Aitor Maritxalar
Online proiektuak

Solr bilatzailea: esperientziak

Bilaketa bakarra erabiltzaileentzat: datu guztiak klik batean

Solr bilatzailea: esperientziak

- Aurkezpena
- Zer da Solr?
- Bilatzaile baten garapen esperientziak

Aurkezpena

Solr bilatzailea: esperientziak

- Bilatzaile aurreratuak
- Eduki (dokumental) digitalak
- Web garapena
- Funtzioak
  - Aholkularitza
  - Proiektuen kudeaketa
  - Garapena

Aitor Maritxalar

Espezializazioa

Zer da Solr?

Solr bilatzailea: esperientziak

Zer da Solr?

- CBS Interactive-k sortua, 2004an
- *Full-text* bilaketa **motorea**
- Datu-eskema malgua
- Indexatzailea + berreskuratzailea
- Apache lizentzia: kode irekia

Erabilpen eszenarioa

- Testu zentrikoa
- Bilaketa nagusi (ez eguneraketa)
- Dokumentua oinarri: ez erlazionala
- Datu-egitura neurrikoa

Ezaugarriak: orokorrak

- Azkarra
- Eskalablea
- Egongorra
- Milioika dokumentu
- Hedagarria (plugin-ak)
- Hizkuntza anitz

Ezaugarriak: erabilpen ikuspegitik

- Emaitzen errelebantzia parametrizagarria
- Faceting-a
- Paginazioa eta ordenazioa
- Emaitzen testuinguruak
- Bilaketa geoespaziala

Ezaugarriak: query-ak

- Operadore logikoak: AND, OR, NOT
- *Wildcard* bilaketak
- Termino bat baino gehiago bilaketan
- Interbaloak
- *Fuzzy* bilaketak
- Bilaketa geoespaziala

Solr dokumentua

- Oinarrian ia edozein formatu: PDF, Word, LibreOffice, TXT,…
- *Schema* definitu daiteke.
- *Dokumentua*k guk nahi dugun egitura eduki dezake
- Ezaugarri-balio egitura da.
- Dokumentua/*schema* ongi definitzen dago gakoa

Solr: alternatibak

Motoreak

- Amazon ElasticSearch
- Sphinx
- OpenSearchServer
- Lucee

Softwarea

- vuFind
- samvera
- blackLight
- Fulcrum

Noiz erabili Solr

- Edukia edo datu-basea “berezia” denean
- Eduki hetereogeneoa daukagunean
- Erabiltzaile-esperientzia bereziki landu nahi denean
- Bilaketaren eta emaitzen kontrol osoa eduki nahi denean

Bilatzaile baten garapen esperientziak

Solr bilatzailea: esperientziak

Datu-basean oinarritutako bilaketa

- Esfortzu handia
- Kostu altua
- Abiadura arazoak
- Eduki heterogeneoak elkarrekin erakusteko arazoak

Eta emaitzak, gehienetan…

FullText bilatzailea behar dut!!!

Indexazioaren eskema tipikoa

Bilatzailea definitzen: lehen lanak

OHAR GARRANTZITSUA!!!

- Erabiltzaileak ez du gure egitura ezagutzen!!!
- Erabiltzaileak ez du jakiten zehazki zer bilatzen duen (kasu askotan).

Helburua: zertarako da bilatzailea?

- Megabilatzailea ala bilatzaile gidatua?
- Estrategia: bilatu eta iragazi vs. gidatua
- Espezializatua ala deskubrimendua?
- Salmentarako al da?
- Eduki konkretuak topatzeko ala zerrendak sortzeko?
- …

Helburuak

BDB

- Datu-base desberdinen emaitzak elkartu
- Edukien ikusgarritasuna

Euskaltzaindia.eus

- Edukien ikusgarritasuna
- Erabiltzailearen esperientzia hobetu
- Eduki oso heterogeneoak elkartu

Erabiltzaile profila definitu

- Ze erabiltzaile motari zuzenduko zaion bilatzailea.
- Web estatistikak aztertu!!!
- Askotan erabiltzaileek guk uste ez duguna bilatzen dute.
- Kontuz: ERABILTZAILEAK EZ DIRA DOKUMENTALISTAK!!!

Erabiltzaile profila

BDB

- Gaia ezagutzen du
- Kontzeptuak ezagutzen ditu

Euskaltzaindia.eus

- Erabiltzaile arrunta
- Profesionala
- Ikertzailea

Prezisioa vs estaldura

- **Prezisioa**: ematen diren emaitza guztiak zuzenak dira
- **Estaldura**: atera behar duten emaitza guztiak ateratzen dira.

Prezisioaren eta estalduraren arteko proportzioa topatzea da gakoa!!!

Prezisioa vs estaldura

BDB

- Prezisioa: ertaina
- Estaldura: ertaina

Euskaltzaindia.eus

- Prezisioa: bajua/ertaina
- Estaldura: altua

Baliabideen inbentarioa egin

- Indexatu beharreko baliabideen inbentarioa egin.
- Formatuak
- Datu-baseak
- Hizkuntzak
- Konpuruak (dokumentuak edota erregistro kopurua)

Inbentarioa

BDB

- Datu-baseak
- PDFak

Euskaltzaindia.eus

- Datu-base dokumentalak
- Datu-base espezializatuak
- PDF solteak
- Hiztegiak
- Liburuak
- Albisteak

Puntu honetan softwarea aukeratu behar da…

Solr aukeratu dugu. Segi aurrera!!!

Hemendik aurrera informatikarien esku zaudete...

Dokumentua (*schema*) diseinatu

- Bilatzailea diseinatzerako orduan urrats garrantzitsuenetakoa.
- Indexatzerako orduan kudeatuko den egitura amankomuna.
- Laburrean: SOLRera bidaltzeko dokumentu “birtual” bat sortzen da.
- Eremuak eta eremu-motak finkatu.
- Facet-ak
- GUZTIA INDEXATU ORDUAN!!! Ez da oso ideia ona…

Facet-ak

- Oso garrantzitsuak
- Iragazteko aukera ematen dute
- Dimentsioaren ideia ematen dute
- Gure egitura erakusten dute
- Erabiltzailea ohituta dago

Eskema

BDB

- Eremu guztiak biltzen duen dokumentua.
- Facet asko
  - Autoritateak
  - Deskriptorea
  - Zenbait eremu
  - Facet “fiktizioak”
- Bilaketarako eremu testual bereziak

Euskaltzaindia.eus

- Dokumentu fiktizioa sortu genuen:
  - Izenburua
  - Egiletza
  - Saila
  - …
- Facet: Bilaketarako thesaurus berezia
- Bilaketarako FullContent eremua

Esportazioak

- Laburrean: baliabideetatik JSON fitxategi erraldoiak sortu.
- Lehen definitu egitura izango du.
- Eduki testualen “garbiketa” asko zaindu.

Bilatzaile baten arrakastaren zati handiena indexatzailean sartzen den edukiaren aurreprozesamenduan oinarritzen da.

Esportazioak

BDB

- Azpiatal bakoitzeko JSON bat
- Eremu fiktizioak: urtea, mendea
- PDFen prozesamendua
- ExtraContent eta FullContent eremuak

Euskaltzaindia.eus

- Lan gehien eman duen atala
- 24 esportazio desberdin
- Baliabide bakoitzak trataera berezia
- Bilaketarako FullContent eremua

Lematizazioa / stemmerra

- Garrantzi berezia euskararen kasuan.
- Ez da beti komenigarria.
- Estaldura asko igotzen du, baina prezisioa jaistsi.
- Erabiltzaileak espero ez ditzakeen emaitzak: “Baiona” eta “begi” adibideak.
- Testu zaharrekin ez da komenigarria izaten.

Lematizazioa / stemmerra

BDB

- Stemmerra aplikatua da: *unspell*
- Hizkuntza guztietako informazioa

Euskaltzaindia.eus

- Bilatzaile nagusian ez da aplikatu
- Bi bilatzaile berezituetan bai

Laburbilduz

Solr bilatzailea: esperientziak

Laburbilduz

1. Bilatzailea zertarako eta norentzat den ondo definitu.
2. Tresna eta hornitzailea aukeratu.
3. Solr aukeratzen baduzue:
    1. Informatikariarekin batera dokumentuaren egitura eta facet-ak finkatu.
    2. Esportazio onak sortu.
4. Solr bilatzailea zuen webgunearekin lotu.

Mila esker!!!

Aitor Maritxalar
Online proiektuak

aitor@maritxalar.eus