Večjezičnost spletnih informacijskih virov: medjezično iskanje 1

Razlogi za razvoj MI,

definicije in pregled postopkov MI.

Motivacija za razvoj MI

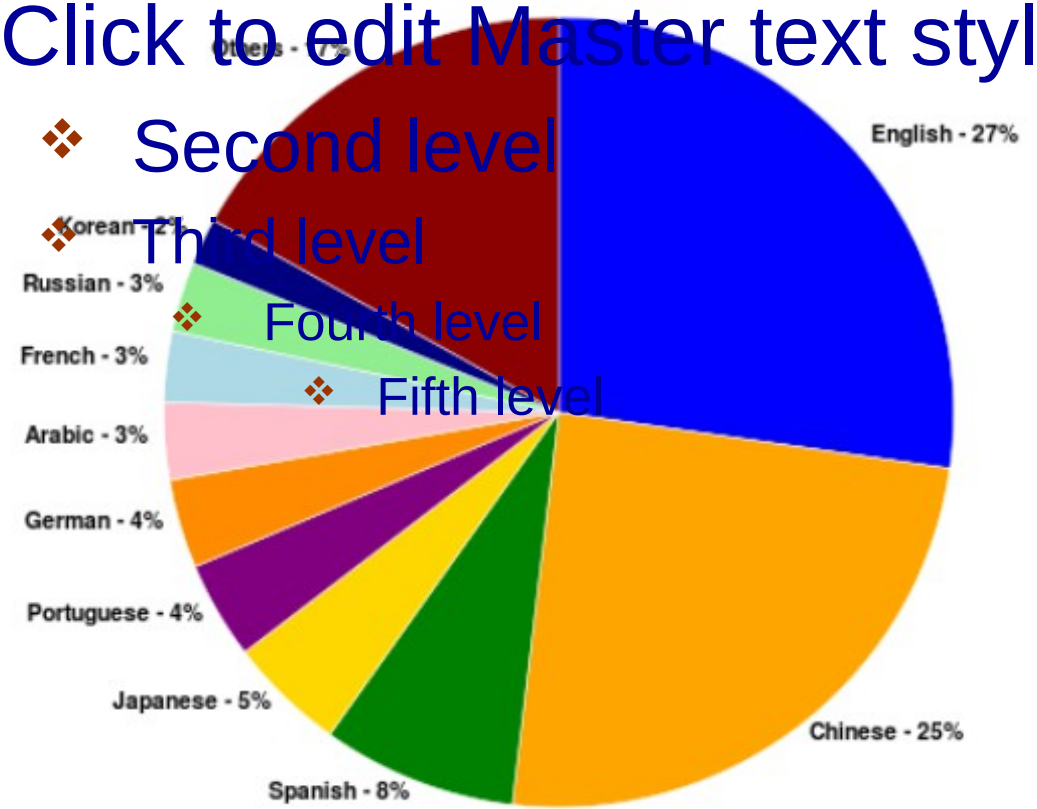
- Vzporedno z razvojem omrežnega (spletnega) publiciranja je potekal razvoj iskalnikov spletnih dokumentov.
- Na začetku skoraj 100% dokumentov v angleščini – vsa metodologija spletnih iskalnikov prilagojena angleščini.
- Danes (konec 2011) porazdelitev jezikov spletnih dokumentov bistveno drugačna:
  - 57% angleščina,
  - 29% evropski, neangleški jeziki,
  - 14% ostalo.

Jeziki Internetnih uporabnikov

31. maj 2011

Rank	Language	Internet users	
1	**English**	565,004,000	27%
2	**Chinese**	509,965,000	25%
3	**Spanish**	164,969,000	8%
4	**Japanese**	99,182,000	5%
5	**Portuguese**	82,587,000	4%
6	**German**	75,423,000	4%
7	**Arabic**	65,365,000	3%
8	**French**	59,779,000	3%
9	**Russian**	59,700,000	3%
10	**Korean**	39,440,000	2%
	Others	350,557,000	17%

- Click to edit Master text styles
  - Second level
    - Third level
      - Fourth level
        - Fifth level

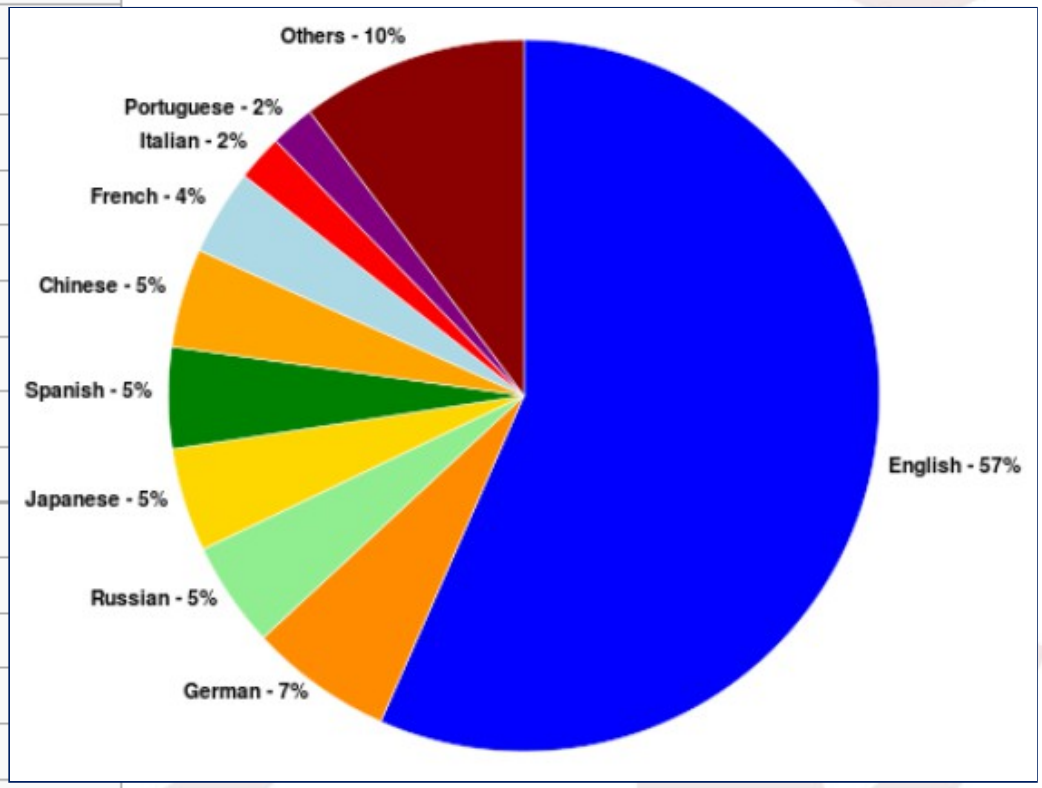


Jeziki Internetnih dokumentov

- Click to edit Master text styles
  - Second level
    - Third level
      - Fourth level
        - Fifth level

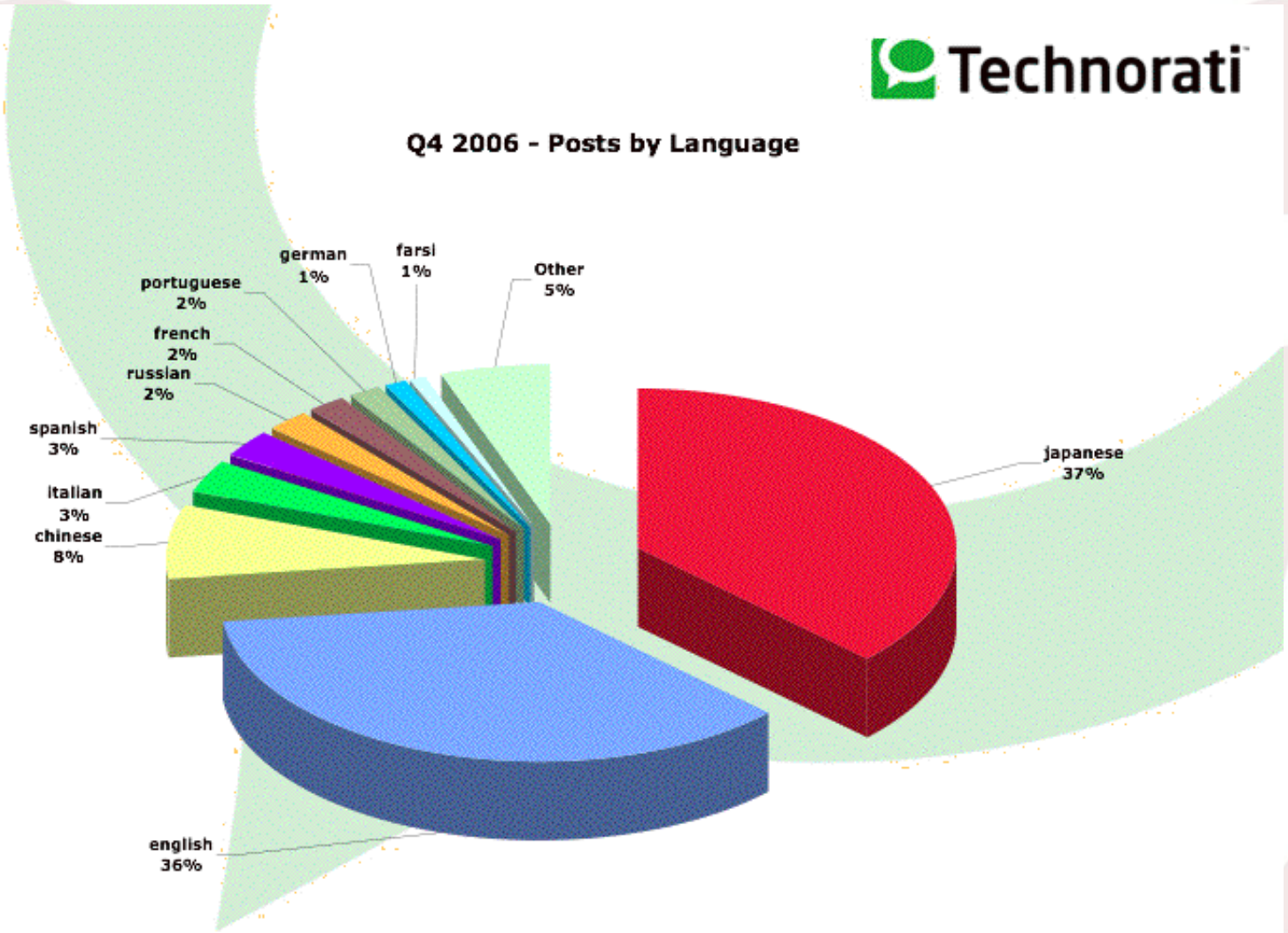
31 December 2011

Language	Percentage
English	56.6%
German	6.5%
Russian	4.8%
Japanese	4.7%
Spanish	4.6%
Chinese	4.5%
French	3.9%
Italian	2.1%
Portuguese	2.0%
Polish	1.4%
Arabic	1.3%
Dutch	1.1%
Turkish	1.1%
Swedish	0.7%
Persian	0.7%
Czech	0.5%
Romanian	0.4%
Korean	0.3%
Greek	0.3%
Hungarian	0.3%
Thai	0.3%
Vietnamese	0.3%
Danish	0.3%
Indonesian	0.3%
Finnish	0.2%
Norwegian	0.2%
Bulgarian	0.2%
Slovak	0.2%
Hebrew	0.1%
Croatian	0.1%
Lithuanian	0.1%
Serbian	0.1%
Catalan	0.1%
Slovenian	0.1%
Ukrainian	0.1%
Norwegian Bokmål	0.1%



Motivacija za razvoj MI

Porazdelitev jezikov v spletnih blogih (technorati.com, april 2007).



Motivacija za razvoj MI

- Taka porazdelitev se neposredno odraža v zbirkah velikih iskalnikov.
- Gradnja zbirk in iskalni algoritmi spletnih iskalnikov so še vedno do neke mere prilagojeni angleščini.
- Sofisticiranost jezikovno-odvisnih metod se zmanjšuje, ker ne morejo biti enako dobre za vse spletne jezike.

Motivacija za razvoj MI

Iskanje z iskalnimi zahtevami v naravnem jeziku je:

- Primerjanje besed ali besednih zvez iz iskalne zahteve z besedami ali besednimi zvezami v dokumentih.
- Iskanje ne more dati rezultatov, če sta iskalna zahteva in dokument v različnih jezikih.

Motivacija za razvoj MI

Kaj pa, če so dokumenti v zbirki v različnih jezikih?

- Iskalec mora sestaviti ločene iskalne zahteve v jezikih dokumentov.
- Težave:
  - iskalec se tekoče izraža le v enem ali dveh jezikih, ostali dokumenti nepoiskani,
  - neizenačena kvaliteta rezultatov zaradi različnega znanja jezikov pri istem iskalcu,
  - multiplikati prevodov istega dokumenta,
  - velik iskalni napor,
  - ...

Definicije medjezičnega iskanja

- Problematiko večjezičnega okolja dokumentov poskušamo reševati s sistemi, ki poskušajo izvajati avtomatsko in enakovredno iskanje neglede na jezike dokumentov.
- Imenujemo jih **medjezični iskalniki**.
- V strokovni literaturi se pojavljajo različni izrazi:
  - cross-language IR,
  - cross-lingual IR,
  - multilingual IR,
  - translingual IR...
- Razmejitev njihovih pomenov včasih ni jasna.

Definicije MI

Medjezično iskanje je iskanje, pri katerem je

- naravni jezik iskalne zahteve lahko različen od jezika ali jezikov dokumentov v zbirki.
- Iskalna zahteva je v jeziku *a* ali *b*,
- dokumenti v zbirki so v jezikih *a* in *b*,
- poiskani relevantni dokumenti so v jezikih *a* in *b*.
- MI je tudi iskanje po enojezični zbirki, če so lahko iskalne zahteve v različnih jezikih.

Definicije MI

- Enojezično ali istojezično iskanje (monolingual IR):
  - Iskalna zahteva in poiskani dokumenti v zbirki so v istem jeziku.
  - Medjezično iskanje z enim delom svoje definicije pokriva tudi enojezično iskanje.

Definicije MI

- Najširši izraz je večjezično iskanje (multilingual IR), ki vključuje
  - enojezično iskanje,
  - medjezično iskanje, in
  - iskanje dokumentov z deli v več jezikih.

- Tudi sisteme s pomnoženo enojezično funkcionalnostjo lahko imenujemo večjezični sistemi:
  - ločene iskalne zahteve v različnih jezikih in priklic dokumentov v teh jezikih.

Definicije MI

Ameriški zorni kot:

- medjezični sistemi so »sistemi, ki iskalcem nudijo dokumente, ki jih ti ne znajo prebrati«.

Splošno o MI: IR vs. MI

Področji IR in MI imata mnogo skupnega:

- načine organiziranja dokumentov v zbirkah,
- metode avtomatskega indeksiranja,
- interpretiranje iskalnih zahtev,
- računanje relevantnosti dokumentov.

Splošno o MI: IR vs. MI

Med področji IR in MI obstaja bistvena razlika:

- klasični IR ne potrebuje prevajanja.

- Vsak avtomatski postopek MI, ki ni vezan na ročno indeksiranje z večjezičnimi tezavri, vključuje neko vrsto računalniškega prevajanja.

Pregled metod MI

Ontologije

Ontologija:

- predstavitev mreže ali hierarhije konceptov in njihovih povezav

Tezaver:

- ontologija namenjena opisovanju in iskanju dokumentov v kontroliranih pogojih

Dvojezični leksikon:

- ontologija namenjena strojnemu prevajanju

Dvojezični slovar:

- ontologija namenjena človeškemu prevajanju