



Információ-visszakeresés (Information Retrieval IR) elmélete és gyakorlata

Góth Júlia

122-es szoba

goth.julia@itk.ppke.hu

(4 kredit)

(Számonkérés módja: szóbeli vizsga)

Az információ-visszakeresés (IR) formális definíciója:

Def. 1: Legyen IR az alábbi:

$$IR = (U; IN; Q; O) \rightarrow R;$$

ahol

- U = felhasználó (user),
- IN = információigény (information need),
- Q = keresőkérdés (query),
- O = keresendő objektumok halmaza,
- R = a Q keresőkérdésre válaszként visszaadott objektumok halmaza.

Az információ-visszakeresés (IR) formális definíciója:

Az információigény mindig több, mint ami a keresőkérdésben megfogalmazódik, ezáltal:




Def.2.: A felhasználói információigény (IN) az alábbi:

$$IN = (Q; I);$$

- *ahol I azt a felhasználó-specifikus információ többletet jelenti, amely nem fogalmazódik meg a keresőkérdésben.*

Tigris - Google Search - Google Chrome

Tigris - Google Search x tigris, micimacko - Google x

← → ↺ <https://www.google.hu/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=Tigris>   

About 1,030,000 results (0.32 seconds)

Tigris (állat) - Wikipédia
[hu.wikipedia.org/wiki/Tigris_\(állat\)](https://hu.wikipedia.org/wiki/Tigris_(állat)) ▾ [Translate this page](#)
A tigris (Panthera tigris) a ragadozók rendjébe és a macskafélék családjába tartozó faj. Valamennyi alfaja veszélyeztetett. A tigris a legnagyobb ma élő ...
[Elterjedése](#) - [Alfajok](#) - [Megjelenése](#) - [Életmódja](#)

Tigris - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Tigris ▾
The Tigris River (/ˈtɑːɡrɪs/) is the eastern member of the two great rivers that define Mesopotamia, the other being the Euphrates. The river flows south from ...
Cities: Diyarbakır, Mosul, Baghdad Source: Lake Hazar
Countries: Turkey, Syria, Iraq


Tigris–Euphrates river system - Wikipedia, the free ...
en.wikipedia.org/wiki/Tigris–Euphrates_river_system ▾
The Tigris and Euphrates, with their tributaries, form a major river system in Western Asia. From sources in the Taurus mountains of eastern Turkey they flow ...
[Geography](#) - [General description](#) - [Ecological threats](#) - [Water dispute](#)

Tigris Restaurant
www.tigrisrestaurant.hu/english ▾
Tigris (Tiger) Restaurant is located in the centre of Budapest, at just a few minutes walking distance both from Deák Square and Roosevelt Circus. The building ...


Tigris Étterem
www.tigrisrestaurant.hu/ ▾ [Translate this page](#)
A Tigris étterem Budapest belvárosában, a Deák tértől és a Roosevelt tértől egyaránt néhány percnyi séta távolságra található. Az étteremnek helyet adó épület ...

See results about

[Tiger \(Animal\)](#)
Lower classifications: Trinil tiger, Siberian Tiger, Caspian...
Lifespan: 20 – 26 y (In captivity)



[Tigris \(River in Asia\)](#)
Cities: Diyarbakır
Source: Turkey



[Tigris–Euphrates river system](#)
The Tigris and Euphrates, with their tributaries, form a major river system in Western Asia. From sources ...

[Feedback](#)

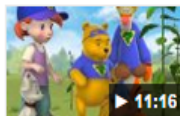
Images for Tigris, micimackó

Report images



More images for Tigris, micimackó

Barátaim Tigris és Micimackó-Zsugorodó Zsebibaba ...

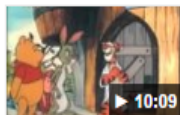


www.youtube.com/watch?v=qR9K0-GHEzk ▼

Apr 26, 2011 - Uploaded by fagurigusz

Szerintem ezek az új filmek sokkal jobbak egyrészt az animációjuk is javult másrészt egy modern világban élünk ...

Micimackó Tigris feltalálja magát - YouTube

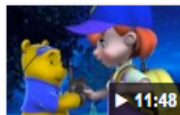


www.youtube.com/watch?v=FZckMhPfHU0 ▼

Feb 6, 2011 - Uploaded by Kovács Ákos

Hát igen ennek volt valami rendes grafikája és története de ez a teljes komputertizált verzió, ami ezzel a címmel ...

Barátaim Tigris és Micimackó-Füles a Holdra utazik - YouTube



www.youtube.com/watch?v=3nLIQ5-a0Fs ▼

Oct 1, 2011 - Uploaded by fagurigusz

Barátaim Tigris és Micimackó-Füles a Holdra utazik. fagurigusz.
SubscribeSubscribedUnsubscribe 631 ...

Rajzfilm - Barátaim Tigris és Micimackó - Mesefilm | Mese @

mesekukac.hu/micimacko-rajzfilm/ ▼ [Translate this page](#)

Az rajzfilmsorozat a százholdas pagonyban játszódik. Főszereplők „züldök”, a



Az információ-visszakeresés modelljei

2. ELŐADÁS

1. Információ-visszakereső modellek.
2. Klasszikus információ-visszakereső modellek.
3. A Boole-féle információ-visszakereső modell.
4. A vektortér modell.

1. Információ-visszakereső modellek

- 1.1. Információ-visszakereső modellek áttekintése
- 1.2. Információ-visszakereső modellek csoportosítása

1.1. Információ-visszakereső modellek áttekintése -1.

Az információ-visszakeresés alapelemei:

- dokumentum (document),
- kérdés (query),
- relevancia (relevance),
- visszakeresés (retrieval).

1. 1. Információ-visszakereső modellek áttekintése -2.

Attól függően, hogy :

- a dokumentumokat,
- a kérdést, és
- a visszakeresést

hogyan modellezzük többféle információ-visszakereső
modellt (information retrieval models) különböztetünk
meg.

1. 2. Információ-visszakereső modellek csoportosítása -1.

Klasszikus modellek tulajdonságai:

- Első (hagyományos) modellek,
- Matematikai módszereken alapulnak,
- *Kérdés (Q)* és a *Dokumentum (D)* távolságának matematikai mérésén alapul
- Mintaillesztési, illetve távolság-alapú modellek
- Könnyű implementálási lehetőség
- Kereskedelmi keresők ezek speciális változatain alapulnak

1. 2. Információ-visszakereső modellek csoportosítása -2.

Nem-klasszikus, alternatív modellek tulajdonságai:

Internetes, hálózati környezet tapasztalatait is felhasználja:

- nemcsak magát a dokumentumot vizsgálja,
- hanem a dokumentumok egymáshoz való viszonyát is felhasználja
- 80-as, 90-es évek modelljei

1. 2. Információ-visszakereső modellek csoportosítása -3.

Nem-klasszikus modellek:

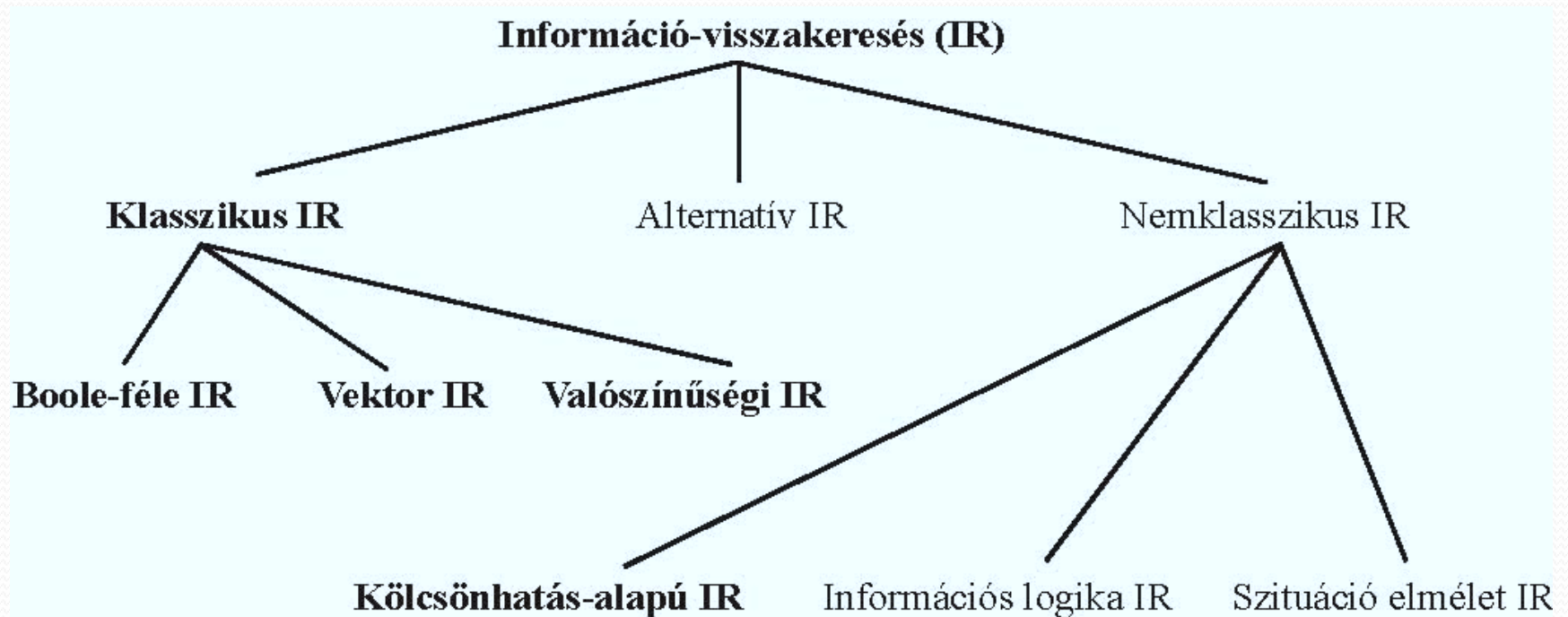
- Információs logika (Information Logic) alapú
- Szituációelmélet (Situation Theory) alapú
- Kölcsönhatás alapú (Associative, Interaction) modell, ahol a dokumentumok (objektumok):
 - nem egymástól elszigetelt egységeket képeznek,
 - hanem egy összekötött hálózatot.

1. 2. Információ-visszakereső modellek csoportosítása -4.

Alternatív modellek:

- Klaszter-alapú (Cluster) modell
- Fuzzy-alapú (Fuzzy halmazelmélet, modell) modell
- Genetikus algoritmus (Genetic algorithm) alapú modell
- Tudásbázis (Knowledge Base) alapú modell

Az információ-visszakereső modellek



2. Klasszikus információ- visszakereső modellek

2. Klasszikus információ-visszakereső modellek -1.

- Hagyományos modellek,
- Matematikai módszereken alapulnak,
- *Kérdés (Q)* és a *Dokumentum (D)* távolságának matematikai mérésén alapul
- Mintaillesztési, illetve távolság-alapú modellek
- Könnyű implementálási lehetőség
- Kereskedelmi keresők ezek speciális változatain alapulnak

2. Klasszikus információ-visszakereső modellek -2.

Klasszikus információ-visszakereső modellek :

- Boole modell (Boolean Model), matematikai logikán és halmazelméleten alapul
- Vektortér modell (Vector Space Model), ami a lineáris algebrán alapul
- Valószínűségi modell (Probabilistic Model), a valószínűségszámításon és a Bayes-statisztikán alapul

3. Boole modell

3.1. Boole modell-bevezetés

3.2. Boole-logika

3.3. Halmazelmélet

3.4. Boole modell formális leírása

3.5. Példa Boole modellre

3. Boole modell-bevezetés-1.

- Első modell
- Széles körben elterjedt
- Kereskedelmi keresők alapjai
- A Boole logikára (Boolean Logic), és a klasszikus halmazelméletre (Set Theory) épül
- A Q kérdést, és a D dokumentumokat is szavak halmazaként (kifejezések halmazaként) kezeli

3. Boole modell-bevezetés-2.

A Boole modellben a visszakeresés azon alapul, hogy:

- a dokumentum tartalmazza-e
- avagy sem a

a keresőkérdésben megadott kifejezéseket.

3.2. Boole logika -1.

Logikai feladatokhoz :

Boole algebrát használunk,
hogy a logikai kapcsolatokat matematikai úton
kezeljük.

A Boole algebra szerint, bármely bonyolult logikai
feladat megadható az alapvető logikai operátorok
(alapoperátorok) segítségével.

3.2. Boole logika -2.

Alapoperátorok:

- Negáció (tagadás, invertálás)
- Logikai ÉS kapcsolat
- Logikai VAGY kapcsolat

3.2. Boole logika -3.

- Negáció (NOT):
 - Valamely esemény, vagy logikai függvény, vagy változó igazságtartalmának az ellenkezőjét vesszük figyelembe
 - Igazságtábla:

A	NOT A
1	0
0	1

3.2. Boole logika -4.

- Logikai ÉS kapcsolat (AND):
 - Konjukció eredménye: csak akkor **1** ha valamennyi változó egyidejűleg **1**.
 - Igazságtábla:

A	B	A AND B
1	0	0
0	1	0
0	0	0
1	1	1

3.2. Boole logika -5.

- Logikai VAGY kapcsolat (OR)
 - Diszjunkció eredménye: ha bármely változó 1-es, akkor az eredmény is 1-es
 - Igazságtábla:

A	B	A OR B
1	0	1
0	1	1
0	0	0
1	1	1

Boole logika példa

Igazolja, hogy a diszjunkció asszociatív, azaz

Bármely p, q, r kijelentésre: $|(p \vee q) \vee r| = |p \vee (q \vee r)|$.

$ p $	$ q $	$ r $	$ (p \vee q) \vee r $	$ p \vee (q \vee r) $
i	i	i	i	i
i	i	h	i	i
i	h	i	i	i
i	h	h	i	i
h	i	i	i	i
h	i	h	i	i
h	h	i	i	i
h	h	h	h	h

3.3. Halmazelmélet -1.

Halmaz: közös tulajdonságú elemek összessége; (A, B, C, ...)

- **Halmazelem:** a halmaz egy eleme; (a, b, c, ...)
- **Üres halmaz:** elem nélküli halmaz; jelölés: \emptyset
- **Alaphalmaz** (halmazuniverzum): az a halmaz, amelynek minden halmaz része; jelölés: H vagy U
- **Halmazrendszer** (halmazcsalád): halmazokból álló nem üres halmaz
- **Halmaz számossága:** a benne lévő halmazelemek száma; jelölés: $|A|$
- **Hatványhalmaz:** egy halmaz összes részhalmazát tartalmazó halmaz; jelölés: $P(A)$

3.3. Halmazelmélet -2.

Halmaz megadása:

- az elemeinek felsorolásával:

pl. $A := \{ 1, 3, 5, 7, \dots \}$

- az elemek közös tulajdonságának segítségével:

pl. $A := \{ \text{Páratlan számok} \}$

- **Eleme**: egy adott elemet tartalmaz az adott halmaz; $a \in A$
- **Nem eleme**: egy adott elemet nem tartalmaz az adott halmaz; $a \notin A$

A halmaz elemeinek megadásánál az elemek sorrendje nem számít.

3.3.Halmazelmélet -3.

- Ha egy A halmaz minden eleme B halmaznak is eleme, akkor az A halmazt a B halmaz **részhalmazának** nevezzük.

Jelölése: $A \subset B$.

- Az A és B halmazok **egyenlőek**,
ha $A \subset B$ és $B \subset A$ egyidejűleg fennáll.
- **Diszjunkt** halmazok: olyan két halmaz, amelynek nincs közös része ($A \cap B = \emptyset$)

3.3.Halmazelmélet -4.

Halmazműveletek:

- Az A és B halmazok *egyesítésén* vagy **unióján** mindazon elemek halmazát értjük, amelyek vagy A-nak, vagy B-nek (vagy mindkettőnek) elemei.

Jelölése: $A \cup B = \{x \mid x \in A \text{ vagy } x \in B\}$.

- Az A és B halmazok *közös részén* vagy **metszetén** azon elemek halmazát értjük, amelyek A-nak és B-nek is elemei.

Jelölése: $A \cap B = \{x \mid x \in A \text{ és } x \in B\}$

- Az A és B halmazok **különbségén** azon elemek halmazát értjük, amelyek A-nak elemei, de B-nek nem.

Jelölése: $A - B = \{x \mid x \in A \text{ és } x \notin B\}$, vagy $A \setminus B$.

3.3. Halmazelmélet -5.

Halmazműveletek (folyt.):

- Az A és B halmazok **szorzatának** (*Descartes-szorzatának*) nevezzük azt a C halmazt, amelynek elemei az A és B halmaz elemeiből az összes lehetséges módon képzett rendezett elempárokból áll.

Jelölése: $C = A \times B = \{(a,b) \mid a \in A \text{ és } b \in B\}$.

- Ha az A halmaz a H *alaphalmaz* részhalmaza, akkor a $H-A$ halmazt az A halmaz (H -ra vonatkozó) **komplementer halmazának** vagy **kiegészítő halmazának** nevezzük.

Jelölése: ha $A \subset H$, $A^- = H - A = \{x \mid x \in H \text{ és } x \notin A\}$

3.3. Halmazelmélet -6.

Tetszőleges A , B és C halmazokra érvényesek a következő összefüggések:

- *idempotencia*: $A \cup A = A$ és $A \cap A = A$
- *kommutativitás*: $A \cup B = B \cup A$ és $A \cap B = B \cap A$
- *asszociativitás*: $A \cup (B \cup C) = (A \cup B) \cup C$ és
 $A \cap (B \cap C) = (A \cap B) \cap C$
- *disztributivitás*: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$,
 $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

Legyen A és B ugyanazon H alaphalmaz két tetszőleges részhalmaza. Érvényesek a következő egyenlőségek.

- $A \cup \emptyset = A$ $A \cap \emptyset = \emptyset$
- $A \cup H = H$ $A \cap H = A$ ha $A \subset H$
- $A \cup A^c = H$ $A \cap A^c = \emptyset$ ha $A \subset H$.

Halmazelmélet-példa

Egy fordítóirodában 52 fordító dolgozik.

- Közülük 20-an beszélik az orosznyelvet,
- 19-en a franciát és
- 35-en az angol nyelvet.
- Az orosz és az angol nyelvet is 11,
- a franciát és az orosz 7,
- a franciát és az angolt pedig 9 fordító beszéli.

a./ Hány fordító beszéli mindhárom nyelvet?

b./ Hányan beszélik közülük csak az orosz nyelvet?

3.4.Boole modell formális leírása -1

Boole modell (Boolean Information Retrieval Model)
formálisan:

Adottak:

- $D = \{D_1, \dots, D_j, \dots, D_m\}$ dokumentumok, ahol $D_j \in \wp(T)$ valamint a
- $T = \{t_1, t_2, \dots, t_i, \dots, t_n\}$ indexkifejezések, amelyek leírják a dokumentumokat,
- A Q keresőkérdés egy Boole kifejezés

3.4.Boole modell formális leírása -2

- A dokumentumok: formálisan indexkifejezések halmaza
- A valóságban a dokumentum és a reprezentációja két külön entitás.
- Matematikailag viszont ekvivalensnek tekinthetők:
 - hiszen a dokumentumokat a rájuk jellemző indexkifejezésekkel reprezentáljuk,
 - valamint a visszakeresés során is a dokumentumok reprezentációját használjuk, és nem magukat a dokumentumokat.

3.4.Boole modell formális leírása -3

A visszakeresés azon alapul, hogy az adott dokumentum

- tartalmazza-e ,

vagy

- nem

a keresőkérdésben megadott kifejezéseket.

3.4.Boole modell formális leírása -4

A visszakeresés két fő lépésből áll, amelyek a következők:

1. Lépés:

Meghatározzuk a dokumentumok azon S_i halmazát, amely tartalmazza vagy nem a t_i kifejezést:

- t_i kifejezésre: $S_i = \{D \mid t_i \in D\}$
- *negált* t_i ($\neg t_i$) kifejezésre : $S_i = \{D \mid t_i \notin D\}$

3.4.Boole modell formális leírása -5

2. Lépés:

A Q kérdésre válaszként visszaadott dokumentumokat a logikai műveleteknek megfelelő halmazműveletek adják meg:

- \cap (metszet) megfelel a \wedge (logikai ÉS) kapcsolatnak,
- \cup (unió) megfelel a \vee (logikai VAGY) kapcsolatnak

3.4.Boole modell formális leírása -6

Például:

$$Q = t_1 \text{ OR } (t_2 \text{ AND } t_3)$$

- S_1 eredményhalmaz t_1 indexkifejzésre,
- S_2 eredményhalmaz t_2 indexkifejzésre,
- S_3 eredményhalmaz t_3 indexkifejzésre,

A Q keresőkérdésre visszkapott eredményhalmaz:

$$S_1 \cup (S_2 \cap S_3)$$

Példa Boole modellre

- Legyen a valós dokumentumok halmaza O az alábbi:
- $O = \{O_1, O_2, O_3\}$

O_1	O_2	O_3
Még nyílnak a völgyben a kerti virágok , Még zöldell a nyárfa az ablak előtt, De látod amottan a téli világot? Már hó takará el a bérci tetőt.	Fenyő ága Hósubában , Mire vársz a Hófúvásban ? Hideg az a Kristály bunda , Gyere haza Kis házunkba.	Fekete Pont Fehér Ágon: Varjú károg: Fázom Fázom.

Példa Boole modellre

Legyen T indexkifejezések halmaza az alábbi:

$T = \{t_1; t_2; t_3; t_4; t_5; t_6\}$, ahol:

- $t_1 = \text{virág},$
- $t_2 = \text{tél},$
- $t_3 = \text{hó},$
- $t_4 = \text{fenyő},$
- $t_5 = \text{bunda},$
- $t_6 = \text{varjú}.$

Példa Boole modellre

Adja meg a $T = \{t_1; t_2; t_3; t_4; t_5; t_6\}$ indexkifejezésekkel az

- $O = \{O_1; O_2; O_3\}$ objektumokat reprezentáló
- $D = \{D_1; D_2; D_3\}$, dokumentumhalmazt!
- $D_1 = \{\text{virág, tél, hó}\} = \{t_1; t_2; t_3\}$
- $D_2 = \{\text{hó, fenyő, bunda}\} = \{t_3; t_4; t_5\}$,
- $D_3 = \{\text{varjú}\} = \{t_6\}$.

Példa Boole modellre

- Legyen Q keresőkérdés az alábbi : $\text{hó} \wedge \text{fenyő}$, azaz

$$Q = t_3 \wedge t_4$$

Ekkor az S_3 és S_4 a következő:

- $S_3 = \{D_1; D_2\}$, azon dokumentumok halmaza, amelyek tartalmazzák a $t_3 = \text{hó}$ indexkifejezést,
- $S_4 = \{D_2\}$, azon dokumentumok halmaza, amelyek tartalmazzák a $t_4 = \text{fenyő}$ indexkifejezést.

A Q keresőkérdésre válaszként adott dokumentumhalmaz a következő:

$$S_3 \cap S_4 = \{D_2\}$$

Tehát a Q keresőkérdésre adott válasz az O_2 objektum lesz.

Példa Boole modellre

Legyen Q keresőkérdés az alábbi:

- $Q_1 = \neg t_1 \vee \neg t_2$
- $Q_2 = (t_1 \vee t_3) \wedge (t_2 \vee t_3)$
- $Q_3 = (t_1 \vee (t_2 \wedge t_3))$

4. Vektortér modellek

4.1. Vektortér modell helye az információ-visszakereső modellek közt

4.2. Vektortér modell formális leírása

Vektorok normalizálása (*Normalization*)

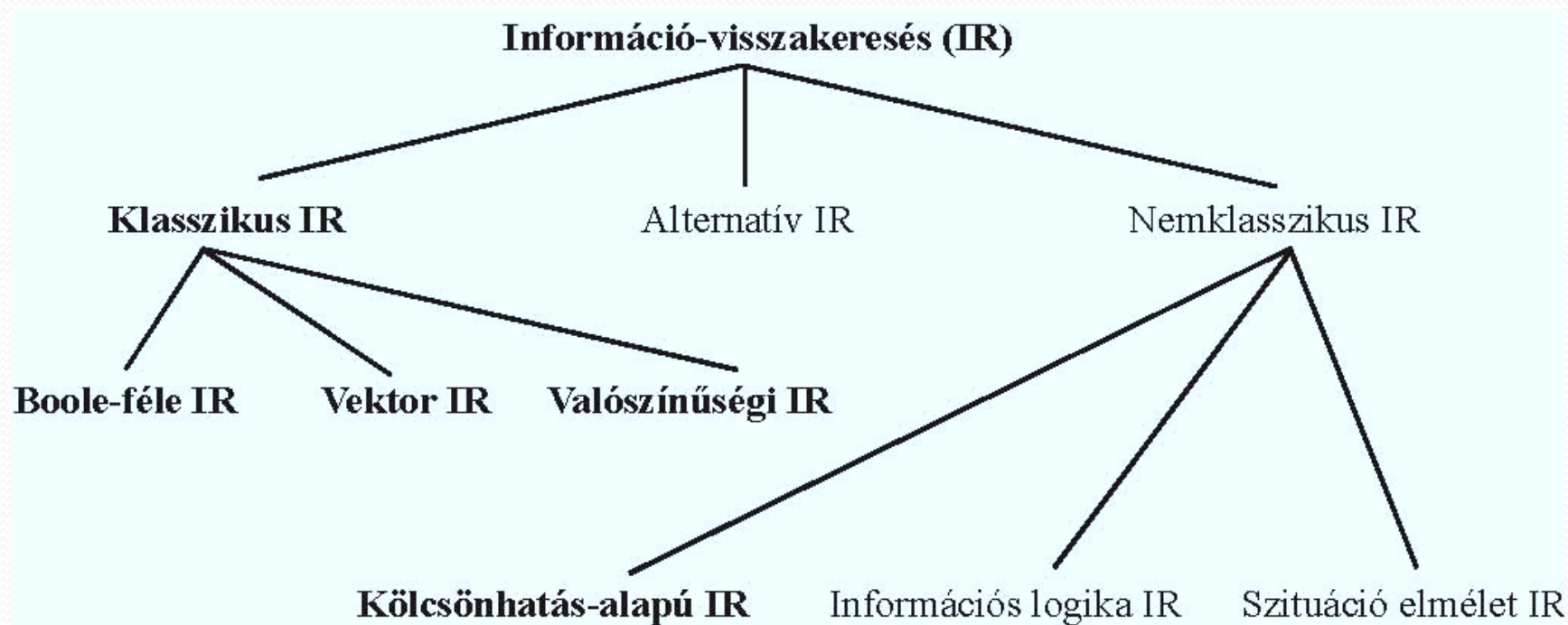
- vektor iránya nem változik,
- viszont a hossza egy lesz.
- A normalizált vektort úgy kaphatjuk meg, hogy az eredeti vektort elosztjuk a hosszával.

4.1. Vektortér modell az információ-visszakereső modellek közt

Klasszikus információ-visszakereső modellek :

- Boole modell (Boolean Model), matematikai logikán és halmazelméleten alapul
- **Vektortér modell (Vector Space Model), a lineáris algebrán alapul**
- Valószínűségi modell (Probabilistic Model), a valószínűségszámításon és a Bayes-statisztikán alapul

4.1. Vektortér modell az információ-visszakereső modellek közt -5



4.2. Vektortér modell formális leírása -1.

A vektortér modell:

- egy fontos,
- jól érthető, és
- széles körben kutatott és használt klasszikus modell
- amelyet szöveges objektumok feldolgozására, és információ-visszakeresésre már régóta használnak (Salton, 1966).

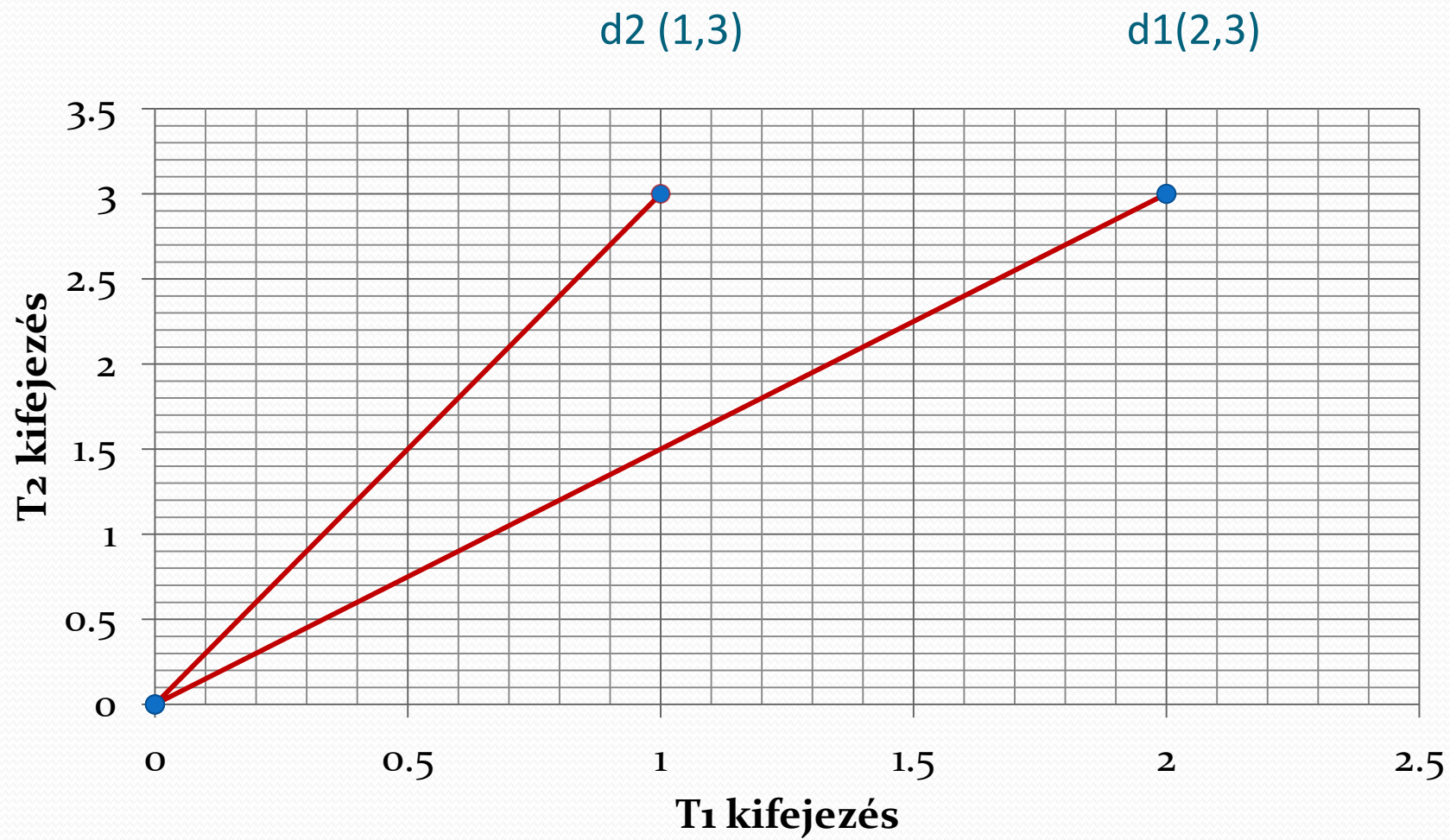
4.2. Vektortér modell formális leírása -2.

Ezt vektortér modellnek nevezik,

- mert minden
 - dokumentum és a
 - kérdés is atér egy pontjába van leképezve,
- amely tér alapját a dokumentumokban található kifejezések adják.

4.2. Vektortér modell formális leírása -3.

- A tér matematika modellje:
 - egy orthonormált euklideszi tér,
 - amelyben a tengelyek páronként egymásra merőlegesek.
- A tér dimenzióit az indexkifejezések adják.
- A visszakeresés azon alapul, hogy a
 - kérdés-vektor és a
 - dokumentum-vektormennyire van „közel” egymáshoz.



4.2. Vektortér modell formális leírása -4.

Legyen

- D egy véges halmaz, melynek elemei a dokumentumok:
$$D = \{D_1, \dots, D_j, \dots, D_m\}$$
- T egy véges halmaz, melynek elemei az indexkifejezések:
$$T = \{t_1, \dots, t_i, \dots, t_n\}$$
- Minden D_j dokumentumhoz hozzárendelünk egy n hosszú \mathbf{v}_j súlyvektort. A vektor elemeit súlyoknak nevezzük:
$$\mathbf{v}_j = (w_{ij})_{i=1, \dots, n} = (w_{1j}, \dots, w_{ij}, \dots, w_{nj})$$
 - általában $0 \leq w_{ij} \leq 1$
 - a w_{ij} súllyal azt fejezzük ki, hogy a t_i kifejezés milyen mértékben tükrözi a D_j dokumentum tartalmát.

4.2. Vektortér modell formális leírása -4.

A \mathbf{v}_j súlyvektorokból megadható a **TxD** (term-by-document) kifejezés-dokumentum mátrix:

- m (dokumentumok száma) oszlopa van
- n (indexkifejezések száma) sora van
- amelynek elemei a súlyok,
- $\mathbf{TD} = (w_{ij})_{n \times m}$, ahol $i=1 \dots n$, $j=1 \dots m$

4.2. Vektortér modell formális leírása -5.

• TD Mátrix

$$\begin{pmatrix} w_{11} & \cdots & w_{1j} & \cdots & w_{1m} \\ \vdots & & \vdots & & \vdots \\ w_{i1} & \cdots & w_{ij} & \cdots & w_{im} \\ \vdots & & \vdots & & \vdots \\ w_{n1} & \cdots & w_{nj} & \cdots & w_{nm} \end{pmatrix}$$

4.2. Vektortér modell formális leírása -6.

A kifejezések kiválasztása és a súlyok meghatározása :

- nehéz elméleti (nyelvészeti, szemantikai) és
- gyakorlati probléma.

Ennek számos lehetséges megoldása van.

- A legnyilvánvalóbb az, hogy az index- kifejezéseket magukban a dokumentumokban keressük.
- Feltételezzük, hogy a szavak előfordulási gyakorisága a dokumentumokban jelentőséggel bír, és ezért azonosítóként használhatók.

4.2. Vektortér modell formális leírása -7.

Indexkifejezések meghatározása:

- automatikus (a dokumentumból)
- manuális (szakértők által).