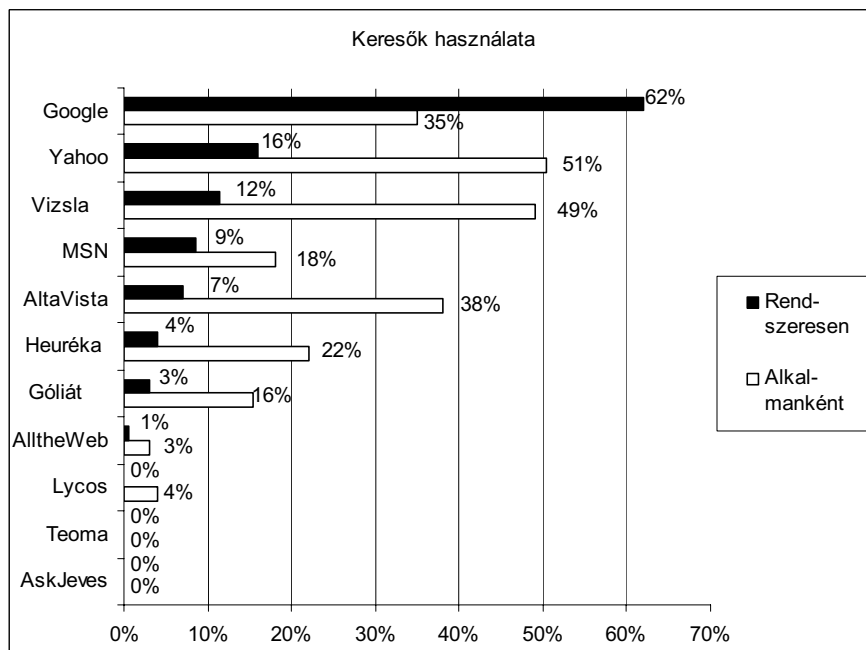
15. táblázat. A kereső szolgáltatók ismertsége és igénybe vétele a DE hallgatóinál

A hallgatók között a legismertebb keresők között szerepel az USA népszerűségi listájának jelenlegi első 3 helyezettje (Google, Yahoo, MSN), de jó az ismertsége az USA-ban egykor oly népszerű AltaVistának és Lycosnak is. Meglepően magas a hallgatók között az AltaVistát használók aránya, miközben jelenleg az amerikai felhasználók esetében a kereséseknek kevesebb, mint 1%-át hajtják végre e szabadszavas keresőt használva [Search Engine Watch]. Erre magyarázat lehet, hogy a 90-es évek második felében, amikor hazánk iskoláiban is megjelent az Internet, a tanárok egy része az AltaVistát, az akkor még világszerte a legnépszerűbbek közé számító keresőt ismerte meg, s kezdte el tanítani.

Érdekes módon az indexállományok nagysága és a népszerűség, ismertség között nincs összefüggés, hiszen a méret szerint a 2. helyezett AlltheWeb nem szerepel a legkedveltebb keresők között sem az USA-ban, sem a kanadai diákok között, és a DE diákjai között is minimális ismertséggel rendelkezik. Az amerikai, kanadai középiskolások között népszerű Ask Jeeves a debreceni egyetemisták között alig ismert. A legismertebb magyar nyelvű kereső a Vizsla24 (85%-os ismertséggel), majd a Heuréka és a Góliát a sorrend.



20. ábra. A DE hallgatóinak ismeretei az Internetes kereső szolgáltatókról



21. ábra. A legnépszerűbb kereső szolgáltatók használatának aránya a DE hallgatói között

A felmérésünk szerint a hallgatók átlagosan 2,35 keresőt használnak alkalmanként és 1,13 keresőt rendszeresen, ami 3,48 kereső használatát jelenti valamilyen rendszerességgel. Ez a szám magasabb, mint a kanadai felmérésben, ahol átlag 1-2 kereső szolgáltatást használnak a középiskolások. A hallgatók közül 61-en (30,5%) válaszolták, hogy nem használnak rendszeresen keresőt.

A DE hallgatói körében a leggyakrabban használt, azaz a legnépszerűbb keresési szolgáltatók arányát mutatja a 21. ábra. Miután egy hallgató több szolgáltatást is megjelölhetett, így a százalékos arányokat az összes *Alkalmanként*, illetve az összes *Rendszeresen* jelölés arányában adtuk meg, ezért a százalékok összege nem egyenlő 100-zal. A hallgatók között magasan a legnépszerűbb a Google: rendszeresen 62% ezen a szolgáltatón keres, s hasonló az arány a kanadai diákok között is: 66,7%. A DE hallgatói és dolgozói körében végzett egy évvel későbbi felmérés is [Bujdosó, 2006] a Google elsöprő népszerűségét támasztja alá.

A hallgatókat megkértük arra is, hogy adják meg az általuk leggyakrabban használt kiinduló keresési oldalt, amennyiben a tesztbeli felsorolásban az nem szerepel. A tesztben szereplőkön kívül további 4 kiinduló oldalt neveztek meg, összesen 11 esetben:

| Név       | Százalék |
|-----------|----------|
| Kapu      | 0,5%     |
| Keresőlap | 1,0%     |
| Kurzor    | 1,0%     |
| Startlap  | 3,0%     |

16. táblázat. A hallgatók által kiegészítésként megadott legnépszerűbb kiindulási oldalak

### 7.1.7. Milyen stratégiákat és technikákat használnak a webes információ-visszakeresésnél?

Felmérésünk részletes eredményét a 17. táblázat tartalmazza:

|                                                                                                       | Nem   | Alkal-<br>manként | Rend-<br>szeresen |
|-------------------------------------------------------------------------------------------------------|-------|-------------------|-------------------|
| Használ-e logikai operátorokat (AND, OR, NOT)?                                                        | 76,0% | 19,5%             | 4,5%              |
| Használja a + vagy a - jelet a szükséges ill. a tiltott szavak megadására?                            | 80,0% | 15,5%             | 4,5%              |
| Használ idézőjelet ( “ ”) kifejezések keresésénél?                                                    | 53,5% | 36,5%             | 10,0%             |
| Használ helyettesítő karaktereket csak részlegesen ismert szavak keresésekor?                         | 65,5% | 29,0%             | 5,5%              |
| Használja a Speciális keresés (Advanced Search) ablakot a keresési szavak kapcsolatának beállítására? | 64,0% | 28,5%             | 7,5%              |

**17. táblázat. Ajánlott Internetes keresési stratégiák használata a DE hallgatóinál**

A Boole-operátorok használata hatékony stratégiát eredményez a szabadszavas keresők találati eredményének javítására [Sullivan,2001; Barlow,2001]. A felmérésünkben szereplő hallgatók több, mint háromnegyede nem használja a Boole-operátorokat, míg a kanadaiak 48 %-a nem, 36 %-a alkalmanként, 16 %-a pedig rendszeresen használja ezeket az operátorokat. Hasonló arányok jellemzik a + (kötelező szerepeltetés) és – (kizárás) operátorok használatát is, s ennél a pontnál is magasabbak a kanadaiak között a használat arányai.

Összetartozó szavakat, szóhasználatokat kifejezésként keresni szintén jelentősen szűkíti az eredménylistát. A DE hallgatói ezt a lehetőséget ismerték és használták a leginkább, de az alkalmanként és rendszeresen használók száma sem érte el az 50%-ot, s evvel az eredménnyel alulmaradtak a kanadai diákokkal szemben.

Helyettesítő karaktereket a szavak végén a többes szám vagy a toldalékok helyettesítésére lehet használni, de néhány kereső a szavak közepén is felhasználja a szó különböző helyesírási lehetőségeinek keresésénél. A DE hallgatói e kérdésnél mutattak jobb eredményt a kanadai diákokhoz képest: 37%

nyilatkozott arról, hogy alkalmanként vagy rendszeresen használ joker karaktereket, míg ez az arány a kanadai diákoknál csupán 9%.

A lekérdezések megfogalmazásának könnyítésére ma már szinte minden kereső szolgáltató biztosítja a részletes (vagy speciális) keresés lehetőségét, ahol logikai operátorok használata nélkül, a kereső szavak különböző mezőkbe való beírásával előírhatjuk a köztük lévő logikai kapcsolatokat, illetve különböző szempontok szerint szűrhetjük, szűkíthetjük az eredményt. A részletes keresés ablakig azonban kevesen jutnak el, mert az oda mutató hivatkozást vagy nem veszik észre, vagy nincsenek tisztában a jelentőségével, vagy nem fordítanak időt, figyelmet annak kipróbálására. A DE hallgatóinak többsége is nemleges választ adott e lehetőség használatára. A kanadai diákok között még rosszabb ez az arány: 82,3% nem, 10,1% alkalmanként, 7,6% rendszeresen használja ezt a lehetőséget.

Az Interneten történő hatékony információkereséshez ajánlott stratégiákhoz különböző technikák kapcsolódnak, amelyekkel az eredmények szűkíthetők, finomíthatók, illetve bizonyos területekre fókuszálhatók. E technikák vizsgálatánál a feltett kérdésekre a következő válaszokat kaptuk:

|                                                                                                                                         | Igen | Nem   |
|-----------------------------------------------------------------------------------------------------------------------------------------|------|-------|
| Tudja-e, hogyan zárhatók ki a keresés eredményhalmazából a kereskedelmi webhelyek, vagy egyéb nem kívánatos tárhelyek?                  | 1,5% | 98,5% |
| Tudja-e, hogyan adható meg, hogy a keresés csak az utolsó 3 hónapban módosított vagy a webre elhelyezett dokumentumok között történjen? | 7%   | 93%   |
| Tudja-e, hogyan adható meg, hogy a keresés csak a lapok megnevezésében (ne a teljes szövegben) történjen?                               | 7,5% | 92,5% |
| Használja a keresők speciális szolgáltatását különböző fájltypusok (pl. kép, hang) keresésekor?                                         | 28%  | 72%   |

**18. táblázat. A keresés eredményének szűrésére és finomítására vonatkozó technikák ismerete a DE hallgatóinál**

A .com site-ok kizárása nagyon hasznos keresési technika lehet, ha angol nyelven szeretnénk az iskolai tananyaggal, a legfrissebb kutatási eredményekkel kapcsolatos információkhoz jutni, de más esetekben is szükség lehet kereskedelmi szerverek kizárására. Szintén hasznos lehet, ha a kereső számára előírjuk, mely site-on, domainen keressen (pl. ha valamely egyetem oldalain szeretnénk információhoz jutni). A keresési technikákkal kapcsolatban a hallgatók itt mutatták a legnagyobb tudatlanságot: A megkérdezettek 98,5%-a nem ismerte e lehetőséget, míg a kanadai diákoknál csak 77,8 %. A 20%-os eltérés oka lehet, hogy a magyar nyelvű információkeresésnél a .com kizárásával nem zárható ki a különböző magyar nyelvű kereskedelmi oldalak, ezért ezt a funkciót hallgatóink ritkán használják.

Majdnem hasonló eredmény jött ki a keresést egy adott időintervallumra (pl. az elmúlt 3 hónapra, félévre, 1 évre, stb.), illetve a keresést a dokumentum bizonyos részeire (pl. megnevezés, webcím, hivatkozás, stb.) korlátozó technikák ismerete terén is. A hallgatók 93%, ill. 92,5%-a nem ismerte ezt a lehetőséget, míg a kanadai diákoknál ez az arány 72,7% és 85,4%.

A legjobb eredményt a különböző típusú információk (pl. kép, hang, stb.) keresését támogató szolgáltatások ismeretében mutattak a hallgatók. Ennél a kérdésnél a nemleges válasz 72% volt, s ez alig tér el a kanadai diákoknál mért 70,7%-tól.

### **7.1.8. Mi befolyásolja a hallgatók keresési készségét?**

A kérdőív feldolgozása során szeretnénk volna választ kapni arra a kérdésre, mitől függ a hallgatók keresési készsége, pontosabban befolyásolja-e azt a hallgató neme, az informatika iránti érdeklődése, az Internethez való hozzáférése, illetve van-e különbség a különböző karok hallgatóinak keresési készségében?

A hallgatók keresési készségének vizsgálatakor a különböző keresési stratégiákra és technikákra adott válaszokat faktoranalízis alkalmazásával egy változóba vontuk össze, s e változónak a négy független változótól való függését vizsgáltuk. Az alkalmazott statisztikai próbákat és pontos eredményeiket a **függelék 13.2** fejezete tartalmazza.

Gunn és Hepburn szerint a keresési készséget az otthoni Internethez való hozzáférés befolyásolja, mivel – eredményeik szerint - a különböző keresési

stratégiákat és technikákat gyakrabban használják azok a tanulók, akiknél otthon is elérhető az Internet. Kimutatása szerint azok a tanulók, akik otthonról is el tudják érni az Internetet, nagyobb százalékban maguktól tanulnak meg keresni, és kisebb arányban tanártól vagy osztálytársaktól, mint azok, akik otthon nem rendelkeznek Internet eléréssel. Az otthoni Internet eléréssel bírók kisebb mértékben tartják gyengének keresési készségüket, mint azok, akiknek nincs otthon Internet. Ezek a megállapítások – ha kisebb arányban is – de felmérésünk szerint igazak a DE hallgatói között is:

| Kitől tanult meg az Interneten keresni? | Otthonában rendelkezik Internet eléréssel | Otthonában nem rendelkezik Internet eléréssel |
|-----------------------------------------|-------------------------------------------|-----------------------------------------------|
| Tanár                                   | 18%                                       | 24%                                           |
| szülő / családtag                       | 14%                                       | 9%                                            |
| barát / osztálytárs                     | 39%                                       | 49%                                           |
| magától tanulta meg                     | 65%                                       | 55%                                           |

**19. táblázat. Az otthoni Internet elérés és a keresés tanulása közötti kapcsolat a DE hallgatóinál**

|                                     | Otthonában rendelkezik Internet eléréssel | Otthonában nem rendelkezik Internet eléréssel |
|-------------------------------------|-------------------------------------------|-----------------------------------------------|
| Keresési készségét gyengének tartja | 7%                                        | 15%                                           |

**20. táblázat. Az otthoni Internet elérés és a keresési készségről való vélekedés közötti kapcsolat**

#### **7.1.8.1. Az Internethez való hozzáférés és a keresési készség kapcsolata**

A keresési készség függését az Internethez való hozzáféréstől külön vizsgáltuk az otthoni Internet elérésre, illetve az oktatási intézményben való Internet elérésre adott válaszok függvényében. Mindkét esetben a próbák azt igazolták, hogy a válaszok által meghatározott populációk eloszlása nem

különbözik egymástól. Bár kimutatható, hogy az Internethez otthon is hozzáférő hallgatókra a keresések tanulásánál nagyobb önállóság és nagyobb magabiztosság jellemző, de statisztikailag nem tudtunk szignifikáns összefüggést kimutatni a keresési készség és az Internethez való hozzáférés között (lehet az otthon vagy az oktatási intézményben). Ennek okát abban látjuk, hogy a hallgatók általánosan néhány keresési kulcsszó megadásával keresnek, amit az általános vagy középiskolai tanulmányuk ideje alatt már elsajátítanak egy alapszinten, újabb módszereket, technikákat pedig a későbbiekben sem tanulnak meg.

#### **7.1.8.2. Az informatika iránti érdeklődés és a keresési készség kapcsolata**

Az összefüggés vizsgálatához megkértük a hallgatókat, hogy informatikai érdeklődésük mértékét jelöljék egy 5 fokozatú skálán. Az így kapott populációk eloszlásának összehasonlításakor a különböző próbák szignifikáns különbséget mutattak. Az így kapott eredmények utólagos vizsgálatánál (post hoc teszt) kimutattuk, hogy a keresési készség és az internet iránti érdeklődés csak azon hallgatók esetében mutat szignifikáns összefüggést, akik a közepestől nagyobb mértékben érdeklődnek az informatika iránt.

#### **7.1.8.3. A nemek és a keresési készség kapcsolata**

A hallgatók neme és a keresési készségük között a különböző próbák korrelációt mutattak, de további vizsgálatokkal (kétszemponos varianciaanalízis) igazoltuk, hogy az informatika iránti érdeklődés erőssége különbözik nemenként oly erősen, amely a két nem keresési készségében a különbséget eredményezi. Evvel igazoltuk, hogy a felhasználó neme önmagában nem determinálja a felhasználó keresési készségét.

#### **7.1.8.4. Milyen eltérések vannak az egyes karokon tanulók keresési készségében?**

Felmérésünk során kíváncsiak voltunk arra is, hogy lényegesen különbözik-e az egyes karokon tanuló hallgatók keresési készsége? A különböző próbák eredménye szerint a keresési készség karonként nem homogén. A Kruskal-Wallis-próba rangtáblázatából leolvasható az is, hogy a keresési készségre a legmagasabb rangokat az Informatikai Kar (164), a

Közgazdaságtudományi Kar (122), a Műszaki Főiskolai Kar (120) és az Általános Orvostudományi Kar (104) esetében kaptuk. További post hoc vizsgálattal kimutattuk, hogy szignifikánsan csak az Informatikai Kar hallgatóinak eredménye különbözik az Agrárgazdasági és Vidékfejlesztési Kar (79), valamint az Állam- és Jogtudományi Kar (79) hallgatóinak eredményétől. Ezek az eredmények megerősítik az informatika iránti érdeklődés és a keresési készség kapcsolatának vizsgálatánál kapott eredményeinket.

### **7.1.9. Milyen felhasználói szokások jellemzik a hallgatókat a kereséseknél?**

A felhasználói szokások vizsgálatánál az egyik szempont, hogy a felhasználó hány kereső szolgáltatást használ? A külföldi felmérések szerint [Gunn&Hepburn,2003, Greenspan,2002, Griffiths&Brophy,2002] a felhasználók döntően 1-2 keresőszoftvert használnak. Amíg a kanadai, amerikai, angol felhasználóknál megfigyelhető a választott szabadszavas keresőhöz való lojalitás, addig a Debreceni Egyetem diákjai között ez nem áll fenn. A DE hallgatói - ahogy azt a 6-os kérdésnél már vizsgáltuk – döntően 3-4 keresőt használnak alkalmankénti vagy rendszeresen, s ez a szám magasabb az említett felmérések számainál. Csupán egy kereső szoftvert a hallgatók 22%-a használ: 5% alkalmanként, 17% rendszeresen.

*Az első keresés sikertelensége esetén mit tesz?* kérdésre a hallgatók válaszai: 85 % tovább finomítja a kereső kérdést; 11,5% másik keresővel próbálkozik.; 3,5 % abbahagyja a keresést. Habár hallgatóink átlagban több keresőt használnak, az első keresés sikertelensége esetén mégis kevesebben próbálkoznak egy másik keresővel, mint Greenspan 2002-es felmérése szerint, ahol ez az arány 27%. A 7. kérdés eredményeit ismerve valószínű, hogy a kereső kérdés további finomításán nem a különböző szűresi, szűkítési lehetőségeket értik a hallgatók, hanem újabb, esetleg több keresési kulcsszó megadását.

A hallgatóknak az eredményoldalak áttekintésével kapcsolatos válasza: 4%-a csak az első egy-két eredményt nézi meg, 30,5% végignézi az első eredményoldalt, 65,5 % pedig további eredményoldalakat is átnéz. Greenspan felmérésében ezek az arányok rendre 16%; 32% és 52%.

Az [iProspect,2004] tanulmány megállapítása, hogy a nők kevésbé alaposan nézik át az eredménylistát, mivel 15,2%-kal több nő hagyja abba az áttekintést az első eredményoldal után, mint férfi.. Felmérésünkben ez az arány 8%.

|                                         | <b>férfi</b> | <b>nő</b> |
|-----------------------------------------|--------------|-----------|
| csak az első egy-két eredményt nézi meg | 3%           | 4%        |
| végignézi az első eredményoldalt        | 26%          | 33%       |
| további eredményoldalakat is átnéz.     | 71%          | 63%       |

**21. táblázat. Az eredményoldalak áttekintésével kapcsolatos válaszok férfi/nő bontásban**

### **7.1.10. Hogyan vélekednek a keresők hatékonyságáról?**

A felmérés végén megkérdeztük a hallgatókat, hogy szerintük a legjobb keresők a weben lévő dokumentumok hány százalékában keresnek? Evvel a kérdéssel burkoltan a láthatatlan web fogalmára szerettünk volna rákérdezni, hiszen a magas értékek a szabadszavas keresők iránti bizalmat tükrözik, amely szerint azok a weben át elérhető információk többségét indexelik.

A különböző szabadszavas keresőkön feltett kérdések találati listája nagyon különböző, kevés az átfedés közöttük, a látható web teljes lefedettségét egyik kereső sem éri el. [Gulli & Signorini,2005; Bharat & Broder,1998]. Ha a weben át elérhető dokumentumok közé hozzászámítjuk a különböző online adatbázisokat, intranetes állományokat, s mindazt, ami a láthatatlan web fogalmába beletartozik, akkor a legjobb kereső által indexelt dokumentumok is csak töredékét teszik ki a weben át elérhető összesnek. A 22. táblázat a kérdésre adott válaszokat tartalmazó változó (*XXI. kereshető web*) leíró statisztikáját mutatja. Az első oszlop a karok kódját (Karkod) tartalmazza, majd az N alatti oszlop a kérdésre választ adó hallgatók számát. A harmadik oszlopban a karonkénti átlagok olvashatók

**Descriptives**

XXI. kereshető web

|       | N   | Mean     | Std. Deviation | Std. Error | 95% Confidence Interval for Mean |             | Minimum | Maximum |
|-------|-----|----------|----------------|------------|----------------------------------|-------------|---------|---------|
|       |     |          |                |            | Lower Bound                      | Upper Bound |         |         |
| 1     | 7   | 71,42857 | 19,518001      | 7,377111   | 53,37743                         | 89,47971    | 40,000  | 100,000 |
| 2     | 17  | 48,35312 | 31,232587      | 7,575015   | 32,29480                         | 64,41143    | ,001    | 90,000  |
| 3     | 14  | 59,28571 | 15,794004      | 4,221125   | 50,16653                         | 68,40490    | 35,000  | 80,000  |
| 4     | 33  | 71,72727 | 16,494834      | 2,871382   | 65,87846                         | 77,57609    | 30,000  | 100,000 |
| 5     | 2   | 66,00000 | 1,414214       | 1,000000   | 53,29380                         | 78,70620    | 65,000  | 67,000  |
| 6     | 14  | 65,21429 | 18,622478      | 4,977067   | 54,46199                         | 75,96658    | 20,000  | 90,000  |
| 7     | 2   | 85,00000 | 7,071068       | 5,000000   | 21,46898                         | 148,53102   | 80,000  | 90,000  |
| 8     | 3   | 56,66667 | 15,275252      | 8,819171   | 18,72084                         | 94,61250    | 40,000  | 70,000  |
| 9     | 7   | 69,00000 | 6,082763       | 2,299068   | 63,37438                         | 74,62562    | 60,000  | 80,000  |
| 10    | 11  | 51,36364 | 23,461768      | 7,073989   | 35,60181                         | 67,12547    | 10,000  | 80,000  |
| 11    | 9   | 60,22222 | 31,292083      | 10,430694  | 36,16900                         | 84,27545    | 7,000   | 90,000  |
| 12    | 19  | 62,89474 | 19,812720      | 4,545350   | 53,34531                         | 72,44416    | 40,000  | 100,000 |
| 13    | 14  | 72,21429 | 20,625520      | 5,512402   | 60,30546                         | 84,12311    | 30,000  | 98,000  |
| 14    | 30  | 65,43333 | 21,500094      | 3,925362   | 57,40507                         | 73,46160    | 20,000  | 100,000 |
| Total | 182 | 64,08243 | 21,724210      | 1,610306   | 60,90505                         | 67,25982    | ,001    | 100,000 |

**22. táblázat. A XXI. kereshető web változó leíró statisztikája**

| Karkod | Kar | Karkod | Kar  |
|--------|-----|--------|------|
| 1      | ÁJK | 8      | GYTK |
| 2      | ÁOK | 9      | HPFK |
| 3      | AVK | 10     | IK   |
| 4      | BTK | 11     | KTK  |
| 5      | DK  | 12     | MTK  |
| 6      | EFK | 13     | MTK  |
| 7      | FOK | 14     | TTK  |

**23. táblázat . A Karkod változó és a karok összekapcsolása**

Az egyes keresők által indexelt web méretére vonatkozó karonkénti becslések átlaga 48% és 72% között mozog, ami 64%-os egyetemi átlagot eredményez. Ezek nagyon magas számok, s mindenképpen a szabadszavas keresők teljesítményének túlértékelését tükrözik.

## 8. Egy pedagógiai kísérlet a keresési készség javítására

A webről történő információ-visszakeresés hatékonyságának növelése két jól elkülönülő módon valósítható meg. Az egyik technikai jellegű: A rendelkezésre álló eszközök tökéletesítésével vagy újak fejlesztésével minél egyszerűbbé és hatékonyabbá tenni a keresést. A másik a felhasználók keresési készségének növelésére irányul: Szükséges az információ-visszakeresést tanítani, növelni kell a felhasználók tudatosságát az információ-forrásokkal illetően.

### 8.1. Az információ-visszakeresés tanítása

Hazánkban az Internetnek az oktatásba történő bevezetése közel 10 éves múltra tekint vissza: Kezdetben csak a felsőoktatásban, majd fokozatosan a közép és általános iskolákban is elérhetővé vált a Világháló használata. Míg a kezdeti időkben a helyi hálózatok hiánya, a hálózatépítés költségei, a számítógépek száma korlátozta az Internet korlátlan elérését, addig napjainkban gyakorlatilag minden diáknak lehetősége nyílik iskolai éve alatt az Internet használatára, s nagyon sok tanulónak otthonában is biztosított a Világháló elérése. Az iskolai tantervekben megjelent az információ-visszakeresés, a tanmenetek tartalmazzák a weben történő böngészés, visszakeresés tanítását [Mozaik, Nemzeti Tankönyvkiadó]. A tanmenetek között böngészve jó néhányban szerepel az összetett és a részletes keresés tanítása is.

A Debreceni Egyetem Műszaki Főiskolai Karán a *Műszaki Informatika* tárgy keretein belül történik azon számítógépes programok oktatása, valamint ezen programokból a hallgatóknak egy olyan szintre való felzárkóztatása, amelyre a későbbi tanulmányaik folyamán szükségük lesz. A tárgyat a hallgatók többsége első éves korában tanulja.

E tárgy egyik modulja a prezentáció, valamint az információ webes megjelenítése. E modul keretében évek óta vizsgáljuk, mennyire képesek hallgatóink tájékozódni a weben. A gyakorlati órákon feladott néhány keresési

feladat során tapasztalhatjuk, hogy hallgatóink szinte kizárólagosan néhány keresési kulcsszó megadásával keresnek, és nem használják a szükséges információ pontosítását elősegítő, a kereső szoftverek biztosította lehetőségeket, valamint nem tudják az eredményhalmazt szűkíteni, szűrni. Nem ismerik a tematikus keresők előnyeit egy-egy nagyobb téma, tárgyszerűség áttekintésekor. Számos tudományos folyóiratot, cikket tartalmazó adatbázis érhető el a weben keresztül, amelyek létezéséről általában nem tudnak a hallgatók. A hiányokat érzékelve a webes információ-visszakeresés tanítását felvettük a gyakorlati tananyagba, ami egy gyakorlati foglalkozást (két tanítási órát) köt le a rendelkezésre álló időkeretből. A modulra eső időkeret összesen 12-14 óra.

## 8.2. A kísérlet

Felmérésünk egyik eredménye szerint az információ visszakeresését a hallgatók legtöbbször magukra, illetve társaikra hagyatkozva tanulják, s csak a megkérdezett hallgatók ötöde nyilatkozta, hogy keresési készségének kialakításánál, azaz a keresések tanulásánál tanárának is szerepe volt.

**Hipotézisünk** szerint tanári irányítás mellett, alkalmas keresési feladatok elvégzése, az alkalmazott stratégiák és technikák megbeszélése által a hallgatók keresési készsége jelentősen növelhető.

### **A kísérlet független változói:**

- az általunk kidolgozott, a keresési stratégiák és technikák megismertetésére alkalmas feladatsor;
- a kísérlet céljának és a didaktikai követelményeknek megfelelő tanítási stratégia.

**A kísérlet függő változója** a különböző információs igények megválaszolásában nyújtott teljesítmény.

Kísérletünk egycsoportos kísérlet [Falus et al.,2000], ahol a kiindulási állapotot a tanítást megelőzően, a különböző keresési stratégiák és technikák alkalmazását igénylő, információs igények megválaszolásában nyújtott teljesítmény alapján mérjük. A kísérlet zárásakor az eredményt a modulban tanult ismereteket számonkérő zárthelyi dolgozatban mutatott teljesítmény adja.

### 8.3. Didaktikai alapok

A stratégiák a célok elérését szolgáló módszerek, eljárások, eszközök rendszere.

A célok szerint megkülönböztetünk

- az információ tanítására;
- a fogalomtanításra;
- az elemi készségek tanítására;
- a szociális és tanulási készségek tanítására, valamint
- a gondolkodás fejlesztésére

szolgáló tanulási stratégiákat [Falus, 2003]. Az öt, céljaiban egymástól jól elkülönülő stratégia az alkalmazott oktatási módszereket és eljárásokat illetően is különbözik egymástól. Az alkalmazott módszerek a célok felsorolásának sorrendjében: A bemutatás, a magyarázat és megbeszélés, a direkt oktatás, a kooperatív tanulás és a felfedezéssel tanulás.

A keresési készség fogalma alatt a kereséskor megfigyelhető, alkalmazott ismeretek összességét értjük. A keresési készség tanítására – mint ahogy az alapvető készségek és az elemi ismeretek elsajátítására – a direkt oktatás módszere a legmegfelelőbb. A direkt oktatás alapelve az, hogy a tanulók akkor sajátítják el leginkább az alapvető ismereteket és készségeket, ha a kitűzött célokat elemeire bontjuk, s határozott tanári irányítás mellett végigvezetjük a tanulókat az elsajátítás menetén. E módszernél a tanulók egyéni gyakorlásainak van jelentős szerepe.

### 8.4. Keresési stratégiák

Az információ-visszakeresés tanításakor fontosnak tartjuk tudatosítani a hallgatókban, hogy különböző jellegű információk igény más-más keresési stratégiát, valamint a stratégiának megfelelő keresőeszköz használatát igényli.

Nem elég egyetlen keresőt megismernünk, hiszen a témától vagy esetleg a földrajzi elhelyezkedéstől függően más-más keresőt érdemes használnunk a kívánt információ eléréséhez. Számos keresőprogram áll rendelkezésre (lásd <http://kereso.lap.hu/>). A sikeres kereséshez azonban a megfelelő keresőszoftver kiválasztásán túl fontos annak tulajdonságainak ismerete is: Nem mindegy,

hogyan fogalmazzuk meg a lekérdezést, hiszen az az eredményhalmazt döntően befolyásolja.

### 8.4.1. Tematikus keresők

A gyakorlati órákon feladott kereséseknél és a felmérésünk eredményéből (6. kérdés) is látszik, hogy a hallgatók nagyrészt szabadszavas keresőt használnak minden keresési feladat végrehajtásakor, ami megnehezíti egy-egy szélesebb tárgyterülethez tartozó információ visszakeresését.

Példák tematikus kereső használatát igénylő információs igényekre:

- *Soroljon fel legalább 10 operációs rendszert! Adjon meg egy-két szempontot, mi szerint lehet csoportosítani e szoftvereket!*
- *Adjon meg a biológiai kutatások terén minél több irányzatot!*
- *Milyen témákkal foglalkoznak az anyagtudományok?*

A kérdések nem tartalmazzák azt a szót, fogalmat, amire keresünk, így keresési kulcsszóként nem lehet szerepeltetni őket. (Egy extrém esetben egy hallgató az egész kérdést beírta a kereső mezőbe. Miután magyar nyelvű természetes nyelvi kereső még nem működik, így nem meglepő, hogy üres lett az eredményhalmaza.) A kérdésekben szereplő fogalmak (pl. operációs rendszer, biológia, stb.) túl általánosak, így csak messziről való közelítést tesznek lehetővé, ez pedig rengeteg találati oldalt jelent. Gondot okoz a témasodródás is: Például az operációs rendszerekre való lekérdezésnél a kevésbé népszerű operációs rendszerekről csak az eredménylista végén találunk oldalakat. Az ilyen típusú lekérdezések tematikus keresők használatát igénylik.

A tematikus keresőknél az összegyűjtött anyag jellege szerint lehet akadémiai vagy egyéb szakgyűjtemény, illetve a közönség széles rétegeinek szánt, minél több szolgáltatást nyújtó üzleti portál. E keresőknél is - éppúgy, mint a web egészén - jellemző az angol nyelv dominanciája. Kevés nemzeti nyelvű tudományos gyűjtemény létezik. Magyarországon leginkább a Magyar Elektronikus Könyvtár áll közel a katalógus rendszerekhez.

A legismertebb tematikus keresők:

Magyar Elektronikus Könyvtár <http://mek.oszk.hu/>

Wikipédia <http://www.wikipedia.org/>

MiMi <http://www.mimi.hu/>

Lap.hu oldalak <http://fogalom.lap.hu/>  
Elektronikus Információs szolgáltatás <http://www.eisz.hu/>  
Sulinet <http://www.sulinet.hu/>  
Librarians's Index to the Internet <http://lii.org/>  
Infomine - Scholarly Internet Resource Collections <http://infomine.ucr.edu/>  
Educational Subject Directory <http://www.academicinfo.net/>  
Open Directory Poject <http://www.dmoz.org/>  
About – What you need to know <http://www.about.com/>  
Encyclopedia Britannica <http://www.britannica.com/>  
Yahoo <http://www.yahoo.com/>

#### **8.4.1.1. A tematikus keresők használatának előnyei, hátrányai**

##### **Előnyök:**

- A kiválogatott dokumentumok témakörönként rendezve, csoportosítva találhatók meg, ami elősegíti egy adott téma áttekintését.
- A szakgyűjteményekben a dokumentumokat a témák szakértői válogatják be, így biztosított, hogy minőségileg kontrolált anyagot találunk.
- Az összegyűjtött anyag sokszor kiértékelve, magyarázatokkal ellátva érhető el.
- A legfrissebb dokumentumok (pl. hírek) itt találhatók meg, hiszen a keresőprogramoknak idő szükséges, míg adatbázisukba beemelik, s így kereshetővé teszik ezen oldalakat.

##### **Hátrányok:**

- A webnek szűkebb részét fedik le, mint a szabadszavas keresők adatbázisai.
- Szubjektív döntés áll a mögött, hogy mely oldalt hova soroljanak.
- A témakörönkénti csoportosítás sokszor nem elég részletes.
- Előfordulhat, hogy egy lap a hierarchia több kategóriájába is beletartozik.
- A felhasználók nem tudják, a hierarchia mely részén keressék kérdésükre a választ. Probléma lehet az is, ha a felhasználók a kategóriákra különböző fogalmat használnak, mint a katalógus rendszer.

#### 8.4.1.2. A jónak vélt URL cím beírása

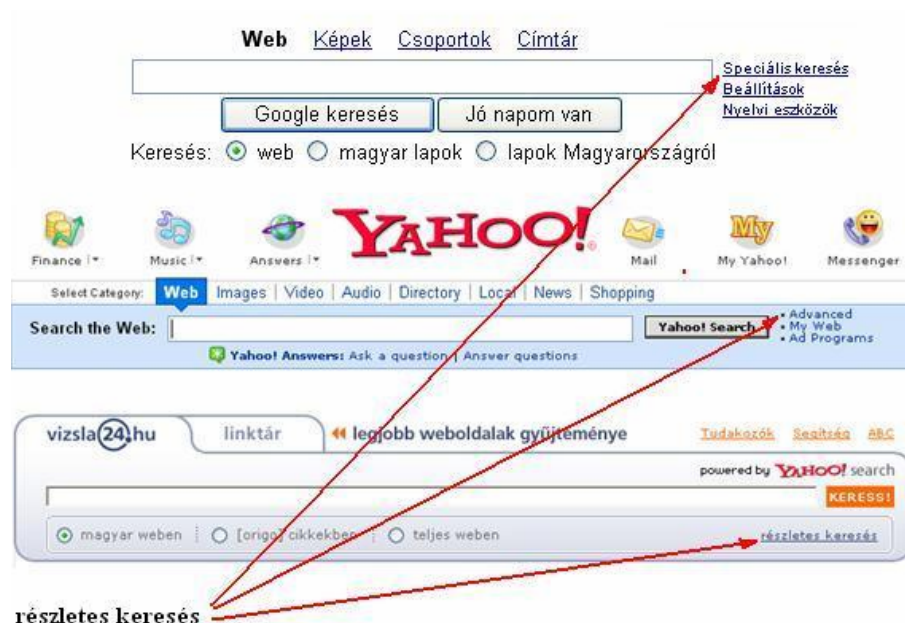
Egy-egy céggel, intézménnyel kapcsolatos információ keresésekor jó taktika a nem hivatalos `www.társaság_neve.com` általános címet kipróbálni (pl. `www.oracle.com`). Ekkor a társaság nyitólapjára kerülve, onnan a hiperhivatkozások, elágazások követve megtalálható a céggel, intézménnyel kapcsolatos információ. A `.com` mellett használatosak még az `.edu`, `.org`, `.net`, `.int`, valamint a `.gov`, `.mil` általános domainek, egyéb esetekben a cím legtöbbször az ország kódjára végződik.

#### 8.4.2. Szabadszavas keresők

A hallgatók az információ visszakeresésekor szinte mindig szabadszavas keresőt használnak, de - felmérésünk szerint - a többség nem ismeri és nem használja az összetett és/vagy a részletes keresést, így az alábbi típusú kérdések megválaszolására legfeljebb csak hosszadalmas böngészés után képesek:

- *Milyen speciálkollégiumokat hirdetnek az ELTE-n?*
- *Milyen adatokat publikáltak hazánkban az ipari növekedésről az elmúlt 3 hónapban?*
- *Milyen adatbázisokkal kapcsolatos PowerPoint fájlok találhatóak a weben?*
- *Keressen egy képet, amely egy darab piros almát ábrázol!*

Miután a hallgatók többsége csak a keresők kezdő oldalán található beviteli mezőt használja a lekérdezés megadásakor – ennek oka lehet, hogy nem tudnak a részletes keresés lehetőségéről –, szükségesnek látjuk, hogy egy-két keresőnél példaként megmutassuk a kereső nyitólapján található hivatkozást, amelyen át a részletes keresés előhívható.



22. ábra. A részletes keresés elérése különböző keresők esetén

A direkt oktatás elvének megfelelően fontosnak látjuk, hogy a keresési kulcsszavak közötti kapcsolatok előírására, a dokumentum nyelvére, fájl típusára, a megjelenés dátumára, az előfordulás helyére vonatkozó szűrési, szűkítési lehetőségeket a tanulók keresési feladatokon keresztül ismerjék meg. A leggyakoribb típusú összetett keresések a részletes keresés ablakban adhatók meg a legegyszerűbben, de célszerűnek tartjuk hangsúlyozni, hogy az összetett keresésekkel lehet legfinomabban az információs igényt megfogalmazni. Erre hasznos egy-két példát is bemutatni. (Pl. a következő keresési feladat utolsó sora csak összetett kereséssel adható meg.)

A keresések tanításánál fontosnak tartjuk egy keresési feladaton keresztül azt is bemutatni, hogy az információs igény pontosításával, a szűkítések megadásával az eredményhalmaz, s ez által a keresésre szánt idő mily mértékben csökkenthető.

**Keresési feladat:** *Keressünk információt Nagy Sándor századfordulós magyar festő munkásságáról, a nevéhez köthető (szecessziós) üvegablakokról!*

| A lekérdezés megadása                                                | A találatok száma* |
|----------------------------------------------------------------------|--------------------|
| Nagy Sándor                                                          | 3 260 000          |
| “Nagy Sándor”                                                        | 414 000            |
| festő "Nagy Sándor"                                                  | 17 900             |
| festő "Nagy Sándor" –hódító.                                         | 15 500             |
| festő "Nagy Sándor" –hódító. –makedón –macedón                       | 12 300             |
| Nagy Sándor üvegablak.                                               | 484                |
| "Nagy Sándor" üvegablak.                                             | 124                |
| "Nagy Sándor" üvegablakok                                            | 147                |
| (üvegablakok OR üvegablak) "Nagy Sándor "                            | 183                |
| (üvegablakok OR üvegablak) "Nagy Sándor " –hódító”                   | 147                |
| szecessziós (üvegablakok OR üvegablak) "Nagy Sándor””: kb.           | 79                 |
| ("szecessziós üvegablakok" OR "szecessziós üvegablak") "Nagy Sándor" | 3                  |

A tanórák alatt érdemes rávilágítani a kereséskor elkövetett gyakori hibákra is:

#### 8.4.2.1. Általános hibák

- Sokszor előforduló hiba, hogy a hallgatók magyar nyelvű szövegek keresésekor begépeléskor nem használják az ékezetes betűket (pl. *építész*et helyett). Ez esetben számos jelentős honlap kimarad az eredménylistából, s csak azok az oldalak jelennek meg, amelyeknél az URL címben szerepel a keresési kifejezés (pl. *epiteszet.html*), esetleg a szöveg létrehozója meta adatként keresési kulcsként hibásan is szerepelteti (gondolva az ékezet nélküli begépelésekre), vagy a dokumentum szövegében is helytelenül írva jelenik meg a keresett szó.
- Szintén kizárunk releváns oldalakat, hibásan szűkítjük a találati listát rossz helyesírással megadott keresési kulcsszóval. (Pl. *színész* – 1200 találat, *színész* – 37600 találat a Google esetében.)

---

\* A keresés a Google alkalmazásával történt.

- Sok esetben túl messziről közelítenek a hallgatók egy-egy keresésnél. Ha a gótikus építészet stílusjegyeire vagyunk kíváncsiak, akkor kevés az *építészet* keresési kulcs megadása.
- Érdeemes tudatosítani, hogy keresőnként különböző eredményeket kapunk, ha a szavakat egyes vagy többes számban szerepeltetjük. Ekkor célszerű lehet helyettesítő karaktert használni, bár nem minden kereső alkalmazza e technikát (pl. Vizsla24). A helyettesítő jel keresőnként különbözhet (pl. \*, % karakter).
- Sokan nem tesznek különbséget szavak és kifejezések keresése között. Főlegesen bővítjük az eredményhalmazt, ha kifejezéseket keresési kulcsszavaként adunk meg (osztott rendszerek “osztott rendszerek” helyett). Szókapcsolatok keresése esetén érdemes az idézőjelet kitenni vagy a részletes lekérdezés ablakánál a megfelelő mezőbe írni a szavakat, mert így az eredménylistát lényegesen szűkíthetjük.
- Megszorítások adhatók a dokumentum előfordulási helyére domainenként, site-onként. Szűkíthetjük az eredménylistát, ha megadjuk a dokumentum nyelvét, típusát, méretét, a dokumentumok létrehozására vonatkozó időkorlátot, stb. Természetesen keresőszoftverenként változik, hogy melyik milyen szűkítést enged meg, s milyen módon fogadja el a megadást, ezért egy-egy kereső használata előtt célszerű annak tulajdonságait megismerni (Segítség menü).

#### **8.4.2.2. A szabadszavas keresők használatának előnyei, hátrányai**

A tanítás folyamán fontos tudatosítani a hallgatókban a szabadszavas keresők használatának előnyeit és hátrányait.

##### **Mikor célszerű szabadszavas keresőt használni?**

- Ha nagyon speciális téma, különleges kulcsszavak iránt keresünk.
- Ha a webhelyek “mélyén” található lapok megtalálása a célunk.
- Ha nagy számú, érdeklődési körünknek megfelelő web oldalt szeretnénk visszakapni.

- Ha keresett dokumentumok típusára, forrására, nyelvére, keletkezésének időpontjára megszorításokat akarunk tenni.

#### **A szabadszavas keresők használatának hátrányai:**

- Túl sok találati oldal.
- Az információs igény pontos megfogalmazását igényli.
- Sokszor gond a rosszul megadott lekérdezés eredményének értelmezése.

Lényegesnek tartjuk hangsúlyozni azt is, hogy a különböző típusú információs igény más-más keresési stratégiát, és annak megfelelő eszköz (tematikus kereső, illetve szabadszavas kereső) használatát igényli.

### **8.4.3. Webes adatbázisok**

Számos hasznos információ különböző adatbázisokból nyerhető (pl. telefonszámok, menetrend, időjárási adatok, gazdasági információk, céginformációk, szaknévsor, adatbankok, stb.).

#### **Típuskérdések:**

- *Mekkora a távolság közúton Miskolc – Békéscsaba között?*
- *Mennyibe kerül a vonatjegy Egerből Debrecenbe?*
- *Milyen idő várható Londonban a következő 5 napra?*
- *Hány forintba kerül 1 norvég korona?*

A webes adatbázisok a láthatatlan web részét képezik. Egy-egy adatbázis elérését segítik a rendszerezett hivatkozások gyűjteményei, a témakatalógusok (pl. [www.invisible-web.net](http://www.invisible-web.net)), de egy-egy adatbázis kezdőlapjának megtalálásához a szabadszavas keresők is jól használhatók. Ekkor a keresés megfogalmazásakor célszerű a kulcsszavak megadása mellett a lekérdezést bővíteni az *adatbázisok* OR *archívumok* OR *gyűjtemények* megadásával, majd a kapott nyitóoldalon keresni az adatbázis saját lekérdező eszközeivel.

Érdeemes tudatosítani a diákokban, hogy az egyetemi könyvtárak nyitólapjáról számos tudományos adatbázis, folyóirat-gyűjtemény érhető el.

A metakeresők használatával számos weben keresztül elérhető adatbázist kérdezhetünk le egy időben, egységes lekérdező interfészen keresztül, így a web nagyobb részéből kapunk találatokat.

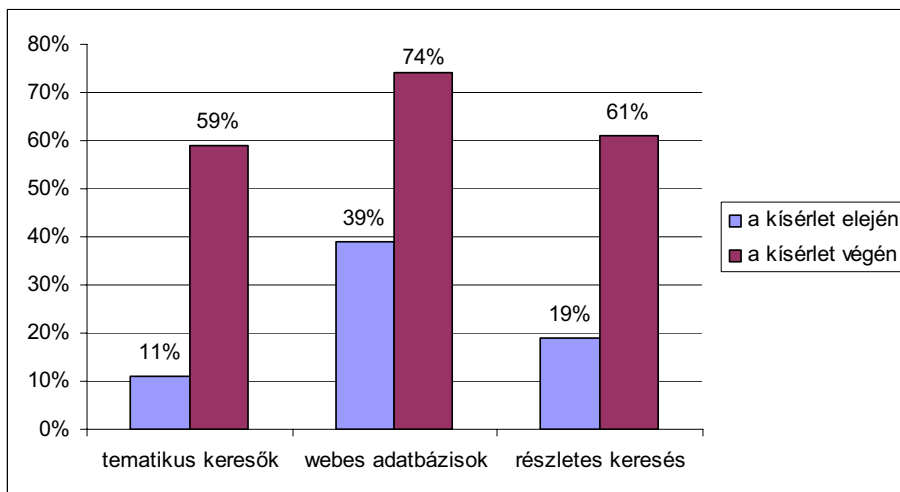
## 8.5. A kísérlet leírása és eredménye

A kísérlet moduljában a hallgatók a weblapkészítés alapjait sajátítják el, ahol az elkészítendő weblapoknak mindig előzetesen megadott kérdésekre kell választ adniuk. A kérdések összeállításánál törekszünk arra, hogy a válaszadások a különböző információ-visszakereső stratégiák és technikák, valamint az ezeknek megfelelő eszközök használatát igényeljék.

A modul elején, a kísérlet kezdetekor a hallgatók kérdéseket kaptak, amelyekre a válaszokat a weben kellett megkeresniük. Egy darabig önállóan kerestek a weben, és mentették a kérdésekre felelő oldalakat. Bizonyos idő után a kérdésekre vonatkozó válaszokat a csoport tagjai tanári irányítás mellett egyeztetették. Ekkor kérdeztünk rá, hogy ki milyen módon jutott eredményhez. Ha szükséges volt, ismertettük az adott kérdés megválaszolásához leginkább használható keresési stratégiát és technikát, amelyet a csoport tagjai a megbeszélés alatt ki is próbáltak. Lényegesnek tartottuk kiemelni, hogy mely esetekben mely keresési stratégiát, s az annak megfelelő kereső eszközt célszerű használni, valamint ezen eszközök előnyeit és hátrányait is ismertetni.

A kísérlet eredményességét a modult záró zárthelyi dolgozatokban mutatott teljesítmény alapján mértük. A számonkérés során is különböző kérdésekre válaszoló weboldalt kellett a hallgatóknak létrehozniuk. A kérdések hasonlatosak voltak a kereső eszközök ismertetésénél felsorolt típuskérdésekhez. A rendelkezésre álló idő rövideje miatt egy zárthelyi feladatsor tartalmazott mindhárom kereső eszköz (tematikus kereső, szabadszavas kereső, webes adatbázis) alkalmazását igénylő kérdéseket, de az eredményhalmaz szűkítési technikáit igénylő feladatokból általában csak egyet.

Egy-egy csoportban többféle zárthelyi feladatsort adtunk ki. A kísérlet értékelésekor csak azon hallgatók dolgozatait vettük figyelembe, akik jelen voltak azon a gyakorlaton, amelynek témája a webről történő információ-visszakeresés tanítása volt. A feltett kérdések megválaszolását igénylő kereső eszközök, technikák használatának arányát mutatja a következő ábra:



**23. ábra. Kereső eszközök használatának aránya a hallgatók között**

Az eredmények alátámasztják, hogy a webről történő információ-visszakeresés tanítása után jelentősen javult az információs igény megválaszolását lehetővé tevő kereső eszközök és technikák ismerete és használata. A kísérlet eredményét torzíthatja, ha a hallgatónak kevés volt az idő arra, hogy a megkeresett információkat a weblapján is megjelenítse.

## 9. Összegzés

A globalizált világ elvárása, hogy a tanulók ne konkrét ismereteket, hanem tudásuk gyarapításának képességét tanulják meg. Az információ elérése kiemelt fontosságú szerepet kap. Az Internet az emberi tudás határtalan tárházaként tekinthető, amelyben az információ keresése alapvető készsége kell legyen korunk minden polgárának.

Általános szabály, hogy egy gyakorlati feladat sikeres megvalósítását elősegíti a feladattal kapcsolatos elméleti háttér ismerete, így az információ-visszakeresés elméletének, a webes keresési eszközök tulajdonságainak ismerete is javíthatja keresési készségünket.

A hallgatók jelentős része nem ismeri a lehetőségeket egy-egy információ keresésekor, ezért növelni kell a diákok tudatosságát a keresési stratégiák, s a stratégiának megfelelő keresőeszközök kiválasztásánál és használatánál. A tapasztalat azt mutatja, hogy a többség szabadszavas keresőt használ akkor is, amikor tematikus kereső használata lenne indokolt. Az információs igénynek megfelelő stratégiát alkalmazva hatékonyabbá tehető a keresés.

A sikeres információ-visszakereséshez a megfelelő kereső kiválasztásán túl fontos annak tulajdonságainak ismerete is: nem mindegy, hogyan fogalmazzuk meg a lekérdezést, hiszen az az eredményhalmazt döntően befolyásolja. Felmérésünk szerint a hallgatók többsége nem ismeri az összetett és részletes keresés nyújtotta lehetőségeket. Az információ-visszakeresés hatékonysága érdekében meg kell ismertetni velük a visszakeresésekor alkalmazható különböző technikákat, amellyel az információs igény pontosítható, s ezáltal az eredményhalmaz szűkíthető, a keresésre szánt idő csökkenthető.

A szabadszavas keresők tulajdonságainak ismerete nemcsak az információs igény optimális megfogalmazásánál segít, de jól hasznosítható a webre szánt dokumentumok elkészítésénél is. A weblapok elkészítésének tanításakor hasznos azt is ismertetni, hogy az egyes szabadszavas keresők a dokumentumok mely tulajdonságait veszik figyelembe az eredményoldalak rangsorolásánál. Weboldalaink optimalizálásával javíthatjuk weboldalunk rangsorbeli helyezését.

A keresési készség kialakításánál növelni kell a tanárok szerepét. Kísérletünk igazolta, hogy már egy-két, a didaktikai elvárásoknak megfelelő tanítási óra után is jelentősen javítható a hallgatók keresési készsége.

## 10. Irodalomjegyzék

- [1] Amazon <http://www.amazon.com/>
- [2] Arocena,G., Mendelzon,A., Mihaila,G. (1997) Applications of a Web Query Language, *Computer Networks and ISDN Systems archive Volume 29*
- [3] Attar,R., Fraenkel,A.S. (1977) Local feedback in full-text retrieval systems. *Journal of the ACM*
- [4] Baeza-Yates,R., Ribeiro-Neto,B., (1999) Modern Information Retrieval, Addison Wesley
- [5] Barabasi, A-L., Albert, R., Jong H.(1999) Diameter of the World Wide Web. In: *Nature*
- [6] Barlow, L. (2001). The spider's apprentice: A helpful guide to Web search engines. Monash Information Services from <http://www.monash.com/spidap.html>
- [7] Bergman, M.K. (2001). The Deep Web: Surfacing Hidden Value. *Journal of Electronic Publishing*, 7(1).
- [8] Beyond General World Wide Web Searching, UC Berkeley Library <http://www.lib.berkeley.edu/>
- [9] Bharat,K., Broder. A.Z. (1998) *A technique for measuring the relative size and overlap of public web search engines*. In WWW7
- [10] Bilal,D. (2000) Children's use of Yahoo!igans! Web Search Engine: 1. Cognitive, physical, and affective behaviors on fact-based search tasks. *Journal of the American Society for Information Science*, 51 (7)
- [11] Blackwell Online <http://bookshop.blackwell.co.uk/>
- [12] Blelloch,G. (1997) Algorithms in the Real World <http://www.cs.cmu.edu/~guyb/real-world/indexing/>
- [13] Breadth first search and depth first search (1996) ICS 161: Design and Analysis of Algorithms, Lecture notes for February 15, 1996 <http://www1.ics.uci.edu>
- [14] Brewington, B.E. & Gybenko, G. (2000) How dynamic is the Web? WWW9.org. <http://www9.org/w9cdrom/264/264.html>
- [15] BrightPlanet <http://www.brightplanet.com>

- [16] Brin,S. és Page,L. (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine WWW7 / Computer Networks 30(1-7), pp. 107-117, <http://www.stanford.edu>
- [17] Brooks, T. (2003) Web search: how the Web has changed information retrieval, Information Research, Vol. 8 No. 3, April 2003
- [18] Bujdosó Gy., Students should learn to use word processors, Teaching Mathematics and Computer Science, Debrecen (megjelenés alatt)
- [19] Children and their Use of the Internet <http://www.pages.drexel.edu/~ab352/>
- [20] Cho,J. és Garcia-Molina,H. (2000) The Evolution of the Web and Implications for an Incremental Crawler. To appear in the 26th Intl. Conf. on Very Large Databases
- [21] Cockburn, A. & McKenzie, B. (2001) 'What do Web users do? An empirical analysis of Web use' *International Journal of Human-Computer Studies*, 54(6), <http://www.cosc.canterbury.ac.nz/andrew.cockburn/papers/ijhcsAnalysis.pdf>
- [22] Croft, W. B.és Harper,D. (1979). *Using probabilistic models of document retrieval without relevance information*. Journal of Documentation 35
- [23] Development of a European Service for Information on Research and Education <http://www.lub.lu.se/desire>
- [24] Dictionary of HTML META Tags <http://vancouver-webpages.com/META>
- [25] Dublin Core Metadata <http://dublincore.org>
- [26] Environics Research Group (2001) *Young Canadians in a wired world: the students' view* Report prepared for the Media Awareness Network and the Government of Canada. <http://www.media-awareness.ca/>
- [27] European Commission (2000) Memorandum az egész életen át tartó tanulásról, <http://www.om.hu/main.php?folderID=1027>
- [28] Falus,I. et al. (2000) Bevezetés a pedagógiai kutatás módszereibe, Műszaki Könyvkiadó
- [29] Falus,I. et al. (2003) Didaktika : elméleti alapok a tanítás tanulásához, Nemzeti Tankönyvkiadó
- [30] Friedman,E.,Uher,M.,Windhager,E.(2003) Keresés a Világhálón, Híradástechnika 58(3)

- [31] Greenspan, R. (2002) Search Engine Usage Ranks High, Cyberatlas,  
[http://cyberatlas.internet.com/markets/advertising/article/0,,5941\\_1500821,00.html](http://cyberatlas.internet.com/markets/advertising/article/0,,5941_1500821,00.html)
- [32] Griffiths, J., Brophy, P. (2002) Student searching behaviour in the JISC Information Environment, <http://www.ariadne.ac.uk/issue33/edner/>
- [33] Gulli, A., Signorini, A. (2005) The Indexable Web is more than 11.5 billion pages, WWW14, <http://citeseer.ist.psu.edu/context/2655616/0>
- [34] Gunn, H. és Hepburn, G. (2003) Seeking Information for School Purposes on the Internet Canadian Journal of Learning and Technology
- [35] How Much Information: Internet (2000),  
<http://www.sims.berkeley.edu/research/projects/how-much-info/internet.html>
- [36] Hu, Chen, Schmaly, Ritter (2001) An Overview of World Wide Web Search Technologies,  
<http://www.eng.auburn.edu/users/wenchen/publication/overview.ps>
- [37] Information Seeking Behavior <http://www.librarysupportstaff.com/infoseek.html>
- [38] Inktomi WebMap. <http://www.inktomi.com/webmap/>
- [39] Invisible Web <http://www.lib.berkeley.edu/TeachingLib/Guides/Internet>
- [40] iProspect (2004) Search Engine User Attitudes Survey <http://www.iprospect.com>
- [41] Kleinberg, Kumar, Raghavan, Rajagopalan, Tomkins (1999) The Web as a graph: Measurements, models and methods. Invited survey at the International Conference on Combinatorics and Computing
- [42] Kleinberg, J. (1998) Authoritative sources in a hyperlinked environment, 9th ACM-SIAM Symposium on Discrete Algorithms
- [43] Koehler, W. (1999) Digital libraries and the World Wide Web sites and page persistence. Information Research, 4, (4)  
<http://informationr.net/ir/4-4/paper60.html>
- [44] Lawrence, S., Giles, L. (1998) Searching the World Wide Web, Science, April 3
- [45] Lawrence, S., Giles, L. (1999) Accessibility and Distribution of Information on the Web, *Nature*, Vol. 400, pp. 107-109
- [46] Markwell, J. & Brooks, D.W. (2002) Broken links: just how rapidly do science education hyperlinks go extinct? Lincoln, NE: University of

- Nebraska-Lincoln. Department of Biochemistry.  
[http://www.class.unl.edu/biochem/url/broken\\_links.html](http://www.class.unl.edu/biochem/url/broken_links.html)
- [47] Melnik,S., Raghavan,S., Yang,B. és Garcia-Molina,H. (2001) Building a Distributed Full-Text Index for the Web WWW10, Hong-Kong
  - [48] Morris,C. So What's the Problem?  
[http://www.wdvl.com/Internet/Dead\\_SearchEngines](http://www.wdvl.com/Internet/Dead_SearchEngines)
  - [49] Mozaik Kiadó  
<http://www.mozaik.info.hu/HOMEPAGE/HUNMWRI/HIREK.HTM>
  - [50] Navarro, G.(1998) Approximate Text Searching , PhD tézis, Dept. of Computer Science, Univ. Of Chile
  - [51] Nemzeti Tankönyvkiadó <http://www.nettankonyv.hu/web/guest/home>
  - [52] Nielsen, J. (2001) Search: Visible and Simple *Alertbox*, <http://useit.com>
  - [53] On the size of the World Wide Web (2001) <http://www.pandia.com/sw-2001/57>
  - [54] O'Neil,E., Lavoie,B., Bennett (2003) Trends in the Evolution of the Public Web, D-Lib Magazine
  - [55] Országos Széchényi Könyvtár [http://www.oszk.hu/index\\_hu.htm](http://www.oszk.hu/index_hu.htm)
  - [56] Perinotti, T. (1997) How Search Engines Work, *Microsoft Interactive developer*
  - [57] Princz, M. (2001) Knowledge Discovery in Databases, ICAI, Eger
  - [58] Princz, M., Rutkovszky, E. (2004) Content Discovery of Invisible Web, ICAI, Eger
  - [59] Princz, M. (2006) A Survey on How Students Seek Information on the Internet, Teaching Mathematics and Computer Science, Debrecen
  - [60] Recommended Sites and Search Techniques  
<http://library.albany.edu/internet/search.html>
  - [61] Rutkovszky, E.,Princz, M. (2001) XML - the ideal language for weaving a better web, MicroCAD International Computer Science Conference
  - [62] Salton,G. and Buckley, C. (1990) *Improving retrieval performance by relevance feedback*, Journal of the American Society for Information Science, vol. 41.
  - [63] Schneiderman,B.(1997) Designing the User Interface: Strategies for Effective Humancomputer Interaction, Addison-Wesley
  - [64] SearchEngines <http://www.searchengines.com/>
  - [65] SearchEngineShowdown <http://www.searchengineshowdown.com>
  - [66] SearchEngineWatch <http://searchenginewatch.com>

- [67] Shapiro,Y.,Lehoczky,E. Factors that influence search engines rankings  
*<http://www.searchengines.com>*
- [68] Shivakumar, N.; Garcia-Molina, H.(1998) Finding near-replicas of documents on the web, Workshop on Web Databases
- [69] Sullivan,D. (2001) Search engine math. *Search Engine Watch*,  
*<http://searchenginewatch.com/facts/math.html>*
- [70] The Web Robots FAQ...*<http://www.robotstxt.org/wc/faq.html#log>*
- [71] Vargha, A. (2000) Matematikai statisztika, Pólya Kiadó
- [72] W3 <http://www.w3.org/>
- [73] W3C Metadata Area *<http://www.w3.org/Metadata/>*
- [74] WebReference *<http://www.webreference.com>*
- [75] What Is Natural-Language Searching?  
*<http://www.nwc.com/1120/1120f1side2.html>*
- [76] Wikipedia <http://www.wikipedia.org/>
- [77] Witten,I., Moffat,A., Bell,T. (1999) *Managing Gigabytes: Compressing and Indexing Documents and Images*, Morgan Kaufmann Publishing, San Francisco, ISBN 1-55860-570-3.
- [78] Zobel,J.,Moffat,A.,Ramamohanarao,K.. (1998) Inverted Files Versus Signature Files for Text Indexing

# 11. Publikációs jegyzék

## Referált dolgozatok

- [1] Princz M. Knowledge Discovery in Databases  
5<sup>th</sup> International Conference on Applied Informatics, Eger, 2001, p. 201-212, [Zentralblatt pre02065284]
- [2] Princz M. A Survey on How Students Seek Information on the Internet  
Teaching Mathematics and Computer Science,  
Debrecen, 4/1 (2006), p. 153-165
- [3] Princz M., Rutkovszky E. Content Discovery of Invisible Web  
6<sup>th</sup> International Conference on Applied Informatics, Eger, 2004
- [4] Princz M. Computer Aided Problem Solving  
Teaching Mathematics and Computer Science,  
Debrecen, megjelenés alatt

## Lektorált dolgozatok

- [5] Princz M. A weben lévő információk hozzáférhetősége  
Debreceni Műszaki Közlemények, 2004/1, p. 51-66
- [6] Princz M. E-learning  
Debreceni Műszaki Közlemények 2005/1, p. 93-113

## Egyéb cikk

- [7] Princz M. Keresési stratégiák a Web-en  
DE MFK Tudományos Közlöny 2000, p. 209-212., Debrecen

## Nemzetközi konferencia kiadványban megjelent cikkek és előadások

- [8] Princz M. Systems to access information in the Web  
MicroCAD'2000 International Computer Science Conference, p 169-173, Miskolc, 2000
- [9] Princz M., Rutkovszky E. Datamining  
MicroCAD'2001 International Computer Science Conference, p 117-125, Miskolc, 2001

- [10] Rutkovszky E., Princz M. XML - the ideal language for weaving a better web  
MicroCAD'2001 International Computer Science Conference, p 209-214, Miskolc, 2001
- [11] Princz M. Információkeresési stratégiák a weben és tanításuk tapasztalatai  
14<sup>th</sup> International Conference in Computer Science and Education  
ISBN 973-86097 p 28-34, Kolozsvár, 2004.
- [12] Princz M. Hallgatók webes információ keresési készségének felmérése  
16<sup>th</sup> International Conference in Computer Science and Education  
ISBN-10 973-7840-12-7 p.56-68, Szováta, 2006.

#### **Hazai konferencia kiadványban megjelent cikkek és előadások**

- [13] Princz M. Az informatika oktatásának helyzete a KLTE MFK-n  
Informatika a felsőoktatásban II. kötet p. 641-645., Debrecen, 1999
- [14] Princz M., Rutkovszky E. Trendek az adatbányászatban  
NetworkShop 2001, Sopron, CD melléklet  
<http://www.niif.hu/rendezvenyek/networkshop/>
- [15] NetworkShop 2001, Sopron, CD melléklet  
<http://www.niif.hu/rendezvenyek/networkshop/>
- [16] Princz M., Rutkovszky, E. Ismeret reprezentáció a weben  
NetworkShop 2002, Eger, CD melléklet  
<http://www.niif.hu/rendezvenyek/networkshop/>
- [17] Princz M., Husi G. A webes keresők használatának tanítása  
Informatika a felsőoktatásban, Debrecen, 2002, CD melléklet
- [18] Husi G., Princz M. Informatikával támogatott oktatási projekt a Hajdú-Bihar Megyei Rendőr-főkapitányságon  
Informatika a felsőoktatásban, Debrecen, 2002, CD melléklet
- [19] Princz M. A weben lévő információk hozzáférhetősége  
NetworkShop 2003, Pécs, CD melléklet  
<http://www.niif.hu/rendezvenyek/networkshop/>
- [20] Princz M. A webes keresők tanításának tapasztalatai  
E-learning alkalmazások a hazai felsőoktatásban,

Budapest, 2003., p. 185-190

<http://www.szamalk.hu/okk/E-learning/Program.htm>

- [21] Princz M. Tájékozódás a weben

NetworkShop 2004, Győr, CD melléklet

<http://www.niif.hu/rendezvenyek/networkshop/>

- [22] Princz M., Nagy J. A hallgatók keresési szokásai

Informatika a felsőoktatásban, Debrecen, 2005, CD melléklet

### **Elektronikus formában elérhető tananyagok**

- [23] Műszaki informatika II (MF MUI 02M03) tárgy előadásainak és gyakorlatainak tananyaga (DE MFK)

<http://delfin.unideb.hu/~pmaria/> (folyamatos fejlesztés alatt)

### **Előadások**

- [24] Nagyszámítógépes és mikrogépes rendszerek működése a SZÜV-ben  
Csillag Pál Szimpózium, Balatonlelle, 1988

- [25] Az Internet Kukac

SZÜV, Debrecen, 1990

- [26] Adatbányászat

INFO ÉRA 2000, <http://www.szikszi.hu/info/adatbanyaszat.html>

Békéscsaba, 2000

- [27] Információkeresési stratégiák

DE MFK ÁMI Intézeti tudományos nap, Debrecen, 2000

- [28] Adatbázisrendszerek jövője

DE MFK ÁMI Intézeti tudományos nap, Debrecen, 2000

- [29] Keresőgépek

INFO ÉRA 2001, <http://www.infoera.hu/infoera2001/info4.shtml>

- [30] E-business

DE MFK Gépészeti Szakmai Napok, Debrecen, 2003

- [31] A hallgatók keresési szokásai

INFO ÉRA, Békéscsaba, 2005

- [32] Felmérés a hallgatók keresési készségéről

Menedzsment és Alkalmazott Informatika Szeminárium, Cserépfalu, 2006

## Szoftverek

### *A Szerzői Jogvédő Hivatalnál bejegyzett szoftvertermékek*

1983-1991 között a SZÜV, Debrecen munkatársaként az alábbi szoftverek fejlesztésében vettem részt programozóként. A szoftverek PL1-ben, illetve Clipper programnyelven íródtak.

- [33] Elő- és utókalkulációs rendszer – Medicor (kb. 3500 sor)
- [34] Bérrendszer – Hajdú Megyei Tejipari Vállalat (kb. 1900 sor)
- [35] Termelésirányítási rendszer – Hajdú Megyei Tejipari Vállalat (kb. 1500 sor)
- [36] Értékesítési rendszer – Barnevál (kb. 2700 sor)
- [37] Üzletelszámoltatási rendszer – Hajdú Megyei Vendéglátóipari Vállalat (kb. 1500 sor)
- [38] Termelés elszámolás – Hajdú Megyei Húsipari Vállalat (kb. 1700 sor)
- [39] Állóeszköz nyilvántartási rendszer – TÜZÉP, Nyiregyháza (kb. 7000 sor)
- [40] Osztott-keret nyilvántartás – DOTE (kb. 2500 sor)

### *Egyéb szoftvertermékek*

Az alábbi rendszereket 100%-ban magam készítettem:

- [41] Menetlevél nyilvántartási, számlázási rendszer – HOMEX Kft, 1992 (kb. 5000 sor)
- [42] Fuvar-elszámoltatási rendszer – AGROS Kft, 1993 (kb. 4000 sor)
- [43] Menetlevél nyilvántartási, számlázási rendszer – Herczeg Kft, 1993 (kb. 5000 sor)
- [44] Hardver nyilvántartási rendszer – Ybl Miklós Műszaki Főiskola, 1994 (kb. 2000 sor)
- [45] Főiskolai Gopher adatbázis létrehozása – Ybl Miklós Műszaki Főiskola, 1994
- [46] Általános és Menedzserképző Intézet honlapja – DE MFK, 2000

## 12. Summary

### 12.1. Introduction, Verification of the Choice of the Topic

Today's knowledge centered society supposes needs and encourages its members' continuous learning. One condition of the long term and dynamic development of the global economy is that a greater and greater majority of people should have access to intellectual capital and use it. The already acquired intellectual capital devolves fast, expertise and knowledge become outdated. Thus individual success does not depend on actual specific knowledge or the amount of information but on the ability to acquire more, that is, on learning above all.

A Memorandum on Lifelong Learning [EC, 2000] issued by the Commission of European Communities in October 2000 puts the emphasis on a new learning culture and also includes forms of learning outside school. The emphasized basic skills are the following: competence in information technology, foreign languages, technical culture, enterprises and social skills (like self confidence, independence and risk taking).

Another basic and very important skill is the ability to learn and acquiring this ability also means adapting to changes and finding one's way in the ever more extended information flow. New methods and techniques help people to learn and one of the most important of these is the use of the computer and the internet and the possibilities provided by web interfaces. The Internet is becoming more and more widespread in everyday life and thus in education as well. E-learning systems supply carefully chosen and designed knowledge in formal and less formal ways. Of course there are question related to the topic under discussion to which the answers are not contained in the material provided, but there are questions coming up in everyday life the answers to which are searched for according to individual interest and previous knowledge. When doing so, apart from consulting a competent person (teacher, librarian, expert), information available on the web is more and more frequently used. The

ability to deal with this huge amount of information is a vital need for the members of our information society and a necessary condition of successful informal web based learning.

## **12.2. Research Goals**

The web can be considered a huge mass of information, navigating on which is what our time expects from us. When seeking information on the web, the most widely used tool, free word search engines, are the best known of all information seeking systems.

The research contains three parts regarding information seeking on the web:

### **12.2.1. Review of the Theory of Information Retrieval**

Information retrieval is a very wide area which includes information storage, display and retrieval as well. It incorporates many research fields (e.g. algorithms, use of metadata, multimedia retrieval, artificial intelligence, and natural language processing, etc.).

The literature of information retrieval related research fields is rich [Blackwell, Amazon], but due to the large extension of the field, there are only a few books written in English [Baeza-Yates&Ribeiro-Neto,1999] to give a systematic overall view of the topic, and there are no books available in Hungarian. [OSZK]. In the first part of our dissertation we undertake to give an overall review of this wide and rich field, from our point of view focusing on information retrieval from the web, and make it available in Hungarian as well.

The development and maintenance of the main tool used for searching the web, that of free word search engines, is more a business and also involves patented technology on a strong market. This implies that we know about how they work only what the developers consider safe to publish and this usually means few technical details. The main tools for learning are articles available on the web, conference material, information available on special web ports as well as trials and inferences.

- In the first part of our thesis discuss the basic concepts of information retrieval, how search engines work, models of information retrieval, and the characteristics of documents. We consider the languages used to describe the form and content of documents. We review query types and we also discuss some techniques used in queries. Through examples we explain methods developed for indexing documents and searching them. We also show the possibilities of searching in multimedia objects.
- At the beginning of the chapter *Information Retrieval from the Web* we overview the problems occurring while searching the web for information and the distribution of the web. It is widely known that free word search engines map only a smaller part of the web, so we need to mention the existence of then invisible web and the possibilities to search it. We discuss three basic types of web search: search based on key-word, topic hyper-reference and the tools used. We overview the structure of free word search engines, being the most widely used tools, and the general and specific characteristics of the gathering, indexing, querying and ranking parts. We discuss multiple and detailed searches that make information needs more specific. We also show solutions to improve ranking. We give an example of searches based on hyper-references, and discuss the taxonomy of web-mining.
- In the chapter *Optimizing Web* pages, we study what influences the order of the search results. If we know what features determine where our web page is put in the ranking, then optimizing pages designed for the web can improve the place taken by our page and with that better ranking the information can reach more people. Thus knowing about the characteristics of free word search engines can be useful in designing documents for the web.

### **12.2.2. A Survey on How Students Seek Information on the Internet**

Navigating among the information available on the Internet has become an expectation for the members of the information society we are living in. This especially applies to students of higher education, the intellectuals of the future.

There are a large number of papers discussing users' searching strategies and behaviour [iProspect,2004, Greenspan,2002, Nielsen,2001, Environics Research Group, 2001]. A Canadian study [Gunn&Hepburn,2003] undertaken among 12<sup>th</sup> grade high school students examining their Internet search skills was of special interest to us.

We made a survey among the 1<sup>st</sup> year students of the Faculty of Technical Engineering at Debrecen University on how they can navigate on the web. During the seminars we could see that:

- our students search using only a few key words;
- the majority is not familiar with the possibilities provided by free word search engines to specify the required information, and they cannot narrow down or filter the result set;
- they are not aware that certain information needs require different search tools.

Seeing the flaws we were also curious to find out how widespread this phenomenon is, that is how students at other faculties of the university search the web.

1. With this object we developed questionnaire that was answered by 200 full time students of the University of Debrecen (UD) representative of the ratio of the number of students at the faculties. We were seeking to answer the following questions in connection with the students' knowledge of information location:
2. What Internet services do the students use and how often?
3. What information sources do they prefer in their study related assignments next to their textbooks?
4. How do they perceive their own ability to locate information on the Internet?
5. What Internet access possibilities do students have?

6. How do students learn to seek information on the Internet?
7. What knowledge do students have of the various Internet search engines?
8. What strategies and techniques do students use to enhance information seeking on the Internet?
9. What influences students' search skills?
10. What searching habits do students have?
11. How do they rate the effectiveness of search engines?

We compiled the answers given to our questions and analysed them and also compared them with other similar survey results.

### **Main Results:**

- Among the students of Debrecen University the most frequently used Internet services are the web (99%) and the e-mail (86%). These percentages are the same as those in international surveys [Greenspan,2002, Environics Research Group, 2001]. Our students search for information on the web for personal purposes more often than for their studies.
- Most students at the University of Debrecen equally are in favour of the World Wide Web and books, but all in all there is a majority of those who prefer print sources in addition to textbooks. We find this fact a bit surprising as textbooks contain the course material, while the corresponding research results are mostly available on the World Wide Web first: it takes a considerably longer time to publish print material than publishing on the World Wide Web, and due to the limited financial resources many print materials aren't available in the libraries of the institutions.
- Although most students use few effective strategies when searching World Wide Web for information and use even fewer services to limit and refine the search results they consider their search skills to be average or good. Positive self assessment is even more obvious with Canadian students.

- Most students at UD pleased possibilities to connect to the Internet in their educational institutions. Half of students reported that they had Internet connection at home.
- The results of the surveys show that students rely mostly on themselves (60%) and their mates (44%) when learning how to locate information on the World Wide Web, and teachers have a less important role (21%).
- The 3 most popular search engines in the USA (Google, Yahoo, MSN) [SearchEngineWatch] are among the best-known among the students. Ask Jeeves, which is popular among American and Canadian high school students, is fairly unknown among our students. The most popular search engine in Hungarian is Vizsla24 (with 85%), then come Heuréka and Góliát.
- According to our survey students use 2.35 search engines on average at a time and 1.13 search engines regularly, which gives an average of 3.48 search engine use with some regularity. This number is higher, than the corresponding one in the Canadian survey, where this number is 1-2 among the users. [Greenspan,2002, Gunn&Hepburn,2003]
- When specifying their information requirements 76% of the students at Debrecen University do not use logical operators, 54% do not use expressions, 66% do not use substitute characters, and 64% do not use refined search. The results among Canadian students were very similar.
- 90 % of our students are not familiar with techniques for narrowing down search results (e.g. filter by file format, date, occurrence, domain and language).
- When analysing our survey we would have liked to answer the question what students' search skills depend on, more precisely if it is influenced by the student's sex, their interest in computer science, their access to the Internet or at which faculty they study. Based on the answers of the students we could not statistically find a significant correspondence between access to the Internet (either at home or in the educational institution) and search skills. Statistical tests have

shown, on the other hand, that search skills are influenced by the interest in computer science. Various tests have shown a correlation between the students' sex and their search skills, but further investigations (two-conditional variation analysis) have proved that sex influences search skills through the interest in computer science, but on its own, it is independent from it. We have come to the same conclusion examining the search skills of students studying at the various faculties: there was a significant difference in results among the students studying at the Faculty of Informatics and the students studying at the other faculties.

- When the first search is not successful is typical for our students to refine the search query (85%) and only 12% will try another search engine. Our student overview a larger portion of the search results than average users do.
- Various estimates are available on the web on the size of the invisible web, which say that it is more than 10 times larger than the visible web [Bergman, 2001, BrightPlanet]. There have been technical efforts taken to make the invisible web visible [Princz, Rutkovszky, 2004]. The invisible web is growing very fast, so making its content available is an urgent need. Using free word search engines users can find a sufficient amount of information on the web in almost any topic, so they are likely to suppose that they can reach all information available on the web with this software. This is what our survey also shows [Princz, 2006]. The estimates, given by students at the various faculties regarding the size of indexed web sites, ranged from 48% to 72%, which gives an average of 64% at the university. These are very high percentages and they show an overestimation of search engines.

### **12.2.3. A Pedagogical Experiment to Improvement Searching Skill**

One of our survey results is that in most cases students rely on themselves and their peers when learning search skills, and only 5% of them claimed that their teachers also helped them.

Our hypothesis is that completing suitable search tasks with teacher guidance, discussing the applied strategies and techniques can improve students' search skills.

The independent variables of the experiment were:

- a task sheet we developed to familiarize search strategies and techniques,
- a teaching strategy suitable for the goals and didactic needs of the experiment.

The dependant variable in the experiment is the varying performance depending on the various information requirements.

Our experiment is a single group experiment [Falus et al.,2000], where the initial state is measured on the basis of performance given while searching for information requiring various search strategies and techniques.

For the teaching of search skills, just like for basic skills and knowledge, direct methods are the best [Falus et al.,2003]. The main idea of direct instruction is that students acquire the basic skills and knowledge if we break the goals down into sub-goals and with firm teacher guidance students are led along the process. In this method students' individual practicing is vital. At the end of the experiment the result is provided by the performance given in a test asking for the acquired knowledge.

The result of the experiment supports that after teaching information retrieval from the web the use and familiarity with search tools and techniques needed to gain the required information have considerably improved.

### **12.3. Conclusion**

Our globalized world expects the students to learn how to widen their knowledge rather than acquiring specific knowledge. Finding information is of vital importance. The Internet is an endless warehouse of human knowledge, where information search skills are basic for everybody.

It is a rule of the thumb that knowing the theoretical background to a practical problem helps solving it, so knowing about the theory of information retrieval and the characteristics of web search tools can improve our search skills.

Most students are not familiar with the possibilities, so their awareness needs to be increased in search strategies and in choosing and using the appropriate search tools. Experience shows that the majority uses free word search engines even when a thematic engine would be appropriate.

According to our survey most of our students do not know about multiple and refined searches. So to increase their efficiency they need to learn about the various techniques applicable in information retrieval, with which the information need can be made more specific and the result set narrowed down, and thus the time consumption decreased.

Knowing about the characteristics of free word search engines helps not only with the better specification of the information requirement but also with designing web documents. When teaching web page design it is also worth discussing what features of the document influence their ranking in certain free word search engines. By optimizing our web pages, we can improve their ranking.

The role of the teacher in developing search skills should be increased. Our results have proved that one or two didactically appropriate classes can considerably improve students' search skills.

# 13. Függelék

## 13.1. Kérdőív

**A KÖVETKEZŐ KÉRDŐÍV KITÖLTÉSÉVEL KÉREM SEGÍTSE MUNKÁNKAT!**

**A VÁLASZADÁS ÖNKÉNTES ÉS NÉVTELEN, VÁLASZAIT BIZALMASAN KEZELJÜK!**

Egyes kérdéseknél értelemszerűen több válasz is lehetséges! ( III. illetve X. kérdés)

**I. A Debreceni Egyetem melyik karán tanul?**

-----  
**II. Neme**

1. férfi            2. nő

**III. Érdeklődik az informatika iránt?**

1. nem    2. korlátozottan    3. közepesen    4. igen    5. a kedvencem

**IV. Milyen Internet szolgáltatásokat használ? *Több válasz is lehetséges!***

1. www    2. e-mail    3. ftp    4. telnet    5. chat

**V. Tanulmányaihoz szokott az Interneten információt, tananyagot gyűjteni?**

1. nem    2. alkalmanként    3. rendszeresen

**VI. Szokott az Interneten keresni személyes (nem oktatási) célból?**

1. nem      2. alkalmanként      3. rendszeresen

**VII. Tanulmányaihoz tankönyvein kívül mely információs forrásokat részesít előnyben?**

1. web      2. könyvek      3. egyformán mindkettő      4. egyik sem

**VIII. Hogyan jellemezné keresési készségét az Interneten?**

1. gyenge      2. közepes      3. jó

**IX. Oktatási intézményében rendelkezésére állnak olyan számítógépek, amelyeken át az Internet elérhető?**

1. nem      2. korlátozottan      3. igen

**X. Otthonában rendelkezik Internet eléréssel?**

1. nem      2. igen

**XI. Kitől tanult meg az Interneten keresni? *Több válasz is lehetséges!***

1. tanár      2. szülő / családtag      3. barát / osztálytárs      4. magától tanulta meg

## XII.

| Hallott már az alábbi keresőkről? | Ha a válasz igen → |     | Milyen gyakran használja? |              |              |
|-----------------------------------|--------------------|-----|---------------------------|--------------|--------------|
| Kereső szolgáltató                | Igen               | Nem | Soha                      | Alkalmanként | Rendszeresen |
| AlltheWeb                         | 1                  | 2   | 1                         | 2            | 3            |
| AltaVista                         | 1                  | 2   | 1                         | 2            | 3            |
| AskJaves                          | 1                  | 2   | 1                         | 2            | 3            |
| Google                            | 1                  | 2   | 1                         | 2            | 3            |
| Inktomi                           | 1                  | 2   | 1                         | 2            | 3            |
| Lycos                             | 1                  | 2   | 1                         | 2            | 3            |
| MSN                               | 1                  | 2   | 1                         | 2            | 3            |
| Teoma                             | 1                  | 2   | 1                         | 2            | 3            |
| Yahoo                             | 1                  | 2   | 1                         | 2            | 3            |
| Góliát                            | 1                  | 2   | 1                         | 2            | 3            |
| Heuréka                           | 1                  | 2   | 1                         | 2            | 3            |
| Vizsla                            | 1                  | 2   | 1                         | 2            | 3            |

**XIII. Ha a fenti felsorolásban nincs benne az Ön által leggyakrabban használt kereső, kérem, adja azt meg!**

-----

**XIV. A keresés megadásakor:**

|                                                                               | nem | alkalmanként | rendszeresen |
|-------------------------------------------------------------------------------|-----|--------------|--------------|
| Használ-e logikai operátorokat (AND, OR, NOT, NEAR)?                          | 1   | 2            | 3            |
| Használja a + vagy a - jelet a szükséges ill. a tiltott szavak megadására?    | 1   | 2            | 3            |
| Használ idézőjelet ( “ ”) kifejezések keresésénél?                            | 1   | 2            | 3            |
| Használ helyettesítő karaktereket csak részlegesen ismert szavak keresésekor? | 1   | 2            | 3            |

**XV. Tudja-e, hogyan zárhatók ki a keresés eredményhalmazából a kereskedelmi webhelyek vagy egyéb nem kívánatos tárhelyek?**

1. nem      2. igen

*Ha a válasza igen, akkor kérem, írja le azt is, hogyan!*

.....

**XVI. Tudja-e, hogyan adható meg, hogy a keresés csak az utolsó 3 hónapban módosított vagy a Webre elhelyezett dokumentumok között történjen?**

1. nem      2. igen

*Ha a válasza igen, akkor kérem, írja le azt is, hogyan!*

.....

**XVII. Tudja-e, hogyan adható meg, hogy a keresés csak a lapok megnevezésében (ne a teljes szövegben) történjen?**

1. nem      2. igen

*Ha a válasza igen, akkor kérem, írja le azt is, hogyan!*

---

**XVIII. Használja a keresők speciális szolgáltatását különböző fájltypusok (pl. kép, hang) keresésekor?**

1. nem
2. igen

*Ha a válasza igen, akkor kérem, írja le azt is, hogyan!*

---

**XIX. Az első keresés sikertelensége esetén mit tesz?**

1. tovább finomítja a kereső kérdést
2. másik keresővel próbálkozik.
3. abbahagyja a keresést

**XX. Mi jellemző az Ön keresési szokására?**

1. csak az első egy-két eredményt nézi meg
2. végignézi az első eredményoldalt
3. további eredményoldalakat is átnéz.

**XXI. Mit gondol, a legjobb keresők a weben lévő dokumentumok hány százalékában keresnek?**

.....

**Köszönjük közreműködését!**

## 13.2. A hallgatók keresési készségét befolyásoló tényezők statisztikai vizsgálata

A hallgatók keresési készségét meghatározza, hogy mennyire alkalmazzák a különböző keresési stratégiákat és technikákat, amelyeket a kérdőív XIV. – XVIII. kérdései mérnek. A XIV. kérdés 5 alkérdésében a keresési stratégiák használatára kérdeztünk rá, a XV.-XVIII. kérdés a szűrési lehetőségek ismeretére kérdez.

| A kérdés száma | Tartalma                                                    | A leíró változó neve |
|----------------|-------------------------------------------------------------|----------------------|
| XIV/1.         | logikai operátorok használata                               | XIV.logikai          |
| XIV/2.         | + - operátorok használata                                   | XIV.+ -              |
| XIV/3.         | kifejezés használata                                        | XIV.kifejezés        |
| XIV/4.         | helyettesítő karakterek használata                          | XIV.joker            |
| XIV/5.         | speciális keresés ablak használata                          | XIV.speciális        |
| XV.            | szűrés webhely vagy tartomány szerint                       | XV.domain            |
| XVI.           | szűrés időpont szerint                                      | XVI.idő              |
| XVII.          | szűrés a keresett kifejezések előfordulásának helye szerint | XVII.hely            |
| XVIII.         | szűrés a dokumentum típusa szerint                          | XVIII.formátum       |

**24. táblázat. A keresési készségre vonatkozó kérdéseket tartalmazó változók**

A további vizsgálatokhoz szükséges e 9 kérdésre adott választ tartalmazó változókat “összevonni” úgy egy változóvá, hogy a műveletvégzés a lehető legkevesebb információvesztéssel járjon, vagyis a transzformált sokaságról az eredeti sokaságéval azonos következtetéseket lehessen levonni. Az összevont változónak a hallgatók keresési készségét kell leírnia. Az összevonás előtt mérlegeltük, hogy egy-egy kérdésre adott nemleges válasz azért született-e, mert

habár a szabadszavas kereső lehetővé teszi, de a hallgató mégsem használja a kínált lehetőséget, vagy a használt kereső nem is kínálja fel a vizsgált stratégiát, technikát. A keresési készség vizsgálatánál a szabadszavas keresőknek csak azon tulajdonságait látjuk célszerűnek figyelembe venni, amelyek a leggyakrabban használt keresők többségének egyaránt sajátjai.

A 6. kérdésre adott válaszok vizsgálatánál már láttuk, hogy a hallgatóknak több, mint a fele 3 szabadszavas keresőt használ különböző rendszerességgel: a Google-t, a Yahoo-t és a Vizslát. A keresési készséget leíró változó képzésénél csak azon stratégiákat, technikákat vesszük figyelembe, amelyeket az említett 3 kereső mindegyike felkínál.

A Yahoo nem használ helyettesítő karaktereket (joker karaktereket), e keresőnél nem lehetséges kizárni a keresésből bizonyos domainekeket, site-okat, valamint nem lehetséges annak a megadása, hogy a keresés csak a dokumentum különböző részeiben (pl. TITLE, címsorok, stb.) történjen. Ezen megfontolásokból az összevonásból kizárjuk *XIV.joker*, *XV.domain* valamint a *XVII.bely* változókat.

A változók számának további csökkentésére **faktoranalízist** alkalmazunk a maradék 6 változóra.

### 13.2.1. Faktoranalízis

**Correlation Matrix<sup>a</sup>**

|                 |                 | XIV. logikai | XIV. +- | XIV. kifejezés | XIV. speciális | XVI. idő | XVIII. formátum |
|-----------------|-----------------|--------------|---------|----------------|----------------|----------|-----------------|
| Correlation     | XIV. logikai    | 1,000        | ,229    | ,412           | ,238           | ,470     | ,514            |
|                 | XIV. +-         | ,229         | 1,000   | ,262           | ,101           | ,208     | ,346            |
|                 | XIV. kifejezés  | ,412         | ,262    | 1,000          | ,260           | ,384     | ,506            |
|                 | XIV. speciális  | ,238         | ,101    | ,260           | 1,000          | ,246     | ,206            |
|                 | XVI. idő        | ,470         | ,208    | ,384           | ,246           | 1,000    | ,396            |
|                 | XVIII. formátum | ,514         | ,346    | ,506           | ,206           | ,396     | 1,000           |
| Sig. (1-tailed) | XIV. logikai    |              | ,001    | ,000           | ,000           | ,000     | ,000            |
|                 | XIV. +-         | ,001         |         | ,000           | ,077           | ,002     | ,000            |
|                 | XIV. kifejezés  | ,000         | ,000    |                | ,000           | ,000     | ,000            |
|                 | XIV. speciális  | ,000         | ,077    | ,000           |                | ,000     | ,002            |
|                 | XVI. idő        | ,000         | ,002    | ,000           | ,000           |          | ,000            |
|                 | XVIII. formátum | ,000         | ,000    | ,000           | ,002           | ,000     |                 |

a. Determinant = ,295

#### KMO and Bartlett's Test

|                                                  |                    |         |
|--------------------------------------------------|--------------------|---------|
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy. |                    | ,803    |
| Bartlett's Test of Sphericity                    | Approx. Chi-Square | 239,439 |
|                                                  | df                 | 15      |
|                                                  | Sig.               | ,000    |

#### Communalities

|                 | Initial | Extraction |
|-----------------|---------|------------|
| XIV. logikai    | 1,000   | ,570       |
| XIV. + -        | 1,000   | ,247       |
| XIV. kifejezés  | 1,000   | ,541       |
| XIV: speciális  | 1,000   | ,206       |
| XVI. idő        | 1,000   | ,488       |
| XVIII. formátum | 1,000   | ,613       |

Extraction Method: Principal Component Analysis.

#### Total Variance Explained

| Component | Initial Eigenvalues |               |              | Extraction Sums of Squared Loadings |               |              |
|-----------|---------------------|---------------|--------------|-------------------------------------|---------------|--------------|
|           | Total               | % of Variance | Cumulative % | Total                               | % of Variance | Cumulative % |
| 1         | 2,664               | 44,402        | 44,402       | 2,664                               | 44,402        | 44,402       |
| 2         | ,924                | 15,399        | 59,801       |                                     |               |              |
| 3         | ,794                | 13,231        | 73,032       |                                     |               |              |
| 4         | ,638                | 10,637        | 83,668       |                                     |               |              |
| 5         | ,549                | 9,145         | 92,813       |                                     |               |              |
| 6         | ,431                | 7,187         | 100,000      |                                     |               |              |

Extraction Method: Principal Component Analysis.

#### Component Matrix<sup>a</sup>

|                 | Component |
|-----------------|-----------|
|                 | 1         |
| XIV. logikai    | ,755      |
| XIV. + -        | ,497      |
| XIV. kifejezés  | ,735      |
| XIV: speciális  | ,453      |
| XVI. idő        | ,699      |
| XVIII. formátum | ,783      |

Extraction Method: Principal Component Analysis.

a. 1 components extracted.

#### 25. táblázat. A faktoranalízis eredménye

A korrelációs mátrix a változók közötti kapcsolat erősségét mutatja. E szerint a legkevésbé a *XIV.speciális* és a *XIV.+* - változók között van kapcsolat. E két változó esetében a korreláció mértéke 0,101, a szignifikanciaszint\* (0,077) meghaladja az 5%-ot, ami a két változó bizonyos mértékű lineáris függetlenségére utal.

A faktoranalízisben szereplő Kaiser-Meyer-Olkin szám magas értéke (0,803) azt mutatja, hogy az adatokra a faktoranalízis nagyon jól használható. A korrelációs mátrix alatt a teljes függetlenség Bartlett-féle próbája azt teszteli, hogy van-e egyáltalán valamilyen elméleti lineáris típusú kapcsolat az elemzésbe bevont változók között. Esetünkben 0,000 a szignifikanciaszint, ami azt jelenti, hogy a tesztelt változók többsége lineárisan függ egymástól.

A faktoranalízist végrehajtva egy faktorváltozót kapunk eredményül (Keresési készség). A hat változó esetében az SPSS által kínált módszerek közül a főkomponens módszert alkalmazva jutottunk a legmagasabb variancia százalékhoz, amelynek értéke 44,402.

A komunalitások táblájából leolvasható, hogy a létrehozott faktorváltozó a mért változók szóródásának hány százalékát magyarázza (Extraction), míg a komponens mátrix azt adja meg, hogy a változók milyen mértékben hatnak a faktorváltozóra. Látható, hogy nincs túl nagy különbség a súlyozásban, az egyes változók közel hasonló mértékben befolyásolják a faktorváltozót.

A további elemzések végett megvizsgáljuk a faktoranalízis eredményeként kapott *Keresési készség* nevű faktorváltozó normalitását, hiszen a lineáris regresszió, illetve a variancia analízis feltétele a változók normális eloszlása és szóráshomogenitása.

A *Keresési készség* változó normalitását a nem-paraméteres próbák közül a Kolmogorov-Szmirnov-próbával vizsgáljuk, illetve elkészítjük a vizsgált változó hisztogramját is.

$H_0$ : A *Keresési készség* változó normális eloszlású

$H_1$ : A *Keresési készség* változó nem normális eloszlású

---

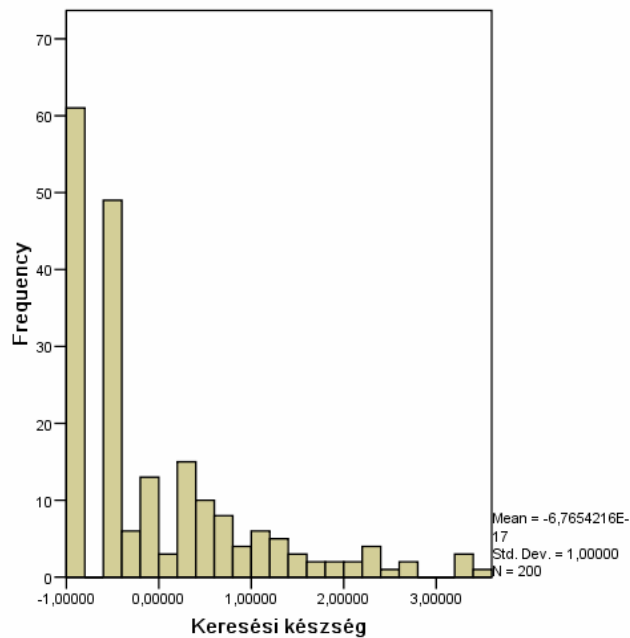
\* A próba terjedelmét az SPSS szignifikanciaszintként szerepelteti. A továbbiakban ezt a jelölést használjuk.

| One-Sample Kolmogorov-Smirnov Test |                |                      |
|------------------------------------|----------------|----------------------|
|                                    |                | Keresési<br>késztség |
| N                                  |                | 200                  |
| Normal Parameters <sup>a,b</sup>   | Mean           | ,0000000             |
|                                    | Std. Deviation | 1,0000000            |
| Most Extreme<br>Differences        | Absolute       | ,216                 |
|                                    | Positive       | ,216                 |
|                                    | Negative       | -,200                |
| Kolmogorov-Smirnov Z               |                | 3,051                |
| Asymp. Sig. (2-tailed)             |                | ,000                 |

a. Test distribution is Normal.  
b. Calculated from data.

26. táblázat. A Keresési készség változó normalitásának vizsgálata

A Kolmogorov-Szmirnov-próba szignifikanciaszintje 0,000, ami a  $H_0$  hipotézis, azaz a *Keresési készség* változó normalitásának elutasítását jelenti.



24. ábra. A Keresési készség változó hisztogramja

A változó eloszlásának nem normális volta a változó hisztogramjáról is egy pillanat alatt megállapítható. A Gauss-görbének ugyan látszanak a körvonalai, de a változó hisztogramja erős ferdeséget mutat. Ennek oka, hogy nagyon sok hallgató alig használja az információk igény megfogalmazását elősegítő operátorokat. Ha a változót leíró statisztikákat is kérjük az SPSS-től, akkor az eloszlásgörbe ferdesége és meredeksége is leolvasható: ferdeség=1,485, meredekség=1,773.

| Descriptive Statistics |           |           |           |           |           |           |            |           |            |
|------------------------|-----------|-----------|-----------|-----------|-----------|-----------|------------|-----------|------------|
|                        | N         | Minimum   | Maximum   | Mean      | Std.      | Skewness  |            | Kurtosis  |            |
|                        | Statistic | Statistic | Statistic | Statistic | Statistic | Statistic | Std. Error | Statistic | Std. Error |
| Keresési készség       | 200       | -,84047   | 3,43349   | ,0000000  | 1,000000  | 1,485     | ,172       | 1,773     | ,342       |
| Valid N (listwise)     | 200       |           |           |           |           |           |            |           |            |

27. táblázat. A Keresési készség változó leíró statisztikája

### 13.2.2. Homogenitás-vizsgálatok

A továbbiakban azt vizsgáljuk, hogy a hallgatók keresési készségére mi van befolyással, függ-e az a hallgató Internethez való hozzáféréstől, az informatika iránti érdeklődésétől, a hallgató nemétől, illetve van-e különbség a különböző karok hallgatóinak keresési készségében?

A homogenitás-vizsgálatoknál figyelembe kell vennünk, hogy a *Keresési készség* változó nem normális eloszlású, ami a paraméteres próbák alkalmazásának előfeltétele. Viszont a normalitás megsértése ritkán vezet súlyos következményekre, ha a mintaelemszámok nem túl kicsik (pl. egyszempontos variancia-analízis esetén mintánként eléri a 10-et). Ha a mintaelemszámok egyenlők, akkor a varianciaanalízis az elméleti varianciák közepes mértékű különbségét még elviseli. A különböző mintaelemszámok és a különböző varianciák esetén robusztus paraméteres próbákat kell alkalmazni, amelyek eltekintenek a változók normalitásának és a szórások egyenlőségének feltételétől. [Vargha, 2000].

Ha feltételeik teljesülnek, a paraméteres próbák nagyobb erejűek, mint a helyettük alkalmazható nem-paraméteres próbák. Amennyiben nem-paraméteres módszereket alkalmazunk, nem követünk el hibát a próba feltételek nem

teljesülése miatt. A nem-paraméteres próba viszonylagos ereje nagyobb, mint a robosztus (korrektül nem alkalmazható) paraméteres próba ereje.

### 13.2.2.1. Az Internethez való hozzáférés és a keresési készség kapcsolata

Az otthoni Internethez való hozzáférésre a kérdőív X. kérdésében (Otthonában rendelkezik Internet eléréssel?) kérdeztünk rá, s az erre adott válaszokat a X. *Internet otthon* változó tartalmazza.

| A válasz      | A válasz kódja | A válaszolók száma | A válaszolók száma %-ban |
|---------------|----------------|--------------------|--------------------------|
| nem           | 1              | 100 fő             | 50,0%                    |
| igen          | 2              | 99 fő              | 49,5%                    |
| nem válaszolt |                | 1 fő               | 0,5%                     |
| összesen      |                | 200 fő             | 100,0%                   |

28. táblázat. A X. Internet otthon változó leíró statisztikája

Kérdésünk megválaszolásához a nem-paraméteres próbák közül a Mann-Whitney-Wilcoxon, valamint a Kolmogorov-Szmirnov-próbát alkalmazzuk, de robosztus t-próbával is megvizsgáljuk a kapcsolatot.

A két populációt azok a mintaelemek alkotják, ahol az X. *Internet otthon*=1 (ha a hallgató nem rendelkezik otthoni Internet hozzáféréssel), s X. *Internet otthon*=2 (ha elérhető az otthonában az Internet.).

H<sub>0</sub>: a két populáció sztochasztikusan ugyanakkora a *Keresési készség* tekintetében

H<sub>1</sub>: a két populáció sztochasztikusan eltér a *Keresési készség* tekintetében

| Test Statistics <sup>a</sup> |                  |
|------------------------------|------------------|
|                              | Keresési készség |
| Mann-Whitney U               | 4858,000         |
| Wilcoxon W                   | 9908,000         |
| Z                            | -,230            |
| Asymp. Sig. (2-tailed)       | ,818             |

a. Grouping Variable: X. Internet otthon

29. táblázat. Mann-Whitney-Wilcoxon-próba az X. Internet otthon és a Keresési készség változókra

**Test Statistics<sup>a</sup>**

|                             |          | Keresési<br>készség |
|-----------------------------|----------|---------------------|
| Most Extreme<br>Differences | Absolute | ,083                |
|                             | Positive | ,083                |
|                             | Negative | -,034               |
| Kolmogorov-Smirnov Z        |          | ,582                |
| Asymp. Sig. (2-tailed)      |          | ,887                |

a. Grouping Variable: X. Internet otthon

**29. táblázat. Kolmogorov-Szmirnov-próba az X. Internet otthon és a Keresési készség változókra**

Mindkét nem-paraméteres próba esetén a szignifikanciaszint (0,818 ill. 0,887) magasabb 5%-tól, ami  $H_0$  hipotézisünk megtartását is jelenti egyben. A paraméteres próbák közül lefuttatjuk a robosztus, független mintás t-próbát is hipotézisünk eldöntésére.

$H_0$ : a két populáció átlaga megegyezik

$H_1$ : a két populáció átlaga nem egyezik meg

**Independent Samples Test**

|                  |                                | Levene's Test for<br>Equality of Variances |      | t-test for Equality of Means |         |                 |                    |                          |                                                 |           |
|------------------|--------------------------------|--------------------------------------------|------|------------------------------|---------|-----------------|--------------------|--------------------------|-------------------------------------------------|-----------|
|                  |                                | F                                          | Sig. | t                            | df      | Sig. (2-tailed) | Mean<br>Difference | Std. Error<br>Difference | 95% Confidence<br>Interval of the<br>Difference |           |
| Keresési készség | Equal variances<br>assumed     | 2,164                                      | ,143 | -,693                        | 197     | ,489            | -,09862450         | ,14225649                | -,379166                                        | ,18191654 |
|                  | Equal variances<br>not assumed |                                            |      | -,693                        | 194,062 | ,489            | -,09862450         | ,14233771                | -,379352                                        | ,18210297 |

**30. táblázat. Független mintás t-próba az X. Internet otthon és a Keresési készség változókra**

A t-próba eredménye a szórások megegyezésének és különbözőségének esetében is ugyanaz: 0,489-es szignifikanciaszint mellett megtartjuk a  $H_0$  hipotézist. Evvel igazoltuk, hogy az Internethez való otthoni hozzáférés és a keresési készség között nincs kapcsolat.

A kérdőív IX. kérdése (Oktatási intézményében rendelkezésére állnak olyan számítógépek, amelyeken át az Internet elérhető?) szintén az Internethez

való hozzáférésre kérdez. E kérdésnél 3 lehetséges válasz közül választhattak a hallgatók, s az erre adott válaszokat a *IX. Internet mh.* változó tartalmazza.

| A válasz      | A válasz kódja | A válaszolók száma | A válaszolók száma %-ban |
|---------------|----------------|--------------------|--------------------------|
| nem           | 1              | 4 fő               | 2,0%                     |
| korlátozottan | 2              | 38 fő              | 19,0%                    |
| igen          | 3              | 158 fő             | 79,0%                    |
| nem válaszolt |                | 0 fő               |                          |
| összesen      |                | 200 fő             | 100,0%                   |

**31. táblázat. Az IX. Internet mh változó leíró statisztikája**

A *IX. Internet mh* változóra adott válaszok három populációt alkotnak. Azt a kérdést, hogy a *Keresési készség* változó értékei a *IX. Internet mh.* változó 3 populációjában megegyeznek-e, a Kruskal-Wallis nem-paraméteres próbával teszteljük.

$H_0$ : A *Keresési készség* változó eloszlása ugyanakkora *IX. Internet mh.* változó mindhárom populációjában

$H_1$ : A *Keresési készség* változó eloszlása különbözik *IX. Internet mh.* változó három populációjában

| Ranks            |                  |     |           |
|------------------|------------------|-----|-----------|
|                  | IX. Internet mh. | N   | Mean Rank |
| Keresési készség | 1                | 4   | 129,13    |
|                  | 2                | 38  | 97,71     |
|                  | 3                | 158 | 100,45    |
|                  | Total            | 200 |           |

**Test Statistics<sup>a,b</sup>**

|             | Keresési készség |
|-------------|------------------|
| Chi-Square  | 1,102            |
| df          | 2                |
| Asymp. Sig. | ,577             |

a. Kruskal Wallis Test

b. Grouping Variable: IX. Internet mh.

**32. táblázat. Kruskal-Wallis-próba az IX. Internet mh és a Keresési készség változókra**

A próba eredményeként  $H_0$  hipotézist 0,577-es szignifikanciaszint mellett megtartjuk, ami az Internethez való intézményi hozzáférés és a keresési készség közötti korreláció hiányát jelenti. Evvel igazoltuk, hogy a hallgatók keresési készsége nem függ az Internethez való hozzáférésüktől.

### 13.2.2.2. Az informatika iránti érdeklődés és a keresési készség kapcsolata

A további elemzésnél szeretnénk azt is megvizsgálni, hogy az informatika iránti érdeklődés mennyire befolyásolja a hallgató keresési készségét? Az informatika iránti érdeklődésre a kérdőív harmadik kérdésében kérdeztünk (Érdeklődik az informatika iránt?). A válaszokat a *III.érdeklődés* változó tartalmazza.

| A válasz      | A válasz kódja | A válaszolók száma | A válaszolók száma %-ban |
|---------------|----------------|--------------------|--------------------------|
| nem           | 1              | 11 fő              | 5,5%                     |
| korlátozottan | 2              | 33 fő              | 16,5%                    |
| közepesen     | 3              | 88 fő              | 44,0%                    |
| igen          | 4              | 55 fő              | 27,5%                    |
| a kedvencem   | 5              | 13 fő              | 6,5%                     |
| nem válaszolt |                | 0 fő               | 0,0%                     |
| összesen      |                | 200 fő             | 100,0%                   |

33. táblázat. A *III.érdeklődés* változó leíró statisztikája

A *Keresési készség* és a *III.érdeklődés* változók homogenitásának eldöntésére egyszempontos robosztus varianciaanalízist , illetve a nem-paraméteres próbák közül a Kruskal-Wallis-próbát hajtjuk végre.

$H_0$ : a keresési készség az informatika iránti érdeklődéstől független

$H_1$ : a keresési készséget az informatika iránti érdeklődés meghatározza

### ANOVA

| Keresési készség |                |     |             |        |      |
|------------------|----------------|-----|-------------|--------|------|
|                  | Sum of Squares | df  | Mean Square | F      | Sig. |
| Between Groups   | 62,940         | 4   | 15,735      | 22,551 | ,000 |
| Within Groups    | 136,060        | 195 | ,698        |        |      |
| Total            | 199,000        | 199 |             |        |      |

### Robust Tests of Equality of Means

| Keresési készség |                        |     |        |      |
|------------------|------------------------|-----|--------|------|
|                  | Statistic <sup>a</sup> | df1 | df2    | Sig. |
| Welch            | 11,521                 | 4   | 39,950 | ,000 |
| Brown-Forsythe   | 17,219                 | 4   | 37,994 | ,000 |

a. Asymptotically F distributed.

**34. táblázat. Robosztus egyszempontos varianciaanalízis a III.érdeklődés és a Keresési készség változókra**

### Test Statistics<sup>a,b</sup>

|             | Keresési készség |
|-------------|------------------|
| Chi-Square  | 45,949           |
| df          | 4                |
| Asymp. Sig. | ,000             |

a. Kruskal Wallis Test

b. Grouping Variable: III. inf. érdeklődés

**35. táblázat. Kruskal-Wallis-próba a III.érdeklődés és a Keresési készség változókra**

Mindkét próba eredményeként  $H_0$  hipotézisünket el kell utasítanunk 0,000-s szignifikanciaszinten, azaz a próbák alapján azt mondhatjuk, hogy a keresési készséget az informatika iránti érdeklődés meghatározza. Ha tovább elemezzük, hogy mi az összefüggés az informatika iránti érdeklődés foka és a keresési készség között, akkor a Post Hoc tesztből leolvasható, hogy valójában csak azon hallgatók esetében van összefüggés, akik a kérdésre az “igen” vagy „a kedvencem” választ adták, a többiek esetében (a megkérdezettek 66%-a) a válaszok és a keresési készség között nem mutatható ki kapcsolat.

# Multiple Comparisons

Dependent Variable: Keresési készség

Games-Howell

| (I) III. inf. érdeklődés | (J) III. inf. érdeklődés | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval |             |
|--------------------------|--------------------------|-----------------------|------------|------|-------------------------|-------------|
|                          |                          |                       |            |      | Lower Bound             | Upper Bound |
| 1                        | 2                        | -,17375168            | ,22889617  | ,939 | -,8774729               | ,5299696    |
|                          | 3                        | -,35078331            | ,21771801  | ,516 | -1,0366294              | ,3350628    |
|                          | 4                        | -,89120856*           | ,24202363  | ,012 | -1,6181303              | -,1642868   |
|                          | 5                        | -2,4165577*           | ,45055969  | ,000 | -3,7823649              | -1,0507504  |
| 2                        | 1                        | ,17375168             | ,22889617  | ,939 | -,5299696               | ,8774729    |
|                          | 3                        | -,17703163            | ,12887073  | ,646 | -,5381122               | ,1840489    |
|                          | 4                        | -,71745689*           | ,16667925  | ,000 | -1,1819201              | -,2529936   |
|                          | 5                        | -2,2428060*           | ,41498262  | ,001 | -3,5405532              | -,9450588   |
| 3                        | 1                        | ,35078331             | ,21771801  | ,516 | -,3350628               | 1,0366294   |
|                          | 2                        | ,17703163             | ,12887073  | ,646 | -,1840489               | ,5381122    |
|                          | 4                        | -,54042526*           | ,15096239  | ,005 | -,9606192               | -,1202313   |
|                          | 5                        | -2,0657744*           | ,40892327  | ,002 | -3,3551501              | -,7763986   |
| 4                        | 1                        | ,89120856*            | ,24202363  | ,012 | ,1642868                | 1,6181303   |
|                          | 2                        | ,71745689*            | ,16667925  | ,000 | ,2529936                | 1,1819201   |
|                          | 3                        | ,54042526*            | ,15096239  | ,005 | ,1202313                | ,9606192    |
|                          | 5                        | -1,5253491*           | ,42236542  | ,019 | -2,8338500              | -,2168482   |
| 5                        | 1                        | 2,41655768*           | ,45055969  | ,000 | 1,0507504               | 3,7823649   |
|                          | 2                        | 2,24280600*           | ,41498262  | ,001 | ,9450588                | 3,5405532   |
|                          | 3                        | 2,06577437*           | ,40892327  | ,002 | ,7763986                | 3,3551501   |
|                          | 4                        | 1,52534911*           | ,42236542  | ,019 | ,2168482                | 2,8338500   |

\*. The mean difference is significant at the .05 level.

36. táblázat. Post Hoc teszt a III.érdeklődés és a Keresési készség változókra

## 13.2.2.3. A nemek és a keresési készség kapcsolata

A felmérésben szereplő hallgatók nemének eloszlását a II. nem változó tartalmazza.

| A válasz      | A válasz kódja | A válaszolók száma | A válaszolók száma %-ban |
|---------------|----------------|--------------------|--------------------------|
| Férfi         | 1              | 62 fő              | 31,0%                    |
| Nő            | 2              | 138 fő             | 69,0%                    |
| nem válaszolt |                |                    |                          |
| Összesen      |                | 200 fő             | 100,0%                   |

37. táblázat. A II. nem változó leíró statisztikája

$H_0$ : a keresési készség a nemtől független

$H_1$ : a keresési készséget a hallgató neme meghatározza

**Test Statistics<sup>a</sup>**

|                        | Keresési készség |
|------------------------|------------------|
| Mann-Whitney U         | 2975,500         |
| Wilcoxon W             | 12566,500        |
| Z                      | -3,496           |
| Asymp. Sig. (2-tailed) | ,000             |

a. Grouping Variable: II. nem

**38. táblázat. Mann-Whitney-Wilcoxon-próba a II. nem és a Keresési készség változókra**

**Test Statistics<sup>a</sup>**

|                          | Keresési készség                 |
|--------------------------|----------------------------------|
| Most Extreme Differences | Absolute<br>Positive<br>Negative |
|                          | ,256<br>,256<br>-,007            |
| Kolmogorov-Smirnov Z     | 1,676                            |
| Asymp. Sig. (2-tailed)   | ,007                             |

a. Grouping Variable: II. nem

**39. táblázat. Kolmogorov-Szmirnov-próba a II. nem és a Keresési készség változókra**

**Independent Samples Test**

|                  |                             | Levene's Test for Equality of Variances |      | t-test for Equality of Means |        |                 |                 |                       |                                           |           |
|------------------|-----------------------------|-----------------------------------------|------|------------------------------|--------|-----------------|-----------------|-----------------------|-------------------------------------------|-----------|
|                  |                             | F                                       | Sig. | t                            | df     | Sig. (2-tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference |           |
|                  |                             |                                         |      |                              |        |                 |                 |                       | Lower                                     | Upper     |
| Keresési készség | Equal variances assumed     | 4,161                                   | ,043 | 3,299                        | 198    | ,001            | ,49230968       | ,14922929             | ,19802691                                 | ,78659246 |
|                  | Equal variances not assumed |                                         |      | 3,017                        | 96,498 | ,003            | ,49230968       | ,16316766             | ,16844578                                 | ,81617358 |

**40. táblázat. Független mintás t-próba a II. nem és a Keresési készség változókra**

A különböző próbák szintje mindenütt 5% alatt marad, ami azt jelenti, hogy  $H_0$  hipotézisünket el kell utasítanunk, azaz a keresési készséget a hallgató neme meghatározza. De valóban különbség tehető a nők és férfiak keresési készsége között? Nem arról van inkább szó, hogy valamely más tényező, mondjuk az informatika iránti érdeklődés erőssége különbözik nemenként oly erősen, amely a két nem keresési készségében a különbséget eredményezi?

Ennek a feltételezésnek az eldöntéséhez kétszemponos független mintás, robusztus varianciaanalízist hajtunk végre, ahol a két csoportosító változó *II. nem* és *III.érdeklődés*, a függő változónk pedig a *Keresési készség*.

$H_0(A)$ : a *Keresési készség* elméleti átlaga független a *II. nem* változótól

$H_0(B)$ : a *Keresési készség* elméleti átlaga független a *III.érdeklődés* változótól

$H_0(AB)$ : a *Keresési készség* elméleti átlaga független a *II. nem* és a *III.érdeklődés* változók interakciójától

$H_1(A)$ : a *Keresési készség* elméleti átlaga függ a *II. nem* változótól

$H_1(B)$ : a *Keresési készség* elméleti átlaga függ a *III.érdeklődés* változótól

$H_1(AB)$ : a *Keresési készség* elméleti átlaga függ a *II. nem* és a *III.érdeklődés* változók interakciójától

#### Tests of Between-Subjects Effects

Dependent Variable: Keresési készség

| Source                  | Type III Sum of Squares | df  | Mean Square | F      | Sig. |
|-------------------------|-------------------------|-----|-------------|--------|------|
| Corrected Model         | 64,092 <sup>a</sup>     | 8   | 8,012       | 11,343 | ,000 |
| Intercept               | 6,387                   | 1   | 6,387       | 9,042  | ,003 |
| II.nem                  | ,003                    | 1   | ,003        | ,004   | ,949 |
| III.érdeklődés          | 50,824                  | 4   | 12,706      | 17,989 | ,000 |
| II.nem * III.érdeklődés | ,792                    | 3   | ,264        | ,374   | ,772 |
| Error                   | 134,908                 | 191 | ,706        |        |      |
| Total                   | 199,000                 | 200 |             |        |      |
| Corrected Total         | 199,000                 | 199 |             |        |      |

a. R Squared = ,322 (Adjusted R Squared = ,294)

41. táblázat. Kétszemponos varianciaanalízis a *II. nem* és a *III.érdeklődés*, valamint a *Keresési készség* változókra

A próba eredményéből leolvashatjuk, hogy a *Keresési készség* változó független a *II. nem* változótól (Sig.=0,949), valamint a *II. nem* és a *III.érdeklődés* változók interakciójától (Sig.=0,772), s függ a *III.érdeklődés* változótól (Sig.=0,000). Ez gyanakvásunkat támasztja alá: A nem az informatikai érdeklődésen keresztül befolyásolja csak a keresési készséget, önmagában azonban független attól.

#### 13.2.2.4. Milyen eltérések vannak az egyes karokon tanulók keresési készségében?

Végezetül szeretnénk tesztelni azt is, hogy az egyes karokon tanuló hallgatók keresési készsége között van-e különbség. A kérdőívet kitöltő hallgatók karonkénti megoszlását az alábbi táblázat tartalmazza:

| Karkod | I.kar | Megnevezés                                | Fő     | Százalék |
|--------|-------|-------------------------------------------|--------|----------|
| 1      | ÁJK   | Állam- és Jogtudományi Kar                | 9 fő   | 4,5%     |
| 2      | ÁOK   | Általános Orvostudományi Kar              | 19 fő  | 9,5%     |
| 3      | AVK   | Agrárgazdasági és Vidékfejlesztési Kar    | 15 fő  | 7,5%     |
| 4      | BTK   | Bölcsészettudományi Kar                   | 36 fő  | 18,0%    |
| 5      | DK    | Debreceni Konzervatórium                  | 2 fő   | 1,0%     |
| 6      | EFK   | Egészségügyi Főiskolai Kar                | 16 fő  | 8,0%     |
| 7      | FOK   | Fogorvostudományi Kar                     | 3 fő   | 1,5%     |
| 8      | GYTK  | Gyógyszerésztudományi Kar                 | 3 fő   | 1,5%     |
| 9      | HPFK  | Hajdúböszörményi Pedagógiai Főiskolai Kar | 8 fő   | 4,0%     |
| 10     | IK    | Informatikai Kar                          | 11 fő  | 5,5%     |
| 11     | KTk   | Közgazdaságtudományi Kar                  | 10 fő  | 5,0%     |
| 12     | MFK   | Műszaki Főiskolai Kar                     | 19 fő  | 9,5%     |
| 13     | MTK   | Mezőgazdaságtudományi Kar                 | 15 fő  | 7,5%     |
| 14     | TTK   | Természettudományi Kar                    | 34 fő  | 17,0%    |
|        |       | Összesen                                  | 200 fő | 100%     |

42. táblázat. A kérdőívet kitöltő hallgatók karonkénti megoszlása

A minta megoszlása megegyezik a 2004. őszén a Debreceni Egyetem nappali tagozatán tanuló hallgatók karonkénti megoszlásának arányával. A minta reprezentativitása miatt az egyes karokat csak néhány fő képviseli. A varianciaanalízis alkalmazhatósága végett [Vargha,2000] a 10 fő alatt képviselt karokat összevonjuk az alábbi módon:

| Karkod | Össze-<br>vont<br>karkod | I.kar | Megnevezés                                | Fő    |
|--------|--------------------------|-------|-------------------------------------------|-------|
| 1      | 1                        | ÁJK   | Állam- és Jogtudományi Kar                | 9 fő  |
| 2      | 2                        | ÁOK   | Általános Orvostudományi Kar              | 19 fő |
| 5      | 1                        | DK    | Debreceni Konzervatórium                  | 2 fő  |
| 7      | 2                        | FOK   | Fogorvostudományi Kar                     | 3 fő  |
| 8      | 2                        | GYTK  | Gyógyszerésztudományi Kar                 | 3 fő  |
| 9      | 1                        | HPFK  | Hajdúböszörményi Pedagógiai Főiskolai Kar | 8 fő  |

43. táblázat. A kis létszámú karok összevonása

$H_0$ : a keresési készség karonként homogén

$H_1$ : a keresési készség karonként nem homogén

Hipotézisünket nem-paraméteres próbával és robusztus varianciaanalízissel is teszteljük.

| Ranks            |                  |     |           |
|------------------|------------------|-----|-----------|
|                  | Összevont karkod | N   | Mean Rank |
| Keresési készség | 1,00             | 19  | 79,13     |
|                  | 2,00             | 25  | 103,54    |
|                  | 3,00             | 15  | 78,90     |
|                  | 4,00             | 36  | 90,00     |
|                  | 6,00             | 16  | 99,53     |
|                  | 10,00            | 11  | 163,64    |
|                  | 11,00            | 10  | 121,85    |
|                  | 12,00            | 19  | 119,92    |
|                  | 13,00            | 15  | 88,73     |
|                  | 14,00            | 34  | 98,94     |
|                  | Total            | 200 |           |

**Test Statistics<sup>a,b</sup>**

|             | Keresési<br>készség |
|-------------|---------------------|
| Chi-Square  | 23,925              |
| df          | 9                   |
| Asymp. Sig. | ,004                |

a. Kruskal Wallis Test

b. Grouping Variable: Összevont karkod

**44. táblázat. Kruskal-Wallis-próba az Összevont karkod és a Keresési készség változókra**

A Kruskal-Wallis-próba szignifikanciaszintje 0,004, amely  $H_0$  hipotézisünk elvetését jelenti. A próba rangtáblázatából leolvasható, hogy mely karokon a legmagasabb a *Keresési készség* értéke: Az első három helyezett az Informatikai Kar, a Közgazdaságtudományi Kar és a Műszaki Főiskolai Kar.

**Test of Homogeneity of Variances**

Keresési készség

| Levene<br>Statistic | df1 | df2 | Sig. |
|---------------------|-----|-----|------|
| 2,398               | 9   | 190 | ,014 |

**ANOVA**

Keresési készség

|                | Sum of<br>Squares | df  | Mean Square | F     | Sig. |
|----------------|-------------------|-----|-------------|-------|------|
| Between Groups | 27,006            | 9   | 3,001       | 3,315 | ,001 |
| Within Groups  | 171,994           | 190 | ,905        |       |      |
| Total          | 199,000           | 199 |             |       |      |

**45. táblázat. Egyszempontos varianciaanalízis az Összevont karkod és a Keresési készség változókra**

A Kruskal-Wallis-próbának megfelelő eredményt kapunk a robosztus Welch, illetve Brown-Forsythe-próbák alkalmazásával is.  $H_1$  hipotézisünket kell tehát elfogadnunk, azaz a keresési készség karonként nem homogén. A Post Hoc-próba alkalmazásával megnézhetjük azt is, mely karok különböznek szignifikánsan egymástól:

# Multiple Comparisons

Dependent Variable: Keresési készség  
Games-Howell

| (I) Összevont karkod | (J) Összevont karkod | Mean Difference (I-J) | Std. Error | Sig.  | 95% Confidence Interval |             |
|----------------------|----------------------|-----------------------|------------|-------|-------------------------|-------------|
|                      |                      |                       |            |       | Lower Bound             | Upper Bound |
| 1,00                 | 2,00                 | -,45332236            | ,24251082  | ,688  | -,12663251              | ,3596804    |
|                      | 3,00                 | ,01569418             | ,21375055  | 1,000 | -,7172574               | ,7486458    |
|                      | 4,00                 | -,25951248            | ,20006221  | ,950  | -,9210978               | ,4020729    |
|                      | 6,00                 | -,40378859            | ,28090663  | ,902  | -,13839500              | ,5763728    |
|                      | 10,00                | -,17082274*           | ,41823906  | ,032  | -,32983479              | -,1181069   |
|                      | 11,00                | -,67000916            | ,35374059  | ,674  | -,20205496              | ,8805313    |
|                      | 12,00                | -,58595892            | ,23643988  | ,318  | -,13898913              | ,2179734    |
|                      | 13,00                | -,24142598            | ,27582631  | ,996  | -,12086443              | ,7257924    |
|                      | 14,00                | -,41963540            | ,22775190  | ,706  | -,11729561              | ,3336853    |
| 2,00                 | 1,00                 | ,45332236             | ,24251082  | ,688  | -,3596804               | 1,2663251   |
|                      | 3,00                 | ,46901654             | ,26479148  | ,748  | -,4204128               | 1,3584459   |
|                      | 4,00                 | ,19380989             | ,25387028  | ,999  | -,6483920               | 1,0360118   |
|                      | 6,00                 | ,04953377             | ,32145570  | 1,000 | -,10401416              | 1,1392092   |
|                      | 10,00                | -,12549050            | ,44648532  | ,214  | -,28877450              | ,3779350    |
|                      | 11,00                | -,21668680            | ,38672682  | 1,000 | -,16205467              | 1,1871732   |
|                      | 12,00                | -,13263656            | ,28342380  | 1,000 | -,10795321              | ,8142590    |
|                      | 13,00                | ,21189638             | ,31702584  | 1,000 | -,8654888               | 1,2892816   |
|                      | 14,00                | ,03368696             | ,27621760  | 1,000 | -,8779026               | ,9452765    |
| 3,00                 | 1,00                 | -,01569418            | ,21375055  | 1,000 | -,7486458               | ,7172574    |
|                      | 2,00                 | -,46901654            | ,26479148  | ,748  | -,13584459              | ,4204128    |
|                      | 4,00                 | -,27520665            | ,22655665  | ,965  | -,10382482              | ,4878349    |
|                      | 6,00                 | -,41948277            | ,30035240  | ,918  | -,14555999              | ,6166344    |
|                      | 10,00                | -,17239216*           | ,43154020  | ,032  | -,33340703              | -,1137728   |
|                      | 11,00                | -,68570334            | ,36937167  | ,695  | -,20615890              | ,8901823    |
|                      | 12,00                | -,58595892            | ,23643988  | ,318  | -,13898913              | ,2179734    |

|       |       |             |           |       |            |           |
|-------|-------|-------------|-----------|-------|------------|-----------|
| 6,00  | 14,00 | -,16012292  | ,23981149 | 1,000 | -,9459745  | ,6257287  |
|       | 1,00  | ,40378859   | ,28090663 | ,902  | -,5763728  | 1,3839500 |
|       | 2,00  | -,04953377  | ,32145570 | 1,000 | -,1392092  | 1,0401416 |
|       | 3,00  | ,41948277   | ,30035240 | ,918  | -,6166344  | 1,4555999 |
|       | 4,00  | ,14427611   | ,29076993 | 1,000 | -,8571584  | 1,1457106 |
|       | 10,00 | -,13044388  | ,46845083 | ,215  | -,29892066 | ,3803290  |
|       | 11,00 | -,26622057  | ,41189158 | 1,000 | -,17372212 | 1,2047801 |
|       | 12,00 | -,18217033  | ,31690075 | 1,000 | -,12634154 | ,8990748  |
|       | 13,00 | ,16236261   | ,34727859 | 1,000 | -,10253435 | 1,3500687 |
| 10,00 | 14,00 | -,01584681  | ,31047254 | 1,000 | -,10699752 | 1,0382816 |
|       | 1,00  | 1,70822740* | ,41823906 | ,032  | -,1181069  | 3,2983479 |
|       | 2,00  | 1,25490503  | ,44648532 | ,214  | -,3779350  | 2,8877450 |
|       | 3,00  | 1,72392157* | ,43154020 | ,032  | -,1137728  | 3,3340703 |
|       | 4,00  | 1,44871492  | ,42492650 | ,089  | -,1486586  | 3,0460884 |
|       | 6,00  | 1,30443881  | ,46845083 | ,215  | -,3803290  | 2,9892066 |
|       | 11,00 | 1,03821824  | ,51543191 | ,600  | -,8017032  | 2,8781397 |
|       | 12,00 | 1,12226848  | ,44321718 | ,325  | -,5066273  | 2,7511643 |
|       | 13,00 | 1,46680142  | ,46542217 | ,116  | -,2121833  | 3,1457862 |
| 11,00 | 14,00 | 1,28859200  | ,43864402 | ,179  | -,3291096  | 2,9062936 |
|       | 1,00  | ,67000916   | ,35374059 | ,674  | -,6805313  | 2,0205496 |
|       | 2,00  | ,21668680   | ,38672682 | 1,000 | -,1871732  | 1,6205467 |
|       | 3,00  | ,68570334   | ,36937167 | ,695  | -,6901823  | 2,0615890 |
|       | 4,00  | ,41049668   | ,36162276 | ,971  | -,9483145  | 1,7693078 |
|       | 6,00  | ,26622057   | ,41189158 | 1,000 | -,12047801 | 1,7372212 |

46. táblázat. Post Hoc teszt az Összevont karkod és a Keresési készség változókra

A táblázatból leolvasható, hogy a keresési készség vonatkozásában az összevont ÁJK-DK-HPFK hármas és az Informatikai Kar, valamint az Agrárgazdasági és Vidékfejlesztési Kar és az Informatikai Kar között van szignifikáns különbség.

# Információkeresés a weben és tanítása

Értekezés a doktori (Ph.D.) fokozat megszerzése érdekében az informatika  
tudományágban

Írta: Krauszné Princz Mária okleveles matematikus és matematika szakos  
középiskolai tanár

Készült a Debreceni Egyetem TTK Matematika Doktori Iskola Informatika  
programja keretében

Témavezető: Dr. Fazekas Gábor

A doktori szigorlati bizottság:

Elnök: Dr. ....

Tagok: Dr. ....

Dr. ....

A doktori szigorlat időpontja: 200.....

Az értekezés bírálói:

Dr. ....

Dr. ....

Dr. ....

A bíráló bizottság:

Elnök: Dr. ....

Tagok: Dr. ....

Dr. ....

Dr. ....

Dr. ....

Az értekezés védésének időpontja: 200 .....