JUL U HERAKLIONU

letnja škola o pretraživanju informacija

Anđelka Zečević, MATF

andjelkaz@matf.bg.ac.rs

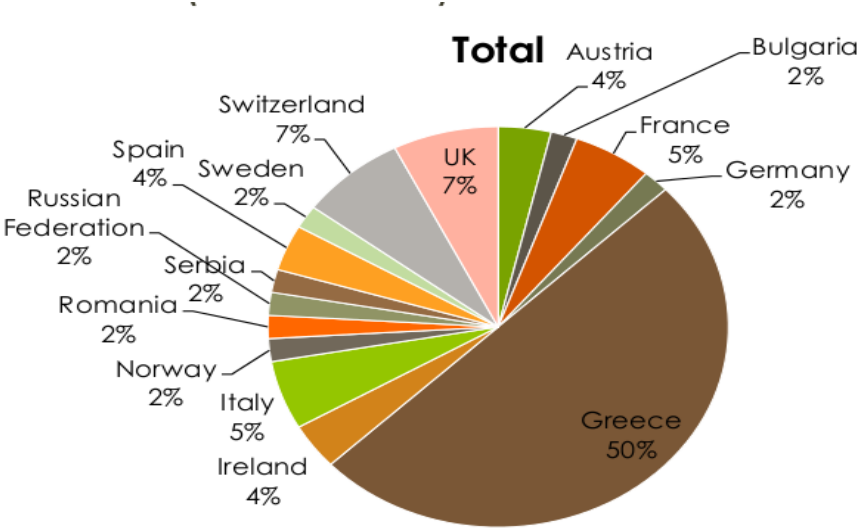
Letnja škola o pretraživanju informacija

- 3rd MUMIA Training School

  http://www.mumia-network.eu/index.php/training-school-2014
- od 21. do 25. jula 2014. godine
- Heraklion, Krit, Grčka
- FORTH:
  - The Foundation for Research and Technology-Hellas
  - Institut za računarstvo i informatiku (ICS)
  - Institut za primenjenu matematiku (IACM)

Letnja škola o pretraživanju informacija

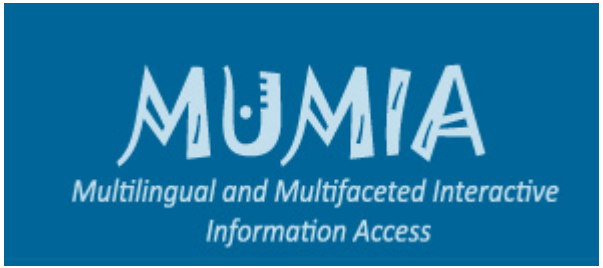
- 38 polaznika
- 16 polaznika volontera iz FORTHa



Letnja škola o pretraživanju informacija

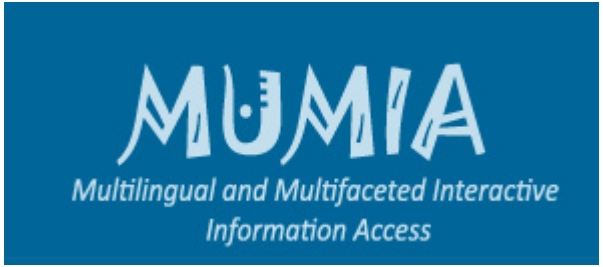
- **Keith van Rijsbergen** (University of Glasgow)
- **Kalervo Järvelin** (University of Tampere)
- **Berkant Barla Cambazoglu** (Yahoo Research)
- **Martin Braschler** (Zurich University of Applied Sciences)
- **Sébastien Ferré** (University of Rennes 1, IRISA)
- **Preben Hansen** (Stockholm University, SICS)
- **Gabriella Pasi** (Milano Bicocca)
- **Kostas Stefanidis** (FORTH-ICS)
- **Mihai Lupu** (TU-WIEN)
- **Michail Salampasis** (ATEI of Thessaloniki)
- **Yannis Tzitzikas** (University of Crete, FORTH-ICS)
- **George Paltoglou** (University of Wolverhampton)
- **Stefanos Vrochidis** (Informatics and Telematics Institute, Thessaloniki)

MUMIA COST akcija



- IC1002: Multilingual and Multifaceted Interactive Information Access
- datum početka: 30. novembar 2010.
- datum završetka: 29. novembar 2014.
- 32 zemlje članice (28+4)
- koordinator: Mike Salampasis, Technological Educational Institute of Thessaloniki, Greece
- http://www.mumia-network.eu/

MUMIA COST akcija



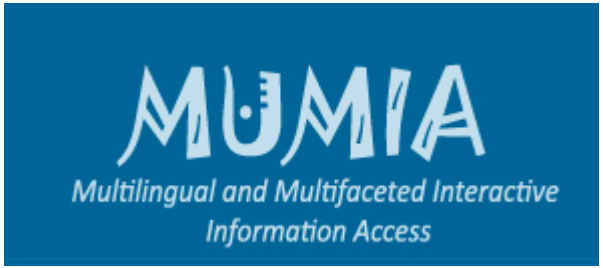
Motivacija:

Bogatstvo i kompleksnost informacija i njihove interne povezanosti prevazilaze mogućnosti postojećih pretraživačkih sistema

Cilj:

Stimulisati i koordinisati saradnju između istraživačkih centara u Evropi iz oblasti višejezičnog i višefasetnog interaktivnog pristupa informacijama čiji bi zaključci mogli da doprinesu razvoju sledeće generacije pretraživačkih sistema

MUMIA COST akcija



- 5 radnih grupa:
  - **WG1**: Integrating and Managing Language Resources
  - **WG2**: Processing Infrastructures for IR and MT
  - **WG3**: User Centred Aspects of MUMIA
  - **WG4**: Semantic Search and Faceted Search, Visualization
  - **WG5**: Distributed and Social Search

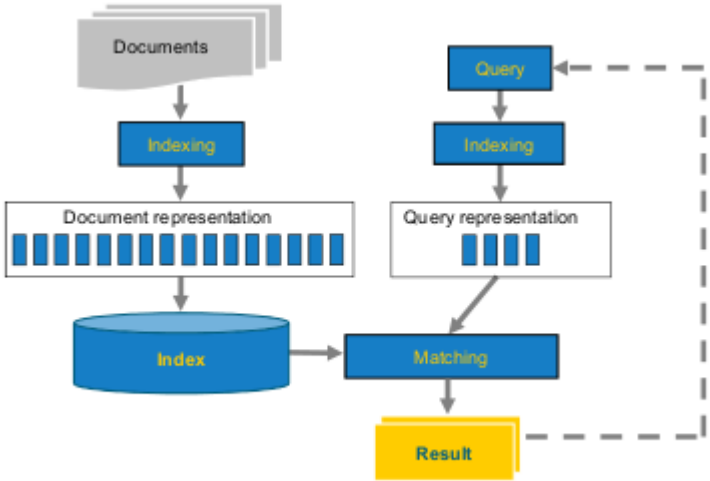
Uvod u IR, Keith van Raisbergen

... information is not a stuff contained in books as marbles might be contained in a bag - even though we sometimes speak

of it in that way. It is, rather, a relationship. The impact of a given message on an individual is relative to what he already knows, and, of course, the same message could convey different amounts of information to different receivers, depending on each one's internal model or map.

Uvod u IR, Keith van Raisbergen

- kreiranje kolekcije dokumenata
- postavljanje upita
- obrada kolekcije i upita i svođenje na formu koja je podesna za dalje korišćenje
- kreiranje indeksa tj. strukture u kojoj se efikasno čuvaju dokumenti
- upoređivanje upita i sadržaja indeksa
- generisanje liste relevantnih rezultata



Opinion retrieval in social media, Georgios Paltoglou

- Opinion retrieval:
  npr. *Šta ljudi misle o Bus-Plus-u?*
  pretraga sa ciljem da se pronađe nečije mišljenje
  opšti princip: “**Who** thinks/feels **how** about **what**?”
  fokus su primarno socijalni mediji: Twitter, blogovi, Facebook, ...
- U vezi je sa:
  analizom sentimenata ili istraživanjem sentimenata
- U odnosu na klasičan model razlikuje se pojam relevantosti:
  tema + mišljenje

Opinion retrieval in social media, Georgios Paltoglou

Nevolja sa socijalnim medijima (konkretno Twitter-om):

- neobičan jezik: skraćenice, :-), x-|, svega 140 karaktera
- sadržaj je efemeran
- autorstvo je važno
- bitan je vremenski tok

- prilikom obrade bitna je normalizacija tokena
  - prvođenje *out-of-vocabulary* tokena u normaliizovane forme
  - neki zadaci su laki: “helloooo” -> “hello”
  - za ostale se razvijaju tehnike mašinskog učenja (specijalan slučaj prevođenja) koje koriste morfosintaksičke sličnosti i kontekst

Opinion retrieval in social media, Georgios Paltoglou

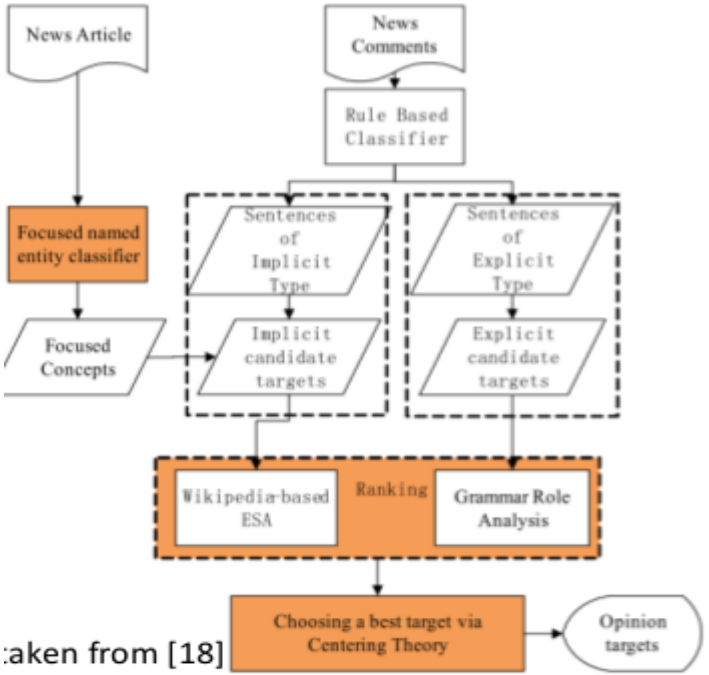
- **Who**: entitet (osoba, organizacija, neko ko može da *oseća, misli*)
- može da bude sam autor teksta (direktno) ili treće lice (indirektno)
- u svakom slučaju je zanimljiv zadatak:
  - pronaći odgovarajući entitet (Gate, IdentiFinder)
  - vezati ostatak testa za njega - anafora (OpenNLP)

Opinion retrieval in social media, Georgios Paltoglou

- **What:** objekat praćen atributima
- obično se vrši analiza imenica

  koje se grupišu pod istim entitetima

  (tzv. explicit semantic analysis)

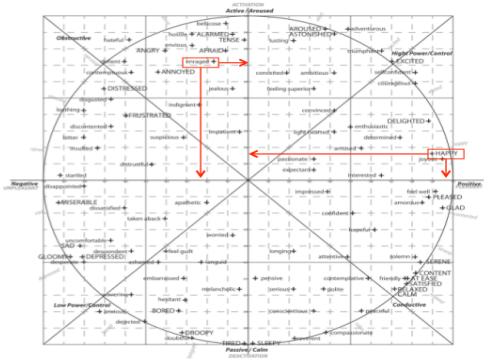


:aken from [18]

Opinion retrieval in social media, Georgios Paltoglou

- **How:** ciljno mesto
- moguće su ocene:
  - pozitivno/negativno/neutralno
  - na skali od -n do n
  - prepoznavanje univerzalne emocije: ljubav, strah, tuga…
  - detektovanje emocije u Raselovom krugu
- kako se emocije uče:
  - tehnike mašinskog učenja

    (SVM, Naïve Bayes, Entropija..)



Opinion retrieval in social media, Georgios Paltoglou

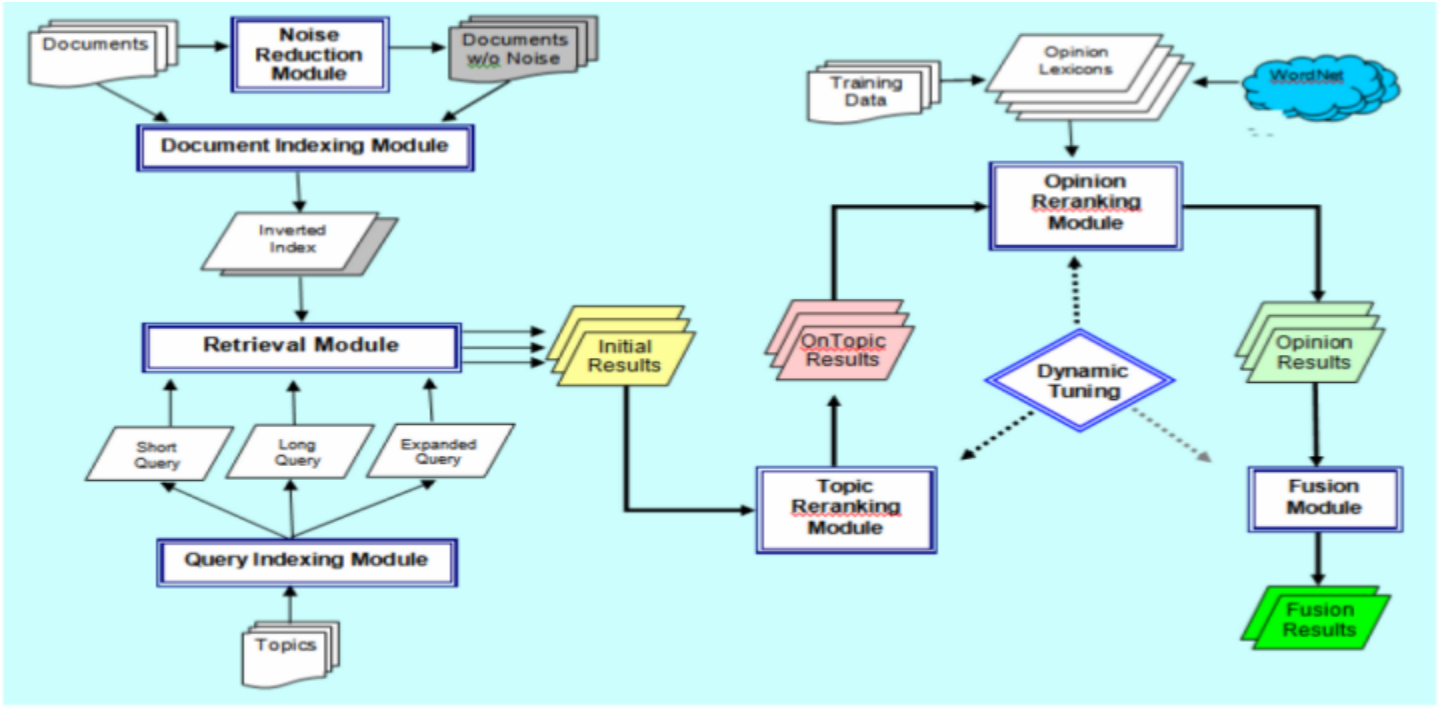
- prostor za poboljšanja:
  - bag-of-words model: nije idealan jer se gube veze između reči
    - n-grami (n dovoljno veliko)
  - skupovi za učenje:
    - trenutno se koriste se pregledi sa direktnim ocenama korisnika (npr. Amazon, TREC kolekcije)
  - razvijanje afektivnih leksikona
    - SentiWordNet, Affective norms for English words

Opinion retrieval in social media, Georgios Paltoglou

- sve zajedno:
  - primenjuje se klasično pretraživanje po temi
  - lista dobijenih rezultata se ili filtrira ili se menjaju pozicije dokumenata u skladu sa traženim mišljenjem (mašinsko učenje ili na osnovu leksikona)
  - kombinacija ova dva daje konačan skor dokumentu

- proširivanje upita:
  - terminima koji su vremenski vezani za teme bliske upitu
  - korišćenjem leksikona

Opinion retrieval in social media, Georgios Paltoglou



IR Evaluation ++, Mihai Lupu

- definisati eksperimentalne procedure i mere za upoređivanje rezultata sistema i korisnikovih očekivanja
- barijere:
  - subjektivnost
  - nemogućnost manuelne provere u celosti
- mere efikasnosti: prostorne i vremenske zahteve pretrage
- **mere efektivnosti**: korisnost samog pretraživanja
  - **interne mere**: sa stanovišta skupa rezultata
  - eksterne mere: sa stanovišta korišćenja alata pretrage
- šta meriti: **tematsku relevantnost**, šarolikost skupa, kvalitet, pouzdanost, razumljivost informacija….

IR Evaluation ++, Mihai Lupu

Test kolekcije:

- TREC
  - Text REtrieval Conference
  - http://trec.nist.gov/
  - od 1992. godine u US
  - svake godine su aktuelne nove teme
  - npr. spanish retrieval, medical retrieval, blog retrieval….

IR Evaluation ++, Mihai Lupu

Test kolekcije:

- CLEF
  - Cross Language Evaluation Forum
  - http://www.clef-initiative.eu/
  - od 2000. godine u EU
  - svi resursi su višejezični
  - npr. cross language geographical retrieval

IR Evaluation ++, Mihai Lupu

Test kolekcije:

- NTCIR
  - NII Test Collection for IR systems
  - http://research.nii.ac.jp/ntcir/index-en.html
  - od 1997. na svakih 1.5 godina u JP
  - npr. medical NLP for clinical documents, temporal information access

IR Evaluation ++, Mihai Lupu

Često korišćene mere na nivou upita:

- preciznost P (en. precision)
- odziv R (en. recall)
- recipročne pozicije RR (en. reciprocal rank)
- prosečna preciznost AP (en. average precision)
  - P i R zavise od nivoa odsecanja
- kumulativni doprinos CG (en. cumulative gain)
  - da li su svi dokumenti isto relevantni?

$$\mathrm{CG_p} = \sum_{i=1}^{p} rel_i$$

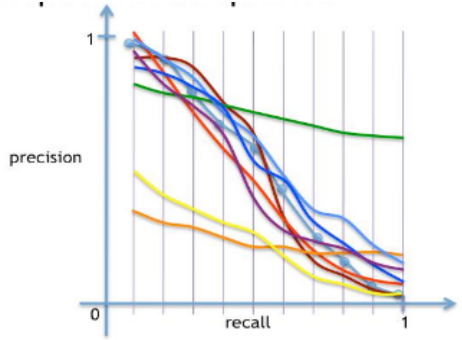
$$AP = \frac{\sum_{i=1}^{k} (P(i) \cdot rel(i))}{R}$$

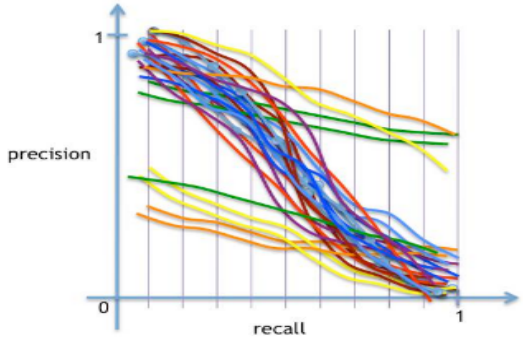
- R= # relevant documents
- i = rank
- k = # retrieved documents
- P(i) precision at rank *i*
- rel(i)=1 if document at rank *i* relevant, 0 otherwise

IR Evaluation ++, Mihai Lupu

Često korišćene mere na nivou sistema:

- P-R krive
- upoređivanje n-preciznosti (P@n)
- upoređivanje recipročnih pozicija
- MAP (en. mean average precision)





IR Evaluation ++, Mihai Lupu

Zaključak:

- postoji puno mera

  "there is a measure to make anyone a winner"

- upoređivanje samih mera i ispitivanje poželjnih karakteristika (ograničenost, monotonost, kompletnost, ...), korelacije sa drugim merama,...

Integrating IR Technologies for professional search, Mike Salampasis

- profesionalna pretraga:
  - ugovorom predviđen posao
  - pretraga sa profesionalnom motivacijom i ciljevima
- razlike u odnosu na veb pretragu:
  - fokus na domenskom znanju (specijalni NLP alati)
  - koriste se podaci sa kompleksnim shemama klasifikacije i velikom količinom metapodataka
  - dugotrajne sesije sa potrebom da se pauziraju i nastave
  - relevantnost je strožije definisana
  - proces dobijanja rezultata treba da bude jasan
  - rezultate je moguće reprodukovati

Integrating IR Technologies for professional search, Mike Salampasis

Pretraga patenata:

- broj patenata u US u 2000. godini: 380 000
- broj patenata u US u 2012. godini: 580 000
- zbog velikog broja predloga patenata vrši se automatska klasifikacija (pridruzivanje standardnih IPC kodova) i kasnije se manuelno proverava validnost od strane ispitivača
- veliki patent centri:
  - European Patent Office
  - United States and Patent Trademark office
  - Japanese File Index

Integrating IR Technologies for professional search, Mike Salampasis

- IPC - International Patent Classification
- svaki patent ima svoj kod (jedan ili više)

IPC	Code	Title
Section	B	Performing operations.Transporting.
Class	B64	Aircraft. Aviation. Cosmonautics.
Subclass	B64C	Aeroplanes. Helicopters.
Main Group	B64C 25/00	Alighting gear
Subgroup	B64C 25/02	Undercarriages

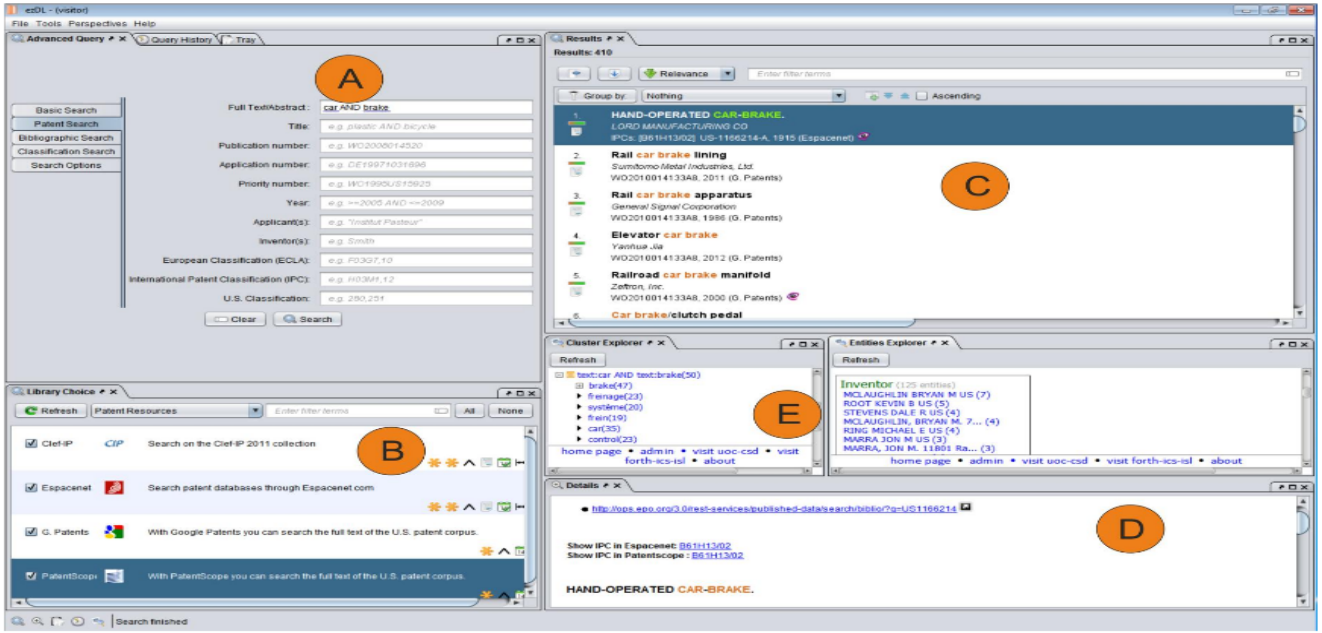
Integrating IR Technologies for professional search, Mike Salampasis

Specijalni zahtevi pretrage:

- pretraga sadržaja, meta podataka, slika
- odziv je jako bitan - jedan pronalazak može da odbaci patent
- sesije su dugotrajne
- ceo proces i rezultati se mogu, po potrebi, predstaviti na sudu

Integrating IR Technologies for professional search, Mike Salampasis

PerFedPat: http://www.perfedpat.eu/



Integrating IR Technologies for professional search, Mike Salampasis

Zaključak:

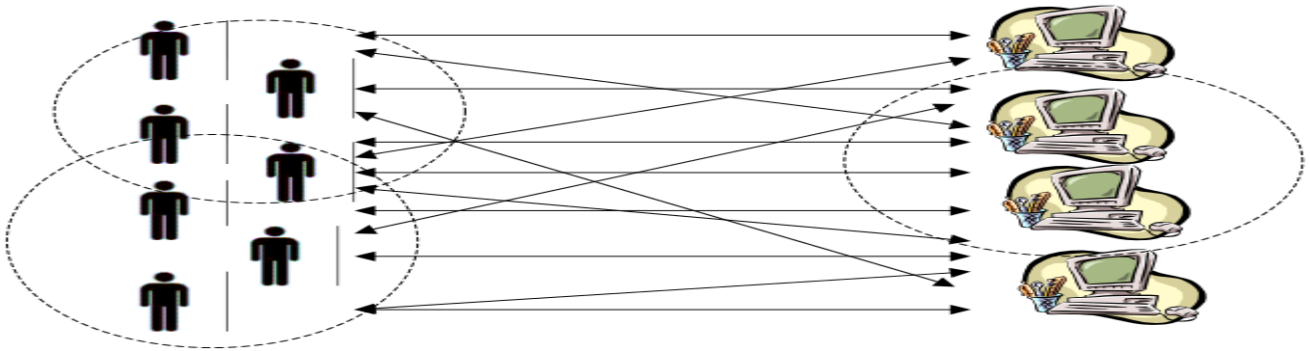
- sistemi za profesionalnu pretragu su kompleksni
- integrišu veliki broj alata i resursa
- arhitekture su višeslojne
- ostavljen je prostor za definisanje standarda i metodologija

Collaborative information seeking, Preben Hansen

- *CIS is an information access activity related to a specific problem-solving activity that, implicitly or explicitly,involves human beings interacting with (an) other human(s) directly and/or through texts (documents, notes, figures, etc.) or other media as information sources in a work-task-related information seeking process either in a specific workplace setting or in a more open community.* (Hansen, 2000)

- Npr. pretraga patenata, planiranje putovanja, istraživanje dokumentacije

Collaborative information seeking, Preben Hansen



akcenat je na timskom radu:

- definisanje zajedničkih informacionih potreba
- formulisanje zajedničkih upita
- pretraživanje informacija zajednički
- razgovor o procesu pretraživanja i propratnim aktivnostima
- dogovor u vezi razultata i odlučivanje kako ih upotrebiti

Collaborative information seeking, Preben Hansen

- neophodno je postojanje aplikacija i hardvera koji omogućava harmonizovanje toka:
  - sinhrona/asinhrona komunikacija
  - lokalizovana/na daljinu/lokalizovana+na daljinu komunikacija

- akcenat je na interaktivnim sistemima:

  omogućiti ugodan pristup informacijama
- akcenat je na traženju:

  kontekst ponašanja, kreiranja i korišćenja novih informacija

Collaborative information seeking, Preben Hansen

Interesantna pitanja:

- kako organizovati tok informacija: identifikovati dokumente koji su od interesa nekoj grupi, podeliti informacije pravilno među članovima tima, izvršiti sumarizaciju prema interesima grupe

- kako izvršiti evaluaciju:
  - upitnici, elektronski dnevnici, monitoring… (visoka cena)
  - analiza log datoteka, pracenje istorije dokumenata i istorije linkova izmedju njih

Collaborative information seeking, Preben Hansen

Slične discipline:

- **socijalna pretraga** (en. social search): pretraga sadržaja generisanog na socijalnim mrežama
- **zajednička pretraga:** pretraga grupe ljudi korošćenjem klasičnih sistema pretrage
- **filtriranje vođeno saradnjom** (en. collaborative filtering): predviđanje interesa korisnika na osnovu prikupljanja informacija o korisnicima sa kojima sarađuje

Contextual Search and Contextual Factors Aggregation, Gabriella Pasi

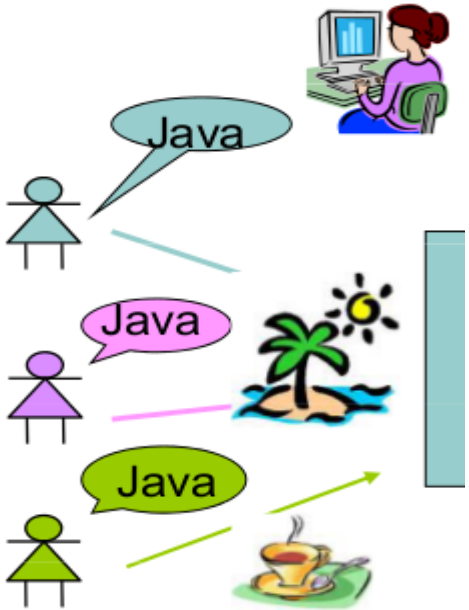
- pretraživačke mašine rade po principu "*one size fits all*"
- kontekst pretrage se ignosriše
- šta je kontekst:
  - korisnik:

    godine, pol, ekspertiza, socijalni kontekst, ...
  - zadatak:

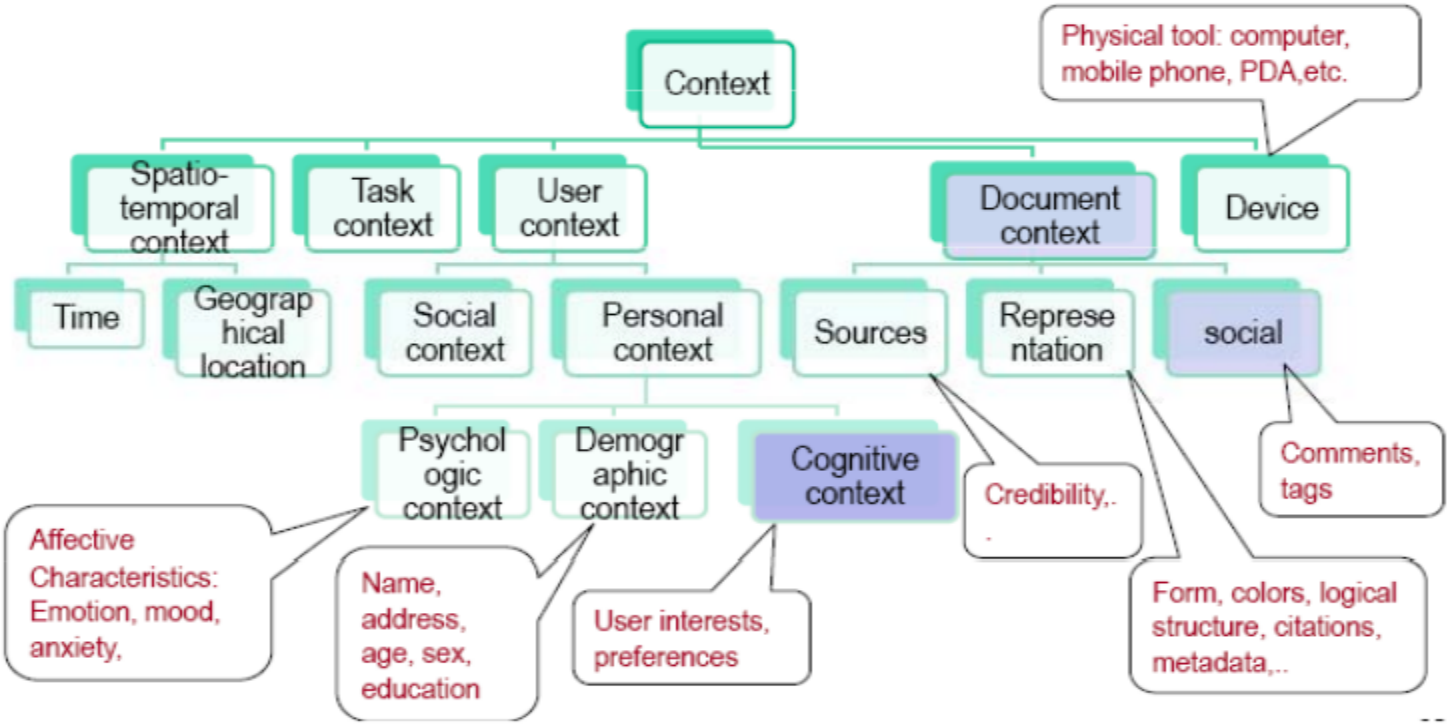
    planiranje puta, istraživanje, proizvod, ...
  - fiziči parametri:

    uređaj za pretragu, vreme, mesto…
  - dokument:

    autori, žanr, tagovi….



Contextual Search and Contextual Factors Aggregation, Gabriella Pasi



Contextual Search and Contextual Factors Aggregation, Gabriella Pasi

Načini uključivanja konteksta u pretragu:

- modelovanje korisnika:

  npr. informacije korisnikovog profila tipa godina, pola ili jezika

- modelovanje kognitivnih preferencija korisnika

  npr. lista ključnih reči, informacije o Desktop-u, istorija pretrage (posećene strane + raniji upiti + vreme provedeno na stranama), socijalne aktivnosti tipa postova, tagova, …

Contextual Search and Contextual Factors Aggregation, Gabriella Pasi

- način prikupljanja podataka:
  - eksplicitno: upitnici, profili, ...
  - implicitno: istorija pretrage, socijalne informacije, markeri …
- trajnost podataka:
  - kratkoročne potrebe korisnika (vezane za sesije)
  - dugoročne potrebe korisnika (uvek)
  - interesantno pitanje: kako pratiti sesije? kako detektovati prelaz korisnika sa jedne na drugu sesiju?
    - vremenska ograničenja
    - sličnost upita
    - sličnost rezultata pretrage
    - ....

Contextual Search and Contextual Factors Aggregation, Gabriella Pasi

podesne strukture za predstavljanje konteksta:

- vektori reči sa težinskim koeficijentima
- izvedene taksonomije na osnovu čestih termina u ličnim dokumentima ili direktnim izborom korisnika

  npr. Open Directory Project http://www.dmoz.org/
- kreiranje personalnih ontologija ekstrakcijom ključnih imenovanih entita iz korisnikovih dokumenata i uklapanjem u neku od postojećih ontologija

  npr. Yago ontologija (Wikipedia+WordNet+GeoNames)

  http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago//

Contextual Search and Contextual Factors Aggregation, Gabriella Pasi

Kontekst u pretrazi:

- na nivou formiranje upita: proširivanje upita, promena težinskih koeficijenata na osnovu profila
- na nivou generisanja liste rezultata: klasifikovanje rezultata, preraspoređivanje dokumenata u listi
  - pitanje dominantne skale relevantnosti
  - isti upit, isti korisnik, ista kolekcija i moguća nova lista rezultata ukoliko korisnik promeni dimenziju
  - tzv. pitanje agregacije (objedinjavanja) rezultata npr. linearna kombinacija, fazi operatori, … (problem sličan objedinjavanju rezultata u kolaborativnoj pretrazi)

Multilingual IR and Cross-Language Retrieval, Martin Braschler

- Monojezično pretraživanje: A -> A
- Dvojezično pretraživanje: A -> B
- Višejezično pretraživanje:
  - A -> A, B, ...
  - AB -> A, B, AB, AC, ...

- Međjujezično pretraživanje: barem dvojezično pretraživanje

- potreba za prevođenjem je očigledna

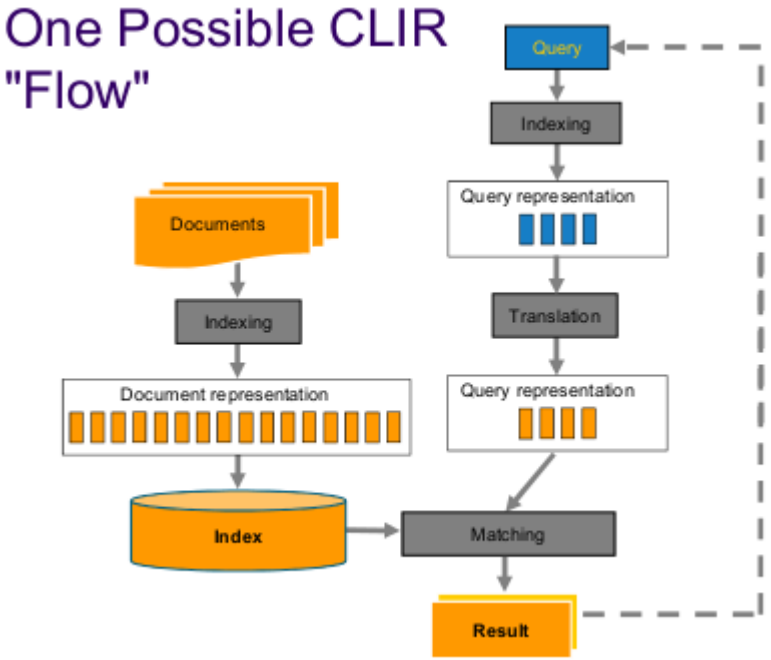
Multilingual IR and Cross-Language Retrieval, Martin Braschler

Mogući modeli:

- prevođenje upita
- prevođenje dokumenata
- prevođenje i dokumenata i upita
- bez prevođenja

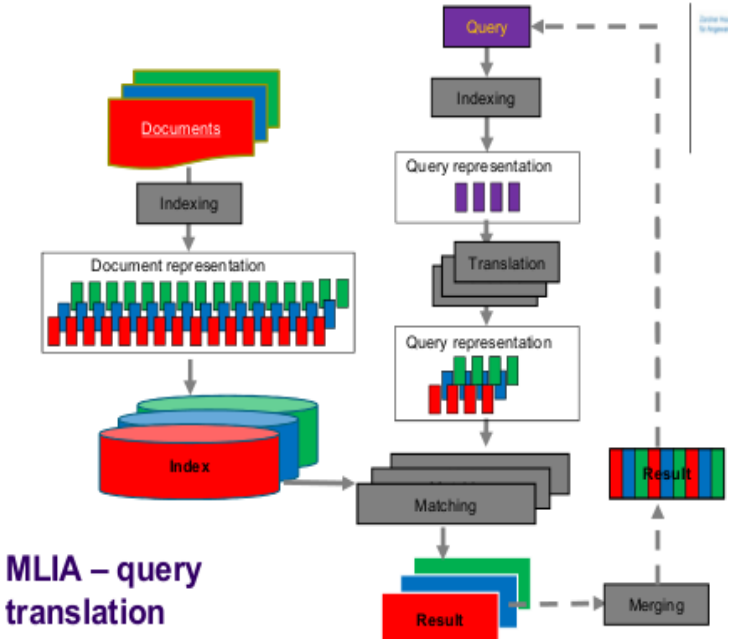
Multilingual IR and Cross-Language Retrieval, Martin Braschler

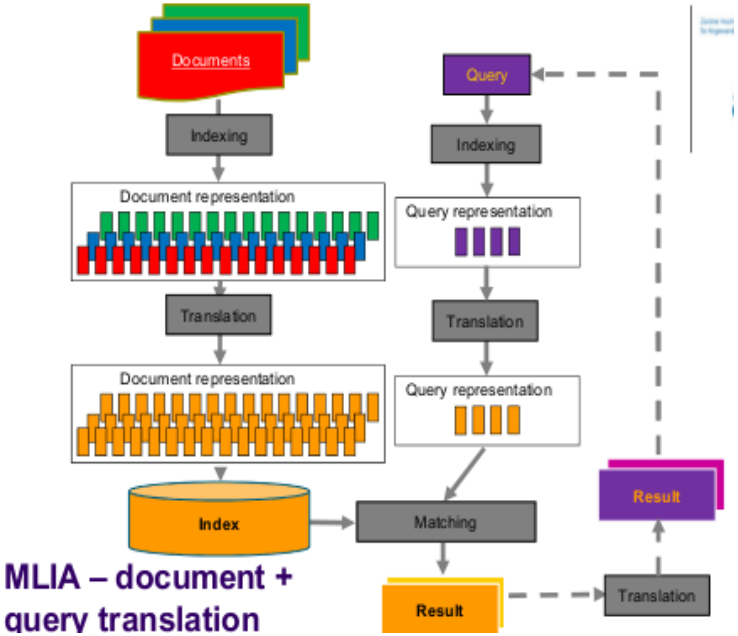
Dvojezični model:



Multilingual IR and Cross-Language Retrieval, Martin Braschler

Višejezični modeli:





Multilingual IR and Cross-Language Retrieval, Martin Braschler

Realnost:

- 6 800 jezika
  - Evropa: 230
  - Azija: 2 197
  - Afrika: 2 092
- 600 njih postoji u pisanoj formi

Izvori informacija:

- www.ethnologue.com
- www.omniglot.com

Multilingual IR and Cross-Language Retrieval, Martin Braschler

Osnovni koraci:

- priprema kolekcije
  - unifikovanje kodnih shema i formata
  - eliminacija duplikata, “opštenamenskih sadržaja”, ….

- identifikacija jezika
  - neophodna zbog primene alata za obradu kako upita tako i dokumenata
  - uključuje: distribuciju slova, statistike trigrama, …
  - predlog: skupiti što je moguće veći broj prediktora jezika i primeniti princip većinskog glasanja

Multilingual IR and Cross-Language Retrieval, Martin Braschler

- tokenizacija
  - izdvajanje sekvenci podesnih za indeksiranje
  - da li su IBM360, H&P, data-base (systems) tokeni?
  - pitanje složenica (npr. za nemački ili danski jezik)

    Bundesbankpräsident, Computersicherheit….
  - pitanje segmentacije istočnih jezika 我不是中国人
    - jezički nezavisan pristup: n-grami

Multilingual IR and Cross-Language Retrieval, Martin Braschler

- eliminacija funkcijskih (stop) reči:
  - članovi, predlozi, veznici, …
  - može da bude osetljivo: who ili WHO (World Health Organization)
  - šta je sa upitom “to be or not to be”?

- opšti zaključci:
  - nema jasnih pravila, moguće su greške
  - smanjuje se veličina indeksa

Multilingual IR and Cross-Language Retrieval, Martin Braschler

Normalizacija:

- glavni cilj je kontrola vokabulara
- dogovor oko:
  - dijakritika
  - leksičkih varijanti ("analyzing", "analysis")
  - pravopisnih pravila (“color”, “colour”)

Multilingual IR and Cross-Language Retrieval, Martin Braschler

Stemming:

- fleksija - kako interpretirati broj, rod, ...
- derivacija - sufiksi, prefiksi, ...
- mogu se koristiti gramatička pravila:

  npr. IF ((" *-ing ") & (length>3)) → remove –ing
- mogu se koristiti naučeni algoritmi
  - npr. ako je dostupna POS anotacija
- neki opšti zaključci:
  - bolje je svoditi reči na stem, nego ne
  - “light” varijante npr. za imenice i prideve

Multilingual IR and Cross-Language Retrieval, Martin Braschler

Stemming:

- CLEF kolekcija

  +4% with the English language
  +4% Dutch
  +7% Spanish
  +9% French
  +15% Italian
  +19% German
  +29% Swedish
  +34% Bulgarian
  +40% Finnish
  +44% Czech

- alternativa: n-grami i indeksiranje preko n-grama

Multilingual IR and Cross-Language Retrieval, Martin Braschler

- “fini” detalji pre indeksiranja:
  - prepoznavanje imenovanih entiteta
  - korišćenje informacija korisničkih profila…

- ulaz u indeks je vektor

Doc. ids	Text
1	Pease porridge hot, pease porridge cold
2	Pease porridge in the pot
3	Nine days old
4	Some like it hot, some like it cold
5	Some like it in the pot
6	Nine days old

Feature #	Feature	df, document ids, positions
1	cold	2; (1, 6), (4, 8)
2	days	2; (3, 2), (6, 2)
3	hot	2; (1, 3), (4, 4)
4	in	2; (2, 3), (5, 4)
5	it	2; (4, 3,7), (5, 3)
6	like	2; (4, 2,6), (5, 2)
7	nine	2; (3, 1), (6, 1)
8	old	2; (3, 3), (6, 3)
9	pease	2; (1, 1,4), (2, 1)
10	porridge	2; (1, 2,5), (2, 2)
11	pot	2; (2, 5), (5, 6)
12	some	2; (4, 1,5), (5, 1)
13	the	2; (2, 4), (5, 5)

Position
Document id
Document frequency

Multilingual IR and Cross-Language Retrieval, Martin Braschler

problemi prevođenja:

- prevođenje upita: obično su kratki i kontekst je neuhvatljiv

  “post” :

  Mail? Position? Other? Post office? An entry in a blog?

  post-mortem examination? Post Emily? Washington Post?

Multilingual IR and Cross-Language Retrieval, Martin Braschler

problemi prevođenja:

- prevođenje dokumenata:
  - korišćenje rečnika
    - neophodne leme
    - out-of-vocabulary problem
    - višeznačnosti
  - statističko modelovanje
    - neophodni paralelni/uporedni korpusi, veb?
    - razni algoritmi učenja
    - MOSES

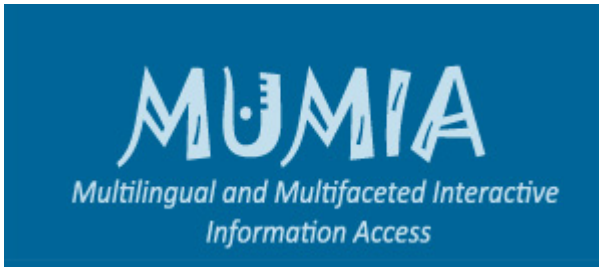
Multilingual IR and Cross-Language Retrieval, Martin Braschler

- izazovi prevođenja
  - prevođenje fraza

    npr. Final Four Results (EN) -> final quatre résultat (FR)

  - prepoznavanje vlastitih imena, imena geo lokacija, …

Hvala na pažnji!



http://www.mumia-network.eu/