Ződi Zsolt

A relevancia rejtélye

SZIGORÚAN ELLENŐRZÖTT KAPUŐRÖK

AZ INTERNETES FORGALOMIRÁNYÍTÓ SZOLGÁLTATÁSOK SZABÁLYOZÁSI

KÉRDÉSEI

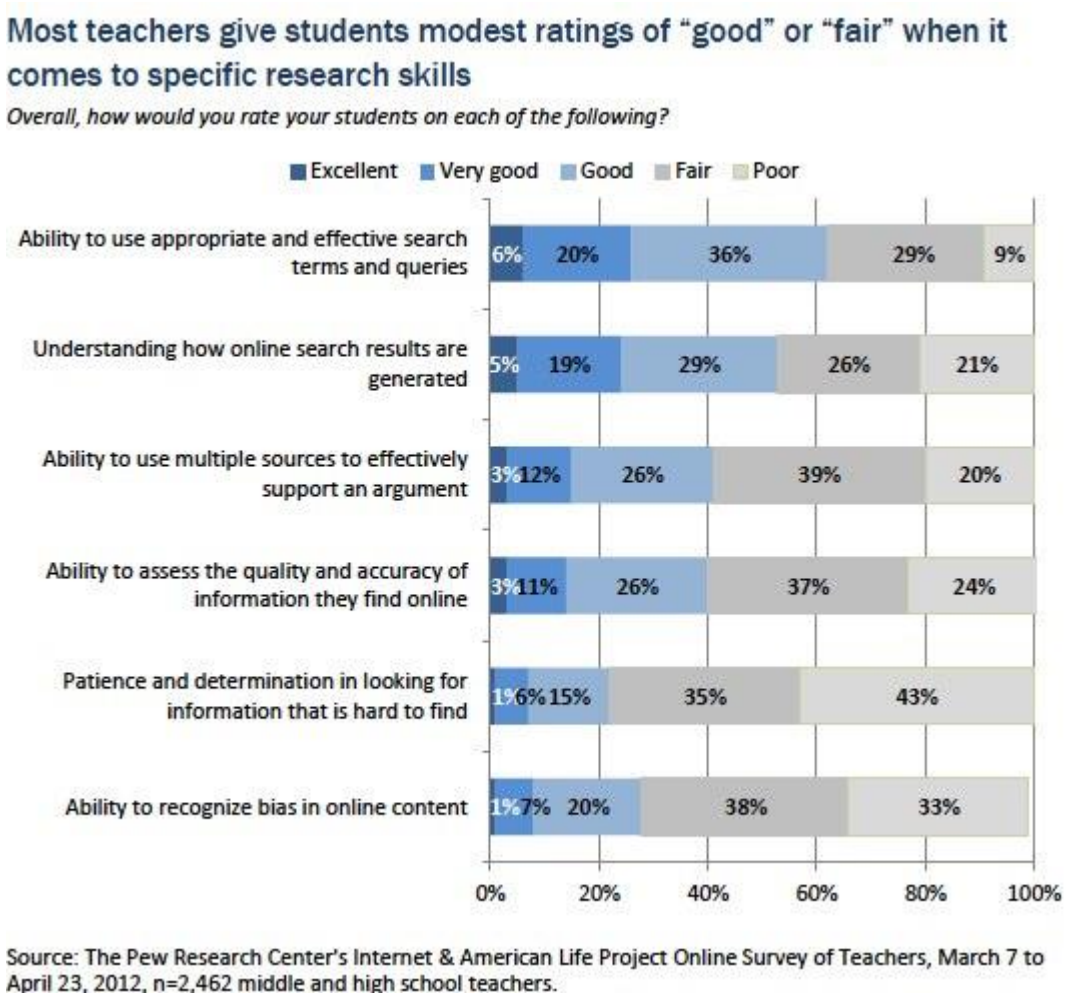
2015. október 16.

MTA Társadalomtudományi Kutatóközpont
Jogtudományi Intézet

mtatkjti

A keresőszolgáltatások jelentősége

- A keresőszolgáltatások mára megkerülhetetlenek – ami nem szerepel a „kedvencek" között azt kizárólag ezeken keresztül érjük el.
- Nemcsak „kutatunk", hanem problémákat is ezzel oldunk meg, és egyszerű információkat is ezzel keresünk.
- Az internetes keresőszolgáltatások saját algoritmusaik segítségével rangsorolják az internetes tartalmakat.
- A rangsorolás az általuk elképzelt „fontosság", „odaillőség" vagy „relevancia" szerint történik.
- A relevancia a kérdés és a válasz egymáshoz képesti távolságát méri.
- „Nem releváns" valami, ha nincsen a kérdés és a válasz között összefüggés, míg „nagyon releváns" ha az összefüggés szoros.



Source: The Pew Research Center's Internet & American Life Project Online Survey of Teachers, March 7 to April 23, 2012, n=2,462 middle and high school teachers.

mtatkjti

A keresés története

- A keresés nagyobb információ-halmazokban legelőször **szurrogátumokkal** (helyettesítőkkel) történt – pl. katalógus-cédula, tárgyszójegyzék, tartalomjegyzék, taxonómia – a teljes szöveg helyett annak „kivonata”. (Ennek utóda a metaadat, az adatról szóló adat.)
- A számítógép feltalálása után megjelent a **teljes szövegű index.** Minden szót felleltároztak, (géppel) és egy hatalmas listába rendezték.
- A fentebbi módszerek nem mondanak semmit **a találatok közötti fontossági sorrendről.** Erre újabb módszerek alakultak ki.
  - Szógyakoriság mérése (ha egy szó gyakrabban fordul elő egy szövegben, akkor az inkább arról szól).
  - A szavak helyének súlyozása (ha egy szó a címben van, akkor az fontosabb).
  - A dokumentumok fontosságának objektív mérése a rá **hivatkozó** dokumentumok számával. (Shepard - 1875, SCI - 1961)



mtatkjti

A „fontosság” új fogalma a Google-nél

„The citation (link) graph of the web is an important resource that has largely gone unused in existing web search engines. We have created maps containing as many as 518 million of these hyperlinks, a significant sample of the total. These maps allow **rapid calculation of a web page's "PageRank", an objective measure of its citation importance that corresponds well with people's subjective idea of importance.** Because of this correspondence, PageRank is an excellent way to prioritize the results of web keyword searches.” (Sergey Brin and Lawrence Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine, http://infolab.stanford.edu/~backrub/google.html







A relevancia több, mint objektív (statikus) fontosság

A relevancia fogalma jóval a keresők előtt is létezett más területeken

- Nyelvészet, információelmélet:

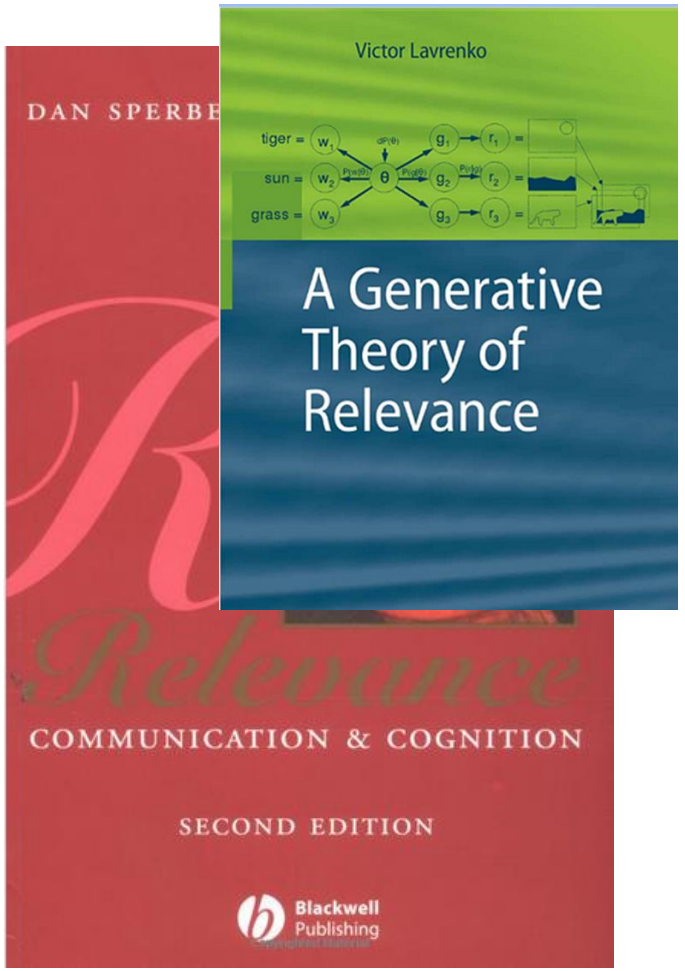
  *„the search for relevance is a basic feature of human cognition (…) utterances **automatically create expectations which guide the hearer towards the speaker's meaning**, (…) an input (a sight, a sound, an utterance, a memory) is relevant to an individual when it connects **with background information** he has available to yield conclusions that matter to him" (Sperber-Wilson)*

- Számítástudomány (Lavrenko):

  *„a binary relation between a given information item (document D) and the **user's request** (query Q)."*

- Jog

  „Evidence is relevant if:
  (a) it has any tendency to make a fact more or less probable than it would be without the evidence; and
  (b) the fact is of consequence in determining the action." (Federal Rules of Evidence, Rule 401.)



A relevancia nemcsak az információról szól, hanem az információ és a kérdező viszonyáról is

mtatkjti

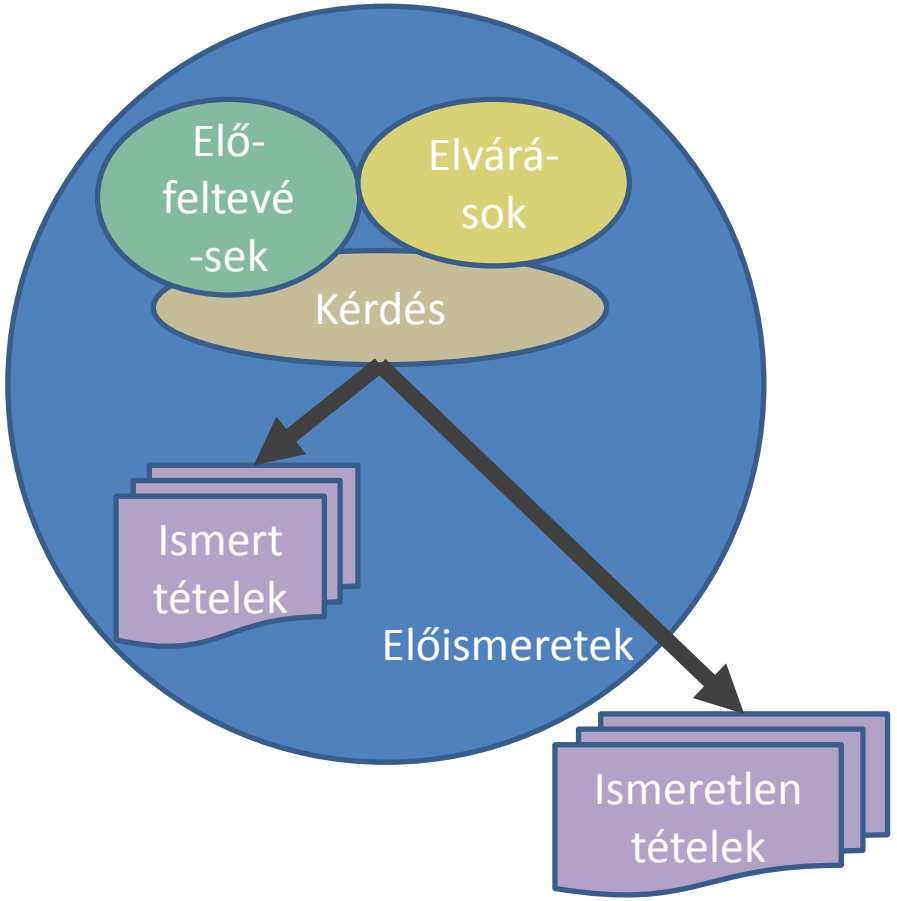
Mire gondol a kérdező? – a dinamikus relevancia

A relevancia tehát a beszélgető, információ-kereső, probléma megoldó **előfeltevéseinek, elvárásainak, információ-igényének** egy információ-egységgel való megfelelőségét, egybevágását méri, fejezi ki.

- A relevancia bármilyen külvilágban létező „tétel” (item) távolságát jelenti a belső **elvárásainkkal** és **előfeltevéseinkkel**.
- Csak arról a témáról tudunk kérdést feltenni, amiről valamilyen **előismereteink** már vannak.
- A kérdéseinket az **előfeltevéseink** tükrében tesszük fel.
- A válasszal kapcsolatban vannak bizonyos **elvárásaink**.

Pl. Mennyi a minimálbér összege 2015-ben?

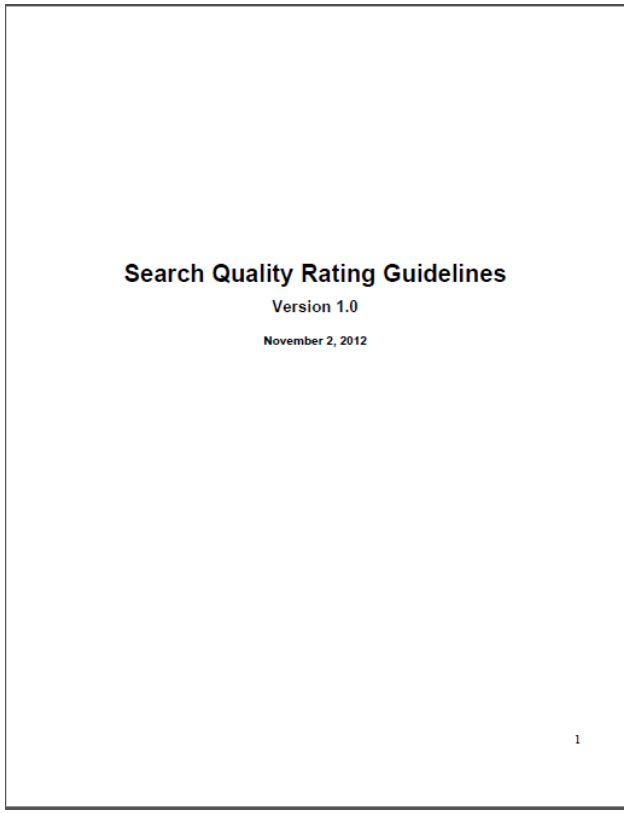
- előismeret: minden ami a „világunkban” tudott, (a problémára vonatkozó és nem vonatkozó ismeretek, pl. jogi ismeretek)
- előfeltevések: a konkrét problémára vonatkozó előzetes tudás (a problémára vonatkozó szűkebb tudáshalmaz pl. évente változik)
- elvárások: pontosan tudni szeretném mennyi a minimálbér összege
- kérdés: az a kérdés, amelyet megfogalmazunk „Mennyi a minimálbér összege?” vagy „minimálbér 2015”



mtatkjti

Hogyan próbálja a Google kitalálni, mire gondolunk #1 – user intent*

- **Action intent** – Users want to accomplish a goal or engage in an activity, such as download software, play a game online, send flowers, find entertaining videos, etc. These are "do" queries: users want to do something.
- **Information intent** – Users want to find information. These are "know" queries: users want to know something.
- **Navigation intent** – Users want to navigate to a website or webpage. These are "go" queries: users want to go to a specific page.

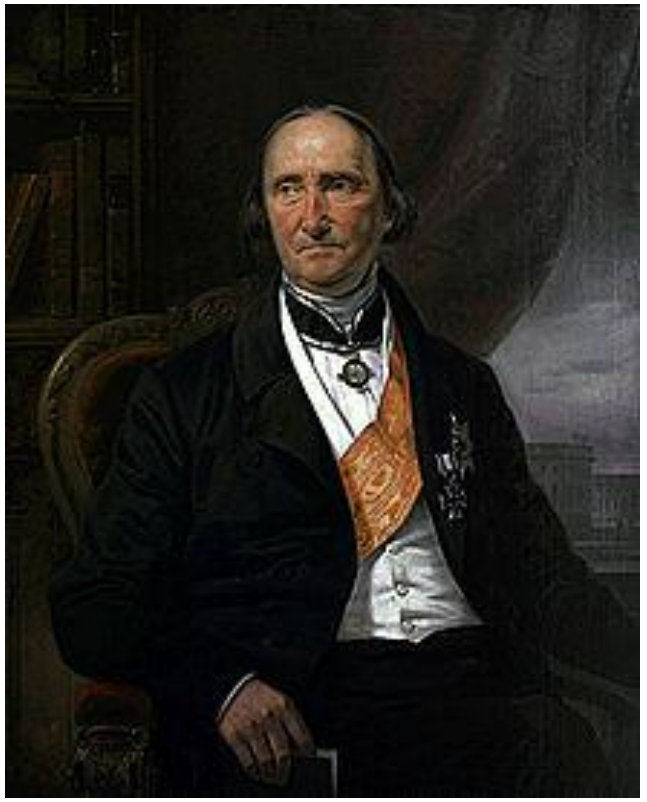

Search Quality Rating Guidelines
Version 1.0
November 2, 2012
1

* A Google „Search Quality Rating Guidelines Version 1.0 November 2, 2012" alapján - https://static.googleusercontent.com/media/www.google.com/hu//insidesearch/howsearchworks/assets/searchqualityevaluatorguidelines.pdf

mtatkjti

Hogyan próbálja a Google kitalálni, mire gondolunk #2 – query interpretations*

- **Dominant Interpretation**: The dominant interpretation of a query is the interpretation that most users have in mind when they issue the query. For example, most users typing [windows], English (US) want results on the Microsoft operating system, rather than the glass windows on a wall. The dominant interpretation should be clear to you, especially after doing a little web research.
- **Common Interpretations**: In some cases, there is no dominant interpretation. The query [mercury], English (US) might refer to the car brand, the planet, or the chemical element (Hg). While none of these is clearly dominant, all are common interpretations. Many or some people might want results related to these interpretations.
- **Minor Interpretations:** Sometimes you will find less common interpretations. These are interpretations that few users have in mind. We will call these minor interpretations. Consider again the query [mercury], English (US). Possible meanings exist that even most English (US) users probably do not know about, such as Mercury Marine Insurance and the San Jose Mercury News. These are minor interpretations.

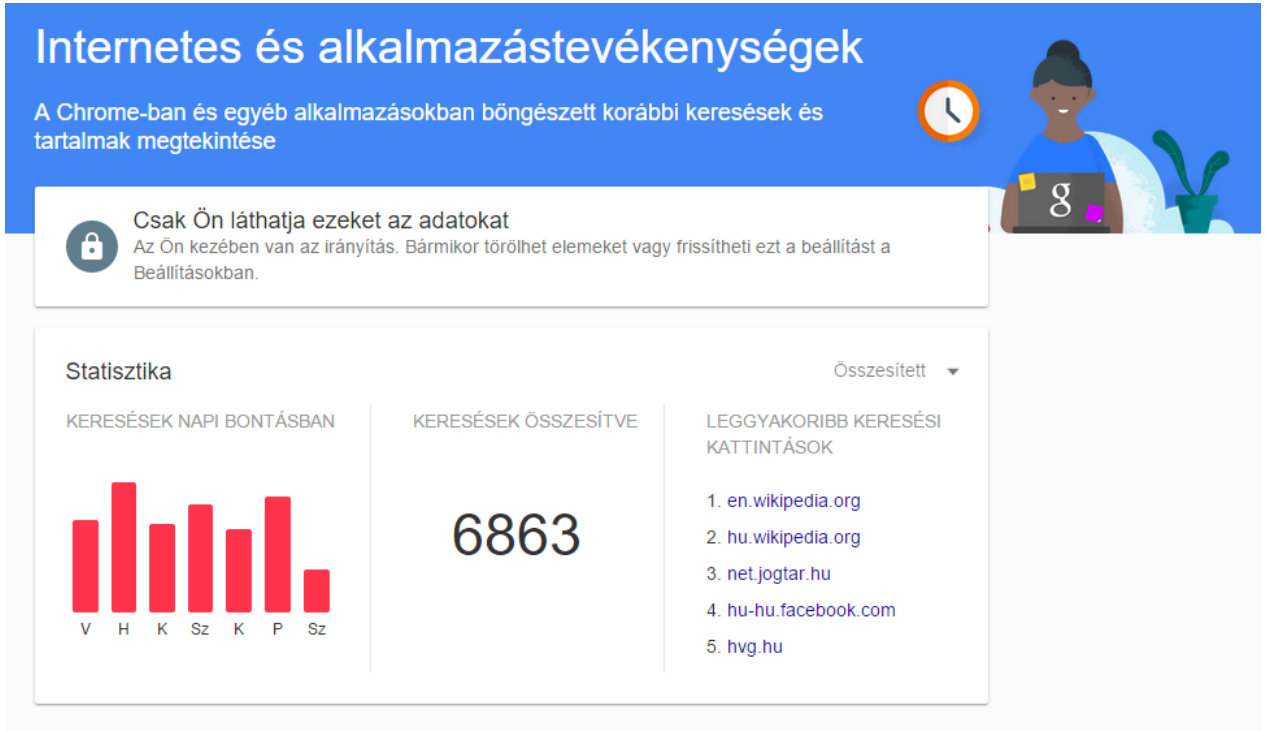


* A Google „Search Quality Rating Guidelines Version 1.0 November 2, 2012” alapján - https://static.googleusercontent.com/media/www.google.com/hu//insidesearch/howsearchworks/assets/searchqualityevaluatorguidelines.pdf

mtatkjti

Hogyan próbálja a Google kitalálni, mire gondolunk #3 – geographical location and past searches*

- A földrajzi hely és a korábbi keresések szintén felülsúlyozzák a találatok között azokat az oldalakat, amelyekre már korábban kerestem
- Mindkettő kikapcsolható ugyan, de én alig tudok olyan emberről, aki ezt valaha kikapcsolta volna.



https://www.google.com/insidesearch/howsearchworks/algorithms.html

mtatkjti

A relevancia fokozatai*

Vital - The Vital rating is used for these very special situations:
1) The dominant interpretation of the query is navigation, and the landing page is the target of the navigation query.
2) The dominant interpretation of the query is an entity (such as a person, place, business, restaurant, product, company, organization, etc.), and the landing page is the official webpage associated with that entity.

Useful - A rating of Useful is assigned to pages that are very helpful for most users.

Relevant - A rating of Relevant is assigned to pages that are helpful for many or some users. Relevant pages have fewer valuable attributes than were listed for Useful pages. Relevant pages should still "fit" the query, but they might be less comprehensive, less up-to-date, come from a less authoritative source, or cover only one important aspect of the query.

Slightly Relevant - A rating of Slightly Relevant is assigned to pages that are not very helpful for most users, but are somewhat related to the query.

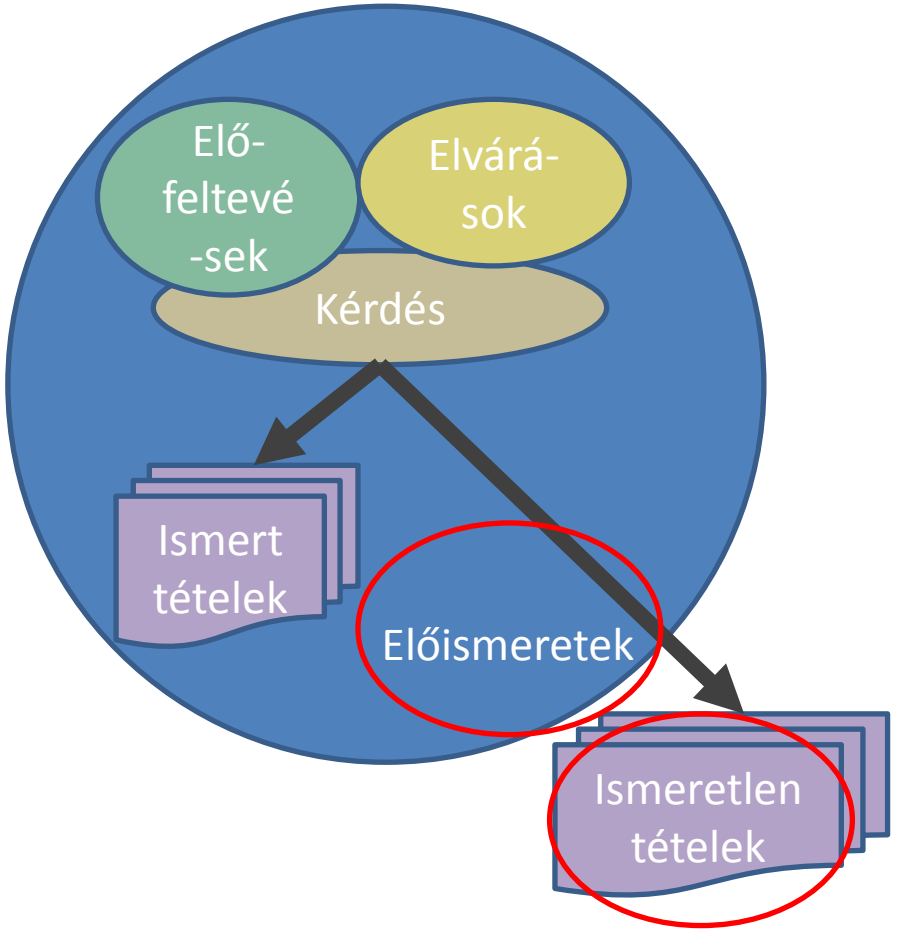
Off-Topic - A rating of Off-Topic or Useless should be assigned to pages that are helpful for very few or no users.

Useless -

mtatkjti

Vissza a relevanciához

- A relevancia statikus fogalmától eljutottunk a relevancia dinamikus fogalmáig.
- Már figyelembe van véve a „keresési szándék”, de még csak néhány egyszerű paraméterrel
- A kereső nem tud visszakérdezni…
- De három probléma van:
  - Az előismeretek
  - Az ismeretlen tételek
  - A „know” típusú keresés kétféle jellege



mtatkjti

Az előismeretek

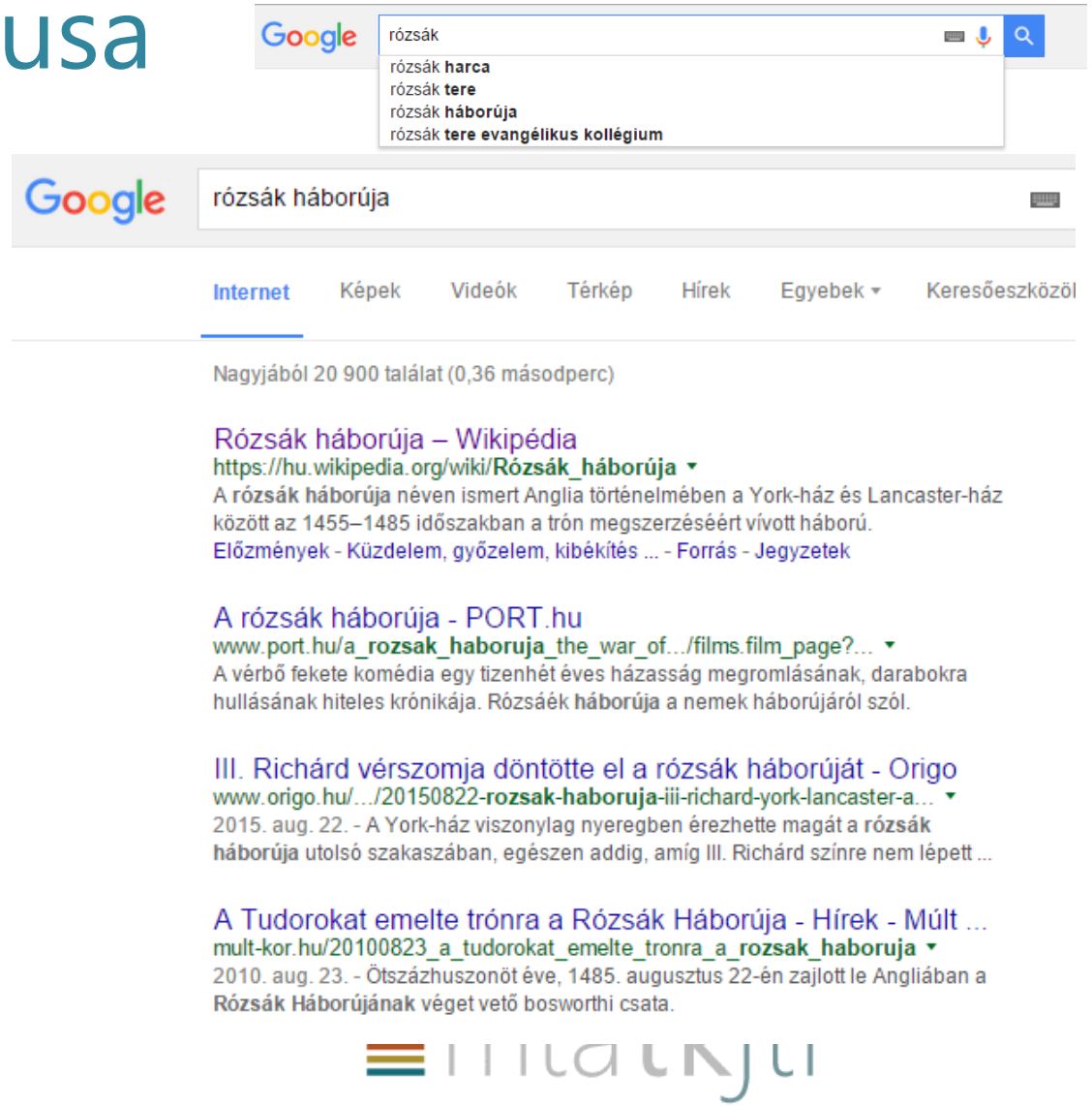
- Amiről nincsenek előismereteink, azt nem tudjuk elhelyezni a rendszerbe
- Ha a Google egyre jobban igyekszik kitalálni, hogy mire gondoltunk, hogyan bővülnek az előismereteink?
- Szükség van a Google-ön kívüli világ tágítására is, hogy olyan asszociációkra és kapcsolatokra is szert tehessünk, amelyek még nem léteznek
- A legjobban ezt a jogi problémákkal kapcsolatos kutatás támasztja alá
- Ebben kiderült, hogy a keresők jellemzően a típusproblémákban érnek el jó eredményt. Ami kilóg abban téves találatokat és eredményeket adnak



mtatkjti

A nem ismert tételek és a „know” kérdezés kétféle típusa

- A know kérdezéseket kétféle típusba lehet sorolni:
  - Tudott tétel keresése [minimálbér összege 2015-ben]; [lavrenko generative theory]
  - Kutatás – inkább címként lehet ábrázolni. „mindent tudni akarok a [rózsák háborúja]-ról”
- A nem ismert tételek közül a kutatás során talált tételek megbízhatósága és kontextusba helyezése óriási problémát jelent, ha az előismereteink hiányosak
- De ilyet nem is nagyon kapunk, mert „kiszolgálják az ízlésünket”
- Ha vannak megfelelő előismereteink, akkor a keresők hatalmas segítséget jelentenek, de csak a keresők segítségével az előismeretek nehezen bővíthetőek.
- A találatok mindenkinek ugyanazok lesznek: az iskolai „kutatási projektek” ugyanúgy néznek ki.



Problémák, következtetések

- A keresők, ahelyett, hogy kinyitnák a világot, miközben görcsösen igyekeznek kiszolgálni a keresők igényeit valójában egyre jobban bezárnak a saját világunkba.
- Nem ismerik a személyes előismereteinket, előfeltevéseinket, és valódi elvárásainkat, hanem homogenizálják a kérdéseket is, (automatikus kiegészítés funkció) és értelemszerűen a kérdésre adott válaszokat is.
- Nem tudnak „visszakérdezni", így valójában már a várakozásainkat és az előfeltevéseinket is homogenizálják. („Az emberek amikor ezt keresik, erre szoktak gondolni" – de mi van, ha én másra gondolok?)
- A kereső jó dolog – de felvet egy sor olyan új problémát, amely eddig ismeretlen volt.
- Néhány új probléma, amely teljesen új dimenzióba került az internet miatt.
  - szerzői jogsértés (korlátlan és azonnali tömeges másolás, kontrollálhatatlanság
  - iskolai piszkálódás (kikapcsolhatatlanság, a nyilvános tér kitágulása)
  - menekültkérdés (az irigyelt világ kézzelfogható közelségbe kerülése, és a régi világ jelenléte, amely az adaptálódást nagyon nehézzé teszi)
- A keresők esetén ez a probléma a megtalálható információk homogenizálása.

mtatkjti