



Web-es keresők és kiértékelése

7. ELŐADÁS

1. Webes keresés relevanciahatékonyságának mérése
2. Leighton-módszer

1. Webes keresés relevanciahatékonyságának mérése

1.1. Az információ-visszakereső rendszerek minősítése

1.1.1. A webes információ-visszakereső rendszerek vizsgálata -1. lépés

1.1.2. A webes információ-visszakereső rendszerek vizsgálata -2. lépés

1.1.3. A webes információ-visszakereső rendszerek vizsgálata -3. lépés

1.1.4. A webes információ-visszakereső rendszerek vizsgálata -4. lépés

1.1.5. A webes információ-visszakereső rendszerek vizsgálata -5. lépés

1.2. Webes visszakereső rendszer relevanciahatékonysága mérésének sajátosságai

1.1. Az információ-visszakereső rendszerek minősítése

- Az információ-visszakereső rendszereket, mint bármely szoftver rendszert minősíteni szokás.
- Az információ-visszakereső rendszereknél elsősorban azt vizsgáljuk, hogy milyen precíz a találati lista, azaz mennyire relevánsak a válaszok.
- A találati lista a keresőkérdésre a kereső által visszaadott linkeket (oldalakat) tartalmazza.
- Ezt a vizsgálati módszert visszakeresési teljesítmény kiértékelésnek (becslésnek) hívjuk.

1.1.1. A webes visszakereső rendszerek vizsgálata -1. lépés

A vizsgálat első lépése:

egy elfogulatlan kérdéssort kell felállítani a rendszer teszteléséhez. A tesztkérdések kiválasztása két lépésből áll:

1. Ki kell jelölni a témát, amelyet keresni kívánunk,
2. pontosan meg kell határozni a keresőkérdéseket.

Az elfogult értékelés elkerülésére általános és speciális kérdéseket is választani kell.

1.1.2. A webes visszakereső rendszerek vizsgálata -2.lépés

A vizsgálat 2. lépése:

A tesztelés végrehajtása előtt kell meghatározni, hogy a kérdésekre kapott válaszokat hogyan fogjuk minősíteni.

- Fontos, hogy a válaszok megismerése előtt alakítsuk ki az értékelési módszert.
- Ellenkező esetben elfogultabban, szubjektívebben értékeljük a keresőt, hiszen ismerjük erős és gyenge pontjait. Megtudjuk, melyik kérdésre ad több választ, mennyire pontos válaszokat ad, stb.

1.1.2. A webes visszakereső rendszerek vizsgálata -2.lépés

Nem minősíthetünk az alapján, hogy egy kérdésre hány választ adnak vissza, mert:

- A válaszok közt általában inaktív,
- és a felhasználók számára felhasználhatatlan oldalak is vannak.

A felhasználók által megfogalmazott információ igény sokszor határozatlan, így a visszakeresett dokumentumok nem lehetnek pontos válaszok.

Másrészt különböző felhasználóknak más a megítélésük, hogy melyik válasz releváns és melyik nem.

1.1.3. A webes visszakereső rendszerek vizsgálata -3.lépés

A kiértékelés előtt meg kell határozni, hogy a keresőkérdésekre kapott válaszokat mikor tekintjük relevánsnak.

A relevancia két entitáshoz kapcsolódik:

- az egyik az információforrás (maga a dokumentum, vagy az azt helyettesítő összegzés) valamilyen formája,
- a másik az információigény (ami magában foglalja a felhasználó szavakkal megfogalmazott igényét, és az információ-visszakereső rendszerrel közölt keresőkérdést) formája

1.1.4. A webes visszakereső rendszerek vizsgálata -4. lépés

A vizsgálat 4. lépése:

A találati oldalakat kategóriákba soroljuk.

- A kategóriák meghatározzák, hogy egy választ mikor tekintek relevánsnak.
- Kétféle kategória van:
 - 1. kategória
 - 0. kategória

Kategóriák meghatározása

1. kategória: Releváns oldalak (linkek). Nehéz eldönteni, hogy melyik oldalt tekintjük relevánsnak, hiszen a különböző felhasználóknak más a megítélésük, hogy melyik válasz releváns és melyik nem. Általában akkor tekintjük relevánsnak a találatot, ha

- az oldal címe tartalmazza a keresett kifejezést,
- az oldalon szerepel a keresett kifejezés vagy
- kérdés által meghatározott tématerülethez tartozik az oldal.

Kategóriák meghatározása

o. kategória: Az oldal irreleváns, semmilyen szempontból nem tesz eleget a kereső kérdésnek, vagyis nem sorolható az 1.kategóriába.

- dupla link: A keresőkérésre visszakapott találati listában többször is előfordul ugyanaz az oldal.
- inaktív link: Egy linkre kattintva különböző hibaüzeneteket kaphatunk (pl. File not found, Forbidden, Server not responding.)

Kategóriák meghatározása

- A kategóriák által az oldalak besorolása egyértelmű, minden egyes oldal egyik kategóriába tartozik
- Természetesen több relevancia kategóriát is választhatunk. A releváns oldalakat különböző szempontok szerint további kategóriákba sorolhatjuk.

1.1.5. A webes visszakereső rendszerek vizsgálata -5. lépés

A vizsgálat 5. lépése:

A kategóriák meghatározása után el kell dönteni, hogy a kereső által visszaadott találati listából hány dokumentumot soroljunk be a kategóriákba.

- Ez is fontos kérdés, hiszen Webes keresőknél a visszakapott találatok száma tízezres nagyságrendű is lehet.
- Gyakorlatilag lehetetlen több ezer linket elolvasni, és megvizsgálni, hogy releváns-e a kérdésre vagy sem.
- A létező tanulmányok általában az első 10 vagy 20 választ veszik figyelembe.

1.2. Webes visszakereső rendszer relevanciahatékonysága mérésének sajátosságai

A **visszakereső módszer** relevancia-hatékonyságának laboratóriumi mérése alapmértékeken (pontosság, teljesség) nyugszik.

A világhálón élesben működő **webes visszakereső rendszer** relevanciahatékonysága mérésének sajátosságai vannak, mégpedig a világháló sajátosságaiból adódóan.

1.2. Webes visszakereső rendszer relevanciahatékonysága mérésének sajátosságai

Magának a Világhálónak a jellemzői:

- a Web-oldalak sokféle formátuma (HTML stb.),
- a Web-oldalak „laza” szerkezete (nem olyan értelemben precízen definiáltak, mint pl. egy relációs adatbázis esetében a táblázatok),
- a Web-oldalak dinamikusak (megjelennek, eltűnnek, módosulnak),
- a Web-oldalak hálózatot alkothatnak (linkek révén, Web-gráf),
- a Web-felhasználók száma óriási, a felhasználók különböznek egymástól beszélt nyelv, érdeklődési terület stb. szempontjából,
- a Web-oldalak száma exponenciálisan növekszik.

1.2. Webes visszakereső rendszer relevanciahatékonysága mérésének sajátosságai

Élesben működő visszakereső rendszer esetében:

- A keresőkérdés előre nem ismert.
- Nem ismerjük a web tartalmát, ezért nem tudjuk, hogy egy kérdésre hány releváns válasz van ezért a felidézés nem számítható ki, relevancialista nem adható meg.
- Ha nagyon sok választ kapunk a kérdésre, akkor a pontosságot sem tudjuk meghatározni. Gyakorlatilag lehetetlen az összes linket elolvasni, és megvizsgálni, hogy melyik releváns és melyik nem.

2. Leighton módszer

- 2.1. Leighton módszer tulajdonságai
- 2.2. Csoportok megadása
- 2.3. Súlyszámok megadása
- 2.4. Mértékek megadása

2. Leighton módszer

Web-es keresők relevancia hatékonyságának kiértékelésére nagyon jól használható a Leighton-módszer .

- A módszer a mérőszám meghatározásához csak az első 5 vagy 10 (a kísérlet szempontjait figyelembe véve tetszőlegesen megválasztható számú) találatot veszi figyelembe.

2. Leighton módszer

Leighton módszere:

- olyan mérőszámot használ, amely a pontosságot méri a rendezési hatékonysággal súlyozva.
- A két mérőszámot egy értékben egyesíti.
- A mérőszám értéke egy 0 és 1 közé eső szám.
- 0 a mérőszám értéke, ha nem kapunk választ a kérdésre, vagy az összes válasz irreleváns.
- 1 a mérőszám értéke, ha az összes válasz releváns, és megfelelő számú választ kaptunk vissza.

2.1. Leighton módszer tulajdonságai

A módszer fontos tulajdonságai:

- Minden kérdés esetén megvizsgáljuk a kereső által visszaadott első 5 (10) választ, és a vizsgált kritériumnak megfelelően a fent említett kategóriákba soroljuk őket.
- A relevanciát bináris skálán mérjük: ha a linket az 1. kategóriába soroltuk, akkor 1 pontot, ha a 0. kategóriába, akkor 0 pontot ér a válasz. A dupla linkeket a mérőszám meghatározásánál vesszük figyelembe.

2.1. Leighton módszer tulajdonságai

- Figyelembe vesszük a releváns linkek rangsorbeli elhelyezkedését. Minél előrébb helyezkednek el a releváns linkek, annál nagyobb lesz a mérőszám értéke.
- A mérőszám tükrözi, ha a kereső kevesebb találatot ad vissza ugyanannyi jó találat esetén, vagyis magasabb a precizitása. Így persze könnyebb megtalálni a releváns linket. (Ez persze nem jelentheti azt, hogy csak nagyon kevés találatot adjon vissza egy jó kereső, netán egyet se.)

2.2. Csoportok megadása

- A rangsor figyelembe vételére a pontosság kiszámításakor súlytényezőket használunk.
- A visszakapott találatokat csoportokba soroljuk.
- A csoportokhoz súlyszámokat rendelünk.
- Az azonos csoportban található linkek azonos súlyozással szerepelnek a mérőszám értékének kiszámításakor.

A csoportok kialakítása és a súlytényezők megválasztása természetesen többféleképpen történhet.

2.2. Csoportok megadása

Csoportok 5, illetve 10 pontosság esetében:

Az **5-pontosság** esetében az első öt találatot osztjuk fel 2 csoportra a következő módon:

- 1. csoport: az első 2 találat.
- 2. csoport: a következő 3 találat.

2.2. Csoportok megadása

A **10-pontosság** esetében az első tíz találatot osztjuk fel 3 csoportra a következő módon:

- 1. csoport: az első 2 találat. .
- 2. csoport: a következő 3 találat.
- 3. csoport: a következő 5 találat.

A felhasználó számára fontos, hogy a találati lista elején legyenek a releváns találatok. Feltételezhető, hogy a felhasználó számára általában az első 5 link a legfontosabb, közülük kiemelten fontos az első kettő.

2.3. Súlyszámok megadása

- Az ugyanabban a csoportban lévő linkek ugyanolyan súlyozást kapnak.
- Az 5-pontosság esetében a következő módon súlyozunk:
 - Az első csoport, vagyis az első 2 link súlya: 10.
 - A második csoport, vagyis a következő 3 link súlya: 5.
- A súlyszámok alapján elégedettebbek vagyunk, ha rögtön a kezdő képernyőn megkapjuk a releváns linkeket. Sok felhasználó nem vizsgálja meg a további találati oldalakat.

2.3. Súlyszámok megadása

- A 10-pontosság esetében a következő módon súlyozunk:
 - Az első csoport, vagyis az első 2 link súlya: 20.
 - A második csoport, vagyis a következő 3 link súlya: 17.
 - A harmadik csoport, vagyis a további 5 link súlya 10.

2.3. Súlyszámok megadása

- A kitartó felhasználó ugyan több linket megvizsgál, de számára is fontos, hogy a találati lista elején legyenek a releváns linkek.
- Azt feltételezzük, hogy a kitartó felhasználó számára az első 5 link a legfontosabb. Közülük is kiemelten fontosak az első találati oldalon található linkek. Ezt a szemléletet tükrözi a csoportok kialakítása, és a súlyszámok meghatározása.

Steps of Leighton method

1. Select search engine to be measured.
2. Define relevance categories.
3. Define groups.
4. Define weights.
5. Give queries q_i ($i = 1, \dots, s$).
6. Compute P_{5_i} and/or P_{10_i} for q_i ($i=1, \dots, s$).
7. The first 5/10-precision of the search engine is:

$$P_k = \frac{1}{s} \sum_{i=1}^s P_{k_i} \text{ where } k = 5 \text{ or } k = 10.$$

Leighton módszer mértékei

- A csoportok és súlyszámok ismeretében határozzuk meg minden egyes kérdés esetén a mérőszámok értékét.
- Az **5-pontosság** mértékét a következő módon számítjuk ki. A számláló az első öt válasz súlyozott összege:

$$\frac{(Pontok_{1.-2.link} \cdot 10) + (Pontok_{3.-5.link} \cdot 5)}{35 - [(5 - találatok_száma_{első 5!}) \cdot 5]}$$

Az 5-pontosság esetében a következő módon súlyozunk:

- Az első csoport, vagyis az első 2 link súlya: 10.
- A második csoport, vagyis a következő 3 link súlya: 5.

Leighton módszer mértékei

- 10 pontosság mértéke:

$$\frac{(Pontok_{1.-2.link} \cdot 20) + (Pontok_{3.-5.link} \cdot 17) + (Pontok_{6.-10.link} \cdot 10)}{141 - [(10 - találatok_száma_{első 10!}) \cdot 10]}$$

A 10-pontosság esetében a következő módon súlyozunk:

- Az első csoport, vagyis az első 2 link súlya: 20.
- A második csoport, vagyis a következő 3 link súlya: 17.
- A harmadik csoport, vagyis a további 5 link súlya 10.

Rövid példa 1.

Egy adott q keresőkérdésre 3 választ kaptunk, amelyek mind relevánsak.

- A számláló (csak a relevanciát nézi) :
 - $(2 \times 10) + (1 \times 5) = 25.$
 - Az első két találat releváns, az első csoportba tartozik, tehát a súlya:10
 - A harmadik találat a 2. csoportba tartozik, tehát a súlya: 5.
- A nevező (nem vesszük figyelembe, hogy a válasz releváns-e vagy sem, csak a válaszok számát) :

$$35 - (2 \times 5) = 25.$$

$$P_5 = 25 : 25 = 1.$$

Rövid példa 2.

Egy adott kérdésre 5 találatot kapunk:

- amelyek közül: a 2., 3., és a 4. releváns.

A számláló:

$$(1 \times 10) + (2 \times 5) = 20,$$

És így a $P_5 = 20 : 35 = 0.571$.

- Ha az első 3 találat lenne releváns:

$$\text{akkor } P_5 = ((2 \times 10) + (1 \times 5)) : 35 = 0.714.$$

Példa leighton módszerre –P10

Query	No.of hits	Relevant hits	Irrelevant hits	π
1.	16	1-4	5-16	
2.	18	-	1-18	
3.	27	1-6	7-27	
4.	16	1-2	3-16	
5.	18	1-2	3-18	
6.	25	1-10	11-20	
7.	33	1-11	12-30	
8.	20	1-10	11-20	
9.	20	0	1-20	
10.	28	1-12	13-28	
Average				

Példa leighton módszerre –P10

Query	No.of hits	Relevant hits	Irrelevant hits	π
1.	16	1-4	5-16	0.525
2.	18	-	1-18	0
3.	27	1-6	7-27	0.716
4.	16	1-2	3-16	0.284
5.	18	1-2	3-18	0.284
6.	25	1-10	11-20	1
7.	33	1-11	12-30	1
8.	20	1-10	11-20	1
9.	20	0	1-20	0
10.	28	1-12	13-28	1
Average				0.58