

Mészáros Tamás

Ágensalapú információbeszerzés és -megosztás

Témavezető:

Dr. habil. Dobrowiecki P. Tadeusz

ágensek

Információk a digitális univerzumban

- Hatalmas mennyiségű digitális tartalmat halmozunk fel. Becslések szerint...

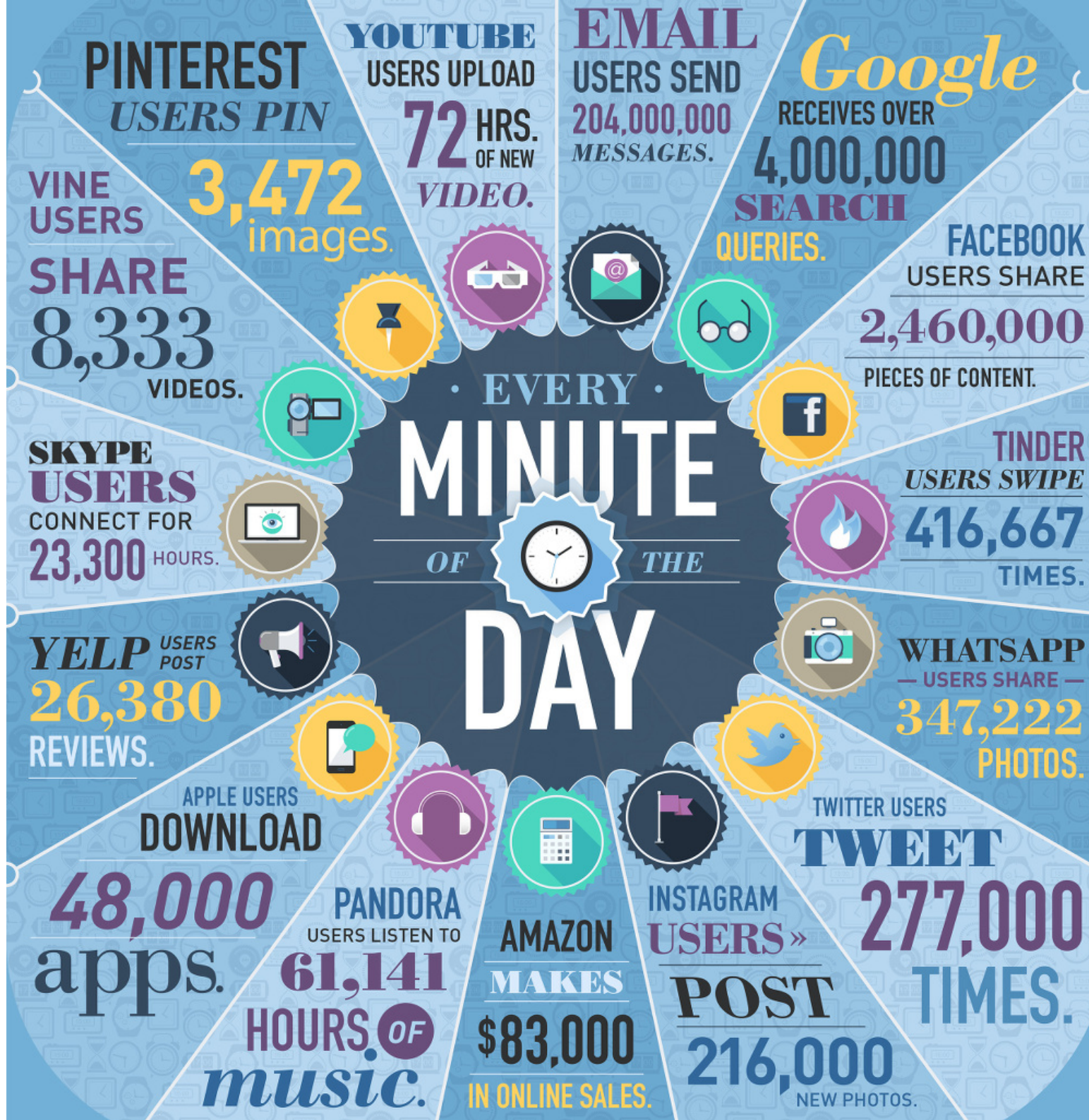
[Forbes](#) (2015) eladott tárolókapacitás 2014 – 2019: 0,5 – 2 ZB

[IDC/EMC](#) (2014) teljes tárolt adatmennyiség 2013 – 2020: 4.4 – 44 ZB

1 ZB ([zettabájt](#)) = 1 milliárd TB - minden [földlakóra](#) kb. 1 TB tárolt adatmennyiség jut

- Ezzel a világgal ma már szinte mindenki kapcsolatban áll
 - egyének (nem csak szakemberek), vállalatok, kormányzati szervek stb.
- Az információszerzés alapvetően ebbe a világba helyeződött át.
 - digitális könyvtárak, gyűjtemények
 - digitalizált vagy már elektronikusan készülő dokumentumok
 - weblapok, tweetek, fórumbejegyzések, Facebook üzenetek, email
- Nemcsak fogyasztói vagyunk az információknak, hanem egyre inkább termelői és megosztói is.

Minden percben ...



[illegible]

Az információelérés problémái

- Meg tudjuk fogalmazni, mire van szükségünk? Ismerjük a keresőket?
Google operátorok: site: filetype: intext: OR -kulcsszó ... (20+)
- Megkapjuk, amit szeretnénk? (*felidézés, recall*)
 - Az összes releváns dokumentum közül mennyit ad vissza a rendszer
 - A web mekkora részét „látja” például a Google? 10%? 30%?
- Azt kapjuk, amit szeretnénk? (*pontosság, precision*)
 - A visszakapott találatok között hány releváns tétel van?
Releváns a „Térítési díjak – II. Rákóczi Ferenc Általános Iskola” találat?
 - Mennyi időt kell(ene) szánnunk a tartalmak kiértékelésére?
266 000 találatot (Google: „Rákóczi betegség”) át tudunk nézni?
- Nem új problémák, már a könyvtártudomány is foglalkozott velük
 - a digitális világ súlyosbítja a helyzetet
 - ugyanakkor új megoldásokat is kínál

A digitális világ lehetőségei

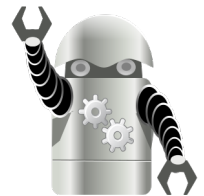
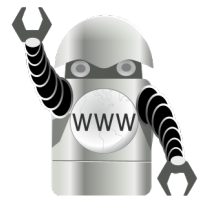
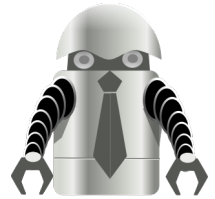
- **Automatizált** rendszereket használhatunk
 - a forrástartalmak felkutatására és elemzésére
 - kb. ötévente **egy nagyságrenddel növekedő** számítási kapacitással
 - alapvetően statisztikai módszerekkel (egyre nyilvánvalóbb korlátokkal)
- **Tudásalapú** működéssel szélesíthetők, átléphetők a korlátok (példák)
 - források és kapcsolataik automatizált feltérképezése (**Pagerank**)
 - tartalmak kiegészítése metaadatokkal (**RDFa, Microdata** stb.)
 - adatforrás-modellezés (pl. Internet of Things szenzorok **adatmodelljei**)
 - szemantikus tartalomkezelés (W3C **Szemantikus Web** koncepció)
 - gépi tartalomelemzés (számítógépes nyelvészet, adat- és szövegbányászat)
 - tárgyterületi modell (fogalmi hálók, ontológiák stb.) gépi reprezentációja
 - gépi tanulás (rangsorolás, paraméterhangolás, rejtett nyelvi modell építés)

Kutatásaim tárgya



Kutatási célkitűzésem

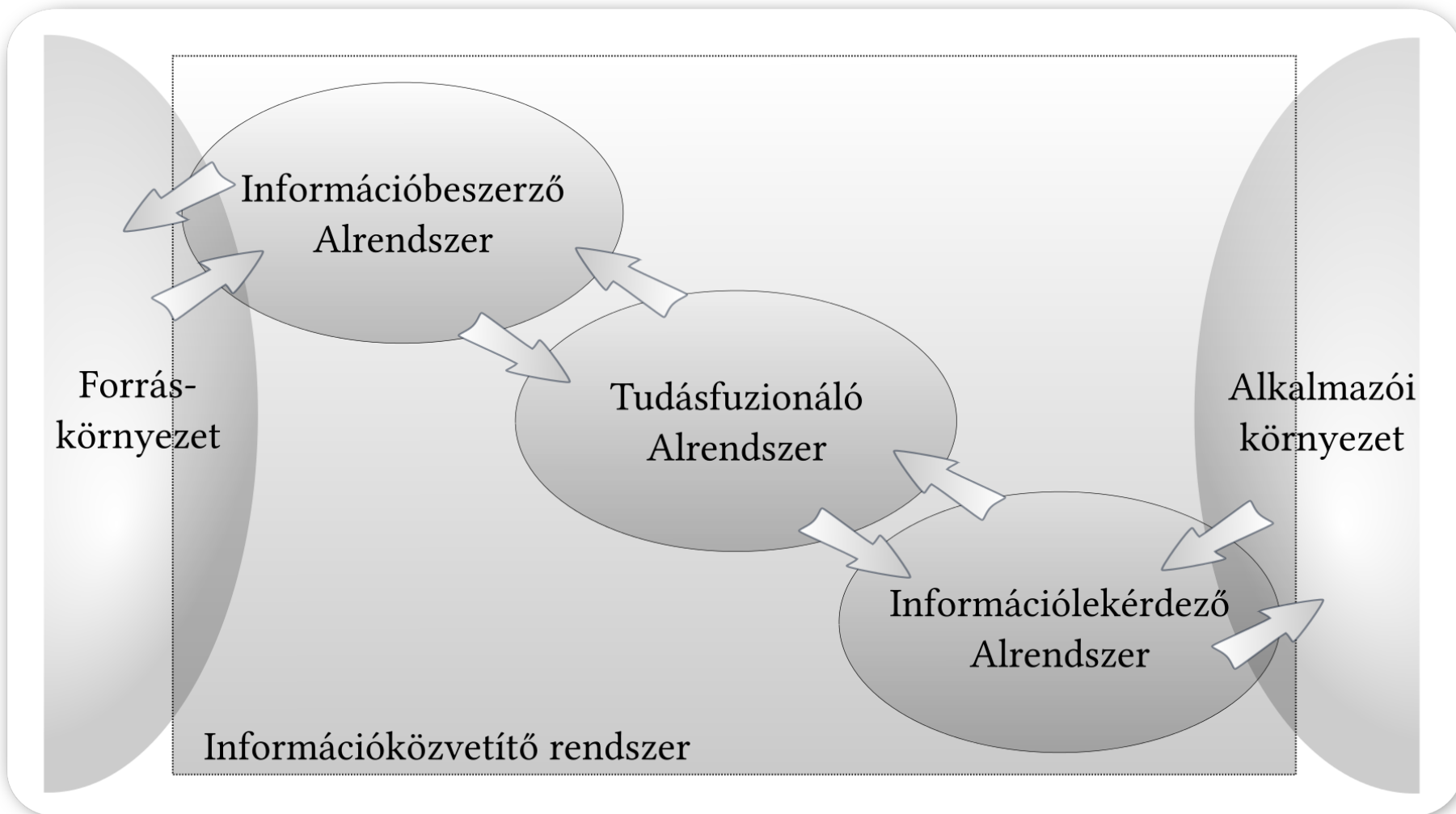
- Intelligens ágensek ...
 - önállóan működő, tudásalapú entitások
 - környezetbe ágyazott rendszerek, azt észlelik és befolyásolják
 - proaktívak: céljaiknak megfelelően kezdeményeznek
 - szociálisak: kommunikálnak és együttműködnek másokkal
- ... alkalmazása az információbeszerzés és -megosztás terén ...
 - forrásfelkutatás és információbeszerzés
 - tartalomelemzés
 - tárgyterület- és kontextusmodellezés
 - emberi és gépi interfészek megvalósítása
- ... egységes tervezési és implementációs módszertan kidolgozásával tudásalapú rendszerek (ágensek) funkcionális és tartalomintegrációjára az információbeszerzés és szolgáltatás területén



Kutatási és mérnöki célkitűzéseim

- Tudásalapú módszerek az információelérés teljes folyamatában
 - az információs igény egyszerű, de pontos megfogalmazása
 - az elosztott, heterogén információbeszerzés támogatása
 - a beszerzett tartalmak gépi „megértésének” elősegítése
 - különféle elemzési módszerek egységbe integrálása
 - integrált tudás- és adattár létrehozása, és erre építve
 - a felhasználók felkéréseinek megfelelő válasz előállítása
- A használhatóság (önállóság, integráció, kezelhetőség) elősegítése
 - az autonóm működés hangsúlyozása a folyamat minden lépésében
 - az emberi (tárgyterületi, kezelői) tudás integrálása
 - platform- és alkalmazásfüggetlen funkcionális, adat- és tudásintegráció
 - természetes nyelvű felhasználói interfészek alkalmazása

A kialakított rendszer főbb részei



Téziscsoportok

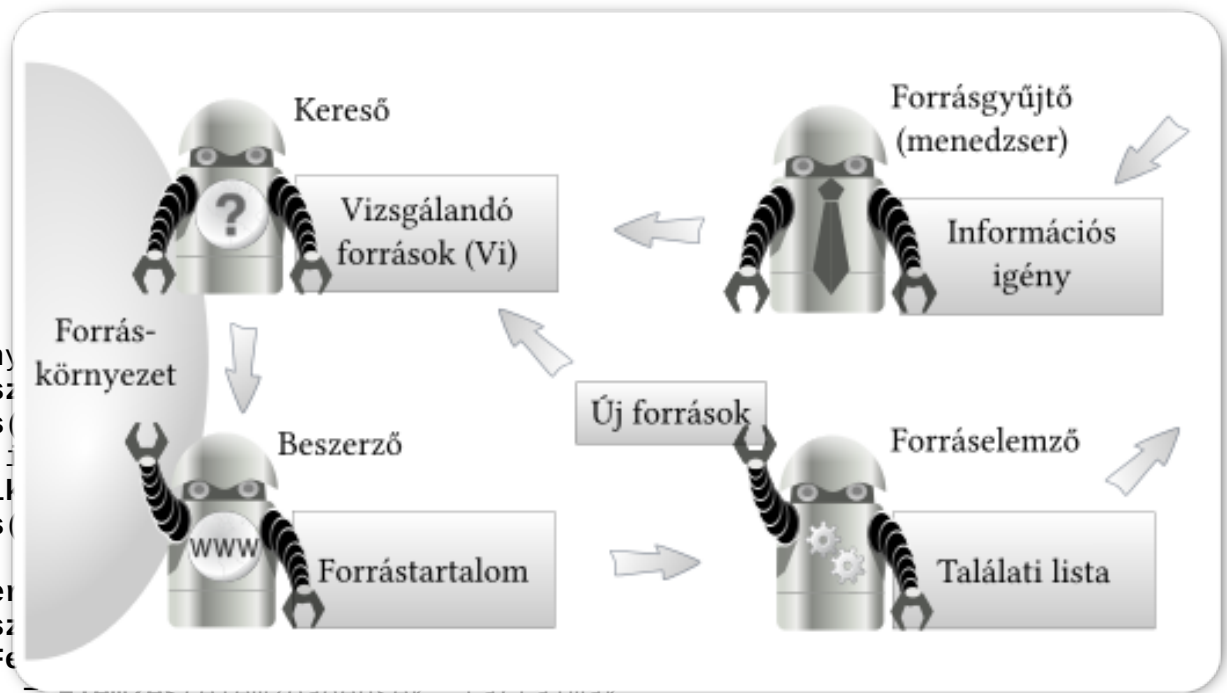
- Ágensalapú architektúra és működés
 - Információbeszerző Alrendszer: forráskeresés és tartalomelemzés
 - a Tudásfuzionáló Alrendszer komponenseinek ágensalapú integrációja
 - az Információlekérdező Alrendszer interfész ágensei
- Az elemzőrendszer elvi működése és tartalomreprezentációja
 - tartalmak és elemzési adatok univerzális, XML-alapú reprezentációja
 - modelltranszformációra épülő elosztott tartalomelemzés
 - elemzett tartalomváltozatok összekapcsolása XLink hivatkozáshálózattal
- Kontrollált természetes nyelvű felhasználói felület
 - az információs igények leírása és gépi reprezentációra fordítása
 - az elégséges kifejezőerejű kontrollált nyelv létrehozásának módszere
 - a természetes nyelvű felület működési módja

1. téziscsoport: ágensalapú felépítés

- 1.1. Kidolgoztam egy több-ágens architektúrát és működési módot az Információbeszerző Alrendszer számára, amelynek feladata információforrások felkutatása és forrástartalmak beszerzése. A vállalkozó hálók protokoll alkalmazásával egy menedzser ágens fogadja az információs igény leírását, majd forráskereső, tartalombeszerző és -elemző vállalkozó ágenseket bíz meg a dekomponált részfeladatok végrehajtásával.
- 1.2. Kialakítottam egy megoldást a tudásalapú komponensek, a Dokumentumtár és a Konceptcionális Modell integrációjára. A fuzionált tudástár az architekturális integrációt interfész ágensek alkalmazásával valósítja meg. Az egységes ágenskommunikációs protokoll mellett tartalmi szinten a közös koncepcionális modellre és az XML tartalomformátumra épít. Részletesen a Bayes-hálók integrációját ismertettem.
- 1.3. Kidolgoztam az Információlekérdező Alrendszer ágensalapú megvalósítását, amelynek kereső, lekérdező, böngésző és riportgeneráló funkcióit menedzser ágensek valósítják meg a fuzionált tudástár ágenseire támaszkodva, míg az alkalmazói környezet rendszereivel és felhasználóival interfész ágensek kötik össze és kontextusmodellező ágensek segítenek a külső felkérések pontosabb értelmezésében.

Forráskereső ágensrendszer

- Vállalkozási hálók protokoll alkalmazásával működő több-ágens rendszer
 - menedzser: Forrásgyűjtő ágens
 - vállalkozók: kereső, tartalombeszerző és elemzőágensek

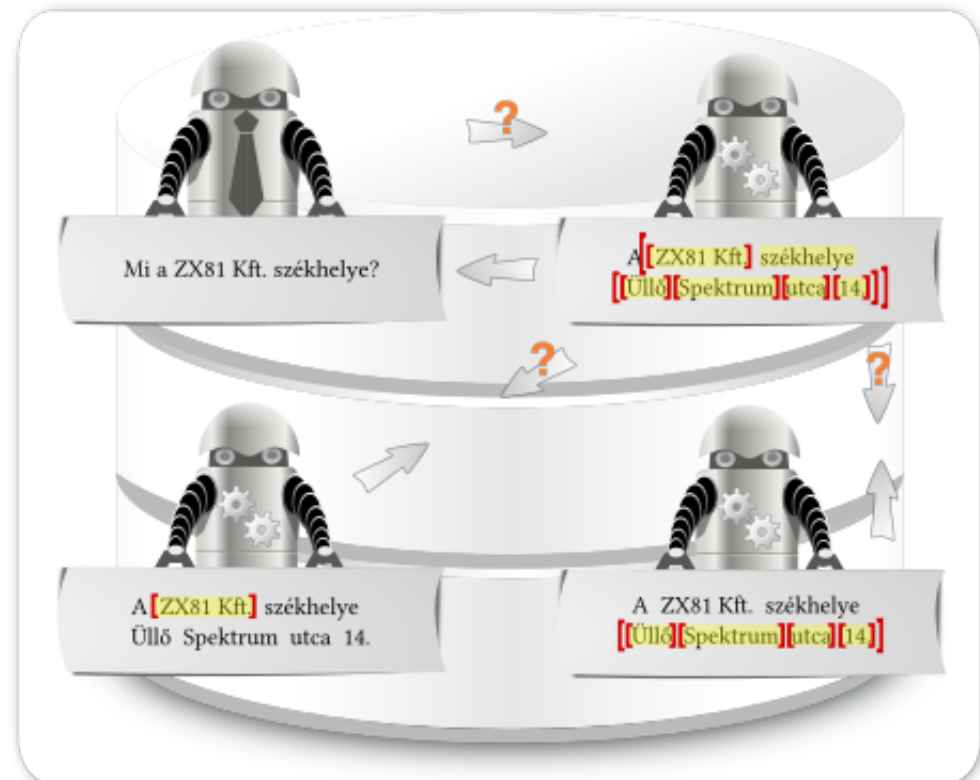


```

ForrásgyűjtőÁgensProgram(környezet)
    információs_igény = IgényÉszlelés(környezet)
    források[] = ForrásÉszlelés(környezet, információs_igény)
    keresőminta = Igényelemzés(információs_igény)
    keresőágensek[] = KeresőFeladat(környezet, keresőminta)
    források[] += Forráskeresés(környezet, források, keresőágensek)
    Ciklus {
        beszerzőágensek[] = Beszerző(környezet, források)
        tartalmak[] = Tartalombeszámítás(környezet, beszerzőágensek)
        elemzőágensek[] = ElemzőFeladat(környezet, tartalmak)
        források[], relevanciák[] = Elemzés(környezet, elemzőágensek, tartalmak,
                                            információs_igény)
        találatok[] += (források, relevanciák)
    } Ciklus amíg ! LeállásiFeltétel(találatok, források)
    találatok[] = TalálatRangasoroló(találatok)
    TalálatokVisszaadásaCselekvés(környezet, találatok)
}
    
```

Tartalomelemző ágensrendszer

- Elemzőágens
 - adott tartalomtípust képes értelmezni, elemezni, és új tartalomtípust állít elő
 - az XML technológia elemeire támaszkodik (feldolgozás és transzformáció)
- Az elemzőkből felépülő több-ágens rendszer
 - transzformációsorozatokat hajt végre bemenetei és felkérései alapján
 - kooperatív és versengő ágensok
 - adat- és célvezérelt működésű
 - futásidőben adaptívan változhat
 - dinamikusán skálázható
 - alkalmazásfüggetlen megoldás



2. téziscsoport: elemzés és tartalomreprezentáció

- 2.1. Kidolgoztam egy XML-alapú formátumot, amely egy dokumentumon belül képes a forrástartalmak és azok különbözőképpen elemzett (különböző struktúrájú) változatainak tárolására. Az implementációs eszközöktől független formátum két részből áll: az elemzett tartalomból (amely az alkalmazás igényei szerint tetszőleges módon kialakítható), valamint a beszerzési és az elemzési metaadatokból.
- 2.2. A dokumentumformátumra alapozva kialakítottam a tartalomelemző-rendszer elvi működését, amelyet modelltranszformátor-hálózatként határoztam meg. A módszer dekomponálja a feladatot az ágensalapú megvalósítás számára, valamint lehetőséget teremt az elemzés formális leírására, elemzési hálózatok tervezésére és verifikálására is. Kidolgoztam az ágensalapú megvalósítást, és meghatároztam az ágensek programját, amely kiválasztja a releváns bemeneti tartalomelemeket, előállítja az elemzés eredményét, majd azt a kívánt kimeneti formába önti.
- 2.3. Kidolgoztam egy módszert az elemzések által előállított tartalomváltozatok összetartozó részeinek (attribútumok, elemek és azokon belüli szövegrészletek) összekapcsolására. A kidolgozott megoldás egy XLink hivatkozáshálózat alkalmazásával írja le a strukturált dokumentumrészek között kapcsolatokat. Az így kialakított nézethálózat alapjául szolgálhat magyarázatgeneráló funkciók megvalósításának is.

Tartalomreprezentáció és elemzés

metaadatok
XTL nyelven

tartalmi rész
XTM nyelven

XML Tartalom Dokumentum (XTD)

Beszerezés ideje: 2015. 04. 13. 13:12:12 +0200

Forrás: <http://www.zx81kft.hu/hir/20150320/hitelfelvetel.html>

Elemzések: ENT:Cégnév, ENT:Szám, ENT: Adószám, ENT: mennyiség, ENT: személynév

beszerzési
metaadatok

elemzési
metaadatok

jelölt
tartalomelemek

elemzett

forrástartalom

eredeti

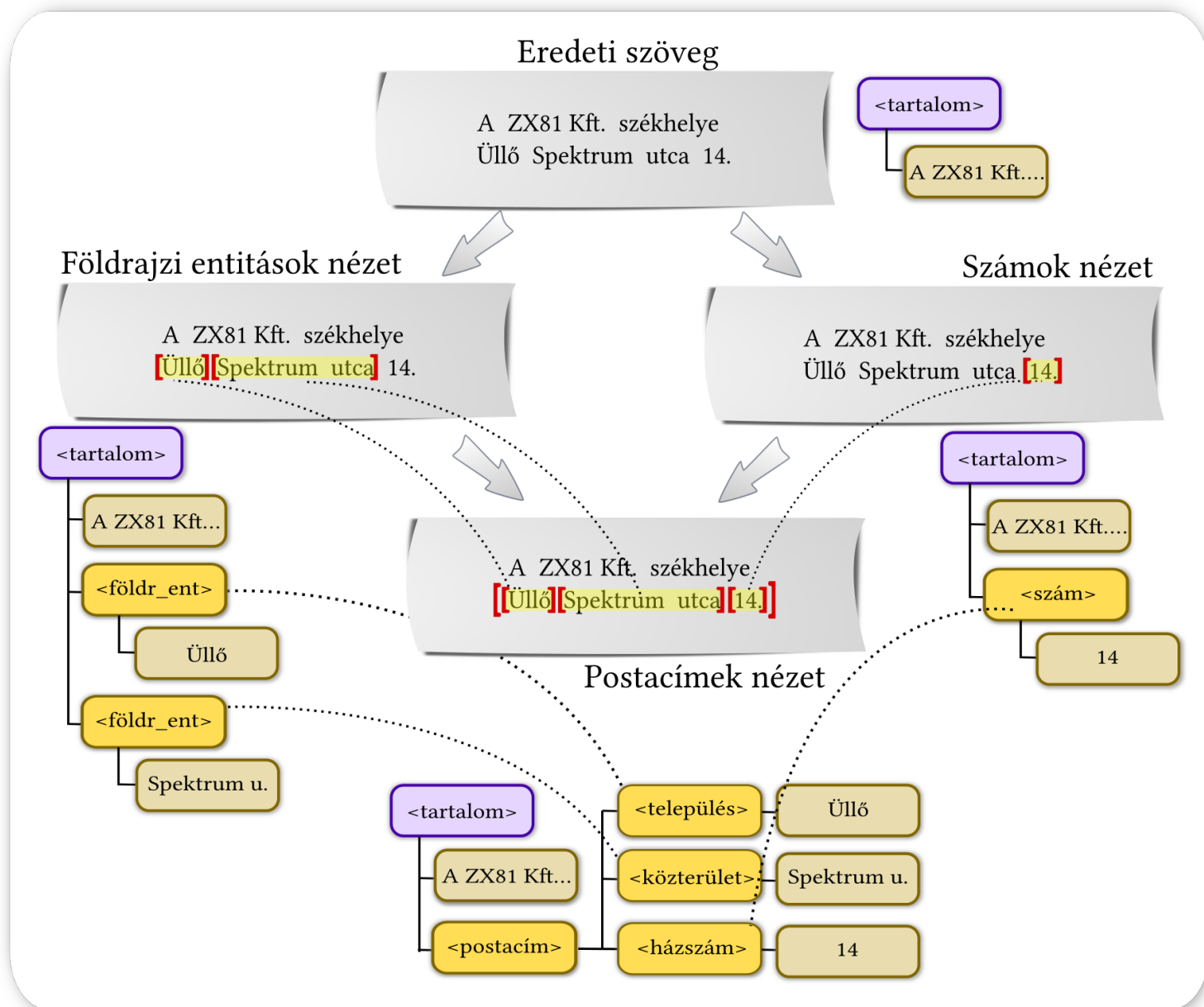
forrástartalom

A **ZX81 Kft.** a Magyar Nemzeti Bank **2015**ös növekedési hitelprogramja keretében **140 millió forint**os vállalati hitelt vett fel. A Kft ügyvezetője **Háté Zénó** kijelentette, hogy vállalkozásuk nagyon sikeresen indult a program pályázatán, és biznak benne, hogy az összeg jelentősen hozzájárul termelékenységük javításához.

A **ZX81 Kft.**(adószáma **1 234 578-1-41**) székhelye **Üllő, Spektrum utca 14.****2015**évi adózott eredménye **68 millió forint**tett. A vállalkozás **12**főállású dolgozót foglalkoztat.

```
Elemzés(XTD) {
  XMLrészletek = BemenetKiválasztás(XTD, elemző_minta)
  XMLeredmény = TartalomElemzés(XMLrészletek)
  kimeneti_nézet = NézetElőállítás(XMLeredmény, XSLT)
  NézetEllenőrzés(kimeneti_nézet, elvárt_nézet_deklarációja)
  újXTD = NézetBeillesztés(XTD, kimeneti_nézet)
  Visszaadás újXTD
}
```

Az elemzési eredmények hivatkozáshálózata



3. téziscsoport: természetes nyelvű felület

- 3.1. Kidolgoztam egy kontrollált természetes nyelvű megoldást az alkalmazói környezet humán felhasználói által megfogalmazott információs igények leírására. Segítségével lehetővé válik egy könnyen használható és rugalmas felhasználói felület kialakítása. A nyelvtani szabályok, a szókincs és a szemantika rögzítésével valósítottam meg a természetes nyelvű mondatok feldolgozását és átalakítását (lefordítását) az információs igények belső reprezentációjára.
- 3.2. Meghatároztam az elégséges kifejezőerejű nyelv megalkotásának rendszeren belüli forrásait. Kialakítottam egy eljárást környezetfüggetlen kontrollált nyelvek modellalapú elkészítésére, amely a tárgykörnyezet koncepcionális modelljének nyelvi réteggel történő kiegészítésére épít.
- 3.3. Kidolgoztam a kontrollált nyelvű felhasználói felület működési módját. Bemutattam, hogy javaslattevő funkció megvalósításával hogyan alakítható ki egy prediktív szövegbeviteli rendszer, amely segíti a felhasználókat a nyelv korlátos szabályainak és szókincsének alkalmazásában. Kidolgoztam egy nyelvi elemzésre és javaslattevőre alkalmas reprezentációt, és részletesen bemutattam az elemzés és javaslattevő algoritmusának működését környezetfüggetlen nyelvtanokra.

Keresés a főnévi vonzattárban kérdés segítségével

Az alábbi mezőben magyar nyelven lehet feltenni főnevekkel és vonzataikkal kapcsolatos kérdéseket. A rendszer csak bizonyos szerkezetű kérdéseket tud megválaszolni, ezért kontrollálja a kérdés feltevését.

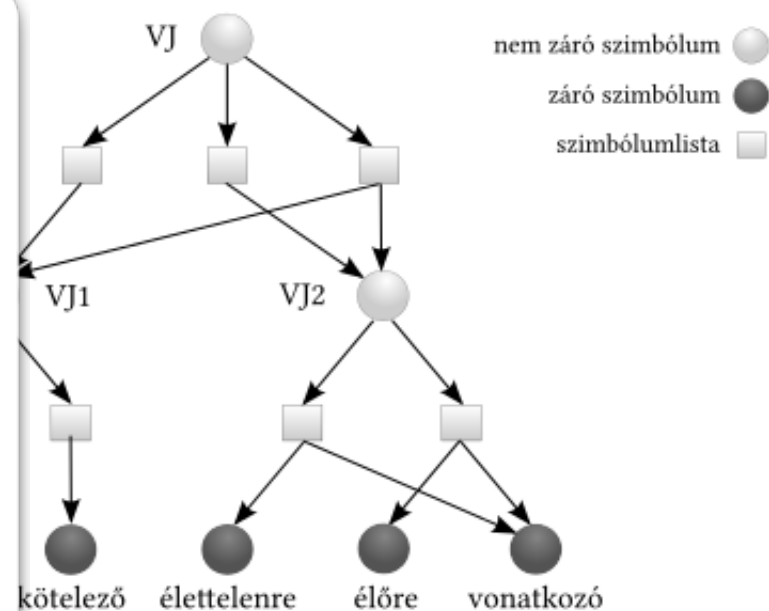
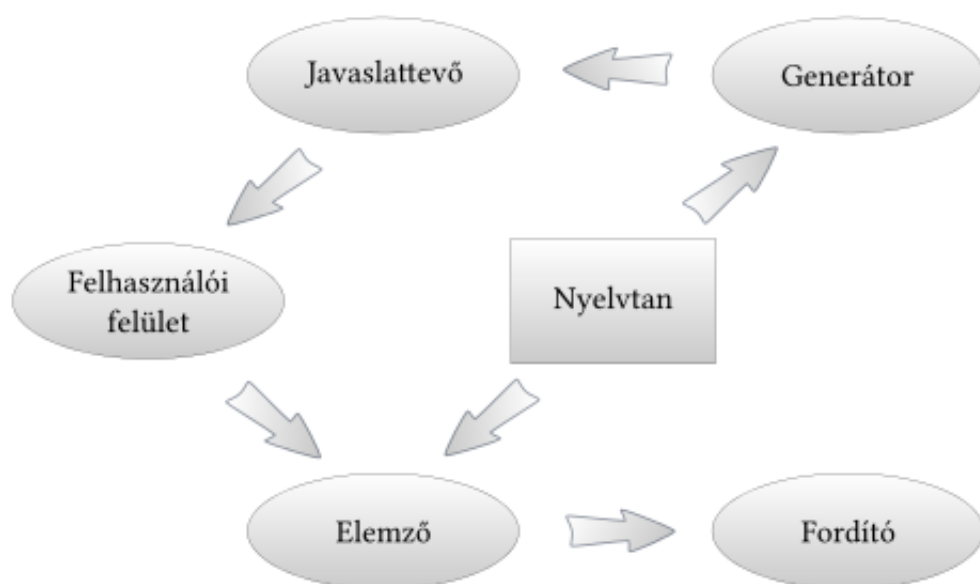
Első használat előtt érdemes elolvasni a [prediktív szövegbevitellel kapcsolatos tájékoztatót](#).

Kérdés:

m



Kérem a választ!



Az eredmények összefoglalása

- Az ágens megközelítésmód alkalmazása az információbeszerzés területén
 - tulajdonságaik, alapvető működési módjuk illeszkedik az elvárásokra
 - automatizált működéssel a feladatok széles körét képesek ellátni
 - automatizált kereső és letöltőrendszerek
 - elosztott, skálázható, rugalmas elemzőrendszerek
 - kommunikációs és kooperációs képességeik segítik
 - interfészek kialakítását belső és külső rendszerek felé is
 - a felhasználói kapcsolattartást
- XML technológiák alkalmazása az elemzés és az adatintegráció terén
 - segítségével leírhatók és megvalósíthatók a modellalapú elemzések
 - képes a különféle tartalmak és elemzéseik egységes tárolására
- Kontrollált természetes nyelvek támogatják az egyszerű felhasználást
 - a komplexitást az emberekhez közel álló módon kezelik
 - információlekérdezés mellett a tudásbevitel támogatására is felhasználhatók

Az eredmények gyakorlati alkalmazása

- OTKA F030 763 (témavezető)
 - ágensek lehetséges szerepkörei, alkalmazásuk vizsgálata
 - DocMan webrendszer: dokumentumtár + statisztikai elemzők
- IKF: Információ és Tudástárház projekt
 - ágensalapú tartalombeszerzés (webes letöltőrendszer)
 - XML-alapú tartalomreprezentáció és elosztott elemzőrendszer
- Főnévlista webes felület
 - kontrollált természetes nyelvű beviteli felület és SQL fordító
- CLABS: kontrollált nyelvű szövegkivonatok készítése
 - humán szakértők egyszerű integrálása
 - természetes nyelvű szövegekből Prolog tudásbázis építése
- R5-COP: természetes nyelvű, modellalapú robotinterfészek
 - robotok képességeit leíró modellekből felhasználói interfész generálása

Hogyan készült a háttérkép?

A **DHmine** rendszeremben feldolgoztam a dolgozatom szövegét az RStudio interaktív webes elemzőben az alábbi programmal:

```
library(tm)
library(SnowballC)
library(wordcloud)
corpus <- Corpus(DirSource("~/R/doktori", encoding="UTF-8"))
corpus <- tm_map(corpus, stripWhitespace)
corpus <- tm_map(corpus, tolower)
corpus <- tm_map(corpus, removeWords, stopwords("english"))
corpus <- tm_map(corpus, removeWords, stopwords("hungarian"))
corpus <- tm_map(corpus, stemDocument, "hungarian")
corpus <- tm_map(corpus, PlainTextDocument)
wordcloud(corpus, scale=c(4,.7), max.words=350, random.order=TRUE,
  rot.per=0.35, use.r.layout=FALSE, colors=brewer.pal(8, "Dark2"))
```

A szerző elérhetőségei

Mészáros Tamás

BME Méréstechnika és Információs Rendszerek Tanszék

Web: <http://www.mit.bme.hu/~meszaros/>

Email: meszaros@mit.bme.hu



Elérhetőségek

Iroda: 1117 Budapest, Magyar tudósok krt. 2. I. ép. IE437

Tel.: +36 1 463-2899

Fax: +36 1 463-4112

Email: meszaros (*) mit * bme * hu

Személyes honlap: <http://home.mit.bme.hu/~meszaros/>

Bemutakozás

Oktatóként, kutatóként, és nem utolsó sorban alkotó mérnökként dolgozom a tanszéken.

Oktatási munkám az informatikus képzésekre koncentrálok, az operációs rendszerek és az intelligens rendszerek szakirányok oktatásában veszek részt.

Kutatási területeim: tudásalapú rendszerek, szövegelemzés és -bányászat, kontrollált természetes nyelvű felhasználói felületek, valamint mindezek alkalmazásai az információszolgáltatásban és beszerzésben.

Mérnökként a fentiekén kívül webes és XML technológiákkal, valamint Linux rendszerekkel foglalkozom.

Aktuális: Digitális bölcsészet konferenciát szervezek ősszel az MTA-n.



Europass CV



Önálló labor témáim összefoglalása (2015)